Language Models and the Magic of Metaphor: A Comparative Evaluation with Human Judgments

Simone Mazzoli[1], Alice Suozzi[1,*] and Gianluca E. Lebani[1,2]

[1]QuaCLing Lab, Dipartimento di Studi Linguistici e Culturali Comparati, Università Ca' Foscari Venezia, Dorsoduro 1075, 30123 Venice, Italy
[2]European Centre for Living Technology (ECLT), Ca' Bottacin, Dorsoduro 3911, 30123 Venice, Italy

Abstract

This study evaluates whether Italian-trained Large Language Models (LLMs) can interpret metaphors by comparing their performance to both human judgments and human-produced interpretations. Using three datasets containing metaphors, human interpretations, and implausible alternatives, we assess model performance via log-likelihood scores. Results show that LLMs partially replicate human understanding and are influenced by expression conventionality and linguistic context.

Keywords

Metaphor Interpretation, Linguistic Evaluation, Benchmark, Italian, Language Models

1. Introduction

Metaphor is counted among the violations of the principle of compositionality, according to which the meaning of a linguistic expression can be determined based on the meaning of its individual parts and their syntactic structure [1]. It is configured as a syntactically well-formed sentence that is semantically incongruent when interpreted literally, based on the lexically-encoded meanings of its components. Its definitions have undergone numerous variations, ranging from the idea of simple lexical substitution of a literal term to that of a constitutive principle of the human conceptual system [2]. This is because, although there is general agreement that an interaction occurs between the two concepts evoked by the metaphor in determining the meaning of the metaphorical expression, a comprehensive formalization of the nature of this interaction has yet to be achieved. In fact, understanding metaphors requires the integration of linguistic, contextual, and cultural knowledge, thus representing a challenge not only for humans but also for Large Language Models (LLMs).

LLMs have seen significant growth in recent years, demonstrating excellent performance across a wide range of interpretation and language production tasks. Their ability to understand and generate textual information has revolutionized many areas of natural language processing and numerous other fields. Since their introduction, a central question has been whether these models construct plausible representations of meaning or merely memorize patterns of form [3], as captured by the well-known *stochastic parrots* metaphor [4]. Given their success, there has been growing interest in the development of LLMs optimized for contexts in which languages other than English are predominant. Although multilingual models or those primarily trained on English are capable of processing and generating text in Italian, they are often considered less capable of capturing the nuances and specific characteristics of the language [5]. The recent introduction of LLMs trained from scratch on Italian data, together with models subsequently adapted through optimization processes for a specific language, makes it particularly interesting to verify whether their ability to understand metaphors can approach that of humans.

In light of this, this study aims to examine the extent to which interpretations and related inferences produced by humans in response to metaphorical stimuli are favored by LLMs, as opposed to implausible interpretations that are either meaningless or convey the opposite of the intended meaning. A systematic preference for human-generated interpretations would suggest that the semantic representations of LLMs are sufficiently robust to produce accurate interpretations and replicate human inferential processes. More broadly, this would imply that the distributional information in text, which underpins the internal representations of these models [6], is sufficient to construct a semantic and common-sense knowledge framework capable of generating valid inferences about figurative language.

Another promising line of research at the intersection of psycholinguistics and computational linguistics explores the cognitive plausibility of LLMs, that is, the extent to which metrics derived from these models can predict human performance on cognitive tasks. This project takes a step in that direction by collecting human judgments on the conventionality of linguistic stimuli and the

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
Simone.mazzoli@unive.it (S. Mazzoli); Alice.suozzi@unive.it (A. Suozzi); Gianluca.lebani@unive.it (G. E. Lebani)
https://www.unive.it/data/persone/29007635 (S. Mazzoli); https://www.unive.it/data/persone/24102251 (A. Suozzi); https://www.unive.it/data/persone/21257857 (G. E. Lebani)

adequacy of sentence-level context for comprehending expressions. It then investigates the correlation between these human ratings and LLM performance, with the aim of evaluating the models' sensitivity to such aspects.

2. Related Works

Metaphor interpretation tasks can be grouped into three categories [7]: property extraction, word-level paraphrasing, and explanation matching. Property extraction involves identifying shared attributes between the metaphor's Topic and Vehicle (e.g., *Love is a tide* → *Love is unstoppable*), inspired by comparison-based theories such as the Salience Imbalance Theory [8, 9] and the Career of Metaphor Theory [10]. Word-level paraphrasing replaces the metaphorical term with a literal counterpart (e.g., *She devoured the novels* → *She read the novels very quickly*), though this is limited when metaphors include multiple figurative terms or when idioms are involved. Explanation matching pairs metaphors with dictionary-like glosses (e.g., *A red-letter day* → *A day of significance*), but struggles with extended metaphors.

Previous works have leveraged such tasks to assess the models' ability of interpreting metaphors. This project fits within current research efforts aimed at testing the semantic capabilities of large language models in processing metaphors, combining several innovative aspects inspired by the following studies.

Pedinotti et al. [11] tested BERT on a dataset of 100 metaphors across four syntactic types. BERT successfully distinguished between metaphorical, literal, and nonsensical variants based on pseudo-log-likelihood. Embedding analysis showed alignment with metaphorical senses, suggesting that BERT encodes metaphor-relevant features. Following this example in the organization of stimuli, the present study ensures that metaphorical expressions are balanced across fine-grained syntactic groups. This design choice addresses an often overlooked aspect in related work, which tends to rely on examples with limited structural variation or narrow contextual constraints. Furthermore, as in the aforementioned work, the stimuli and the models tested are in Italian, offering a perspective on metaphor that differs from the more commonly adopted anglocentric approach.

