

University of Bari Aldo Moro  
Department of Computer Science

# **LLM e allucinazioni: Tentativo di costruzione di un classificatore generale**



**Emanuele Fontana**

November 23, 2025



# Contents

<b>1 Metodologia</b>	<b>1</b>
1.1 Dataset: BeliefBank . . . . .	1
1.2 Raccolta Dati ed Estrazione delle Attivazioni . . . . .	1
1.3 Probing e Allineamento . . . . .	2
1.3.1 Linear Probing . . . . .	2
1.3.2 Allineamento Cross-Model . . . . .	2
<b>2 Risultati</b>	<b>3</b>
2.1 Analisi PCA delle Attivazioni . . . . .	3
2.2 Risultati Quantitativi . . . . .	3
2.2.1 Scenario 1: Qwen → Falcon . . . . .	3
2.2.2 Scenario 2: Falcon → Qwen . . . . .	4
2.3 Matrici di Confusione . . . . .	4
<b>Appendices</b>	<b>9</b>
<b>A Visualizzazioni PCA</b>	<b>11</b>
A.1 Qwen2.5-7B . . . . .	11
A.2 Falcon3-7B-Base . . . . .	15



# List of Figures

2.1	PCA delle Attivazioni Hidden Layer per Qwen2.5-7B (sinistra) e Falcon3-7B-Base (destra). I punti rossi indicano le allucinazioni. . . . .	4
2.2	Matrici di Confusione per Hidden Layers. Sinistra: Qwen (Teacher). Destra: Falcon su Qwen (Student). . . . .	6
A.1	PCA 2D - Qwen2.5-7B - Attention Layers . . . . .	12
A.2	PCA 2D - Qwen2.5-7B - MLP Layers . . . . .	13
A.3	PCA 2D - Qwen2.5-7B - Hidden Layers . . . . .	14
A.4	PCA 2D - Falcon3-7B-Base - Attention Layers . . . . .	16
A.5	PCA 2D - Falcon3-7B-Base - MLP Layers . . . . .	17
A.6	PCA 2D - Falcon3-7B-Base - Hidden Layers . . . . .	18



# List of Tables

2.1	Risultati per lo Scenario 1 (Qwen → Falcon)	5
2.2	Risultati per lo Scenario 2 (Falcon → Qwen)	5



# Chapter 1

## Metodologia

In questo capitolo viene descritta la metodologia impiegata per rilevare le allucinazioni nei Large Language Models (LLM) utilizzando le attivazioni interne. Vengono dettagliati il dataset utilizzato, il processo di raccolta dati e le tecniche di probing e allineamento applicate.

### 1.1 Dataset: BeliefBank

Il dataset principale utilizzato per questo studio è BeliefBank **TODO: citare**, un dataset strutturato di credenze progettato per valutare la coerenza e la veridicità dei modelli di IA. BeliefBank è costituito da due componenti principali: fatti e vincoli (constraints).

### 1.2 Raccolta Dati ed Estrazione delle Attivazioni

In questo studio sono stati impiegati due LLM:

- **Qwen2.5-7B**: Un potente modello open-weights.
- **Falcon3-7B-Base**: Un altro modello base allo stato dell'arte.

Per ogni fatto nel dataset, ogni modello è stato interrogato determinare la veridicità dell'affermazione. Durante questo processo di generazione, sono state catturate le attivazioni interne da tre tipi di layer:

- **Hidden States**: L'output dei blocchi transformer.
- **MLP Layers**: Gli output delle reti feed-forward all'interno dei blocchi transformer.
- **Attention Layers**: Gli output dei meccanismi di self-attention.

Le risposte generate sono state confrontate con le etichette ground truth di BeliefBank per determinare se il modello stesse allucinando (ovvero, affermando una falsità come verità o viceversa).

## 1.3 Probing e Allineamento

Per rilevare le allucinazioni, è stato impiegato un approccio di probing lineare.

### 1.3.1 Linear Probing

È stato addestrato un classificatore di Regressione Logistica (il “Probe”) sulle attivazioni di un modello “Teacher”. L’input per il probe è il vettore di attivazione di uno specifico layer, e l’output è un’etichetta binaria che indica se il modello sta allucinando. Il metodo utilizzato per rilevare le allucinazioni è una semplice substring: se la ground\\_truth è contenuta nella risposta generata allora l’istanza è etichettata come “corretta” (0), altrimenti come “allucinata” (1). Al modello è stato fornito il seguente prompt:

Answer the following question with just the essential information, without explanations.

### 1.3.2 Allineamento Cross-Model

Per indagare la trasferibilità del rilevamento delle allucinazioni tra modelli, è stata eseguita una fase di allineamento: un modello “Teacher” e l’altro “Student”. È stato addestrato un modello di Regressione Ridge per mappare le attivazioni del modello Student nello spazio delle attivazioni del modello Teacher.

