Creation and Validation of a Monolingual Spanish NLI Dataset for Metaphor Interpretation via Model-in-the-Loop

Alec Sánchez-Montero[1], Gemma Bel-Enguix[2], Sergio-Luis Ojeda-Trueba[2]
[1]Universitat Pompeu Fabra, [2]Universidad Nacional Autónoma de México
alecmisael.sanchez@upf.edu, {gbele, sojedat}@iingen.unam.mx

Abstract

Large Language Models (LLMs) can easily generate fluent text, but assessing whether they truly understand metaphors requires moving beyond English-centric datasets and binary token classification tasks. To test if current state-of-the-art models perform genuine structural alignment and analogical reasoning rather than just echoing statistical token co-occurrence, we introduce a new monolingual Spanish Natural Language Inference (NLI) dataset specifically built for metaphor interpretation. Using a Model-in-the-Loop approach, we reconstruct the literal truth conditions of metaphors sourced from science texts. Before human experts curated the data, we performed an ablation study—evaluated via BERTScore and Cross-Entropy—to test whether explicit symbolic scaffolding improves analogical reasoning. While automated evaluations suggested that forcing models to follow explicit metaphorical rules diminished their fluency and increased text surprisal, human evaluation revealed the opposite: this explicit guidance produced far more accurate and strictly literal outputs. This reveals a limitation in how we evaluate NLU: automated metrics consistently penalize the cognitive 'heavy lifting' required to resolve a metaphor, simply because they are built to reward surface-level statistical fluency. By releasing this resource, we aim to shift the focus from surface-level generation to real cognitive alignment and metaphorical understanding in Spanish NLU.

Keywords: metaphor, natural language inference, automatic metahor interpretation, language resources

1. Introduction

Figurative language expressions, and metaphors in particular, remain a persistent challenge for Natural Language Understanding (NLU). Within this subject, the focus in Natural Language Processing (NLP) has traditionally been on metaphor detection, which has been understood as a classification problem and a sequence labeling task (Rai and Chakraverty, 2021). However, in recent years, literature has addressed metaphor interpretation as a key task for measuring true 'understanding' by Large Language Models (LLMs), arguing that comprehending a metaphor requires going beyond mere lexical detection or contrasting meanings, as proposed by dominant annotation protocols such as MIP/MIPVU (Pragglejaz, 2007; Steen et al., 2010), to involve structural alignment and analogical reasoning (Bowdle and Gentner, 2005). In other words, the focus has shifted from identifying which tokens are used metaphorically to ensuring that models can grasp the underlying semantic conditions that validate a figurative expression (i.e., as opposed to mimicking statistical patterns based on token co-occurrence).

In this context, the Natural Language Inference (NLI) framework has emerged as the benchmark method for assessing metaphorical competence in LLMs. Following established protocols such as FLUTE (Chakrabarty et al., 2022), this task is typically framed as determining whether a literal context (Premise) entails or contradicts a metaphorical expression (Hypothesis) (Stowe et al., 2022; Tong et al., 2024; Sengupta et al., 2025). This approach is a step up from simple statistics or lexical bias; rather than simply paraphrasing a metaphor, the model must demonstrate that it can discern the literal state of affairs that makes a metaphor semantically valid. Reliable metaphor understanding by LLMs is vital for downstream tasks such as machine translation and summarization (Rana et al., 2025).

Nevertheless, advances in this field are constrained by two major challenges. First, most high-quality NLI tasks and datasets (such as FLUTE or MUNCH) focus on the English language. In contrast, the scarcity of comparable resources in languages such as Spanish makes it difficult to evaluate LLMs in specific linguistic and cultural contexts, thereby limiting analysis to imperfect transfer phenomena or reliance on multilingual models (Sanchez-Bayona and Agerri, 2025). Secondly, creating NLI datasets for metaphor understanding involves a complex and time-consuming annotation task.

This paper addresses the scarcity of resources by introducing a monolingual Spanish NLI dataset specifically designed for metaphor interpretation. We adopt a Model-in-the-Loop (MITL) methodology for efficient data annotation: LLMs generate candidate premise-hypothesis pairs, which are then curated by human experts. Furthermore, we conduct an ablation study to probe whether explicit metaphorical guidance triggers better reasoning capabilities in LLMs compared to pure statistical

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 77–87
11 May 2026.

inference, contributing to the discussion on emergent analogical reasoning in small-data scenarios. This work represents the foundational phase of a broader research agenda.

The remainder of this paper is organized as follows: Section 2 reviews the shift from metaphor identification to interpretation within the NLI framework. Section 3 details the MITL methodology we used to annotate our dataset, along with the experimental design and the ablation study regarding explicit metaphorical guidance. Section 4 presents the results and analysis derived from human evaluation, followed by a discussion on the implications for analogical reasoning. Finally, Section 5 summarizes our contributions and future directions.

