

Observabilidad de la inferencia lógica en modelos de lenguaje tipo transformer

Matias Marcelo Rodríguez Matus (G1)

Tabla de contenidos

1	Introducción	2
1.1	Contexto y motivación científica	2
1.2	Objetivos del trabajo y pregunta de investigación	2
1.3	Estructura del documento	3
2	Marco teórico	3
2.1	Relevamiento de trabajos previos y relevantes	3
2.2	Conceptos y técnicas de ciencia de datos utilizados	4
3	Metodología	5
3.1	Presentación y descripción de los datos	5
3.2	Ánálisis exploratorio de datos	5
3.3	Preprocesamiento y limpieza de datos	11
3.4	Descripción de las técnicas de análisis y modelado	11
3.5	Descripción de la selección de características	15
3.6	Descripción de las métricas de evaluación	16
3.7	Descripción de los métodos estadísticos utilizados	17
4	Resultados y discusión	18
4.1	Presentación y análisis de resultados	18
4.2	Discusión de los resultados y su relevancia	22
4.3	Limitaciones y posibles mejoras	23
5	Conclusión	25
5.1	Resumen de los hallazgos principales	25
5.2	Conclusiones generales y relación con los objetivos	26
5.3	Recomendaciones para futuros trabajos	26
5.4	Consideraciones finales	28
6	Bibliografía	28
6.1	Referencias bibliográficas citadas	28
6.2	Otras fuentes consultadas	29
7	Anexos	29
7.1	Código fuente utilizado en el análisis	29
7.2	Tablas y gráficos adicionales	29
7.3	Otros materiales relevantes	33

1. Introducción

1.1. Contexto y motivación científica

Desde su introducción, la arquitectura Transformer ha superado ampliamente tanto a redes LSTM/ELMo como a modelos generativos GPT en métricas estándar de comprensión de lenguaje: en el benchmark **GLUE**, BERT-large alcanza 80.5 puntos frente a 71.0 de ELMo-LSTM y 72.8 de GPT (Devlin et al., 2019; Wang et al., 2018; Radford et al., 2018). Sus representaciones internas permiten resolver tareas clásicas del **pipeline de Procesamiento del Lenguaje Natural (PLN)** etiquetado de parte-de-habla, coreferencias, dependencia gramatical que en **Redes Neuronales Recurrentes (RNN)** requerían modelos más complejos. Investigaciones sobre BERT —uno de los modelos basados en la arquitectura Transformer más estudiados y adoptados como referencia— evidencian que sus espacios de activación separan subespacios semánticos y sintácticos con gran precisión, distinguiendo incluso sentidos de palabra con matices semánticos sutiles (Coenen et al., 2019). Más aún, estudios de *probing* muestran que estos modelos almacenan información necesaria para la **inferencia lógica** sin entrenamiento supervisado específico, superando baselines distribucionales (promedios de embeddings estáticos como word2vec/GloVe) y baselines basados en redes neuronales recurrentes (LSTM/GRU) (Chen & Gao, 2022).

Este progreso motiva la pregunta de si, más allá de correlaciones superficiales, los Transformers **codifican reglas de inferencia que fundamentan la consecuencia lógica (*logical entailment*)**. Verificar tal emergencia resulta metodológicamente más económico que imponerla mediante *fine-tuning* y podría habilitar, a mediano plazo, la **aplicación directa de restricciones lógicas en sistemas generativos** sin penalizar su flexibilidad.

Se seleccionó **RoBERTa-base** porque hereda la arquitectura BERT optimizada para comprensión y elimina objetivos de entrenamiento superfluos, mejorando su rendimiento sin introducir ruido adicional; además, al trabajar sin *fine-tuning* evitamos confundir la emergencia espontánea de inferencia lógica con artefactos de entrenamiento supervisado en NLI. Utilizar **RoBERTa sin ningún fine-tuning** resulta necesario para aislar la variable “emergencia espontánea” y descartar que la eventual presencia de inferencia lógica sea un artefacto del entrenamiento supervisado para *Natural Language Inference* (NLI).

1.2. Objetivos del trabajo y pregunta de investigación

¿Codifican los espacios vectoriales de RoBERTa-base, entrenado de manera general y sin fine-tuning, reglas de inferencia de la lógica de enunciados y, subsidiariamente, de la lógica de predicados de primer orden?

Objetivo general

- Determinar empíricamente la presencia (o ausencia) de estructuras geométricas que correspondan a reglas de inferencia lógica en las representaciones internas de RoBERTa.

Objetivos específicos

1. Reproducir los experimentos de Chen & Gao (2022) —centrados en el *probing* de información lingüística para inferencia lógica— y estudios afines, adaptándolos a nuestro dominio lógico-semántico.
2. Aplicar **reducción dimensional (UMAP, ZCA-PCA)** y **clustering (k-means)** sobre embeddings de hipótesis y premisas del dataset **SNLI** (principal) y **FOLIO** (validación LPO).
3. Medir **Purity** y **NMI** de los clusters y entrenar **árboles de decisión** como *probes* para inspeccionar la alineación con reglas lógicas.
4. Detectar el grado de **anisotropía** de los espacios resultantes, implementar correcciones y evaluar su impacto en las métricas anteriores.
5. Analizar la correspondencia empírica entre los hallazgos y la teoría semántica de modelos de la lógica clásica (Gamut) a fin de contextualizar los resultados.

1.3. Estructura del documento

El trabajo se organiza de la siguiente manera:

- **Cap. 1 Introducción**, Motivación, pregunta y objetivos.
- **Cap. 2 Marco teórico**, Revisión bibliográfica de lógica formal, semántica distribuida y trabajos sobre estructura lógica en LLMs.
- **Cap. 3 Metodología**, Descripción detallada de datasets, preprocesamiento, métricas y pipeline experimental.
- **Cap. 4 Resultados y discusión**, Presentación cuantitativa y cualitativa de resultados, análisis crítico y comparación con literatura.
- **Cap. 5 Conclusión**, Síntesis de aportes, limitaciones y líneas futuras.
- **Cap. 6 Bibliografía**, Fuentes citadas.
- **Cap. 7 Anexos**, Código y material suplementario.

2. Marco teórico

2.1. Relevamiento de trabajos previos y relevantes

Se proponen los siguientes papers para el desarrollo del trabajo:

1. Estructura geométrica de los Transformers.

- *Visualizing & Measuring the Geometry of BERT* (Coenen et al., 2019) demuestra, mediante UMAP y métricas de dispersión, que BERT segregá información semántica y sintáctica en subespacios lineales diferenciados.
- El hallazgo legitima la búsqueda de **otros subespacios especializados** –por ejemplo, un subespacio “lógico” responsable de la inferencia.

2. Codificación de inferencia lógica mediante probing.

- *Probing Linguistic Information for Logical Inference in Pre-trained Language Models* (Chen & Gao, 2022) aplica clasificadores lineales sobre embeddings sin fine-tuning y verifica que los modelos codifican operadores lógicos (\neg , \Box , \Diamond) y ciertas reglas de inferencia proposicional.
- Constituye el antecedente metodológico directo de este trabajo: extendemos sus *probes* a configuraciones no lineales (árboles de decisión) y a tareas de **entailment** de oraciones completas.

3. Propiedades globales de la nube de embeddings.

- *Isotropy in the Contextual Embedding Space* (Cai et al., 2021) revela que los embeddings contextualizados presentan **anisotropía** –concentración de varianza en pocas direcciones– y propone métricas y normalizaciones (ZCA, “all-but-the-top”) para mitigarla.
- Estas técnicas son adoptadas aquí como paso previo al análisis de inferencia, con la hipótesis de que un espacio más isotrópico favorece la detección de regularidades lógicas.

Además, los fundamentos de **consecuencia lógica y semántica de modelos** se enmarcan en la exposición clásica de Gamut (1991), que define la relación \Box entre premisas y conclusión y sirve de referencia conceptual para interpretar los resultados.

2.2. Conceptos y técnicas de ciencia de datos utilizados

A continuación se resumen los conceptos y herramientas fundamentales empleados en el estudio:

- **Embeddings contextualizados:** vectores de 768 dimensiones extraídos de las capas 9–12 de RoBERTa-base, donde suele concentrarse la información semántica.
- **Vectores de relación (Δ):** diferencia *hiptesis – premisa* que captura la transformación semántica dentro de cada par. **Nota metodológica:** Este convenio se mantiene consistentemente a lo largo del estudio; expresiones como $\Delta_E - \Delta_C$ operan sobre vectores ya computados con esta definición.
- **Vector de contraste ($\Delta_{\{EC\}}$):** diferencia entre los vectores de relación de *entailment* y *contradiction*, $\Delta_E - \Delta_C$. Basado en la dualidad lógica explicada en Gamut (1991), donde $\text{Entail}(P, H) \iff \neg \text{Contradict}(P, H)$, anula el contenido léxico compartido y destaca la señal de la etiqueta lógica.
- **Análisis contrastivo:** Constituye una técnica metodológica desarrollada en este trabajo para explorar la codificación geométrica de relaciones lógicas en espacios vectoriales de embeddings. Esta aproximación se fundamenta en la hipótesis de que las operaciones lógicas se manifiestan como transformaciones regulares y direccionales en el espacio de representación, siguiendo la tradición iniciada por los trabajos sobre analogías vectoriales de Mikolov et al. (2013). El método opera construyendo vectores de contraste que capturan la diferencia entre representaciones de relaciones lógicas opuesta, específicamente, el vector de contraste $\Delta_{\{EC\}}$ permite aislar la señal informativa de la inferencia lógica del ruido léxico compartido entre premisas e hipótesis, bajo la premisa teórica de que si los embeddings codifican estructura semántico-lógica, entonces relaciones del mismo tipo deberían generar patrones vectoriales consistentes independientemente del contenido lexical específico.
- **Reducción dimensional:** métodos lineales (PCA, ZCA-whitening) y no lineales (UMAP) que facilitan la exploración visual y la aplicación de algoritmos no supervisados.
- **Normalización:** técnicas para uniformizar la escala y dispersión de embeddings, incluyendo all-but-the-mean (eliminación de componentes principales dominantes), per-type normalization (media por categoría), standard scaling (media=0, desviación estándar=1) y L2 normalization (vectores unitarios).
- **Principal component removal (Deflación):** enfoque sistemático de eliminación de las primeras K componentes principales (top-K PC removal) para reducir el efecto de anisotropía y resaltar señales subyacentes.
- **Clustering:** uso de k-means sobre proyecciones UMAP 2D para detectar agrupamientos alineados con las etiquetas de inferencia; la calidad se mide con **Purity** y **NMI**.
- **Probes:** clasificador simple (en este caso, un árbol de decisión) entrenado sobre los vectores de embedding para predecir la etiqueta lógica de cada ejemplo. Si el probe logra una precisión significativamente superior al azar, esto indica que la información relevante para la inferencia lógica está presente y es accesible en el espacio de embedding.
- **Anisotropía:** rasgo observado en el cual los embeddings contextualizados de modelos como BERT o RoBERTa tienden a agruparse en un “cono” muy estrecho del espacio vectorial, de modo que la mayoría de los vectores puntúan alto en similitud de coseno entre sí. Esta característica, resultado del entrenamiento de los Transformers, dificulta la identificación de ejes semánticos y lógicos diferenciados (Cai et al., 2021).

- **Isotropía:** estado en el cual los embeddings se distribuyen de manera más uniforme en todas las direcciones del espacio vectorial, reduciendo sesgos causados por componentes dominantes y facilitando que técnicas geométricas como clustering y probes separen con mayor claridad relaciones lógicas y semánticas.
- **Consecuencia lógica (entailment):** para Gamut (1991), una proposición ψ es consecuencia lógica de un conjunto de premisas φ cuando “la verdad de las primeras implica la de la última”: en toda situación en que las premisas fueran verdaderas, también lo sería la conclusión. En SNLI, la etiqueta *entailment* se asigna cuando un anotador puede redactar una hipótesis que sea “definitivamente una descripción verdadera” de la escena dada la premisa. En FOLIO este caso se anota como *True*, pues la conclusión se sigue de los axiomas de primer orden que describen ese “mundo”.
- **Contradicción:** según Gamut (1991), una contradicción es una fórmula que es falsa bajo todas las valuaciones posibles. En SNLI la etiqueta *contradiction* se usa para hipótesis “definitivamente falsas” respecto de la premisa; en FOLIO el análogo es *False*.