La pipeline sperimentale procede come segue:

1. Addestramento di un Probe sulle attivazioni del Teacher.
2. Addestramento di un Aligner per mappare le attivazioni dello Student su quelle del Teacher.
3. Proiezione delle attivazioni dello Student utilizzando l’Aligner.
4. Test del Probe del Teacher sulle attivazioni proiettate dello Student.

Ciò ci consente di valutare se la rappresentazione interna della veridicità (o dell’allucinazione) è condivisa tra diverse architetture LLM.

# Chapter 2

## Risultati

Questo capitolo presenta i risultati degli esperimenti di rilevamento delle allucinazioni. Vengono analizzate le prestazioni dei probe lineari e l'efficacia dell'allineamento cross-model.

### 2.1 Analisi PCA delle Attivazioni

Per visualizzare la separabilità degli stati allucinati e non allucinati, è stata eseguita l'Analisi delle Componenti Principali (PCA) sulle attivazioni.

Come mostrato in Figura 2.1, esiste una notevole separazione tra le due classi nello spazio delle attivazioni, in particolare per il modello Qwen. Nei layer intermedi e finali è possibile osservare diversi cluster. Per il modello Falcon, la separazione è meno marcata ma ancora evidente in alcuni layer.

### 2.2 Risultati Quantitativi

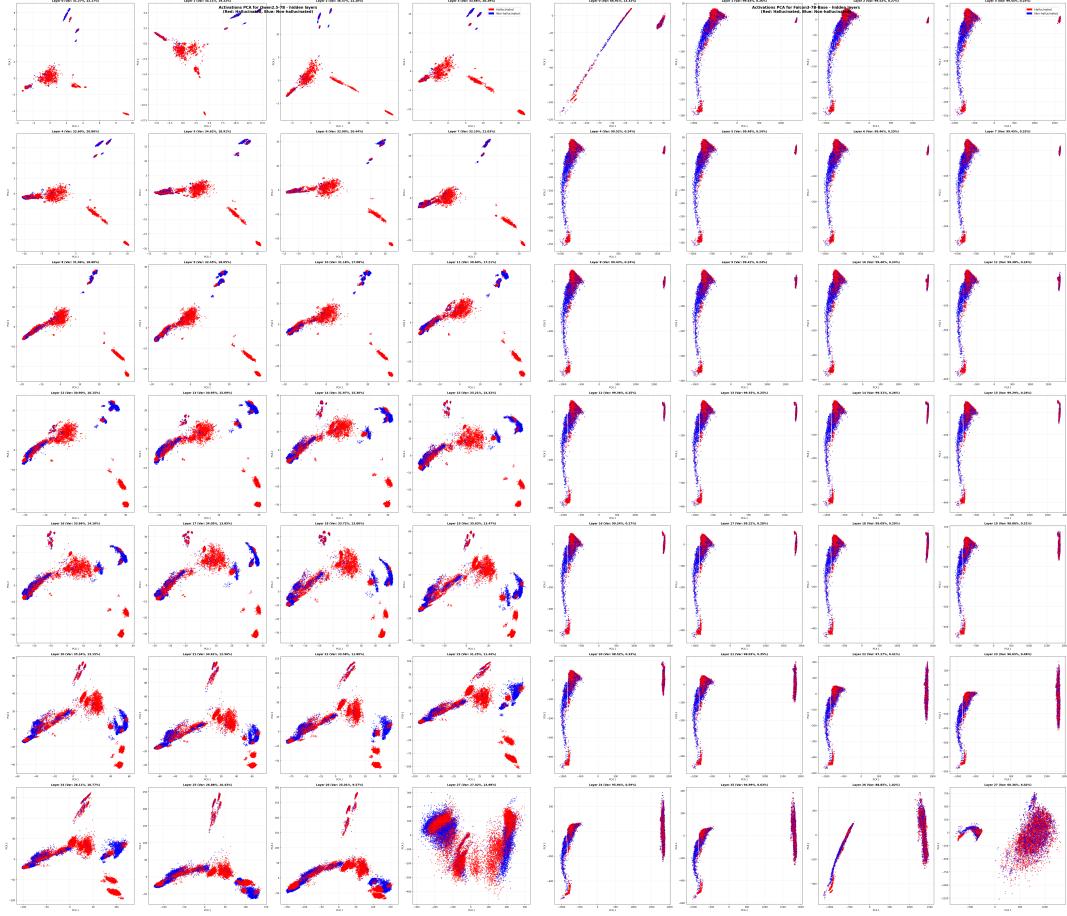
Sono state valutate le prestazioni dei probe in due scenari:

1. **Scenario 1:** Qwen2.5-7B come Teacher → Falcon3-7B-Base come Student.
2. **Scenario 2:** Falcon3-7B-Base come Teacher → Qwen2.5-7B come Student.

