



UNIVERSIDAD DE BUENOS AIRES  
FACULTAD DE CIENCIAS EXACTAS Y NATURALES

# Degradación controlada en un Gran Modelo de Lenguaje (LLM)

Tesis de Licenciatura en Ciencias de Datos

Alejandro Wainstock

Director: Dr. Enzo Tagliazucchi  
Ciudad de Buenos Aires, 2025



# DEGRADACIÓN CONTROLADA EN UN LLM

La investigación en Grandes Modelos de Lenguaje (LLMs) reveló su potencial no solo como herramientas de procesamiento de lenguaje natural, sino también como análogos computacionales para comprender la cognición humana. Esta tesis exploró los efectos de la degradación controlada en un LLM, específicamente el modelo *Gemma 3 4B-IT*, con el objetivo de caracterizar los patrones de deterioro y establecer analogías con trastornos del lenguaje y del pensamiento en humanos.

Para ello, se implementaron tres métodos de perturbación (ruido gaussiano multiplicativo, ablación de pesos y cuantización uniforme) aplicados sistemáticamente sobre componentes clave del modelo: las matrices de *Embeddings*, de las capas de atención y de las redes *Feed-Forward* (MLP). Se diseñó un protocolo experimental riguroso para evaluar el rendimiento del modelo degradado en un conjunto de tareas que incluyen relatos de sueños, la descripción de la lámina del *Cookie Theft*, razonamiento lógico-matemático y comprensión verbal. Las métricas de evaluación se obtuvieron mediante análisis lingüísticos computacionales y la evaluación de un LLM juez externo (GPT-4o-mini).

Los resultados revelaron en algunos casos una vulnerabilidad diferencial de los componentes del modelo. Por ejemplo, la degradación en las capas de atención produjo un deterioro gradual, caracterizado por una pérdida progresiva de la concisión y coherencia discursiva. El análisis del orden de deterioro mostró que, dependiendo del componente afectado, emergen distintas jerarquías de vulnerabilidad. Notablemente, el daño en la matriz de *Embeddings* afectó primero a la riqueza léxica, mientras que el daño en la atención y el MLP impactó de forma más temprana en las tareas de razonamiento.

Este trabajo subrayó el valor de los LLMs como modelos *in silico* para la neurociencia computacional, permitiendo simular y estudiar la desintegración de funciones cognitivas de manera controlada. Asimismo, ofreció *insights* sobre la arquitectura funcional de los LLMs, identificando vulnerabilidades y contribuyendo al desarrollo de sistemas de IA más robustos y comprensibles. Las implicaciones futuras abarcan desde el desarrollo de herramientas de diagnóstico temprano hasta la formulación de hipótesis sobre la etiología de los trastornos del lenguaje y del pensamiento.

**Palabras clave:** Grandes Modelos de Lenguaje, Degradación Controlada, Simulación Cognitiva, Neurociencia Computacional, Trastornos del Lenguaje, Gemma, Modelos In Silico.



## CONTROLLED DEGRADATION IN AN LLM

Research on Large Language Models (LLMs) has revealed their potential not only as natural language processing tools but also as computational analogs for understanding human cognition. This thesis explored the effects of controlled degradation on an LLM, specifically the *Gemma 3 4B-IT* model, with the aim of characterizing its patterns of deterioration and establishing analogies with language and thought disorders in humans.

To this end, three perturbation methods (multiplicative Gaussian noise, weight ablation, and uniform quantization) were systematically applied to key components of the model: the *Embedding* matrix, the attention layers, and the *Feed-Forward* networks (MLP). A rigorous experimental protocol was designed to evaluate the performance of the degraded model on a set of tasks including dream narratives, the description of the *Cookie Theft* picture, logical-mathematical reasoning, and verbal comprehension. Evaluation metrics were obtained through computational linguistic analysis and the assessment by an external judge-LLM (GPT-4o-mini).

The results revealed a differential vulnerability of the model’s components. For example, degradation in the attention layers produced a gradual deterioration, characterized by a progressive loss of conciseness and discursive coherence. The analysis of the order of deterioration showed that, depending on the affected component, different hierarchies of vulnerability emerge. Notably, damage to the *Embedding* matrix first affected lexical richness, whereas damage to the attention and MLP layers had an earlier impact on reasoning tasks.

This work underscored the value of LLMs as *in silico* models for computational neuroscience, allowing for the controlled simulation and study of the disintegration of cognitive functions. Furthermore, it offered insights into the functional architecture of LLMs, identifying vulnerabilities and contributing to the development of more robust and understandable AI systems. Future implications range from the development of early diagnostic tools to the formulation of hypotheses about the etiology of language and thought disorders.

**Keywords:** Large Language Models, Controlled Degradation, Cognitive Simulation, Computational Neuroscience, Language Disorders, Gemma, In Silico Models.



## AGRADECIMIENTOS

En primer lugar, agradezco a mi director de tesis, Enzo Tagliazucchi. Valoro enormemente la confianza y la libertad que me dio para explorar mis propios intereses. Gracias a su generosa orientación, acompañamiento y sabiduría fue posible transformar una intención en un proyecto de investigación concreto.

Extiendo mi gratitud a la Facultad de Ciencias Exactas y Naturales y a la Universidad de Buenos Aires en general. Mi reconocimiento a cada uno de los docentes que trabajan comprometidos con una formación de excelencia, que reivindica a la educación pública, gratuita y de calidad. Esto fue un pilar fundamental para mi desarrollo y el de la sociedad en su conjunto.

A mis compañeros y amigos de cursada, con quienes compartí incontables horas de estudio, chats, discusiones, resoluciones de ejercicios, debates y mates. Su inteligencia y buen humor hicieron que el camino fuera aun más interesante y divertido de lo imaginado.

Finalmente, a mi familia y amigos/as, por su apoyo, su aliento y su paciencia, que fueron indispensables para poder completar esta etapa.





## Índice general

1. Introducción . . . . .	1
1.1 Lenguaje en diagnósticos de salud mental . . . . .	1
1.2 Trabajos previos en la simulación de trastornos del lenguaje en LLMs . . . . .	2
1.3 Objetivos y motivación de la tesis . . . . .	3
2. Degradación inducida en un LLM . . . . .	5
2.1 Modelo de lenguaje seleccionado . . . . .	5
2.2 Fundamentos de LLMs . . . . .	5
2.3 Formas de degradación artificial estudiadas . . . . .	8
2.3.1 Ruido multiplicativo Gaussiano ( <b>mult_gauss</b> ) . . . . .	8
2.3.2 Ablación de pesos ( <b>ablation</b> ) . . . . .	9
2.3.3 Cuantización uniforme ( <b>uni_quant</b> ) . . . . .	9
2.4 Diseño experimental . . . . .	10
3. Evaluación del deterioro y sus efectos . . . . .	13
3.1 Tareas pedidas al modelo y <i>prompts</i> utilizados . . . . .	13
3.2 Preprocesamiento de textos . . . . .	16
3.3 Métricas discursivas y de performance . . . . .	16
3.4 Criterio de deterioro significativo . . . . .	19
4. Resultados . . . . .	21
4.1 Análisis preliminar . . . . .	21
4.2 Indicadores discursivos y psicológicos . . . . .	28
4.3 Razonamiento y capacidades cognitivas . . . . .	32
4.4 Análisis de métricas y orden de deterioro . . . . .	33
5. Discusión . . . . .	39
5.1 Interpretación de los hallazgos principales . . . . .	39
5.2 Limitaciones . . . . .	40
5.3 Implicancias y conclusión . . . . .	41
5.4 Líneas futuras de investigación . . . . .	42
Bibliografía . . . . .	47
Apéndice . . . . .	49
Apéndice A: Ejemplos de <i>prompts</i> de evaluación del LLM-Juez . . . . .	51
A.1 System Prompt . . . . .	51
A.2 Instrucciones ( <i>prompt</i> ) para el LLM-Juez . . . . .	51
A.2.1 Bloque 1: Calidad del relato ( <b>quality_cookies</b> ) . . . . .	51
A.2.2 Bloque 2: Detección de Unidades de Contenido ( <b>component_detection</b> ) . . . . .	52



# 1. INTRODUCCIÓN

## 1.1. Lenguaje en diagnósticos de salud mental

El análisis del lenguaje natural se consolida como una herramienta eficaz, accesible y con alta validez ecológica para el diagnóstico automático de trastornos neurológicos y psiquiátricos. Su capacidad para detectar marcadores discursivos relevantes ofrece una alternativa costo-efectiva a los métodos tradicionales. La literatura científica evidencia diferencias lingüísticas significativas entre diversas patologías, sentando las bases para la modelización computacional.

En la literatura disponible se identifica un conjunto de marcadores lingüísticos específicos que se correlacionan con distintas condiciones psiquiátricas y neurológicas. Por ejemplo, las alteraciones del pensamiento como la tangencialidad (desvío del tema central) y el descarrilamiento (cambios abruptos de tópico) son indicadores clásicos en la esquizofrenia, ya que reflejan una pérdida de coherencia discursiva [6].

En el plano afectivo, un marcador clave es la alteración del *arousal*, que se refiere al nivel de activación fisiológica y psicológica de una persona. Una disminución significativa de este arousal da lugar al aplanamiento afectivo, un síntoma central en la esquizofrenia. En el discurso, esto puede manifestarse como una reducción en la intensidad emocional del contenido, con un tono afectivo atenuado, produciendo narrativas más planas y factuales [12, 3].

En la dimensión léxica, se observa que una baja diversidad (variedad de palabras) y, en muchos contextos, una menor densidad léxica (proporción de palabras de contenido) son rasgos presentes en la esquizofrenia [5] y, en cierta medida, también en el declive cognitivo [14]. De forma complementaria, una elevada frecuencia léxica —el uso de palabras muy comunes— es un marcador sensible en individuos con riesgo genético de Alzheimer, lo que sugiere una dificultad para acceder a un vocabulario más rico [19]. Asimismo, una menor profundidad léxica, es decir, el uso de términos generales en vez de específicos (e.g., “animal” en lugar de “leopardo”), es típica de la demencia semántica por el deterioro del conocimiento conceptual [2]. Sanz et al. (2022) mostraron que los pacientes con Alzheimer tienden a exhibir una menor profundidad léxica y mayor variabilidad en su discurso, marcadores que permiten discriminar eficazmente entre esta enfermedad y el Parkinson [15].

Finalmente, se considera el rol de la personalidad, a menudo evaluada mediante el modelo de los “Cinco Grandes” (*Big Five*): Apertura a la experiencia (*Openness*), Responsabilidad (*Conscientiousness*), Extraversión (*Extraversion*), Amabilidad (*Agreeableness*) y Neuroticismo (*Neuroticism*). La literatura establece correlaciones entre estos rasgos y la salud mental; por ejemplo, niveles bajos de apertura a la experiencia se asocian consistentemente con un mayor riesgo de depresión resistente al tratamiento [17]. Lingüísticamente, la apertura tiende a manifestarse en un estilo más analítico y sofisticado (mayor uso de artículos y preposiciones, y vocabulario de contenido intelectual), y en mayor complejidad discursiva [20, 16].

Para elicitarse discurso en los pacientes, se recurre a tareas específicas como la narración de sueños y a herramientas estandarizadas. Entre estas últimas, una de las más emblemáticas es la tarea de descripción de la lámina del “Robo de Galletas” (*Cookie Theft Task*),

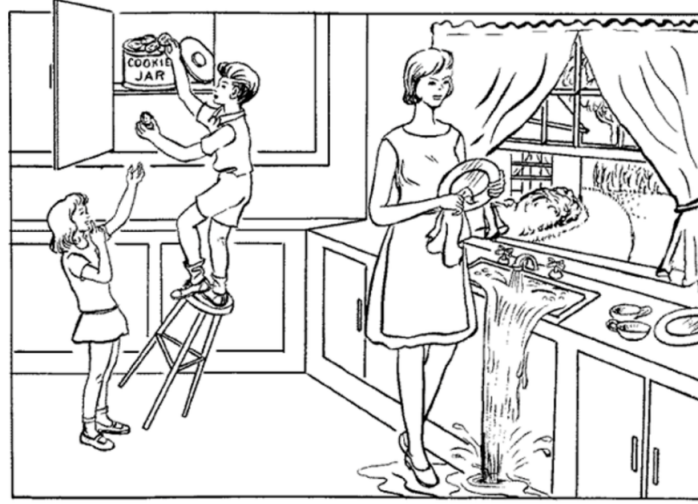


Fig. 1.1: Imagen de *Cookie Theft Task*, originaria del *Boston Diagnostic Aphasia Examination* [8]

como se ve en la Figura 1.1 [8]. Esta imagen presenta una escena doméstica de cocina con múltiples personajes y acciones simultáneas (unos niños intentando robar galletas, una mujer distraída, un banco cayendo, entre otras), que fue diseñada para provocar un discurso narrativo complejo. Tradicionalmente, el análisis clínico de estos relatos se centra en cuantificar “Unidades de Contenido” (UCs), es decir, elementos informativos clave de la escena mencionadas por el paciente, y evaluar la coherencia discursiva, para diagnosticar déficits en la afasia y las demencias. Estudios recientes muestran que los LLMs pueden funcionar como evaluadores (“*judge-LLMs*”), puntuando la presencia de UCs y la coherencia, con una fiabilidad que rivaliza o en algunos casos supera a la de los evaluadores humanos [9].

Otros estudios, como el de Colla et al. (2022), utilizaron la métrica de perplejidad (*perplexity*) —derivada de modelos de lenguaje— para distinguir entre pacientes con Alzheimer y sujetos sanos, destacando su potencial en el diagnóstico precoz de trastornos neurodegenerativos [4]. Estas investigaciones establecen un sólido fundamento clínico para explorar cómo los grandes modelos de lenguaje (LLMs) pueden no solo reflejar, sino también simular patrones patológicos del lenguaje humano mediante manipulaciones controladas.

## 1.2. Trabajos previos en la simulación de trastornos del lenguaje en LLMs

Basados en los hallazgos clínicos, diversos estudios comenzaron a explorar cómo los grandes modelos de lenguaje (LLMs) pueden presentar espontáneamente patrones lingüísticos análogos a los observados en trastornos humanos, así como las formas de inducirlos artificialmente para simular distintos tipos de daño neurológico. Palaniyappan et al.

(2023) propusieron utilizar estos modelos como sistemas *in silico* para estudiar los patrones lingüísticos propios de la psicosis, argumentando que los errores generados por los LLMs —como la degeneración textual o la invención de información— se asemejan al lenguaje desorganizado y las falsas creencias observadas en pacientes psicóticos [13].

La investigación reciente se enfocó en distintas metodologías de “degradación controlada”:

- **Manipulación de parámetros:** Fradkin, Nour y Dolan (2023) exploraron cómo la variación de parámetros en un modelo *GPT-2* puede recrear patrones del trastorno formal del pensamiento (*FTD*), característico de la esquizofrenia. Al manipular la temperatura (que controla la aleatoriedad de la predicción) y la capacidad de memoria (*memory span*), que limita el uso del contexto global, lograron generar narrativas menos coherentes. Específicamente, un aumento en la temperatura incrementó la distancia semántica entre palabras, mientras que una reducción en la memoria la aumentó entre oraciones consecutivas, emulando rasgos típicos del *FTD* como el descarrilamiento y la tangencialidad [7].
- **Deterioro estructural interno:** Li et al. (2022, 2024) indujeron un “deterioro artificial” en modelos *GPT-2* para simular las anomalías lingüísticas de la demencia tipo Alzheimer. Utilizaron el enmascaramiento selectivo de la matriz de *embeddings* y de las matrices de valor (*V*) en las capas de atención para evaluar el modelo en su capacidad para diferenciar transcripciones de pacientes (obtenidas de tareas como la descripción de la lámina “Cookie Theft”) de las de controles sanos, usando la perplejidad como métrica principal. Su primera investigación (2022) observó que el daño en las capas de atención era más efectivo para simular los déficits. Posteriormente (2024), demostraron que la resistencia a este tipo de daño es significativamente mayor en modelos de mayor tamaño (*GPT-2-XL* vs. *Small*), sugiriendo una relación entre la escala del modelo y su resiliencia cognitiva, un análogo a la reserva cognitiva en humanos [10, 11].
- **Simulación de neurodegeneración:** Alexos et al. (2024) utilizaron el concepto de “erosión neuronal” para simular procesos neurodegenerativos en un modelo *LLaMA 2*. Tras descartar la ablación por no producir un declive gradual, optaron por la inyección de ruido gaussiano en los bloques de atención y *feed-forward*. Evaluaron el desempeño del modelo con preguntas de exámenes de IQ derivados de MENSA, que incluían tareas de razonamiento verbal y cuantitativo. Los resultados mostraron un deterioro progresivo que emula patrones clínicos humanos: un declive inicial en las habilidades matemáticas (pérdida de reconocimiento de patrones), seguido por el lenguaje y, finalmente, una afectación de la comprensión profunda, donde el modelo respondía de forma incoherente o en bucles [1].

### 1.3. Objetivos y motivación de la tesis

Esta tesis se inscribe en la intersección de la neurociencia computacional y la Inteligencia Artificial, utilizando los Grandes Modelos de Lenguaje (LLMs) como un laboratorio *in silico* para explorar los mecanismos del lenguaje y su desintegración. El objetivo principal es caracterizar sistemáticamente los efectos de la degradación controlada en un LLM de última generación, para lo cual se establecen los siguientes objetivos específicos:

1. **Implementar y aplicar métodos de degradación:** Se estudiará la aplicación de distintas técnicas de perturbación artificial sobre diferentes componentes clave de la arquitectura del modelo.
2. **Evaluar el impacto en el rendimiento:** Se medirá el efecto de cada tipo de daño en un conjunto diverso de tareas, que abarcan desde el razonamiento lógico-matemático y verbal hasta la generación de discurso y la descripción de escenas complejas.
3. **Analizar la vulnerabilidad diferencial:** Se investigará si los distintos componentes del modelo exhiben diferentes niveles de sensibilidad al daño y si existe una jerarquía consistente en el deterioro de las capacidades cognitivas evaluadas.