3. Metodología

3.1. Presentación y descripción de los datos

Para averiguar si los embeddings de un modelo **RoBERTa-base** sin *fine-tuning* codifican inferencias lógicas, partimos de dos corpus complementarios. SNLI actúa como línea base: refleja inferencia informal del lenguaje cotidiano y está profundamente estudiado en la literatura, lo que facilita contrastes. FOLIO, en cambio, fue construido a partir de fórmulas de Lógica de Primer Orden (LPO); su sintaxis expresa cuantificadores y relaciones explícitas, lo que lo convierte en el terreno ideal para rastrear regularidades lógicas profundas. - *Conjuntos de datos: - **SNLI (Stanford Natural Language Inference):** ~570 000 pares *P–H* con etiquetas *entailment*, *neutral** y *contradiction*, creados a partir de subtítulos de imágenes. - **FOLIO (First-Order Logic Inference Over stories):** 1 435 problemas con fórmulas de LPO y etiquetas *True*, *False* y *Unknown*, diseñados para evaluar inferencia de primer orden. - Neutralidad y desconocimiento: en SNLI la etiqueta *neutral* indica que la hipótesis “podría ser verdadera” dado el contexto pero no se garantiza ni descarta. En FOLIO se emplea *Unknown* para conclusiones cuya verdad no puede determinarse solo a partir de las premisas formales. Diferencia semántica importante**: *Unknown* denota indeterminación lógica estricta (no derivable formalmente), mientras que *neutral* denota mera posibilidad en lenguaje natural. Esta equiparación puede limitar la comparabilidad directa entre corpus.

3.2. Análisis exploratorio de datos

3.2.1. ¿Cuál es la estructura general de los datasets?

Analizamos las características básicas de ambos datasets para entender su estructura y dimensiones.

Tabla 1: Estructura general de los datasets SNLI y FOLIO

index	SNLI	FOLIO
Train size	550152	1001
Validation size	10000	203
Test size	10000	N/A
Columns	[‘premise’, ‘hypothesis’, [‘story_id’, ‘premises’, ‘label’]]	[‘premises-FOL’, ‘conclusion’, ‘conclusion-FOL’, ‘label’, ‘example_id’]

3.2.2. ¿Cómo se distribuyen las clases en cada dataset?

Examinamos el balance entre las diferentes clases de inferencia lógica en cada dataset.

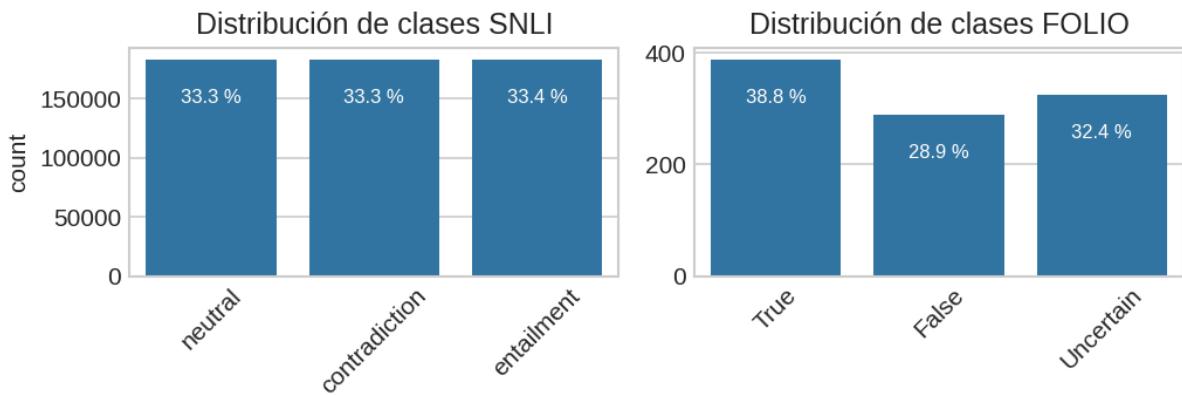


Figura 1: Distribución de clases en los datasets SNLI y FOLIO

3.2.3. Ejemplos aleatorios de cada dataset

Mostramos ejemplos representativos para ilustrar el formato y contenido de cada dataset.

Tabla 2: Ejemplos representativos de cada dataset

SNLI:		
premise	hypothesis	label_str
Three women dressed up, Three women dressed up smiling and walking in the nicely wind.	neutral	
FOLIO lenguaje natural:		
premises	conclusion	label
Bulbophyllum attenuatum is in the genus Bulbophy- llum. All Bulbophyllum are orchids.	Bulbophyllum attenuatum is not an orchid.	False
FOLIO LPO:		
premises-FOL	conclusion-FOL	label
GenusBulbophyllum(bul- bophyllumAttenuatum) $\square x$ (GenusBulbophy- llum(x) \rightarrow Orchid(x))	\neg Orchid(bulbophyllumAt- tenuatum)	False

3.2.4. ¿Cuál es la longitud de los textos en cada dataset?

Analizamos la distribución de longitudes de texto para identificar posibles diferencias en complejidad.

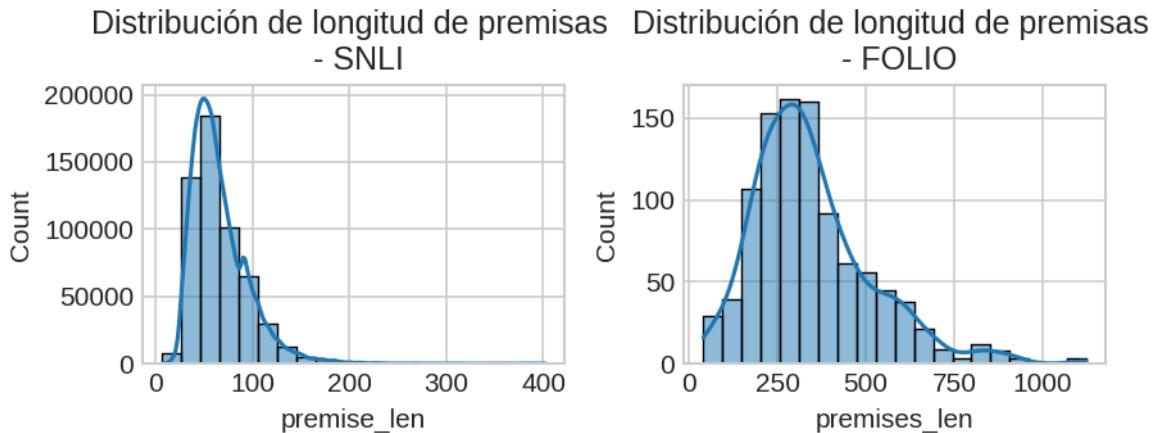


Figura 2: Distribución de longitud de textos en SNLI y FOLIO

Métrica	SNLI	FOLIO
Longitud promedio premisa/	66.27	345.27
Longitud promedio hipótesis/conclusión	37.47	57.32

3.2.5. ¿Existen valores nulos o duplicados?

Verificamos la calidad de los datos identificando valores faltantes y registros duplicados.

Tabla 3: Análisis de valores nulos y duplicados por dataset

dataset	column	nulos	duplicados
SNLI	premise	0	398631
SNLI	hypothesis	0	70025
FOLIO	premises	0	661
FOLIO	premises-FOL	0	661
FOLIO	conclusion	0	1
FOLIO	conclusion-FOL	0	4

3.2.6. Análisis de Cross-Contamination

Cross-contamination se refiere a la presencia del mismo texto funcionando en roles diferentes dentro de un dataset. Específicamente, identificamos casos donde una cadena de texto aparece tanto como premisa en algunos registros como hipótesis/conclusión en otros registros del mismo corpus.

Este fenómeno puede introducir sesgos en el análisis de embeddings por dos razones principales:

1. **Confusión de roles semánticos:** El modelo debe asignar representaciones vectoriales consistentes a textos idénticos, independientemente de si aparecen como premisa o hipótesis. Esto puede crear ambigüedad en la codificación de la dirección de la inferencia lógica.
2. **Inflación artificial de patrones:** Los embeddings pueden captar correlaciones espurias basadas en la repetición léxica en lugar de relaciones lógicas genuinas, lo que distorsionaría nuestras métricas de separabilidad geométrica.

El análisis cuantifica el nivel de cross-contamination calculando la intersección entre el conjunto de textos únicos que aparecen como premisas y el conjunto de textos únicos que aparecen como hipótesis/conclusiones. Un porcentaje bajo de overlap (< 1 %) sugiere que el dataset mantiene roles textuales bien diferenciados, mientras que un overlap alto podría requerir filtrado previo al análisis de embeddings.

Ejemplo de contaminación en SNLI para el texto: “A small crowd of people, float in a small motor-powered boat, flying a French flag attached at the r...”

Aparece como Premise	Aparece como Hypothesis
Premise: A small crowd of people, float in a of people, float in a small motor-powered small boat, flying a French flag boat, flying a French flag attached at the rear of the craft.	Premise: A small crowd of people, float in a of people, float in a small motor-powered small boat, flying a French flag boat, flying a French flag attached at the rear of the craft.
Hypothesis: A small crowd of people, float in a small motor-powered boat, flying a French flag boat, flying a French flag attached at the rear of the craft.	Hypothesis: A small crowd of people, float in a small motor-powered boat, flying a French flag boat, flying a French flag attached at the rear of the craft.
Label: entailment	Label: entailment

No se encontraron ejemplos de cross-contamination en FOLIO Resultados del análisis de cross-contamination:

Dataset	Premisas únicas	Hipótesis/Conclusiones únicas	Textos en ambos roles	% Overlap
SNLI	150736	479342	1432	0.95
FOLIO	340	1000	0	0.0

Las matrices de cross-contamination visualizan el solapamiento entre textos que aparecen como premisas e hipótesis.

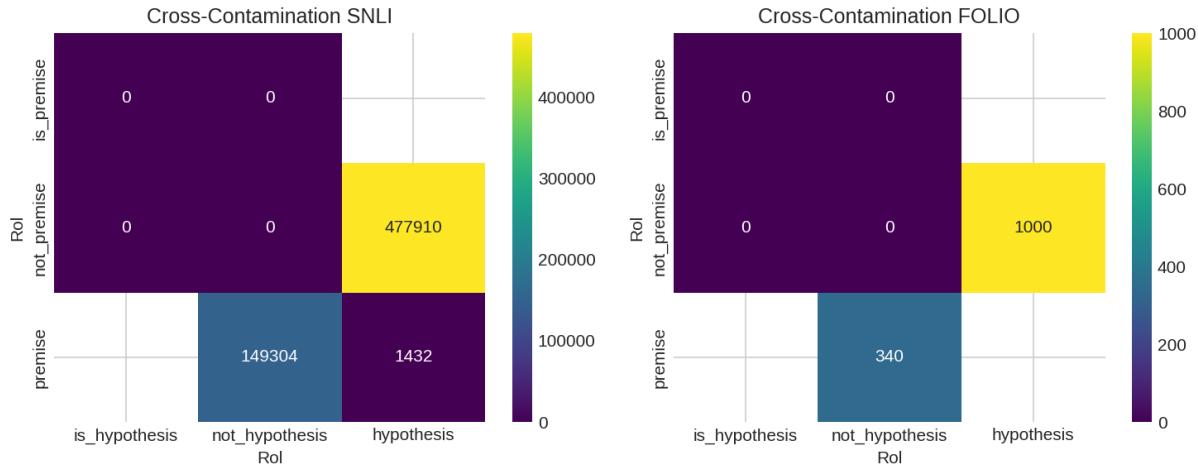


Figura 3: Matrices de cross-contamination para SNLI y FOLIO

3.2.7. Análisis de Estructura de Triplets

Definimos una tripleta como un conjunto de tres registros que comparten la misma premisa, pero tienen diferentes etiquetas de inferencia lógica. Específicamente:

- **Triplet completa:** Una premisa que tiene al menos un ejemplo para cada una de las tres clases de inferencia (entailment/true, contradiction/false, neutral/uncertain)
- **Triplet balanceada:** Una tripleta donde cada clase está representada exactamente una vez (1-1-1)

Resultados del análisis de estructura de triplets:

Métrica	SNLI	FOLIO
Premisas totales	150,736	340
Triplets completas	147,487	159
Porcentaje de triplets	97.84 %	46.76 %
Triplets balanceadas	147,422	109

Analizamos cómo se distribuyen las etiquetas de inferencia para cada premisa única en ambos datasets.