2. Related Work

2.1. The Concept of Metaphor

Traditionally, linguistic research in NLP has approached metaphor primarily through either Conceptual Metaphor Theory (CMT) (Lakoff and Johnson, 1980) or MIP/MIPVU methodologies (Pragglejaz, 2007; Steen et al., 2010). CMT posits metaphors as fixed, static mappings between a source and a target domain (e.g., ARGUMENT IS WAR), while MIP/MIPVU focuses on a lexical procedure to identify these mappings by contrasting "basic" dictionary meanings with contextual usage. While these frameworks have been instrumental for metaphor detection and corpora annotation, they treat metaphoricity as a binary property of text, largely ignoring the real-time cognitive and communicative mechanisms required for metaphoric competence.

However, when evaluating LLMs, the question is not merely whether a metaphor exists, but *how* it is processed and represented by the system. Unlike the static view of CMT or the binary token-based framework of MIP/MIPVU, the Career of Metaphor (CoM) theory suggests that metaphor is a dynamic process depending on how expressions become conventionalized (Bowdle and Gentner, 2005). According to this framework, metaphor processing shifts from structural alignment (an active, computationally expensive comparison or analogy) in novel metaphors to categorization (retrieval of stored associations of lexical items) as they become conventional.

This distinction is particularly relevant for the Reasoning vs. Statistics debate in NLP (Webb et al., 2023). While LLMs demonstrate prowess in completing figurative patterns, it remains unclear whether this capability stems from genuine structural generalization—akin to the alignment process—or merely from exploiting higher-order statistical correlations found in pre-training data, effectively treating all metaphors as conventionalized categories. If models rely exclusively on distributional priors, they may fail to interpret metaphors that require active mapping between distant domains. Therefore, adopting the CoM framework enables us to probe whether LLMs are simulating reasoning or merely retrieving frequent patterns.

2.2. Literature Review

Recent work has highlighted significant gaps in how NLP systems model figurative language. A fundamental issue lies in the definition of the task itself. As noted by Fuoli et al. (2025), metaphoricity is not a binary label (0/1) but a "radial category with more or less prototypical examples". Consequently, while recent studies evaluating LLMs report high F1 scores in binary identification via fine-tuning, these models operate as "black boxes" that replicate labels without modeling the underlying process. Zero-shot prompting performance remains inconsistent, suggesting that models lack a robust internal representation of metaphoricity (Fuoli et al., 2025).

While early benchmarks like GLUE (Wang et al., 2018) treated NLI as a broad, sentence-level semantic task, Chakrabarty et al. (2022) introduced FLUTE as a framework for figurative language processing, which validated that a model understands a metaphor (Hypothesis) only if it can reconstruct the literal situation (Premise) that sustains its truth conditions. However, recent findings have questioned LLMs' ability to perform this mapping genuinely. Although LLMs perform above chance in Winograd-style tasks, they rely on distributional surface cues rather than deep semantic interpretation, according to Liu et al. (2022).

On the other hand, Chen and Mao (2023) identified that even specialized architectures, such as MelBERT (Choi et al., 2021), exhibit "analogical blindness," which means they can effectively identify the Target domain (the topic) but fail to retrieve the Source domain (the image) of a metaphor, indicating that mere detection does not imply structural mapping or conceptual grounding. In the context of Spanish, Puraivan et al. (2024) observe a similar trend: while GPT-4 excels at disambiguating polysemous verbs, its "interpretations" often collapse into static dictionary definitions—a form of categorization—rather than context-aware conceptual mappings.

Therefore, the core debate is whether LLMs are capable of transcending statistical association through explicit scaffolding. Recently, Sengupta et al. (2025) analyzed how the explicit identification of source and target domains acts as a cognitive "boost" for improving NLI performance in few-shot settings. This dependence on external guidance contrasts with psycholinguistic findings in humans

by Ahrens et al. (2024), who found that explicit metaphor signaling does not necessarily aid in the comprehension of novel expressions.

This behavioral divergence forces us to confront a fundamental hypothesis about the inner workings of current language models. Unlike embodied human cognition for metaphor processing, LLMs may require explicit symbolic markers to emulate the structural alignment process and avoid the pitfalls of statistical categorization or conventionality biases, as observed in Italian LLMs by Mazzoli et al. (2025). While Webb et al. (2023) argue that analogical reasoning has emerged as a zero-shot capability in large-scale models, if LLMs truly possessed the flexible structural alignment seen in human cognition, they would not require the explicit symbolic scaffolding observed in recent NLI tasks. Therefore, the reported "analogical prowess" of these models may be less an emergent cognitive faculty and more a reflection of their ability to navigate higher-order statistical correlations within the lexicon.

3. Resource Creation

The development of this resource stems from the need to move beyond surface-level label detection toward the resolution of underlying semantics in Spanish. To this end, we adapted the NLI task structure from FLUTE (Chakrabarty et al., 2022; Sengupta et al., 2025) to a curated set of metaphorical expressions. The final annotated dataset is publicly available on a GitHub repository.

3.1. Sample Selection and Curation Process

The source data consists of Spanish tweets focused on Popular Communication of Science (PCS), originally annotated by Sánchez-Montero et al. (2025). To ensure experimental validity and control, we applied a strict filtering criterion: we selected only the 200 instances that achieved "perfect human agreement" for the overall metaphor category (i.e., a soft-label score of 1.0). We hypothesized that using these prototypical examples would isolate the model's reasoning capacity by eliminating the semantic noise inherent in borderline cases of metaphoricity. This sample includes various scientific and colloquial metaphors used in PCS, such as personifications (e.g., telescopes making new scientific discoveries), direct comparisons (e.g., presenting evolution as a tree), and lexical metaphors (e.g., black holes or digital threads).

For each selected instance, we identified the specific tokens that were previously annotated as metaphorical. We then manually authored a brief natural language explanation (NLE) for each metaphor, adhering to the classification schema used in the original dataset (e.g., Direct, Indirect, or Personification). These NLEs constitute the ground-truth metaphorical knowledge, which is a critical component of our experimental design, as it provides the explicit symbolic guidance required for the ablation study, where the presence or absence of this expert-level information serves as the primary independent variable.

3.2. NLI Annotation via Model-in-the-Loop

Following the paradigm established by Chakrabarty et al. (2022), the NLI resource we introduce is structured as a reconstruction of truth conditions. For every metaphorical Hypothesis (H), we generated two types of literal Premises (P):

- **Entailment (E):** A literal description of the real-world situation that validates the figurative expression.
- **Contradiction (C):** A literal scenario that renders the analogical mapping proposed in the hypothesis false or impossible.

We utilized a *model-in-the-loop* approach to generate these premises, employing `gpt-4.1-mini` (OpenAI et al., 2024), `Qwen2.5-7B-Instruct` (Quen: An Yang et al., 2025), `Mistral-7B-Instruct-v0.3` (Jiang et al., 2023), and `gemma-3-12b-it` (Team et al., 2025) as the generative engines. With this selection, we aimed to evaluate whether metaphorical reasoning scales across different architectures and paradigms. While GPT-4.1 serves as a highly aligned proprietary baseline, the inclusion of robust open-weight models (Qwen, Mistral, and Gemma) allows us to investigate whether the capacity for structural alignment in Spanish is a generalized emergent capability or is heavily dependent on proprietary instruction-tuning pipelines.

The premise generation followed two distinct configurations for our ablation study:

- **Baseline Configuration:** The model received only the metaphorical Hypothesis (H) and was tasked with generating E and C without any external guidance.
- **Scaffolded Configuration:** The model received H along with the NLE regarding the identified metaphors. This setup forces the model to integrate external symbolic knowledge before performing the inference task.

We implemented a few-shot prompting strategy, providing two prototypical examples of (H, E, C) triplets. We included these examples to anchor the model's output to a literal register, explicitly forbidding the use of figurative language in the premises.

Drawing on the experimental framework of Sengupta et al. (2025), who utilized a temperature of 0.8 for metaphorical NLI tasks, we opted for a more constrained temperature of 0.7. This setting facilitates sufficient linguistic variation for Spanish paraphrasing while maintaining the logical precision required for the premises. During our initial pilot tests, higher temperatures led to a degradation in literalness, causing the models to hallucinate secondary metaphorical associations rather than produce strictly factual premises. Each prompt was executed in an independent session to eliminate the risk of context-window leakage or intra-experimental bias.[1]

3.3. Automatic Quality Estimation

Given the labor-intensive nature of the human validation step in the MITL pipeline, manually reviewing the outputs from every model and configuration would be highly impractical. Therefore, to focus our annotators' efforts and identify the most reliable option for the final curation, we implemented a preliminary automatic evaluation phase. We assessed the generated entailments (E) using two complementary metrics:

- **Semantic Coherence (BERTScore):** We used BERTScore (Zhang et al., 2020) with the multilingual `xlm-roberta` checkpoint (Conneau et al., 2020) to measure how well the generated entailment (E) preserved the underlying meaning of the original metaphorical text (H). The metric ranges from 0 to 1, with higher values reflecting a stronger semantic overlap between E and H. Additionally, we implemented a strict penalty for "language drift"; any premise generated in English received a significant deduction to prevent high scores from being assigned to non-Spanish outputs.
- **Conditioned Fluency and Information Loss (Cross-Entropy):** We calculated the average Cross-Entropy (CE) loss to evaluate the predictability and linguistic quality of the literal E conditioned on the original metaphorical H. Rather than evaluating the generated text in isolation, we measured the surprisal of E given H. To avoid self-preference bias, we employed two independent causal models—`Llama-3.1-8B` (Grattafiori et al., 2024) and `Mistral-7B` (Jiang et al., 2023)—as external judges. Lower CE scores represent lower surprisal when transitioning from the metaphorical source to the literal target, indicating a more predictable and fluent reconstruction of truth conditions and a minimal loss in Spanish.

The results of this automatic evaluation across both the baseline and scaffolded configurations are detailed in Table 1.

