UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

INTERPRETABILIDAD MECANICISTA DE LLMS APLICADA AL ANÁLISIS DE SENTIMIENTOS Y EMOCIONES

TESIS PARA OBTENER EL GRADO DE:

MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

EDUARDO ANDER QUINTERO SÁNCHEZ

BAJO LA DIRECCIÓN DE:

DRA. CRISTINA LÓPEZ RAMÍREZ

EN CODIRECCIÓN:

DR. RICARDO RAMOS AGUILAR

CUNDUACÁN, TABASCO, A: AGOSTO 2026

UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO

DIVISIÓN ACADÉMICA DE CIENCIAS Y TECNOLOGÍAS DE LA INFORMACIÓN

INTERPRETABILIDAD MECANICISTA DE LLMS APLICADA AL ANÁLISIS DE SENTIMIENTOS Y EMOCIONES

TESIS PARA OBTENER EL GRADO DE:

MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

EDUARDO ANDER QUINTERO SÁNCHEZ

BAJO LA DIRECCIÓN DE:

DRA. CRISTINA LÓPEZ RAMÍREZ

EN CODIRECCIÓN:

DR. RICARDO RAMOS AGUILAR

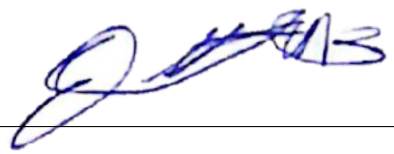
CUNDUACÁN, TABASCO, A: AGOSTO 2026

Declaración de Autoría y Originalidad

En la Ciudad de Cunduacán el día siete del mes de julio del año 2026, el que suscribe **Eduardo Ander Quintero Sánchez**, alumno del Programa de la **Maestría en Ciencias de la Computación** con número de matrícula **242H21008**, adscrito a la **División Académica de Ciencias y Tecnologías de la Información**, de la Universidad Juárez Autónoma de Tabasco, como autor de la Tesis presentada para la obtención de Grado de Maestro en Ciencias de la Computación y titulada **Interpretabilidad Mecanicista de LLMs aplicada al Análisis de Sentimientos y Emociones**, dirigida por la Dra. Cristina López Ramírez y el Dr. Ricardo Ramos Aguilar.

DECLARO QUE: La Tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la LEY FEDERAL DEL DERECHO DE AUTOR (Decreto por el que se reforman y adicionan diversas disposiciones de la Ley Federal del Derecho de Autor del 01 de Julio de 2020 regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita. Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad o contenido de la Tesis presentada de conformidad con el ordenamiento jurídico vigente.

Cunduacán, Tabasco a 7 de julio de 2026.



Estudiante: Eduardo Ander Quintero Sánchez

Cunduacán, Tabasco, a 03 de julio de 2026
Oficio No. 1402/2026/DACYTI/D

Asunto: Autorización de impresión de Tesis

C. Eduardo Ander Quintero Sánchez
Egresado de la Maestría en Ciencias de la Computación

En virtud de que cumple satisfactoriamente los requisitos establecidos en el Reglamento General de Estudios de Posgrado vigente en la Universidad, informo a Usted que se autoriza la impresión del trabajo recepcional **"Interpretabilidad mecanicista de LLMS aplicada al análisis de sentimientos y emociones"**, para presentar examen y obtener el Grado de Maestro en Ciencias de la Computación.

Sin otro particular, aprovecho la ocasión para enviarle un afectuoso saludo.

Atentamente

Dra. Laura Beatriz Vidal Turrubiates
Directora

División Académica de Ciencias y Tecnologías de la Información

C.c. p. Dr. Eduardo Hernández de la Cruz. – Encargado del despacho de la Coordinación de Posgrado.
Archivo

DRA.*LBVT/EHC

Carta de Cesión de Derechos

Villahermosa, Tabasco a 7 de julio de 2026.

Por medio de la presente manifiesto haber colaborado como AUTOR en la producción, creación y/o realización de la obra denominada: **Interpretabilidad Mecanicista de LLMs aplicada al Análisis de Sentimientos y Emociones**.

Con fundamento en el artículo 83 de la Ley Federal del Derecho de Autor y toda vez que, la creación y/o realización de la obra antes mencionada se realizó bajo la comisión de la Universidad Juárez Autónoma de Tabasco; entendemos y aceptamos el alcance del artículo en mención de que tenemos el derecho al reconocimiento como autores de la obra, y a la Universidad Juárez Autónoma de Tabasco mantendrá en un 100 % la titularidad de los derechos patrimoniales por un período de 20 años sobre la obra en la que colaboramos, por lo anterior, cedemos el derecho patrimonial exclusivo en favor de la Universidad.

COLABORADOR

Estudiante: Eduardo Ander Quintero Sánchez

TESTIGOS

Dra. Cristina López Ramírez

Dr. Ricardo Ramos Aguilar

Agradecimientos

Agradezco a la Universidad Juárez Autónoma de Tabasco y a la División Académica de Ciencias y Tecnologías de la Información por el espacio académico brindado para cursar la Maestría en Ciencias de la Computación y desarrollar este trabajo de investigación.

Agradezco a la Dra. Cristina López Ramírez, directora de esta tesis, y al Dr. Ricardo Ramos Aguilar, codirector, por el acompañamiento brindado durante el desarrollo y revisión del proyecto.

Asimismo, agradezco a los integrantes del jurado revisor, Dra. Juana Canul Reich, Dr. Jesús García Ramírez y Dr. Miguel Antonio Wister Ovando, por el tiempo dedicado a la revisión del documento y por las observaciones realizadas, las cuales contribuyeron a mejorar la versión final de la tesis.

Agradezco también al personal académico y administrativo del programa de posgrado por el apoyo brindado durante el proceso de formación y los trámites correspondientes.

Finalmente, agradezco a mi familia, compañeros y personas cercanas que, de distintas formas, me acompañaron durante esta etapa. En particular, agradezco a mis amigos Luis Enrique Ramón Pedrero y Andrés Alberto Góngora Ramos por su apoyo y compañía durante este proceso.

Índice general

Índice de figuras

Índice de tablas

Resumen

Los modelos de lenguaje grande han mostrado un desempeño relevante en tareas de procesamiento de lenguaje natural, incluido el análisis y clasificación de emociones. Sin embargo, su creciente complejidad dificulta comprender qué representaciones internas influyen en sus predicciones. Esta tesis aborda este problema mediante el uso de sparse autoencoders para descomponer activaciones internas en características latentes potencialmente interpretables.

La metodología consistió en obtener activaciones internas de un modelo de lenguaje de la familia LLaMA a partir de un corpus etiquetado con categorías emocionales, entrenar un sparse autoencoder y analizar la relación entre sus dimensiones latentes y las predicciones del modelo. Posteriormente, se realizaron intervenciones controladas sobre características latentes seleccionadas y aleatorias, midiendo sus efectos mediante cambios en probabilidad y log-odds.

Los resultados muestran que el steering latente mediante sparse autoencoders produce cambios medibles y estructurados en la evidencia emocional del modelo. Algunas características individuales presentan perfiles de respuesta diferenciados entre clases emocionales, lo que permite inspeccionar cómo ciertas intervenciones modifican la salida del modelo. No obstante, las comparaciones agregadas no muestran una ventaja cuantitativa robusta frente a características aleatorias.

En conjunto, la tesis muestra que los sparse autoencoders pueden utilizarse para analizar e intervenir representaciones internas de modelos de lenguaje en tareas de análisis de emociones. Aunque los resultados no implican características emocionales perfectamente especializadas, sí evidencian que el espacio latente aprendido contiene direcciones cuya intervención modifica de manera sistemática la evidencia emocional del modelo.

Palabras clave: sparse autoencoders, interpretabilidad mecanicista, modelos de lenguaje, steering latente, análisis de emociones

Abstract

Large language models have shown strong performance in natural language processing tasks, including emotion analysis and emotion classification. However, their increasing complexity makes it difficult to understand which internal representations influence their predictions. This thesis addresses this problem through the use of sparse autoencoders to decompose internal activations into potentially interpretable latent features.

The methodology consisted of extracting internal activations from a LLaMA-family language model using an emotion-labelled corpus, training a sparse autoencoder, and analyzing the relationship between its latent dimensions and the model's predictions. Controlled interventions were then applied to selected and random latent features, measuring their effects through changes in probability and log-odds.

The results show that latent steering with sparse autoencoders produces measurable and structured changes in the model's emotional evidence. Some individual features exhibit class-differentiated response profiles, allowing inspection of how specific interventions modify the model output. However, aggregate comparisons do not show a robust quantitative advantage over random features.

Overall, this thesis shows that sparse autoencoders can be used to analyze and intervene on internal representations of language models in emotion analysis tasks. Although the results do not imply perfectly specialized emotional features, they provide evidence that the learned latent space contains directions whose intervention systematically modifies the model's emotional evidence.

Keywords: sparse autoencoders, mechanistic interpretability, language models, latent steering, emotion analysis

Capítulo 1

Protocolo de tesis

1.1. Introducción

En los últimos años, el avance de los modelos de lenguaje grande (LLMs, por sus siglas en inglés *Large Language Models*) ha permitido mejoras significativas en tareas como el análisis y clasificación de emociones (AE) (Devlin et al., 2019; Sun et al., 2019). Estas tareas, pertenecientes al procesamiento del lenguaje natural, buscan identificar y clasificar las expresiones afectivas en texto (Feldman, 2013; Strapparava y Mihalcea, 2007). Esta tarea es crucial en aplicaciones como el marketing, atención al cliente y análisis de redes sociales, donde es importante entender las emociones subyacentes en los mensajes escritos (Diwali et al., 2024). Los LLMs, como los modelos de la familia LLaMA o ChatGPT, han demostrado un rendimiento sobresaliente en estas tareas, con potencial para superar a los enfoques tradicionales al capturar matices más complejos como el sarcasmo y la ironía (Zhang et al., 2024).

Sin embargo, a medida que los modelos se vuelven más poderosos, también se hacen más difíciles de interpretar, lo que plantea una preocupación crítica sobre su uso en aplicaciones sensibles (Samek et al., 2019). Esta falta de interpretabilidad ha llevado a que los LLMs sean considerados “cajas negras”, donde no está claro por qué el modelo toma ciertas decisiones (Molnar, 2025). Entender las razones detrás de estas decisiones es crucial no solo para aumentar la confianza en los resultados, sino también para cumplir con normativas y reducir sesgos en sectores clave como la salud, finanzas y servicios al cliente (Lipton, 2018; Rudin, 2019).

En respuesta a esta necesidad, la interpretabilidad mecanicista ha emergido como un campo

que busca descomponer estos modelos en componentes más simples y comprensibles, ayudando a entender fenómenos como la superposición, lo que permite analizar cómo procesan la información y generan predicciones (Elhage et al., 2022). Trabajo reciente en este campo ha demostrado que técnicas como el aprendizaje de diccionarios con sparse autoencoders (SAE) pueden abordar el problema de la superposición al descomponer los modelos en características más interpretables, ofreciendo una forma práctica de explorar y comprender las representaciones internas de los LLMs (Cunningham et al., 2024).

1.2. Marco teórico

El marco conceptual de esta investigación se enfoca en definir los conceptos clave que sustentan la idea de aplicar aprendizaje de diccionarios con sparse autoencoders para la interpretación a nivel mecanicista de Llama 3 en el análisis de emociones.

1.2.1. Modelos de lenguaje grande (LLMs) y LLaMA 3

Los LLMs son redes neuronales profundas que se entrenan en grandes cantidades de datos textuales para realizar una amplia gama de tareas de procesamiento de lenguaje natural (Brown et al., 2020; Touvron et al., 2023). LLaMA 3, desarrollado por Meta, es la tercera generación de una serie de modelos de código abierto que se destacan por su eficiencia y precisión en tareas complejas de PLN. A diferencia de otros modelos propietarios, LLaMA 3 es accesible a la comunidad científica para su personalización y mejora, lo que lo convierte en una plataforma ideal para investigar cómo los LLMs procesan y clasifican emociones (Dubey et al., 2024).

1.2.2. Interpretabilidad mecanicista

La interpretabilidad mecanicista es un enfoque emergente en inteligencia artificial que busca descomponer modelos complejos, como los LLM, en componentes más simples y comprensibles. A diferencia de la interpretabilidad post-hoc, que analiza las salidas del modelo después de su entrenamiento, la interpretabilidad mecanicista se centra en entender directamente los mecanismos internos del modelo. Esto implica examinar cómo las diferentes partes del modelo—como ca-

pas, neuronas o pesos—contribuyen colectivamente a su comportamiento final, permitiendo una comprensión más detallada y transparente. Este enfoque ha ganado relevancia ante la creciente complejidad de los LLMs, ya que permite desentrañar cómo procesan la información y generan predicciones. Al identificar las estructuras internas que determinan el comportamiento del modelo, la interpretabilidad mecanicista ofrece un marco valioso para diagnosticar, ajustar y mejorar estos sistemas, además de aumentar la confianza en su uso en aplicaciones críticas (Bereska y Gavves, 2024).

1.2.3. Steering

El steering en interpretabilidad mecanicista se refiere a las técnicas utilizadas para guiar o intervenir en los procesos internos de los modelos de inteligencia artificial, con el objetivo de comprender y manipular cómo estos producen sus predicciones. Este enfoque se centra en modificar activaciones neuronales, intervenir en representaciones latentes o ajustar rutas de cálculo internas para observar cómo estas alteraciones afectan el comportamiento del modelo. Por ejemplo, técnicas de steering pueden ser utilizadas para forzar la activación de neuronas específicas asociadas con conceptos monosemánticos o para suprimir aquellas que generan comportamientos no deseados (Olah et al., 2018).

En el contexto de los modelos generativos, como los LLM, el steering permite una exploración más profunda de cómo se representan conceptos abstractos y cómo se combinan para producir salidas. Al integrar estas herramientas con análisis mecanicistas, los investigadores pueden no solo diagnosticar fallas, sino también guiar el modelo hacia comportamientos alineados con los objetivos del usuario o eliminar sesgos indeseados (Bau et al., 2020; Vig y Belinkov, 2019).

1.2.4. Superposición, polisemanticidad y monosemanticidad

Un problema central en la interpretabilidad de modelos de lenguaje es que las activaciones internas no suelen estar organizadas de forma directamente interpretable para los humanos. En modelos neuronales de alta dimensionalidad, una misma neurona o coordenada de activación puede participar en la representación de múltiples conceptos, mientras que un mismo concepto puede estar distribuido entre varias neuronas. Este fenómeno se conoce como *superposición* y

dificulta establecer una correspondencia directa entre unidades individuales del modelo y conceptos semánticos específicos (Elhage et al., 2022).

