A framework for annotating and modelling intentions behind metaphor use

Anonymous ACL submission

Abstract

Metaphors are part of everyday language and shape the way in which we conceptualize the world. Moreover, they play a multifaceted role in communication, making their understanding and generation a challenging task for language models (LMs). While there has been extensive work in the literature linking metaphor to the fulfilment of individual intentions, no comprehensive taxonomy of such intentions, suitable for natural language processing (NLP) applications, is available to present day. In this paper, we propose a novel taxonomy of intentions commonly attributed to metaphor, which comprises 9 categories. We also release the first dataset annotated for intentions behind metaphor use. Finally, we use this dataset to test the capability of large language models (LLMs) in inferring the intentions behind metaphor use, in zero- and in-context few-shot settings. Our experiments show that this is still a challenge for LLMs.

1 Introduction

Metaphors are pervasive in literary and political discourse, but they are also frequently used in our everyday language. Therefore, they need to be interpreted by natural language understanding systems. Consider the following quote from the "I Have A Dream" speech by Dr. Martin Luther King, Jr.: *Now is the time to <u>rise</u> from the <u>dark</u> and <u>desolate valley</u> of segregation to the <u>sunlit path</u> of <u>racial justice</u>.* In this sentence, several words are used metaphorically, among which are *dark* and *sunlit*. According to the Conceptual Metaphor Theory (CMT), a single *conceptual metaphor* may underpin these diverse linguistic manifestations (Lakoff and Johnson, 1980). Conceptual metaphors are mappings that allow one to conceptualize a TARGET domain (often more complex or abstract) based on prior knowledge of a SOURCE domain (more concrete). For instance, the conceptual metaphor JUSTICE IS LIGHT allows one to understand the abstract domain of racial justice (the TARGET) in terms of the more concrete domain of light (the SOURCE). Thus, segregation is associated with a dark, gloomy place, while social justice is a bright, sunny one. On a higher level of analysis, these metaphors are used in Dr. King's speech with specific communicative goals: making complex issues intelligible, appealing to the audience's emotions and calling them to action, and possibly more.

Following the rise of CMT, metaphor theorists have increasingly focused their research on the many and varied effects that metaphor has on cognitive processes. It has been observed that in some recurring contexts, *e.g.* in political speech, metaphorical language tends to be preferred to literal language due to its effects on the receivers (Musolff, 2004). This has led some researchers, most notably Steen (2008, 2023), to emphasise the *communicative dimension* of metaphor, a dimension in which metaphors are sometimes used deliberately to produce specific effects. Computational linguists have also investigated pragmatic aspects of metaphor, such as its affective component (Piccirilli and Schulte Im Walde, 2022) and argumentative potential (Beigman Klebanov and Flor, 2013). The communicative role of metaphors can be explained in terms of the intentions (*viz.* discourse goals) that they are supposed to achieve. The literature relating metaphor and intention is rich but generally quite fragmented. With some exceptions (Roberts and Kreuz, 1994), metaphor scholars tend to focus only on isolated intentions. Hence, there is still a lack of a systematic and comprehensive account of intentions behind metaphor use and an operationalised framework enabling annotation of such intentions in linguistic data.

In this paper, we fill in this gap by systematising the existing literature on metaphor and intention, and proposing a first-of-a-kind unified taxonomy of intentions behind metaphor use. We further propose an annotation procedure and release a first dataset annotated for intentions behind metaphor use. We show that the proposed taxonomy is thus suitable for annotating metaphors in unrestricted text. We make our dataset publicly available.[1]

Using this dataset, we test GPT-4 Turbo and two Llama2-Chat models (the 13B and 70B versions) on their ability to infer the intentions behind metaphor use. The task requires the models to select one category from the taxonomy for a given metaphorical expression in a sentence. The best-performing model, GPT-4, reaches an average accuracy of 42.99% in the zero-shot setting and a slightly higher accuracy of 44.68% in the five-shot setting, demonstrating that inferring the intentions behind metaphor use is a challenging task for state-of-the-art LLMs.

2 Related work

2.1 Conceptual and deliberate metaphor

In a seminal paper, Ortony emphasized the necessary role of metaphor in everyday language (Ortony, 1975). Proponents of CMT reinforced this idea, highlighting that our own conceptual system is, at least partly, metaphorically structured (Lakoff and Johnson, 1980). Abstract concepts, *e.g.* emotions like love, are conceived of through various kinds of conceptual metaphors.

While CMT revealed the pervasive nature of metaphor in human cognition, several authors emphasized the importance of communicative aspects in the analysis of metaphors. In particular, Steen (2008) stressed the significance of discerning deliberate metaphors from non-deliberate ones. Deliberate Metaphor Theory (DMT) departs from CMT by recognizing that only metaphors intentionally used *as* metaphors in communication involve online cross-domain mappings (Steen, 2017, 2023), and non-deliberate metaphors can be processed differently–by lexical disambiguation. DMT faced criticisms, however. For instance, Gibbs (2011) highlighted the difficulty of identifying deliberate metaphors without specific linguistic markers and the unreliability of producers' conscious judgments on their own intentions. In order to address these challenges, advocates of DMT developed the Deliberate Metaphor Identification Procedure (DMIP) and clarified the distinction between deliberate and conscious use of metaphors (Reijnierse et al., 2018; Steen, 2014).