#### 2.2.1 Scenario 1: Qwen → Falcon

In questo scenario, il probe è stato addestrato sulle attivazioni di Qwen. Qwen ha raggiunto un'elevata accuratezza nel rilevare le proprie allucinazioni.

I risultati nella Tabella 2.1 mostrano che mentre il probe è estremamente efficace sul Teacher (Qwen), le prestazioni calano significativamente quando applicato allo Student allineato (Falcon), con un'accuratezza che varia dal 65% al 73%.



**Figure 2.1:** PCA delle Attivazioni Hidden Layer per Qwen2.5-7B (sinistra) e Falcon3-7B-Base (destra). I punti rossi indicano le allucinazioni.

### 2.2.2 Scenario 2: Falcon → Qwen

Nello scenario inverso, Falcon è stato utilizzato come Teacher.

La Tabella 2.2 indica che Falcon è un Teacher meno efficace, con un'accuratezza di base inferiore. Anche il trasferimento a Qwen è limitato, con accuratezze intorno al 62-65%.

## 2.3 Matrici di Confusione

Gli errori sono stati ulteriormente analizzati utilizzando le matrici di confusione.

La Figura 2.2 illustra che l'applicazione cross-model comporta un numero maggiore di Falsi Positivi e Falsi Negativi rispetto all'autovalutazione del Teacher.

Qwen è un modello da 28 layer ciascuno con una hiddensize di 3584, mentre Falcon ha 28 layer con una hiddensize di 3,072. Questo potrebbe suggerire che usare un modello Teacher più grande e complesso aiuti a catturare rappresentazioni più ricche

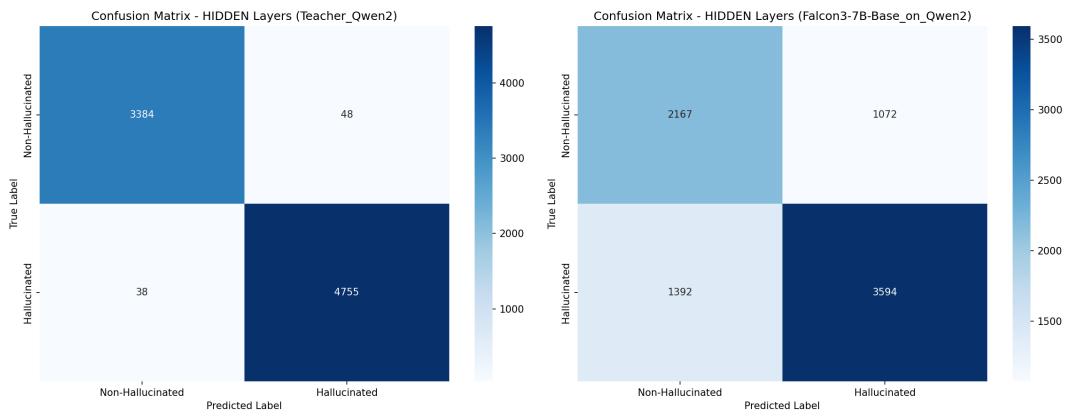
<b>Tipo Layer</b>	<b>Acc Teacher</b>	<b>Acc Student</b>	<b>F1 Teacher</b>	<b>F1 Student</b>
Attention	0.9880	0.7328	0.9897	0.7642
MLP	0.9889	0.6546	0.9905	0.7245
Hidden	0.9895	0.7004	0.9910	0.7447

**Table 2.1:** Risultati per lo Scenario 1 (*Qwen* → *Falcon*)

<b>Tipo Layer</b>	<b>Acc Teacher</b>	<b>Acc Student</b>	<b>F1 Teacher</b>	<b>F1 Student</b>
Attention	0.8930	0.6534	0.9110	0.7197
MLP	0.7970	0.6283	0.8270	0.7045
Hidden	0.8737	0.6530	0.8944	0.7231

**Table 2.2:** Risultati per lo Scenario 2 (*Falcon* → *Qwen*)

per il rilevamento delle allucinazioni, facilitando il trasferimento allo Student.



**Figure 2.2:** Matrici di Confusione per Hidden Layers. Sinistra: Qwen (Teacher). Destra: Falcon su Qwen (Student).

# Bibliography



# Appendices



# Appendix A

## Visualizzazioni PCA

In questa appendice sono riportate le visualizzazioni complete dell’Analisi delle Componenti Principali (PCA) a 2 dimensioni per tutte le tipologie di layer (Attention, MLP, Hidden) e per entrambi i modelli analizzati (Qwen2.5-7B e Falcon3-7B-Base). I punti rossi rappresentano le istanze allucinate, mentre i punti blu rappresentano le istanze corrette.

