Prompting Metaphoricity: Soft Labeling with Large Language Models in Popular Communication of Science Tweets in Spanish

Alec Sánchez-Montero alecm@comunidad.unam.mx
Gemma Bel-Enguix gbele@iingen.unam.mx
Sergio-Luis Ojeda-Trueba sojedat@iingen.unam.mx
Gerardo Sierra gsierram@iingen.unam.mx
Universidad Nacional Autónoma de México

Abstract

In this paper, we explore how large language models (LLMs) can be used to assign soft labels for metaphoricity in Popular Communication of Science (PCS) tweets written in Spanish. Instead of treating metaphors as a binary yes/no phenomenon, we focus on their graded nature and the variability commonly found in human annotations. Through a combination of prompt design and quantitative evaluation over a stratified sample of our dataset, we show that GPT-4 can assign probabilistic scores not only for general metaphoricity but also for specific metaphor types with consistency (Direct, Indirect, and Personification). The results show that, while LLMs align reasonably well with average human judgments for some categories, capturing the subtle patterns of inter-annotator disagreement remains a challenge. We present a corpus of 3,733 tweets annotated with LLM-generated soft labels, a valuable resource for further metaphor analysis in scientific discourse and figurative language annotation with LLMs.

1 Introduction

Automatic metaphor detection has undergone a significant evolution over the last decades, transitioning from traditional rule- and knowledge-based methods to statistical and machine learning methods, including supervised, semi-supervised, and unsupervised techniques (Zayed, 2021). Most recently, due to considerable advances in deep learning, the utilization of large language models (LLMs) has shown promising results in various natural language understanding (NLU) tasks, including metaphor detection and figurative language processing (Wachowiak and Gromann, 2023; Tian et al., 2024; Jia et al., 2025; Xu et al., 2024; Lin et al., 2024). However, as we have previously noted (Sánchez-Montero et al., 2025), research specifically focused on metaphor detection in Spanish based on annotation disagreement remains significantly limited.

This study addresses the intricate nature of metaphor annotation from an exploratory LLM perspective, a task we found to be characterized by inherent subjectivity and consequent disagreements among human annotators. We assume that these disagreements not only reflect the complexity of the task, but may also be symptomatic of the gradable nature of "metaphoricity", where expressions possess different degrees of metaphorical quality (Hanks, 2006). Particularly, we believe that advanced LLMs, such as `gpt-4o` and `gpt-4.1`, could be suitable for multi-label classification of metaphors, given their ability to handle datasets exhibiting an uneven distribution of multiple categories (Cloutier and Japkowicz, 2023; Kostina et al., 2025), allowing us to distinguish between different types or degrees of metaphoricity that contribute to variability in annotation.

To explore this variability of interpretation and the ability of LLMs to reflect it, we rely on a prompted-based methodology using optimized GPT-4 models, chosen for their demonstrated capabilities in annotating textual data (Yu et al., 2023, 2024; Yu, 2025). Our main objective is to investigate how different prompting strategies may influence the LLM's ability to identify cases where metaphoricity is ambiguous or susceptible to multiple interpretations, paralleling the disagreements found in human annotation of Mexican Spanish Public Communication of Science (PCS) tweets. This approach deepens our understanding of metaphor gradability, a core concept in analogical reasoning, while also holding practical value for NLU in Spanish, where AI systems must grasp metaphor and figurative language to more accurately interpret and respond to human communication. A key contribution of our resource lies in its incorporation of soft labeling and the use of LLM-based reasoning to complement human annotation. This paper is structured as follows: Section 2 provides the necessary background on key concepts

Proceedings of the 2nd Workshop on Analogical Abstraction in Cognition, Perception, and Language (Analogy-Angle II), pages 45–56

and related work; Section 3 details the characteristics of the dataset used in this study; Section 4 outlines the methodology employed, including the prompt design, the experimental setup, and the results obtained, followed by concluding remarks in Section 5.

2 Background

2.1 Foundational Concepts

Linguistic Metaphor. According to Conceptual Metaphor Theory (CMT), linguistic metaphor is the manifestation in natural language of conceptual metaphors, where one conceptual domain (source) is used to understand another (target) through a structured mapping of entities and relationships (Lakoff and Johnson, 1980). Metaphors are not mere stylistic devices or figures of speech, but fundamental phenomena shaping human cognition and grounded in our bodily experiences.

Metaphor Annotation. Metaphor annotation presents challenges due to a lack of methodological consistency and variability in intuitions, making comprehensive corpora characterization and comparison across studies difficult (Veale et al., 2016). To address this, the Metaphor Identification Procedure Vrije Universiteit Amsterdam (MIPVU) (Steen et al., 2010) (developed initially by (Praggle-jaz, 2007) as Metaphor Identification Procedure, or MIP) offers a widely adopted systematic methodology for identifying potentially metaphorical linguistic units or metaphor-related words (MRWs), which encompass indirect, direct, and implicit metaphorical expressions, as well as explicit signals of metaphor and instances of personification. Beyond MIPVU, there are other approaches, such as the Deliberate Metaphor Identification Procedure (DMIP) (Reijnierse et al., 2017), focusing on deliberate metaphors from a semiotic and communicative perspective (Steen, 2008), and annotation schemes that extend identification to conceptual metaphors, annotating source and target domains (Shutova and Teufel, 2010).