La motivación detrás de este trabajo es doble: por un lado, profundizar en el funcionamiento interno de los LLMs, utilizando la degradación como una herramienta para inferir la función de sus componentes; y por otro, explorar el potencial de estos sistemas como modelos computacionales para la neurociencia, que permitan formular y testear hipótesis sobre la cognición, el lenguaje y sus alteraciones.

Para abordar estos objetivos, el trabajo se estructura de la siguiente manera. El Capítulo 2 detalla el modelo de lenguaje seleccionado y los fundamentos teóricos de su arquitectura, así como los métodos de degradación implementados y el diseño experimental. El Capítulo 3 describe el protocolo de evaluación, incluyendo las tareas propuestas al modelo, el preprocesamiento de los datos y el conjunto de métricas diseñadas para cuantificar el rendimiento y el deterioro. El Capítulo 4 presenta los resultados obtenidos, desde un análisis cualitativo de los fallos hasta un estudio cuantitativo del orden de deterioro de las métricas. Finalmente, el Capítulo 5 discute la interpretación y las implicancias de estos hallazgos, reconoce las limitaciones del estudio y propone futuras líneas de investigación.

## 2. DEGRADACIÓN INDUCIDA EN UN LLM

### 2.1. Modelo de lenguaje seleccionado

Para esta investigación se eligió trabajar con la familia de modelos *Gemma*, desarrollada por Google. La principal ventaja de estos modelos es su naturaleza de código abierto, que garantiza el acceso público a sus pesos internos, un requisito indispensable para aplicar las técnicas de degradación controlada que constituyen el núcleo de este estudio. Además, la familia *Gemma 3* es multimodal (por lo que permite trabajar con imágenes) y se ofrece en múltiples tamaños (1B, 4B, 12B y 27B parámetros) y dos variantes principales: las pre-entrenadas (*PT*), ideales para *fine-tuning*, y las ajustadas por instrucciones (*IT*), optimizadas para seguir consignas complejas.

Dentro de este ecosistema, se seleccionó específicamente el modelo *Gemma 3* de 4 mil millones de parámetros (4B) en su variante ajustada por instrucciones (*IT*). Esta elección se debe a que la versión *IT* es fundamental para evaluar el rendimiento en las tareas de razonamiento y generación de texto diseñadas para la tesis. A su vez, el tamaño de 4B parámetros ofrece un equilibrio ideal entre la capacidad para exhibir comportamientos cognitivos complejos y la eficiencia computacional para la experimentación iterativa en el *hardware* disponible (GPU NVIDIA RTX 3090). El modelo se operó con una precisión de 32 bits (*float32*).

### 2.2. Fundamentos de LLMs

Los Grandes Modelos de Lenguaje (LLM), como la familia de modelos *Gemma*, representan un hito en el campo de la inteligencia artificial y el procesamiento del lenguaje natural. Sin embargo, su complejidad se cimienta sobre principios computacionales que evolucionaron durante las últimas décadas.

La red neuronal artificial, un modelo matemático inspirado en el cerebro humano, consiste en nodos interconectados (neuronas) organizados en capas que aprenden a reconocer patrones complejos, ajustando los pesos (ponderaciones) de sus conexiones a través de un proceso de “aprendizaje” a partir de ejemplos de entrenamiento.

Las primeras arquitecturas para el procesamiento de secuencias —como las Redes Neuronales Recurrentes (RNN) o *LSTM*— procesan el texto de manera secuencial, por lo que tienen dificultad para manejar dependencias contextuales a larga distancia. El mecanismo de atención, y en particular la arquitectura de *Transformer* [18], fue el hito que permitió el procesamiento paralelo de toda la secuencia y así un manejo del contexto drásticamente superior, sentando las bases para la escala y el rendimiento de los LLMs modernos.

Para que una red neuronal o un *Transformer* pueda procesar lenguaje humano, este debe ser convertido previamente a una representación numérica, a partir de un proceso que se realiza en dos etapas: la *tokenización* y la generación de *embeddings*.

1. La ***tokenización*** es el proceso por el cual el texto de entrada se segmenta en unidades más pequeñas (denominadas “*tokens*”), que —según la técnica— pueden corresponder con las palabras o con sub-unidades significativas. Esto último garantiza que haya un vocabulario (o *tokens*) de tamaño fijo (llamado *k*) y la capacidad de manejar

palabras raras o fuera de vocabulario al construirlas a partir de *tokens* conocidos. Por ejemplo, la palabra “inmanejable” podría ser *tokenizada* como [“in”, “manej”, “able”]. El modelo *Gemma 3* cuenta con un vocabulario de  $k = 262\,208$  *tokens*.

2. Una vez que se *tokenizó* el texto, es necesario convertir cada *token* en un vector numérico llamado ***embedding***. La longitud de cada uno de estos vectores o *embeddings* es fija para cada modelo y se suele denominar “dimensión del modelo” ( $d_{\text{model}}$ ). Para esto, almacena internamente una gran tabla de consulta denominada “matriz de *Embeddings*”,  $E \in R^{k \times d_{\text{model}}}$ , que está compuesta por el *embedding* de cada *token* del vocabulario, acomodados en forma de fila. Los valores de esta matriz, que contiene la representación vectorial de cada *token* posible, son ponderaciones que se aprenden durante el entrenamiento y, al igual que los otros parámetros del modelo, quedan fijas al momento de la inferencia (producción de lenguaje). En *Gemma 3*,  $d_{\text{model}} = 2560$ .

Cuando se procesa una secuencia de entrada, el modelo utiliza  $E$  para buscar el vector correspondiente a cada uno de esos *tokens* que la componen. Acomodando el *embedding* de cada *token* en forma de fila, el resultado es la creación de una segunda matriz, la matriz de *embeddings* de entrada de la secuencia,  $X \in R^{n \times d_{\text{model}}}$ , donde  $n$  es la longitud (en *tokens*) de la secuencia. Esta matriz  $X$ , que representa semánticamente la secuencia de entrada completa, es la que finalmente se utiliza como entrada inicial para la primera capa del *Transformer*.

Si bien la arquitectura original de *Transformer* propuesta en 2017 [18] consistía en un *encoder* y un *decoder*, lo más habitual en el momento son arquitecturas de solo *decoder*, que se organiza en un apilamiento de bloques o capas con la misma arquitectura. *Gemma 3*, por ejemplo, es *decoder-only*, y contiene 34 capas.

El componente distintivo en el *decoder* es el mecanismo de auto-atención (*self-attention*), que permite ponderar la importancia relativa de los *tokens* de una secuencia entre sí. Para cada *token*, intenta identificar a qué otras partes del texto debe “prestar atención” para ajustar mejor su propio significado en ese contexto. El ejemplo típico en español es la palabra “banco”, que, según el contexto, se interpretará como un objeto en donde sentarse (“banco de plaza”), como una institución bancaria (“entré al banco”), como una capacidad de sostener algo (“no se lo va a bancar”), como una manera informal rioplatense de apoyar una propuesta (“banco eso”), etc.

El mecanismo de *self-attention* opera a través de tres matrices de pesos que son parte fundamental de los parámetros aprendidos por el modelo para cada capa: la de Consulta o *Query* ( $W_Q \in R^{d_{\text{model}} \times d_Q}$ ), la de Clave o *Key* ( $W_K \in R^{d_{\text{model}} \times d_K}$ ) y la de Valor o *Value* ( $W_V \in R^{d_{\text{model}} \times d_V}$ ). Los valores  $d_Q$ ,  $d_K$  y  $d_V$  son hiperparámetros fijos de diseño de cada modelo (por ejemplo, en *Gemma 3*,  $d_Q = 2048$ ,  $d_K = d_V = 1024$ ) y estas tres matrices son estáticas para un modelo ya entrenado. Para cada secuencia de entrada  $X$ , se generan dinámicamente otras tres matrices específicas para esa secuencia: *Query* ( $Q$ ), *Key* ( $K$ ) y *Value* ( $V$ ), que se obtienen mediante la multiplicación de la matriz de entrada  $X$  por cada una de las matrices de pesos:

$$Q = X \cdot W_Q$$

$$K = X \cdot W_K$$

$$V = X \cdot W_V$$



De esta forma, las matrices  $Q$ ,  $K$  y  $V$  de cada capa contienen las representaciones de la secuencia de entrada proyectadas en los subespacios de consulta, clave y valor, respectivamente. El puntaje de atención se calcula entonces mediante una operación de similitud entre las consultas ( $Q$ ) y las claves ( $K$ ), seguida de una normalización y una ponderación de los valores ( $V$ ). La fórmula canónica es

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

donde la raíz cuadrada de  $d_k$  actúa como un factor de escala para estabilizar el entrenamiento. El resultado es una nueva matriz donde cada vector de *token* ha sido enriquecido con información contextual ponderada del resto de la secuencia.

Un LLM apila entonces múltiples bloques *Transformer*. Cada bloque contiene dos sub-componentes principales:

- **Atención multi-cabeza (*Multi-Head Attention*):** En lugar de realizar la auto-atención una sola vez, se ejecuta en paralelo múltiples veces (cada una es una “cabeza”) con diferentes matrices de proyección ( $Q$ ,  $K$ ,  $V$ ). Esto permite al modelo capturar simultáneamente diferentes tipos de relaciones contextuales (sintácticas, semánticas, etc.).
- **Red *Feed-Forward* (*MLP*):** Una red neuronal densa, también conocida como Perceptrón Multicapa (*MLP*), que procesa la salida de la capa de atención de forma independiente para cada posición de *token*. Esta red típicamente consiste en dos transformaciones lineales con una función de activación no lineal (como *GELU* o *SwiGLU*) entre ellas. Cada una de estas transformaciones lineales se define por sus propias matrices de pesos, que son parámetros clave del modelo y añaden capacidad de representación no lineal.

Además, cada sub-componente está envuelto en una conexión residual y una normalización de capa, técnicas cruciales que previenen la desaparición del gradiente y permiten entrenar modelos con cientos de estas capas apiladas.

Mientras que el mecanismo de atención permite comprender el texto de entrada, la generación de nuevo texto en modelos como *Gemma* se realiza de forma auto-regresiva. Este proceso es un bucle iterativo:

1. El modelo recibe una secuencia de entrada (el *prompt*).
2. Procesa esta secuencia a través de sus capas *Transformer* para generar una representación contextualizada.
3. Una capa final predice una distribución de probabilidad sobre todo el vocabulario para el siguiente *token*. Se modela la probabilidad condicional  $P(t_{i+1}|t_1, t_2, \dots, t_i)$ .
4. Se selecciona un *token* de esta distribución. Este proceso se controla a partir de un parámetro llamado “temperatura”: un valor de temperatura = 0 hace que la selección sea determinística, eligiendo siempre el *token* con la probabilidad más alta. A medida que la temperatura aumenta, se introduce una mayor variabilidad en la selección, lo que permite generar respuestas más diversas y creativas.
5. Este nuevo *token* se anexa a la secuencia de entrada.

6. El proceso se repite desde el paso 2 con la secuencia extendida, generando la respuesta *token* por *token* hasta que se alcanza una condición de parada (e.g., un *token* especial de fin de secuencia o una longitud máxima).

En resumen, el funcionamiento de un LLM moderno es un flujo de transformación de datos: el texto se convierte en *tokens*, los *tokens* en vectores de *embedding* (formando una matriz), esta matriz es procesada por una pila de capas *Transformer* que utilizan la auto-atención para enriquecer la información contextual, y finalmente, un proceso auto-regresivo genera nuevo texto *token* a *token*.

Es importante destacar que todos estos pesos o parámetros del modelo se ajustan o “aprenden” a partir de vastos corpus de texto durante una primera etapa de entrenamiento. En este proceso, conocido como pre-entrenamiento, el modelo aprende las reglas gramaticales, las relaciones semánticas entre palabras y un conocimiento general del mundo, dando lugar a la versión pre-entrenada (PT).

Posteriormente, para mejorar su capacidad de seguir instrucciones y dialogar, el modelo se somete a una segunda fase de post-entrenamiento. Esta comienza con un ajuste fino supervisado (*Supervised Fine-Tuning*), donde se utilizan conjuntos de datos con ejemplos de preguntas y respuestas de alta calidad. A continuación, para refinar aún más su comportamiento, se aplican diversas técnicas de aprendizaje por refuerzo, en particular, para alinear las predicciones del modelo con las preferencias humanas, para potenciar el razonamiento matemático y para mejorar sus habilidades de escritura de código computacional. El resultado de estas fases de ajuste es la versión final del modelo, a veces denominada “ajustada por instrucciones” (IT, por sus siglas en inglés), que es la utilizada en este trabajo.

Para el propósito de esta tesis, el análisis de degradación se centra en el impacto de perturbar tres conjuntos de matrices fundamentales: la matriz de *Embeddings* ( $E$ ), que codifica el conocimiento léxico inicial; las 34 matrices de Valor ( $W_V$ ), una por cada capa, dentro del mecanismo de atención; y las tres matrices de pesos que componen las redes *Feed-Forward* (*MLP*) en cada capa (*up*, *down* y *gate*), siendo todas estas responsables de una parte importante del procesamiento en cada bloque del modelo.

### 2.3. Formas de degradación artificial estudiadas

Para evaluar la robustez y la resiliencia del modelo de red neuronal ante diferentes tipos de daño, se implementó una estrategia de degradación controlada. Esta consiste en la aplicación sistemática de tres métodos de perturbación distintos, cada uno con su parámetro que controla la magnitud, y diseñados para simular una forma específica de deterioro que podría ocurrir en sistemas biológicos o computacionales. La selección de estos métodos permite analizar el comportamiento del modelo no solo ante una pérdida genérica de información, sino ante fallos de naturaleza fundamentalmente diferente, como la pérdida de precisión, la degradación o la pérdida estructural. A continuación, se describe en detalle cada uno de los métodos implementados.

#### 2.3.1. Ruido multiplicativo Gaussiano (*mult\_gauss*)

Este método simula una pérdida progresiva de la precisión sináptica o la presencia de ruido en las vías de señalización neuronal. El procedimiento consiste en modificar los pesos (parámetros) de los tensores seleccionados del modelo. Cada peso individual es

multiplicado por un factor de escala aleatorio, que se extrae de una distribución Normal (Gaussiana) con una media de 1 y una desviación estándar, denotada como  $\sigma$ , que actúa como el parámetro de intensidad de la degradación. La fórmula de actualización para un peso  $w$  es

$$w' = w \cdot N(1, \sigma),$$

donde  $w'$  es el nuevo valor del peso y  $N(1, \sigma)$  se refiere a la realización de una variable aleatoria gaussiana. Al utilizar una media de 1 (el neutro multiplicativo), la modificación no introduce un sesgo sistemático, sino que el resultado fluctúa aleatoriamente alrededor del valor original del peso. El parámetro  $\sigma$  controla la degradación: un valor de  $\sigma = 0$  implica que todos los factores de escala son exactamente 1, sin causar ningún cambio; a medida que  $\sigma$  aumenta, la distribución de los factores de escala se ensancha, provocando que los pesos se alejen cada vez más de su valor original. Este tipo de perturbación es análogo a un deterioro funcional donde la conexión sináptica se vuelve menos fiable, pero sigue presente. Se eligió multiplicación en lugar de suma, para que el efecto sea parejo en los distintos tensores, independientemente de la escala de los valores presentes originalmente en cada uno.

### 2.3.2. Ablación de pesos (ablation)

La ablación es un método de degradación estructural que simula la muerte neuronal o la pérdida completa de conexiones sinápticas. Este método anula selectivamente un porcentaje de los pesos del modelo. Para cada tensor de pesos a modificar, se genera una máscara binaria de la misma dimensionalidad. Cada elemento de esta máscara tiene una probabilidad  $p$  de ser 0 y una probabilidad  $1 - p$  de ser 1. Posteriormente, el tensor de pesos original se multiplica elemento a elemento por esta máscara:

$$W' = W \odot M,$$

donde  $W$  es el tensor de pesos original,  $M$  es la máscara binaria y  $\odot$  denota la multiplicación de Hadamard (elemento a elemento). Esta técnica representa una forma más drástica y categórica de daño. A diferencia del ruido gaussiano, que modifica sutilmente el valor de una conexión, la ablación la elimina por completo (la “apaga”) al llevar su peso a cero. El parámetro  $p$  controla directamente el nivel de deterioro estructural; por ejemplo, un valor de  $p = 0.1$  significa que aproximadamente el 10 % de las conexiones en las capas afectadas son eliminadas de forma aleatoria. Es un modelo que busca estudiar la pérdida irreparable de componentes.

### 2.3.3. Cuantización uniforme (uni\_quant)

La cuantización es un método de compresión que degrada el modelo al reducir la precisión de sus representaciones numéricas, emulando limitaciones ‘computacionales’ o una pérdida en la riqueza de la información interna. El proceso reduce el conjunto ‘continuo’ de valores que puede adoptar cada peso a un conjunto finito de  $N_q$  niveles discretos. Para cada tensor:

1. Se identifican los valores mínimo ( $w_{min}$ ) y máximo ( $w_{max}$ ) de sus pesos.
2. El rango  $[w_{min}, w_{max}]$  se divide en  $N_q$  niveles equiespaciados.

3. Cada peso original del tensor es reemplazado por el valor del nivel cuantizado más cercano.

Si bien los modelos digitales ya operan con una cuantización implícita (e.g., punto flotante de 32 bits), este procedimiento fuerza una reducción mucho más agresiva de la resolución. El parámetro  $N_q$  determina la severidad de la compresión: un valor alto (e.g., 256) permite una representación relativamente fiel, mientras que un valor muy bajo (e.g., 2) reduce cada peso a solo uno de esos valores posibles, forzando un redondeo extremo. Esta técnica busca simular un deterioro en la incapacidad del sistema para mantener representaciones internas sutiles, complejas y matizadas. En este caso, el parámetro que controla el daño es la cantidad  $N_q$  de niveles discretos permitidos por tensor, por lo que el daño aumenta al disminuir este valor.