Tong et al. [12] developed the MUNCH dataset, which included 10,000 metaphorical sentence paraphrases and 1,500 triplets (metaphor, correct paraphrase, incorrect paraphrase). They proposed two tasks: paraphrase selection and paraphrase generation. GPT-3.5 outperformed other models but often diverged from human responses, highlighting challenges in capturing metaphorical nuance. A notable strength of this work was its attempt to accommodate the presence of multiple correct responses produced by humans, which served as an effective strategy to address the variability and intrinsic originality of linguistic expression. Similarly, the present study aims to reflect, as much as possible, the originality of speakers in generating the stimuli on which models are evaluated. To this end, multiple correct interpretations are collected and systematically compared against incorrect ones, so that subjectivity and individuality in metaphor interpretation are explicitly taken into account. Moreover, particular attention was paid to the ecological validity of the stimuli: metaphorical expressions were directly extracted from a linguistic corpus, with minimal alterations to the original excerpts. Correct interpretations used for evaluation were produced by human annotators.

With a more explicit focus on the relationship between metaphor and its interpretation, Liu et al. [13] introduced Fig-QA, a Winograd-style task that requires models to pair metaphoric expressions with their appropriate literal reformulations. Incorrect pairings may involve either mismatched metaphors or literal paraphrases that convey the opposite meaning of the original metaphor. GPT-3 performed best in zero-shot settings, though still below human level. Fine-tuned models like RoBERTa approached human accuracy, particularly when inferring literal meaning from figurative language. In Liu et al.'s setup, choosing the correct metaphor-meaning pair was equivalent to assigning a higher probability to that pair, which is the same principle used in the present study. Each metaphor in their dataset was paired with both a correct and an opposing interpretation, forming the positive and negative instances, respectively. Similarly, in this study, a distinction is drawn between plausible interpretations, which are formulated by humans, and implausible ones, represented by two distractors carefully constructed according to two distinct semantic rules. This approach prevents inflated accuracy due to models consistently rejecting only one type of distractor, thus supporting a more balanced and accurate assessment of their interpretative abilities.

3. The Magic of Metaphor: our Study

3.1. Dataset

As previously mentioned, the linguistic data used in this study include metaphors, human-generated interpretations and ratings, as well as strings functioning as distractors. The following section describes the methods employed for data collection.

3.1.1. Metaphors

The metaphors included in the dataset were manually extracted from the official records of the Italian Parliament,

specifically from debates in the Chamber of Deputies during the 16th, 17th, and 18th legislatures (covering a time span from 2008 to 2022)[1]. These records, consisting of stenographic transcripts and committee summaries, were consulted to identify metaphorical expressions, with only minimal edits. Selected text segments include variable amounts of syntactic context (e.g., coordination and subordination) to preserve interpretability of the metaphor.[2]

A political discourse corpus was selected over literary or general-purpose corpora for two main reasons. First, although poetic texts contain rich and frequent figurative language, poetic metaphors often involve extended networks of interrelated expressions, making them hard to isolate for individual analysis. In contrast, metaphors in political language are typically employed to emphasize conceptual content and are more concise due to the oral nature of parliamentary discourse. These characteristics make them easier to isolate, interpret, and analyze without compromising semantic coherence.

Second, political speech allows for more efficient metaphor identification and clearer estimation of figurative-to-literal usage ratios. For example, the word *scheletro* 'skeleton' is more likely to appear figuratively (e.g., *scheletro normativo*) in political language than in medical contexts, where it retains a purely literal meaning. A specialized corpus thus offers a clearer view of metaphor usage patterns than a general corpus, where both uses may be equally distributed.

Metaphors were annotated using the Metaphor Identification Procedure (MIP) by the Pragglejaz Group [14]. MIP operates at the word level and requires annotators to compare the contextual meaning of a lexical unit with a more basic, concrete, and historically prior meaning. A word is tagged as metaphorical if its contextual meaning contrasts with its basic meaning but can still be understood via it.

To ensure syntactic and lexical variety, the dataset was balanced across seven groups, defined by three key variables, as detailed in Table 1: (1) *pattern*, or the syntactic relation between the metaphorical term and its context marker; (2) *valency*, or the number of syntactic arguments of the metaphorical verb; and (3) *metaphorical element class*, indicating whether the metaphor is expressed by a noun, verb, or adjective. Subscript indices were used to distinguish items when two elements shared the same lexical class. An example of a metaphor from each group is provided in Table 7 in Appendix A.

The final dataset contains 140 metaphorical items, systematically balanced across syntactic patterns, valency, and lexical class of the metaphorical term, thereby offering a robust foundation for experimental and computational studies on metaphor interpretation.

Table 1
Balanced groups in the metaphor dataset

Pattern	Valency	Metaphorical Element (PoS)	Group Size (n = 140)
N_1 di N_2	None	Noun_1	20
$N \sim$ Adj	None	Noun	20
$N \sim$ Adj	None	Adjective	20
$N_1 = N_2$	None	Noun_2	20
$V \sim N$	Intransitive	Verb	20
$V \sim N$	Transitive	Verb	20
$V \sim N$	Transitive	Verb and Noun	20

3.1.2. Human Interpretations and Ratings

We collected metaphor interpretations through a questionnaire structured into four sections: informed consent, demographic data, completion instructions (in both video and text format) and the experimental section containing the metaphors. Each questionnaire included 14 metaphors, two for each balancing group, presented in random order. A total of 10 different questionnaires were created to cover the dataset of 140 metaphors.