En este contexto, una unidad o característica se considera *polisemántica* cuando responde a varios patrones, conceptos o funciones no necesariamente relacionados entre sí. Por ejemplo, una dimensión interna podría activarse ante textos con carga emocional negativa, pero también ante ciertos patrones gramaticales o estilos de escritura. Esta mezcla de factores limita la interpretación directa de las activaciones del modelo, ya que la activación de una unidad no necesariamente indica la presencia de un único concepto.

Por el contrario, una característica *monosemántica* se entiende como una dimensión cuya activación está asociada predominantemente con un patrón semántico o funcional identificable. En términos ideales, una característica monosemántica respondería de manera consistente ante ejemplos que comparten un mismo concepto y permanecería relativamente inactiva ante ejemplos no relacionados. Sin embargo, la monosemanticidad absoluta es difícil de demostrar empíricamente, especialmente en modelos de lenguaje grandes, por lo que en esta tesis el término se utiliza de manera operativa.

En particular, esta investigación no asume que las características latentes extraídas por los sparse autoencoders sean perfectamente monosemánticas. En cambio, se consideran *características potencialmente interpretables* aquellas dimensiones latentes que cumplen dos condiciones parciales: primero, presentan patrones de activación estadísticamente diferenciados respecto a ciertas clases emocionales; segundo, al ser intervenidas durante la inferencia, producen cambios medibles en la evidencia asociada con las predicciones emocionales del modelo. Esta definición permite distinguir entre una asociación descriptiva y una posible influencia funcional, sin afirmar que cada dimensión latente corresponda de manera exclusiva a una emoción humana.

1.2.5. Autoencoders dispersos y aprendizaje de diccionarios

Los autoencoders dispersos, o *sparse autoencoders* (SAEs), son modelos entrenados para reconstruir activaciones internas de una red neuronal a partir de una representación latente de mayor dimensionalidad pero con activación dispersa. En el contexto de interpretabilidad mecanicista, se utilizan como una forma de aprendizaje de diccionarios: el objetivo es descomponer una

activación interna densa en una combinación de características latentes más simples, selectivas y potencialmente interpretables (Bricken et al., 2023; Templeton et al., 2024).

Sea $x \in \mathbb{R}^d$ una activación interna extraída de una capa del modelo de lenguaje, por ejemplo del flujo residual de un transformer. Un SAE aprende un codificador f_{enc} que transforma x en un vector latente $z \in \mathbb{R}^{d'}$, y un decodificador f_{dec} que reconstruye la activación original a partir de dicho vector latente. Esta relación puede expresarse como

$$z = f_{\text{enc}}(x), \tag{1.1}$$

y

$$\hat{x} = f_{\text{dec}}(z). \tag{1.2}$$

De forma general, el codificador de la Ecuación (1.1) puede expresarse mediante la transformación afín

$$a = W_{\text{enc}}(x - b_{\text{dec}}) + b_{\text{enc}}, \tag{1.3}$$

seguida de la activación esparsa

$$z = g(a), \tag{1.4}$$

donde W_{enc} es la matriz del codificador, b_{enc} es el sesgo del codificador, b_{dec} representa un sesgo asociado a la reconstrucción y $g(\cdot)$ es una función de activación diseñada para inducir esparsidad. Posteriormente, el decodificador de la Ecuación (1.2) reconstruye la activación mediante

$$\hat{x} = W_{\text{dec}} z + b_{\text{dec}}. \tag{1.5}$$

El entrenamiento busca que la reconstrucción $\hat{x}$ de la Ecuación (1.2) conserve la mayor parte de la información contenida en x, pero utilizando solo un subconjunto pequeño de dimensiones latentes activas. Para ello se minimiza la función de pérdida

$$\mathcal{L} = \frac{1}{N} \sum_{j=1}^{N} \|x_j - \hat{x}_j\|_2^2 + \lambda\, \Omega(z_j), \tag{1.6}$$

donde el primer término de la Ecuación (1.6) corresponde al error cuadrático medio de recons-

trucción, $\Omega(z_j)$ controla la esparsidad del código latente y λ regula la importancia relativa de dicha penalización. En variantes clásicas de SAEs, $\Omega(z_j)$ suele implementarse como una penalización L_1, dada por

$$\Omega(z_j) = \|z_j\|_1 . \tag{1.7}$$

La penalización L_1 de la Ecuación (1.7) favorece que muchas coordenadas del vector latente sean cercanas a cero, permitiendo que cada activación se represente mediante pocas características. No obstante, implementaciones recientes de SAEs para modelos de lenguaje también emplean reglas de activación que imponen la esparsidad directamente, como Top-K, reduciendo la dependencia de penalizaciones continuas y proporcionando un control más explícito del número de características activas (Gao et al., 2025; He et al., 2024).

Un aspecto importante de los SAEs usados en interpretabilidad es que el espacio latente suele estar sobredimensionado. Si la activación original tiene dimensión d, se define un espacio latente de dimensión $d' = \alpha d$, donde α es un factor de expansión, por ejemplo $4\times$, $8\times$ o mayor. Esta sobrecompletitud permite que el SAE represente patrones que en el modelo original podrían encontrarse superpuestos en las mismas coordenadas. En términos intuitivos, si una activación del modelo mezcla varios factores semánticos o funcionales, el SAE proporciona un espacio más amplio donde dichos factores pueden separarse parcialmente en dimensiones latentes distintas.

En esta tesis, los SAEs se utilizan con dos propósitos complementarios. Primero, permiten obtener una representación latente dispersa de las activaciones internas del modelo de lenguaje. Segundo, proporcionan un espacio intervenible: al modificar una dimensión latente específica y reconstruir la activación correspondiente, es posible evaluar si dicha dimensión tiene influencia funcional sobre la salida del modelo. Por ello, el SAE no se emplea únicamente como herramienta descriptiva, sino como componente central de una estrategia de intervención.

1.2.6. Variantes de funciones de activación y esparsidad en SAEs

La función de activación $g(\cdot)$ de la Ecuación (1.4) desempeña un papel central en los SAEs, ya que determina qué dimensiones latentes se activan para cada ejemplo y cómo se impone la esparsidad. Existen distintas variantes utilizadas en la literatura reciente, cada una con ventajas y limitaciones en términos de estabilidad, control de esparsidad e interpretabilidad.

Una variante clásica consiste en utilizar una activación ReLU junto con una penalización L_1. En este caso, el código latente se define como

$$z = \mathrm{ReLU}(a), \tag{1.8}$$

y la esparsidad se promueve mediante el término $\lambda\|z\|_1$ en la función de pérdida ((1.6)). Esta formulación es simple y diferenciable casi en todas partes, pero el grado de esparsidad depende del valor de λ, por lo que puede ser necesario ajustar cuidadosamente el equilibrio entre reconstrucción y dispersión.

Otra alternativa es la activación Top-K, en la cual solo se conservan las K dimensiones latentes con mayor activación para cada ejemplo y el resto se fija en cero. Formalmente, si $\mathrm{TopK}(a)$ denota el conjunto de índices correspondientes a las K mayores preactivaciones, entonces el código satisface

$$z_i = \begin{cases} a_i, & \text{si } i \in \mathrm{TopK}(a), \\ 0, & \text{en otro caso.} \end{cases} \tag{1.9}$$

Esta variante impone una esparsidad estricta, ya que cada ejemplo activa exactamente, o como máximo, K características latentes. Esto facilita la comparación entre ejemplos y evita que el número de dimensiones activas varíe de forma excesiva. Sin embargo, un valor de K demasiado pequeño desde el inicio puede dificultar el entrenamiento, por lo que algunas implementaciones emplean esquemas de *K-annealing*: se inicia con un valor alto de K, menos restrictivo, y se reduce progresivamente hasta alcanzar el valor final deseado (He et al., 2024).

Una tercera variante reciente es JumpReLU, utilizada en trabajos como Gemma Scope. Esta activación introduce un umbral aprendido para cada dimensión latente, de modo que una característica solo se activa si su preactivación supera dicho umbral. De manera simplificada, puede escribirse como

$$z_i = a_i \, \mathbf{1}(a_i > \theta_i), \tag{1.10}$$

donde θ_i es un umbral aprendido para la dimensión i y $\mathbf{1}(\cdot)$ es una función indicadora. La Ecuación (1.10) permite controlar de forma más flexible qué características se activan, pero introduce dificultades adicionales de optimización debido al comportamiento discontinuo de la función indi-

cadora (Lieberum et al., 2024). La Tabla 1.1 resume cualitativamente estas variantes.

En esta investigación se utiliza una variante basada en Top-K ((1.9)), ya que permite controlar directamente el número de características latentes activas durante la codificación. Esta elección es adecuada para el objetivo del trabajo porque facilita comparar intervenciones entre dimensiones latentes y evita que los efectos observados dependan de códigos excesivamente densos. Además, el uso de un esquema progresivo de K permite iniciar el entrenamiento con una representación menos restrictiva y avanzar hacia una codificación más dispersa, conservando un equilibrio entre fidelidad de reconstrucción e interpretabilidad.

Variante	**Mecanismo de esparsidad**	**Ventaja principal**	**Limitación principal**
ReLU + L_1	Penalización continua sobre la magnitud del código latente.	Es simple, estable y fácil de implementar.	Requiere ajustar cuidadosamente el peso L_1; no fija directamente el número de latentes activos.
Top-K	Conserva únicamente las K activaciones latentes más grandes.	Controla explícitamente cuántas características se activan por ejemplo.	Puede dificultar el entrenamiento si K es demasiado restrictivo desde etapas tempranas.
Top-K con *K-annealing* [1]	Reduce progresivamente el número de latentes activos durante el entrenamiento.	Facilita una transición gradual hacia representaciones más dispersas.	Introduce un hiperparámetro adicional: el calendario de reducción de K.
JumpReLU	Activa cada dimensión solo si supera un umbral aprendido.	Permite umbrales específicos por característica y mayor flexibilidad.	Su optimización puede ser más compleja por la discontinuidad del umbral.

Tabla 1.1. Comparación cualitativa entre variantes de activación y esparsidad en sparse autoencoders.

Es importante señalar que la esparsidad no garantiza por sí misma la monosemanticidad. Una dimensión latente puede ser poco frecuente y, aun así, responder a varios patrones distintos. Por ello, en esta tesis las características latentes extraídas por el SAE se evalúan posteriormente mediante dos criterios adicionales: su selectividad estadística respecto a clases emocionales y su efecto funcional cuando son intervenidas mediante steering. De esta manera, la interpretación

[1] Función de activación seleccionada para el entrenamiento del SAE en esta investigación.

de una característica no se basa únicamente en su activación, sino también en su relación con las etiquetas emocionales y en el cambio que produce sobre la evidencia del modelo.

1.2.7. Conjuntos de datos para análisis de emociones en texto

La selección del corpus es un aspecto relevante en tareas de análisis de emociones, ya que diferentes conjuntos de datos pueden variar en origen, granularidad emocional, esquema de anotación, dominio textual y nivel de balance entre clases. En esta investigación se requiere un corpus textual que permita estudiar representaciones emocionales en modelos de lenguaje sin incorporar información visual, auditiva o multimodal, debido a que el objetivo del trabajo se centra en activaciones internas generadas a partir de texto.

Entre los recursos disponibles para análisis de emociones se encuentran conjuntos de datos especializados como ISEAR, EmoBank y GoEmotions, así como benchmarks unificados construidos a partir de múltiples corpus. ISEAR contiene descripciones auto-reportadas de experiencias emocionales asociadas con emociones básicas, lo que lo convierte en un recurso útil para estudiar eventos emocionales desde una perspectiva psicológica. Sin embargo, su dominio está centrado en relatos personales y su cobertura textual es menos diversa que la de un benchmark integrado. EmoBank, por su parte, utiliza un esquema dimensional basado en valencia, activación y dominancia, por lo que resulta útil para estudiar emociones desde una representación continua; no obstante, dicho formato requeriría un proceso adicional de discretización o mapeo para convertirlo en clases emocionales categóricas. GoEmotions ofrece una taxonomía fina de emociones y un gran número de ejemplos provenientes de Reddit, pero su naturaleza multi-etiqueta y su granularidad de 27 categorías pueden introducir mayor complejidad experimental cuando se busca controlar de forma estable la respuesta de un modelo de lenguaje mediante etiquetas candidatas discretas.

En este contexto, se eligió utilizar *UnifyEmotion* como base experimental, debido a que integra y estandariza múltiples corpus de emociones en texto, permitiendo trabajar con una colección más heterogénea que un único dataset individual. Esta característica resulta conveniente para el objetivo de la tesis, ya que el interés no es entrenar un clasificador especializado en un dominio particular, sino analizar si las activaciones internas de un modelo de lenguaje contienen

direcciones latentes relacionadas con evidencia emocional. La Tabla 1.2 resume la comparación cualitativa entre los principales recursos considerados.

Dataset	Tipo de recurso	Ventaja principal	Limitación para esta tesis
UnifyEmotion[2]	Benchmark unificado de corpus textuales de emociones.	Integra fuentes heterogéneas y permite trabajar con etiquetas emocionales estandarizadas.	Requiere seleccionar un subconjunto estable de emociones para el diseño experimental.
GoEmotions	Comentarios de Reddit anotados con una taxonomía fina de emociones.	Gran tamaño y cobertura emocional amplia.	Su granularidad y naturaleza multi-etiqueta pueden complicar el control experimental con etiquetas candidatas discretas.
EmoBank	Corpus anotado mediante dimensiones emocionales como valencia, activación y dominancia.	Permite estudiar emociones como variables continuas.	No está formulado principalmente como clasificación categórica directa, por lo que requeriría mapeos adicionales.
ISEAR	Relatos autoreportados de experiencias emocionales.	Tiene fundamento psicológico y emociones básicas claramente definidas.	Está centrado en experiencias personales y ofrece menor heterogeneidad textual que un benchmark unificado.

Tabla 1.2. Comparación de datasets de emociones considerados durante la selección del corpus experimental.

1.2.8. Análisis de sentimientos y emociones

El análisis de sentimientos es el proceso de identificar la polaridad de un texto, clasificándolo como positivo, negativo o neutro (Diwali et al., 2024). Por otro lado, el análisis de emociones busca detectar emociones más específicas como alegría, tristeza, ira, miedo o sorpresa, basándose en teorías psicológicas como las de Ekman (Diwali et al., 2024). La capacidad de los LLMs para realizar estas tareas depende de cómo representen internamente las emociones, lo que se hace difícil de interpretar sin técnicas avanzadas de descomposición (Belinkov y Glass, 2019).