## 2.4. Diseño experimental

Los experimentos se ejecutaron en un entorno local controlado, utilizando una estación de trabajo equipada con GPU NVIDIA RTX 3090. Para garantizar la consistencia en la generación de texto a lo largo de todos los experimentos, se establecieron los siguientes parámetros de inferencia fijos para el modelo:

- Temperatura: 1.0
- Top-k: 64
- Top-p: 0.95
- Penalización por repetición: 1.0
- Máximo de *tokens* nuevos: 350

El núcleo del experimento consiste en un protocolo sistemático que se repite para cada combinación de método de degradación (ej. `mult_gauss`, `ablation`, etc.) y grupo de parámetros objetivo (matrices  $W_V$  de atención, matrices del MLP y matriz  $E$  de *Embeddings*).

1. **Definición del rango de degradación:** Para cada método, se establece un rango de intensidad de daño, definiendo un valor mínimo, un máximo y entre 5 y 15 pasos intermedios. Esto permite observar la evolución del rendimiento del modelo desde una degradación leve hasta una severa.
2. **Replicación y control de la aleatoriedad:** Con el fin de obtener resultados estadísticamente robustos, para cada nivel de intensidad se generan 10 instancias de modelo degradado. Esto se logra utilizando 10 semillas de aleatoriedad predefinidas y fijas (en este caso se usaron los enteros del 42 al 51). Una semilla de aleatoriedad es un número que se utiliza para inicializar un generador de números pseudoaleatorios; al fijarla, se garantiza que la secuencia de operaciones que dependen del azar sea siempre la misma, permitiendo la reproducibilidad de los experimentos. De este modo, se obtienen 10 réplicas independientes para cada punto experimental (en los métodos aleatorios). El uso de una semilla constante para cada una de las 10 trayectorias tiene una implicación metodológica clave:

- Para los métodos estocásticos (`mult_gauss`, `ablation`), el uso de una misma semilla a través de los distintos niveles de intensidad asegura una progresión de daño correlacionada y monótona. Por ejemplo, si un peso específico es puesto a cero durante una ablación del 10 %, se garantiza que ese mismo peso también será cero cuando la intensidad aumente al 20 % bajo la misma semilla. Esto crea 10 trayectorias de degradación consistentes donde el daño es acumulativo.
- En contraste, el método de cuantización uniforme (`uni_quant`) es determinista. Su resultado depende únicamente de los valores mínimo y máximo del tensor de pesos, no de una semilla aleatoria. Por lo tanto, para un nivel de cuantización dado, solo se genera un modelo perturbado.

**3. Evaluación del rendimiento:** Cada una de las instancias de modelo degradado generadas (una por cada nivel de intensidad y cada una de las 10 semillas) es sometida al conjunto completo de tareas y *prompts* definidos para la evaluación.

La variabilidad inherente a los procesos estocásticos se gestiona mediante la agregación de los resultados. Para cada nivel de intensidad de un método de degradación, las métricas de rendimiento obtenidas de las 10 instancias de modelo (cada una con su semilla) se promedian. Este procedimiento permite estimar el efecto típico de dicho nivel de degradación, mitigando la influencia del azar de una única ejecución y proporcionando una medida más robusta y generalizable del impacto en el modelo de determinada magnitud de daño. La fijación de las semillas en cada paso garantiza, a su vez, la replicabilidad total del estudio.



### 3. EVALUACIÓN DEL DETERIORO Y SUS EFECTOS

#### 3.1. Tareas pedidas al modelo y *prompts* utilizados

Para evaluar el rendimiento de los modelos, se diseñó un conjunto de tareas agrupadas en cuatro categorías temáticas: generación de discurso espontáneo a través de relatos de sueños, descripción de escenas complejas mediante la tarea *Cookie Theft*, razonamiento lógico-matemático, y tareas de comprensión y razonamiento verbal.

Para las tareas de Lengua y Matemática, si bien existen esquemas de evaluación estandarizados y validados (como tests de coeficiente intelectual o *benchmarks* específicos para LLMs), se optó por la elaboración de preguntas propias para esta fase exploratoria inicial, inspirada en los tests utilizados regularmente (como en [1]). Esta decisión metodológica responde a dos criterios fundamentales. Primero, el uso de preguntas novedosas aumenta la probabilidad de que el modelo no haya sido entrenado específicamente para ellas, a diferencia de lo que podría ocurrir con *benchmarks* públicos y ampliamente conocidos, especialmente considerando el reciente lanzamiento del modelo utilizado. Segundo, el diseño de las tareas se ajustó a los tiempos y recursos disponibles para esta investigación.

En todos los casos, las consignas fueron formuladas con un nivel de complejidad accesible para un estudiante de los primeros años de educación secundaria. Se utilizó el idioma inglés para todas las interacciones, ya que es el lenguaje principal en el que fue entrenado el modelo de Google. A continuación, se presentan las listas de todas las preguntas realizadas a los modelos. A cada instancia del modelo se le presentó una única pregunta por vez.

#### Tareas de razonamiento lógico-matemático

Se presentó una serie de ejercicios matemáticos con distintos niveles de dificultad, diseñados para evaluar desde la aritmética básica hasta la resolución de problemas algebraicos y de proporciones:

- “What is the result of  $3 + 48 / 6 - 7$ ?”
- “What number completes the sequence: 93, 94, 95, \_\_, 97?”
- “How many hours are there in four and a half days?”
- “If  $3/4$  of a number is 27, what is the number?”
- “What number comes next in the sequence: 120, 113, 106, \_\_?”
- “What number is missing in the arithmetic sequence: 99, \_\_, 71, 57?”
- “What number comes next in the sequence: 1.2, 4.2, 7.2, \_\_?”
- “What number makes the equation true:  $3(x - 5) = 4x + 7$ ?”
- “A car travels 150 km in 2 hours and 30 minutes. What is its average speed in km/h?”

- “What number comes next in the sequence: 4, 5, 7, 10, 14, \_\_?”
- “What is the smallest positive integer greater than 100 that is divisible by both 6 and 15?”
- “A rectangle has a length that is 5 times its width. If its perimeter is 72 cm, what is its area?”
- “A water tank is  $\frac{4}{7}$  full. After removing 36 litres, it is  $\frac{2}{7}$  full. What is the full capacity of the tank in litres?”
- “In a class of 40 students, 60 % are girls. If 50 % of the boys and 25 % of the girls wear glasses, how many students wear glasses?”
- “The sum of three consecutive odd numbers is 111. What is the smallest of these numbers?”

### Tareas de comprensión y razonamiento verbal

Se solicitó a los modelos que resolvieran problemas de lógica verbal y semántica, justificando sus respuestas. Estas tareas requerían identificar patrones, encontrar anomalías y determinar relaciones de significado:

- “Which word comes next: January, March, May, \_\_\_?”
- “Which word does not belong: carrot, apple, banana, grape”
- “Which word does not belong: spoon, tea, coffee, juice”
- “Find the odd one out: walk, blue, run, jump”
- “Find the odd one out: red, green, square, blue”
- “Select the word that is different: star, moon, chair, sun”
- “Which word does not belong: prism, sphere, pyramid, circle”
- “Choose the word that does not fit: cello, violin, hammer, trumpet”
- “Which word is closest in meaning to ‘enormous’? huge, small, narrow, little”
- “Which of these is different: shout, orbit, spin, rotate”
- “Find the antonym of ‘scarce’ among these: abundant, limited, rare, minimal”

### Tareas de relato de sueño

Se instruyó a los modelos que generaran el relato de un sueño. Para este objetivo, se utilizaron las siguientes variaciones de consignas (*prompts*), con el fin de observar la consistencia en la generación ante estímulos semánticamente similares:

- “Just narrate a dream with as much detail as you can.”
- “Just narrate a dream.”



- “Just describe a dream.”
- “Share a dream.”
- “Tell a dream in as much detail as you can.”
- “Tell what happened in a dream, with detail.”

### Tareas de descripción de imagen (*Cookie Theft Task*)

Se presentó a los modelos la imagen del *Cookie Theft Task* y se les solicitó que describieran la escena con las siguientes variaciones de consignas:

- “Tell me everything you see going on in this picture.”
- “Tell me everything you see going on in this picture. Describe everything that is happening.”
- “Describe everything happening in this picture.”
- “Tell me what you see going on in this scene.”
- “List everything you see in the image.”
- “Describe in detail everything that is happening in the image.”
- “Describe this picture in as much detail as possible.”
- “Provide a complete description of the image.”
- “Tell me everything you see happening within this scene.”
- “Give a thorough account of the visual information presented in the drawing.”
- “Recount the entire contents of the image.”
- “Offer a comprehensive summary of what is depicted in this illustration.”
- “Present an exhaustive description of the picture’s content.”
- “Please detail all the elements and activities shown in the picture.”
- “Articulate everything that is visibly occurring in this drawing.”
- “Describe this picture.”
- “Describe what you see in this image.”
- “Provide a detailed description of the scene shown.”
- “Give a comprehensive account of everything depicted in this picture.”
- “Describe the contents of this image in detail.”

### 3.2. Preprocesamiento de textos

Se implementó una secuencia para preparar para el análisis a los textos generados por cada LLM. Un efecto observado en las generaciones de texto de los LLMs deteriorados es la aparición de bucles repetitivos, donde una secuencia de muy pocas palabras se reitera de forma consecutiva hasta el final del texto. Para mitigar el impacto de este ruido, se desarrolló un procedimiento para identificar y truncar dichas repeticiones en todos los casos.

En el caso de análisis de relatos de sueño, además, el siguiente paso consistió en su segmentación en oraciones individuales, utilizando la biblioteca `spaCy` en `Python`. Para facilitar y mejorar este proceso, previamente se aplicó una normalización del texto que incluyó la unificación de espacios en blanco y la expansión de contracciones comunes en inglés (como “*don’t*” a “*do not*”).

Luego, se realizó un análisis léxico de cada oración. Primero, se identificó la categoría gramatical de cada palabra (sustantivo, verbo, adjetivo, adverbio, etc.), en un proceso conocido como etiquetado morfosintáctico (*Part-of-Speech Tagging*, o *POS Tag*). Luego, cada una fue reducida a su forma canónica o ‘lema’, que utiliza la categoría gramatical (*POS tag*) previamente identificada para realizar una conversión morfológica precisa: por ejemplo, el verbo “*saw*” se lematiza como “*see*”, y el sustantivo “*feet*” como “*foot*”. Para cada oración, se mantuvo una versión con *stopwords* o palabras vacías (términos de alta frecuencia y bajo contenido semántico como “*the*”, “*a*”, “*in*”), y una versión con estos eliminados.

### 3.3. Métricas discursivas y de performance

Para el análisis cuantitativo de los textos generados por cada modelo en las distintas tareas, se diseñó una estrategia que combina dos grandes enfoques:

- **Métricas lingüísticas computacionales:** Se calcularon diversos indicadores objetivos del discurso mediante la implementación de *scripts* en `Python` y el uso de librerías especializadas en procesamiento de lenguaje natural (PLN). Estas métricas se enfocaron en cuantificar aspectos como la coherencia semántica, la diversidad y la complejidad léxica.
- **Evaluación mediante LLM externo:** Se empleó un Gran Modelo de Lenguaje (LLM) como evaluador externo para obtener métricas de desempeño y calidad. Para ello, se adoptó una metodología de evaluación conocida en la literatura como “LLM-como-Juez” (*LLM-as-a-Judge*) [21]. Este paradigma constituye una alternativa costo-efectiva a la evaluación humana, que si bien es el *gold standard*, en este caso resultaba inviable para el volumen de datos generados. Específicamente, se utilizó la API de *GPT-4o-mini* (versión *gpt-4o-mini-2024-07-18*), de la compañía *OpenAI*.

El proceso se realizó en un paradigma *zero-shot* (es decir, sin provisión de ejemplos modelo), donde para cada evaluación, este LLM “juez” recibió un *prompt* que contenía la consigna original de la tarea, la respuesta correcta esperada (en el caso de tareas de Lengua y Matemática) y una sola respuesta generada por un modelo a evaluar. Se le proveyeron instrucciones detalladas sobre su rol como evaluador imparcial, solicitándole que puntuara diversas dimensiones en escalas predefinidas o

que emitiera juicios booleanos (Verdadero/Falso). Para garantizar la fiabilidad, consistencia y reproducibilidad, todas las evaluaciones se realizaron fijando el parámetro de temperatura en 0, asegurando así respuestas deterministas, y exigiendo una salida estructurada en formato JSON. A modo de ejemplo, en el Apéndice se detallan los *prompts* y uno de los esquemas JSON utilizados para la evaluación de la tarea *Cookie Theft*.

A continuación, se describen las métricas específicas utilizadas, agrupadas según la tarea en la que fueron aplicadas.

### Métricas generales (aplicadas en todas las tareas)

- **fue\_recortado** (Booleano): Indica si el texto generado fue truncado debido a la detección de un patrón de repetición infinita. Este es un indicador clave de un fallo catastrófico en la generación.
- **tokens** (Entero): Cantidad de *tokens* en la respuesta original del modelo. El máximo de *tokens* permitidos para la generación fue de 350.
- **grammatical\_correctness** (Escala 0-5): Nivel de corrección gramatical y sintáctica del texto.

### Métricas para tareas de Lengua y de Matemática

- **answer\_correct** (Booleano): Determina si la respuesta final proporcionada por el modelo (explícita o implícita) coincide con la respuesta correcta esperada, sin distinción de mayúsculas y minúsculas. En las tareas de Lengua, se evalúa además si la justificación es lógicamente válida.
- **argument\_coherence** (Escala 0-5): Calidad y solidez lógica de la argumentación o explicación presentada, donde 0 indica un razonamiento incoherente o contradictorio y 5 una explicación clara y lógicamente consistente.
- **conciseness\_level** (Escala 0-5): Grado en que la respuesta es directa y evita información superflua, donde 0 es excesivamente largo o verboso y 5 es breve y al punto, sin perder información importante.

### Métricas específicas para tareas de Matemática

- **knows\_method** (Escala 0-5): Evalúa si el modelo demuestra identificar el método o procedimiento matemático para resolver el problema, donde 0 indica que no se identifica un método válido y 5 que se identifica un método completamente válido, incluso si no se aplica correctamente.
- **applies\_method** (Escala 0-5): Evalúa si el modelo ejecuta correctamente el método identificado para llegar a la solución, donde 0 indica que no aplica correctamente un método válido y 5 que lo aplica y ejecuta de forma impecable.

## Métricas específicas para tareas de relato de sueños y de descripción de imágenes

Métricas de coherencia y estructura narrativa (evaluadas por LLM externo)

- **derailment\_score** (Escala 0-5): Mide la frecuencia de saltos abruptos o cambios de tema que rompen el hilo narrativo, donde 0 indica ausencia de este problema y 5 una severidad muy alta.
- **tangentiality\_score** (Escala 0-5): Evalúa la tendencia del discurso a desviarse hacia detalles irrelevantes o asociaciones indirectas, donde 0 indica ausencia y 5 una severidad muy alta.
- **narrative\_coherence** (Escala 0-5): Valora la cohesión y la estructura lógica general del relato, donde 0 es muy pobre y 5 es excelente.

## Métricas específicas para tareas de relato de sueños

Métricas afectivas y de personalidad (evaluadas por LLM externo)

- **valence** (Escala -3 a 3): Tono emocional del relato, desde muy negativo (-3) a muy positivo (3).
- **arousal** (Escala 1 a 5): Nivel de activación o intensidad emocional del contenido, desde muy bajo (1) a muy alto (5).
- **openness** (Escala 0-5): Rasgo de personalidad (*Big Five*) que refleja imaginación y curiosidad (0=muy bajo, 5=muy alto).
- **conscientiousness** (Escala 0-5): Rasgo de personalidad (*Big Five*) que indica organización y disciplina (0=muy bajo, 5=muy alto).
- **extraversion** (Escala 0-5): Rasgo de personalidad (*Big Five*) asociado a la sociabilidad y la expresividad (0=muy bajo, 5=muy alto).
- **agreeableness** (Escala 0-5): Rasgo de personalidad (*Big Five*) que refleja cooperación y empatía (0=muy bajo, 5=muy alto).
- **neuroticism** (Escala 0-5): Rasgo de personalidad (*Big Five*) relacionado con la inestabilidad emocional (0=muy bajo, 5=muy alto).

Métricas lingüísticas computacionales

- **mean\_sim\_consec\_nomic** (Similitud media entre oraciones consecutivas) (Rango teórico -1 a 1): Mide la coherencia semántica local, reflejando la fluidez temática. Se calcula como la media de las similitudes coseno entre los *embeddings* de oraciones consecutivas. Estos *embeddings* o vectores numéricos son generados por el modelo de código abierto *nomic-embed-text-v1*, que transforma el texto en un espacio vectorial donde la proximidad semántica se corresponde con la distancia angular entre vectores. Valores cercanos a 1 indican alta continuidad temática, mientras que valores cercanos a 0 sugieren transiciones abruptas.