### A.1 Qwen2.5-7B

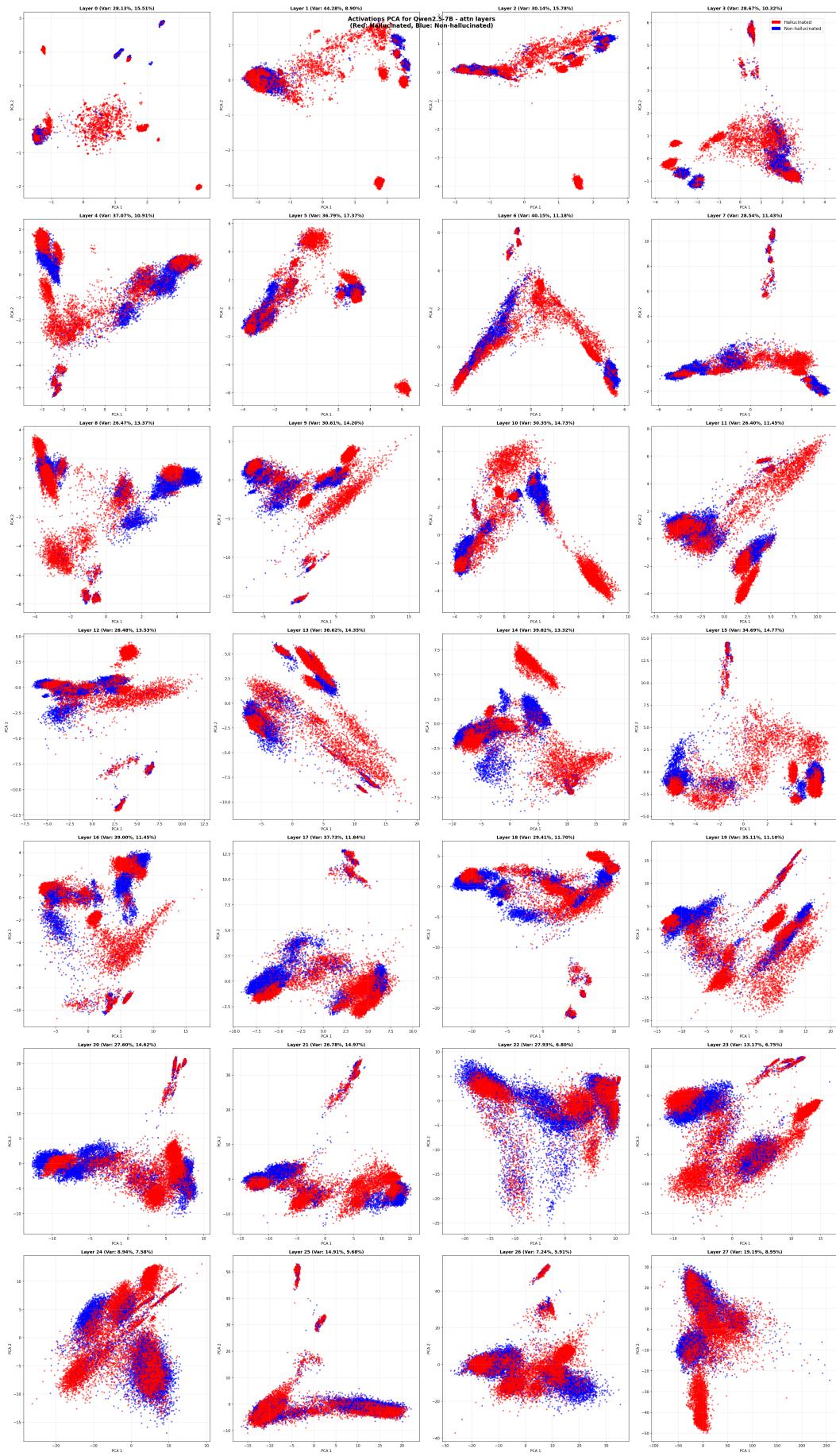


Figure A.1: PCA 2D - Qwen2.5-7B - Attention Layers

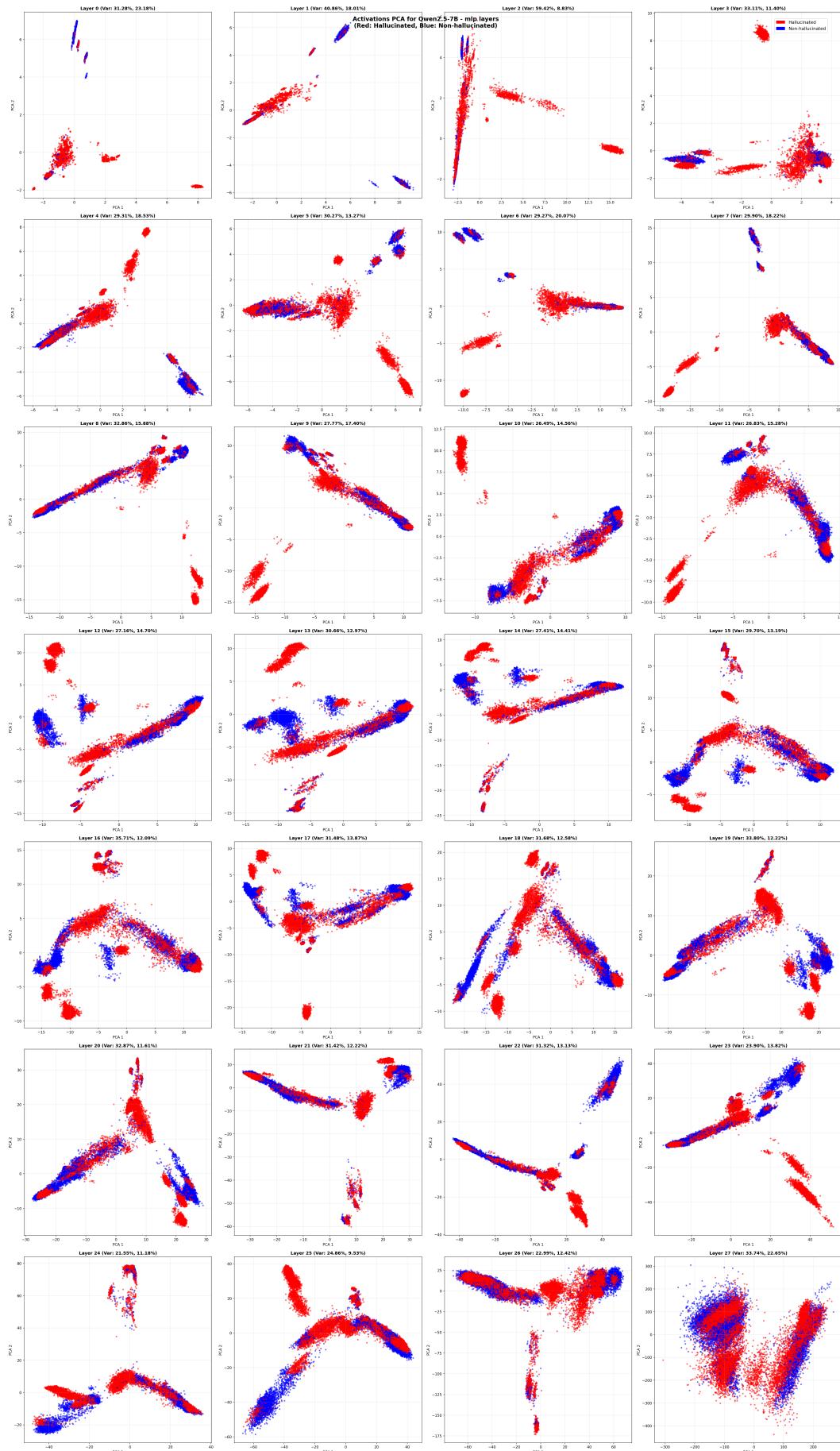
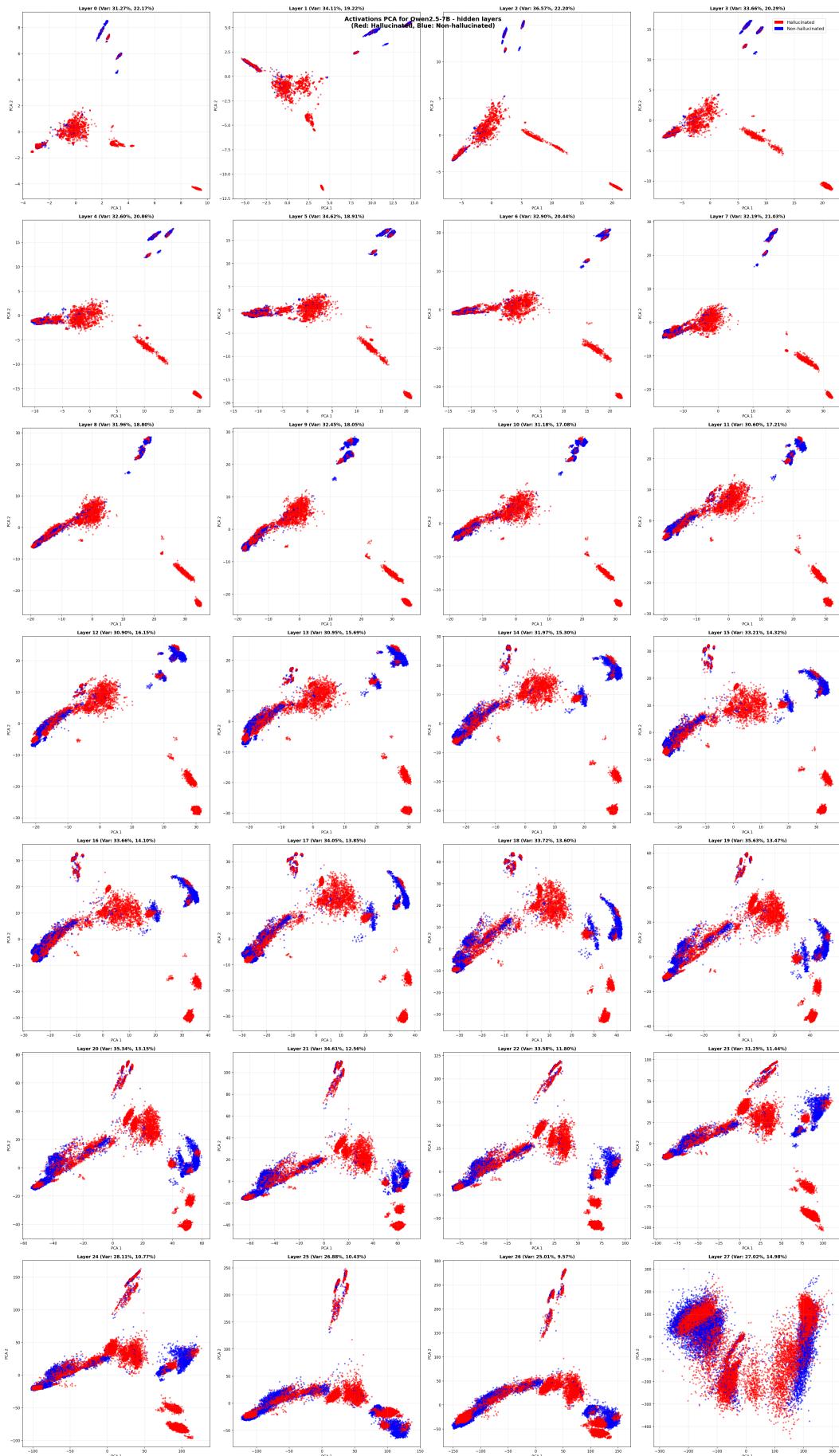