Participants were presented with sentence prompts that followed a fixed syntactic structure and pragmatic function, deliberately designed by the researchers to ensure consistency and reduce interpretive bias stemming from linguistic variation (see Tab. 8 in Appendix A). For each metaphor, participants were asked to write one or more possible completions based on the provided standardized sentence frame. The layout of the questionnaire as viewed by the participants is provided in Appendix B. A total of 121 Italian-speaking adults ($M_{age} = 32.8$ years, $SD = 13.6$) participated in the experiment. Only one participant reported a different native language, and their responses were excluded from the analysis.

The responses were corrected for grammatical consistency where necessary, including verb agreement, merging of prepositions and articles, and the addition of copulas. Grammatically incorrect interpretations were discarded. In total, 2,540 interpretations were collected, of which 2,117 were unique[3]. The distribution of interpretations per metaphor was described using descriptive statistics: mean (18.14), median (17), standard deviation (4.57), minimum (10) and maximum (31).

[1] Official records consulted from the website of the Italian Chamber of Deputies: https://www.camera.it/leg18/221

[2] The metaphor collection process involved using a database search tool to identify lexical units in parliamentary debates by querying word roots. Each occurrence whose metaphorical nature was confirmed was subsequently added to our database.

[3] This means that 0.83% of all collected interpretations consist of duplicates, that is, identical interpretations provided by different participants in response to metaphors that tend to elicit higher agreement.

In addition, the conventionality of each metaphor was evaluated on a scale of 1 to 5, how frequently the participant hears the expression used with the same meaning. The adequacy of the context was also evaluated on the same scale, measuring whether the provided sentence context was sufficient for understanding the metaphor.

The rating collection described above allowed us to obtain an average conventionality score for each metaphor. This score reflects the degree of conventionality or novelty of the metaphor perceived by the participants. To illustrate, we report one metaphor rated as novel (e.g., (1), with an average score of 2.40) and one rated as conventional (e.g., (2), with an average score of 4.86):

(1) La Repubblica italiana con questo Governo sta diventando lo *zampirone* per l'impresa.
'The Italian Republic, with this Government, is becoming like a mosquito coil for businesses.'

(2) È un dramma determinato a sua volta dall'*esplosione* demografica dell'Africa subsahariana.
'It is a crisis caused in turn by the demographic explosion in sub-Saharan Africa.'

3.1.3. Distractors

To create implausible interpretations for the collected metaphors (i.e., distractors), inspiration was drawn from the APL Medea test [15], a standardized tool designed to assess pragmatic skills in children aged 5 to 14. One of its subtests presents a figurative metaphor, and the child must choose the image that best represents it among one correct and three distractors. These include a literal interpretation, a semantically related image, and one showing elements of the sentence without integrating them meaningfully.

In this study, a similar approach was used: two distractors were created for each of the 140 metaphors, totaling 280 distractors. They were based on alternative completions of the sentences presented to human participants (see Tab. 8), following two specific criteria: (i) Literal Distractors (LD) are plausible only if the metaphorical word is taken literally. For instance:

(3) Dei numeri *aridi* sono dei numeri che sono privi di umidità.
'Dry numbers are numbers that are devoid of moisture.'

(4) Dicendo *elefante* burocratico si intende qualcosa che ha una lunga proboscide come un elefante.
'By saying bureaucratic elephant, one means something that has a long trunk, like an elephant.'

These distractors use predicates or attributes that belong solely to the metaphor's Vehicle and not the intended Topic. (ii) Opposite Metaphorical Distractors (OMD) express the opposite meaning of the most frequently given human interpretation. For example:

(5) Si intende che il risultato è molto importante come una briciola.
'It is meant that the result is very important, like a crumb.'

(6) Dicendo *cassaforte* di eccellenze si intende qualcosa che contiene cose di poco valore come una cassaforte.
'By saying safe of excellences, it is meant something that contains things of little value, like a safe.'

In (5), *molto importante* contradicts the typical interpretation of *briciola* (small, insignificant). Similarly, in (6), *cose di poco valore* is the opposite of *preziose*, which was the dominant human interpretation of the metaphorical *cassaforte*.

3.2. Models

We evaluated six autoregressive models based on three different architectures (LLaMA, GPT-2, Mistral), trained on Italian data using two distinct approaches: LLaMAntino-2-7b (adapted model) [16], and GePpeTto [17] and Minerva (trained from scratch) [18]. Information about the models' architectures can be found in Table 2, while their training data are summarized in Table 3.

We also include several baselines for comparison. The first baseline is the accuracy level expected from random selection among interpretations (0.33). Additionally, we test two simple models based on input string length: Longest String, which always selects the interpretation with the highest number of characters, and Shortest String, which chooses the interpretation with the fewest characters. Furthermore, we adopted a model based on the Gulpease index, a readability metric designed to assess the complexity of Italian texts. The index considers the number of sentences, letters, and words in a given text segment [19]. This model consistently selects the interpretation with the highest Gulpease score.