- **mean\_wordnet\_depth** (Profundidad léxica) (Rango teórico 0 a ~25): Es la media de la profundidad de los lemas en la jerarquía de *WordNet*. *WordNet* es una gran base de datos léxica del inglés donde las palabras se agrupan en conjuntos de sinónimos cognitivos (*synsets*), interconectados por relaciones semánticas. La “profundidad” de un término indica su especificidad en esta estructura jerárquica. Una mayor profundidad indica términos más especializados (e.g., *poodle*), mientras que una profundidad baja sugiere conceptos generales (e.g., *animal*).
- **mean\_zipf\_score** (Frecuencia léxica) (Rango teórico 0 a ~8): Calcula el promedio de los *scores Zipf* de cada lema. Este puntaje se deriva de la Ley de Zipf, un principio empírico que postula que la frecuencia de una palabra es inversamente proporcional a su rango en una tabla de frecuencias. Se utiliza la librería **wordfreq**, que provee valores pre-calculados a partir de grandes corpus (como *Wikipedia* y *Google Books*) para obtener una medida estandarizada de cuán común es un término. Valores altos (hasta 7–8) corresponden a lemas muy comunes como “the” o “and”, mientras que valores bajos (cerca de 0), a términos extremadamente raros.
- **lexical\_density** (Densidad léxica) (Rango teórico 0 a 1): Es el porcentaje de lemas de contenido (excluyendo *stopwords* como “the”, “and”, “is”) sobre el total de lemas en el texto; refleja cuán “informativo” o “sustantivo” es el lenguaje.
- **type\_lemma\_ratio** (Diversidad léxica) (Rango teórico 0 a 1): Proporción entre la cantidad de lemas distintos y el número total de lemas (*tokens*) en el texto, excluyendo *stopwords*. Valores más altos indican mayor variedad léxica. El mínimo teórico tiende a 0 (si todo el texto es la repetición de un mismo lema) y el máximo es 1 (si todos los lemas son diferentes).

### Métricas específicas para tareas de descripción de imagen

Métricas de unidades de contenido (evaluadas por LLM externo)

- **descriptions** (Recuento de Unidades de Contenido) (Rango teórico 0 a 22): Cantidad de Unidades de Contenido (UCs) detectadas en la descripción de la imagen. Para estandarizar esta métrica, se utilizó como referencia la lista de 22 preguntas dicotómicas (sí/no) propuestas en el apéndice del trabajo de Kleiman (2025) [9]. Dicha lista desglosa la escena en sus elementos y acciones fundamentales, permitiendo un recuento objetivo del nivel de detalle del relato.

Finalmente, para facilitar la comparación visual en gráficos, un subconjunto de estas métricas fue normalizado utilizando la técnica de escalado Mín-Máx. Este método reescala cada serie de datos a un rango común (generalmente de 0 a 1), basándose en los valores mínimo y máximo observados en dicha serie. Aquellas consideradas ‘negativas’ o ‘indeseadas’ (como tangencialidad y descarrilamiento), además son invertidas, de forma tal que el deterioro tienda a disminuir todas las métricas que tienen un sentido preferido.

### 3.4. Criterio de deterioro significativo

Se estableció un criterio objetivo para cuantificar y comparar la progresión del daño a través de las distintas condiciones experimentales, definiendo puntos de referencia equivalentes en la trayectoria del deterioro. Con este fin, se definió una métrica para identificar el

umbral en el que una alteración en el rendimiento se considera significativa, independientemente de la forma de degradación (ruido gaussiano, ablación, cuantización), su localización (atención, MLP, *Embeddings*) o las unidades de la métrica de rendimiento.

El criterio se basa en el tamaño del efecto de la diferencia entre el rendimiento del modelo en un estado de degradación y su rendimiento en el estado inicial (sin daño). Para ello, se utiliza la  $d$  de Cohen, que cuantifica la magnitud de la diferencia entre dos grupos. Se estableció como umbral de deterioro significativo el nivel de degradación en el cual la  $d$  de Cohen, al comparar el grupo de resultados del modelo dañado con el del modelo intacto, alcanza o supera por primera vez el valor de 0.8. Para estimar el valor exacto de este parámetro entre dos puntos de medición discretos, se aplica interpolación lineal.

La elección de esta medida es fundamental: A diferencia de las pruebas de significancia estadística, donde diferencias triviales pueden volverse significativas simplemente al aumentar el tamaño de la muestra, la  $d$  de Cohen es independiente del número de simulaciones. Esto enfoca el análisis en cambios de gran magnitud, ya que un valor de  $d \geq 0.8$  se interpreta convencionalmente como un tamaño del efecto grande, lo que garantiza una diferencia sustancial y no ambigua entre ambos grupos.

La  $d$  de Cohen ( $d$ ), que es la diferencia entre las medias de los dos grupos, estandarizada por la desviación estándar agrupada, se calcula como

$$d = \frac{M_1 - M_2}{SD_{pooled}},$$

donde  $M_1$  y  $M_2$  son las medias de los dos grupos y  $SD_{pooled}$  es la desviación estándar agrupada. Esta última es un promedio ponderado de las desviaciones estándar de los dos grupos (el grupo sin degradar, grupo 1, y el grupo en un nivel de degradación específico, grupo 2), que se define como

$$SD_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}},$$

donde  $n_1$  y  $n_2$  son los tamaños de las muestras de cada grupo, y  $s_1$  y  $s_2$  son sus respectivas desviaciones estándar. Este criterio permite, por lo tanto, determinar el punto de inflexión en el que cada métrica sufre un deterioro sustancial.

A partir de estos umbrales, es posible construir un ranking de vulnerabilidad que ordena las métricas desde la más a la menos frágil ante un tipo de daño específico, considerando en cada una el valor del parámetro de degradación para el cual alcanza por primera vez el umbral de deterioro significativo.

Adicionalmente, para validar que las métricas de rendimiento seleccionadas capturan de manera efectiva una característica del declive del modelo, se evaluó la existencia de una correlación monótonica entre el parámetro de degradación y la media de los valores normalizados de cada métrica. Para ello, se utilizó el coeficiente de correlación de Spearman ( $\rho$ ). La hipótesis nula ( $H_0$ ) de este test es que no existe correlación entre las dos variables ( $\rho = 0$ ). Dado que se espera que el rendimiento disminuya a medida que aumenta el daño, se realizó un test unilateral para probar la hipótesis alternativa ( $H_1$ ) de una correlación negativa ( $\rho < 0$ ). El propósito de este análisis es confirmar la consistencia de las métricas, asegurando que su respuesta al daño es sistemática. Cabe aclarar que, si bien una correlación monótonica confirma que la medición no es puramente aleatoria, la ausencia de esta no es concluyente, ya que un patrón de deterioro más complejo (no-monótono) podría no ser detectado por este test.

## 4. RESULTADOS

### 4.1. Análisis preliminar

Como ejemplo de análisis preliminar y cualitativo, se muestran a continuación las respuestas obtenidas para la pregunta “What number completes the sequence: 93, 94, 95, \_\_, 97?”, bajo la modalidad de multiplicación gaussiana en la matriz  $W_V$  de atención, para valores crecientes del parámetro de deterioro ( $\sigma$ ), en todos los casos utilizando la semilla `random_seed=42`:

#### Sin deterioro ( $\sigma = 0$ )

The missing number is 96. The sequence is simply consecutive integers.

La respuesta es correcta y, dado que se trata de una pregunta sencilla, la justificación también es muy breve.

#### Con $\sigma = 0.6$

The sequence is increasing by 1.

- $93 + 1 = 94$
- $94 + 1 = 95$
- $95 + 2 = 97$

Therefore, the missing number is 96.  
The sequence is 93, 94, 95, 96, 97.

En este caso, el *output* obtenido pierde concisión respecto al obtenido con el modelo sin deteriorar, aunque la respuesta sigue siendo breve y correcta.

**Con**  $\sigma = 1.0$

The sequence is 93, 94, 95, 97. The pattern is that we are adding 1 to the previous number.

- $93 + 1 = 94$
- $94 + 1 = 95$
- $95 + 2 = 97$

So the next number in the sequence is 96.

Therefore, the answer is 96.

Here's the reasoning:

1. We are looking for the next number in the sequence.
2. The sequence is number values.
3. The numbers are consecutive integers starting from 93.
4. The next number in the sequence is 96.

To find the next number, we determine the difference between consecutive terms.

- $95 - 94 = 1$
- $96 - 95 = 1$

Therefore, 96 is the next number in the sequence.

Final Answer: 96

La respuesta pierde concisión y puede contener imprecisiones o errores leves. Aun así, se identifica que el modelo parece comprender la pregunta que se le hizo, conoce un método para resolverla y lo aplica correctamente, obteniendo también una respuesta final correcta.



Con  $\sigma = 1.1$

The sequence is 93, 94, 95, 96, 97. The pattern is that the numbers increase by 1, 2, 3, 4, 5 as the numbers increase.

The expected sequence is: 93, 94, 95, 96, 97. The next number in the sequence is 96. Here's how the sequence works:

- The number in the sequence is 95.
- The next number in the sequence is 96.
- The number of steps between numbers in the sequence are 1, 2, 3, 4, 5.

The sequence of numbers are: 93, 94, 95, 96, 97. The difference between consecutive terms is 1, 2, 3, 4, 5. The next number is 97.

The difference between consecutive numbers in the sequence is 1. The sequence is 95.

So, the sequence is: 93, 94, 95, 96, 97. The next expected number is 98. The pattern is:

- 95, 97.

The general form of the expected number is: If the number of steps between the numbers is 1, the sequence is 98.

The number of the sequence is 98. The list of steps is:

- 93 is the number in the sequence.
- The number of steps between them is 1

En este caso, se identifica que el modelo parece haber comprendido la pregunta que se le hizo y de alguna manera conoce cuál es el método para resolverla, lo cual comienza por identificar el patrón en la sucesión. Sin embargo, no logra llevarlo a cabo correctamente: Si bien todavía produce oraciones en inglés coherentes, la respuesta en general pierde estructura y no alcanza una respuesta final correcta.

**Con  $\sigma = 1.3$** 

Let's analyze the pattern in the number sequence 93, 94, 95, and 96. The number 95 is the expected output, and let's determine the pattern. The sequence is: 93, 94, 95, 96, 97, ... and 98, street, 99, etc. are expected to be the sequence. Given the numbers, the number 95 is in the sequence. The number 95 is the number in the number sequence. The number 94 is the number in the number sequence. The numbers 95, 96 and 97 are associated with this particular number sequence. The following algorithm is used to determine the number of possible number of number in this case. The number 94 is in the number sequence. The number 94 is the first number in the sub-algorithm for the number of type expected, which are number of the number of the number in the number sequence. The number 95 is the number in the number sequence. The number of repetitions of the number associated with the algorithm is a number associated with the function being used. The number of numbers that are in number series are already known. The number de number of functions in the number of number sequences.

The number 94 is a number associated with the number of functions in the algorithm. The number of numbers in the algorithm are already registered in the number of lines of code, de number of tables and data,

Con este nivel de deterioro, el modelo aun parece haber comprendido la consigna y también comprender que debe hallar un patrón en la sucesión para responder correctamente. Sigue habiendo oraciones en inglés, con gramática correcta en algunos casos y, aunque pierden coherencia, se siguen refiriendo de alguna manera al problema. No se alcanzan argumentos o razonamientos coherentes, tampoco una respuesta correcta.

**Con  $\sigma = 1.4$** 

Okay, let's analyze the sequence. The sequence is 95, 96, 98, and the game, and the following number patterns.

The numbers 95, 97, and 99 will be the numbers on the sequence.

Let's determine the specific number that will occur.

The general pattern will have to do with the starting number. Looking at the sequence of numbers, I can see that the number will be 97.

If the number is 98, the sequence will be 99, 97, 97, 98, Age, Primes, and unpredicted numbers, the function will not be in a particular sequence.

The number 9, 97, is the corresponding possible hit count. If the number is 9, the function is completed.

If the sequence is not observed, la fonction ne sera pas degressive. (Drift, par exemple, prend une nouvelle fonction et un double-me-surface, dans la fonction. (Drift, la fonction sur, la fonction, la fonction, et le 9.

(Source: ccm, la fonction "Grand Terre, par exemple, la fonction (3, fonction incrémente, la fonction. - le fonction). (1/375, la fonction ocuse, dans la fonction. (3, le fonction. météorologique fonction les intervalles). (I, le fonction. (3, le fonction. (1/3. (Drift. - le fonction. (1. les fonction. (3. l'information. et (1' identifiant

Las primeras oraciones producidas en la respuesta se refieren aun al problema, de

manera generalmente coherente y comprensible, aunque con ciertos saltos abruptos (como “and the game”), para luego volver al problema planteado. Sigue utilizando oraciones en inglés, con gramática correcta en la mayoría de los casos, aunque sin obtener la respuesta correcta. Luego de aproximadamente 10 oraciones, comienza a producir texto en otro idioma, que luego se convierte en lo que pareciera ser palabras o ruido casi aleatorio.

**Con  $\sigma = 1.5$**

Okay, let's analyze the provided problem. Since the sequence is always moving upwards, this looks like a good time for me to provide you with a solution.  
 The sequence will likely be a whole number.  
 The pattern will be a number.  
 In addition, I am very good at this.  
 Looking for a pattern.  
 I'm a great looking for the number.  
 if you are asking at what point, you are looking for a non-sequential, discrete sequence, I will proceed in a bit.  
 I' \$relative \backslash incorrect \backslash finished. [...]

Al comienzo de la respuesta se refiere al problema planteado, aunque de manera mucho menos específica: identifica que se trata de un entero, y que debe hallar un patrón, pero no lo identifica. Al igual que en el ejemplo anterior, luego de aproximadamente 9 oraciones deja de producir lenguaje comprensible, y procede a emitir *output* aleatorio (que, en este caso, se omite por brevedad).

**Con  $\sigma = 1.6$**

Okay, let's analyze the sequence of numbers produced by the output.  
 The sequence of numbers eshion-ed met based on the following assumptions: 1. 2. 3. 4jostred, nagrara- 4. 3. 5. \$a, an-chishon the sequence that starts with igh, \$d, a. s. \$ 0. \$ dan \$ \_ . \$ not mentioned in the set of random digits that we provide, g, e. g. le us provide, ej one the e done dan gedent based game ogottjjs (becemerped in a den' c on er the sequence. eS. e-a corner daon the To resutance code. over grid. le blar e. oin i-t of correspond ngedon-end erott in a standar chot se (bds (Er's- aedit dan' e cheg- \_ -etje- er det- thas onideeper- o-a- (Jett ng unerior- (editor;do- ehi- (1) o- e- oget- tery- med- dre o- o- pro otio (J. fror sodome- ot- t- caterryingetrove- orangtietegjene- (pere lo o- ot- espeti- e-So- utet storeetlandis- ordetingin fact tors- Stert- mong- 6. jinder. ad- jw- er attr- support-

Con este nivel de deterioro, parece comprender la consigna pero solo logra producir una única oración coherente, para luego pasar a generar texto casi aleatorio.

**Con  $\sigma = 1.7$**

Con este nivel de deterioro, el modelo ya solo produce una primera palabra en inglés (“Okay”), para luego comenzar a emitir en otro alfabeto (hindi).

Dado este análisis previo, se decide por lo tanto analizar los *outputs* generados por los modelos sometidos a multiplicación gaussiana en la matriz  $W_V$  de atención, desde  $\sigma = 0$

hasta  $\sigma = 1.4$ , en el cual todavía produce cierto texto coherente (aunque, los resultados de estlos últimos deterioros deben ser tomados con precaución). Este es un valor que se mantiene similar para las distintas consignas y semillas utilizadas.

Vale aclarar que se considera el primer valor en el que el deterioro ya no permite generar texto en inglés hasta el final, para asegurarse de abarcar todos los valores en los que todavía sí lo genera (si se tomara como final  $\sigma = 1.3$ , se podría eventualmente estar ignorando un *output* interesante entre 1.3 y 1.4).

De igual forma se procedió con las otras combinaciones de deterioro y lugar de deterioro, para determinar los siguientes valores máximos de deterioro analizado en cada caso.

La degradación en la matriz de MLP presenta otras características cualitativas. Se muestran algunos ejemplos a continuación:

#### Con $\sigma = 0.3$ en matriz de MLP

93, 94, 95, \*\*96\*\*, 97  
The sequence is consecutive numbers. The missing number is 96.

Hasta aquí, la respuesta del modelo sigue siendo correcta, coherente y concisa.

#### Con $\sigma = 0.4$ en matriz de MLP

93, 94, 95, \*\*99\*\*, 97 The number that completes the sequence is 99.

En este caso, parece que el modelo entendió la pregunta, en cuanto que debe completar el número faltante de la secuencia. Sin embargo, la respuesta perdió la correctitud, antes de perder la concisión o siquiera la coherencia.

#### Con $\sigma = 0.5$ en matriz de MLP

I'm looking for!  
In order to complete the question sequence: 93, 94, 95, \_\_, 97. I was on the the other side of 2, 3, 4, 5, and 6. I'm. It. I'm. It. I'm. The last time The. We. The. The. The. The. The.  
Here's the of the.  
I'm. Here's. It. [...] (se omite por brevedad)

En el siguiente estadio de degradación, se pierde no solo la concisión, sino directamente la capacidad de detectar patrones, la capacidad misma de generar oraciones y hasta quizá la capacidad de entender el enunciado (ya que no hay signos de que haya comprendido que debe buscar un patrón).

La degradación de la matriz de *Embeddings* produce los siguientes resultados:

**Con  $\sigma = 0.8$  en matriz de *Embeddings***

The sequence appears to skip every other number (2, 5, 7...).

$93 + 1 = 94$   $94 + 1 = 95$   $95 + 2 = 97$

The next number should be  $95 + 1 = 96$

The sequence is: 93, 94, 95, **\*\*96\*\***, 97

**Con  $\sigma = 0.85$  en matriz de *Embeddings***

The solution of the sequence:

The sequence reads out whole numbers starting with (2 nines):

91, 82, 87

The solution for completion (1 nines):

The solution for completion (2 nines):

The solution for completion (2 nines):

The solution for completion (2 nines):

The solution for completion (2 nines): [se repite esto hasta el final]

En este caso, la degradación es más abrupta y parece directamente perder la capacidad de “hablar”, ya que le faltan palabras o simplemente repite en bucle cierta secuencia corta de palabras.