Figure A.2: PCA 2D - Qwen2.5-7B - MLP Layers



**Figure A.3: PCA 2D - Qwen2.5-7B - Hidden Layers**

## A.2 Falcon3-7B-Base

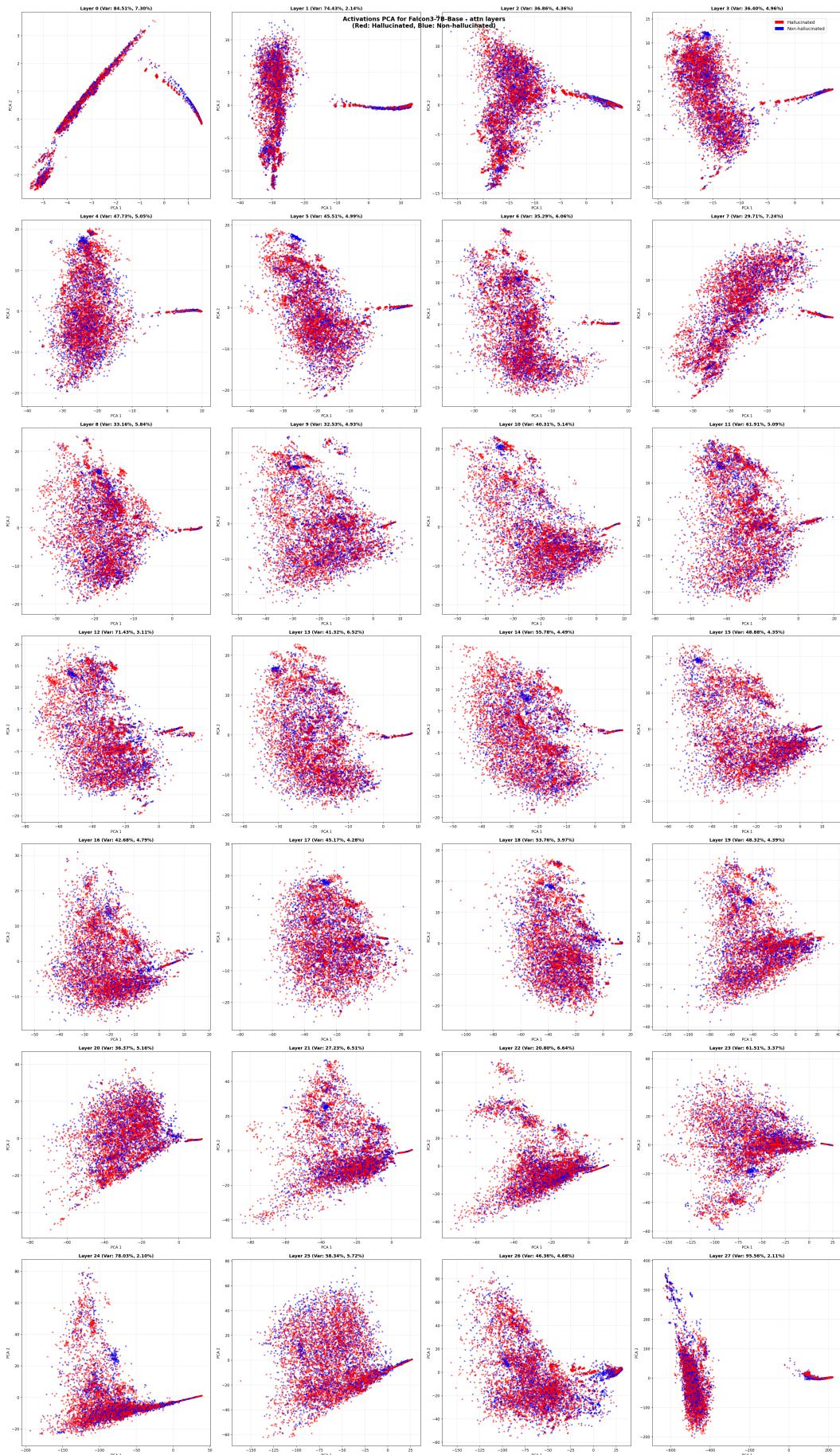