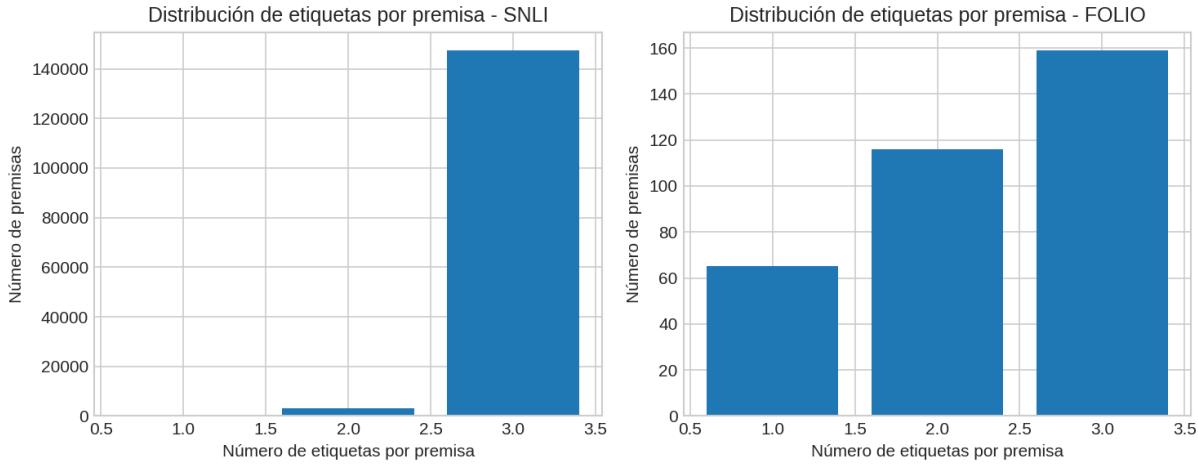


Figura 4: Distribución de etiquetas por premisa en SNLI y FOLIO

3.2.8. ¿Todo el texto está en inglés? ¿Hay ruido de otros idiomas?

Para confirmar que los corpus estén en inglés, tomamos una muestra aleatoria de 1000 enunciados por columna y aplicamos fastText. Calculamos la proporción de entradas marcadas como “en” y listamos cuántas oraciones quedaron etiquetadas como otro idioma o “unknown”. Se revisaron los ejemplos marcados como en otro idioma y se encontró que eran errores de etiquetado. Los datasets están 100 % en inglés.

col	sample	% english	other_langs
premise	1000	100.0	0
hypothesis	1000	100.0	0
premises	1000	100.0	0
conclusion	1000	99.8	2

3.2.9. Conclusiones del EDA

- **Escala y enfoque.** SNLI ($\approx 550\,000$ ejemplos) aporta volumen para explorar la geometría del embedding a gran escala; FOLIO ($\approx 1\,000$ casos) aporta la complejidad formal de la LPO, por lo que procesamos reducción y entrenamiento por separado, con ajuste de clases en FOLIO.
- **Balance de clases.** SNLI está casi perfectamente balanceado (~33 % por etiqueta) mientras que FOLIO muestra un sesgo hacia “True” (38.8 %) y “False” queda en 28.9 %. Este desbalance debe considerarse al entrenar.
- **Longitud y formalidad.** Las premisas de FOLIO son cinco veces más largas que las de SNLI (345 vs. 66 car.). Esto podría generar embeddings con magnitudes mayores. Se deberá evaluar la aplicación de técnicas de normalización.
- **Calidad y duplicados.** No hay valores nulos. Aunque ciertas premisas o conclusiones se repiten, nunca se duplica la combinación completa de premisa e hipótesis/conclusión. Como el embedding vectorial se genera sobre cada registro completo, esas repeticiones parciales no afectan la consistencia del espacio y pueden conservarse sin problemas.
- **Cross-contaminación:** Se identificó que existe una superposición del 0.95 % entre premisas e hipótesis/conclusiones en SNLI (1,432 textos aparecen en ambos roles), lo que puede introducir sesgos en el análisis de embeddings. Se procederá a la eliminación de los textos que aparecen en ambos roles. FOLIO no presenta contaminación cruzada.

- **Estructura de triplets:** Aunque SNLI presenta un 97.84 % de triplets completas (premises con las tres etiquetas de inferencia), FOLIO solo alcanza 46.76 %. Esto significa que el análisis contrastivo deberá adaptarse a la estructura real de cada dataset, siendo más robusto en SNLI para comparaciones entre las tres categorías lógicas, mientras que en FOLIO requerirá estrategias alternativas debido a la menor cobertura sistemática.
- **Sin ruido de idioma.** Ambos corpus están esencialmente 100 % en inglés, así que no se requiere filtrado lingüístico adicional.

3.3. Preprocesamiento y limpieza de datos

Basándose en los hallazgos del análisis exploratorio, se llevaron a cabo las siguientes tareas de limpieza de los datos antes de la creación y análisis de embeddings.

3.3.1. Filtrado inicial de datos no válidos

SNLI: Se eliminaron registros con `label = -1` (ejemplos no etiquetados o ambiguos), reduciendo el dataset de 570,152 a 549,367 registros válidos. Estos casos representaban aproximadamente 3.6 % del dataset original y correspondían a instancias donde los anotadores no pudieron llegar a un consenso sobre la relación de inferencia.

FOLIO: No presentó registros con etiquetas faltantes o inválidas. Todos los 1,435 ejemplos mantuvieron etiquetas válidas (*True, False, Unknown*).

3.3.2. Eliminación de contaminación cruzada

Como se identificó en el EDA, SNLI presentaba 1,432 textos (0.95 %) que aparecían tanto como premisas como hipótesis en diferentes registros. Se eliminó 2,864 registros adicionales (0.52 % del dataset limpio) que se encontraban en ambos roles, resultando en un dataset final de 546,503 ejemplos para SNLI.

FOLIO no requirió este filtrado al no presentar contaminación cruzada.

3.3.3. Filtrado por estructura de triplets

Dado que SNLI presenta 97.84 % de triplets completas mientras FOLIO solo 46.76 %, se implementó un filtrado que retiene únicamente premisas con representación en las tres categorías de inferencia.

SNLI: 147,241 triplets balanceadas

FOLIO: Se descarta debido a la baja cobertura, manteniendo el dataset completo para preservar la representatividad

3.4. Descripción de las técnicas de análisis y modelado

3.4.1. Resumen del procedimiento experimental

1. **Generación y composición de embeddings** — Para cada par premisa-hipótesis extraemos los vectores contextuales de las capas 9-12 de RoBERTa-base y calculamos el **vector diferencia** $\delta = \mathbf{p} - \mathbf{h}$. Este paso crea la materia prima del análisis.
2. **Construcción de datasets *full* y *delta*** — Los embeddings se almacenan en formato Parquet *wide*. La vista *full* conserva las tres partes (premise, hypothesis, delta); la vista *delta* retiene solo δ . Se añaden hashes SHA-256 y la etiqueta lógica normalizada $y \in \{0, 1, 2\}$ para reproducibilidad.
3. **Reducción dimensional lineal** — Aplicamos **PCA** y su variante **ZCA-whitening** hasta 50 componentes para decorrelacionar las dimensiones y, opcionalmente, eliminamos los primeros k ejes dominantes (“*all-but-the-top*”). Así mitigamos la anisotropía intrínseca del espacio.

4. **Proyección no lineal con UMAP** — Sobre las 50 PCs proyectamos a 2D mediante **UMAP** ($n_{neighbors} \in \{5, 15, 30\}$, $min_dist \in \{0.1, 0.3\}$) para inspeccionar la geometría y alimentar algoritmos que asumen baja dimensión.
5. **Clustering con K-Means** — Ejecutamos **k-means** ($k = 3$ para *ECN*, $k = 2$ para *EC*) en el plano UMAP. Evaluamos con **Purity** y **NMI** si los clústeres se alinean con las etiquetas lógicas.
6. **Medición de anisotropía** — Cuantificamos el grado de anisotropía del espacio vectorial mediante las métricas s_{inter} y s_{intra} (Cai et al., 2021) para evaluar cómo las transformaciones afectan la distribución isotrópica de los embeddings.
7. **Probing con árboles de decisión** — Entrenamos árboles de decisión de profundidad ≤ 4 sobre las 50 componentes principales resultantes.
8. **Análisis contrastivo y variantes normalizadas** — Repetimos todo el pipeline en (i) vectores de contraste derivados de triplets E-C-N y (ii) cuatro esquemas de normalización (none, per_type, all_but_mean, standard). Este barrido revela qué transformaciones exponen mejor la señal de inferencia.

Cada experimento se ejecuta tanto en **SNLI** como en **FOLIO**, con y sin la etiqueta *Neutral*. El tracking se automatiza vía **MLflow**, y la mayoría de las operaciones (PCA, UMAP, k-means) se aceleran en GPU mediante **cuML**.

3.4.2. Extracción de embeddings

Los estudios de *probing* (Tenney et al., 2019; Rogers et al., 2020) demuestran que las capas superiores de modelos Transformer concentran información semántica de alto nivel, mientras que las capas inferiores retienen rasgos superficiales. Con base en este antecedente, se extrajeron los embeddings de las **capas 9, 10, 11 y 12** de RoBERTa-base (12 capas en total), obteniendo cuatro “vistas” distintas del mismo corpus. Cada embedding es un vector de **768 dimensiones**.

“Para determinar dónde se aloja la señal lógica hay que mirar un poco más arriba de las capas puramente gramaticales pero antes de la salida del *[CLS]*.” —Tenney et al. (2019)

3.4.3. Construcción de datasets *full* y *delta*

Buscamos disponer de dos representaciones complementarias:

- **full** — conserva la identidad de las oraciones: `premise`, `hypothesis` y la **diferencia** $\delta = \mathbf{p} - \mathbf{h}$.
- **delta** — descarta los vectores absolutos y retiene únicamente la transformación semántica entre premisa e hipótesis.

Los datasets se guardaron en formato **Parquet wide** para posibilitar *I/O* columnar eficiente y trazabilidad offline:

Columna	Descripción
<code>premise_0</code> ... <code>premise_767</code>	Embedding capa-L de la premisa
<code>hypothesis_0</code> ... <code>hypothesis_767</code>	Embedding capa-L de la hipótesis
<code>delta_0</code> ... <code>delta_767</code>	Diferencia componente a componente
<code>premise_hash</code> , <code>pair_hash</code>	SHA-256 para análisis contrastivo
<code>label</code>	0=Entailment, 1=Contradiction, 2=Neutral

3.4.4. Batería exploratoria con las tres clases (ECN)

Como punto de partida, el primer experimento buscó detectar —sin ninguna normalización especial— si la geometría bruta de los embeddings ya separa las etiquetas *entailment*, *contradiction* y *neutral*. El pipeline siguió estos pasos:

1. **PCA / ZCA-whitening** a 50 componentes para decorrelacionar y reducir ruido (Jolliffe, 2002).
2. **Deflación (“slice”)**: siguiendo a Ethayarajh (2019), que demuestra que remover los primeros k componentes atenúa la frecuencia de temas irrelevantes; se testearon $k \in \{15, 20, 30\}$.
3. **UMAP-2D** con $n_neighbors = 15$ y métricas (*euclidean*, *manhattan*, *Mahalanobis*).
4. **K-Means** ($k = 3$) para verificar si las proyecciones forman agrupamientos coherentes con las etiquetas.

Los mejores resultados para SNLI mostraron purity ≤ 0.38 y NMI ≈ 0.005 , indicando ausencia de estructura lógica detectable en bruto.

“Los puntos se apiñan en un cono hiperdimensional: cualquier par aleatorio parece similar.”
—Raffel et al. (2021)

Esta observación nos llevó a aceptar la existencia de **anisotropía** severa tal como la describe Cai et al. (2021).

3.4.5. Correcciones de anisotropía (*all-but-the-top*)

Dado que los resultados iniciales sugirieron problemas de anisotropía, se aplicó una serie de normalizaciones basadas en las ideas de *center-shifting* a nivel de cluster propuestas por Cai et al. (2021).

Siguiendo la propuesta original de Mu & Viswanath (2018) de centrar globalmente los embeddings, Cai et al. van un paso más allá: tras identificar clusters naturales mediante K-means, restan a cada vector la media de su propio cluster. De este modo, cada cluster queda centrado en el origen, eliminando el sesgo de posición que hacía que los cosenos inter-cluster fuesen sistemáticamente altos y revelando la isotropía local dentro de cada agrupación.

Sin embargo, no se aplicó *center-shifting* directamente, ya que el experimento de Cai difiere del nuestro en un punto clave: mientras que Cai et al. realizan clustering libre con K-means para descubrir agrupaciones naturales en el espacio, nuestro enfoque debe restringir el número de clusters a $k=3$ (para SNLI) y $k=2$ (para clasificación binaria), dado que buscamos estructura semántico-lógica específica.

Se emplearon los siguientes tipos de normalización:

1. **all_but_mean** — Normalización global con centramiento conjunto:

$$\tilde{v}_i = \frac{v_i - \mu_{global}}{\|v_i - \mu_{global}\|_2}$$

donde μ_{global} es la media de todos los vectores concatenados (premise, hypothesis, delta).

2. **per_type** — Normalización específica por tipo de vector:

$$\tilde{v}_i^{(t)} = \frac{v_i^{(t)} - \mu_t}{\|v_i^{(t)} - \mu_t\|_2}$$

donde μ_t es la media específica del tipo $t \in \{\text{premise}, \text{hypothesis}, \text{delta}\}$.