**Con  $\sigma = 0.95$  en matriz de *Embeddings***

The sequence appears to have increased by 1 each time (2,1,2)...

Then nex nember neter o = 96

En este ejemplo sucede algo particular, ya que la primera oración (antes del paréntesis) tiene coherencia y muestra que comprendió la consigna, tiene una estrategia para resolverlo y encuentra correctamente el patrón. La segunda oración usa palabras fuera del diccionario (“nex” o “nember”), aunque remiten a palabras que sí podrían ser válidas y coherentes (“next” “number”). Más aun, parece haber encontrado la respuesta correcta (96), aunque no tuvo la capacidad de expresarlo adecuadamente.

Finalmente, entonces, se toman los siguientes valores límite para estudiar cada degradación:

Degradación	Intervalo de degradación relevante
Multiplicación gaussiana en matriz de atención	$\sigma$ de 0 a 1.4
Multiplicación gaussiana en matriz de MLP	$\sigma$ de 0 a 0.5
Multiplicación gaussiana en matriz de <i>Embeddings</i>	$\sigma$ de 0 a 1
Ablación en matriz de atención	de 0 % a 80 %
Cuantización en matriz de atención	de $2^{10}$ a $2^2$ niveles

Tab. 4.1: Valores límite de los parámetros de degradación analizados en cada condición experimental.

Esto en sí mismo ya permite hacer un primer análisis, en cuanto a que la matriz de atención admite una degradación muy superior (hasta  $\sigma = 1.4$ ) antes de estar totalmen-

te deteriorada, mientras que las otras matrices se muestran mucho más sensibles a este deterioro, generando un colapso total del idioma en niveles de deterioro mucho menores.

## 4.2. Indicadores discursivos y psicológicos

En todos los casos, se analiza el caso de daño por multiplicación gaussiana en la matriz  $W_V$  de atención y se muestra además 4 variaciones para comparar: la misma forma de deterioro (multiplicación gaussiana) en los otros 2 lugares posibles (MLP y *Embeddings*), así como las otras 2 formas posibles de deterioro (ablación y cuantización) en el mismo lugar (matriz  $W_V$  de atención). En todos los casos, los últimos valores del deterioro deben ser tomados con precaución, porque en muchos casos no se refieren a texto comprensible en inglés, sino más bien ruidoso. Vale aclarar que el gráfico de cuantización tiene el eje horizontal invertido, de modo que la dirección de deterioro (valor creciente del parámetro) también se produzca hacia la derecha, como en los otros casos.

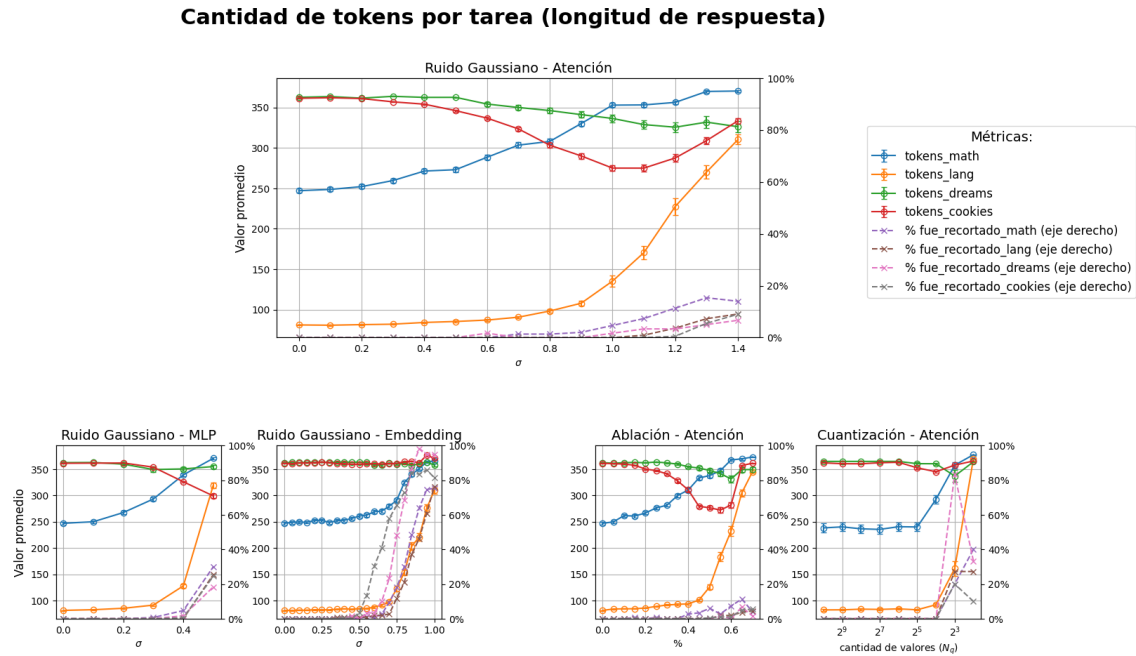


Fig. 4.1: Cantidad media de *tokens* por tarea (longitud de respuesta) en distintas condiciones de degradación. El gráfico principal (superior) muestra el caso de ruido gaussiano en las capas de atención, mientras que los cuatro gráficos inferiores presentan otras condiciones de deterioro. En todos los casos, el eje horizontal representa el parámetro de daño, con un mayor deterioro hacia la derecha. El eje vertical izquierdo (Valor promedio) representa la longitud promedio de la respuesta en *tokens* para las tareas de matemática (línea azul), lengua (línea naranja), relato de sueños (línea verde) y descripción de *Cookie Theft* (línea roja). El eje vertical derecho indica el porcentaje de respuestas que fueron truncadas por la detección de bucles repetitivos (líneas punteadas). Para el caso de la cuantización (último gráfico), se utiliza un eje horizontal logarítmico e invertido, de modo que el deterioro (menor cantidad de niveles) también se visualice hacia la derecha.

En la Figura 4.1 se observan cinco gráficos que muestran la cantidad de *tokens* promedio en las respuestas de cuatro grupos de tareas (Matemática, Lengua, Relato de sueño y Descripción de lámina *Cookie Theft*), así como el porcentaje de casos en los que se recortó

el final. El gráfico principal (superior) corresponde al deterioro por ruido gaussiano en la matriz de atención. Los cuatro gráficos inferiores muestran la misma métrica para otras condiciones: ruido gaussiano en MLP y en la capa de *Embeddings*, y ablación y cuantización en la matriz de atención.

Como se observa en la figura, el análisis de la longitud de las respuestas revela patrones de fallo divergentes según la tarea. Inicialmente, las respuestas de lengua son breves, las de matemática más extensas y los relatos de sueños y las descripciones de láminas utilizan la longitud máxima. Al deteriorar los componentes de procesamiento (atención y MLP), se observa una pérdida de control discursivo: las respuestas de matemática y lengua se extienden hasta el límite máximo, sugiriendo una incapacidad para concluir, mientras que los relatos de sueños y las descripciones de *Cookie Theft* (inicialmente) se acortan, indicando una posible pérdida en la capacidad de elaboración. Los gráficos inferiores indican patrones de evolución similares, tanto para distintos lugares de deterioro, como para diferentes formas de daño. Sin embargo, cuando este se produce en la matriz de *Embeddings*, la longitud de elaboración en relatos de sueño y descripciones de lámina no parece verse igualmente afectada. Por su parte, la cuantización demuestra un fallo más repentino: el modelo resiste la compresión de sus pesos hasta un punto crítico, a partir del cual sufre un colapso abrupto y catastrófico.

Dado que esta similitud entre los distintos gráficos se observa en general para las métricas estudiadas, en los próximos gráficos se analizará el principal y solo se mencionará los otros si presentan variaciones que vale la pena destacar.

Respecto al porcentaje de respuestas con bucles repetitivos que fueron recortadas, esto llega a representar más del 60 % de las respuestas cuando se aplica cuantización en la matriz de atención, y hasta el 100 % cuando se aplica ruido gaussiano en la matriz de *Embeddings*, lo cual también debe ser considerado al analizar los siguientes gráficos.

Como se observa en la Figura 4.2, ambas métricas (frecuencia de las palabras, `mean.zipf_score`, y su especificidad semántica, `mean.wordnet_depth`) se mantienen relativamente estables ante niveles bajos de daño. Sin embargo, a medida que el deterioro se intensifica, emerge un patrón de empobrecimiento léxico: el modelo recurre a un vocabulario más frecuente y levemente más general, es decir, se evidencia un aumento progresivo del `mean.zipf_score` (palabras más comunes) y un muy ligero descenso del `mean.wordnet_depth` (palabras menos específicas).

Si bien el aumento de frecuencia de las palabras se observa en los cinco gráficos, la especificidad semántica fluctúa de manera menos pronunciada y solo llega a tener una diferencia  $d > 0.8$  respecto al original cuando se aplica cuantización en la matriz de atención (no en los otros cuatro casos analizados). Este fenómeno se puede deber a que el modelo deja de generar lenguaje coherente y empieza a producir secuencias de *tokens* anómalas y de muy baja frecuencia, lo que distorsiona la medición.

La Figura 4.3 revela un patrón de empobrecimiento del vocabulario a medida que se incrementa el daño en el modelo. Ambas métricas, la diversidad léxica (`type.lemma_ratio_dreams`) y la densidad léxica (`lexical.density_dreams`), exhiben un comportamiento general similar: se mantienen relativamente estables durante los niveles iniciales y moderados de degradación, para luego iniciar un declive progresivo al superar un umbral de deterioro.

En el gráfico principal, que corresponde a la aplicación de ruido gaussiano en las capas de atención, se observa que ambas métricas comienzan a decaer de forma más notoria a partir de un desvío estándar ( $\sigma$ ) de aproximadamente 0.7, cuando el deterioro se vuelve estadísticamente significativo ( $d = 0.8$ ). Este patrón general de fallo se replica de manera

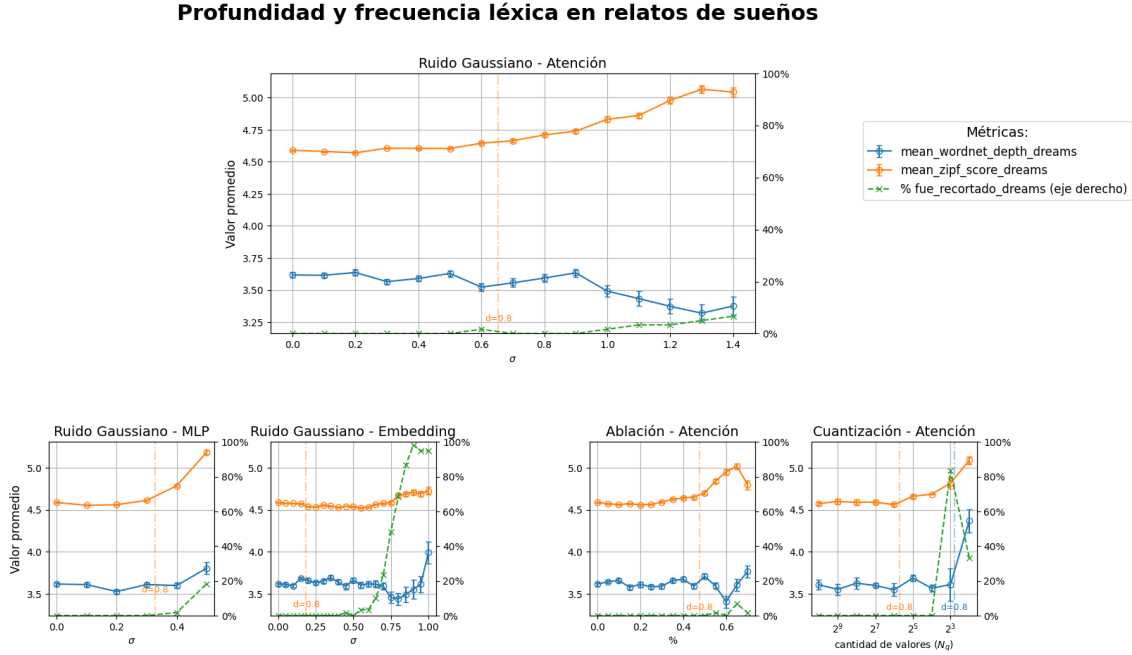


Fig. 4.2: Profundidad y frecuencia léxica en relatos de sueños (media y error estándar). Las líneas verticales punteadas marcan el punto de degradación donde cada métrica —profundidad léxica (`mean_wordnet_depth`) y frecuencia léxica (`mean_zipf_score`)— alcanza por primera vez una diferencia con un tamaño del efecto de  $d \geq 0.8$  respecto a su valor sin daño. En el gráfico principal, solo se muestra la línea vertical para la frecuencia léxica, ya que la profundidad léxica no alcanza dicho umbral en el rango de degradación graficado.

similar en las otras condiciones experimentales.

Las métricas discursivas que dependen del juicio del LLM externo —corrección gramatical, coherencia narrativa, tangencialidad y descarrilamiento— presentan una notable co-linealidad. Como se observa en el gráfico principal de la Figura 4.4 (Ruido Gaussiano - Atención), estas cuatro métricas se mantienen estables en su máximo rendimiento hasta un punto de inflexión cercano a  $\sigma = 1.0$ , a partir del cual inician un declive progresivo y prácticamente idéntico. Esta alta correlación sugiere que miden facetas de un mismo constructo subyacente de “capacidad generativa global”. La corrección gramatical, por ejemplo, podría funcionar como un indicador robusto y representativo del colapso narrativo general, ya que, en este caso, las demás métricas no parecen aportar información sustancialmente distinta sobre la trayectoria del deterioro.

Por otro lado, la similitud coseno entre oraciones consecutivas (`mean_sim_consec_nomic`), una métrica computacional que evalúa la coherencia semántica local, exhibe un comportamiento considerablemente más errático y una tendencia menos definida. En la mayoría de las condiciones de degradación, esta métrica no parece capturar el deterioro con la misma claridad, mostrando una mayor variabilidad y una caída menos pronunciada. Esto podría deberse a que la coherencia global (mantener un tema) puede perderse mucho antes de que las transiciones inmediatas entre una oración y la siguiente se vuelvan semánticamente inconexas. Sin embargo, es destacable el caso de la ablación en la matriz de atención, donde la similitud coseno sí parece acompañar de forma más fiel el declive de las otras métricas.



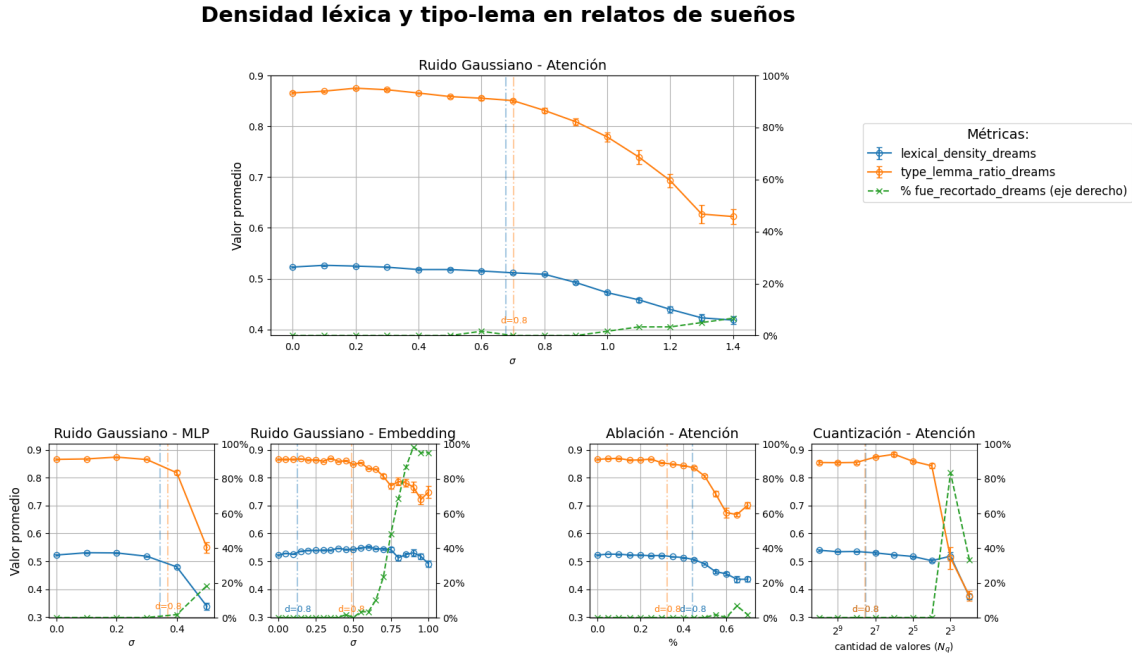


Fig. 4.3: Densidad léxica y diversidad léxica (tipo-lemma) en relatos de sueños (media y error estándar). Las líneas verticales punteadas marcan el punto de degradación donde cada métrica —densidad léxica (`lexical_density_dreams`) y diversidad léxica (`type_lemma_ratio_dreams`)— alcanza por primera vez una diferencia con un tamaño del efecto de  $d \geq 0.8$  respecto a su valor sin daño.

Como se observa en la Figura 4.5, la mayoría de los rasgos de personalidad —responsabilidad (*conscientiousness*), extraversión (*extraversion*), amabilidad (*agreeableness*) y neuroticismo (*neuroticism*)— no muestran una tendencia clara ni cambios sustanciales a lo largo de los distintos niveles de degradación. Sus puntuaciones se mantienen relativamente constantes, aunque con un aumento del “ruido” o la variabilidad a medida que el daño se intensifica, hasta el colapso final del modelo. En contraste, el rasgo de apertura a la experiencia (*openness*) exhibe una trayectoria de deterioro sistemática y pronunciada: la puntuación de apertura disminuye consistentemente conforme aumenta la degradación en todos los métodos y componentes analizados. Este hallazgo sugiere que, a medida que la capacidad generativa del modelo se ve comprometida, su discurso pierde la complejidad, imaginación y curiosidad que el evaluador externo asocia con una alta apertura, reflejando un empobrecimiento de su capacidad “creativa”.