Metaphoricity. The notion of metaphoricity refers to the gradual quality of a linguistic expression perceived as metaphorical, moving away from a strict binary categorization (Julich-Warpakowski and Jensen, 2023). This theoretical perspective recognizes the fuzzy boundaries between literal and figurative language, suggesting that some metaphors are "more metaphorical" than others (Hanks, 2006). The degree of metaphoricity can depend on factors such as conventionality, the semantic or conceptual distance between source and target domains or conceptual frames (Bierwiaczonek, 2024), situational context, and inter-speaker variation (Julich-Warpakowski and Jensen, 2023). Understanding metaphoricity as a gradable phenomenon in NLP allows for modeling the subtleties and ambiguities that manifest in human annotation disagreement, derived from different interpretations of the potential metaphorical meaning of a linguistic expression.

LLMs and Prompt Engineering for Linguistic Annotation. Large language models (LLMs) represent a significant advancement in artificial intelligence, characterized by their ability to process and generate human-like text at scale, while becoming "the de facto baseline models to be used" in most NLP tasks (Zubiaga, 2024). Prompt engineering has emerged as a crucial technique for harnessing the capabilities of these models without extensive fine-tuning, involving the strategic design of textual inputs to guide desired outputs (Sahoo et al., 2024). Common strategies include zero-shot prompting (no examples), few-shot prompting (few examples) and instruction-based prompting. LLMs are increasingly being explored for linguistic annotation, including in tasks with significant human disagreement (Brown et al., 2025), and for processing phenomena such as figurative language and metaphor (Ichien et al., 2024). By providing LLMs with clear instructions and relevant context through well-designed prompts, researchers have shown that these models can perform various annotation tasks, sometimes achieving performance comparable to human annotators or outperforming them (Gilardi et al., 2023).

Learning from Disagreement Moving away from the traditional assumption of a single gold standard with hard labels and a single objective truth, the 'learning from disagreement' approach considers annotation discrepancies as valuable information, particularly for subjective linguistic tasks, such as figurative language annotation, where multiple interpretations coexist and intrinsic subjectivity generates variability (Uma et al., 2021). Rather than simply aggregating annotation disagreements into a single label and biasing models in favor of some linguistic theory, embracing disagreements allows for a richer representation of the inherent variability and gradability of subjective linguistic phenomena (Plank et al., 2014). Capturing this variability requires going beyond traditional hard

labels, using soft labels that represent the distribution or degree of human judgment. As probabilities or degrees of belief, soft labels can capture the inherent uncertainty and gradience of human annotation. Linguistic annotation, particularly for semantic interpretation and figurative language like metaphor, is inherently subjective due to variations in annotators' backgrounds, interpretations, and biases. As observed in our previous work on annotating metaphor in Spanish PCS tweets, this subjectivity resulted in significant inter-annotator disagreement (Sánchez-Montero et al., 2024, 2025). Unlike traditional hard metrics (e.g., F1, accuracy), soft evaluation metrics (e.g., cross-entropy, Manhattan distance, Euclidean distance, Jensen-Shannon divergence) are designed to compare probability distributions (Rizzi et al., 2024). This makes them suitable for evaluating models that produce soft or probabilistic outputs, which are necessary to capture the variability and gradable nature of subjective linguistic phenomena.

3 Dataset

As discussed by (Sanchez-Mora, 2016), Public Communication of Science (PCS) is a multidisciplinary field that encompasses a range of scientific disciplines and media platforms. It prioritizes accessibility and relevance for non-specialist audiences, often relying on metaphors to communicate complex ideas (Taylor and Dewsbury, 2018; Cormick, 2019). Our focus on Mexican Spanish PCS tweets stems from the scarcity of resources in this variety and genre. There is a limited pool of active science communicators on Twitter/X in Mexico, which necessarily constrains corpus size but also defines a domain that is underexplored and culturally meaningful.

The dataset utilized in this study comprises a corpus of 3733 Mexican Spanish tweets from the domain of Public Communication of Science (PCS), specifically annotated for metaphor detection from a multi-label annotation system. We have compiled this dataset from the timelines of 19 science communicators based in Mexico (January 2020 - May 2023). The information collected from these user accounts was obtained without targeting any specific scientific domain.

To our knowledge, this corpus is the first publicly documented effort to annotate linguistic metaphors specifically in Mexican Spanish PCS tweets. Although there are limited resources for metaphor detection in Spanish, such as the CoMeta corpus (Sanchez-Bayona and Agerri, 2022), and previous work has explored other variants or domains (Martínez Santiago et al., 2014; Richi Pons-Sorolla, 2020; Alvarez Mouravskaia, 2020; Uribe and Mejía, 2024), there is a gap in publicly available linguistic metaphor corpora for Mexican Spanish that are suitable for exploring nuances and variability in metaphor annotation beyond simple binary classification of metaphorical expressions. While we acknowledge the existence of larger corpora in English, our decision to focus on Mexican Spanish is both intentional and necessary. We consider Spanish PCS tweets to offer a unique intersection of scientific and colloquial registers that pose rich and often ambiguous metaphorical constructions, ideal for examining gradable metaphoricity and inter-annotator variation in this language variant.