Figure A.4: PCA 2D - Falcon3-7B-Base - Attention Layers

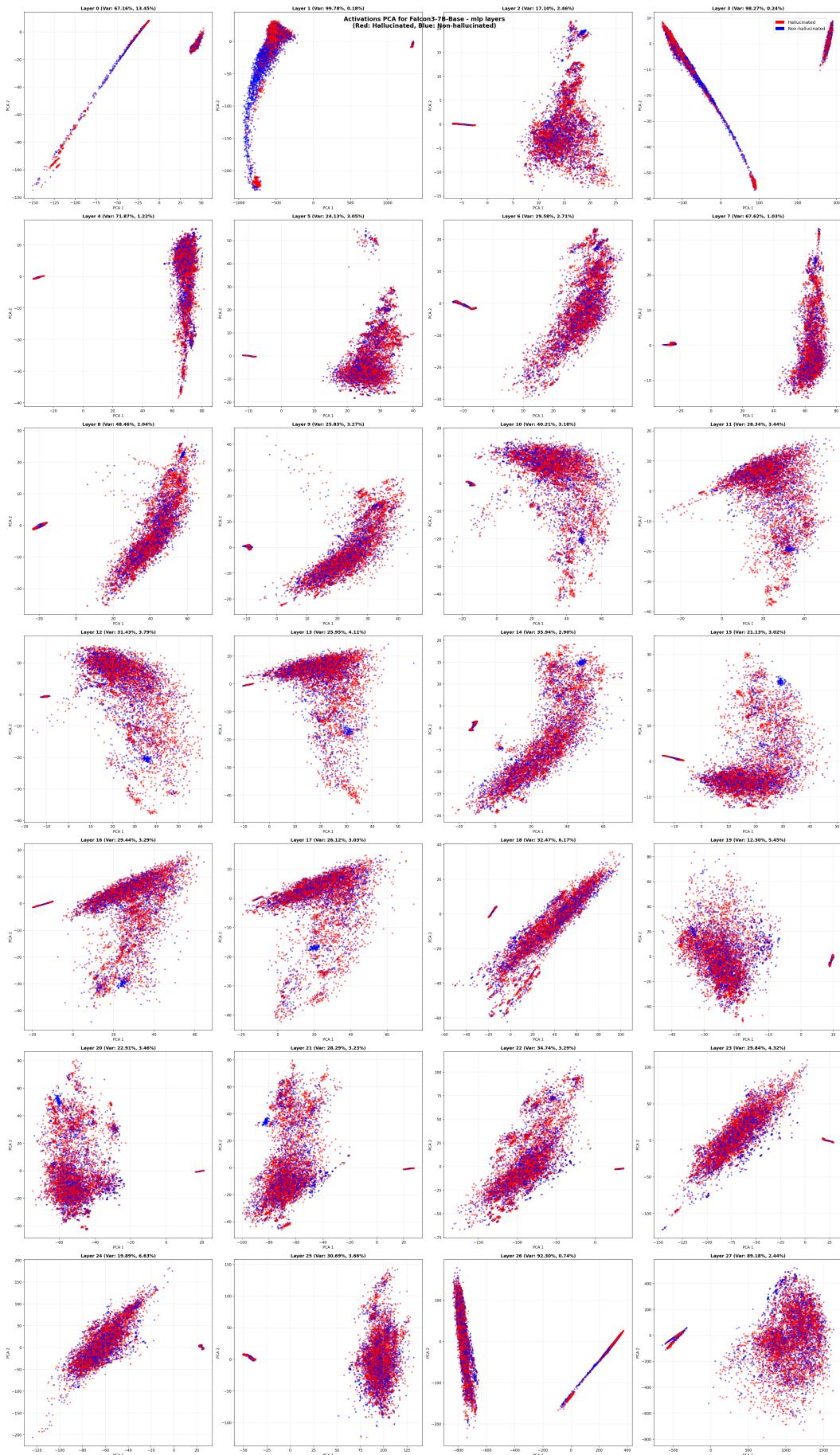


Figure A.5: PCA 2D - Falcon3-7B-Base - MLP Layers