3.3. Data analysis

This study uses log-likelihood as a measure comparable to human preference, already employed in studies on grammaticality and semantic plausibility judgments [20, 21, 22], assuming that a model capable of understanding metaphorical expressions assigns a higher probability to human-generated interpretations than to the two distractors. Autoregressive language models define a probability distribution over subsequent tokens conditioned on the sequence of prior tokens. Consequently, the probability of an entire sentence can be obtained by computing

Table 2
Models hyperparameters

Model	Architecture	Params	Layers×Heads	Hidden size	Training
GePpeTto	GPT-2 Small	117M	12×12	768	From scratch
Minerva-350M	Mistral	350M	16×16	1,152	From scratch
Minerva-1B	Mistral	1.01B	16×16	2,048	From scratch
Minerva-3B	Mistral	2.89B	32×32	2,560	From scratch
Minerva-7B	Mistral (full-context)	7.4B	32×32	4,096	From scratch
LLaMAntino-2-7B	LLaMA 2	7B	32×32	4,096	QLoRA (adapted)

Table 3
Training method and dataset composition

Model	Data size	Training data composition
GePpeTto	13GB	Italian Wikipedia + ItWac
Minerva-350M	70B tokens	$\geq$ 50% Italian: Wikipedia, EurLex, Gazzetta Ufficiale, Gutenberg, web
Minerva-1B	200B tokens	$\geq$ 50% Italian: Wikipedia, EurLex, Gazzetta Ufficiale, Gutenberg, web
Minerva-3B	660B tokens	$\geq$ 50% Italian: Wikipedia, EurLex, Gazzetta Ufficiale, Gutenberg, web
Minerva-7B	2.48T tokens	$\geq$ 50% Italian: Wikipedia, EurLex, Gazzetta Ufficiale, Gutenberg, web
LLaMAntino-2-7B	135GB	Filtered OSCAR (Italian "medium" split, 50M documents)

the product of the conditional probabilities of each token at its respective time step:

$$\tilde{P}(w_1 \dots w_N) = p(w_1) \prod_{i=2}^{N} p\big(w_i \mid w_1 \dots w_{i-1}\big) \quad (1)$$

We consider a metaphor m from the dataset of 140 metaphors, a set of interpretations of m produced by participants denoted as I, a literal distractor LD, and an opposite metaphorical distractor OMD. For each interpretation i belonging to I, the log-likelihoods of the strings i^*, LD*, and OMD* are extracted, where * indicates that the metaphor is concatenated before each string[4]. Accuracy is calculated by taking the ratio of the number of cases in which the string i^* receives a log-likelihood greater than or equal to the highest probability among the two distractors, and the cardinality of I.

$$\mathbf{ACC} = \frac{\sum_{i \in I} \mathbf{1}\big\{\tilde{P}(i^*) \geq \max\big[\tilde{P}(\text{LD}^*),\ \tilde{P}(\text{OMD}^*)\big]\big\}}{|I|}$$

$$\text{where} \quad \mathbf{1}(\phi) = \begin{cases} 1 & \text{if } \phi \text{ is true,} \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

The comparison among the three strings, as illustrated by Equation 2, was therefore carried out for all interpretations provided by human participants along with their corresponding distractors.

3.4. Results

We report in Table 4 the accuracy values achieved by the models[5], highlighting an improvement for the larger models, with the exception of LLaMAntino-2-7b, which achieves higher accuracy only compared to GePpeTto.

A chi-square test revealed that all models exhibit distributions that are significantly different from those expected for the four baselines. As shown in Figure 1, there is a trend within the Minerva family models to disfavor OMDs, and this trend is directly proportional to the size of the model. This makes it necessary to test whether, in cases where this type of distractor does not receive a higher probability, the choice between human interpre-

[4]The existence of a significant difference between the proportions of strings (interpretations and distractors) preferred by the models, comparing the two conditions, PRESENTED IN ISOLATION versus PRECEDED BY THE METAPHOR, was confirmed through chi-square tests, demonstrating the effectiveness of this manipulation and ensuring the soundness of the experimental paradigm.

[5]An additional metric, *weighted accuracy*, was computed using the full set of 2,540 interpretations, including repeated responses from multiple participants. This metric captures the model's ability to assign higher probabilities to more frequently produced interpretations. Weighted accuracy increased by 0.02 points for all LLMs except GePpeTto, which improved by 0.01, suggesting that retaining repeated interpretations has minimal impact on model comparisons.

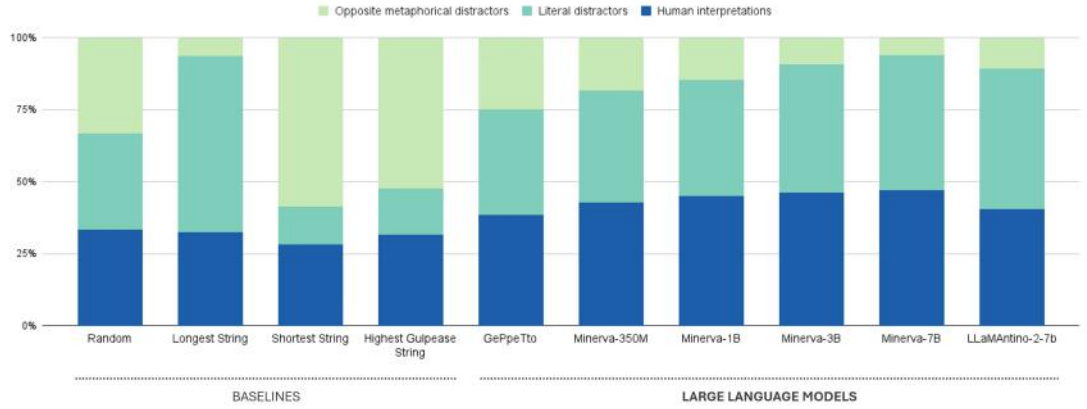

Figure 1: Percentage of sentences with the highest probability by type (model general preferences).

Table 4
Accuracy scores achieved by models

Model	Accuracy
Random	.33
Longest String	.32
Shortest String	.28
Highest Gulpease String	.32
GePpeTto	.39
Minerva-350M	.43
Minerva-1B	.45
Minerva-3B	.46
Minerva-7B	.47
LLaMAntino-2-7b	.40

tations and LDs is due to chance or to one of the simple strategies represented by the baselines.

