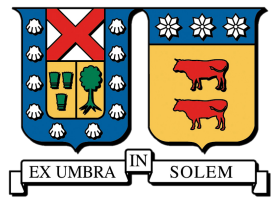
Detección de alucinaciones en LLM mediante la dinámica estructural de los grafos de atención

Nicolás Schiaffino Mellado & Gabriel Venegas Ortiz

PUC & UTFSM

IIC 3641 Aprendizaje Basado en Grafos
Prof. Marcelo Mendoza



Outline

Alucinaciones en LLMs

El impacto de los Large Language Models (LLMs) se expande rápidamente en dominios críticos como la salud, la educación y el desarrollo de software. Sin embargo, esta adopción masiva eclipsa una vulnerabilidad fundamental: las **alucinaciones**.

Este fenómeno puede llevar a resultados inesperados y potencialmente perjudiciales, resaltando la necesidad de una investigación más profunda hacia la detección y mitigación de estas alucinaciones.

- Se entrenará un **Variational Auto Encoder (VAE)** para el grafo obtenido en base a la matriz de activaciones y las activaciones de cada token (nodo) en una capa determinada. Esto con el fin de obtener una representación estructural latente ($\mathbf{Z}$) de cada grafo.
- Una vez se obtengan todas las representaciones estructurales latentes a lo largo de todas las capas, se entrenará un MLP para identificar anomalías en la estructura considerando la capa anterior y posterior.

Generación del Dataset

El dataset con el que se trabajará será una recolección del estado interno de *Llama-3.2-1B* (activaciones y atenciones **por cada capa**).

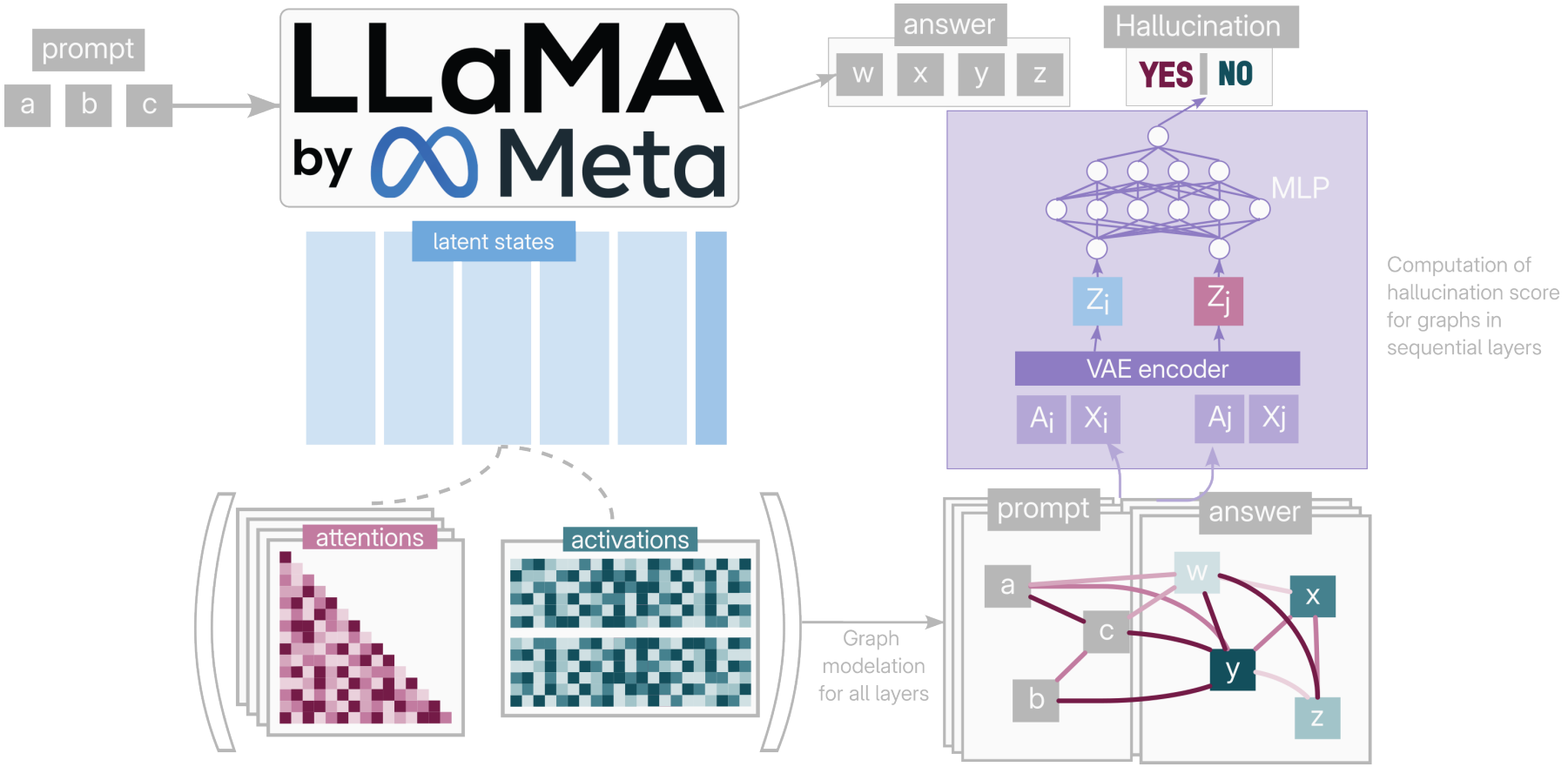
Los inputs para recuperar cada estado interno constan de las preguntas presentes en el dataset *TruthfulQA*.