The dataset was annotated by six native Mexican Spanish-speaking linguistics students (2 female and 4 male), who independently labeled each tweet for the presence and type of metaphor. We divided the dataset into two halves, annotated independently by groups of three annotators each. Given the absence of a public adaptation of MIPVU for Spanish, we developed specific annotation guides for PCS tweets in Mexican Spanish, incorporating concepts from CMT (source/target domains, mapping). Annotators applied labels for the three types of metaphor that we established: Direct Metaphor (3 labels: source unit, target unit, signal), Indirect Metaphor (1 label: source unit) and Personification Metaphor (2 labels: personified object, personifier). Non-metaphorical tweets were indicated by saving them without annotations. Both our annotation guidelines and dataset are publicly available on our GitHub repository. The principles of the Belmont Report were followed in the data annotation process (Belmont, 1978).

In a previous study (Sánchez-Montero et al., 2025), we presented a binary soft-labeled dataset of PCS tweets (metaphorical vs. non-metaphorical). In this paper, we introduce a new layer of analysis through fine-grained metaphor type soft annotation (Direct, Indirect, and Personification) and LLM-generated soft labels with reasoning traces. This expanded dataset provides a complementary perspective, which enriches the original binary setup with gradable metaphor judgments and interpretability signals from LLMs. A more detailed explanation of our multi-label annotation schema can be found

in Sánchez-Montero et al. (2024).

The inherent subjectivity of metaphor identification led to varying degrees of agreement among the annotators. To capture the spectrum of agreement and disagreement at the individual tweet level, we generated soft labels. For each tweet and each metaphor type (Direct, Indirect, Personification), a soft label was calculated as the proportion of annotators who assigned that label at both the binary and multi-label levels.

- For the **binary level**, the soft label represents the probability of a tweet being metaphorical, calculated as the proportion of annotators who labeled it as such (ranging from 0/3 to 3/3).

- For the **multi-class level**, for each tweet and each metaphor type (Direct, Indirect, Personification), a soft label was calculated as the proportion of annotators who assigned that specific label (ranging from 0/3 to 3/3). This allows for tweets to potentially have soft labels across multiple metaphor categories, reflecting the possibility of containing more than one type of metaphor.

Table 1 presents the distribution of tweets according to annotators' levels of agreement, represented by the soft labels, for both binary classification (metaphorical vs non-metaphorical tweets) and specific metaphor types. Our findings reveal a remarkable level of disagreement among annotators in all rating categories. Looking at the binary level, we see that almost half of the tweets (1780 tweets or 47.7%) showed some level of disagreement among annotators (1229 with 1/3 agreement and 551 with 2/3 agreement), in contrast to 1953 tweets (52.3%) where there was perfect consensus, although class 0 (non-metaphorical) is the most prevalent.

When examining specific metaphor types, the data in the 'Disagreement' columns (Soft Labels 0.33 and 0.66) further highlight the difficulty in consistently identifying and classifying Direct (with disagreement in 291 tweets, or 7.8%), Indirect (1340 tweets, or 35.9%, showed disagreement), and Personification (disagreement in 597 tweets, or 16.0%) metaphors. Furthermore, the relatively low number of tweets where annotators achieved perfect positive agreement (Label 1.0) for Direct (only 8), Indirect (99), and Personification (15), especially when considering the total size of the dataset, underscores how challenging it is to reach complete consensus on the specific type of metaphor in each tweet.

Notably, across all metaphor categories analyzed, the distribution of inter-annotator agreement follows a consistent pattern from highest to lowest frequency in the soft labels: $0.0 > 0.33 > 0.66 > 1.0$. This distribution, where perfect positive agreement (Label 1.0) is consistently the least frequent outcome, may suggest that instances of what would be unanimously considered a clear metaphor are relatively rare in this corpus. Taken together, these distributions could support the perspective that metaphoricity exists along a spectrum of gradability, rather than conforming to strict, boundary-defining categories. This rich information about human disagreement, captured by our soft labels, served as the basis for our prompting experiments with LLMs.

4 Soft Metaphor Detection through Prompting

Our research employed a multiphase experimental methodology to explore and model the gradability of metaphoricity and conceptual mapping in the context of linguistic annotation with LLMs. This graded approach allowed us to refine LLM interaction strategies prior to large-scale evaluation and final corpus annotation, aiming to understand the LLM's ability to perform nuanced analogical abstractions on metaphor.

4.1 Qualitative Exploration and Prompt Design

The initial phase of our experimental approach consisted of a qualitative exploration of various prompting strategies to assess their potential ability to elicit GPT-4o responses that reflected the complexity and variability of metaphoricity. For this purpose, we selected a reduced set of 30 tweets from our dataset and applied the following prompt settings to the GPT-4o model:

- Zero-Shot (ZS): We asked the LLM for a binary classification (metaphorical/non-metaphorical tweets) with probability and reasoning process without prior definitions or examples.