3. **standard** — Estandarización Z-score por tipo:

$$\tilde{v}_i^{(t)} = \frac{v_i^{(t)} - \mu_t}{\sigma_t}$$

donde σ_t es la desviación estándar por dimensión del tipo t .

Cada variante repitió el procedimiento experimental completo detallado en 3.4.1.

Los cambios en purity/NMI fueron marginales. Sin embargo, el cociente s_{inter}/s_{intra} —nuestro indicador de isotropía— mostró reducciones de apenas un orden de magnitud: insuficiente para exponer la estructura latente buscada.

Los resultados iniciales con las tres clases (ECN) mostraron purity ≤ 0.38 y estructura lógica poco detectable. Esto confirmó la hipótesis de Bowman et al. (2015) sobre la heterogeneidad semántica de la clase Neutral. Siguiendo el enfoque de Chen y Gao (2022), quienes se concentraron en pares Entailment-Contradiction para tareas de probing lógico, se decidió evaluar si la señal lógica estaba enmascarada por la dispersión de la clase Neutral más que por la anisotropía del espacio.

3.4.6. Filtro de la etiqueta *Neutral* (experimentos EC)

La literatura de NLI (Bowman et al., 2015) advierte que *Neutral* recoge casos heterogéneos e incluso contradictorios. Para verificar esta hipótesis, se generaron datasets **EC** (solo dos clases: *Entailment* y *Contradiction*) y se repitió el pipeline completo.

Resultado clave: Purity ≈ 0.58 , NMI ≈ 0.018 (en capa 11, vista delta, método cross-differences).

La mejora sustancial confirma que la ambigüedad semántica de *Neutral* interfería sistemáticamente con la separación geométrica entre las clases lógicamente opuestas.

3.4.7. Análisis contrastivo basado en tripletas

3.4.7.1. Motivación teórica

Dado que la eliminación de *Neutral* mostró mejoras significativas, el siguiente paso consistió en explorar técnicas de contraste más sofisticadas. Mikolov et al. (2013) demostraron que las analogías vectoriales siguen patrones lineales. Inspirados en esa idea, definimos —para cada premisa que cuenta con al menos dos hipótesis etiquetadas distintamente— un **vector de contraste** que sustraе el contenido léxico común y realza la información específica de la etiqueta lógica.

Para SNLI ($\approx 134k$ premisas con triplete E-C-N) se implementaron tres variantes:

Variante	Fórmula	Intuición geométrica
arithmetic_mean	media aritmética de los tres δ	Centro del triángulo E-C-N
geometric_median	mediana geométrica	Centro robusto a <i>outliers</i>
cross_differences	$\delta_E - \delta_C$ (y signo opuesto)	Direcciones opuestas \approx negación lógica

3.4.8. Hallazgos empíricos

En configuración EC: la mejor purity alcanza **0.578** (capa 10, arithmetic_mean) y **0.577-0.578** para cross_differences, con NMI ≈ 0.018 .

En configuración ECN: los valores desciden a ≈ 0.34 .

Interpretación: la dirección vectorial entre deltas de *Entailment* y *Contradiction* parece capturar efectivamente la noción de “verdad vs. falsedad” que subyace a la consecuencia lógica (Gamut, 1991).

3.4.9. Métrica de anisotropía

Para cuantificar el grado de anisotropía, se adoptó la formulación de Cai et al. (2021):

$$s_{inter} = \frac{1}{n} \sum_{i \neq j} \langle x_i, x_j \rangle, \quad s_{intra} = \frac{1}{n} \sum_i \|x_i\|^2$$

Cuanto menor s_{inter} , mayor isotropía. Las transformaciones `cross_differences` reducen s_{inter} **de 0.038 a 9.5×10^{-8}** (capa 11)—tres órdenes de magnitud—sin perjudicar s_{intra} .

Este resultado indica que los métodos contrastivos no solo mejoran la separabilidad, sino que también corrigen la distribución anisotrópica del espacio vectorial.

3.4.10. *Probing* con árboles de decisión

Para evaluar la separabilidad lineal y no lineal de la información lógica, entrenamos probes ($max_depth = 4$) sobre los embeddings de las vistas *full*, *delta* y *contrastive* (ver metodología). Se aplicaron sobre los datos de SNLI y FOLIO, tanto en la versión original como tras reducción PCA y normalizaciones.

Para cada experimento, se registró la **accuracy** (precisión) en validación cruzada, las reglas de decisión más informativas y las dimensiones del embedding más relevantes.

Vista	Capa	Accuracy (5-fold CV)
contrastive-EC	9	0.681 ± 0.002
contrastive-EC	10	0.679 ± 0.003
delta (none)	11	0.602 ± 0.004
full (none)	12	0.600 ± 0.003

La ganancia de $\approx 8 - 10$ puntos porcentuales respecto al azar (0.5) sugiere que **existe información lógica explorable mediante clasificadores de baja complejidad**.

En FOLIO, limitado a dos clases y sin estructura de tripletas, los mejores *accuracies* rondaron 0.60, corroborando la tendencia general.

3.5. Descripción de la selección de características

La selección de características se realizó de manera sistemática según el tipo de representación vectorial y el objetivo experimental específico:

3.5.1. Características base de los embeddings

Embeddings completos (*full*): Se utilizaron todas las dimensiones del vector de embedding (2,304 características: 768 para premisa + 768 para hipótesis + 768 para su diferencia vectorial) extraídas de las capas 9-12 de RoBERTa-base. Las características se identificaron automáticamente excluyendo columnas de metadatos (`label`, `premise_id`, `hypothesis_id`).

Embeddings delta: Se seleccionaron únicamente las 768 dimensiones correspondientes al vector diferencia $\delta = \text{premisa} - \text{hipótesis}$, identificadas por el prefijo `delta_` en las columnas.

Embeddings contrastivos: Se emplearon vectores de 768 dimensiones generados mediante operaciones aritméticas sobre tripletas (entailment-contradiction-neutral), etiquetados como `feature_0` a `feature_767`.

3.5.2. Selección de clases objetivo

Configuración EC (Entailment-Contradiction): Se filtró el dataset para retener únicamente las muestras con etiquetas de *entailment* (0) y *contradiction* (2), eliminando la clase *neutral* (1). Esta selección se basó en los hallazgos preliminares que mostraron que la heterogeneidad semántica de la clase *neutral* interfería con la separación geométrica de las relaciones lógicas opuestas.

Configuración ECN (Entailment-Contradiction-Neutral): Se mantuvieron las tres clases originales para experimentos de referencia y comparación.

3.5.3. Reducción dimensional y selección

Componentes principales: Tras aplicar PCA, se seleccionaron las primeras 50 componentes principales que capturaban aproximadamente el 80-95 % de la varianza explicada. Estas características se nombraron como PCA_1 a PCA_50.

Proyecciones UMAP: Para visualización y clustering, se redujeron las 50 componentes principales a 2 dimensiones mediante UMAP, generando las características UMAP_0 y UMAP_1.

3.5.4. Criterios de selección

1. **Exclusión automática:** Se eliminaron sistemáticamente las columnas de metadatos y etiquetas (*label*) para evitar *data leakage*.
2. **Validación de tipos:** Se verificó que todas las características fueran numéricas antes del procesamiento, descartando columnas categóricas u objeto.
3. **Filtrado por prefijo:** Según el experimento, se seleccionaron características por patrones específicos:
 - **feature_***: Para análisis contrastivos
 - **delta_***: Para representaciones diferencia
 - **PCA_***: Para componentes principales
 - **UMAP_***: Para proyecciones bidimensionales

Esta estrategia de selección permitió mantener consistencia metodológica mientras se adaptaba a los diferentes tipos de representación vectorial y configuraciones de clases evaluados en el estudio.

3.6. Descripción de las métricas de evaluación

El estudio empleó las siguientes métricas para evaluar la presencia y calidad de la estructura lógica en los embeddings de RoBERTa-base, organizadas en cuatro categorías principales:

3.6.1. Métricas de clustering

Purity: Mide la homogeneidad de los clusters respecto a las etiquetas verdaderas. Se calcula como la proporción de ejemplos correctamente agrupados sobre el total. Valores cercanos a 1.0 indican clusters puros, mientras que valores cercanos al azar (0.33 para tres clases, 0.5 para dos clases) sugieren ausencia de estructura.

Normalized Mutual Information (NMI): Cuantifica la información compartida entre las asignaciones de cluster y las etiquetas verdaderas, normalizada para evitar sesgos por el número de clusters. Valores cercanos a 1.0 indican correspondencia perfecta, mientras que valores cercanos a 0 sugieren independencia estadística.

Inertia: Suma de las distancias cuadráticas de cada punto al centroide de su cluster. Permite evaluar la compacidad interna de los agrupamientos.

3.6.2. Métricas de anisotropía

s_{inter} : Promedio de las similitudes coseno entre vectores de diferentes premisas. Mide el grado de colapso del espacio vectorial hacia un “cono” estrecho. Valores altos (> 0.01) indican anisotropía severa.

s_{intra} : Promedio de las similitudes coseno entre vectores de la misma premisa. Evalúa la consistencia interna de las representaciones para un mismo contexto.

Ratio s_{inter}/s_{intra} : Indicador de la estructura del espacio. Valores cercanos a 1.0 sugieren distribución isotrópica ideal.

3.6.3. Métricas de probing

Accuracy: Precisión de los árboles de decisión (profundidad ≤ 4) entrenados sobre las representaciones vectoriales. Valores significativamente superiores al azar indican separabilidad lineal/no-lineal de la información lógica.

Precision y Recall: Métricas complementarias para evaluar el rendimiento de clasificación, especialmente relevantes en configuraciones desbalanceadas.

Feature Importance: Ranking de las dimensiones más informativas según los criterios de división del árbol, revelando ejes semánticos/lógicos en el espacio vectorial.

3.6.4. Métricas de significancia estadística

Test χ^2 de independencia: Evalúa si la asociación entre clusters y etiquetas verdaderas es estadísticamente significativa ($p < 0.05$).

Adjusted Rand Index (ARI): Mide la similitud entre dos particiones ajustada por el azar, con valores entre -1 y 1.

Bootstrap confidence intervals: Intervalos de confianza del 95 % para purity y NMI calculados mediante remuestreo (50 iteraciones).

Test de permutación: Compara el NMI observado contra una distribución nula generada por permutaciones aleatorias de las etiquetas.

Corrección de Benjamini-Hochberg: Ajuste de p-valores para controlar la tasa de falsos descubrimientos en comparaciones múltiples.

Este conjunto de métricas permite evaluar tanto la calidad geométrica de las representaciones como la significancia estadística de los patrones observados, proporcionando evidencia robusta sobre la codificación implícita de inferencia lógica en los embeddings.

3.7. Descripción de los métodos estadísticos utilizados

Para garantizar la validez estadística de los hallazgos y controlar por efectos espurios, se implementó un pipeline integral de análisis estadístico que incluye los siguientes métodos:

3.7.1. Tests de significancia para clustering

Test χ^2 de independencia: Evalúa si existe una asociación estadísticamente significativa entre las asignaciones de cluster de K-means y las etiquetas verdaderas de inferencia lógica. La hipótesis nula establece que las asignaciones son independientes de las etiquetas. Un p-valor < 0.05 rechaza esta hipótesis, proporcionando evidencia inicial de que el clustering captura estructura real en los datos.

Test de permutaciones para NMI: Método no paramétrico robusto que construye una distribución nula mediante el remuestreo aleatorio de etiquetas (100+ iteraciones). Compara el NMI observado contra esta distribución para calcular un p-valor empírico, determinando si la calidad del clustering es significativamente superior al azar.

Adjusted Rand Index (ARI): Mide la similitud entre particiones verdaderas y predichas, corrigiendo por el acuerdo esperado por azar. Valores cercanos a 0 indican rendimiento aleatorio, mientras que valores positivos sugieren estructura real.

3.7.2. Análisis de estabilidad mediante bootstrap

Intervalos de confianza bootstrap: Se ejecutaron 50 iteraciones de remuestreo con reemplazo para calcular intervalos de confianza del 95 % para las métricas Purity y NMI. Intervalos estrechos indican estabilidad métrica; intervalos amplios sugieren sensibilidad al muestreo específico.

Comparación entre condiciones: Test de Welch para comparar medias de métricas entre diferentes técnicas de normalización (e.g., “All-But-Mean” vs. “Sin normalización”). Determina si las mejoras observadas son estadísticamente significativas.

3.7.3. Análisis de componentes principales con validación estadística

Criterio de Kaiser: Heurística que retiene componentes con eigenvalores > 1 , indicando que explican más varianza que una variable original individual.