As detailed in Table 1, performance varied depending on the models' ability to maintain the target language. Mistral-7B, for instance, exhibited language drift in over 20% of its outputs. Due to our strict penalty for code-switching, its overall BERTScore fell to 0.7589, well below the rest of the group. At the other end of the spectrum, the GPT baseline delivered the most stable results; it achieved the highest semantic overlap (BERTScore = 0.847) and produced the most predictable Spanish phrasing, reflecting the lowest CE values from both judges (1.2842 and 1.2578).

A critical takeaway from the ablation setup is the behavior of the scaffolded configuration, which constitutes an unexpected pattern regarding explicit guidance. Except for Mistral-7B, providing the explicit NLE led to higher cross-entropy and lower BERTScores across all models. This might indicate that processing explicit metaphorical constraints disrupts their fluency and leads to more literal metaphor paraphrases that could be perceived as less natural. Consequently, we discarded the 'noisier' outputs and exclusively forwarded the GPT-generated premises to the human curation phase, which meant that our expert annotators could focus entirely on evaluating complex logical and analogical errors instead of filtering out basic translation or grammatical failures.

4. Manual Validation

4.1. Overall Assessment

Although automated metrics like BERTScore and Cross-Entropy effectively filter out low-quality outputs, they ultimately act as proxies for surface-level fluency and lexical overlap. These scores cannot definitively prove whether a model has genuinely 'understood' a metaphor. Therefore, to assess the top performer's metaphorical competence beyond statistical measurements, we conducted a manual validation via expert annotation. Based on the automated quality estimations, we isolated the outputs from our top-performing model (GPT-4.1) to build our core NLI dataset.

During this phase, we annotated the generations from both the baseline and the scaffolded configurations. We anticipated that a side-by-side comparison of these setups would reveal how injecting external symbolic knowledge (via the NLEs) shifts the model's underlying interpretative process. Still, since expert manual evaluation of semantics and

[1] The complete translated text of the prompts, including the few-shot examples used for both configurations, is detailed in the Appendix.

Model / Configuration	BERTScore (↑)	CE (Llama 3.1) (↓)	CE (Mistral 7B) (↓)
GPT Baseline	**0.8470**	**1.2842**	**1.2578**
GPT Scaffolded	0.8421	1.3724	1.3178
Qwen Baseline	0.8449	1.4192	1.3745
Qwen Scaffolded	0.8400	1.5986	1.5333
Mistral Baseline	0.7589	1.4852	1.2911
Mistral Scaffolded	0.7777	1.5630	1.3684
Gemma Baseline	0.8332	1.4928	1.4488
Gemma Scaffolded	0.8305	1.6026	1.5286

Table 1: Automatic Evaluation of E premises in relation to H: BERTScore and Cross-Entropy

pragmatics is incredibly time-consuming, we narrowed the scope of this initial validation phase. We focused our qualitative analysis on the 200 items from the dataset, restricting our review entirely to the Entailment (E) premises.

We opted to leave the Contradiction (C) premises for future work due to the limitations in how current models handle metaphorical negation. In our preliminary experiments, we observed that the model couldn't reliably negate just the metaphors in H. Instead, it usually ended up bluntly negating the entire situation. Validating the C premises requires checking whether a model can actively negate a metaphorical mapping without breaking basic logic—a much more demanding task that could have derailed the manual validation, given the quality of these outputs. Therefore, isolating the entailments ensured our annotators could zero in on one core question: did the model accurately and literally reconstruct the truth conditions that validate the metaphor?

We graded the generated entailments using a 3-point scale (0, 0.5, and 1). We deliberately avoided a simple binary system because LLM outputs are highly nuanced; a model might successfully unpack one metaphor but stumble on another within the same sentence. Our annotation criteria were defined as follows:

- **Score 1 (Fully Correct):** To get a perfect score, the generated text had to completely strip away all figurative language (i.e., it had to be 100% literal) while accurately and comprehensively explaining the underlying meaning of the metaphor (or all the metaphors, if multiple were present) from the original hypothesis.

- **Score 0.5 (Partially Correct):** Assigned when the model grasped the core meaning but failed to execute the task flawlessly. This happened mostly in two scenarios: either the model explained the metaphor but accidentally snuck in a new figurative expression, or it successfully resolved one metaphor but missed others present in the same context.

- **Score 0 (Incorrect):** Reserved for clear failures in metaphorical reasoning. We assigned a zero if the model hallucinated facts, logically contradicted the source text, or just "parroted" the original figurative words instead of translating them into a literal reality.

The distribution of human-annotated scores across the two experimental configurations is detailed in Table 2.

The manual validation seems to contradict our preliminary automated metrics. On paper, the unguided baseline configuration appeared superior: it generated text with higher statistical fluency (lower cross-entropy) and better lexical overlap (higher BERTScore). Yet, when evaluated for actual semantic validity, it failed to capture the literal truth conditions in roughly 24.5% of its generations. The scaffolded configuration, on the other hand, despite scoring lower on automated metrics, proved vastly superior in human evaluations: fully correct entailments climbed from 67.0% to 88.5%, and complete conceptual errors were reduced by nearly two-thirds, dropping from 24.5% to 8.5%.