2.2 Intentions in language use

The notion of communicative intention (CI) holds a central position in the field of pragmatics. CI is the speaker's intention to convey non-natural meaning through their utterances (Grice, 1957). In its original formulation, it comprises three sub-intentions (Recanati, 1986): (i) the intention to produce a specific effect in the hearer, (ii) the intention that the hearer recognizes (i), and (iii) the intention for (i) to be fulfilled, at least in part, by its recognition. Subsequent research has shown that one can distinguish among different intentions, varying in nature–prior intention *vs* intention in action (Searle, 1983), temporal aspect–proximal *vs* prospective (Haugh and Jaszczolt, 2012), and social dimension–individual *vs* social (Ciaramidaro et al., 2007).

As stressed by Gibbs (1999), conceiving of intentions as individual mental states makes them opaque since agents are not always aware of the causes of their behaviour. Intentions should be viewed as a *social judgment* instead. Inspired by Anscombe's philosophy of action (Anscombe, 1957), in this paper we follow a "grammatical" approach and conceive of intentions as features attributed to linguistic acts. More specifically, intentions are those *reasons* that speakers may provide once asked why they resorted to certain metaphors. Ultimately, they serve as hermeneutic tools for understanding human behaviour broadly, and linguistic behaviour in particular.

2.3 Intentions and metaphor

Although there is not a common notion of intention shared among all metaphor scholars, in the literature intentions are typically formalized as prior intentions, that is, as representations in the speaker's mind of their goals. A paper by Roberts and Kreuz builds a first taxonomy of intentions for various forms of figurative language, including metaphor (Roberts and Kreuz, 1994). This taxonomy was developed through experiments where participants were asked to provide reasons for using each figure of speech. We believe that the taxonomy has some limitations. First, the participants' judgments may have been biased by a comparativist definition of metaphor (*i.e.* metaphor as implicit comparison). Second, the number of participants assigned to metaphor was too low. As a result, there is still

[1] Anonymous URL

the need for an improved, metaphor-specific taxonomy.

Previous work has explored to some extent the relation between metaphor and individual intentions. Researchers have observed how metaphors can convey emotions (Katz Fainsilber and Ortony, 1987; Fussell and Moss, 2014), persuade (Sopory and Dillard, 2002; van Stee, 2018), contribute to argumentation (Wagemans, 2016; van Poppel, 2021), serve didactic purposes (Cameron, 2003), add humor (Attardo, 2015), and cultivate intimacy among speakers and comprehenders (Cohen, 1978; Goatly, 1997). These studies highlight the multifaceted nature of intentions behind the use of metaphor in communication and inspired our novel taxonomy.

3 Taxonomy of intentions

We introduce the individual intention categories, motivating them through theoretical considerations, previous literature and examples adapted from published material.

Lexicalized metaphor. These metaphors are associated with a plain communicative intention, and the utterance is judged as meant to convey just its propositional message. For lexicalized metaphors, the question of why a metaphor was preferred over a literal paraphrase does not arise in interpretation. In Cameron's words, the metaphoric expression is "just the way to say it" (Cameron, 2003).

(1) a. I fell in love.
b. Summer bedding is looking tired.

Sentence (1a) is an example of how the language of emotions often relies on metaphors. This observation, already noted by Katz Fainsilber and Ortony (1987), aligns with the idea that emotions may be conceptualized metaphorically, as maintained by CMT. Example (1b), instead, shows how the language we use to talk about some activities tends to have its own metaphorical jargon. This is true for academic domains such as mathematics, physics and the like, but also for non-academic domains like sports or hobbies. Both examples are cases of lexicalized metaphors which constitute the most conventional way of talking about the TARGET.

Artistic use of metaphor. These metaphors are used to attribute at once a whole set of features to the TARGET. These features need not be clearly determined in advance. Ultimately, the intention is to stimulate the receiver's creative interpretation.

(2) a. It is the east, and Juliet is the Sun.
b. Fermi's mantle in physics had fallen on his young shoulders.

Some metaphors are not easily paraphrasable because they could be paraphrased in a number of different, yet equally valid, ways. The ambiguity of the metaphorical meaning can be inherent to the TARGET of the metaphor or it can be related to the set of features that the metaphor attributes. At least in poetry and literature, interpreters tend to activate multiple mappings at once (Rasse et al., 2020) and ambiguity in interpretation is shown to correlate with aesthetic liking (Jacobs and Kinder, 2017).

Visualization. The utterer might resort to a metaphor whose SOURCE is easier to visualize than the TARGET. The intention is to help the receiver to form an intuitive representation of the latter.

(3) a. It was like a very bright light was just shining outward.
b. It would bounce up and down like a yo-yo.

Metaphors often hinge on a highly concrete/imaginable SOURCE to address an abstract TOPIC[2]. This is particularly true for subjective feelings, as in example (3a). Fussell and Moss (2014) provide evidence for the ability of metaphors to express precise emotional states. More recently, Broadwell et al. (2013) developed a prototype model for automated metaphor identification partly based on imageability.

Some metaphors do not constitute mappings from the concrete to the abstract, but just from the familiar to the unfamiliar (3b). As already stressed by Ortony (1975), metaphoric expressions are often perceived as more vivid than their literal paraphrases. Thus, they can foster the formation of a more insightful mental image. Vivid metaphors can be instrumental not only for descriptive purposes. As reported in (Cameron, 2003), they can also be used to express more clearly some commands (*cf.* a PE teacher explaining their pupils how to perform a dance: *you are spokes in a wheel*).

Persuasiveness. Using a metaphor to refer to the TARGET—in a political speech, for instance—the

[2] In psycholinguistics literature, imageability refers to the property of words to easily evoke a mental image of their meaning (Paivio et al., 1968). Imageability and concreteness, thought positively correlated, might be two distinct constructs (Dellantonio et al., 2014; