

# 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

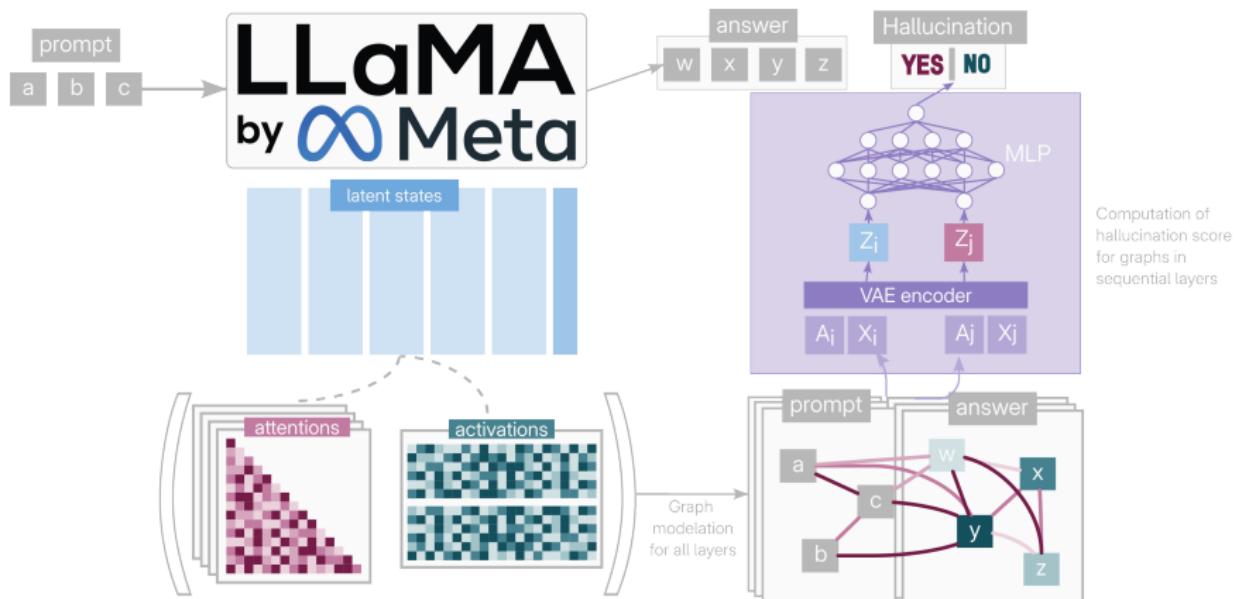


# Outline

1 Avance en Actividades

2 Entregables

# Graphical Abstract



# Asociado al LLM

Recordatorio de las actividades:

- ① Generar las respuestas del Llama-1B para truthfulQA.
- ② Implementar un script de *autopsy* para extraer hidden states y scores de atención.

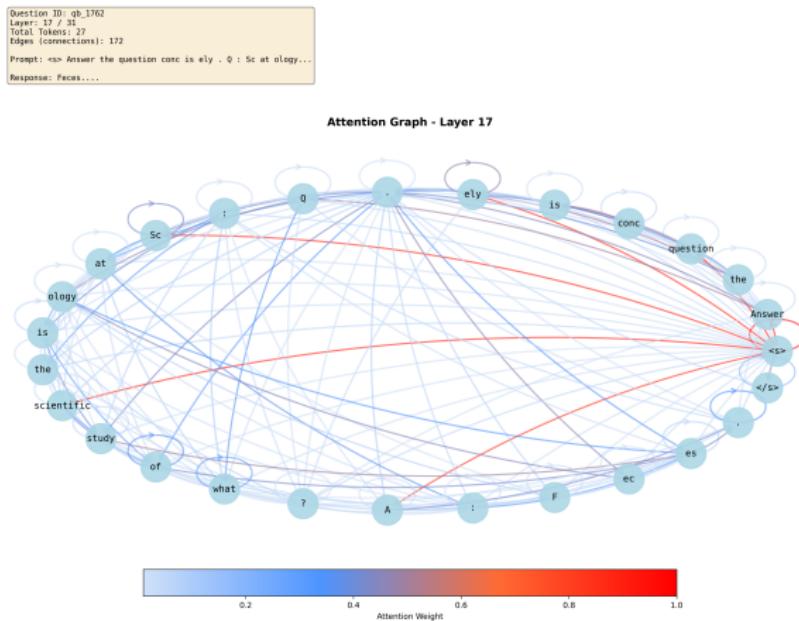
# Asociado al LLM

Detalles que se  
cambiaron:

- ➊ Generar las respuestas del **Llama-1B** → **Llama2-7B-chat-vf** para **truthfulQA** → **triviaQA**.
- ➋ Implementar un script de *autopsia* para extraer hidden states y scores de atención.  
(En el repositorio corresponde a `src/trace_extractor.py`)

# Generación del grafo

Gracias al script `src/dataloader.py` se genera. Un ejemplo es:



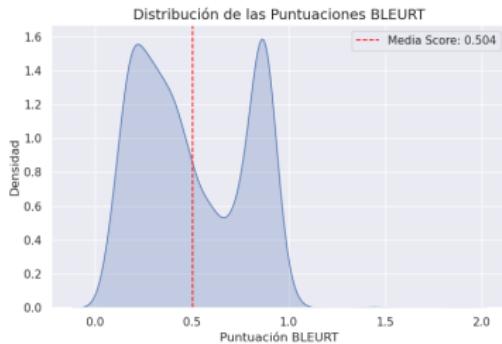
# Implementación de VAE

Antes de pasar de lleno a la implementación de VAE, se debió realizar tareas adicionales:

- Comparar respuestas generadas por Llama2-7B-chat-vf con el ground-truth del dataset. Como no se puede hacer directamente, se debió implementar **bleurt-20** para comparar. Se encuentra en `src/trace_to_gt.py`
- Implementar baselines que **justifiquen** las ganancias de la arquitectura planteada (tipo ablaciones). Los 3 modelos adicionales se encuentran en `src/baseline.py`

Con esto se busca confirmar que agregar incertidumbre (VAE) es realmente valioso.

# Detalles Arquitectura VAE



**Figure:** Bleurt Scores de las generaciones del LLM. Mientras más positivo significa una mayor similitud. Se considera 0.5 en adelante como una respuesta no alucinada.

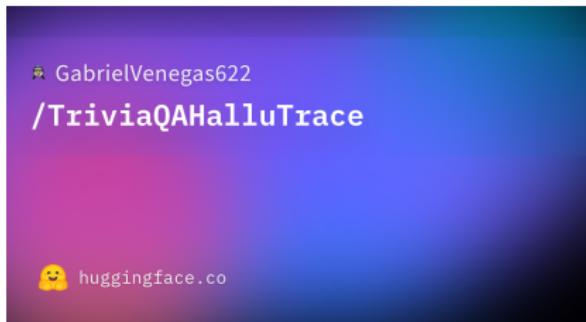
Se utiliza Graph Isomorphism Network with Edges (GINE) a modo de decoder mediante la función de GINEconv de pytorch geometric.

Los parámetros default son: un vector de dimensión 64 para su salida ( $Z \sim \mathcal{N}(\mu, \sigma)$ ) y se considera un dropout de 0.3. Luego se pasa por un biLSTM de 2 capas que verificará la dinámica de la estructura de los grafos. [0]

# Links Entregables

Para cumplir con las actividades y los entregables:

- El dataset con activaciones y atenciones ya está disponible en:  
[huggingface/GabrielVenegas622/TriviaQAHalluTrace](https://huggingface.co/GabrielVenegas622/TriviaQAHalluTrace).
- Los scripts mencionados (`src/trace_extractor.py`,  
`src/dataloader.py`, `src/trace_to_gt.py`) están disponibles en el  
repositorio: [GabrielVenegas622/HallucinationsDetectionGML](https://github.com/GabrielVenegas622/HallucinationsDetectionGML)



GabrielVenegas622/  
**HallucinationsDetection...**