- One-Shot with Definition for Binary Classification (1S-Def-Bin): We included a metaphor definition (source-target domain connection) and an example, requesting binary classification with probability and reasoning process.

Metaphor Category	Perfect Agreement			Disagreement		
	Label 0.0	Label 1.0	Total Tweets	Label 0.33	Label 0.66	Total Tweets
Binary Classification	1753	200	1953	1229	551	1780
Direct (D)	3434	8	3442	236	55	291
Indirect (I)	2294	99	2393	1000	340	1340
Personification (P)	3121	15	3136	498	99	597

Table 1: Soft-Label Distribution by Levels of Inter-Annotator Agreement in the Dataset

- Few-Shot with Definitions for Multi-label Classification (FS-Def-Multi): We provided definitions and examples for metaphor types (Direct, Indirect, Personification), requesting multi-label classification with probabilities for metaphorical tweets and reasoning process.

- Few-Shot with Definitions and Chain-of-Thought for Metaphorical Tweets (FS-Def-CoT-M): We added a step-by-step reasoning protocol (chain-of-thought) for multi-label classification, requesting the LLM to perform this process on tweets assumed to be metaphorical. The CoT sought to break down and guide the steps of LLM analogical reasoning for the fine classification of metaphor types.

- Few-Shot with Definitions and Chain-of-Thought for Binary and Multi-label Classification (FS-Def-CoT-BM): Combined the initial binary classification with multi-label classification by metaphor type, asking for probabilities and reasoning process, while applying the chain-of-thought protocol for metaphor identification. This integrated strategy sought to simulate a more complete analogical reasoning process (similar to that of human annotators), from the binary identification of the concept mapping to the detailed categorization of its type.

- Few-shot with Definitions, Chain-of-Thought, and Human Simulation (FS-Def-CoT-Sim): A variation of the previous prompt where an explicit instruction was added to the LLM to simulate the average of three human annotations when determining the binary probability and, if metaphorical, to follow the CoT process for multi-label classification with probabilities considering the same simulation.

The evaluation in this phase was primarily qualitative. We manually reviewed the LLM's responses to observe how it interpreted the instructions, its ability to identify potential metaphorical language and classify metaphor types according to the provided definitions, the quality and structure of its reasoning processes, and its capacity to assign probabilities that seemed to reflect uncertainty or degrees of metaphoricity.

Qualitatively, the ZS prompt showed some ability to assign non-binary probabilities and note subtle or conventionalized metaphors (e.g., assigning 0.6 to "plasticidad cerebral" [neuroplasticity] or 0.5 to "agujero de gusano" [wormhole], with justifications acknowledging their metaphorical origin or technical use). However, without explicit guidance, the consistency and alignment with our specific theoretical framework were less assured. The 1S-Def-Bin prompt appeared to guide the LLM more directly towards applying the provided definition based on source-target domain connection. Yet, interestingly, this prompt seemed to introduce more 'doubt' in the model for some tweets perceived as clearly non-metaphorical by human annotators (Soft Label 0.0), leading it to assign small but non-zero probabilities more frequently than the ZS prompt. For instance, in the case of "agujero de gusano," which had perfect human agreement as metaphorical (Soft Label 1.0), the ZS prompt assigned a higher probability (0.5) and a justification more open to the term's metaphorical origin than the 1S-Def-Bin prompt, which assigned a lower probability (0.3) arguing its technical use made it less metaphorical. This suggests that while a definition provides structure, it may sometimes override other signals the LLM captures in a zero-shot setting that are relevant to human judgment, leading to unexpected deviations.

For multi-label classification, the FS-Def-Multi prompt successfully elicited probabilities across different metaphor types, demonstrating the LLM's capacity for multi-label soft assignment and for differentiating between distinct forms of analogical manifestation. The addition of a Chain-of-Thought (CoT) protocol in subsequent few-shot prompts

(like FS-Def-CoT-M and FS-Def-CoT-BM) generally led to more structured and detailed explanations, where the LLM explicitly broke down its reasoning based on domain identification and type characteristics, simulating the steps involved in the analogical reasoning for the classification of conceptual mappings by type. The most comprehensive strategy, FS-Def-CoT-BM, showed promise in simulating a multi-stage annotation process, reflecting a more complete analogical process.

For instance, for the tweet "Es como si solo tuvieras 93 tipos diferentes de piezas de Lego y con ellas pudieras armar todo el universo" ("It's as if you only had 93 different types of Lego pieces and with them you could assemble the entire universe") (Human labels: D=0.66, I=0.66), the FS-Def-Multi prompt assigned high probability to D (1.0) but low to Indirect (0.2), while a Few-Shot + CoT variant (referring to FS-Def-CoT(BM) here) assigned slightly lower to D (0.9) and higher to I (0.6), more closely reflecting the human annotators' equal emphasis on both types. On another example, "Por primera vez, los científicos detectan los «gritos» de las plantas cuando son cortadas" ("For the first time, scientists detect the 'screams' of plants when they are cut"), there was a qualitative difference in interpretation: while human annotators saw a strong Indirect Metaphor and no Personification (Human labels: I=1.0, P=0.0), the LLMs (using FS-Def-Multi, FS-Def-CoT(BM), and FS-Def-CoT-Sim prompts) consistently assigned high probability to Personification (1.0, 1.0, 0.95 respectively) and low to Indirect (0.2, 0.4, 0.05), highlighting a divergence in how the models perform this specific analogical mapping compared to the human consensus in this instance.