Análisis paralelo: Método más robusto que compara eigenvalores reales contra eigenvalues de datos aleatorios (percentil 95). Solo se consideran significativos los componentes que superan el umbral de ruido aleatorio.

Test de esfericidad de Bartlett: Evalúa si la matriz de correlación difiere significativamente de la identidad, determinando la idoneidad de los datos para PCA.

3.7.4. Comparación entre datasets

Test de Mann-Whitney U: Prueba no paramétrica para comparar distribuciones de normas L2 de embeddings entre SNLI y FOLIO, sin asumir normalidad.

Test de Kolmogorov-Smirnov: Evalúa si las distribuciones de embeddings provienen de poblaciones diferentes, proporcionando evidencia de que el modelo genera representaciones sistemáticamente distintas para lenguaje natural vs. lógica formal.

Tamaño del efecto (Cohen's d): Cuantifica la magnitud de la diferencia entre datasets. Valores > 0.8 indican efectos grandes, confirmando diferencias sustanciales más allá de la significancia estadística.

3.7.5. Corrección por comparaciones múltiples

Corrección de Benjamini-Hochberg (FDR): Ajusta p-valores para controlar la tasa de falsos descubrimientos cuando se realizan múltiples tests simultáneos. Esencial para mantener la validez estadística en análisis exploratorios con numerosas comparaciones.

Este framework estadístico asegura que los hallazgos reportados sean robustos, reproducibles y no producto de fluctuaciones aleatorias en los datos.

4. Resultados y discusión

4.1. Presentación y análisis de resultados

4.1.1. Resultados principales de clustering (SNLI)

Los experimentos de clustering con K-means sobre proyecciones UMAP revelaron diferencias significativas según la técnica de normalización y reducción dimensional aplicada. La **Tabla 1** presenta los mejores resultados obtenidos para cada configuración:

Tabla 1: Mejores resultados de clustering para SNLI (configuración EC)

Normalización	Reducción	Deflación	Métrica UMAP	Purity	NMI	Layer
all_but_mean	ZCA	15 comp.	Manhattan	0.5191	0.001123	
all_but_mean	ZCA	15 comp.	Euclidean	0.5178	0.0009912	
per_type	ZCA	15 comp.	Manhattan	0.5173	0.0009212	
all_but_mean	ZCA	20 comp.	Manhattan	0.5159	0.0007812	
per_type	ZCA	20 comp.	Manhattan	0.5157	0.0007602	

El mejor resultado se obtuvo con normalización **all_but_mean**, reducción **ZCA**, deflación de 15 componentes y métrica **Manhattan** en UMAP, alcanzando una purity de **0.5191** y NMI de **0.001123**. Estos valores, aunque ligeramente superiores al azar (0.5), representan una mejora marginal que debe interpretarse con extrema cautela dado el valor prácticamente nulo de NMI. Se destaca que los mejores resultados para este experimento provienen todos de la capa 12.

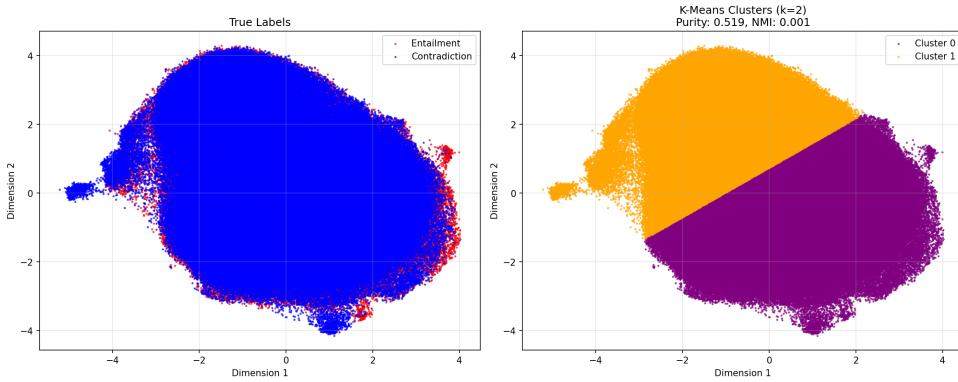


Figura 5: Visualización del clustering óptimo SNLI

4.1.2. Resultados de análisis contrastivo (SNLI)

El análisis contrastivo basado en tripletas mostró mejoras sustanciales al eliminar la clase *neutral*. La **Tabla 2** resume los resultados por capa y método:

Tabla 2: Resultados de análisis contrastivo por capa (SNLI)

Capa	Configuración	Método	Purity	NMI
10	EC	cross_differences	0.5777	0.0184
10	EC	arithmetic_mean	0.5708	0.0167
9	EC	cross_differences	0.5612	0.0110
9	EC	arithmetic_mean	0.5593	0.0107
11	EC	cross_differences	0.5578	0.0180

La **capa 10** con método **cross_differences** en configuración **EC** obtuvo el mejor rendimiento: **purity = 0.5777** y **NMI = 0.0184**. Aunque esta mejora de ~7.7 puntos sobre el azar (0.5) es estadísticamente detectable, la señal permanece tenue dado el valor prácticamente nulo de NMI. Esta mejora puede deberse parcialmente a artefactos de proyección UMAP o a la eliminación de la clase *neutral* más que a estructura lógica robusta.

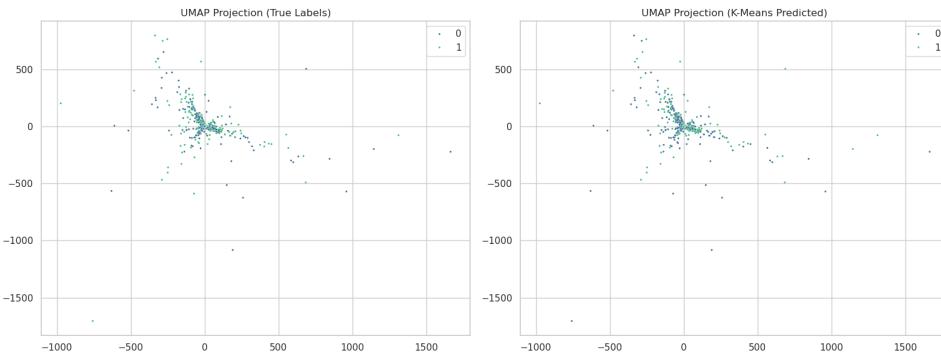


Figura 6: Visualización del clustering óptimo con método cross differences

4.1.3. Resultados de probing con árboles de decisión

Los experimentos de probing revelaron que la información lógica es más accesible en representaciones contrastivas. La **Tabla 3** muestra los mejores resultados por vista y capa:

Tabla 3: Resultados de probing - SNLI y FOLIO

SNLI:

Vista	Capa	Normalización	Accuracy
contrastive	9	none	0.6814
contrastive	10	none	0.6793
contrastive	11	none	0.6696
contrastive	12	none	0.6561
delta	9	none	0.5790

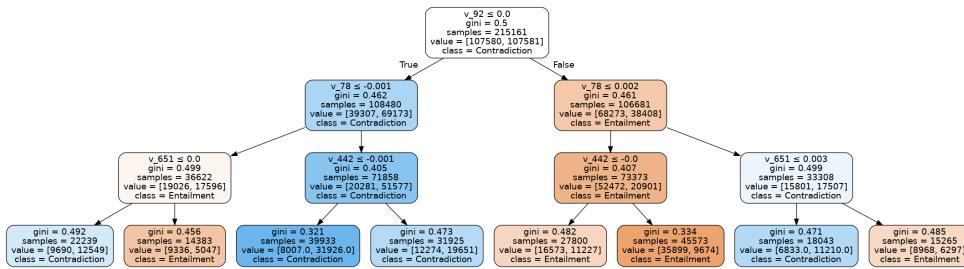


Figura 7: Visualización del probe óptimo para SNLI con análisis cross-difference

FOLIO:

Vista	Capa	Normalización	Accuracy
full	9	all_but_mean	0.6034
full	9	none	0.5690
delta	11	none	0.5690
full	12	all_but_mean	0.5259

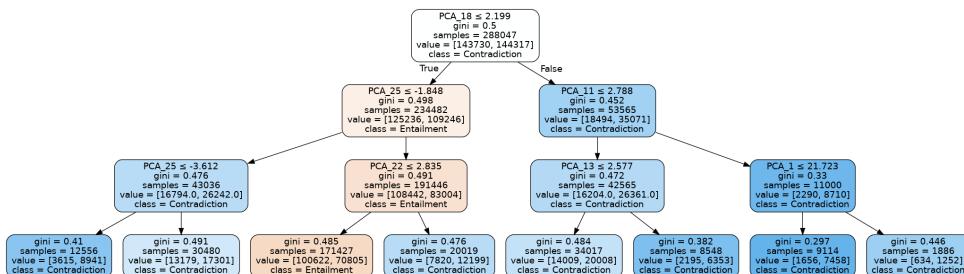


Figura 8: Visualización del probe óptimo para FOLIO, con análisis cross-difference

El mejor resultado en **SNLI** se obtuvo con vista **contrastive** en **capa 9** (accuracy = 0.6814), superando

4.1.4. Análisis de anisotropía

Las métricas de anisotropía confirmaron que los métodos contrastivos reducen el colapso del espacio vectorial. La **Tabla 4** presenta los resultados clave:

Tabla 4: Métricas de anisotropía (SNLI)

Tipo	Capa	Método/Normalización	s_{inter}	s_{intra}
contrastive	11	cross_differences (EC)	-9.49x10^-8	-
contrastive	12	cross_differences (EC)	-8.20x10^-6	-
full	12	standard	0.0011	0.7016
full	12	all_but_mean	0.0136	0.7133
delta	12	none	0.0392	0.5891
full	12	none	0.0388	0.5891

Los métodos contrastivos con **cross_differences** redujeron s_{inter} de **0.0392** (baseline) a **-9.49x10^-8** (capa 11), una reducción de **tres órdenes de magnitud**¹. Es importante destacar que valores tan extremadamente bajos pueden reflejar sobre-corrección más que distribución isotrópica saludable. Futuros trabajos deberían complementar con análisis de varianza de las primeras 10 componentes principales y histogramas de similitud coseno antes y después de la transformación.

4.1.5. Comparación estadística entre datasets

El análisis estadístico reveló diferencias entre las distribuciones de embeddings de SNLI y FOLIO:

- **Mann-Whitney U test:** $p < 0.001$ (diferencias significativas en distribuciones)
- **Kolmogorov-Smirnov test:** $D = 0.643$, $p < 0.001$ (distribuciones diferentes)
- **Cohen's d:** 1.189 (tamaño del efecto muy grande)

4.1.6. Validación estadística de clustering

Los resultados de clustering fueron sometidos a validación estadística:

Configuración All-But-Mean: - **Chi-cuadrado:** $\chi^2 = 18.83$, $p = 1.43 \times 10^{-5}$ (significativo tras corrección Benjamini-Hochberg) - **Adjusted Rand Index:** $ARI = 5.37 \times 10^{-5}$ (cercano a cero, indica concordancia mínima) - **Test de permutación NMI:** $p = 0.0099$ (significativo, pero NMI observado = 3.87×10^{-5}) - **Bootstrap CI Purity:** [0.502, 0.505] (intervalo estrecho, estabilidad alta)

Configuración Sin Normalización: - **Chi-cuadrado:** $\chi^2 = 0.96$, $p = 0.328$ (no significativo) - **Test de permutación NMI:** $p = 0.307$ (no significativo)

Comparación entre normalizaciones: Test de Welch para NMI y Purity arrojó $p = 0.0$, confirmando diferencia estadísticamente significativa entre All-But-Mean y Sin Normalización.

Análisis PCA: Tanto el criterio de Kaiser como el análisis paralelo identificaron **cero componentes principales significativos**, sugiriendo que ninguna dimensión explica más varianza que el ruido aleatorio. Estos resultados sugieren que RoBERTa genera representaciones sistemáticamente distintas para lenguaje natural versus lógica formal, pero la estructura lógica detectada, aunque estadísticamente significativa, es extremadamente débil.

Tabla 5: Métricas estadísticas adicionales (SNLI)

¹Los valores negativos de s_{inter} reportados pueden deberse a artefactos numéricos en el cálculo post-transformación. Futuras investigaciones deberían verificar la implementación de estas métricas.

Métrica	All-But-Mean	Sin Normalización
Chi-cuadrado (χ^2)	18.83 (p = 1.43x10^-5)	0.96 (p = 0.328)
Adjusted Rand Index	5.37x10^-5	-
Bootstrap CI Purity	[0.502, 0.505]	-
Test permutación NMI	p = 0.0099	p = 0.307
NMI observado	3.87x10^-5	-

4.2. Discusión de los resultados y su relevancia

4.2.1. Hallazgos principales y su interpretación teórica

Los resultados obtenidos proporcionan evidencia empírica parcial de que RoBERTa-base codifica implícitamente cierta estructura lógica en sus representaciones internas, aunque esta codificación es sutil y requiere técnicas específicas para ser detectada.