## APPENDIX A. VISUALIZZAZIONI PCA

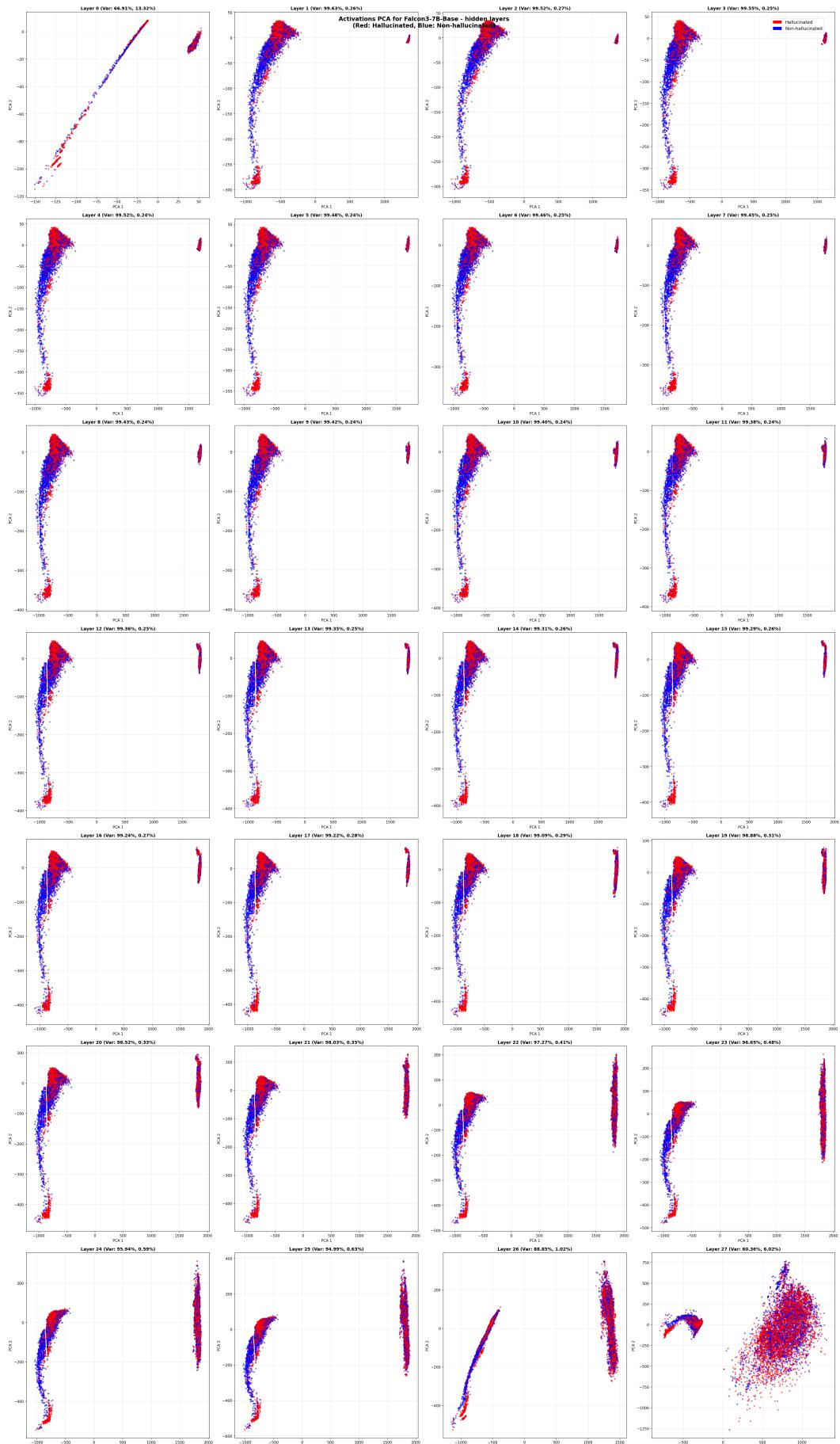


Figure A.6: PCA 2D - Falcon3-7B-Base - Hidden Layers