The FS-Def-CoT-Sim prompt showed particular promise in its attempt to model the outcome of collective human judgment. Qualitatively, it sometimes produced binary probabilities that reflected intermediate levels of human disagreement. For instance, for the tweet "Cuando nace una estrella sigue agregando materia de la nube que se formó..." ("When a star is born it continues to add matter from the cloud that formed it..."), which had a human binary agreement of 0.66, this prompt assigned a binary probability of 0.40, providing a score within the disagreement range. Furthermore, this prompt's multi-label assignments sometimes aligned well with human multi-label distributions even when the binary was intermediate. For the tweet "La dopamina interfiere en la función de tu reloj interno..." ("Dopamine interferes with the function of your internal clock..."), while the human binary was 1.0, this prompt assigned 0.75; however, its multi-label score for Indirect (0.85) aligned closely with the human score (1.0), suggesting it could capture the specific type of analogical mapping even when its overall certainty differed.

To complement the qualitative exploration of prompting strategies, we calculated the Mean Absolute Difference (MAD) between the soft label assigned by the LLM and the corresponding human soft label for each tweet, averaging this value across the set of tweets tested with each prompt. This simple metric gives us an initial indication of the LLM's closeness to human judgments on these examples. It is crucial to emphasize that these results are based on very small samples and are not generalizable to the full corpus. Table 2 presents the MAD for the prompting strategies evaluated in this phase, for both binary classification and the multi-label categories. A lower MAD indicates better preliminary alignment with human soft labels for that category and prompt strategy on the tested samples.

According to preliminary results, for binary classification, the strategy incorporating the Human Simulation instruction shows the lowest MAD (0.157), suggesting it may capture the overall presence/absence judgment with potentially better alignment to human consensus levels in this preliminary sample. For multi-label classification, the picture is more nuanced across categories. Looking at the Average MAD (Multi-label) across all three types, the "Few-shot with Definitions + CoT (Binary & Multi-label)" prompt shows a slightly lower average MAD (0.139) compared to the "Few-shot with Def. + CoT + Human Simulation" prompt (0.160) and the simpler multi-label prompts without CoT. While these results provide initial quantitative justification for selecting the most promising prompting strategies for larger-scale evaluation, an important consideration when implementing complex strategies like Chain-of-Thought (CoT) is the increased token consumption. This, in turn, translates to higher computational cost. However, the qualitative observation of more structured reasoning and the logical appeal of guiding the LLM through complex classification steps strongly suggest that CoT could lead to a more robust and interpretable model in its analogical processing, particularly for capturing the nuances and variability of metaphoricity. Similarly, the human simulation

Prompting	Binary	Direct	Indirect	Personif.	Avg. Multi
ZS	0.285	—	—	—	—
1S-Def-Bin	0.250	—	—	—	—
FS-Def-Multi	—	0.388	0.378	0.321	0.362
FS-Def-CoT-M	—	0.118	0.337	0.351	0.269
FS-Def-CoT-BM	0.198	**0.084**	0.185	0.149	**0.139**
FS-Def-CoT-Sim	**0.157**	0.198	**0.180**	**0.102**	0.160

Table 2: Preliminary quantitative results comparing the Mean Absolute Deviation (MAD) between LLM predictions and human annotators across prompting strategies.

strategy showed potential for eliciting responses that more closely approximated patterns of human agreement/disagreement.

4.2 Quantitative Evaluation on a Larger Sample

Following the qualitative exploration and preliminary quantitative analysis, the second phase of our methodology focused on conducting a more rigorous quantitative evaluation of promising prompting strategies, model configurations, and parameters on a larger sample of the corpus. The primary objective was to obtain statistically more robust metrics to assess the LLM's ability to generate soft labels that could align with human annotation, capture disagreement patterns, and model the gradability of metaphoricity, with the aim of informing the selection of the final approach for full corpus annotation.

For this phase, a stratified random sample of 750 tweets (ca. 20% of the corpus) was selected from the total 3733 tweets. Stratification ensured that the sample represented the distribution of soft labels observed in the full dataset, reflecting the varying levels of human agreement encountered in the data, from clear cases to instances of significant disagreement. We conducted several experiments by applying different configurations to this sample, including baseline zero-shot prompting, few-shot prompting with and without human simulation instruction and a brief reasoning protocol, prompts that included more extensive elements from the annotation guide and additional few-shot examples, as well as model and temperature tuning. For all experiments in this phase, only the probabilistic soft labels (binary and multi-label) were requested as output from the LLM; reasoning processes were not included in the output.

For each category (Binary, Direct, Indirect, Personification), we computed the Mean Absolute Difference (MAD), Pearson Correlation, and Binary Cross-entropy between the LLM's soft labels and the corresponding human soft labels across the 750-tweet sample. Table 3 presents these metrics for all tested configurations.