This clash between statistical metrics and human judgment reveals a troubling blind spot in NLU evaluation for figurative language. The explicit scaffolding appears to 'force' the model to engage in active structural mapping rather than relying on highly probable word associations. Doing this cognitive "heavy lifting" naturally increases text surprisal and lowers lexical overlap metrics. Nonetheless, as the human scores demonstrate, accepting this higher surprisal is precisely what is required to generate factually accurate and literal premises and, ultimately, to verify whether current models can actually understand and interpret real metaphorical language.

4.2. Qualitative Analysis

When we manually reviewed the outputs where the zero-shot baseline failed but the scaffolded model succeeded, two distinct behavioral patterns emerged. Without explicit guidance on how to handle figurative language, the baseline GPT model

Configuration	Score 1	Score 0.5	Score 0
GPT Baseline	134 (67%)	17 (8.5%)	49 (24.5%)
GPT Scaffolded	**177 (88.5%)**	6 (3%)	**17 (8.5%)**

Table 2: Human evaluation results for generated entailments ($N = 200$), comparing the zero-shot baseline against the explicitly guided scaffolded configuration.

essentially fell into two traps: it either ignored the metaphor completely (contextual distraction) or it tried to explain it but accidentally used another metaphor (analogical blindness).

1. Contextual Distraction and Omission

The first common error was simple omission. Faced with a metaphorical expression it did not quite know how to map, the unguided model often took the easy way out: it ignored the figurative phrase entirely or just summarized the literal context surrounding it. Since the actual truth conditions of the metaphor were lost, these outputs received 0.

Consider this example of personification from the dataset:

> **Original Text (ES):** *"...la hazaña fue posible luego de que el orbitador de la misión Tianwen-1 analizara el suelo marciano durante tres meses."*
> **Original Text (EN):** *"...the achievement was made possible after the Tianwen-1 mission orbiter analyzed the Martian soil for three months."*
> **Metaphor:** Personification (attributing the cognitive action of 'analyzing' to a machine).
>
> **Baseline Generation (Score 0) (ES):** *"China envió una nave espacial que aterrizó con éxito en Marte y desplegó un rover llamado Zhurong para explorar la superficie..."*
> **Baseline Generation (Score 0) (EN):** *"China sent a spacecraft that successfully landed on Mars and deployed a rover named Zhurong to explore the surface..."*
>
> **Scaffolded Generation (Score 1) (ES):** *"Un satélite en órbita alrededor de Marte recopiló y transmitió datos sobre la composición y características del suelo marciano..."*
> **Scaffolded Generation (Score 1) (EN):** *"A satellite orbiting Mars collected and transmitted data on the composition and characteristics of Martian soil..."*

In this instance, we observe a clear case of semantic displacement. Although the baseline model recognizes the general topic (China's Mars mission), it fails to resolve the specific metaphorical action of the orbiter. Instead of translating the verb *analizara* (analyzed) into its physical equivalent, the model defaults to generating a high-probability summary of the overall event. In contrast, the scaffolded configuration uses the explicit NLE to anchor its attention: by being told that "analyzing" in this context refers to a mechanical process, the model is forced to identify that an orbiter "analyzes" soil by *collecting and transmitting data*.

This omission behavior was not limited to scientific concepts; it also affected digital metaphors. In another instance involving a Twitter thread (*"Te comparto un hilo con información..."* = Here's a thread with some information...), the baseline completely omitted the concept of the thread, generating a premise about showing an image instead. As one of our evaluators explicitly flagged in the annotation logs: *"[The model] completely ignores 'thread'."* The scaffolded version, conversely, successfully mapped the metaphor to its literal reality: *"Se publicaron varios mensajes relacionados entre sí..."* (Several interconnected messages were published).

When faced with sentences containing multiple layered metaphors, the unguided baseline often suffered from "metaphorical fatigue." It successfully unpacked the first figurative expression but completely ignored the subsequent ones. Consider this example regarding the neurochemistry of love:

> **Original Text (ES):** *"...tu cerebro produce oxitocina, dopamina y serotonina, un verdadero cóctel de felicidad. Además, se activa constantemente el 'circuito de recompensa'..."*
> **Original Text (EN):** *"...your brain produces oxytocin, dopamine, and serotonin, a real cocktail of happiness. In addition, the 'reward circuit' is constantly activated..."*
> **Metaphors:** Metaphorical comparison (*cóctel*), personification (*produce*), and lexicalized metaphor (*circuito de recompensa*).
>
> **Baseline Generation (Score 0.5) (ES):** *"...el cerebro libera oxitocina... y se activa el **circuito de recompensa** asociado con sensaciones placenteras."*
> **Baseline Generation (Score 0.5) (EN):** *"...the brain releases oxytocin... and activates the **reward circuit** associated with pleasurable sensations."*
>
> **Scaffolded Generation (Score 1) (ES):** *"...el cerebro libera simultáneamente oxitocina... además, se activa una **red neuronal** asociada a la recompensa."*
> **Scaffolded Generation (Score 1) (EN):**

> *"...The brain simultaneously releases oxytocin... In addition, a **neural network** associated with reward is activated."*

As our annotator log points out: *"it completely failed to find the second metaphor."* While the baseline correctly translates the "cocktail" into a literal chemical release, it becomes distracted and leaves the lexicalized "reward circuit" intact.