[2]Dataset seleccionado como corpus principal para la realización de los experimentos de esta investigación.

1.2.9. Literatura relacionada

El campo de la interpretabilidad en inteligencia artificial ha avanzado significativamente en los últimos años, con varios enfoques dirigidos a descomponer modelos complejos para hacerlos más comprensibles. Estos avances han permitido a los investigadores identificar cómo los LLMs representan conceptos abstractos y concretos, lo que ha abierto la puerta para la manipulación de estas características con el fin de influir en el comportamiento del modelo (Dalvi et al., 2019).

Anthropic ha sido uno de los líderes en este campo, aplicando técnicas avanzadas de autoencoders dispersos y aprendizaje de diccionarios para descomponer modelos de lenguaje como Claude 3 en millones de características interpretables (Bricken et al., 2023; Templeton et al., 2024).

Por su parte, OpenAI ha implementado activaciones *TopK*, una técnica que optimiza la eficiencia de los autoencoders dispersos, eliminando penalizaciones innecesarias y mejorando la estabilidad del entrenamiento (Gao et al., 2025).

Siguiendo esta línea, Google DeepMind presentó Gemma Scope en 2024, un conjunto de autoencoders dispersos diseñados para analizar múltiples capas de modelos Gemma-2, incluyendo arquitecturas de hasta 27 mil millones de parámetros. Este trabajo introdujo los JumpReLU Sparse Autoencoders, los cuales combinan dispersión con linealidad para facilitar el análisis de circuitos internos y tareas de seguridad en inteligencia artificial (Lieberum et al., 2024).

Finalmente, los investigadores de Fudan University y Shanghai AI Lab desarrollaron Llama Scope, una suite de autoencoders dispersos para el modelo Llama-3.1-8B. Este trabajo implementó variantes mejoradas de los Top-K Sparse Autoencoders, junto con técnicas innovadoras como el “K-annealing”, lo que permitió analizar características latentes en posiciones específicas de cada capa del modelo. Llama Scope también exploró la generalización de estas características a contextos y tareas distintas, reforzando su utilidad en el análisis profundo de modelos masivos (He et al., 2024).

El análisis de emociones es un área de investigación activa con un interés creciente en cómo los LLMs, como LLaMA 3, pueden mejorar la precisión de las predicciones emocionales. Esto es particularmente importante en industrias como el marketing y la atención al cliente, donde las empresas dependen de la interpretación emocional para tomar decisiones estratégicas (Diwali et

al., 2024). Un avance reciente en análisis de emociones usando como base LLaMA 3 es *Emotion-LLaMA*, un modelo que combina la arquitectura de LLaMA con *instruction tuning* para realizar tanto reconocimiento como razonamiento emocional de manera multimodal. Este enfoque permite al modelo identificar no solo la polaridad de los sentimientos, sino también una gama más amplia de emociones complejas, basadas en teorías emocionales como las de Ekman y Plutchik (Cheng et al., 2024). *Emotion-LLaMA* ha demostrado ser efectivo en la detección de emociones complejas en textos y contenido multimedia, lo que lo convierte en un referente dentro de este campo en rápido desarrollo.

1.2.10. Comparación cualitativa con métodos de interpretabilidad post-hoc

Los métodos de interpretabilidad post-hoc buscan explicar el comportamiento de un modelo ya entrenado a partir de sus entradas, salidas o gradientes, sin modificar necesariamente sus mecanismos internos. Entre los enfoques más utilizados se encuentran LIME, SHAP e Integrated Gradients. LIME explica predicciones individuales aproximando localmente el comportamiento del modelo mediante un modelo interpretable; SHAP asigna valores de importancia a las características de entrada a partir de una formulación basada en valores de Shapley; e Integrated Gradients atribuye la predicción de una red profunda a sus características de entrada mediante la acumulación de gradientes a lo largo de una trayectoria entre una línea base y la entrada evaluada.

A diferencia de estos métodos, la interpretabilidad mecanicista busca analizar componentes internos del modelo, como activaciones, neuronas, capas o direcciones latentes, con el objetivo de comprender cómo dichas representaciones participan en el comportamiento observado. En esta tesis se adopta un enfoque mecanicista basado en sparse autoencoders, ya que el interés principal no es explicar únicamente qué tokens o características de entrada contribuyen a una predicción, sino estudiar si ciertas dimensiones latentes aprendidas a partir de activaciones internas pueden intervenirse y producir cambios medibles en la evidencia emocional del modelo. La tabla 1.3 resume la comparación entre los métodos post-hoc y la interpretabilidad mecanicista basada en sparse autoencoders con steering.

Método	Tipo de explicación	Ventaja principal	Limitación principal
LIME	Aproximación local mediante un modelo interpretable alrededor de una predicción.	Es flexible y puede aplicarse a distintos clasificadores.	La explicación depende de perturbaciones locales y puede ser inestable.
SHAP	Atribución de importancia a características de entrada mediante valores aditivos inspirados en teoría de juegos.	Ofrece una formulación teórica unificada para atribución de características.	Puede ser costoso computacionalmente y no revela directamente mecanismos internos.
Integrated Gradients	Atribución basada en gradientes acumulados desde una línea base hasta la entrada.	Es útil para redes diferenciables y cuenta con axiomas formales de atribución.	Requiere definir una línea base y explica sensibilidad de entrada, no necesariamente estructura interna.
Sparse Autoencoders + steering	Descomposición de activaciones internas en dimensiones latentes dispersas e intervención sobre dichas dimensiones.	Permite analizar e intervenir direcciones latentes para evaluar efectos funcionales.	La interpretación de las dimensiones latentes no siempre es monosemántica ni claramente superior a controles aleatorios.

Tabla 1.3. Comparación cualitativa entre métodos post-hoc e interpretabilidad mecanicista basada en sparse autoencoders.

1.3. Justificación

La investigación propuesta aborda una necesidad emergente en el campo de la inteligencia artificial: el desarrollo de modelos de lenguaje que no solo sean precisos en sus predicciones, sino que también sean interpretables. La creciente complejidad de los modelos de lenguaje grande (LLMs), como los modelos de la familia LLaMA, ha intensificado la preocupación sobre su naturaleza de “caja negra”. Este aspecto limita su aplicabilidad en contextos críticos, donde es fundamental comprender los motivos detrás de las decisiones del modelo, especialmente en tareas sensibles como el análisis de emociones.

Aplicaciones en sectores como la atención al cliente, la salud y la educación requieren modelos que puedan interpretar y clasificar correctamente el espectro emocional de los usuarios, generando respuestas que no solo sean precisas, sino también éticamente responsables. En estos contextos, la falta de interpretabilidad puede conducir a malentendidos, pérdida de confianza

y, en casos extremos, daños para los usuarios o las organizaciones.

El uso de autoencoders dispersos y aprendizaje de diccionarios en esta investigación se justifica porque permite analizar las representaciones internas de los modelos de lenguaje desde una perspectiva distinta a los métodos de explicabilidad post-hoc. En lugar de explicar únicamente la relación entre entradas y salidas, este enfoque busca descomponer activaciones internas en dimensiones latentes y evaluar si dichas dimensiones pueden ser intervenidas para producir cambios medibles en la evidencia emocional del modelo. De esta manera, la investigación se ubica dentro de la interpretabilidad mecanicista, al estudiar no solo qué predice el modelo, sino qué representaciones internas participan funcionalmente en dichas predicciones.

El aporte concreto de esta tesis consiste en proponer y evaluar un flujo experimental para el análisis de emociones en modelos de lenguaje mediante sparse autoencoders. Dicho flujo incluye la obtención de activaciones internas, el entrenamiento de un autoencoder disperso, la selección de características latentes potencialmente asociadas con señales emocionales, la intervención controlada de dichas características durante la inferencia y la medición de sus efectos mediante cambios en probabilidad y log-odds. Este procedimiento permite estudiar si el espacio latente aprendido contiene direcciones cuya modificación afecta de manera sistemática la evidencia emocional del modelo.

A diferencia de trabajos que se enfocan únicamente en identificar características interpretables de manera descriptiva, esta investigación incorpora una evaluación funcional mediante steering latente. Además, compara las características seleccionadas con características aleatorias como control experimental, lo cual permite matizar la interpretación de los resultados y evitar atribuir especialización emocional fuerte a dimensiones latentes que no demuestren una ventaja robusta frente a controles. Por ello, el aporte no se limita a aplicar sparse autoencoders, sino a evaluar empíricamente qué tan informativas y funcionalmente relevantes son las características obtenidas bajo las métricas propuestas.

Este proyecto no busca demostrar una mejora directa en la precisión de clasificación emocional, sino aportar una metodología de análisis e intervención que permita comprender mejor cómo los modelos de lenguaje representan y utilizan señales afectivas. En este sentido, la investigación contribuye a la construcción de herramientas más transparentes y analizables para el estudio de LLMs, al mismo tiempo que establece límites claros sobre el grado de especialización

e interpretabilidad que puede atribuirse a las características latentes aprendidas.

1.4. Preguntas de investigación

1. ¿Cómo permite el uso de sparse autoencoders la descomposición de representaciones internas en modelos de lenguaje aplicados al análisis de emociones?
2. ¿Qué dimensiones latentes obtenidas mediante esta descomposición muestran relación con señales emocionales y cómo influye su intervención directa en la evidencia asociada con las predicciones emocionales del modelo?

1.5. Hipótesis

El aprendizaje de diccionarios mediante sparse autoencoders sobre activaciones internas de un modelo de lenguaje permite proyectar dichas activaciones a un espacio latente disperso, cuya intervención controlada produce cambios sistemáticos y medibles en la evidencia asociada a las predicciones emocionales del modelo.

1.6. Objetivo general

Desarrollar y validar un enfoque basado en sparse autoencoders para extraer y analizar características latentes interpretables, potencialmente relacionadas con información emocional, en modelos de lenguaje, y evaluar su influencia funcional mediante intervenciones en el espacio latente.

1.6.1. Objetivos específicos

1. Obtener activaciones internas representativas de una capa del modelo de lenguaje a partir de un corpus de textos con diversidad emocional equilibrada.
2. Entrenar un sparse autoencoder que aprenda un espacio latente disperso, de alta calidad y estructurado, capaz de capturar de manera estable la estructura interna de las activaciones

del modelo de lenguaje.

3. Analizar la relación entre las dimensiones latentes y las categorías emocionales mediante una métrica de selectividad estadística.

4. Evaluar la influencia funcional de las características seleccionadas mediante intervenciones controladas en el espacio latente y medir su impacto en la evidencia asociada con las predicciones emocionales del modelo.

1.7. Metodología general

La metodología de esta investigación se fundamenta en trabajos recientes de interpretabilidad mecanicista de modelos de lenguaje mediante aprendizaje de diccionarios con sparse autoencoders. En particular, busca extraer características latentes a partir de activaciones internas de un modelo de lenguaje, analizar su relación con categorías emocionales y evaluar su influencia funcional mediante intervenciones controladas en el espacio latente.

A diferencia de una evaluación orientada únicamente al desempeño clasificatorio, el objetivo metodológico de esta tesis es estudiar si las representaciones internas del modelo contienen direcciones latentes que, al ser intervenidas, producen cambios sistemáticos y medibles en la evidencia asociada con predicciones emocionales. Para ello, la metodología se organizó en seis etapas: selección y preparación del corpus, extracción de activaciones internas, entrenamiento del sparse autoencoder, construcción del conjunto latente etiquetado, selección e intervención de características latentes y evaluación de los efectos de steering.

1.7.1. Diseño general del estudio

El estudio se diseñó como un análisis experimental de interpretabilidad mecanicista sobre un modelo de lenguaje de la familia LLaMA. Primero, se construyó un corpus textual balanceado con ejemplos asociados a categorías emocionales. Después, cada texto fue procesado por el modelo de lenguaje para extraer activaciones internas de una capa seleccionada. Estas activaciones se utilizaron para entrenar un sparse autoencoder de manera no supervisada.

Una vez entrenado el sparse autoencoder, las activaciones fueron codificadas en un espacio latente disperso. Las etiquetas emocionales del corpus se emplearon posteriormente, no durante el entrenamiento del autoencoder, sino durante el análisis de las dimensiones latentes. A partir de este conjunto latente etiquetado, se seleccionaron características candidatas mediante una métrica de diferencia de medias entre clases emocionales. Finalmente, se realizaron intervenciones controladas sobre dimensiones latentes individuales para medir su efecto en la salida del modelo.

1.7.2. Selección y preparación del corpus

Para vincular las representaciones latentes con emociones humanas explícitas se utilizó como base el benchmark textual *UnifyEmotion* (Bostan y Klinger, 2018). Aunque el recurso original contempla una mayor diversidad de emociones, en esta investigación se trabajó únicamente con cuatro clases: *Happiness*, *Anger*, *Sadness* y *Fear*. Esta reducción tuvo un propósito metodológico. Utilizar un número mayor de emociones incrementaba la ambigüedad entre etiquetas cercanas, dificultaba mantener un balance controlado entre clases y hacía menos estable la evaluación basada en probabilidades y log-odds de etiquetas candidatas. Por ello, se optó por un subconjunto reducido de emociones suficientemente diferenciadas.

El corpus fue filtrado y balanceado antes de la extracción de activaciones. En primer lugar, se conservaron únicamente textos cuya longitud estuviera entre 10 y 80 tokens, de acuerdo con el tokenizador del modelo utilizado. Este filtro permitió descartar ejemplos demasiado cortos, que podrían carecer de evidencia emocional suficiente, y textos excesivamente largos, que podrían introducir variabilidad adicional durante el procesamiento por lotes. En segundo lugar, se construyó un subconjunto balanceado de aproximadamente 8000 ejemplos por clase, dando un total de 32000 ejemplos.

El balanceo tuvo como finalidad reducir sesgos en el ranking de características latentes y facilitar comparaciones entre clases emocionales. Todas las activaciones empleadas en este estudio provienen del mismo conjunto final filtrado y balanceado. El sparse autoencoder se entrenó con las activaciones no etiquetadas de estos ejemplos, mientras que las etiquetas emocionales se utilizaron únicamente en las etapas posteriores de análisis, selección de características e interpretación.

1.7.3. Modelo base y extracción de activaciones internas

El modelo de lenguaje utilizado fue *Llama-3.2-1B-Instruct*, obtenido a través de HuggingFace Hub mediante el identificador `meta-llama/Llama-3.2-1B-Instruct`. Se eligió esta variante por ser un modelo compacto y accesible, suficiente para capturar patrones internos relevantes para el análisis emocional y, al mismo tiempo, viable para la extracción masiva de activaciones y el entrenamiento del autoencoder con recursos computacionales moderados.