To analyze this hypothesis, an additional chi-square test was conducted, excluding OMDs from the observations. The results allow us to reject the hypothesis that Minerva-350M randomly chooses between human interpretations and LDs ($\chi^2(1) = 14.618, p < .001$), however this is not possible for any other model in the same family. The same hypothesis can also be rejected for LLaMAntino-2-7b ($\chi^2(1) = 11.132, p < .001$) and for GePpeTto ($\chi^2(1) = 4.713, p < .05$). Yet, only for Minerva-350M and GePpeTto is it true that human interpretations are non-randomly favored, whereas LLaMAntino-2-7b, in contrast, shows a stronger preference for LDs.

In addition to the inability to reject the hypothesis of random choice between human interpretations and LDs, for Minerva-7B it was not possible to reject the hypothesis that the model always chooses the longer string between LDs and OMDs. The opposite is true for the smaller Minerva-3B model, whose results differ significantly from the expected distribution of preferences between the two distractors if it follows the "longer string" strategy ($\chi^2(1) = 18.833, p < .001$).

Table 5
Correlation between model accuracy and conventionality

Model	Pearson's *r*	sig.
GePpeTto	.131	
Minerva-350M-base-v1.0	.328	***
Minerva-1B-base-v1.0	.281	***
Minerva-3B-base-v1.0	.253	**
Minerva-7B-base-v1.0	.207	*
LLaMAntino-2-7b-hf-ITA	.187	*

$^{*}\, p < .05, ^{**}\, p < .01, ^{***}\, p < .001$

The correlation analysis in Table 5 shows a positive relationship between metaphor conventionality and model accuracy, confirming that models tend to achieve better performance on more conventional metaphors. However, the strength of this correlation varies across models. Minerva-350M shows the highest correlation. Other Minerva models follow a similar trend, with correlation values gradually decreasing as model size increases, from Minerva-1B to Minerva-7B.GePpeTto shows the lowest and non-significant correlation, whereas LLaMAntino-2-7b shows a weak but significant correlation, in line with the larger Minerva models.

The correlation analysis in Table 6 shows a positive relationship between contextual appropriateness and model accuracy, although the strength of this correlation is very low or nearly negligible for some models. Minerva-350M exhibits the highest correlation, suggest-

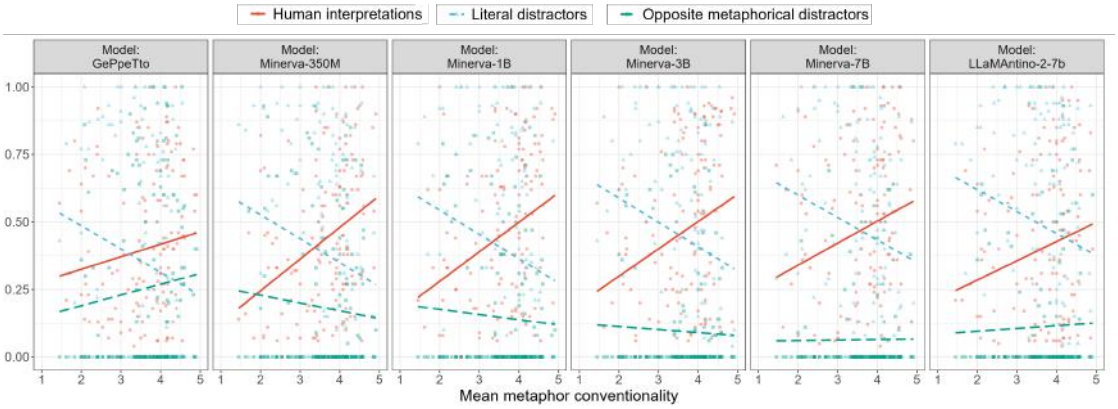

Figure 2: Proportion of sentence choices across mean metaphor conventionality

Table 6
Correlation between model accuracy and context adequacy

Model	Pearson's *r*	sig.
GePpeTto	.055	
Minerva-350M-base-v1.0	.255	**
Minerva-1B-base-v1.0	.191	*
Minerva-3B-base-v1.0	.213	*
Minerva-7B-base-v1.0	.160	
LLaMAntino-2-7b-hf-ITA	.079	

$^{*}\, p < .05, ^{**}\, p < .01, ^{***}\, p < .001$

ing that this model benefits the most from more appropriate context in determining correct interpretations. Minerva-1B and Minerva-3B show significant correlations, indicating a positive but weaker effect compared to Minerva-350M. Interestingly, the correlation observed for the larger model (3B) exceeds that of the smaller one (1B), representing an exception to the previously noted trend in which larger models tend to be less sensitive to variables derived from human judgments. Minerva-7B does not reach the threshold for significance, suggesting that in larger models, the relationship between contextual relevance and accuracy may be less relevant. The same holds for GePpeTto and LLaMAntino-2-7B with negligible correlations.

The correlation between average conventionality and model accuracy offers a solid foundation for investigating how preferences are distributed across the three string types. It enables an analysis of how increasing conventionality affects the likelihood assigned to human interpretations, to OMDs, and to LDs.