Analysis of the metrics reveals that replicating human judgments varies significantly across configurations and metaphor categories. While several few-shot configurations achieved low MADs and Binary Cross-entropy for Direct metaphor (indicating good average alignment), the Pearson Correlation across all categories and configurations remains relatively low. This highlights the challenge in getting an advanced LLM to replicate the specific tweet-level patterns of human disagreement.

Overall, the `gpt-4.1 (Few-shot + Gradable Examples + Human Simulation)` configuration stands out in terms of capturing the overall linear trend and variability of human judgments, particularly for the crucial binary classification (highest Pearson Correlation). While some `gpt-4o` configurations, especially with temperature tuning, show competitive or slightly better MAD and CE for certain categories, the superior binary correlation of the `gpt-4.1` configuration makes it the most promising for modeling the gradability of metaphoricity and aligning with human soft labels. Given the importance of the binary decision as a precursor to multi-label classification, and the potential for better capturing the spectrum of agreement, we selected the `gpt-4.1` configuration for the final corpus annotation.

4.3 Corpus Annotation with LLM

After deciding on the LLM configuration identified and validated in the previous phase, we instructed GPT-4.1 to annotate the full corpus of 3733 PCS tweets. The objective was to generate a comprehensive dataset annotated with LLM-assigned soft labels for metaphoricity, capturing both binary presence and multi-label classification across different types, while also incorporating elements to facil-

Prompt Strategy / Model	**Pearson Correlation (↑)**				**MAD (↓)**				**Binary Cross-entropy (↓)**			
	Bin	Dir	Ind	Per	Bin	Dir	Ind	Per	Bin	Dir	Ind	Per
gpt-4.1 (Few-shot + Gradable Ex + HumSim)	**0.392**	0.289	**0.216**	**0.185**	0.216	0.048	**0.165**	**0.073**	4.589	**0.797**	3.529	**1.877**
gpt-4o + Few-shot + Guide Details	0.293	0.133	0.168	-0.018	0.236	**0.031**	0.245	0.077	2.290	0.863	1.848	1.986
gpt-4o Few-shot - Human Simulation	0.214	**0.304**	0.121	0.015	0.252	**0.031**	0.214	**0.073**	4.857	0.849	3.603	1.994
gpt-4o Few-shot + Human Simulation	0.240	0.291	0.137	0.024	0.244	**0.031**	0.211	0.076	4.511	0.860	3.666	1.985
gpt-4o temp 0.2 Few-shot + Human Simulation	0.276	0.145	0.109	-0.016	0.232	0.036	0.193	**0.073**	4.134	0.845	3.132	2.002
gpt-4o temp 0.5 Few-shot + Human Simulation	0.293	0.221	0.141	0.088	0.229	0.034	0.184	0.075	4.014	0.819	**2.985**	1.926
gpt-4o temp 0.7 Few-shot + Human Simulation	0.303	0.161	0.158	0.045	**0.227**	0.034	**0.183**	**0.073**	**3.888**	0.833	3.183	1.981
gpt-4o Zero-shot	0.254	0.013	0.150	0.033	0.242	0.034	0.179	**0.073**	5.791	0.887	3.951	1.982

Table 3: Quantitative Soft Evaluation Metrics Comparison (LLM vs Human Soft Labels on 750 Tweets)

itate potential semi-supervised refinement in the future.

The design of the final prompt, refined through experimentation in previous phases, aimed to improve the LLM's sensitivity in automatic metaphor detection, particularly for nuances in Direct and Personification metaphors (which were the most difficult to identify consistently during the previous phases). We also sought to model responses that reflected the inter-annotator variability observed in the human soft labels by incorporating intermediate examples and the explicit simulation instruction.

The optimized Few-shot prompt, including clear definitions and examples for Direct, Indirect, and Personification metaphors, along with the human simulation instruction and an internal structured reasoning process, was applied to each tweet. The model was also instructed to simulate the average of three human annotations and provide a binary probability between 0 (non-metaphorical) and 1 (metaphorical). If the binary probability was ≥ 0.5 (classified as metaphorical), soft probabilities between 0 and 1 for each of the three metaphor types were also requested. To facilitate potential future analysis or semi-supervised manual review, a brief justification for the classification was included in the output only for tweets with binary probability ≥ 0.3, corresponding to the lowest probability for considering a tweet as having some degree of perceived metaphoricity by human annotators.

The resulting LLM-annotated corpus consists of 3733 tweets, each associated with a binary soft label, multi-label soft labels (if classified as metaphorical), and a brief reasoning text (for tweets with a perceived metaphoricity ≥ 0.3). For the chosen `gpt-4.1` configuration, the metrics comparing LLM predictions to human soft labels were:

- **Binary**: Pearson Correlation: **0.382**, MAD: **0.215**, Cross-Entropy: **4.229**
- **Direct**: Pearson Correlation: **0.295**, MAD: **0.053**, Cross-Entropy: **0.769**
- **Indirect**: Pearson Correlation: **0.279**, MAD: **0.165**, Cross-Entropy: **3.322**
- **Personification**: Pearson Correlation: **0.124**, MAD: **0.069**, Cross-Entropy: **1.735**

These metrics indicate that while the LLM's soft labels show a degree of alignment with human soft labels (particularly low MAD for Direct and Personification, and the highest correlation for Binary), its ability to precisely replicate the tweet-level variability and complex patterns of human disagreement remains limited, as evidenced by the low Pearson correlations across all categories. Direct metaphors showed the best average alignment

(lowest MAD) and lowest probabilistic error (lowest CE). However, based on our evaluation, Direct and Personification categories, while sometimes having low MAD, presented significant challenges for the LLM in achieving high correlation with human judgments, indicating difficulty in consistently capturing the nuances of disagreement for these specific types.