2. Analogical Blindness

The second error pattern identified is a form of analogical blindness (Chen and Mao, 2023), where the model "interprets" a metaphor by replacing it with another equally figurative expression.

Consider this example of personification from the dataset:

> **Original Text (ES):** *"El telescopio espacial Spitzer ha descubierto 120 000 nuevas estrellas recién nacidas..."*
> **Original Text (EN):** *"The Spitzer Space Telescope has discovered 120,000 new newborn stars..."*
> **Metaphors:** Personification (attributing the action of 'discovering' to a machine and the property of being 'newborn' to stars).
>
> **Baseline Generation (Score 0.5) (ES):** *"El telescopio espacial Spitzer detectó radiación infrarroja proveniente de nubes de polvo en el plano galáctico y, a partir de esos datos, identificó 120 000 estrellas **jóvenes** recién formadas."*
> **Baseline Generation (Score 0.5) (EN):** *"The Spitzer Space Telescope detected infrared radiation coming from dust clouds in the galactic plane and, based on this data, identified 120,000 newly formed **young** stars."*
>
> **Scaffolded Generation (Score 1) (ES):** *"Un telescopio espacial detectó señales de radiación infrarroja provenientes del polvo en las nubes interestelares, lo que permitió a los científicos identificar 120 000 estrellas que se formaron recientemente en el plano de la galaxia."*
> **Scaffolded Generation (Score 1) (EN):** *"A space telescope detected infrared radiation signals coming from dust in interstellar clouds, allowing scientists to identify 120,000 stars that formed recently in the plane of the galaxy."*

In this paradigmatic case, the baseline (0.5) replaced the term 'newborn' with 'young' stars. Although this is a common astronomical description, from a strictly linguistic perspective, "young" is still a metaphor of personification projected onto an inanimate object. As the evaluators point out in the logs: "it replaces newborn with young, another metaphor." In contrast, the scaffolded model (1.0) managed to break the analogical cycle by describing the physical process as "stars that were recently formed."

Nonetheless, our analysis uncovered a blurred line where sometimes absolute literalness seems to border on the impossible. In cases of highly lexicalized or abstract metaphors, such as "black holes" or the very concept of technological 'tools,' even the scaffolded model struggled. A critical finding in manual validation was the persistence of personification in scientific instruments, attributing them the capabilities of "analyzing," "discovering," and "observing." Technological tools, such as telescopes, are essentially extensions of the human body; we attribute human characteristics to them because they allow us to perform human actions on a non-human scale. In these cases, the only way to achieve literalness was to shift the agency from the object to the subject, as can be seen in the following example:

> **Original Text (ES):** *"El telescopio Webb analizó la estrella Fomalhaut..."*
> **Original Text (EN):** *"The Webb telescope analyzed the star Fomalhaut..."*
> **Metaphors:** Personification (attributing the action of 'analyzing' to a machine).
>
> **Scaffolded Generation (Score 1) (ES):** *"Los científicos utilizaron el telescopio Webb para observar la estrella Fomalhaut..."*
> **Scaffolded Generation (Score 1) (EN):** *"Scientists used the Webb telescope to observe the star Fomalhaut..."*

Sometimes, as can be seen in the example, interpretive success did not consist of finding a literal synonym for a particular verb, but rather of reconstructing the entire scene. The symbolic guidance via NLE seems to be the only mechanism that forces the model to recognize this hierarchy of human agency, preventing the model from merely "parroting" the linguistic convention of personifying technology.

Finally, there were a few remaining cases where the model simply failed to decipher the text-context relationship. Without symbolic scaffolding, models tend to "hallucinate" explanations based on the statistical frequency of adjacent words, losing sight of the truth condition that underpins the analogy. Despite their fluency, LLMs often get stuck on the surface of language, needing a symbolic "push" to understand metaphorical language.

4.3. Discussion

We have observed that forcing the model to integrate explicit knowledge of NLEs increases the cross-entropy of the resulting text and reduces its lexical overlap (BERTScore). For traditional evaluation systems, this increase in statistical surprise is a

symptom of degradation; however, human evaluation shows that it is precisely the cognitive cost necessary to achieve semantic accuracy when dealing with metaphor understanding.