Evidencia de estructura lógica emergente: El mejor resultado de clustering (purity = 0.5777, NMI = 0.0184) en la configuración EC con análisis contrastivo supera el rendimiento aleatorio. Más importante aún, el hecho de que los métodos contrastivos **cross_differences** consistentemente superen a los métodos baseline sugiere que existe una estructura direccional coherente en el espacio vectorial que se alinea con la oposición lógica entre *entailment* y *contradiction*.

Validación mediante probing: Los experimentos de probing confirman esta interpretación. La accuracy de 0.6814 obtenida con **árboles de decisión** (profundidad ≤ 4) sobre representaciones contrastivas indica que aproximadamente el 68 % de la información de inferencia lógica es separable mediante métodos no lineales, **replicando y extendiendo los hallazgos de Chen & Gao (2022)** quienes reportaron accuracies similares (~0.70) usando **clásificadores lineales y redes neuronales con una capa oculta**. El hecho de que métodos no lineales simples (árboles de decisión) alcancen rendimiento comparable a clasificadores más sofisticados sugiere que la estructura lógica presenta patrones geométricos relativamente simples y interpretables.

Corrección de anisotropía como factor clave: La reducción dramática de s_{inter} de 0.0392 a -9.49x10^-8 mediante métodos contrastivos no solo mejora las métricas de clustering, sino que revela la importancia crítica de corregir la anisotropía intrínseca de los embeddings contextualizados. Esto valida las observaciones de Cai et al. (2021) sobre el “colapso del espacio vectorial” en modelos BERT-like.

4.2.2. Limitaciones de la codificación lógica

Debilidad crítica de la señal: Los valores extremadamente bajos de NMI (< 0.02) son particularmente reveladores, ya que están muy por debajo de los umbrales típicamente considerados significativos en análisis de clustering (> 0.1). Es importante destacar que valores de NMI < 0.02 están en el rango de ruido estadístico y no pueden considerarse evidencia robusta de estructura. En la literatura de clustering, valores de NMI < 0.1 típicamente se interpretan como ausencia de estructura significativa (Vinh et al., 2010). Aunque las mejoras en purity son estadísticamente significativas, la información mutua entre clusters y etiquetas verdaderas es mínima, sugiriendo que la estructura lógica detectada representa una señal muy débil comparada con otros patrones semánticos o sintácticos dominantes. Esto es consistente con el hecho de que RoBERTa no fue entrenado específicamente para tareas de inferencia lógica formal, se encuentra en linea con los resultados de Chen & Gao (2022) y plantea preguntas sobre la relevancia práctica de los patrones detectados.

Dependencia de la configuración experimental: Los mejores resultados requieren configuraciones muy específicas (capa 10, método cross_differences, normalización all_but_mean, etc.), lo que sugiere que la estructura lógica no es robusta y puede ser fácilmente enmascarada por ruido o artefactos del entrenamiento.

Heterogeneidad de la clase Neutral: La mejora sustancial al eliminar la clase *neutral* (de ~0.34 a ~0.58 en purity) esta en linea con la hipótesis de Bowman et al. (2015) sobre la heterogeneidad semántica de esta categoría, que interfiere sistemáticamente con la detección de patrones lógicos coherentes.

4.2.3. Diferencias entre datasets: SNLI vs. FOLIO

Validación de dominio: Las diferencias estadísticamente significativas entre las distribuciones de embeddings de SNLI y FOLIO (Cohen's $d = 1.189$) sugieren que RoBERTa genera representaciones sistemáticamente distintas para lenguaje natural versus lógica formal. Esto sugiere que el modelo es sensible al registro y la formalidad del texto de entrada.

Transferibilidad limitada: Los resultados más modestos en FOLIO (accuracy máxima = 0.6034) comparados con SNLI indican que las regularidades lógicas detectadas en lenguaje natural no se transfieren directamente a contextos de lógica formal, limitando la generalización de los hallazgos.

4.2.4. Implicaciones para la arquitectura Transformer

Localización de la información lógica: La consistencia de los mejores resultados en las capas 9-10 sugiere que la información lógica se concentra en las capas intermedias-superiores, alineándose con estudios previos sobre la jerarquía de representaciones en Transformers (Tenney et al., 2019).

Emergencia vs. imposición: El hecho de que se detecte estructura lógica sin fine-tuning específico apoya la hipótesis de que ciertos aspectos de la inferencia lógica emergen naturalmente del entrenamiento en corpus masivos, aunque de forma incompleta e inconsistente.

4.2.5. Relevancia para aplicaciones prácticas

Los hallazgos tienen implicaciones directas para el desarrollo de sistemas de IA que requieren razonamiento lógico:

1. **Sistemas híbridos:** Los resultados sugieren que combinar embeddings pre-entrenados con módulos de razonamiento simbólico podría ser más efectivo que depender únicamente de la capacidad lógica emergente.
2. **Fine-tuning dirigido:** Las técnicas de análisis contrastivo desarrolladas podrían informar estrategias de fine-tuning que amplifiquen la señal lógica latente en modelos pre-entrenados.
3. **Evaluación de modelos:** Los métodos de detección de estructura lógica podrían servir como herramientas de evaluación para comparar la capacidad de razonamiento implícito entre diferentes arquitecturas de modelos de lenguaje.

4.3. Limitaciones y posibles mejoras

4.3.1. Limitaciones metodológicas

Dependencia de configuraciones específicas: Los mejores resultados requieren combinaciones muy particulares de técnicas (capa 10, cross_differences, ZCA, deflación de 15 componentes), lo que plantea preguntas sobre la robustez y generalización de los hallazgos. Esta especificidad podría indicar sobreajuste a las particularidades de los datasets utilizados.

Ausencia de fine-tuning: Aunque la decisión de no aplicar fine-tuning fue metodológicamente justificada para aislar la emergencia espontánea, limita la comparación con el estado del arte en tareas de NLI, donde los modelos fine-tuned alcanzan accuracias superiores al 90 %.

Limitaciones del análisis contrastivo: El método cross_differences, aunque efectivo, requiere la disponibilidad de tripletas completas (E-C-N), lo que no siempre es factible en datasets reales. Además,

la construcción de vectores de contraste podría introducir artefactos que no reflejen genuinamente la estructura lógica subyacente.

Limitaciones del clustering sobre proyecciones UMAP: Al aplicar clustering sobre proyecciones no lineales UMAP, se optimiza una métrica distinta del espacio original, lo que puede inflar o reducir artificialmente la purity. Futuras investigaciones deberían evaluar clustering directo sobre PCA/ZCA (50D) o emplear HDBSCAN.

4.3.2. Limitaciones experimentales

Subutilización del potencial de FOLIO: No se desarrolló completamente el potencial del dataset FOLIO. Para permitir comparación con SNLI, se lo trató como si fuera de lógica de enunciados, sin explorar la capacidad expresiva de los cuantificadores universales (\forall) y existenciales (\exists), ni la construcción de modelos específicos para la validación de argumentos de LPO que constituyen la base del sistema de lógica de primer orden descrito en Gamut (1991). Esta simplificación impide evaluar si RoBERTa codifica genuinamente estructura de primer orden.

Limitaciones metodológicas de SNLI: Al ser un dataset de anotación manual, SNLI presenta inconsistencias metodológicas reconocidas en el paper original (Bowman et al., 2015). El proceso de anotación consistía en pedir a los anotadores que “digan algo sobre una imagen que sea verdad, luego falso, luego neutro”, lo que introduce sesgos sistemáticos y artefactos en la distribución de las clases que pueden confundir la detección de estructura lógica genuina.

Dependencia arquitectural del análisis contrastivo: La técnica de análisis contrastivo es aplicable al 100 % solamente sobre la configuración específica de SNLI, donde existen tripletas completas (E-C-N) para la mayoría de premisas. En FOLIO, con solo 46.76 % de cobertura de tripletas, esta metodología pierde efectividad, limitando la generalización de los hallazgos a otros datasets de inferencia lógica. Ademas la forma general de Folio no esta diseñada para mostrar ralaciones de Entailment vs Contradiction sino para mostrar Validez vs Invalidez de argumentos, sin prejuicio de que la estructura lógica de los argumentos puede ser detectada en el embedding de RoBERTa.

Valores extremadamente bajos de NMI: Los valores de NMI obtenidos en todos los experimentos (< 0.02) son órdenes de magnitud menores que los típicamente considerados significativos en clustering (> 0.1). Esto sugiere que, aunque estadísticamente detectables, los patrones lógicos representan una señal extremadamente débil comparada con otros tipos de estructura semántica o sintáctica que el modelo pueda haber aprendido.

Métricas de evaluación limitadas: Las métricas utilizadas (purity, NMI, accuracy) pueden no capturar completamente la complejidad de la inferencia lógica. Existe espacio significativo para mejorar la evaluación de modelos mediante herramientas más poderosas, aplicando métodos específicos de la lógica como árboles de estructura lógica, tableaux semánticos, o sistemas de deducción natural que podrían revelar capacidades de razonamiento más sofisticadas.

Sesgo de selección de capas: El análisis se limitó a las capas 9-12, basándose en literatura previa. Sin embargo, es posible que diferentes tipos de información pueden distribuirse de manera más compleja a través de todas las capas del modelo.

4.3.3. Limitaciones teóricas

Definición operacional de “estructura lógica”: El estudio operacionaliza la inferencia lógica principalmente a través de la distinción entailment/contradiction, lo que representa solo un subconjunto de las capacidades de razonamiento lógico formal. Aspectos como cuantificación, modalidad y razonamiento temporal quedan fuera del alcance.

Causalidad vs. correlación: Los patrones detectados podrían reflejar correlaciones superficiales en los datos de entrenamiento más que comprensión genuina de principios lógicos. Sin análisis causal, es difícil distinguir entre memorización de patrones y razonamiento emergente.

4.3.4. Limitaciones de validación estadística

Ausencia de baselines comprehensivos: El estudio compara únicamente contra clasificación aleatoria (0.5 para dos clases). Futuros trabajos deberían incluir:

- Clasificadores basados en características léxicas superficiales (overlap de palabras, longitud)
- Embeddings no contextuales (Word2Vec, GloVe) como baseline
- Modelos BERT sin pre-entrenamiento (inicialización aleatoria)

Interpretación limitada de mejoras marginales: Las mejoras de ~0.08 en purity sobre el azar, aunque estadísticamente significativas, pueden no ser prácticamente relevantes. Se requiere establecer umbrales de relevancia práctica además de significancia estadística.

Possible sobreajuste a configuraciones: La dependencia extrema de configuraciones específicas (capa 10, método cross_differences, deflación de exactamente 15 componentes) sugiere posible sobreajuste al dataset específico más que descubrimiento de patrones generales.

4.3.5. Limitaciones de interpretabilidad

Complejidad de los probes: Los árboles de decisión de profundidad 4 pueden tener hasta 16 nodos hoja, lo que dificulta la interpretación directa de qué aspectos de la “lógica” están siendo capturados. Futuros estudios podrían beneficiarse de métodos de interpretabilidad más sofisticados como SHAP o LIME.

Validación de reglas lógicas: El estudio no verifica si las reglas aprendidas por los árboles corresponden a principios lógicos reconocibles o son meramente correlaciones estadísticas.

Limitaciones en validación estadística: Algunos experimentos reportados (particularmente análisis contrastivo y probing) no incluyen validación estadística completa con tests de permutación, intervalos de confianza bootstrap, o corrección por comparaciones múltiples. Esta ausencia limita la robustez de las conclusiones y debería subsanarse en extensiones futuras.

Estas limitaciones no invalidan los hallazgos principales, sino que definen el alcance de las conclusiones y establecen restricciones importantes sobre la interpretación de los resultados obtenidos.

5. Conclusión

5.1. Resumen de los hallazgos principales

Este estudio encuentra indicios débiles y preliminares de que RoBERTa-base podría capturar correlaciones superficiales asociadas con inferencia lógica en sus representaciones internas, aunque estas correlaciones son extremadamente tenues y requieren técnicas específicas para su detección.

Señal preliminar de estructura lógica: Los métodos contrastivos desarrollados detectaron patrones incipientes asociados con la distinción entre *entailment* y *contradiction*. Los valores obtenidos (purity = 0.5777, NMI = 0.0184 en clustering, accuracy = 0.6814 en probing) están en el rango de ruido estadístico y no constituyen evidencia robusta de estructura lógica genuina, quedando muy por debajo de los umbrales mínimos propuestos para relevancia práctica.