Para cada entrada textual, se capturaron activaciones internas correspondientes al flujo residual de una capa del modelo. En particular, se seleccionó la capa 7 como capa principal de análisis. Esta decisión permitió reducir la complejidad experimental y concentrar el estudio en un único nivel de representación, en lugar de extender el análisis a múltiples capas. Las activaciones extraídas constituyeron el conjunto de datos base para entrenar el sparse autoencoder.

Durante la extracción, cada texto fue procesado mediante un prompt de clasificación emocional diseñado para restringir la salida del modelo a un conjunto fijo de etiquetas candidatas. El prompt utilizado fue:

> 'Task: Classify the emotion in the following text. You must answer with exactly one word from this list: Happiness, Anger, Sadness, Fear.
>
> Text to classify: "text"
>
> Your answer (one word only):'

Este formato permite que el modelo produzca una única etiqueta emocional como respuesta y facilita la comparación entre clases mediante probabilidades y log-odds asociadas a los tokens candidatos. Para el entrenamiento del autoencoder se utilizaron activaciones a nivel de token, excluyendo elementos no informativos como padding o componentes fijos del prompt. Posteriormente, durante los experimentos de steering, las intervenciones se aplicaron sobre la activación correspondiente al token final del prompt de clasificación, ya que esta posición condiciona directamente la decisión del modelo sobre la etiqueta emocional.

1.7.4. Entrenamiento del sparse autoencoder

Una vez recopiladas las activaciones internas, se entrenó un sparse autoencoder de una sola capa para aprender una representación latente dispersa capaz de reconstruir las activaciones originales. El autoencoder recibió como entrada vectores de activación de dimensión $d = 2048$ y produjo un código latente de dimensión $d' = 16384$, correspondiente a un factor de expansión de $8\times$.

El entrenamiento se realizó de manera no supervisada; es decir, las etiquetas emocionales no fueron utilizadas para optimizar el autoencoder. El objetivo fue aprender un diccionario latente que permitiera descomponer las activaciones internas en características dispersas, manteniendo una reconstrucción fiel de la activación original. Para inducir esparsidad se utilizó una activación Top-K, iniciando con un valor amplio de K y reduciéndolo progresivamente hasta un valor final de $K = 128$. Esta estrategia permitió comenzar con una representación menos restrictiva y avanzar hacia códigos latentes más dispersos.

Al finalizar el entrenamiento, se aplicó una reparametrización del decodificador para normalizar las columnas a norma unitaria y absorber el escalamiento correspondiente en el codificador. Este procedimiento preserva el mapeo entrada–salida del autoencoder, pero facilita la comparación entre dimensiones latentes durante el análisis posterior. La calidad del autoencoder se evaluó mediante métricas de reconstrucción y activación efectiva de las dimensiones latentes.

1.7.5. Construcción del conjunto latente etiquetado

Después del entrenamiento del sparse autoencoder, las activaciones del corpus fueron codificadas para obtener representaciones latentes asociadas a cada ejemplo textual. A esta etapa se le denominó construcción del conjunto latente etiquetado, ya que cada ejemplo conservó su etiqueta emocional original, pero ahora fue representado mediante el vector latente producido por el codificador del autoencoder.

El uso del mismo corpus para entrenar el autoencoder y construir el conjunto latente etiquetado no introduce supervisión directa en el entrenamiento del SAE, debido a que las etiquetas emocionales no participan en la optimización del modelo. Su uso se limita al análisis posterior, donde permiten calcular asociaciones estadísticas entre dimensiones latentes y categorías emocionales.

De esta forma, el autoencoder aprende una representación no supervisada de las activaciones, y las etiquetas se emplean únicamente para interpretar y seleccionar características candidatas.

1.7.6. Selección de características e intervención latente

Para identificar dimensiones latentes potencialmente relacionadas con señales emocionales, se calculó una métrica de diferencia de medias entre clases. Para cada dimensión latente z_i y cada categoría emocional c, se estimó la diferencia dada por

$$\Delta\mu_{i,c} = \mu(z_i \mid \hat{y} = c) - \mu(z_i \mid \hat{y} \neq c), \tag{1.11}$$

donde $\hat{y}$ corresponde a la etiqueta predicha por el modelo bajo el prompt de clasificación. Esta formulación permite identificar dimensiones que tienden a activarse de manera diferenciada cuando el modelo predice una clase emocional específica frente al resto.

A partir del ranking obtenido con la Ecuación (1.11) se construyó un grupo de características seleccionadas. Como control experimental, también se construyó un grupo de características aleatorias con la misma cardinalidad. Esta comparación permitió evaluar si las características seleccionadas mediante la métrica estadística presentaban efectos de steering más estructurados o intensos que dimensiones latentes arbitrarias.

El steering se implementó como una intervención de una sola característica durante la inferencia. Para cada ejemplo, la activación interna en el punto de intervención fue codificada mediante el sparse autoencoder. Después, una dimensión latente específica fue modificada mediante la actualización

$$z_i \leftarrow z_i + \Delta. \tag{1.12}$$

El vector latente intervenido fue decodificado nuevamente hacia el espacio de activaciones del modelo y reinyectado en el mismo punto del flujo residual. Posteriormente, se observó cómo esta modificación afectaba la distribución de salida del modelo sobre las etiquetas emocionales candidatas. Las magnitudes de intervención Δ de la Ecuación (1.12) se evaluaron en el conjunto siguiente (1.13)

$$\Delta \in \{\pm 5, \pm 10, \pm 15, \pm 20, \pm 25, \pm 30\}, \tag{1.13}$$

incluyendo también el caso base $\Delta = 0$.

1.7.7. Diagnóstico del baseline experimental

Con el fin de transparentar el comportamiento inicial del sistema antes de aplicar intervenciones latentes, se incorporó un diagnóstico del baseline experimental utilizado en este trabajo. En este contexto, el baseline corresponde a la reconstrucción del sparse autoencoder sin aplicar *steering*, evaluada sobre el subconjunto balanceado por etiqueta del dataset.

La Tabla 1.4 presenta las proporciones por fila del baseline experimental. Es decir, dada la etiqueta real del dataset, se muestra qué proporción de ejemplos fue predicha como cada una de las clases candidatas antes de aplicar cualquier intervención latente. Se reportan proporciones, en lugar de conteos absolutos, para facilitar la comparación entre clases y evitar que la interpretación dependa del tamaño específico de cada fila.

Etiqueta real	Anger	Fear	Happiness	Sadness
Anger	**0.3740**	0.2772	0.0000	0.3489
Fear	0.1256	**0.3990**	0.0000	0.4754
Happiness	0.1222	0.2026	**0.0434**	0.6318
Sadness	0.1111	0.1989	0.0000	**0.6900**

Tabla 1.4. Distribución de predicciones del baseline experimental por etiqueta real del dataset.

Los resultados muestran que, aun cuando el subconjunto fue construido buscando balance entre etiquetas del dataset, las predicciones del baseline no fueron balanceadas. En particular, la clase *Happiness* tuvo una representación muy baja en la salida del modelo, mientras que *Sadness* concentró una gran proporción de predicciones. Por ejemplo, solo el 4.34 % de los ejemplos con etiqueta real *Happiness* fueron clasificados como *Happiness*, mientras que el 63.18 % de ellos fueron clasificados como *Sadness*. De forma similar, *Sadness* fue la clase con mayor retención, alcanzando una proporción de 69.00 % sobre su propia fila. Este comportamiento confirma que el baseline experimental presentaba un sesgo predictivo importante antes de aplicar *steering*, lo cual aporta contexto para interpretar los efectos posteriores de las intervenciones latentes.

1.7.8. Métricas de evaluación

Los efectos de las intervenciones se evaluaron mediante cambios en la distribución de salida del modelo, restringida a las etiquetas emocionales consideradas. En particular, se utilizaron dos métricas complementarias: cambio medio en probabilidad y cambio medio en log-odds.

El cambio medio en probabilidad se definió como

$$\Delta p = \mathbb{E}\left[P(y \mid \tilde{x}) - P(y \mid x^{(0)})\right], \tag{1.14}$$

donde $\tilde{x}$ representa la activación intervenida y $x^{(0)}$ corresponde a la activación reconstruida por el autoencoder sin intervención latente. Esta línea base de reconstrucción permite aislar el efecto de modificar una dimensión latente específica, evitando confundirlo con el efecto de reconstruir la activación mediante el SAE.

La segunda métrica fue el cambio medio en log-odds:

$$\Delta \ell = \mathbb{E}\left[\operatorname{logit}(\tilde{p}) - \operatorname{logit}(p^{(0)})\right], \tag{1.15}$$

donde la función logit empleada en (1.15) se define como

$$\operatorname{logit}(p) = \ln\left(\frac{p}{1-p}\right) \tag{1.16}$$

Antes de aplicar la transformación logit de la Ecuación (1.16), las probabilidades fueron recortadas al intervalo $[\varepsilon, 1-\varepsilon]$, con $\varepsilon = 10^{-6}$, para evitar inestabilidades numéricas cuando las probabilidades se aproximan a 0 o 1. El uso de log-odds ((1.15)) permitió observar cambios en la evidencia relativa del modelo que podían ser menos visibles en el espacio de probabilidad definido en (1.14).

Además, se estimó la incertidumbre de los efectos mediante bootstrap pareado sobre los ejemplos de evaluación. Se realizaron 1000 remuestreos y se calcularon intervalos de confianza al 95 % para los cambios en log-odds. Este procedimiento permitió evaluar la estabilidad de los efectos inducidos por las intervenciones sin asumir una distribución paramétrica específica.

1.7.9. Configuración experimental

La Tabla 1.5 resume la configuración principal utilizada en los experimentos.

Componente	Configuración
Modelo base	*Llama-3.2-1B-Instruct*
Capa analizada	Flujo residual, capa 7
Corpus base	*UnifyEmotion*
Clases utilizadas	*Happiness*, *Anger*, *Sadness*, *Fear*
Ejemplos por clase	8000
Total de ejemplos	32000
Filtro de longitud	10 a 80 tokens
Dimensión de activación d	2048
Factor de expansión del SAE	$8\times$
Dimensión latente d'	16384
Top-K final	128
Tasa de aprendizaje	3×10^{-4}
Tamaño de batch	512
Épocas de entrenamiento	30
Magnitudes de intervención	$\{\pm5, \pm10, \pm15, \pm20, \pm25, \pm30\}$
Remuestreos bootstrap	1000

Tabla 1.5. Configuración experimental principal.

Finalmente, los resultados obtenidos fueron interpretados en función de su contribución a la interpretabilidad mecanicista de modelos de lenguaje aplicados al análisis de emociones. En particular, se evaluó si las intervenciones latentes producían efectos sistemáticos y medibles, y si las características seleccionadas mediante criterios estadísticos mostraban diferencias frente a características aleatorias utilizadas como control.

1.8. Cronograma de actividades

El cronograma de actividades se presenta por semestres académicos, correspondientes a los cuatro periodos contemplados para el desarrollo de la Maestría en Ciencias de la Computación (ver tabla 1.6). De manera específica, el Semestre I corresponde al periodo agosto–diciembre de 2024, el Semestre II al periodo enero–junio de 2025, el Semestre III al periodo agosto–diciembre de 2025 y el Semestre IV al periodo enero–junio de 2026. Esta organización permite ubicar temporalmente las principales actividades de investigación, desarrollo experimental, elaboración del

artículo científico y redacción de la tesis.

No.	Actividad	Sem. I	Sem. II	Sem. III	Sem. IV
1	Revisión de literatura	✓	✓	✓	✓
2	Redacción del protocolo	✓			
3	Interpretación de LLMs	✓	✓	✓	✓
3.1	Obtención de los vectores de activación	✓	✓		
3.2	Aplicación del aprendizaje de diccionarios con sparse autoencoders		✓		
3.3	Interpretación de las características obtenidas		✓	✓	
3.4	Documentación y reporte de resultados			✓	✓
4	Redacción y envío de artículo			✓	✓
5	Redacción y revisión de la tesis			✓	✓
6	Presentación y defensa de la tesis				✓

Tabla 1.6. Cronograma de actividades por semestre académico.

Referencias

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, *117*(48), 30071-30078. https://doi.org/10.1073/pnas.1907375117

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, *7*, 49-72.

Bereska, L., & Gavves, E. (2024). Mechanistic Interpretability for AI Safety – A Review. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=ePUVetPKu6

Bostan, L. A. M., & Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, 2104-2119.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., ... Olah, C. (2023). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems 33*, 1877-1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Cheng, Z., Cheng, Z.-Q., He, J.-Y., Sun, J., Wang, K., Lin, Y., Lian, Z., Peng, X., & Hauptmann, A. G. (2024). Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Ins-

truction Tuning. *Advances in Neural Information Processing Systems 37*, 110805-110853. https://doi.org/10.52202/079017-3518

Cunningham, H., Ewart, A., Riggs Smith, L., Huben, R., & Sharkey, L. (2024). Sparse Autoencoders Find Highly Interpretable Features in Language Models. *International Conference on Learning Representations*.

Dalvi, F., Adlakha, V., Rosenthal, S., & Singh, M. (2019). What is one grain of sand in the desert? Analyzing individual neurons in deep NLP models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6309-6317.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, junio). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. En J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Diwali, A., Saeedi, K., Dashtipour, K., Gogate, M., Cambria, E., & Hussain, A. (2024). Sentiment Analysis Meets Explainable Artificial Intelligence: A Survey on Explainable Sentiment Analysis. *IEEE Transactions on Affective Computing*, *15*(3), 837-846. https://doi.org/10.1109/TAFFC.2023.3296373

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. (2024). The Llama 3 Herd of Models.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy Models of Superposition. *arXiv preprint arXiv:2209.10652*.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Commun. ACM*, *56*(4), 82-89. https://doi.org/10.1145/2436256.2436274

Gao, L., Dupré la Tour, T., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., & Wu, J. (2025). Scaling and Evaluating Sparse Autoencoders. *International Conference on Learning Representations (ICLR)*.