Diferencias con otros enfoques

Los elementos a grandes rasgos son 3 y se compara con el State of the Art en detección de alucinaciones en LLMs.

1. HaloScope [2] introduce la noción de una diferencia geométrica en el espacio latente entre contenido no alucinado y si alucinado. No obstante, trabaja con una versión **estática** del espacio latente.
2. CHARM [3] propone un modelamiento de las activaciones y atenciones en un grafo pero **colapsa toda la arquitectura a un solo grafo** promediando la información entre capas.
3. HalluShift [1] realiza la detección basándose en las distancias entre las distribuciones de la matriz de atención a través de las capas. Sin embargo, **no captura información estructural**.

Graphical Abstract



Actividades Avance de proyecto

Fase	**Actividades Principales**
Avance Proyecto	1. Generar las respuestas del Llama-1B para *TruthfulQA*. 2. Implementar el script de "autopsia" para extraer los *hidden states* y *scores* de atención $A_{l,h}$ de las capas L. 3. Procesar y guardar en disco el *dataset* final de secuencias de grafos $\{G_l\}$. 4. Implementar el *dataloader* que carga las secuencias de grafos desde el disco. 5. Implementar *VAE* para el entrenamiento sobre los grafos G_l.

Actividades Entrega Final

Fase	**Actividades Principales**
Entrega	1. Entrenar el VAE de forma no supervisada sobre el *dataset* mixto hasta la convergencia (minimizando $L_{recon} + L_{KL}$) 2. Implementar los *scripts* de *scoring de alucinaciones* (MLP) sobre la secuencia de z_l (representaciones estructurales latentes)
Final	3. Implementar HaloScope & HalluShift para conocer cuáles resultados obtienen en Llama-3.2-1B 4. Ejecutar la evaluación en el dataset de *test* y generar gráficos comparativos.

Entregables

Fase	**¿Qué se entrega?**
Avance	Para las actividades 1., 2. y 3. se entregará un dataset subido a alguna plataforma en la nube (por ej. *Zenodo*) junto con un instructivo especificando cómo fue generado el dataset. Para la actividades 4. y 5. se entregará un enlace al repositorio *GitHub* con las implementaciones de los códigos y las correspondientes instrucciones para su ejecución.
Final	Se entregará un enlace al repositorio de *github* extendiendo las instrucciones Los resultados obtenidos tendrán en consideración la replicabilidad de los experimentos y se detallará en el repositorio.

De todas formas, para ambas entregas se cumplirá adicionalmente con comentar el avance y mencionar todas las modificaciones realizadas.

Referencias

[1] S. Dasgupta, S. Nath, A. Basu, P. Shamsolmoali, and S. Das. Hallushift: Measuring distribution shifts towards hallucination detection in llms, 2025. URL https://arxiv.org/abs/2504.09482.

[2] X. Du, C. Xiao, and Y. Li. Haloscope: Harnessing unlabeled llm generations for hallucination detection. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 102948–102972. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ba92705991cfbbcedc26e27e833ebbae-Paper-Conference.pdf.

[3] F. Frasca, G. Bar-Shalom, Y. Ziser, and H. Maron. Neural message-passing on attention graphs for hallucination detection, 2025. URL https://arxiv.org/abs/2509.24770.