Figure 2 shows the trends in the percentages of sentences selected by the models, broken down by average conventionality. The chart illustrates how the share of strings receiving the highest probability varies with the conventionality of the metaphors. Whereas a positive correlation between human interpretation proportions (i.e., accuracy) and metaphor conventionality has been previously observed across all models (albeit non-significant for Geppetto), a one-tailed test for negative correlation revealed a slight negative correlation between average conventionality and the proportion of LDs that received the highest probability across all models: GePpeTto ($r = -.176, p < .05$), Minerva-350M ($r = -.184, p < .05$), Minerva-1B ($r = -.188, p < .05$), Minerva-3B ($r = -.189, p < .05$), Minerva-7B ($r = -.189, p < .05$), and LLaMAntino-2-7b ($r = -.168, p < .05$).

Similar analyses were conducted to examine how the average contextual adequacy of metaphors relates to the distribution of preferences across the three interpretation options. Figure 3 illustrates the proportions of interpretations that received the highest probability as contextual adequacy varies. A one-tailed test for negative correlation between contextual adequacy and the proportion of LDs with the highest probability revealed a significant relationship in both Minerva-1B ($r = -.142, p < .05$) and Minerva-3B ($r = -.171, p < .05$). Both models also show a positive correlation between contextual adequacy and the proportion of human interpretations, suggesting that these interpretations may gain preference at the expense of LDs, with minimal interference from OMDs. In Minerva-350M, while the proportion of human interpretations positively correlates with contextual adequacy ($r = .255, p < .01$), no significant negative correlation was found for either distractor type.

For further analysis, we report the accuracy of the models grouped by the syntactic pattern of the metaphors (see Fig. 4). Broadly speaking, the lowest performance was found in the group featuring a metaphorical intransitive verb combined with a literal subject. In contrast,

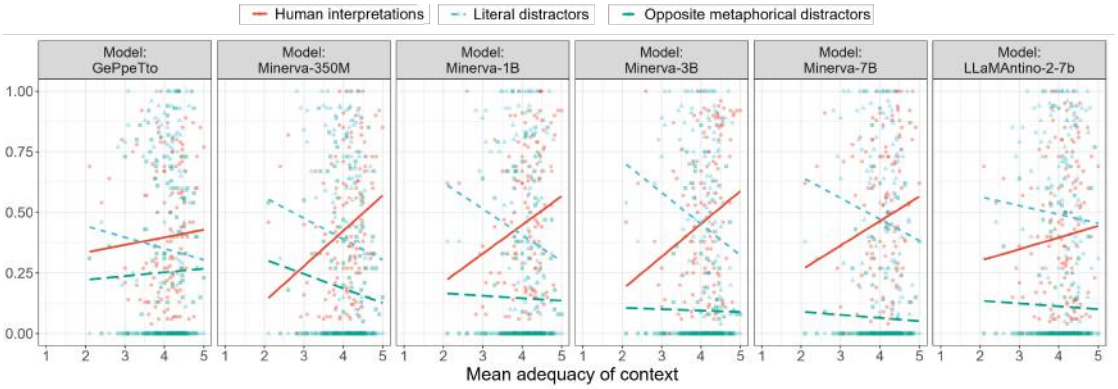

Figure 3: Proportion of sentence choices across mean metaphor adequacy of context.

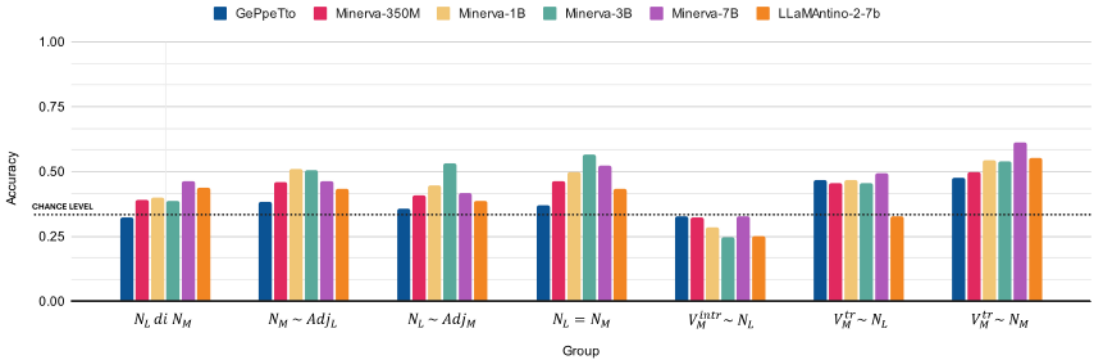

Figure 4: Accuracy grouped by the syntactic balancing group of the metaphors. The dotted line indicates chance level.

the highest accuracy was achieved on metaphors that included both a metaphorical verb and a metaphorical direct object. These trends provide evidence that specific syntactic configurations either disadvantage or support the models' ability to understand metaphors.

4. Discussion

Results highlight distinct preference patterns among language models when choosing between human interpretations and distractors. Notably, Minerva-350M and GePpeTto show a statistically significant preference for human interpretations over LDs, while LLaMAntino-2-7b favors LDs. Larger models in the Minerva family tend to disfavor OMDs, with some exhibiting behavior consistent with simple baseline strategies.

Moreover, model performance is influenced by the conventionality of the metaphor and the adequacy of contextual information. Within the Minerva family, smaller models, such as Minerva-350M, appear more sensitive to these variables, whereas the sensitivity of larger models gradually decreases. This may indicate that larger models are relatively less dependent on perceptual, stimulus-specific variables than smaller ones, likely due to their greater generalization capabilities.

Specifically, considering the results of the positive correlation test between average conventionality and model accuracy, it emerges that for most models, as the metaphors become more conventional, human interpretations are favored while LDs are gradually penalized. GePpeTto, however, does not follow this first trend, but only the second. This suggests that, when LDs are excluded by this model, human interpretations and OMDs exhibit a similar increasing trend, yet they are not equally probable: human interpretations are generally assigned higher probabilities.