He, Z., Shu, W., Ge, X., Chen, L., Wang, J., Zhou, Y., Liu, F., Guo, Q., Huang, X., Wu, Z., Jiang, Y.-G., & Qiu, X. (2024). Llama Scope: Extracting Millions of Features from Llama-3.1-8B with Sparse Autoencoders. *arXiv preprint arXiv:2410.20526*.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramar, J., Dragan, A., Shah, R., & Nanda, N. (2024). Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 278-300. https://doi.org/10.18653/v1/2024.blackboxnlp-1.19

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *ACM Queue*, *16*(3), 31-57. https://doi.org/10.1145/3236386.3241340

Molnar, C. (2025). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (3.ª ed.). https://christophm.github.io/interpretable-ml-book/

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The Building Blocks of Interpretability. *Distill*. https://doi.org/10.23915/distill.00010

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Samek, W., Wiegand, T., & Müller, K.-R. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Vol. 11700). Springer Nature. https://doi.org/10.1007/978-3-030-28954-6

Strapparava, C., & Mihalcea, R. (2007, junio). SemEval-2007 Task 14: Affective Text. En E. Agirre, L. Màrquez & R. Wicentowski (Eds.), *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 70-74). Association for Computational Linguistics. https://aclanthology.org/S07-1013

Sun, C., Huang, L., & Qiu, X. (2019, junio). Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. En J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 380-385). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1035

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., ... Henighan, T. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*. https://transformer-circuits.pub/2024/scaling-monosemanticity/

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. https://doi.org/10.48550/arXiv.2302.13971

Vig, J., & Belinkov, Y. (2019). Analyzing the Structure of Attention in a Transformer Language Model. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://arxiv.org/abs/1906.04284

Zhang, W., Deng, Y., Liu, B., Pan, S., & Bing, L. (2024, junio). Sentiment Analysis in the Era of Large Language Models: A Reality Check. En K. Duh, H. Gomez & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3881-3906). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.246

Capítulo 2

Steering latente de evidencia emocional con sparse autoencoders en modelos de lenguaje

Steering latente de evidencia emocional con sparse autoencoders en modelos de lenguaje

[1]Eduardo Ander Quintero Sánchez *ea.quintero.sanchez@gmail.com*, [1]Cristina López Ramírez *cristina.lopez@ujat.mx*, [2]Ricardo Ramos Aguilar *rramosa@ipn.mx*, [2]Jesús García Ramírez *jegarciara@ipn.mx*, [1]Luis Enrique Ramón Pedrero *242h21005@alumno.ujat.mx*

Resumen: Comprender cómo los modelos de lenguaje codifican información afectiva sigue siendo un desafío central en interpretabilidad mecanicista. En este trabajo se estudia cómo el steering latente con sparse autoencoders (SAEs) afecta la evidencia emocional en un clasificador basado en modelos de lenguaje. Se entrena un SAE sobre activaciones del flujo residual de la capa 7 de Llama-3.2-1B-Instruct y se aplican intervenciones de una sola característica durante clasificación emocional en cuatro clases. Usando características latentes seleccionadas estadísticamente y elegidas al azar, se evalúan los efectos del steering en un rango simétrico de magnitudes de intervención mediante medidas basadas en log-odds respecto a un ancla de reconstrucción del SAE. Los resultados muestran que el steering latente induce cambios estructurados y graduales en la evidencia emocional, con asimetrías direccionales en respuestas agregadas de log-odds y ejemplos representativos a nivel de característica con comportamiento suave y diferenciado por clase. Sin embargo, las comparaciones compactas a nivel de grupo no revelan una ventaja cuantitativa robusta de las características seleccionadas sobre las aleatorias. Estos hallazgos indican que el espacio latente analizado contiene direcciones de intervención que influyen de manera medible en predicciones relacionadas con emociones, pero que dicho control parece distribuido entre dimensiones latentes en lugar de concentrado en un subconjunto especializado pequeño.

Palabras clave: sparse autoencoders, interpretabilidad mecanicista, steering latente, modelos de lenguaje, análisis de emociones.

Institución de adscripción: [1]Universidad Juárez Autónoma de Tabasco, [2]Instituto Politécnico Nacional.

2.1. Introducción

Comprender cómo los modelos de lenguaje representan internamente conceptos de alto nivel sigue siendo un desafío central en la interpretabilidad mecanicista. Aunque los modelos de lenguaje modernos muestran capacidades sólidas en tareas que involucran sentimiento, emoción y razonamiento social, la estructura interna mediante la cual estas señales son codificadas aún no se comprende completamente.

Trabajos recientes han explorado bases representacionales alternativas más allá de las neuronas individuales, particularmente mediante sparse autoencoders (SAEs), los cuales descomponen activaciones en características latentes dispersas. Estos enfoques han mostrado que las representaciones internas pueden analizarse en términos de componentes más interpretables, y que intervenir sobre dichas representaciones mediante *steering* de activaciones puede proporcionar una forma práctica de evaluar su papel causal en el comportamiento del modelo (Bricken et al., 2023; Rimsky et al., 2024; Templeton et al., 2024).

En este trabajo estudiamos el steering latente en un contexto de clasificación emocional. En lugar de enfocarnos en estudios de caso aislados, examinamos cómo responde un modelo de lenguaje cuando dimensiones individuales del SAE son perturbadas a través de un rango simétrico de magnitudes de intervención. Usando activaciones del flujo residual de la capa 7 de Llama-3.2-1B-Instruct, entrenamos un SAE y aplicamos intervenciones de steering sobre características individuales durante la clasificación. Posteriormente evaluamos cómo estas intervenciones modifican la evidencia emocional con respecto a un ancla basada en reconstrucción, usando medidas de efecto agregadas, ejemplos representativos a nivel de característica y estimaciones de incertidumbre mediante bootstrap.

Nuestros resultados muestran que el steering latente induce cambios consistentes, aunque moderados, en la evidencia emocional. A nivel agregado, estos efectos varían suavemente con la magnitud de intervención y exhiben asimetrías direccionales estructuradas, especialmente en el espacio de log-odds, mientras que algunas características individuales producen perfiles de respuesta claros e interpretables. Sin embargo, las comparaciones compactas a nivel de grupo no revelan una ventaja cuantitativa robusta de las características seleccionadas estadísticamente sobre las aleatorias. En conjunto, estos hallazgos apoyan una visión del steering emocional como

un fenómeno real pero distribuido en el espacio latente analizado, y posicionan la contribución principal de este trabajo como una caracterización empírica de cómo las intervenciones latentes basadas en SAEs afectan las predicciones emocionales en la práctica.

El resto del capítulo está organizado de la siguiente manera: la Sección 2.2 revisa el trabajo relacionado; la Sección 2.3 presenta la metodología; las Secciones 2.4–2.6 reportan los principales resultados; la Sección 2.7 discute los hallazgos y limitaciones; y la Sección 2.8 presenta las conclusiones.

2.2. Trabajo relacionado

La interpretabilidad mecanicista busca explicar el comportamiento de los modelos en términos de mecanismos internos que puedan ser validados experimentalmente. Un desafío persistente en esta área es que las neuronas individuales suelen ser polisemánticas debido a la superposición, lo que motiva la búsqueda de bases representacionales alternativas que expongan unidades de cómputo más interpretables (Elhage et al., 2022; Lipton, 2018).

Los sparse autoencoders (SAEs) y enfoques relacionados de aprendizaje de diccionarios han emergido como un método práctico para descomponer activaciones neuronales densas en características latentes dispersas. Trabajos iniciales mostraron que dichas características pueden capturar patrones interpretables que no son visibles a nivel de neuronas individuales (Bricken et al., 2023). Estudios posteriores a gran escala demostraron que este enfoque puede aplicarse a modelos de lenguaje modernos, extrayendo grandes conjuntos de características latentes y permitiendo intervenciones sobre ellas para estudiar efectos conductuales (Templeton et al., 2024). En paralelo, trabajos metodológicos han examinado cómo entrenar y evaluar sparse autoencoders grandes, incluyendo aspectos de escalamiento, calidad de reconstrucción, dispersión y confiabilidad de características (Gao et al., 2025). Suites abiertas como Llama Scope y Gemma Scope reflejan además la creciente madurez de la interpretabilidad basada en SAEs para modelos de lenguaje contemporáneos (He et al., 2024; Lieberum et al., 2024).

Más allá del análisis representacional, los métodos de steering de activaciones proporcionan una vía directa para evaluar la influencia causal de representaciones internas. Contrastive Activation Addition y otros enfoques relacionados de ingeniería de activaciones muestran que

modificar activaciones internas puede modular sistemáticamente el comportamiento del modelo, incluyendo propiedades estilísticas y afectivas como sentimiento y emoción (Konen et al., 2024; Rimsky et al., 2024). En una dirección estrechamente relacionada, se ha mostrado que el sentimiento puede ser capturado aproximadamente por direcciones lineales en espacios de activación de modelos de lenguaje, las cuales pueden identificarse y manipularse mediante intervenciones causales (Tigges et al., 2024).

Nuestro trabajo se ubica en la intersección de estas líneas de investigación. Usamos los latentes de un SAE como variables de intervención en un contexto afectivo, con énfasis en la caracterización conductual agregada en lugar de ejemplos aislados de steering.

2.3. Método

2.3.1. Modelo, datos y fuente de activaciones

Usamos Llama-3.2-1B-Instruct (Dubey et al., 2024) como modelo base. Las entradas provienen de un benchmark etiquetado con emociones, UnifyEmotion (Bostan y Klinger, 2018), y se mapean a cuatro clases: {Happiness, Anger, Sadness, Fear}. Para reducir artefactos relacionados con desbalance en el ranking de características y en las comparaciones por etiqueta, filtramos los ejemplos por longitud, entre 10 y 80 tokens medidos con el tokenizador del modelo, y construimos un subconjunto balanceado por clase con 8,000 instancias por etiqueta.

Las activaciones se extraen del flujo residual en la capa 7 mediante un *forward hook* colocado en la normalización posterior a la atención del bloque. Para el entrenamiento del SAE, almacenamos vectores residuales a nivel de token provenientes del segmento textual de cada prompt, excluyendo el padding y los tokens fijos del prompt.

Durante el steering, las intervenciones se aplican únicamente al vector residual del token final del prompt completo, dejando todas las demás posiciones sin cambios. Esto enfoca la intervención en la posición que condiciona directamente la decisión de clasificación del modelo.

Para el ranking de características y los análisis estratificados, los ejemplos se agrupan de acuerdo con la etiqueta predicha por el modelo bajo el prompt de clasificación, obtenida a partir de los logits del primer token de las palabras candidatas de etiqueta. Estas predicciones se utilizan

en lugar de las anotaciones del corpus para que el ranking y la evaluación permanezcan en el mismo espacio predictivo en el que se aplica el steering.

2.3.2. Sparse autoencoder

Entrenamos un sparse autoencoder (SAE) de una sola capa sobre las activaciones residuales extraídas. En la configuración final, el espacio latente expande la dimensión residual por un factor de $8\times$, de $d = 2048$ a $d' = 16384$. Para un vector de activación $x \in \mathbb{R}^d$, el SAE produce un código latente disperso $z \in \mathbb{R}^{d'}$ y una reconstrucción $\hat{x}$, y se entrena con un objetivo de reconstrucción de error cuadrático medio junto con una regla de activación que induce dispersión sobre z.

La dispersión se impone mediante una regla de activación Top-K. Durante el entrenamiento, el número de latentes activos se reduce progresivamente desde $K = 2048$ hasta un Top-K final de 128, de modo que solo las activaciones latentes positivas más grandes se conservan para cada ejemplo. Después del entrenamiento, las columnas del decodificador se normalizan a norma L_2 unitaria y el escalamiento correspondiente se absorbe en el codificador, siguiendo la práctica estándar de reparametrización de SAEs (He et al., 2024). Esto preserva el mapeo entrada–salida mientras hace que las magnitudes de las características sean más comparables entre dimensiones latentes.

Para mejorar la comparabilidad entre ejemplos, las activaciones se normalizan antes de codificarse usando un factor de norma por ejemplo definido como

$$s(x) = \sqrt{d}/\|x\|_2, \tag{2.1}$$

con un pequeño límite para estabilidad numérica. La codificación y decodificación se realizan sobre la activación normalizada de la Ecuación (2.1), y el vector reconstruido se reescala antes de reinyectarse en el modelo. El SAE se entrena de manera completamente no supervisada; las etiquetas emocionales no se utilizan en esta etapa.

Dado que todas las intervenciones posteriores están mediadas por la reconstrucción del SAE, la calidad del diccionario aprendido afecta directamente la interpretación de los resultados de steering. La Tabla 2.1 reporta tanto la configuración concreta usada en la corrida experimental congelada como los diagnósticos finales de entrenamiento. El SAE resultante alcanza una fide-

lidad de reconstrucción muy alta mientras conserva una estructura latente dispersa no trivial, lo cual respalda su uso como base de intervención para los experimentos posteriores.[1]

Tabla 2.1. Configuración del SAE y diagnósticos finales de entrenamiento.

Componente	Valor
Fuente de activación de entrada	Flujo residual, capa 7
Dimensiones de entrada y latente ($d \to d'$)	$2048 \to 16384$ (expansión $8\times$)
Top-K final	128
Tasa de aprendizaje	3×10^{-4}
Tamaño de lote	512
Épocas de entrenamiento	30
MSE por dimensión	9.7×10^{-5}
Varianza explicada	0.999903
Latentes muertos (post-reparam.)	0.220

2.3.3. Selección de características y steering

Construimos dos grupos fijos y reproducibles de características para comparación. El primer grupo, **Selected**, se obtiene ordenando las dimensiones latentes de acuerdo con una puntuación de diferencia de medias. Para cada etiqueta c, calculamos

$$\Delta\mu_{i,c} = \mu(z_i \mid \hat{y} = c) - \mu(z_i \mid \hat{y} \neq c), \tag{2.2}$$

donde $\hat{y}$ denota la etiqueta predicha por el modelo bajo el prompt de clasificación. Luego seleccionamos las características mejor rankeadas para cada etiqueta y definimos el grupo final Selected como la unión de estos subconjuntos por etiqueta.

El segundo grupo, **Random**, consiste en un subconjunto aleatorio fijo de dimensiones latentes con la misma cardinalidad que Selected, muestreado usando una semilla fija. Este grupo sirve como control para evaluar si el ranking basado en la Ecuación (2.2) identifica direcciones con efectos de steering más estructurados que dimensiones latentes arbitrarias.

El steering se implementa como una intervención de una sola característica durante inferencia. La Figura 2.1 resume el mecanismo: la activación residual en el punto de intervención elegido se codifica mediante el SAE, una coordenada latente se modifica mediante un desplazamiento

[1] Estos diagnósticos se reportan como validación interna de la presente corrida y no como una comparación directa contra estudios previos de SAEs, dado que trabajos recientes suelen evaluar la reconstrucción sobre secuencias retenidas más largas (Gao et al., 2025; He et al., 2024).