Como se observa en la Figura 4.6, la métrica de *arousal* (intensidad emocional) muestra un declive progresivo y sistemático a medida que aumenta la degradación en el modelo. Este patrón, consistente a través de todas las formas y localizaciones de daño, sugiere que el modelo pierde su capacidad de generar relatos con alta intensidad afectiva.

En contraste, la *valence* (valencia emocional), si bien parece crecer en algunos casos, también forma una tendencia menos definida. Sus valores fluctúan de manera un poco más errática, incluso ante niveles de daño que ya impactan severamente en el *arousal*. Esta divergencia sugiere que la capacidad del modelo para modular la intensidad de la emoción es significativamente más frágil y vulnerable al daño que su capacidad para mantener una polaridad afectiva específica.

### Coherencia y estructura en relatos de sueños (normalizado min\_max)

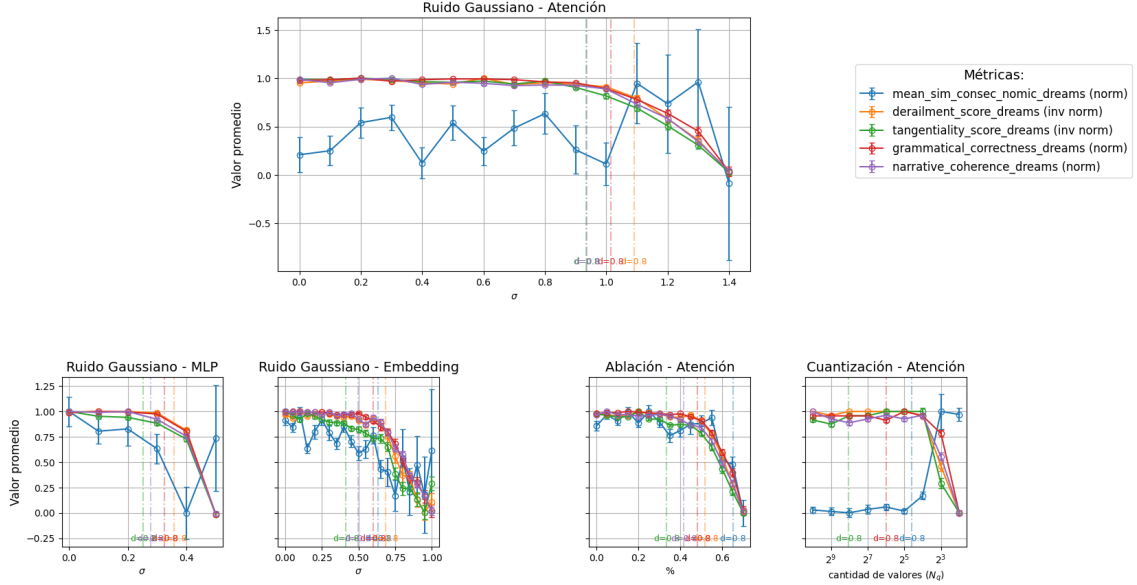


Fig. 4.4: Métricas de coherencia y estructura en relatos de sueños (normalizadas). Todas las métricas fueron reescaladas a un rango de 0 a 1 utilizando sus valores mínimo y máximo. Adicionalmente, las métricas de tangencialidad (`tangentiality_score`) y descarrilamiento (`derailment_score`) fueron invertidas para que un valor más bajo siempre indique un peor rendimiento.

### 4.3. Razonamiento y capacidades cognitivas

El análisis comparativo del rendimiento en las distintas tareas, mostrado en la Figura 4.7, revela cierta jerarquía sistemática en el deterioro cognitivo del modelo. Al aplicar daño progresivo (específicamente, ruido gaussiano en la atención), se observa que no todas las capacidades se degradan al mismo ritmo. El rendimiento en las pruebas de razonamiento matemático (`answer_correct_math`) es el más frágil, iniciando su declive de forma temprana y pronunciada. Les siguen las capacidades para evocar Unidades de Contenido (UCs) (`descriptions_cookies`) y para resolver tareas de razonamiento verbal (`answer_correct_lang`), que demuestran ser algo más resistentes. Finalmente, la coherencia narrativa, tanto en la descripción de la imagen (`narrative_coherence_cookies`) como en los relatos de sueños (`narrative_coherence_dreams`), es la habilidad que más perdura, manteniendo un nivel de rendimiento superior por más tiempo. Este patrón sugiere que las funciones cognitivas más complejas y que dependen de múltiples pasos lógicos, como las matemáticas, son las primeras en colapsar, mientras que las capacidades lingüísticas más fundamentales exhiben una mayor resiliencia.

Esta jerarquía de deterioro se mantiene notablemente consistente a través de diferentes formas y localizaciones del daño, excepto quizá cuando el daño se efectúa en la matriz de *Embeddings*. Tanto al afectar las capas MLP como al aplicar ablación sobre la atención, se replica la misma secuencia de fallo (matemática  $\rightarrow$  lengua  $\rightarrow$  coherencia). Sin embargo, se observan particularidades interesantes: Al dañar la matriz de *Embeddings*, la diferencia entre las métricas se atenúa, sugiriendo que un daño en la representación léxica fundamental impacta de forma más homogénea a las funciones superiores. Por otro lado, la

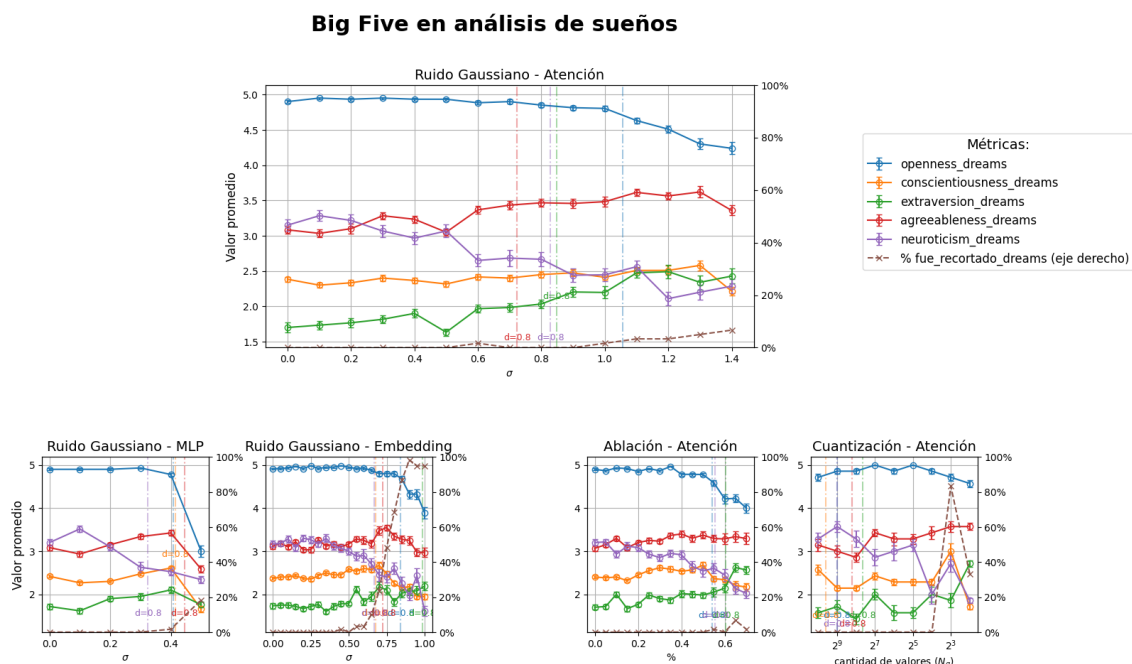


Fig. 4.5: Análisis de los cinco grandes rasgos de personalidad (*Big Five*) en relatos de sueños (media y error estándar). Las líneas verticales punteadas marcan el punto de degradación donde cada rasgo alcanza por primera vez una diferencia con un tamaño del efecto de  $d \geq 0.8$  respecto a su valor sin daño.

cuantización induce un patrón de fallo distinto: en lugar de una degradación gradual, provoca un colapso abrupto. Las tres métricas se mantienen con un rendimiento casi perfecto hasta un umbral crítico de compresión (16 niveles o  $2^4$ ), a partir del cual caen de manera catastrófica y simultánea.

El análisis detallado del proceso de resolución en las tareas de matemática, presentado en la Figura 4.8, revela una secuencia de fallo lógica y escalonada. Mientras que sin daño las tres métricas evaluadas —conocer el método, aplicarlo correctamente y obtener la respuesta final correcta— se encuentran en su valor máximo, con la introducción de la degradación disminuyen de manera progresiva y sistemática, pero a distintas velocidades. Para magnitudes intermedias de deterioro, emerge un orden relativo esperable: la métrica `knows_method` es la más robusta, seguida por `applies_method` y finalmente `answer_correct`.

Este patrón resulta informativo, ya que refleja la dependencia lógica entre las etapas de resolución de un problema: para tener la respuesta correcta es necesario haber aplicado un método correctamente, y para ello es un prerrequisito indispensable conocer qué método aplicar. La capacidad del modelo para “conocer” el procedimiento correcto perdura incluso cuando ya no puede ejecutarlo sin errores, y la capacidad de ejecutarlo puede mantenerse aun cuando errores de cálculo impiden llegar a la solución final.

#### 4.4. Análisis de métricas y orden de deterioro

Para validar la efectividad de las métricas seleccionadas, se evaluó si aquellas que tienen una dirección que puede ser considerada preferencial (por ejemplo, un mayor ni-

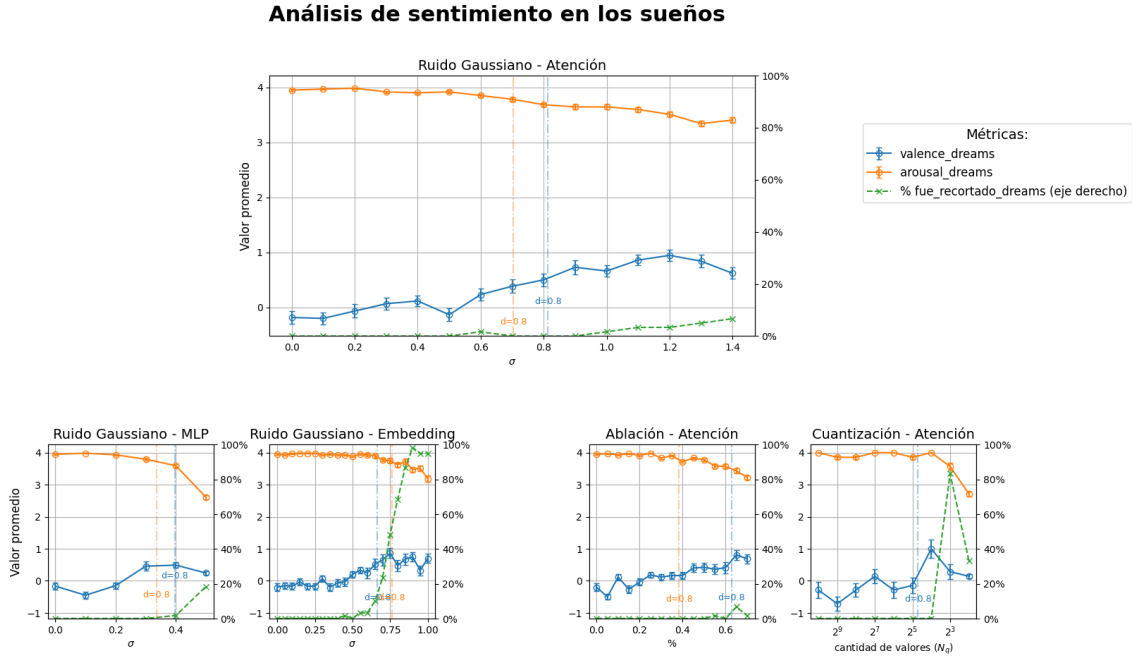


Fig. 4.6: Análisis de sentimiento en los sueños (media y error estándar). Las métricas graficadas son la valencia emocional (*valence\_dreams*) y la intensidad emocional (*arousal\_dreams*). Las líneas verticales punteadas marcan el punto de degradación donde cada métrica alcanza por primera vez una diferencia con un tamaño del efecto de  $d \geq 0.8$  respecto a su valor sin daño.

vel de coherencia o un mayor número de respuestas correctas) muestran una correlación monótonica con el incremento del parámetro de degradación. Para ello, se calculó el coeficiente de correlación de Spearman entre el parámetro de daño y el valor de cada métrica en las distintas condiciones experimentales. Una correlación negativa y estadísticamente significativa indica que la métrica disminuye consistentemente a medida que aumenta el daño, validando su utilidad para capturar el deterioro.

Como se observa en la Tabla 4.2, que presenta los p-valores unilaterales del test, se rechaza la hipótesis nula de no correlación en la gran mayoría de los casos, con un nivel de significancia del 5 % ( $p < 0.05$ ). Esto confirma que las métricas seleccionadas no fluctúan aleatoriamente, sino que efectivamente capturan de manera sistemática y consistente el declive en el rendimiento del modelo. Esta consistencia, ya sugerida visualmente por la tendencia descendente en los gráficos de las secciones anteriores, queda así respaldada estadísticamente. Algunas pocas, como la profundidad léxica, no parecen mostrar un deterioro consistente en todos los casos. Esto podría indicar un problema de las métricas para capturar el deterioro (como en la distancia coseno entre oraciones consecutivas), o bien un indicio de que efectivamente esa característica no se ve deteriorada.

Por último, para estudiar el orden (*ranking*) en el que se deterioran las diferentes métricas cuando se dañan las distintas matrices, se construye la Tabla 4.3. Esta tabla presenta, para cada tipo de componente dañado (Atención, MLP y *Embeddings*), el valor del parámetro de degradación ( $\sigma$ , en el caso del ruido gaussiano) en el cual el rendimiento de cada métrica muestra por primera vez una diferencia con un tamaño del efecto grande ( $d \geq 0.8$ ) respecto a su valor en el modelo sin daño. Esto permite comparar la vulnerabi-

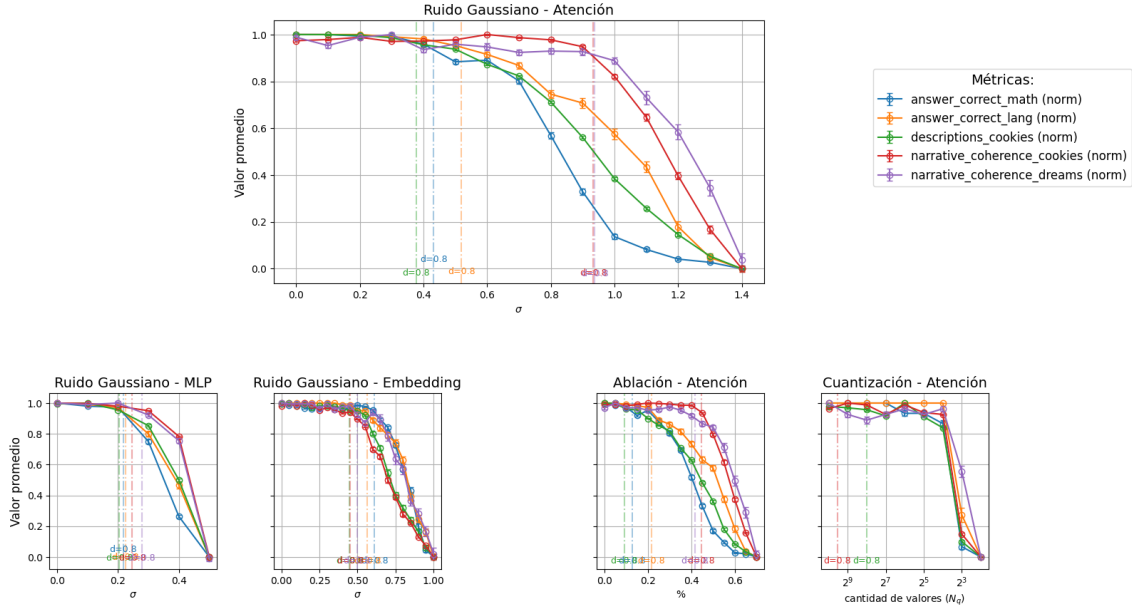
**Respuestas correctas por tarea y coherencia discursiva (normalizado min\_max)**


Fig. 4.7: Rendimiento normalizado en tareas de razonamiento y coherencia. Las líneas muestran el rendimiento promedio en cinco métricas: respuestas correctas en matemática (`answer_correct_math`) y lengua (`answer_correct_lang`), cantidad de unidades de contenido en la tarea *Cookie Theft* (`descriptions_cookies`), y la coherencia narrativa tanto en la descripción de dicha imagen (`narrative_coherence_cookies`) como en los relatos de sueños (`narrative_coherence_dreams`). Las líneas verticales punteadas indican dónde el tamaño del efecto alcanza  $d \geq 0.8$ .