The results regarding the correlation with the ade-

quacy of the sentential context in supporting the comprehension of the metaphorical expression show that, in larger models like Minerva-1B and Minerva-3B, higher contextual adequacy is associated with a reduced preference for literal distractors, and a corresponding increase in the selection of human interpretations. In contrast, Minerva-350M shows a different pattern: while the proportion of human interpretations positively correlates with contextual adequacy, neither distractor type shows a significantly correlated decrease: when human-generated interpretations are not selected, both distractor types contribute equally to the highest-probability outcome.

Furthermore, the observed performance differences across syntactic patterns may reflect underlying biases in the training data. One possible explanation for the poor results on $V_M^{intr} \sim N_L$ constructions is the over-representation in the training data of literal constructions similar to the LDs, such as example (7).

(7) Dicendo *dormire* si intende riposare.
'By saying sleep, one means to rest'

This over-representation may lead the model to favor literal readings, assigning higher probabilities to LDs. Conversely, the higher accuracy on $V_M^{tr} \sim N_M$ constructions may be due to their idiomatic nature and the presence in the training data of explanations that closely resemble human interpretations:

(8) Dicendo *fare lo struzzo* si intende nascondersi.
'By saying burying one's head in the sand, one means to hide.'

These findings collectively underscore the importance of syntactic and idiomatic features in metaphor comprehension, while also pointing to potential limitations in training data diversity.

5. Conclusion

This study explored the capacity of Italian-trained Large Language Models to interpret metaphorical expressions, evaluating their performance based on their ability to choose between human-produced interpretations and systematically designed distractors. Our findings indicate that, while no model fully replicates human-level metaphor comprehension, smaller models, particularly Minerva-350M and GePpeTto, demonstrate a statistically significant preference for human-generated interpretations over distractors.

The observed correlations suggest that distributional semantic representations, though not yet equivalent to human inferential processes, are capable of capturing figurative meaning, particularly for conventional expressions.

These results provide a nuanced picture of the current capabilities and limitations of Italian-specific LLMs in metaphor interpretation. They also underscore the importance of linguistic diversity in model training and evaluation. Future work may benefit from expanding the range of figurative phenomena studied and refining distractor generation to probe more deeply into models' semantic representations. Additionally, collecting a broader set of psychometric judgments could provide valuable insight into how these human factors correlate with model performance.

6. Limitations

This study has several limitations. First, the dataset includes only 140 metaphors, which may constrain the generalizability of the results. Second, all metaphors were drawn from parliamentary discourse, limiting coverage of metaphor use in other domains. Third, conventionality was assessed through subjective ratings, which reflect perceived rather than actual frequency of use and should therefore be considered only a proxy for true conventionality. Finally, limited access to the models' training corpora prevents clear conclusions about whether model performance reflects genuine interpretive ability or memorization of previously seen patterns.

References

[1] J. Pustejovsky, O. Batiukova, The Lexicon, Cambridge University Press, Cambridge, 2019.

[2] G. Lakoff, M. Johnson, Metaphors We Live By, University of Chicago Press, Chicago and London, 1980.

[3] M. Mitchell, D. C. Krakauer, The debate over understanding in AI's large language models, Proceedings of the National Academy of Sciences 120 (2023). doi:10.1073/pnas.2215907120.

[4] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021, pp. 610–623. doi:10.1145/3442188.3445922.

[5] A. Bacciu, C. Campagnano, G. Trappolini, F. Silvestri, DanteLLM: Let's Push Italian LLM Research Forward!, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, 2024, pp. 4343–4355. URL: https://aclanthology.org/2024.lrec-main.388.

[6] A. Lenci, M. Sahlgren, Distributional Semantics, Studies in Natural Language Processing, Cambridge University Press, 2023.

[7] M. Ge, R. Mao, E. Cambria, A survey on computational metaphor processing techniques: From identification, interpretation, generation to application, Artificial Intelligence Review 56 (2023) 1829–1895. doi:10.1007/s10462-023-10564-7.

[8] A. Ortony, Beyond literal similarity, Psychological Review 86 (1979) 161–180. doi:10.1037/0033-295X.86.3.161.

[9] A. Ortony (Ed.), Metaphor and Thought, 2 ed., Cambridge University Press, 1993. doi:10.1017/CBO9781139173865.

[10] B. F. Bowdle, D. Gentner, The career of metaphor, Psychological Review 112 (2005) 193–216. doi:10.1037/0033-295X.112.1.193.

[11] P. Pedinotti, E. D. Palma, L. Cerini, A. Lenci, A howling success or a working sea? testing what bert knows about metaphors, in: Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, 2021, pp. 192–204. doi:10.18653/v1/2021.blackboxnlp-1.13.

[12] X. Tong, R. Choenni, M. Lewis, E. Shutova, Metaphor understanding challenge dataset for LLMs, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, p. 3517–3536. doi:10.48550/arXiv.2403.11810.

[13] E. Liu, C. Cui, K. Zheng, G. Neubig, Testing the ability of language models to interpret figurative language, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 4437–4452. doi:10.18653/v1/2022.naacl-main.330.

[14] P. Group, MIP: A method for identifying metaphorically used words in discourse, Metaphor and Symbol 22 (2007) 1–39. doi:10.1080/10926480709336752.

[15] L. M. LoRusso, APL-Medea – Abilità Pragmatiche Nel Linguaggio, Giunti – OS Organizzazioni Speciali, Firenze, 2009.