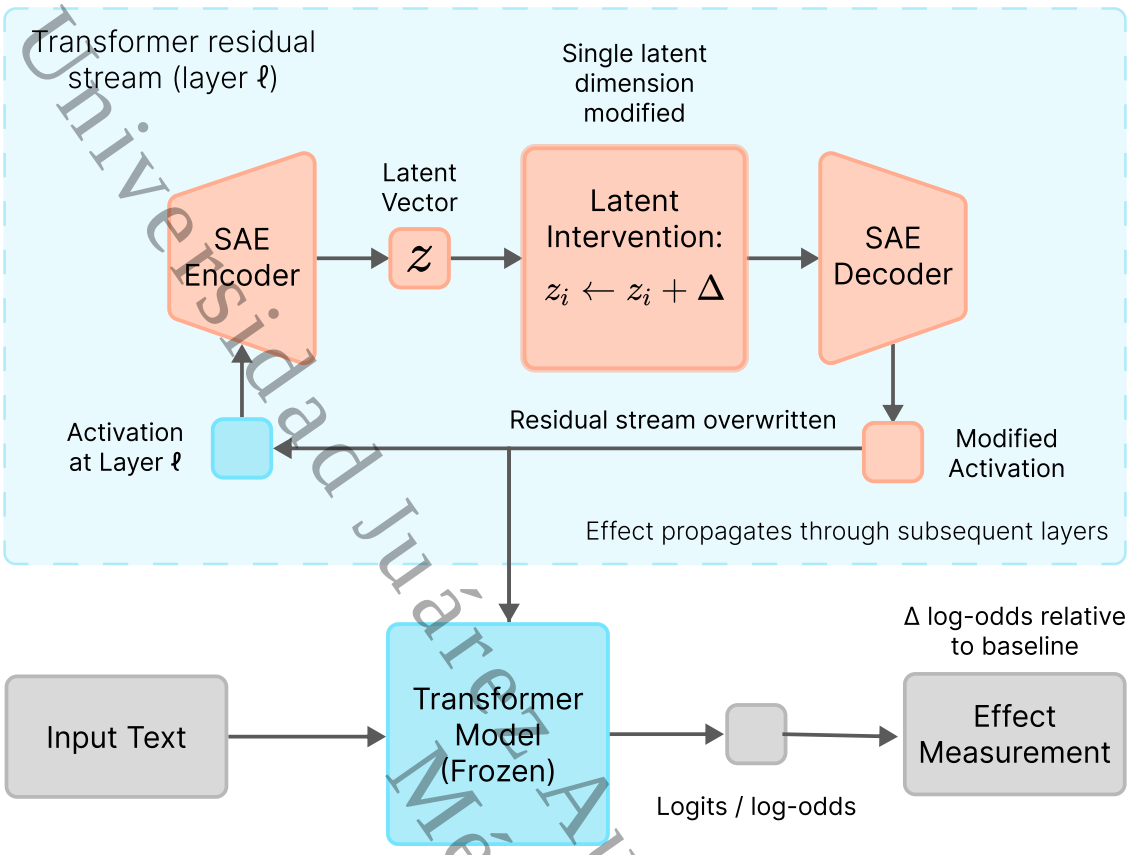


Figura 2.1. Intervención con SAE durante inferencia en la capa ℓ. Una coordenada latente individual se perturba según la Ecuación (2.3), se decodifica de regreso al flujo residual y se evalúa mediante su efecto en los log-odds de salida con respecto a la línea base de reconstrucción ($\Delta = 0$).

aditivo, y la activación decodificada se escribe de regreso en la misma ubicación del flujo residual antes de que continúe el pase hacia adelante.

Concretamente, para un ejemplo dado interceptamos la activación residual en la capa 7 para el token final del prompt, la codificamos con el SAE, modificamos una dimensión latente de acuerdo con la actualización

$$z_i \leftarrow z_i + \Delta, \tag{2.3}$$

decodificamos el código latente modificado y reinyectamos la activación reconstruida en el mismo punto de intervención. Evaluamos un conjunto simétrico de magnitudes de intervención Δ de la Ecuación (2.3) dado por

$$\Delta \in \{\pm 5, \pm 10, \pm 15, \pm 20, \pm 25, \pm 30\}, \tag{2.4}$$

junto con el caso base $\Delta = 0$. El mismo conjunto de intervenciones se utiliza para ambos grupos

de características.

2.3.4. Métricas de evaluación

Los efectos del steering se evalúan a partir de la distribución de siguiente token del modelo en la posición final del prompt, restringida a los tokens de etiqueta. Para cada combinación de característica, magnitud de intervención y etiqueta, comparamos la corrida intervenida contra una corrida ancla sin steering. La línea base o ancla corresponde a la reconstrucción del SAE para el mismo ejemplo con $\Delta = 0$, en lugar de la activación no modificada del modelo. Esto aísla el efecto de la intervención latente del paso de reconstrucción en sí.

Reportamos dos medidas de efecto. La primera es el cambio medio en la probabilidad de etiqueta, definido como

$$\Delta p = \mathbb{E}\Big[P(y \mid \tilde{x}) - P(y \mid x^{(0)})\Big], \tag{2.5}$$

donde $x^{(0)}$ denota la activación ancla reconstruida por el SAE para el mismo ejemplo sin intervención latente. La segunda es el cambio medio en log-odds:

$$\Delta \ell = \mathbb{E}\Big[\operatorname{logit}(\tilde{p}) - \operatorname{logit}(p^{(0)})\Big], \tag{2.6}$$

donde la función logit empleada en (2.6) se define como

$$\operatorname{logit}(p) = \log \frac{p}{1-p}. \tag{2.7}$$

Los log-odds (2.6) se incluyen porque proporcionan una escala de efecto más estable cuando las probabilidades están cerca de 0 o 1. Antes de aplicar la transformación logit de la Ecuación (2.7), las probabilidades se recortan al intervalo $[\varepsilon, 1-\varepsilon]$ con $\varepsilon = 10^{-6}$.

La incertidumbre se estima mediante bootstrap pareado sobre los ejemplos de evaluación usando 1000 remuestreos. Reportamos intervalos de confianza del 95 % para $\Delta\ell$ (2.6); Δp (2.5) se reporta como estimación puntual.

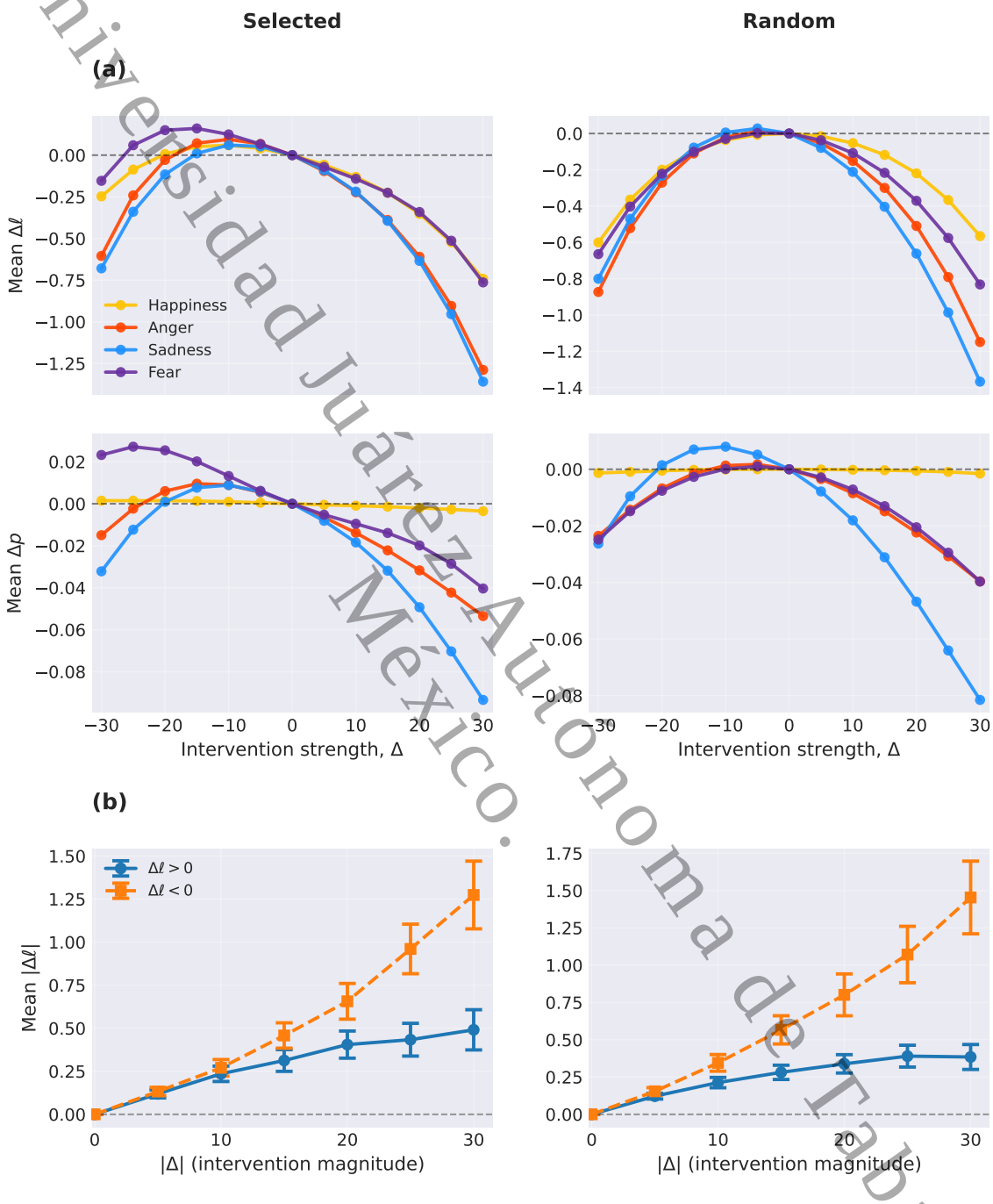


Figura 2.2. Efectos agregados del steering latente para características seleccionadas y aleatorias. El panel (a) muestra el efecto medio con signo del steering como función de la fuerza de intervención Δ. La fila superior reporta el cambio medio en log-odds ($\Delta\ell$), y la fila inferior el cambio medio en probabilidad (Δp), a través de las etiquetas emocionales. El panel (b) muestra el cambio absoluto medio en log-odds, $|\Delta\ell|$, como función de la magnitud de steering $|\Delta|$, separado por el signo del efecto resultante. Los efectos son visualmente más pronunciados en el espacio de log-odds.

2.4. Efectos agregados del steering latente

Primero examinamos la respuesta agregada del modelo al steering latente como función de la fuerza de intervención Δ del conjunto (2.4), resumida en la Fig. 2.2.

En ambos grupos, variar Δ produce desplazamientos suaves y sistemáticos a lo largo del rango de intervención, lo que indica que la respuesta es estructurada y no aleatoria.

La diferencia cualitativa más clara entre grupos aparece en los perfiles de log-odds de la Fig. 2.2a. En el grupo seleccionado, los valores negativos de steering se asocian con un desplazamiento ascendente más pronunciado en la media de $\Delta\ell$, con algunos promedios por clase aproximándose o superando ligeramente el cero, más claramente para Fear. Las características aleatorias exhiben un patrón relacionado pero menos distintivo: sus respuestas promedio se mantienen estructuradas a lo largo del rango de intervención, aunque con un contraste direccional más débil en la región de steering negativo. Esto sugiere una diferencia en el perfil de respuesta agregada más que una simple diferencia en tamaño global del efecto. Al mismo tiempo, el efecto sigue siendo heterogéneo entre etiquetas y debe interpretarse con cautela: no implica que las características seleccionadas individuales se mapeen limpiamente a emociones específicas, ni que el steering aísle representaciones específicas de clase.

Las curvas en espacio de probabilidad de la Fig. 2.2a muestran la misma tendencia de forma atenuada. Los cambios en Δp son visiblemente menores y menos asimétricos que los observados en $\Delta\ell$, lo que es consistente con la compresión inducida por la escala de probabilidad. Esto es especialmente claro para Happiness, cuya curva permanece casi plana a lo largo del rango de intervención. Debido a que la línea base de reconstrucción predice esta etiqueta solo rara vez, desplazamientos sustanciales en la evidencia interna pueden traducirse aun así en cambios modestos de probabilidad. Por esta razón, $\Delta\ell$ ((2.6)) proporciona la vista agregada más informativa de los efectos de steering, mientras que Δp ((2.5)) sirve como una medida complementaria conservadora.

Para evaluar la fuerza de intervención más allá de los promedios con signo, examinamos después el cambio absoluto en log-odds como función de la magnitud de steering, separando los casos de acuerdo con el signo del efecto resultante, como se muestra en la Fig. 2.2b.

La Figura 2.2b confirma una relación dosis–respuesta clara en ambos grupos: conforme au-

menta $|\Delta|$, también aumenta el cambio absoluto medio en log-odds. Sin embargo, este aumento no es direccionalmente simétrico. Tanto en características seleccionadas como aleatorias, los cambios negativos en log-odds ($\Delta\ell < 0$) crecen más rápido y alcanzan magnitudes mayores que los cambios positivos ($\Delta\ell > 0$), por lo que el incremento agregado del efecto absoluto está impulsado principalmente por reducciones más fuertes en la evidencia del modelo. Esta asimetría es visible en ambos grupos y, por tanto, no es exclusiva de las características seleccionadas; de hecho, las características aleatorias tienden a producir desplazamientos negativos ligeramente mayores a magnitudes de intervención altas.

En general, los resultados agregados muestran que el steering latente ejerce un control medible y continuo sobre la evidencia emocional del modelo. Las curvas con signo sugieren algunas diferencias en el perfil de respuesta agregada, pero el análisis de respuesta absoluta muestra que los efectos más fuertes están dominados por cambios negativos y no son exclusivos del grupo seleccionado.

2.5. Perfiles de respuesta a nivel de característica

Los análisis agregados establecen el efecto general del steering latente, pero necesariamente comprimen la heterogeneidad entre características individuales. Para hacer visible esta estructura local, la Figura 2.3 muestra tres casos ilustrativos: una característica seleccionada de alto efecto, una característica seleccionada representativa y una característica aleatoria representativa. Todos los ejemplos fueron elegidos usando la misma métrica preespecificada, es decir, el cambio positivo máximo en log-odds sobre la grilla de intervención.

A través de los ejemplos, las curvas de respuesta son generalmente suaves y organizadas alrededor de $\Delta = 0$, lo que indica cambios direccionales estables a nivel de característica. Su magnitud y signo también varían entre etiquetas dentro de la misma característica, por lo que el steering no amplifica ni suprime uniformemente toda la evidencia emocional. La característica seleccionada de alto efecto proporciona el ejemplo más claro de control local fuerte, mientras que los dos casos representativos muestran que los perfiles suaves y diferenciados por clase no se limitan a ejemplos extremos.