5 Conclusions and Future Work

This study explored the use of large language models (LLMs) to generate soft labels for metaphoricity in Public Communication of Science tweets written in Spanish, aiming to capture gradability and reflect human annotation variability. Through a phased approach involving prompt engineering, model evaluation, and annotation of a 3733-tweet corpus, we demonstrated that LLMs can effectively produce probabilistic soft labels for binary metaphoricity and specific types (Direct, Indirect, Personification Metaphors).

Prompt design significantly impacted performance. Quantitative evaluation revealed that while LLMs achieve reasonable average alignment with human soft labels, they face challenges in consistently replicating the tweet-level patterns of human disagreement. Pearson correlations were relatively low across all categories, highlighting this limitation in modeling human variability. Despite this, the resulting LLM-annotated corpus is a valuable resource for analyzing metaphor and metaphoricity in scientific discourse.

Future work should focus on strategies to improve LLM alignment with the precise patterns of human disagreement, potentially through advanced prompting techniques, fine-tuning on soft-labeled data, or leveraging ensemble annotation strategies. Further analysis and application of the annotated corpus to downstream tasks, such as studying metaphor trends or enabling semi-supervised annotation pipelines, remains a promising direction. This research validates LLMs as a scalable tool for complex linguistic annotation, and can serve as a basis for exploring semi-supervised approaches or future research on LLM capabilities in complex linguistic annotation tasks related to analogical mapping.

Although existing work has shown that LLMs often reflect dominant or surface-level views while failing to capture minority or nuanced perspectives (Santurkar et al., 2023; Sourati et al., 2025), our findings suggest that figurative language presents a more complex challenge than a simple majority/minority opinion divide. Metaphor understanding involves analogical reasoning, cultural grounding, and subjective interpretation—dimensions that do not always align with demographic or opinion group boundaries. Nonetheless, the broader concern about the homogenizing tendencies of LLMs resonates with our observation that LLMs often struggle to model fine-grained human disagreement. As such, we see metaphor annotation as a compelling testbed for probing alignment, interpretability, and diversity in LLM behavior, and advocate for more work at the intersection of linguistic theory, annotation practices, and model development, particularly for figurative understanding in languages beyond English.

Limitations

This study encountered several limitations inherent in the application of large language models (LLMs) to complex linguistic annotation tasks, particularly in replicating the nuances of human soft labels for metaphoricity. A primary limitation is the LLM's demonstrated difficulty in consistently capturing the fine-grained patterns of human disagreement and variability at the tweet level. While quantitative evaluation showed that the LLM could achieve reasonable average alignment with human soft labels for certain metaphor categories (indicated by low Mean Absolute Difference and Binary Cross-entropy), the relatively low Pearson correlation coefficients across all categories highlight that the model did not accurately replicate the specific instances of high or low human consensus for individual tweets.

Furthermore, the performance varied across metaphor types. While Direct metaphors generally showed better average alignment, capturing the variability for both Direct and Personification categories proved challenging, with particularly low correlation observed for Personification. Indirect metaphors also presented difficulties in achieving strong alignment across metrics. This differential performance suggests that certain types of analogical mapping may be harder for current LLMs to model in a way that fully reflects human cognitive processing and social consensus.

Another limitation lies in the inherent constraints of the prompting approach. While prompt engineering significantly influenced the LLM's perfor-

mance, the specific instructions, examples, and simulation requests used may not fully capture the multifaceted cognitive processes and contextual factors that contribute to human metaphorical judgment and inter-annotator variability. The reliance on a specific family of LLMs (GPT models) and the characteristics of the scientific tweet dataset also represent potential limitations to the generalizability of our findings. Future work should address these limitations by exploring alternative methodologies, models, and datasets to improve the replication of human disagreement patterns in LLM-based linguistic annotation.

Aknowledgements

This research was supported by UNAM through the PAPIIT project IG400325 and CONAHCYT (SECIHTI) CF-2023-G-64. Additional support was provided by the Sistema Nacional de Investigadoras e Investigadores (SNII) through a Research Assistant Scholarship.

References

Kevin Alvarez Mouravskaia. 2020. Metaphor identification for spanish sentences using recurrent neural networks. Master's thesis, Pontificia Universidad Católica del Perú.

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*.

Bogusław Bierwiaczonek. 2024. On the gradability of metaphor. *Studies in Logic, Grammar and Rhetoric*, 69(1):31–56.

Megan A. Brown, Shubham Atreja, Libby Hemphill, and Patrick Y. Wu. 2025. Evaluating how llm annotations represent diverse views on contentious topics. *arXiv preprint*.