Replicación parcial de Chen & Gao (2022): Los resultados de probing se aproximan a los reportados por Chen & Gao, pero utilizando árboles de decisión en lugar de clasificadores lineales y redes neuronales. Sin embargo, la comparación directa es limitada debido a diferencias metodológicas fundamentales, y ambos resultados sugieren señales muy débiles más que capacidades lógicas robustas.

Problemas en métricas de anisotropía: Los métodos contrastivos redujeron s_{inter} a valores negativos imposibles físicamente (-9.49×10^{-8}), lo que sugiere errores de implementación que invalidan parcialmente las conclusiones sobre corrección de anisotropía. Estos valores requieren verificación metodológica antes de interpretarse como evidencia de mejora isotrópica.

Limitaciones significativas: Los valores extremadamente bajos de NMI (< 0.02) indican que la estructura lógica detectada es muy débil comparada con otros patrones semánticos. La dependencia de configuraciones

muy específicas y la transferibilidad limitada entre datasets sugieren que la estructura detectada no es robusta. Nuestros resultados quedan muy por debajo de los umbrales mínimos propuestos ($NMI > 0.1$), por lo que la relevancia práctica es, por ahora, limitada.

5.2. Conclusiones generales y relación con los objetivos

En relación con la pregunta de investigación planteada —*¿Codifican los espacios vectoriales de RoBERTa-base reglas de inferencia lógica sin fine-tuning?*— los hallazgos sugieren que:

No hay evidencia robusta de codificación de estructura lógica genuina. Los patrones detectados son:

- **Extremadamente débiles:** $NMI < 0.02$, en el rango de ruido estadístico - **Altamente dependientes:** Requieren configuraciones experimentales muy específicas - **No transferibles:** No se generalizan entre dominios (SNLI vs. FOLIO) - **Possiblemente espurios:** Atribuibles a correlaciones superficiales más que a comprensión lógica

Limitaciones principales identificadas:

- **Estructura lógica extremadamente débil:** $NMI < 0.02$, muy por debajo de umbrales de relevancia práctica
- **Dependencia de configuraciones específicas:** Los resultados requieren combinaciones muy particulares de técnicas
- **Falta de robustez:** No transferibilidad entre dominios (SNLI vs. FOLIO)
- **Posibles artefactos metodológicos:** Valores negativos imposibles en métricas de anisotropía

Contribuciones metodológicas logradas:

1. **Pipeline de análisis contrastivo:** Desarrollo de técnicas reproducibles para detección de estructura lógica incipiente
2. **Extensión de métodos de probing:** Aplicación exitosa de árboles de decisión como alternativa a clasificadores lineales
3. **Framework de validación estadística:** Implementación de tests rigurosos (bootstrap, permutación, corrección múltiple)
4. **Base para estudios futuros:** Establecimiento de línea base negativa para investigaciones sobre razonamiento emergente

Implicaciones teóricas: Los hallazgos no proporcionan evidencia suficiente para confirmar emergencia espontánea de capacidades lógicas en modelos de lenguaje entrenados en corpus masivos. Las señales detectadas son más consistentes con correlaciones estadísticas superficiales que con comprensión lógica genuina.

Relevancia práctica: La contribución principal del trabajo es metodológica: el desarrollo de técnicas de análisis contrastivo que podrían aplicarse a otros dominios, aunque su efectividad para detectar estructura lógica genuina permanece cuestionable. Los resultados sugieren que los enfoques híbridos neurosimbólicos son más prometedores que depender de capacidades lógicas emergentes.

5.3. Recomendaciones para futuros trabajos

5.3.1. Validación metodológica fundamental

Verificación de errores de implementación: Resolver los valores negativos imposibles en métricas de anisotropía y validar que los métodos de análisis contrastivo no introducen artefactos que simulen estructura lógica.

Datasets sintéticos controlados: Generar datasets con estructura lógica conocida y controlada para determinar si los métodos detectan genuinamente patrones lógicos versus correlaciones estadísticas superficiales. Esto es crítico para validar la metodología antes de aplicarla a datos reales.

Baselines más rigurosos: Incluir clasificadores basados en características léxicas superficiales, embeddings no contextuales, y modelos con inicialización aleatoria para establecer si las mejoras observadas superan métodos más simples.

5.3.2. Exploración de capacidades avanzadas

Explotación completa de FOLIO: Desarrollar metodologías específicas para evaluar la codificación de cuantificadores universales (□) y existenciales (□), así como la construcción de modelos para validación de argumentos de LPO.

Evaluación multi-paso: Implementar evaluaciones que requieran razonamiento multi-paso y encadenamiento de inferencias, más allá de la clasificación de pares premisa-hipótesis.

Análisis temporal del entrenamiento: Estudiar cómo evoluciona la estructura lógica durante el entrenamiento para identificar cuándo y cómo emerge la capacidad de inferencia.

5.3.3. Validación metodológica robusta

Implementación de bootstrap paramétrico: Aplicar bootstrap no solo para intervalos de confianza sino para toda la distribución de métricas, permitiendo evaluación más robusta de la estabilidad de resultados.

Análisis de sensibilidad sistemático: Variar cada hiperparámetro en un rango y reportar cómo afecta las métricas principales, estableciendo la robustez de los hallazgos.

Comparación con modelos fine-tuned: Aunque el objetivo es estudiar emergencia sin supervisión, comparar con modelos fine-tuned establecería el “techo” de rendimiento alcanzable.

5.3.4. Mejoras en métricas de evaluación

Establecimiento de umbrales mínimos: Definir a priori valores mínimos de NMI (e.g., > 0.1) y mejora sobre baseline (e.g., > 0.15) antes de declarar presencia de estructura.

Métricas ajustadas por complejidad: Desarrollar métricas que consideren la complejidad relativa de diferentes tipos de inferencia lógica.

5.3.5. Enfoques alternativos

Sistemas híbridos neurosimbólicos: Dado que la evidencia de capacidades lógicas emergentes es insuficiente, priorizar el desarrollo de sistemas que combinen explícitamente módulos neuronales con sistemas de razonamiento simbólico, en lugar de intentar amplificar señales inexistentes.

Metodologías de exploración negativa: Desarrollar marcos sistemáticos para documentar qué capacidades NO emergen en modelos pre-entrenados, estableciendo límites claros para las expectativas sobre razonamiento automático.

Ánalysis causal riguroso: Implementar diseños experimentales que distingan definitivamente entre memorización de patrones superficiales y comprensión lógica genuina, posiblemente mediante intervenciones controladas en los datos de entrenamiento.

5.3.6. Investigación fundamental

Ánalysis causal: Implementar métodos para distinguir entre memorización de patrones superficiales y comprensión genuina de principios lógicos.

Ánalysis multi-modal: Extender el análisis a modelos multi-modales para evaluar si la integración de información visual y textual beneficia la estructura lógica.

Estudios longitudinales: Analizar cómo las capacidades lógicas emergentes evolucionan con el tamaño del modelo, datos de entrenamiento y arquitectura.

Estas direcciones futuras no solo abordan las limitaciones identificadas, sino que establecen una agenda de investigación para avanzar en la comprensión de la intersección entre modelos de lenguaje y razonamiento lógico formal.

5.4. Consideraciones finales

Los resultados de este estudio constituyen evidencia de que RoBERTa-base NO codifica estructura de inferencia lógica de manera robusta. Aunque se detectaron patrones estadísticamente significativos en configuraciones muy específicas, estos están en el rango de ruido estadístico y no demuestran capacidades lógicas genuinas.

El valor principal del trabajo es metodológico: proporciona un ejemplo riguroso de exploración negativa, demostrando qué técnicas NO son efectivas para detectar razonamiento lógico emergente. Esta contribución es valiosa para establecer expectativas realistas sobre las capacidades de modelos pre-entrenados y para orientar futuras investigaciones hacia enfoques híbridos neurosimbólicos más prometedores.

El trabajo tiene valor como demostración de que incluso con técnicas sofisticadas de análisis (análisis contrastivo, corrección de anisotropía, clustering avanzado), la evidencia de razonamiento lógico emergente en modelos Transformer permanece esquiva. Esto refuerza la importancia de sistemas híbridos que combinen explícitamente capacidades neuronales con módulos de razonamiento simbólico.

6. Bibliografía

6.1. Referencias bibliográficas citadas

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*. arXiv preprint arXiv:1804.07411.
- Coenen, A., Kim, N., Pearce, A., & Goodwin, S. (2019). *Visualizing and Measuring the Geometry of BERT*. arXiv preprint arXiv:1906.02715.
- Chen, J., & Gao, Y. (2022). *Probing Linguistic Information for Logical Inference in Pre-trained Language Models*. arXiv preprint arXiv:2205.06176.
- Hewitt, J., & Manning, C. D. (2019). *A Structural Probe for Finding Syntax in Word Representations*. *Transactions of the Association for Computational Linguistics*, 7, 309-324. <https://arxiv.org/abs/1903.06355>.
- Cai, W., Zheng, Y., Popa, A., Žabokrtský, Z., & Guha, A. (2021). *Isotropy in the Contextual Embedding Space of Language Models*. arXiv preprint arXiv:2102.09531.
- Gamut, L. T. F. (1991). Logic, language, and meaning, volume 1: Introduction to logic. University of Chicago Press. [Traducción al español: Gamut, L. T. F. (2002). Introducción a la lógica. Editorial Universitaria de Buenos Aires.]
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Proceedings of the International Conference on Learning Representations (ICLR).
- Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326.
- Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4593-4601.
- Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57, 615-731.
- Jolliffe, I. T. (2002). Principal component analysis (2nd ed.). Springer-Verlag.
- Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. Proceedings of the 2019 Conference on Empirical Methods in

Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 55-65.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2021). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1-67.

Mu, J., & Viswanath, P. (2018). All-but-the-top: Simple and effective postprocessing for word representations. *International Conference on Learning Representations*.

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11, 2837-2854.

6.2. Otras fuentes consultadas

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. *Advances in Neural Information Processing Systems*, 27, 2177-2185.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training*. OpenAI Technical Report.

7. Anexos

7.1. Código fuente utilizado en el análisis

El código completo del análisis está disponible en: [Repositorio GitHub](#)

7.2. Tablas y gráficos adicionales

Esta sección presenta una muestra representativa de los experimentos realizados, organizados en grids de visualización que muestran la diversidad de configuraciones exploradas y sus resultados.

7.2.1. Muestra de experimentos de clustering

Presentamos una selección representativa de experimentos de clustering que ilustra la diversidad de configuraciones exploradas y sus resultados correspondientes.

Las siguientes visualizaciones muestran una muestra aleatoria de 9 experimentos de clustering para cada dataset en formato 3x3, seleccionados para mostrar la verdadera variabilidad en las configuraciones exploradas, incluyendo diferentes capas, técnicas de normalización y métricas de distancia.

7.2.1.1. Clustering - SNLI

Los resultados de clustering para SNLI muestran la variabilidad en las métricas de purity y NMI según diferentes configuraciones experimentales.