This empirical tension supports Liu et al. (2022), as the apparent semantic competence of LLMs is often based on distributional clues rather than deep context understanding or genuine analogical reasoning. When the model operates without scaffolding, its natural inertia is to remain within the statistical comfort zone of figurative language. Replacing the concept of "newborn stars" with "young stars" is not a step toward literalism, but rather a lateral shift within the same semantic domain, which may indicate that it did not completely understand the meaning of the metaphor. In terms of Bowdle and Gentner (2005), this analogical blindness might be explained because models treat metaphors as static categories and retrieve frequent synonyms rather than performing structural alignment between different domains. In other words, metaphorical knowledge might be coded as categorization rather than instances of analogy.

Symbolic scaffolding acts precisely as the mechanism to break this inertia. Forcing the model to transition from statistical categorization to analogical inference also reveals an epistemological boundary in the evaluation task itself. Reaching the "zero degree" of literalness often clashes with how we inherently conceptualize conceptual domains such as science and technology. Historically, we have endowed our scientific instruments with human agency because they function as physical extensions of our own bodies. Consequently, in natural language, telescopes "look," rovers "analyze," and probes "discover." This assimilation of metaphor as a simple static category not only reflects a computational limitation, but also raises the question about the importance of embodied cognition in the architecture of NLP systems and, by extension, in the pre-training data of models.

For the unguided model, the metaphoricity of some expressions remains invisible due to its overwhelming frequency in the pre-training data. The metaphor is so ubiquitous that it masquerades as literal truth. Providing explicit scaffolding, however, forces a pragmatic shift. Rather than just swapping vocabulary, in some cases, the guided model actively rewrites the text to hand agency back to the human researchers (e.g., *scientists used the telescope to observe*...). Interpreting a metaphor in an NLI context goes far beyond simple intralinguistic translation or synonym hunting, as it requires the model to step outside the text and actively reconstruct the real-world hierarchy of who acts and what is acted upon.

5. Conclusion and Future Work

In this work, we have presented the first monolingual Natural Language Inference (NLI) dataset in Spanish, dedicated specifically to metaphor interpretation. Through a Model-in-the-Loop methodology, we have demonstrated that it is possible to generate accurate reconstructions of the truth conditions of figurative expressions, provided that an explicit symbolic scaffolding is supplied.

The method shows promising results and suggests a clear direction for future work through the evaluation of LLM behavior in the detection, classification, and inference of metaphor. Several lines of future research are identified. Also, it will be necessary to investigate different prompting strategies and forms of contextual knowledge that could enhance LLMs efficiency. On the other hand, experiments must be scaled to larger and more general corpora. It will also be essential to develop hybrid human-LLM evaluation methods and metrics capable of accurately assessing model performance on this task and rewarding genuine cognitive alignment over mere statistical fluency.

Our immediate next step involves the manual validation of the Contradiction (C) premises, a logical challenge that will test whether models can actively negate an analogical mapping without violating real-world constraints. Expanding the scope of this research, we also plan to conduct an in-depth qualitative comparison across different LLMs to understand how their baseline interpretative strategies differ when handling figurative language. Furthermore, scaling these experiments beyond science communication will be essential. Future work must investigate whether model performance and "analogical blindness" shift when encountering highly conventionalized or lexical metaphors in general-domain Spanish texts.

Acknowledgements

This paper has been supported by PAPIIT Project IG400325, as well as a Postgraduate Scholarship by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) (CVU 1225477).

Ethical considerations and limitations

The primary drawback of this paper is the limited size of the dataset ($N = 200$) on which the experiment and subsequent manual evaluation were conducted. Although the results were extensively and rigorously validated by expert annotators, the limited number of instances may restrict the broader generalizability of our findings.

The principles of the Belmont Report (Belmont, 1978) were followed in the creation of the dataset

and in the data annotation process.

6. Bibliographical References

Kathleen Ahrens, Christian Burgers, and Yin Zhong. 2024. Making the unseen seen: The role of signaling and novelty in rating metaphors. *Journal of Psycholinguistic Research*, 53(3):36.

Informe Belmont. 1978. Principios éticos y directrices para la protección de sujetos humanos de investigación. *Estados Unidos de Norteamérica: Reporte de la Comisión Nacional para la Protección de Sujetos Humanos de Investigación Biomédica y de Comportamiento*.

Brian F. Bowdle and Dedre Gentner. 2005. The career of metaphor. *Psychological Review*, 112(1):193–216.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159. Association for Computational Linguistics.

Zi-Yuan Chen and Yining Mao. 2023. MetaMapper: Interpretable metaphor detection.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale.

Matteo Fuoli, Weihang Huang, Jeannette Littlemore, Sarah Turner, and Ellen Wilding. 2025. Metaphor identification using large language models: A comparison of rag, prompt engineering, and fine-tuning.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and Aiesha Letman et *al.* 2024. The llama 3 herd of models.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

George Lakoff and Mark Leonard Johnson. 1980. *Metaphors We Live By*. University of Chicago press.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4437–4452. Association for Computational Linguistics.