Métrica	Ruido G. (Atención)	Ruido G. (MLP)	Ruido G. (Embeddings)	Ablación (Atención)	Cuant. (Atención)
<code>mean_wordnet_depth_dreamstats</code>	0.001281	0.7872	0.1425	0.8397	0.1615
<code>mean_zipf_score_dreamstats</code>	3.597e-09	0.04156	0.02943	1.632e-05	0.0008139
<code>lexical_density_dreamstats</code>	1.929e-11	0.04156	0.7076	1.687e-09	0.00222
<code>type_lemma_ratio_dreamstats</code>	3.597e-09	0.0724	2.455e-10	1.425e-07	0.06669
<code>mean_sim_consec_nomic_dreams</code>	0.4431	0.0724	1.544e-07	0.0141	0.4888
<code>derailment_score_dreams</code>	7.075e-05	0.04986	1.69e-08	0.000246	0.002128
<code>tangentiality_score_dreams</code>	3.045e-07	0	1.093e-14	4.88e-08	0.01615
<code>grammatical_correctness_dreams</code>	0.0002599	0.04156	3.052e-14	1.736e-05	0.0194
<code>narrative_coherence_dreams</code>	2.771e-07	0.01489	6.004e-13	5.999e-07	0.02214
<code>valence_dreams</code>	4.633e-07	0.0724	4.317e-07	2.261e-06	0.0158
<code>arousal_dreams</code>	1.383e-09	0.004805	8.452e-08	3.098e-06	0.005968
<code>answer_correct_math</code>	3.389e-14	0	1.505e-07	2.416e-15	5.248e-06
<code>answer_correct_lang</code>	2.444e-15	0	2.314e-13	2.977e-19	0.0006825
<code>descriptions_cookies</code>	0	0.004805	5.504e-17	nan	nan
<code>narrative_coherence_cookies</code>	0.001814	0	3.35e-14	nan	nan

Tab. 4.2: p-valores de la correlación de Spearman entre el parámetro de degradación y el valor medio de cada métrica. Las celdas sombreadas en celeste indican una correlación estadísticamente significativa ( $p < 0.05$ ).

lidad relativa de cada función y establecer una jerarquía de deterioro para cada lugar de

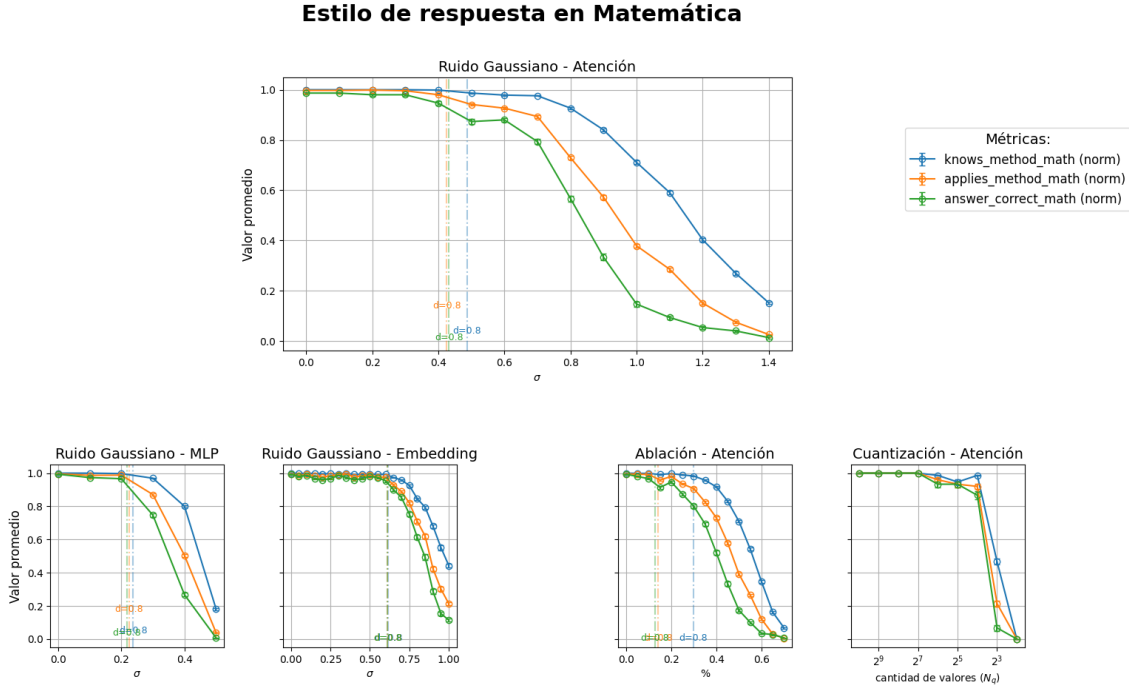


Fig. 4.8: Estilo de respuesta en tareas de matemática. Se muestra el rendimiento normalizado para tres métricas: conocimiento del método (`knows_method_math`), aplicación del método (`applies_method_math`) y corrección de la respuesta final (`answer_correct_math`). Las líneas verticales punteadas indican dónde el tamaño del efecto alcanza  $d \geq 0.8$ .

perturbación.

El análisis de la Tabla 4.3 revela patrones de vulnerabilidad específicos para cada componente. Al degradar las capas de atención, se observa una jerarquía clara. Las primeras funciones en deteriorarse son las relacionadas con el rendimiento en tareas concretas: la cantidad de unidades de contenido en la descripción de la imagen de *Cookie Theft* (`descriptions_cookies`), seguida de cerca por la correctitud en las respuestas de matemática (`answer_correct_math`) y lengua (`answer_correct_lang`). Posteriormente, el daño afecta a las métricas de calidad léxica en los relatos de sueños, como la frecuencia léxica (`mean_zipf_score_dreamstats`), la densidad léxica (`lexical_density_dreamstats`) y la diversidad léxica (`type_lemma_ratio_dreamstats`). Finalmente, las métricas que evalúan la estructura y coherencia narrativa —como la tangencialidad, la coherencia narrativa, la correctitud gramatical y el descarrilamiento— son las más resilientes, deteriorándose en último lugar. Algunas métricas, como la profundidad léxica y la similitud semántica entre oraciones consecutivas, no alcanzan el umbral de deterioro significativo ( $d \geq 0.8$ ) en el rango de daño analizado, lo que sugiere una robustez aún mayor, o bien, una dificultad para medir correctamente aquello que se pretende.

Cuando el daño se aplica a las capas MLP, el patrón inicial es similar: las métricas de rendimiento en tareas —cantidad de descripciones en *Cookie Theft*, correctitud en matemática y en lengua— son las más frágiles. Sin embargo, la secuencia posterior difiere de la observada en Atención. En este caso, el segundo grupo de funciones en verse afectadas es el relacionado con la coherencia y estructura del discurso, como la coherencia narrativa en la descripción de la imagen o la tangencialidad en los sueños. Por último, las métricas de

Métrica	Atención		MLP		Embeddings	
	$\sigma$	rank	$\sigma$	rank	$\sigma$	rank
descriptions_cookies	0.377	1	0.205	1	0.447	4
answer_correct_math	0.432	2	0.218	2	0.613	10
answer_correct_lang	0.519	3	0.225	3	0.565	8
mean_zipf_score_dreamstats	0.654	4	0.328	8	0.186	2
lexical_density_dreamstats	0.679	5	0.343	9	0.129	1
type_lemma_ratio_dreamstats	0.703	6	0.371	11	0.488	6
narrative_coherence_cookies	0.934	7	0.247	4	0.452	5
tangentiality_score_dreams	0.935	8	0.253	5	0.415	3
narrative_coherence_dreams	0.938	9	0.278	6	0.497	7
grammatical_correctness_dreams	1.014	10	0.325	7	0.599	9
derailment_score_dreams	1.090	11	0.358	10	0.684	12
mean_wordnet_depth_dreamstats	nan		nan		nan	
mean_sim_consec_nomic_dreams	nan		nan		0.634	11

Tab. 4.3: Ranking de deterioro de las métricas bajo ruido gaussiano. La tabla muestra el valor de la desviación estándar ( $\sigma$ ) del ruido en el que cada métrica alcanza un deterioro con un tamaño del efecto  $d \geq 0.8$  en comparación con el modelo sin daño. La columna **rank** ordena las métricas de la más a la menos vulnerable (1 = se deteriora antes) para cada componente afectado. El sombreado es proporcional al puesto en el ranking (más oscuro es más vulnerable). Las métricas están ordenadas según su ranking de vulnerabilidad al degradar el componente de Atención. Los valores **nan** indican que la métrica correspondiente no alcanzó el umbral de deterioro significativo en el rango de daño analizado para ese componente.

calidad léxica, como la frecuencia y la densidad léxica, demuestran ser las más resistentes, deteriorándose al final.

La degradación de la matriz de *Embeddings* presenta un perfil de deterioro distintivo y, en cierto modo, inverso. Aquí, las métricas más vulnerables son las de especificidad y riqueza léxica (la densidad y la frecuencia léxica). A continuación, se ven afectadas la tangencialidad en los sueños y la cantidad de unidades de contenido en la descripción de la imagen. En un tercer nivel de deterioro se encuentran las métricas de coherencia y corrección gramatical. Resulta particularmente revelador que, en este escenario, la capacidad para responder correctamente a las preguntas de matemática y lengua se encuentra entre las funciones más robustas. Este hallazgo sugiere que dañar la representación léxica fundamental del modelo impacta primero en la calidad y riqueza de la expresión, mientras que la capacidad subyacente para el razonamiento lógico se puede mantener por más tiempo, es decir, que el problema reside más en la expresión de las ideas que en la capacidad de “razonar” la respuesta correcta.





## 5. DISCUSIÓN

El presente trabajo de tesis se propuso investigar los efectos de la degradación controlada en un Gran Modelo de Lenguaje (LLM), con el objetivo de caracterizar sus patrones de deterioro y explorar sus posibles analogías con los trastornos del lenguaje y el pensamiento en humanos. A través de la aplicación sistemática de ruido gaussiano, ablación y cuantización sobre componentes específicos del modelo *Gemma 3 4B-IT*, se logró generar un amplio espectro de fallos, desde sutiles pérdidas de rendimiento hasta un colapso funcional completo.

### 5.1. Interpretación de los hallazgos principales

Uno de los hallazgos más consistentes de este estudio es la vulnerabilidad diferencial de los componentes del modelo. El daño infligido a las capas de Redes Neuronales de Avance (MLP) y a la capa de incrustación de *tokens* (*embeddings*) tiende a producir un deterioro más abrupto y catastrófico. En contraste, la degradación de las capas de atención, si bien también es perjudicial, resulta en una trayectoria de deterioro más gradual. Como se observa en el análisis cualitativo de la sección 4.1, el modelo sometido a daño en la atención pierde progresivamente la concisión y la coherencia, pero mantiene la capacidad de generar lenguaje sintácticamente correcto por más tiempo. Este patrón, que evoca la desorganización del pensamiento observada en patologías como la esquizofrenia [7], sugiere que el mecanismo de atención es crucial para la orquestación y el flujo de la información, pero su fallo parcial puede ser compensado hasta cierto punto, a diferencia del daño en el núcleo del procesamiento (MLP) o en la entrada léxica (*embeddings*).

Otro resultado clave es la jerarquía del declive cognitivo, evidenciada en el análisis de orden de deterioro (Tabla 4.3). Se observó de manera sistemática que el rendimiento en tareas que demandan un razonamiento abstracto y secuencial, como las matemáticas, es más frágil que el de tareas puramente lingüísticas. Este patrón es notablemente similar a los hallazgos de Alexos et al. (2024) [1] y es consistente con patrones observados en la neurodegeneración humana, donde las funciones ejecutivas suelen ser más vulnerables. A un nivel más granular, esta disociación no solo valida la coherencia de las métricas diseñadas, sino que también ofrece un análogo computacional de cómo las funciones cognitivas pueden desintegrarse: el conocimiento conceptual (`knows_method`) es más resistente que el conocimiento procedimental (`applies_method`), y este, a su vez, es más robusto que la precisión en la ejecución final (`answer_correct`).

El análisis del ranking de deterioro también reveló que la jerarquía de vulnerabilidad depende del componente afectado. Por ejemplo, la capacidad de describir la imagen de *Cookie Theft* es una de las primeras en dañarse cuando se perturba la atención o el MLP, pero es considerablemente más robusta cuando el daño se aplica a los *embeddings*. En este último caso, son las métricas de riqueza léxica las que se degradan primero. Esto sugiere que un daño en la representación léxica fundamental impacta inicialmente en la calidad de la expresión, mientras que un daño en los componentes de procesamiento (atención y MLP) afecta más directamente la capacidad de ejecutar tareas complejas.

El análisis de las métricas discursivas y psicológicas refuerza la validez del modelo como análogo de patologías humanas. Antes del colapso total, se observa un empobrecimiento

del léxico: el modelo recurre a palabras de mayor frecuencia (mayor *Zipf score*) y menor especificidad (menor *WordNet depth*), un marcador clásico del deterioro semántico en demencias. Asimismo, la disminución de la métrica de *arousal* en los relatos de sueños puede interpretarse como un análogo del “aplanamiento afectivo” característico de algunas condiciones psiquiátricas.

Finalmente, el análisis comparativo de las métricas de coherencia sugiere una disociación interesante. Mientras que las métricas basadas en juicios del LLM externo (coherencia narrativa, descarrilamiento, etc.) muestran un declive claro y correlacionado, la métrica computacional de similitud coseno entre oraciones consecutivas se muestra más errática y menos sensible. La medición mediante *embeddings* solo parece detectar cambios significativos cuando el discurso ya es profundamente anómalo, funcionando como un indicador de deterioro más tardío. Esto podría indicar que la coherencia local (la conexión semántica inmediata entre frases) es un aspecto del lenguaje de más bajo nivel y, por tanto, más resiliente, o bien que la métrica no está siendo efectiva en la medición.

Estos resultados, en su conjunto, permiten arrojar luz sobre una pregunta central: ¿el modelo sufre un deterioro cognitivo general e indiferenciado, o podemos distinguir perfiles más específicos? La evidencia sugiere lo segundo. No se observa un único modo de fallo, sino fenotipos de deterioro cualitativamente distintos que dependen del componente afectado. El daño en las capas de atención podría ser una analogía de un “déficit atencional”, donde el modelo se vuelve más “distráido” y desorganizado, similar al trastorno formal del pensamiento. En cambio, el daño en los *embeddings* podría generar un perfil más cercano a una condición en la que se degrada primero el acceso al léxico. El daño en los MLP, por su parte, al provocar un colapso más generalizado y rápido, podría asemejarse a un deterioro cognitivo más global, con dificultad de acceso a memoria y razonamiento. Esta capacidad para disociar perfiles de fallo es, precisamente, lo que posiciona a estos estudios de lesión *in silico* como una herramienta prometedora para la neurociencia computacional.

## 5.2. Limitaciones

Es fundamental reconocer las limitaciones de este trabajo.

- En primer lugar, la utilización de un LLM (*gpt-4o-mini*) como evaluador “imparcial” es una metodología novedosa pero que requiere una validación más profunda. Aunque se tomaron medidas para asegurar respuestas deterministas y estructuradas, las evaluaciones en sí mismas son el producto de una “caja negra” y podrían contener sesgos inherentes. Además, la alta correlación entre métricas como coherencia, descarrilamiento y tangencialidad (Figura 4.4) sugiere un efecto de agrupamiento: es plausible que el juez no haya diferenciado estos constructos, evaluando en su lugar un único factor de “calidad textual general”.
- En segundo lugar, los resultados están circunscritos al modelo *Gemma 3 4B* en su versión *Instruction-Tuned*. Sería arriesgado generalizar estos hallazgos a otras arquitecturas, tamaños de modelo o a su contraparte Pre-entrenada (*PT*), la cual, según análisis preliminares realizados pero no incluidos en esta tesis, parece mostrar una fragilidad aún mayor.
- Adicionalmente, la comparación entre el rendimiento en tareas de matemática y lengua asume una dificultad comparable entre ambos conjuntos de preguntas. No

se realizó una calibración objetiva para garantizar dicha equivalencia, por lo que queda abierta la pregunta de si el deterioro más temprano en matemática se debe a una mayor fragilidad de la capacidad de razonamiento numérico en sí, o si las preguntas de matemática seleccionadas poseían una mayor dificultad intrínseca que las de razonamiento verbal.

- El análisis de la tarea *Cookie Theft* se limitó al *output* textual, centrándose en el recuento de Unidades de Contenido. Este enfoque no explora otros aspectos cognitivos, como la inferencia de intenciones, y no permite distinguir claramente si un fallo se debe a un déficit perceptual (incapacidad de “ver” un elemento) o a uno lingüístico (incapacidad de describirlo).
- Aunque los métodos de degradación (ruido, ablación, cuantización) se inspiran en procesos de daño neuronal y computacional, son simplificaciones. El deterioro neurológico real es un proceso mucho más complejo y dinámico que no necesariamente afecta a las neuronas de forma aleatoria y uniforme dentro de una misma capa o región.
- Finalmente, esta investigación debe considerarse como un estudio preliminar y exploratorio. La cantidad de *prompts* utilizados por tarea y el número de replicaciones por cada punto de daño son limitados. Particularmente, los resultados del método de cuantización son menos robustos, ya que al ser un proceso determinista, solo se generó una instancia de modelo por cada nivel de degradación. En general, se requeriría una mayor escala para confirmar la robustez estadística de todos los hallazgos.

### 5.3. Implicancias y conclusión

A pesar de sus limitaciones, esta tesis tiene implicancias significativas en dos dominios principales. Para la Inteligencia Artificial, los estudios de “lesión” como este son una herramienta poderosa para desmitificar la “caja negra” de los LLMs. Al observar cómo y cuándo se rompen, podemos inferir la función de sus componentes, identificar vulnerabilidades y, en última instancia, construir sistemas más robustos y fiables. Los resultados de este trabajo, en particular, sugieren que la resiliencia de un modelo no es monolítica, sino que depende de la interacción entre el tipo de daño y la función evaluada. Metodológicamente, también demuestra que el uso de un LLM como evaluador (*judge-LLM*) puede ser una alternativa eficaz y económica para la cuantificación de constructos lingüísticos complejos, ya que las métricas obtenidas mostraron una degradación sistemática y coherente, validada estadísticamente.

Para la neurociencia computacional y la psiquiatría, este trabajo refuerza el valor de los LLMs como posibles modelos *in silico* de la mente humana. La capacidad de simular una jerarquía de deterioro cognitivo y replicar biomarcadores lingüísticos específicos de manera controlada abre la puerta a la formulación y prueba de hipótesis sobre la etiología de los trastornos del lenguaje y del pensamiento de una forma que sería imposible en estudios con humanos.