En consistencia con el análisis agregado, el espacio de log-odds proporciona la vista más

clara de este comportamiento. Las curvas correspondientes de Δp conservan tendencias cualitativas similares, pero aparecen más comprimidas, mientras que $\Delta \ell$ hace que las diferencias de sensibilidad por clase sean más fáciles de comparar entre características.

Al mismo tiempo, estos ejemplos no respaldan una interpretación monosemántica fuerte. Incluso cuando una característica está asociada más fuertemente con una etiqueta, su intervención normalmente afecta varias clases a la vez. Por lo tanto, estos perfiles locales muestran que el steering puede producir efectos interpretables a nivel de característica, pero no establecen que las características seleccionadas difieran sistemáticamente de las aleatorias a nivel de grupo.

2.6. Características seleccionadas y aleatorias

Los análisis agregados y los ejemplos representativos a nivel de característica anteriores muestran que el steering latente puede producir desplazamientos estructurados e interpretables en la evidencia emocional. Una pregunta separada, sin embargo, es si las características seleccionadas difieren sistemáticamente de las aleatorias a nivel de grupo. Para abordar esto, comparamos ambos grupos usando estadísticas resumen a nivel de característica diseñadas para capturar magnitud del efecto y forma de respuesta.

La Tabla 2.2 resume la comparación usando un conjunto compacto de métricas en espacio de log-odds y de probabilidad. A través de todas las métricas consideradas aquí, las características seleccionadas y aleatorias muestran valores ampliamente similares, y todos los intervalos de confianza para la diferencia entre grupos se superponen con cero. Bajo estas medidas resumen, las características seleccionadas no exhiben efectos a nivel de característica confiablemente mayores, más asimétricos o más regulares que las aleatorias.

Tabla 2.2. Comparación compacta a nivel de característica entre características seleccionadas y aleatorias. Se reportan medias de grupo por métrica; Dif. denota seleccionadas menos aleatorias con intervalos de confianza bootstrap del 95 %.

Métrica	Seleccionadas	Aleatorias	Dif.	IC 95 %
Media $\lvert\Delta\ell\rvert$	0.521	0.568	-0.047	$[-0.195, 0.096]$
Máx. $\lvert\Delta\ell\rvert$	2.124	2.438	-0.314	$[-0.974, 0.350]$
Asimetría en $\lvert\Delta\ell\rvert$	0.089	0.077	0.012	$[-0.431, 0.435]$
Pendiente local cerca de $\Delta = 0$	-0.013	-0.006	-0.008	$[-0.027, 0.011]$
Media Spearman $\lvert\rho\rvert$	0.794	0.816	-0.022	$[-0.129, 0.075]$
Media $\lvert\Delta p\rvert$	0.031	0.030	0.001	$[-0.005, 0.008]$

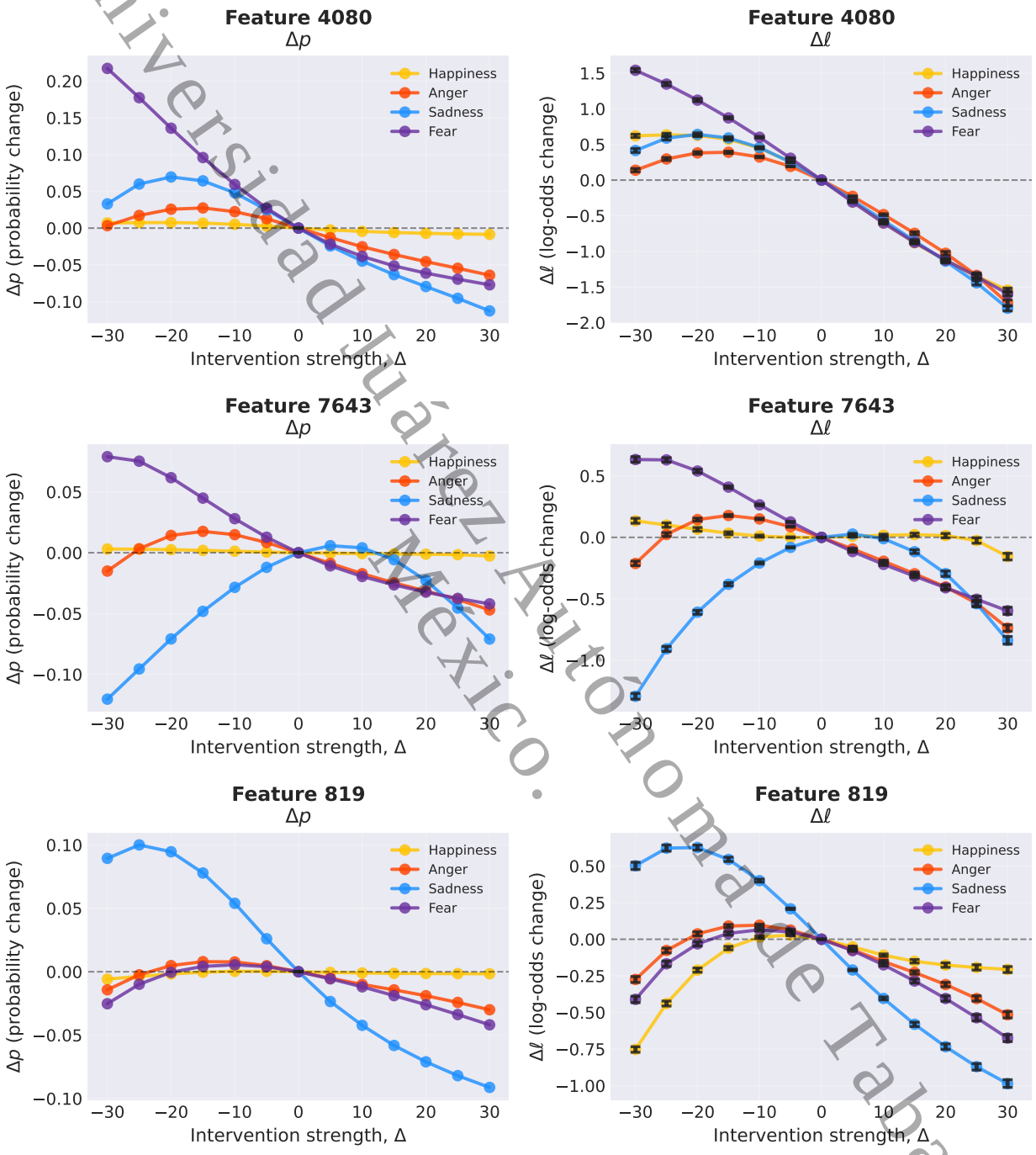


Figura 2.3. Perfiles representativos de steering a nivel de característica para una característica seleccionada de alto efecto, una característica seleccionada representativa y una característica aleatoria representativa. Izquierda: cambio en probabilidad (Δp). Derecha: cambio en log-odds ($\Delta \ell$). Las curvas se muestran a través de las fuerzas de intervención para todas las etiquetas emocionales.

Este resultado ayuda a contextualizar las secciones previas. Aunque las curvas agregadas sugirieron una respuesta con signo algo más direccional para las características seleccionadas, y los ejemplos locales mostraron que algunas direcciones seleccionadas pueden producir efectos fuertes y visualmente interpretables, estos patrones no se traducen en una separación robusta a nivel de grupo una vez que las respuestas se resumen a través de características. En conjunto, los resultados apoyan una visión distribuida de la evidencia emocional en el espacio latente: el procedimiento de selección identifica direcciones útiles para intervención, pero bajo las métricas presentes no aísla un pequeño subconjunto de características con control sistemáticamente más fuerte que alternativas aleatorias.

2.7. Discusión

Los resultados apoyan una visión calificada pero significativa del steering latente en el espacio del SAE. Las intervenciones produjeron cambios sistemáticos en la evidencia emocional a nivel agregado, mientras que los ejemplos representativos mostraron que direcciones latentes individuales pueden producir perfiles locales de respuesta suaves e interpretables. En conjunto, estos hallazgos indican que las intervenciones basadas en SAEs pueden modular de manera medible la evidencia emocional interna del clasificador bajo la configuración estudiada.

Al mismo tiempo, los resultados también limitan la fuerza de las afirmaciones que pueden hacerse. Aunque las características seleccionadas mostraron una respuesta con signo algo más direccional en el análisis agregado, esto no se tradujo en una separación cuantitativa robusta respecto a las características aleatorias bajo las métricas resumen consideradas aquí. Por tanto, la interpretación más plausible no es que el método falle, sino que el control útil se distribuye entre múltiples direcciones latentes en lugar de concentrarse en un subconjunto claramente privilegiado.

En consecuencia, el valor principal del pipeline propuesto reside en proporcionar un método práctico de intervención e inspección. El procedimiento de ranking identifica direcciones que vale la pena explorar, el steering revela cómo dichas direcciones modifican la evidencia por clase, y la comparación entre vistas agregadas y locales ayuda a distinguir casos individuales visualmente convincentes de efectos que generalizan a nivel de grupo.

Varias limitaciones matizan estas conclusiones. Primero, la línea base de reconstrucción permanece desigual entre etiquetas, de manera más notable para Happiness, lo que comprime los efectos en espacio de probabilidad y hace que algunas respuestas a nivel de característica sean más difíciles de interpretar en Δp, incluso cuando los desplazamientos en log-odds siguen siendo sustanciales. Segundo, el presente estudio está restringido a un único modelo, configuración de SAE y esquema de intervención, por lo que queda abierta la medida en que estos patrones generalizan a través de capas, modelos, conjuntos de datos o diccionarios de características. Tercero, el procedimiento de selección de características es intencionalmente simple y podría pasar por alto estructuras más globales o conjuntamente informativas en el espacio latente.

Trabajos futuros podrían, por tanto, evaluar criterios alternativos de selección, extender el análisis a través de capas y modelos, y examinar los efectos de steering con mayor granularidad, por ejemplo a nivel de token o de segmento textual. De manera más amplia, los resultados presentes sugieren que el steering con SAEs es útil para caracterizar cómo las intervenciones latentes modifican la evidencia emocional bajo heterogeneidad realista de características.

2.8. Conclusión

Presentamos un pipeline para seleccionar e intervenir características latentes de un SAE con el fin de estudiar la evidencia emocional en un clasificador basado en un modelo de lenguaje. Los resultados muestran que el steering latente produce cambios sistemáticos y dependientes de la dosis en la evidencia emocional interna, tanto a nivel agregado como en perfiles representativos a nivel de característica. Las comparaciones a nivel de grupo permanecen estrechamente emparejadas entre características seleccionadas y aleatorias bajo las métricas resumen usadas aquí. Estos hallazgos sugieren que el espacio latente del SAE contiene direcciones cuya intervención modifica de manera medible la evidencia emocional, con un control distribuido entre direcciones latentes en lugar de concentrado en un pequeño conjunto de unidades especializadas. En este sentido, el valor principal del pipeline propuesto es proporcionar un marco práctico para explorar y manipular direcciones latentes con efectos interpretables sobre la evidencia emocional.

Referencias

Bostan, L. A. M., & Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, 2104-2119.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., ... Olah, C. (2023). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features/

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. (2024). The Llama 3 Herd of Models.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy Models of Superposition. *arXiv preprint arXiv:2209.10652*.

Gao, L., Dupré la Tour, T., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., & Wu, J. (2025). Scaling and Evaluating Sparse Autoencoders. *International Conference on Learning Representations (ICLR)*.

He, Z., Shu, W., Ge, X., Chen, L., Wang, J., Zhou, Y., Liu, F., Guo, Q., Huang, X., Wu, Z., Jiang, Y.-G., & Qiu, X. (2024). Llama Scope: Extracting Millions of Features from Llama-3.1-8B with Sparse Autoencoders. *arXiv preprint arXiv:2410.20526*.

Konen, K., Jentzsch, S., Diallo, D., Schütt, P., Bensch, O., El Baff, R., Opitz, D., & Hecking, T. (2024). Style Vectors for Steering Generative Large Language Models. *Findings of the Association for Computational Linguistics: EACL 2024*, 782-802. https://doi.org/10.18653/v1/2024.findings-eacl.52

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramar, J., Dragan, A., Shah, R., & Nanda, N. (2024). Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 278-300. https://doi.org/10.18653/v1/2024.blackboxnlp-1.19

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *ACM Queue*, *16*(3), 31-57. https://doi.org/10.1145/3236386.3241340

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. (2024). Steering Llama 2 via Contrastive Activation Addition. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15504-15522. https://doi.org/10.18653/v1/2024.acl-long.828

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., ... Henighan, T. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*. https://transformer-circuits.pub/2024/scaling-monosemanticity/

Tigges, C., Hollinsworth, O. J., Geiger, A., & Nanda, N. (2024). Language Models Linearly Represent Sentiment. *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. https://doi.org/10.18653/v1/2024.blackboxnlp-1.5

Capítulo 3

Conclusiones y recomendaciones generales

3.1. Principales hallazgos

Esta investigación tuvo como objetivo analizar el uso de *sparse autoencoders* como herramienta de interpretabilidad mecanicista para estudiar representaciones internas de modelos de lenguaje en tareas de análisis de emociones. En particular, se buscó evaluar si las activaciones internas de un modelo de lenguaje podían proyectarse a un espacio latente disperso y si la intervención controlada de dicho espacio producía cambios sistemáticos y medibles en la evidencia asociada con las predicciones emocionales del modelo.

A diferencia de una evaluación centrada exclusivamente en el desempeño clasificatorio, esta tesis se enfocó en estudiar el comportamiento interno del modelo a partir de sus activaciones. Para ello, se utilizó un modelo de lenguaje de la familia LLaMA, un corpus textual con categorías emocionales, un *sparse autoencoder* entrenado sobre activaciones internas y un procedimiento de intervención latente orientado a medir cambios en probabilidad y log-odds. En este sentido, el artículo incluido en el Capítulo 2 constituye el núcleo experimental de la investigación, mientras que este capítulo presenta una interpretación integradora de los resultados en relación con los objetivos, la hipótesis y los alcances de la tesis.

Los resultados muestran que el espacio latente aprendido por el *sparse autoencoder* permite

intervenir dimensiones específicas y producir cambios medibles en la evidencia emocional del modelo. Las variaciones observadas fueron más claras al analizar los efectos en términos de log-odds, lo que sugiere que esta métrica resulta más sensible para observar cambios en la evidencia relativa asociada con cada clase emocional. En cambio, los cambios en probabilidad fueron más conservadores y, en algunos casos, menos expresivos debido a la compresión propia de la escala probabilística.