Nicolas Antonio Cloutier and Nathalie Japkowicz. 2023. Fine-tuned generative LLM oversampling can improve performance over traditional techniques on multiclass imbalanced text classification. In *2023 IEEE International Conference on Big Data (BigData)*, pages 5181–5186. IEEE.

Craig Cormick. 2019. *The science of communicating science: the ultimate guide*. CSIRO Publishing.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

Patrick Hanks. 2006. *Metaphoricity is gradable*, page 17–35. Mouton de Gruyter.

Nicholas Ichien, Dušan Stamenković, and Keith J. Holyoak. 2024. Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor and Symbol*, 39(4):296–309.

Kaidi Jia, Yanxia Wu, Ming Liu, and Rongsheng Li. 2025. Curriculum-style data augmentation for llm-based metaphor detection. *Preprint*, arXiv:2412.02956.

Nina Julich-Warpakowski and Thomas Wiben Jensen. 2023. Zooming in on the notion of metaphoricity: Notions, dimensions, and operationalizations. *Metaphor and the Social World*, 13(1):16–36.

Arina Kostina, Marios D. Dikaiakos, Dimosthenis Stefanidis, and George Pallis. 2025. Large language models for text classification: Case study and comprehensive review. *Preprint*, arxiv:2501.08457 [cs].

George Lakoff and Mark Leonard Johnson. 1980. *Metaphors we live by*. University of Chicago press.

Yujie Lin, Jingyao Liu, Yan Gao, Ante Wang, and Jinsong Su. 2024. A dual-perspective metaphor detection framework using large language models. *Preprint*, arXiv:2412.17332.

Fernando Martínez Santiago, Miguel Ángel García Cumbreras, Manuel Carlos Díaz Galiano, and Arturo Montejo Ráez. 2014. Etiquetado de metáforas lingüísticas en un conjunto de documentos en español. *Procesamiento del Lenguaje Natural*, (53):35–42.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

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

W. Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard J. Steen. 2017. Dmip: A method for identifying potentially deliberate metaphor in language use. *Corpus Pragmatics*, 2(2):129–147.

Mateo Richi Pons-Sorolla. 2020. Analizador de lectura fácil 4.0: identificación de metáforas.

Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 84–94, Torino, Italia. ELRA and ICCL.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint*.

Elisa Sanchez-Bayona and Rodrigo Agerri. 2022. Leveraging a new Spanish corpus for multilingual and cross-lingual metaphor detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 228–240, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Alec Sánchez-Montero, Gemma Bel-Enguix, and Sergio-Luis Ojeda-Trueba. 2024. Evaluating the development of linguistic metaphor annotation in Mexican Spanish popular science tweets. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 59–64, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Alec Sánchez-Montero, Gemma Bel-Enguix, Sergio-Luis Ojeda-Trueba, and Gerardo Sierra. 2025. Disagreement in metaphor annotation of Mexican Spanish science tweets. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 155–164, Abu Dhabi, UAE. International Committee on Computational Linguistics.

M. Carmen Sanchez-Mora. 2016. Hacia una taxonomía de las actividades de comunicación pública de la ciencia. *1824 - 2049*, pages 1–9.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source - target domain mappings. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Zhivar Sourati, Farzan Karimi-Malekabadi, Meltem Ozcan, Colin McDaniel, Alireza Ziabari, Jackson Trager, Ala Tak, Meng Chen, Fred Morstatter, and Morteza Dehghani. 2025. The shrinking landscape of linguistic diversity in the age of large language models. *Preprint*, arXiv:2502.11266.

Gerard Steen. 2008. The paradox of metaphor: Why we need a three-dimensional model of metaphor. *Metaphor and Symbol*, 23(4):213–241.

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.

Cynthia Taylor and Bryan M. Dewsbury. 2018. On the problem and promise of metaphor use in science and science communication. *Journal of Microbiology & Biology Education*, 19(1):19.1.46.

Yuan Tian, Nan Xu, and Wenji Mao. 2024. A theory guided scaffolding instruction framework for LLM-enabled metaphor reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7738–7755, Mexico City, Mexico. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Stephany Nieves Uribe and Jorge Mauricio Molina Mejía. 2024. Hacia una extracción semiautomática de metáforas conceptuales en un corpus de economía a partir del procesamiento de lenguaje natural. *Estudios de Lingüística Aplicada*, (76):81–109.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. *Metaphor: a computational perspective*. Number 31 in Synthesis lectures on human language technologies. Morgan & Claypool Publishers.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032. Association for Computational Linguistics.

Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. 2024. Exploring chain-of-thought for multimodal metaphor detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101, Bangkok, Thailand. Association for Computational Linguistics.

Danni Yu. 2025. Towards LLM-assisted move annotation: Leveraging ChatGPT-4 to analyse the genre structure of CEO statements in corporate social responsibility reports. *English for Specific Purposes*, 78:33–49.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2023. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *arXiv*. Publisher: arXiv Version Number: 5.

Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics*, 29(4):534–561.

Omnia Zayed. 2021. Metaphor processing in tweets. Master's thesis, NUI Galway.

Arkaitz Zubiaga. 2024. Natural language processing in the era of large language models. *Frontiers in Artificial Intelligence*, 6.