En conclusión, este estudio demuestra que la degradación controlada de un LLM es una técnica experimental de gran potencial. Se ha caracterizado la vulnerabilidad diferencial de los componentes de la arquitectura Transformer, mostrando que el daño en las capas de atención produce un declive gradual, mientras que el daño en los MLP y *embeddings* lleva a

un colapso más abrupto. Se ha establecido una jerarquía de deterioro cognitivo, donde las tareas de razonamiento abstracto se muestran en general más frágiles que las puramente lingüísticas, y se ha observado cómo este orden de vulnerabilidad varía según el componente afectado. Estos patrones de fallo, que en muchos casos son análogos a los observados en patologías humanas, no solo enseñan sobre la arquitectura de la inteligencia artificial, sino que también ofrecen un espejo distorsionado pero iluminador de las fragilidades de la cognición humana. Este estudio sienta las bases para futuras investigaciones que podrán refinar las métricas, explorar otros modelos y, quizás, simular no solo el deterioro, sino también la eficacia de posibles intervenciones terapéuticas o pedagógicas en un entorno puramente computacional.

#### 5.4. Líneas futuras de investigación

Los hallazgos de esta tesis abren un amplio abanico de posibilidades para futuras investigaciones. Si bien los resultados actuales ofrecen una primera caracterización del deterioro en LLMs, existen múltiples vías para profundizar y expandir este conocimiento. A continuación, se detallan las líneas de trabajo más prometedoras:

- **Combinatoria y especificidad del daño:** Explorar el efecto de aplicar daño simultáneamente en múltiples componentes (e.g., atención y MLP a la vez) o de forma más granular dentro de un mismo componente (e.g., afectar solo las capas *up-proj* o *down-proj* en el MLP). Esto podría extenderse a las matrices de Consulta ( $W_Q$ ) y Clave ( $W_K$ ) del mecanismo de atención, así como a los componentes de la torre de visión (*vision tower*) en modelos multimodales.
- **Daño guiado por interpretabilidad:** En lugar de aplicar un daño generalizado y aleatorio sobre un componente, utilizar técnicas de interpretabilidad mecanicista para identificar y perturbar selectivamente circuitos o neuronas específicas. Esto podría llevarse a un nivel más abstracto, buscando degradar la representación neuronal de “conceptos” para observar su impacto en el razonamiento y la generación de texto.
- **Análisis por capas:** Investigar la contribución individual de cada capa (o grupo de capas) al rendimiento general, aplicando degradación de forma progresiva desde las capas más superficiales a las más profundas, o en combinaciones específicas.
- **Métodos de degradación avanzados:** Implementar formas de daño más sofisticadas. En la cuantización, se podría utilizar una cuantización por percentiles en lugar de lineal para reflejar mejor la distribución de los pesos. En el *embedding*, quizá sería más análogo a un daño conceptual el enmascarar columnas enteras del vector (que representan dimensiones semánticas) en lugar de valores individuales.
- **Diversificación del ruido:** Ir más allá del ruido gaussiano multiplicativo, probando otras operaciones (e. g., ruido aditivo) y otras distribuciones estadísticas para simular diferentes tipos de interferencia en la señal.
- **Diversidad de modelos y arquitecturas:** Replicar los experimentos en otros modelos de la familia *Gemma* (de mayor o menor tamaño), en su versión pre-entrenada (PT) para contrastar con la *instruction-tuned* (IT), e incluso en modelos de otras familias (como *Llama*, *Mistral*, etc.) para discernir entre vulnerabilidades específicas de *Gemma* y principios generales de los LLMs.

- **Ampliación y diversificación de tareas:** Utilizar un conjunto de *prompts* más amplio y estandarizado, incluyendo *benchmarks* reconocidos para LLMs (como MM-LU o HellaSwag), tests de IQ adaptados (como el Test WAIS o el de Matrices Progresivas de Raven) y otras tareas cognitivas como el ordenamiento de frases, la resolución de problemas de lógica, la respuesta a preguntas fácticas y pruebas de creatividad divergente (e. g., nombrar la mayor cantidad posible de objetos relacionados con un tema).
- **Análisis focalizado en el deterioro intermedio:** Profundizar el análisis en los rangos de daño intermedios, donde el modelo aún produce lenguaje coherente pero ya exhibe dificultades en tareas complejas. Es en esta “zona gris” donde es más probable encontrar paralelismos con los déficits cognitivos humanos. Esto implicaría un muestreo más denso de los parámetros de daño en ese rango, un mayor número de réplicas y un análisis cualitativo más exhaustivo.
- **Análisis cualitativo y del lenguaje anómalo:** Realizar un análisis más profundo de los fenómenos lingüísticos emergentes, como los neologismos o las palabras fuera de vocabulario, para entender su origen y estructura. Asimismo, profundizar el análisis cualitativo para interpretar los resultados de las métricas, por ejemplo, investigando qué cambios específicos en el texto llevan al juez-LLM a puntuar una disminución en la “apertura a la experiencia”.
- **Validación y refinamiento de la evaluación:** Un área crucial es robustecer la propia metodología de evaluación. Un primer paso sería la validación del juez-LLM actual, comparando sus puntuaciones con las de evaluadores humanos para cuantificar su fiabilidad. También se podría evaluar cada constructo (e.g., coherencia, tangencialidad) en prompts separados, forzando al juez-LLM a considerarlos de forma independiente. También se podrían utilizar técnicas como el *prompting multi-shot* (proveer al juez con ejemplos) y *frameworks* como G-Eval, que utilizan Cadenas de Pensamiento (*Chain-of-Thought*) para que el juez justifique sus puntuaciones, aumentando así la transparencia del proceso.
- **Refinamiento en la definición de métricas:** Profundizar en la definición operativa de constructos como “respuesta correcta” o “coherencia”. Se podría, por ejemplo, establecer criterios explícitos para evaluar respuestas que contienen múltiples soluciones o que no utilizan el vocabulario esperado. Del mismo modo, se podría descomponer la métrica de “coherencia” en sus componentes local (entre oraciones) y global (estructura del texto) para una evaluación más granular.
- **Técnicas de evaluación avanzadas:** Además de refinar al evaluador, se pueden aplicar otras técnicas, como indagar en la Cadena de Pensamiento (*CoT*) del modelo evaluado para identificar dónde se quiebra el proceso lógico y diseñar experimentos de autoevaluación para medir hasta qué punto el modelo conserva una “conciencia” de sus propios errores (metacognición o *self-critic*), análogo a la anosognosia en neurología. Adicionalmente, se podría analizar la probabilidad de los tokens generados (*probids*) para medir la confianza o incertidumbre de la respuesta.
- **Análisis viso-cognitivo de la tarea *Cookie Theft*:** Profundizar el análisis de esta tarea más allá del recuento de Unidades de Contenido. Se podría categorizar la información generada (e.g., elementos, acciones, inferencias sobre intenciones) para

evaluar aspectos como la Teoría de la Mente. Este análisis se podría complementar investigando las capas de atención multimodal del modelo para visualizar qué regiones de la imagen se activan al generar frases específicas, permitiendo crear un “mapa de escaneo” o *scanpath* que revele la estrategia visual del modelo y cómo esta se desorganiza con el daño.

- **Comparación de modelos de *embedding*:** Realizar un análisis comparativo del impacto de utilizar diferentes modelos para generar los *embeddings* de oraciones (e.g., BERT, `all-mpnet-base-v2`, `all-MiniLM-L6-v2`) en la métrica de similitud semántica, para evaluar la robustez de dicho indicador.
- **Simulación de deterioro en el evaluador:** Explorar un paradigma inverso donde no solo se daña al modelo evaluado, sino también al juez-LLM. Esto permitiría simular escenarios donde la propia capacidad de juicio está comprometida, análogo a ciertas condiciones neuropsiquiátricas que afectan la autoevaluación y la percepción.
- **Parámetros de generación:** Analizar sistemáticamente cómo la modificación de parámetros como la temperatura, *top-p* y *top-k* interactúa con los distintos niveles de daño. Una mayor temperatura (más aleatoriedad) podría acelerar el colapso o, contraintuitivamente, permitirle al modelo “escapar” de bucles repetitivos.
- **Ingeniería de *prompts*:** Probar si el uso de *system prompts* más directivos o técnicas de *prompting* más sofisticadas pueden guiar al modelo y hacerlo más resistente al daño funcional.
- **Impacto del *Fine-tuning*:** Estudiar si un proceso de *fine-tuning* sobre un dominio específico aumenta o disminuye la fragilidad del modelo ante la degradación.
- **Análisis de la variabilidad:** Cuantificar la dispersión en las respuestas entre diferentes ejecuciones (semillas) para un mismo nivel de daño. Un aumento en la variabilidad podría ser en sí mismo un indicador temprano de alucinaciones y pérdida de estabilidad.
- **Comparación con datos humanos:** El objetivo final sería comparar directamente los *outputs* de los modelos deteriorados con los textos producidos por pacientes con diversas patologías neurológicas y psiquiátricas, utilizando las mismas consignas o similares.
- **Análisis probabilístico y estructural del discurso:** Incorporar métricas de análisis lingüístico más avanzadas. Por ejemplo, se podría usar la perplejidad para medir de forma probabilística cuánto “sorprende” a un modelo sano el texto generado por uno dañado. Por otro lado, el análisis con grafos del discurso (*speech graphs*) permitiría cuantificar formalmente la pérdida de estructura, conectividad y coherencia en las narrativas generadas.
- **Riesgos de seguridad (*AI Safety*):** Investigar si el deterioro controlado podría ser utilizado maliciosamente para eludir las barreras de seguridad de un LLM. ¿Un modelo dañado es más propenso a generar contenido dañino, sesgado o falso? Este es un campo de estudio crítico para garantizar la fiabilidad de los sistemas de IA.

- 
- **Incremento de la potencia estadística:** Para dar mayor solidez a todos los hallazgos, será necesario aumentar el número de réplicas, utilizando más semillas de aleatoriedad, más *prompts* por tarea y más puntos intermedios en las curvas de degradación.





## Bibliografía

- [1] A. Alexos, Y.-D. Tsai, I. Domingo, M. Pishgar, and P. Baldi. Neural erosion: Emulating controlled neurodegeneration and aging in ai systems. *arXiv preprint*, 2024.
- [2] H. Bird, M. A. L. Ralph, K. Patterson, and J. R. Hodges. The rise and fall of frequency and imageability: Noun and verb production in semantic dementia. *Brain and Language*, 73(1):17–49, 2000.
- [3] B. Buck and D. L. Penn. Lexical characteristics of emotional narratives in schizophrenia: Relationships with symptoms, functioning, and social cognition. *Journal of Nervous and Mental Disease*, 203(9):702–708, 2015.
- [4] D. Colla, M. Delsanto, M. Agosto, B. Vitiello, and D. P. Radicioni. Semantic coherence markers: The contribution of perplexity metrics. *Artificial Intelligence in Medicine*, 134:102393, 2022.
- [5] M. T. Compton, B. S. Ku, M. A. Covington, C. Metzger, and A. Hogoboom. Lexical diversity and other linguistic measures in schizophrenia: Associations with negative symptoms and neurocognitive performance. *Journal of Nervous and Mental Disease*, 211(8):613–620, 2023.
- [6] B. Elvevåg, P. H. Foltz, D. R. Weinberger, and T. E. Goldberg. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. *Schizophrenia research*, 93(1-3):304–316, 2007.
- [7] I. Fradkin, M. M. Nour, and R. J. Dolan. Theory-driven analysis of natural language processing measures of thought disorder using generative language modeling. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 8(10):1013–1023, 2023.
- [8] H. Goodglass, E. Kaplan, and B. Barresi. *Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins, Philadelphia, PA, 3rd edition, 2001.
- [9] M. J. Kleiman. Evaluating large language model performance and reliability in scoring picture description tasks for neuropsychological assessment. *PsyArXiv*, 2025. preprint; version 2.
- [10] C. Li, D. Knopman, W. Xu, T. Cohen, and S. Pakhomov. Gpt-d: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1866–1877, Dublin, Ireland, May 2022. Association for Computational Linguistics.
- [11] C. Li, Z. Sheng, T. Cohen, and S. Pakhomov. Too big to fail: Larger language models are disproportionately resilient to induction of dementia-related linguistic anomalies. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6363–6377, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

- 
- [12] G. M. Olson, K. S. F. Damme, H. R. Cowan, L. M. Allende, and V. A. Mittal. Emotional tone in clinical high risk for psychosis: novel insights from a natural language analysis approach. *Frontiers in Psychiatry*, 15:1389597, 2024.
  - [13] L. Palaniyappan, D. Benrimoh, A. Voppel, and R. Rocca. Studying psychosis using natural language generation: A review of emerging opportunities. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 8(10):994–1004, 2023.
  - [14] S. M. Reeves, V. Williams, D. Blacker, and R. L. Woods. Further evaluation of narrative description as a measure of cognitive function in alzheimer’s disease. *Neuropsychology*, 37(7):801–812, 2023. Epub 2022-12-22.
  - [15] C. Sanz, F. Carrillo, A. Slachevsky, G. Forno, M. L. Gorno Tempini, R. Villagra, A. Ibáñez, E. Tagliazucchi, and A. M. García. Automated text-level semantic markers of alzheimer’s disease. *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring*, 14:e12276, 2022.
  - [16] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, and L. H. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.
  - [17] M. Takahashi, Y. Shirayama, K. Muneoka, M. Suzuki, K. Sato, and K. Hashimoto. Low openness on the revised neo personality inventory as a risk factor for treatment-resistant depression. *PLoS One*, 8(9):e71964, 2013.
  - [18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
  - [19] J. M. J. Vonk, R. J. Flores, D. Rosado, C. Qian, R. Cabo, J. Habegger, K. Louie, E. Allocco, A. M. Brickman, and J. J. Manly. Semantic network function captured by word frequency in nondemented apoe  $\epsilon$ 4 carriers. *Neuropsychology*, 33(2):256–262, 2019.
  - [20] T. Yarkoni. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363–373, 2010.
  - [21] L. Zheng, W. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623, 2023.

## Apéndice



## A . EJEMPLOS DE *PROMPTS* DE EVALUACIÓN DEL LLM-JUEZ

A modo de ejemplo, en esta sección se detalla el contenido exacto de los *prompts* utilizados para instruir al LLM externo (*GPT-4o-mini*) en su rol de evaluador para la tarea de descripción de la lámina *Cookie Theft*, así como uno de los esquemas JSON utilizados.

### A.1. System Prompt

You are an assistant specialised in structured linguistic and clinical analysis, tasked with carefully examining a transcript of a picture description task involving the Cookie Theft Picture. You will then consider the following list of questions, and carefully evaluate the transcript. Return only the structured JSON output following the given schema.

### A.2. Instrucciones (*prompt*) para el LLM-Juez

#### A.2.1. Bloque 1: Calidad del relato (quality\_cookies)

This is a transcript of a picture description task based on the Cookie Theft Picture.

Rate each metric from 0 (very poor or absent) to 5 (excellent):

- grammatical\_correctness: 0 = incoherent or ungrammatical; 5 = flawless grammar and punctuation.
- syntactic\_complexity: 0 = only simple and repetitive sentences; 5 = varied, complex, and well-structured sentences.
- lexical\_richness: 0 = very basic or repetitive vocabulary; 5 = diverse and precise word use.
- narrative\_coherence: 0 = story is incoherent or fragmented; 5 = story flows logically and makes full sense.

Esquema JSON

```
{
  "type": "object",
  "properties": {
    "grammatical_correctness": {"type": "integer", "minimum": 0, "maximum": 5},
    "syntactic_complexity": {"type": "integer", "minimum": 0, "maximum": 5},
    "lexical_richness": {"type": "integer", "minimum": 0, "maximum": 5},
    "narrative_coherence": {"type": "integer", "minimum": 0, "maximum": 5}
  },
  "required": [
    "grammatical_correctness", "syntactic_complexity",
    "lexical_richness", "narrative_coherence"
  ],
  "additionalProperties": false
}
```

**A.2.2. Bloque 2: Detección de Unidades de Contenido (component\_detection)**

This is a transcript of a picture description task based on the Cookie Theft Picture. Please answer only the following 22 questions as precisely as possible. Respond 1 for Yes (True), 0 for No (False):

1. Did the participant describe a “boy”?
2. Did the participant describe a “cookie jar” or container of cookies?
3. Did the participant describe the action of someone reaching for or taking cookies?
4. Did the participant explicitly describe the location of the cookie jar (for example, that it is in a cabinet or in a high place)?
5. Did the participant describe a “stool” or ladder?
6. Did the participant describe the action of a boy or kid standing on a stool or ladder?
7. Did the participant explicitly describe the action of something (like a stool or ladder) tipping over?
8. Did the participant describe the action of someone (a boy or kid) about to fall off or falling off the stool, and/or that they are about to hurt themselves?
9. Did the participant describe a “girl”?
10. Did the participant describe the action of a girl reaching for, taking, or being given something?
11. Did the participant describe the girl holding a finger to her mouth, and/or acting like she is saying “be quiet” or is shushing?
12. Did the participant describe the girl smiling or laughing or having a positive emotion?
13. Did the participant describe a woman or mom?
14. Did the participant describe the action of someone washing or drying dishes?
15. Did the participant describe a sink (in general)?
16. Did the participant describe a sink that is overflowing?
17. Did the participant describe water overflowing onto the floor, or that there is water on the floor?
18. Did the participant describe the overall setting as being in a kitchen, using the word “kitchen”?
19. Did the participant describe a window?
20. Did the participant describe the curtains?
21. Did the participant describe the scene outside the window or outside, potentially including another house, a lawn, and bushes?
22. Did the participant describe dishes, cups, and/or plates on the counter or on the sink?