También se observó que algunas dimensiones latentes individuales producen perfiles de respuesta suaves y diferenciados entre clases emocionales. Estos perfiles permiten inspeccionar cómo ciertas intervenciones modifican la evidencia del modelo a favor o en contra de determinadas emociones. Sin embargo, las comparaciones agregadas entre características seleccionadas y características aleatorias no mostraron una ventaja cuantitativa robusta de las primeras bajo las métricas utilizadas. Por ello, los resultados no deben interpretarse como evidencia de características emocionales perfectamente especializadas o monosemánticas, sino como evidencia de que el espacio latente aprendido contiene direcciones intervenibles cuya influencia sobre la evidencia emocional parece estar distribuida.

Cumplimiento de objetivos

En relación con el primer objetivo específico, se obtuvieron activaciones internas representativas de una capa del modelo de lenguaje a partir de un corpus textual con categorías emocionales. Este paso permitió construir la base experimental necesaria para analizar representaciones internas más allá de las salidas finales del modelo.

Respecto al segundo objetivo, se entrenó un *sparse autoencoder* capaz de proyectar dichas activaciones a un espacio latente disperso y reconstruirlas posteriormente. Esta representación latente permitió estudiar dimensiones individuales como posibles unidades de análisis e intervención.

En cuanto al tercer objetivo, se analizó la relación entre dimensiones latentes y categorías emocionales mediante una métrica estadística de selectividad basada en diferencias de medias. Este procedimiento permitió seleccionar características candidatas asociadas con determinadas clases emocionales, aunque los resultados posteriores mostraron que dicha selección no implicó

necesariamente una superioridad robusta frente a características aleatorias.

Finalmente, el cuarto objetivo se cumplió mediante intervenciones controladas en el espacio latente. Al modificar dimensiones latentes individuales y reinyectar la activación reconstruida en el modelo, fue posible medir cambios en la evidencia emocional mediante probabilidad y log-odds. Este procedimiento permitió evaluar la influencia funcional de las dimensiones latentes sobre la salida del modelo.

Valoración de la hipótesis

La hipótesis de esta tesis establece que el aprendizaje de diccionarios mediante *sparse autoencoders* sobre activaciones internas de un modelo de lenguaje permite proyectar dichas activaciones a un espacio latente disperso, cuya intervención controlada produce cambios sistemáticos y medibles en la evidencia asociada a las predicciones emocionales del modelo.

Con base en los resultados obtenidos, la hipótesis se considera sustentada en sus componentes principales. En primer lugar, el *sparse autoencoder* permitió proyectar activaciones internas del modelo a un espacio latente disperso. En segundo lugar, la intervención de dimensiones latentes produjo cambios observables y sistemáticos en la evidencia emocional del modelo. En tercer lugar, dichos cambios fueron cuantificables mediante métricas como probabilidad y log-odds.

No obstante, los resultados también delimitan el alcance de la interpretación. La evidencia obtenida no permite afirmar que las características seleccionadas correspondan a unidades emocionales claramente especializadas ni que superen de forma robusta a características aleatorias bajo las métricas agregadas utilizadas. Por tanto, la hipótesis no debe entenderse como una afirmación de monosemanticidad emocional fuerte, sino como una afirmación sobre la existencia de un espacio latente intervenible que afecta de manera medible la evidencia emocional del modelo.

Esta distinción es importante porque permite evitar una sobreinterpretación de los resultados. El principal hallazgo no es que el método haya aislado un conjunto reducido de características emocionales puras, sino que las representaciones latentes aprendidas por el SAE pueden ser utilizadas como un espacio de intervención para estudiar cómo se modifica la evidencia emocional interna del modelo.

3.2. Aportes teóricos, metodológicos o prácticos

El principal aporte teórico de esta tesis consiste en vincular el análisis de emociones en modelos de lenguaje con herramientas de interpretabilidad mecanicista basadas en *sparse autoencoders*. En lugar de abordar el análisis emocional únicamente como una tarea de clasificación, el trabajo lo plantea como un problema de representación interna: cómo ciertas activaciones y dimensiones latentes participan en la evidencia que el modelo utiliza para favorecer una etiqueta emocional.

Desde el punto de vista metodológico, la tesis aporta un flujo experimental que integra extracción de activaciones internas, entrenamiento de un *sparse autoencoder*, análisis de dimensiones latentes, selección estadística de características, intervención mediante *steering* y evaluación de efectos sobre la salida del modelo. Esta integración permite estudiar no solo asociaciones entre características latentes y etiquetas emocionales, sino también el efecto funcional de intervenir dichas características durante la inferencia.

Otro aporte metodológico es la comparación entre características seleccionadas y características aleatorias. Esta comparación actúa como control experimental y permite matizar la interpretación de los resultados. Aunque algunas características seleccionadas muestran efectos visualmente interpretables a nivel individual, la ausencia de una ventaja agregada robusta frente a características aleatorias sugiere que la evidencia emocional no se concentra necesariamente en un conjunto pequeño de dimensiones privilegiadas. Esta conclusión es relevante porque evita atribuir especialización emocional fuerte a características que podrían formar parte de un control distribuido en el espacio latente.

En términos prácticos, la investigación proporciona una base para explorar modelos de lenguaje desde una perspectiva más transparente e intervenible. Aunque el método no se plantea como una estrategia para mejorar directamente la precisión de clasificación, sí permite examinar cómo ciertas intervenciones internas modifican la evidencia asociada con predicciones emocionales. Esto puede ser útil para investigaciones futuras sobre control, análisis de sesgos, seguridad e interpretabilidad de modelos de lenguaje en tareas sensibles.

3.3. Recomendaciones

Como primera recomendación, se sugiere extender el análisis a diferentes capas del modelo. Esta tesis se concentró en una capa específica para mantener un diseño experimental controlado; sin embargo, las representaciones emocionales podrían distribuirse de manera distinta a lo largo de la arquitectura. Evaluar múltiples capas permitiría identificar si existen niveles del modelo más adecuados para el análisis e intervención de información afectiva.

También se recomienda evaluar el método en modelos de lenguaje de mayor tamaño o en arquitecturas distintas. El uso de un modelo compacto permitió realizar experimentos viables con recursos computacionales moderados, pero modelos con mayor capacidad podrían presentar estructuras latentes más ricas o patrones de respuesta diferentes. Esta comparación ayudaría a determinar si los efectos observados son propios de la configuración utilizada o si se mantienen en modelos más amplios.

Otra línea de trabajo futuro consiste en explorar criterios alternativos para seleccionar características latentes. La diferencia de medias utilizada en esta investigación permitió construir un ranking inicial de dimensiones candidatas, pero los resultados sugieren que dicha métrica no fue suficiente para aislar un subconjunto con ventaja cuantitativa robusta frente a características aleatorias. Futuros trabajos podrían incorporar criterios multivariados, métricas basadas en información mutua, análisis causal más fino o métodos que consideren interacciones entre características.

Se recomienda además ampliar el conjunto de emociones estudiadas y evaluar corpus con mayor diversidad lingüística y contextual. En esta tesis se utilizó un subconjunto reducido de emociones para favorecer la estabilidad experimental; sin embargo, fenómenos como ambivalencia emocional, ironía, sarcasmo o emociones mixtas podrían requerir esquemas de análisis más amplios y sofisticados.

Finalmente, se recomienda complementar el análisis cuantitativo con inspección cualitativa de ejemplos representativos. La combinación de métricas agregadas, perfiles de respuesta por característica y análisis de casos individuales permite evitar conclusiones excesivamente generales y ofrece una visión más precisa del papel funcional de las dimensiones latentes en las predicciones emocionales del modelo.

Alojamiento de la Tesis en el Repositorio Institucional	
Título de la tesis:	Interpretabilidad Mecanicista de LLMs aplicada al Análisis de Sentimientos y Emociones
Autor:	Eduardo Ander Quintero Sánchez
ORCID:	https://orcid.org/0009-0007-6372-7257
Resumen:	Los modelos de lenguaje grande han mostrado un desempeño relevante en tareas de procesamiento de lenguaje natural, incluido el análisis y clasificación de emociones. Sin embargo, su creciente complejidad dificulta comprender qué representaciones internas influyen en sus predicciones. Esta tesis aborda este problema mediante el uso de sparse autoencoders para descomponer activaciones internas en características latentes potencialmente interpretables. La metodología consistió en obtener activaciones internas de un modelo de lenguaje de la familia LLaMA a partir de un corpus etiquetado con categorías emocionales, entrenar un sparse autoencoder y analizar la relación entre sus dimensiones latentes y las predicciones del modelo. Posteriormente, se realizaron intervenciones controladas sobre características latentes seleccionadas y aleatorias, midiendo sus efectos mediante cambios en probabilidad y log-odds. Los resultados muestran que el steering latente mediante sparse autoencoders produce cambios medibles y estructurados en la evidencia emocional del modelo. Algunas características individuales presentan perfiles de respuesta diferenciados entre clases emocionales, lo que permite inspeccionar cómo ciertas intervenciones modifican la salida del modelo. No obstante, las comparaciones agregadas no muestran una ventaja cuantitativa robusta frente a características aleatorias. En conjunto, la tesis muestra que los sparse autoencoders pueden utilizarse para analizar e intervenir representaciones internas de modelos de lenguaje en tareas de análisis de emociones. Aunque los resultados no implican características emocionales perfectamente especializadas, sí evidencian que el espacio latente aprendido contiene direcciones cuya intervención modifica de manera sistemática la evidencia emocional del modelo.
Palabras clave:	sparse autoencoders, interpretabilidad mecanicista, modelos de lenguaje, steering latente, análisis de emociones
Referencias citadas:	En la siguiente página se muestran las referencias.

Bibliografía

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2020). Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, *117*(48), 30071-30078. https://doi.org/10.1073/pnas.1907375117

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, *7*, 49-72.

Bereska, L., & Gavves, E. (2024). Mechanistic Interpretability for AI Safety – A Review. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=ePUVetPKu6

Bostan, L. A. M., & Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, 2104-2119.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., ... Olah, C. (2023). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems 33*, 1877-1901. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

Cheng, Z., Cheng, Z.-Q., He, J.-Y., Sun, J., Wang, K., Lin, Y., Lian, Z., Peng, X., & Hauptmann, A. G. (2024). Emotion-LLaMA: Multimodal Emotion Recognition and Reasoning with Instruction

Tuning. *Advances in Neural Information Processing Systems 37*, 110805-110853. https://doi.org/10.52202/079017-3518

Cunningham, H., Ewart, A., Riggs Smith, L., Huben, R., & Sharkey, L. (2024). Sparse Autoencoders Find Highly Interpretable Features in Language Models. *International Conference on Learning Representations*.

Dalvi, F., Adlakha, V., Rosenthal, S., & Singh, M. (2019). What is one grain of sand in the desert? Analyzing individual neurons in deep NLP models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 6309-6317.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, junio). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. En J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Diwali, A., Saeedi, K., Dashtipour, K., Gogate, M., Cambria, E., & Hussain, A. (2024). Sentiment Analysis Meets Explainable Artificial Intelligence: A Survey on Explainable Sentiment Analysis. *IEEE Transactions on Affective Computing*, *15*(3), 837-846. https://doi.org/10.1109/TAFFC.2023.3296373

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., et al. (2024). The Llama 3 Herd of Models.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy Models of Superposition. *arXiv preprint arXiv:2209.10652*.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Commun. ACM*, *56*(4), 82-89. https://doi.org/10.1145/2436256.2436274

Gao, L., Dupré la Tour, T., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., & Wu, J. (2025). Scaling and Evaluating Sparse Autoencoders. *International Conference on Learning Representations (ICLR)*.

He, Z., Shu, W., Ge, X., Chen, L., Wang, J., Zhou, Y., Liu, F., Guo, Q., Huang, X., Wu, Z., Jiang, Y.-G., & Qiu, X. (2024). Llama Scope: Extracting Millions of Features from Llama-3.1-8B with Sparse Autoencoders. *arXiv preprint arXiv:2410.20526*.

Konen, K., Jentzsch, S., Diallo, D., Schütt, P., Bensch, O., El Baff, R., Opitz, D., & Hecking, T. (2024). Style Vectors for Steering Generative Large Language Models. *Findings of the Association for Computational Linguistics: EACL 2024*, 782-802. https://doi.org/10.18653/v1/2024.findings-eacl.52

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramar, J., Dragan, A., Shah, R., & Nanda, N. (2024). Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 278-300. https://doi.org/10.18653/v1/2024.blackboxnlp-1.19

Lipton, Z. C. (2018). The Mythos of Model Interpretability. *ACM Queue*, *16*(3), 31-57. https://doi.org/10.1145/3236386.3241340

Molnar, C. (2025). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (3.ª ed.). https://christophm.github.io/interpretable-ml-book/

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The Building Blocks of Interpretability. *Distill*. https://doi.org/10.23915/distill.00010

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. (2024). Steering Llama 2 via Contrastive Activation Addition. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15504-15522. https://doi.org/10.18653/v1/2024.acl-long.828

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Samek, W., Wiegand, T., & Müller, K.-R. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Vol. 11700). Springer Nature. https://doi.org/10.1007/978-3-030-28954-6

Strapparava, C., & Mihalcea, R. (2007, junio). SemEval-2007 Task 14: Affective Text. En E. Agirre, L. Màrquez & R. Wicentowski (Eds.), *Proceedings of the Fourth International Workshop on*

Semantic Evaluations (SemEval-2007) (pp. 70-74). Association for Computational Linguistics. https://aclanthology.org/S07-1013

Sun, C., Huang, L., & Qiu, X. (2019, junio). Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. En J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 380-385). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1035

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., ... Henighan, T. (2024). Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits Thread*. https://transformer-circuits.pub/2024/scaling-monosemanticity/

Tigges, C., Hollinsworth, O. J., Geiger, A., & Nanda, N. (2024). Language Models Linearly Represent Sentiment. *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. https://doi.org/10.18653/v1/2024.blackboxnlp-1.5

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. https://doi.org/10.48550/arXiv.2302.13971

Vig, J., & Belinkov, Y. (2019). Analyzing the Structure of Attention in a Transformer Language Model. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://arxiv.org/abs/1906.04284

Zhang, W., Deng, Y., Liu, B., Pan, S., & Bing, L. (2024, junio). Sentiment Analysis in the Era of Large Language Models: A Reality Check. En K. Duh, H. Gomez & S. Bethard (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 3881-3906). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-naacl.246