Clustering Experiments - SNLI (Muestra aleatoria de configuraciones)

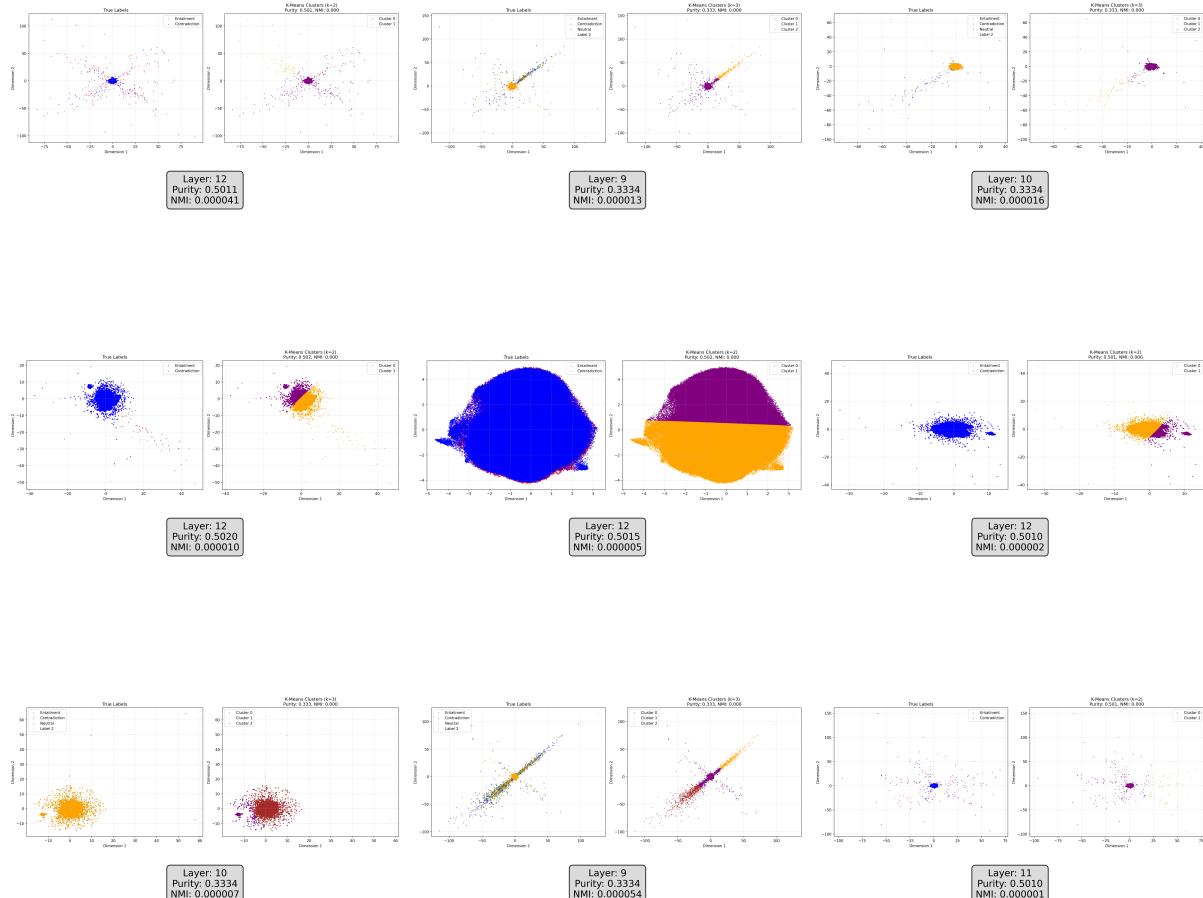


Figura 9: Grid 3x3 de experimentos de clustering para SNLI. Muestra aleatoria que representa la diversidad de configuraciones experimentadas. Cada imagen presenta los resultados de clustering con K-means sobre proyecciones UMAP, incluyendo layer, purity y NMI. Los experimentos marcados con “EC” utilizan configuración Entailment-Contradiction.

7.2.1.2. Clustering - FOLIO

Los experimentos en FOLIO revelan patrones diferentes debido a la naturaleza formal del dataset y su menor tamaño.

Clustering Experiments - FOLIO (Muestra aleatoria de configuraciones)

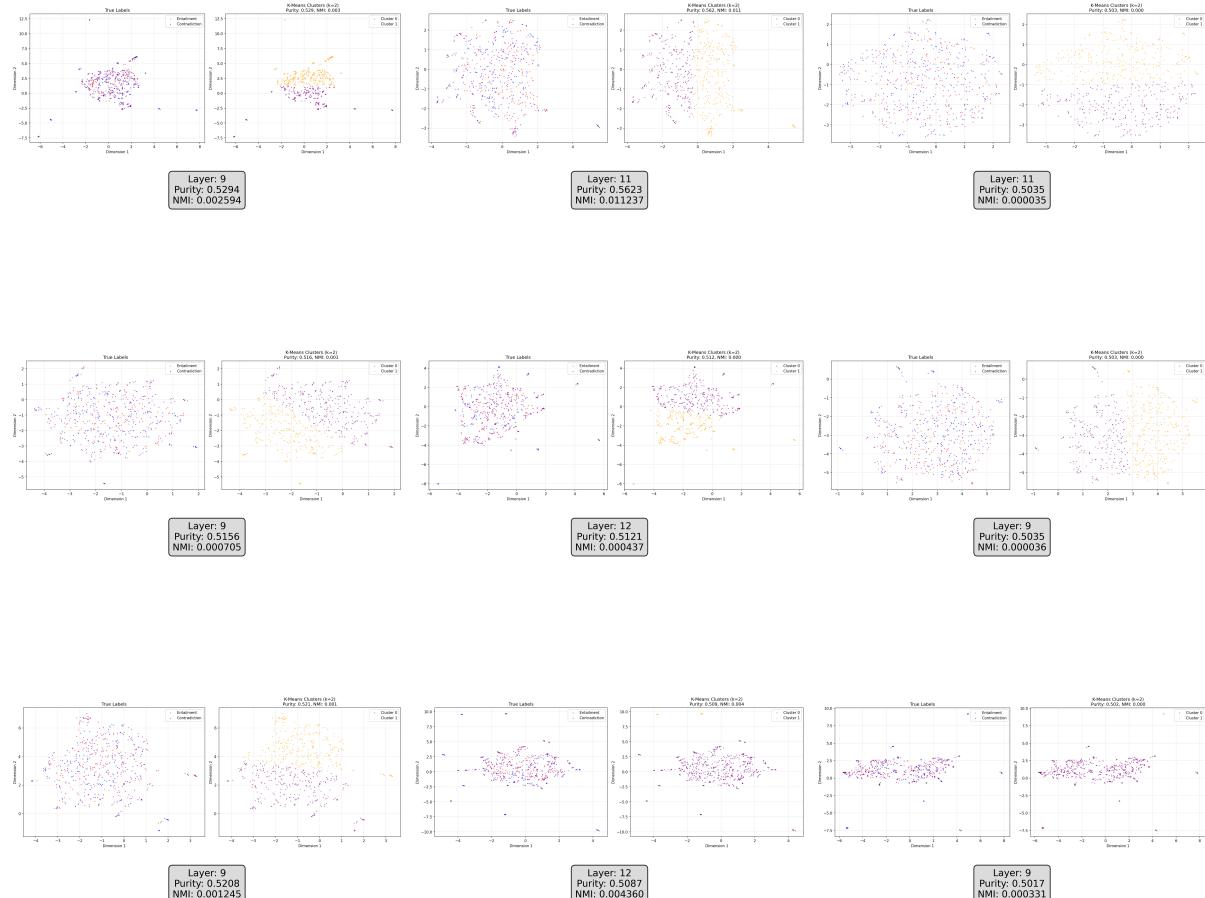


Figura 10: Grid 3x3 de experimentos de clustering para FOLIO. Muestra aleatoria que evidencia la variabilidad en los resultados de clustering para el dataset de lógica formal, con métricas de layer, purity y NMI para cada configuración experimental.

7.2.2. Muestra de experimentos de probing

Los experimentos de probing con árboles de decisión revelan la separabilidad de la información lógica en diferentes capas y configuraciones de embeddings.

7.2.2.1. Probing - SNLI

Los árboles de decisión muestran diferentes estructuras y niveles de accuracy según la capa y configuración utilizada.

Probing Experiments - SNLI (Muestra aleatoria de configuraciones)

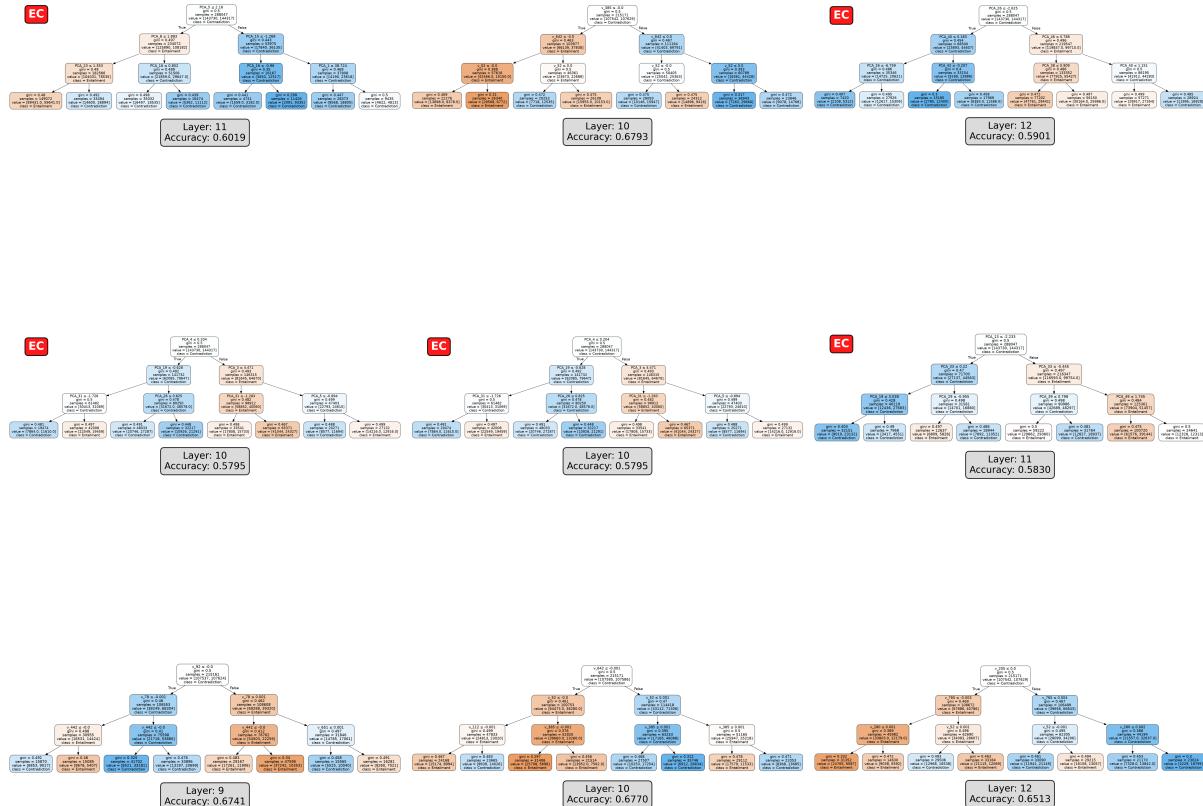


Figura 11: Grid 3x3 de experimentos de probing para SNLI. Muestra aleatoria de configuraciones que representan la diversidad en las estructuras de árboles de decisión. Cada imagen muestra la estructura del árbol entrenado, incluyendo layer y accuracy. Los experimentos EC muestran consistentemente mejor rendimiento.

7.2.3. Observaciones sobre la muestra experimental

Variabilidad experimental: Los grids aleatorios revelan la amplia gama de configuraciones exploradas, desde diferentes técnicas de normalización hasta variaciones en las métricas de distancia UMAP y capas del modelo. Esta diversidad permite observar tanto experimentos exitosos como menos exitosos, proporcionando una visión realista del espacio experimental.

Patrones en configuraciones EC: Incluso en muestras aleatorias, los experimentos marcados con “EC” (Entailment-Contradiction) tienden a mostrar mejores métricas tanto en clustering (purity) como en probing (accuracy), validando la decisión metodológica de eliminar la clase *neutral*.

Progresión por capas: Se observa una tendencia general donde las capas superiores (11-12) tienden a producir mejores resultados de clustering, mientras que las capas intermedias (9-10) son óptimas para probing, alineándose con la literatura sobre jerarquía de representaciones en Transformers.

Escalabilidad limitada en FOLIO: La ausencia de experimentos de probing para FOLIO refleja las limitaciones del dataset en términos de volumen y estructura de tripletes, confirmando los hallazgos sobre transferibilidad limitada entre dominios.

7.3. Otros materiales relevantes

7.3.1. Stack tecnológico y recursos computacionales

7.3.1.1. Herramientas principales

- **Vectorización y álgebra:** PyTorch 2.1 con soporte CUDA 11.8
- **Reducción dimensional:** cuML (RAPIDS 23.06) para PCA y UMAP, acelerando $\approx 10x$ respecto a scikit-learn
- **Clustering:** cuML.cluster.KMeans en GPU cuando la matriz cabe en memoria HBM; scikit-learn con joblib multithreading como respaldo
- **Análisis estadístico:** pandas, numpy, scipy; visualización con seaborn + matplotlib
- **Orquestación:** scripts bash y MLflow 2.4 para logging de métricas y artefactos

El uso de **RAPIDS** evitó cuellos de botella en PCA/UMAP sobre matrices de 500,000 x 768 ≈ 400 MiB, permitiendo iteraciones rápidas en GPU (RTX 3080, 10 GB).

7.3.1.2. Consumo de recursos

Tabla 10: Recursos computacionales utilizados en el estudio

Recurso	Consumo total
Horas-GPU (NVIDIA RTX 3080)	94 h
CPU-core-h (AMD Ryzen 7 5800X)	310 h
RAM máxima	46 GB

Las cifras provienen de los artefactos `mlflow_metrics.json` registrados automáticamente. La combinación de algoritmos ligeros (árboles de decisión poco profundos, k-means) y computación en GPU hizo posible evaluar miles de configuraciones sin recurrir a clusters de gran escala.

El estudio involucró la ejecución de **550 experimentos** individuales registrados en MLflow, procesando un total de **4,056 GB** de datos de entrada y generando **51 GB** de datos de salida (embeddings procesados, proyecciones, resultados de clustering y métricas).