Simone Mazzoli, Alice Suozzi, and Gianluca Lebani. 2025. Language models and the magic of metaphor: A comparative evaluation with human judgments. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 710–721. CEUR Workshop Proceedings.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, and Janko Altenschmidt et *al.* 2024. Gpt-4 technical report.

Pragglejaz. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.

Eduardo Puraivan, Irene Renau, and Nicolás Riquelme. 2024. Metaphor identification and interpretation in corpora with ChatGPT. *SN Computer Science*, 5(8):976.

Quen: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Sunny Rai and Shampa Chakraverty. 2021. A survey on computational metaphor processing. *ACM Computing Surveys*, 53(2):1–37.

Manisha Rana, Rita Chhikara, and Srishti Sharma. 2025. A SURRVEY ON METAPHOR DETECTION AND INTERPRETATION. *International Journal For Multidisciplinary Research*, 7(4):54959.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2025. Meta4xnli: A crosslingual parallel corpus for metaphor detection and interpretation.

Meghdut Sengupta, Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, Eyke Hüllermeier, Debanjan Ghosh, and Henning Wachsmuth. 2025. Investigating the impact of conceptual metaphors on LLM-based NLI through shapley interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 17393–17403, Suzhou, China. Association for Computational Linguistics.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*, volume 14 of *Converging Evidence in Language and Communication Research*. John Benjamins Publishing Company.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Alec Sánchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Gerardo Sierra. 2025. Prompting metaphoricity: Soft labeling with large language models in popular communication of science tweets in spanish. In *Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II)*, pages 45–56. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, and Thomas Mesnard et *al.* 2025. Gemma 3 technical report.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3517–3536, Bangkok, Thailand. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Appendix

Prompt for Baseline Task (English Translation)

You are a linguist specializing in Spanish semantics and pragmatics. Your task is to generate high-quality Natural Language Inference (NLI) pairs.

Instructions:

You are provided with an input sentence containing a metaphorical expression. This sentence serves as the Hypothesis (H).

Your job is to reconstruct the literal context that would generate that hypothesis. You must generate two sentences that function as literal premises:

1. Entailment (E): A description of a 100% literal and realistic situation that, if true, would make the metaphorical Hypothesis true (i.e., what actually happened in the physical world for someone to use that metaphor).

2. Contradiction (C): A description of a 100% literal and realistic situation that, if true, would make the metaphorical Hypothesis false or impossible.

Examples:

Example 1:

Input (H): 'The street was a zoo this morning.'

Output:

Entailment: 'There was deafening noise and people were running around in a chaotic environment.'

Contradiction: 'The urban environment was completely silent and orderly, and everyone was walking along quiet.'

Example 2:

Input (H): 'The candidate tried to shield his public image.' Output:

Entailment: 'The politician took extreme measures to protect his reputation and avoid any criticism or attacks from the press.'

Contradiction: ‘The politician openly exposed himself to criticism and shared compromising information about his private life.’

Important:

- The generated premises (E and C) must not contain metaphors. They should be brief and direct explanations.
- Do not repeat the input text.
- Do not repeat or explain the reasoning.

Input:
{text}
Output format:
[ENTAILMENT]
(Write the E sentence here)
[CONTRADICTION]
(Write the C sentence here)

Prompt for Scaffolded Task (English Translation)

You are a linguist specializing in Spanish semantics and pragmatics. Your task is to generate high-quality Natural Language Inference (NLI) pairs based on the key information provided to you.

Instructions:

You are provided with an input sentence containing a figurative or metaphorical expression. This sentence serves as the Hypothesis (H). You are also provided with the metaphors identified in H as key information.

Your job is to reconstruct the literal context that would generate that hypothesis. You must generate two sentences that function as literal premises:

1. Entailment (E): A description of a 100% literal and realistic situation that, if true, would make the metaphorical Hypothesis true (i.e., what actually happened in the physical world for someone to use that metaphor).
2. Contradiction (C): A description of a 100% literal and realistic situation that, if true, would make the metaphorical Hypothesis false or impossible.

Examples:

Example 1:

Input (H): ‘The street was a zoo this morning.’

Metaphor: ‘The term “zoo” is used here figuratively.’

Output:

Entailment: ‘There was deafening noise and people were running around in a chaotic environment.’

Contradiction: ‘The urban environment was completely silent, orderly, and everyone was walking around quiet.’

Example 2:

Input (H): ‘The candidate tried to shield his public image.”

Metaphor: ‘The verb “shield” is used here in a figurative sense.’

Output:

Entailment: ‘The politician took extreme measures to protect his reputation and avoid any criticism or attacks from the press.’

Contradiction: ‘The politician openly exposed himself to criticism and shared compromising information about his private life.’

Important:

- The generated premises (E and C) must not contain metaphors. They should be brief and direct explanations.
- Do not repeat the input text.
- Do not repeat or explain the reasoning.

Input:
{text}
Key information: {metaphor}
Output format:
[ENTAILMENT]
(Write the E sentence here)
[CONTRADICTION]
(Write the C sentence here)