[16] P. Basile, E. Musacchio, M. Polignano, L. Siciliani, G. Fiameni, G. Semeraro, LLaMAntino: LLaMA 2 Models for Effective Text Generation in Italian Language, arXiv preprint (2023). doi:10.48550/arXiv.2312.09993. arXiv:2312.09993.

[17] L. D. Mattei, M. Cafagna, F. Dell'Orletta, M. Nissim, M. Guerini, Geppetto carves italian into a language model, arXiv preprint (2020). doi:10.48550/arXiv.2004.14253. arXiv:2004.14253.

[18] R. Orlando, L. Moroni, P.-L. H. Cabot, E. Barba, S. Conia, S. Orlandini, G. Fiameni, R. Navigli, Minerva LLMs: The First Family of Large Language Models Trained from Scratch on Italian Data, in: Proceedings of the 10th Italian Conference on Computational Linguistics, 2024, p. 707–719. URL: https://aclanthology.org/2024.clicit-1.77.pdf.

[19] P. Lucisano, M. E. Piemontese, Gulpease: una formula per la predizione della leggibilità di testi in lingua italiana, Scuola e Città (1988) 110–124.

[20] R. Marvin, T. Linzen, Targeted syntactic evaluation of language models, in: E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1192–1202. doi:10.18653/v1/D18-1151.

[21] C. Kauf, E. Chersoni, A. Lenci, E. Fedorenko, A. A. Ivanova, Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models, in: Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, 2024, pp. 263–277. doi:10.18653/v1/2024.blackboxnlp-1.18.

[22] C. Kauf, A. A. Ivanova, G. Rambelli, E. Chersoni, J. S. She, Z. Chowdhury, E. Fedorenko, A. Lenci, Event knowledge in large language models: The gap between the impossible and the unlikely, Cognitive Science 47 (2023) e13386. doi:10.1111/cogs.13386.

A. Appendix A

Table 7
Example metaphors for each syntactic group.

Group	Metaphor
N_L di N_M	Gli agricoltori si trovano in una *giungla* di burocrazia 'Farmers find themselves in a *jungle* of bureaucracy'
$N_M \sim Adj_L$	Ditemi voi se questa è semplificazione, mettere in piedi questo *elefante* burocratico che costerà 30 milioni di euro all'anno. 'You tell me if this is simplification — setting up this *bureaucratic elephant* that will cost 30 million euros per year'
$N_L \sim Adj_M$	L'Italia ha bisogno di una politica estera *trasparente*, matura, lungimirante e programmatica. 'Italy needs a *transparent*, mature, forward-looking, and strategic foreign policy'
$N_L = N_M$	Venezia è una *perla* che racchiude in se stessa quella che è l'identità del popolo veneto. 'Venice is a *pearl* that embodies the identity of the Venetian people'
$V_M^{\text{intr}} \sim N_L$	Il sostegno è necessario a chi oggi ha visto *evaporare*, da un giorno all'altro, il suo reddito. 'Support is needed for those who saw their income *evaporate* overnight'
$V_M^{\text{tr}} \sim N_L$	La disgustosa tappa odierna, di fatto, *narcotizza* il Parlamento. 'Today's disgraceful stage effectively *narcotizes* the Parliament'
$V_M^{\text{tr}} \sim N_M$	Questa regione *affonda le sue radici* in una cultura profonda, in un senso civico importante. 'This region *sinks its roots* into a deep culture and a strong civic spirit'

Table 8
Sample interpretations to be completed.

Group	Interpretation to be completed
N_L di N_M	Dicendo *giungla* di burocrazia si intende qualcosa che … come una giungla 'By saying *jungle* of bureaucracy, one means something that … like a jungle'
$N_M \sim Adj_L$	Dicendo *elefante* burocratico si intende qualcosa che … come un elefante 'By saying *bureaucratic elephant*, one means something that … like an elephant'
$N_L \sim Adj_M$	Una politica estera *trasparente* è una politica estera che … 'A *transparent* foreign policy is a foreign policy that … '
$N_L = N_M$	Si intende che Venezia … come una perla 'One means that Venice … like a pearl'
$V_M^{\text{intr}} \sim N_L$	Dicendo *evaporare* si intende … 'By saying *evaporate*, one means … '
$V_M^{\text{tr}} \sim N_L$	Dicendo *narcotizzare* il Parlamento si intende … il Parlamento 'By saying *narcotize* the Parliament, one means … the Parliament'
$V_M^{\text{tr}} \sim N_M$	Dicendo *affondare le radici* si intende … 'By saying *sink the roots*, one means … '

B. Appendix B

Figure 5 is an example of the actual screens seen by each participant. In the upper part of the screen the sentence containing the metaphor (*Frase* 'Sentence') is provided. In this case, the sentence is *Con questo provvedimento bruciate 80.000 posti di lavoro in dieci anni*, which means 'With this measure, you're destroying 80,000 jobs in ten years'. The metaphorical term is the verb *bruciare*, whose literal meaning is *to burn*.

Below, in a white box the participant must provide their paraphrasis of the metaphorical term (the prompt is 'By saying *to burn* 80.000 jobs you mean ..... 80.000 jobs). After having written their paraphrasis, participants must declare how much they agree with two statements on 5-points scales, whose extremes are *Completely disagree*/*Completely agree*. The first statement is *Sento usare comunemente l'espressione in corsivo con lo stesso significato* 'I often hear the italicized expression used with the same meaning', whilst the second one is *Il resto della frase è sufficiente per interpretare l'espressione* 'The rest of the sentence is enough to interpret the expression'. The relevant expression in both statements is the metaphorical one.

Figure 5: Sample page from the questionnaire employed in the study.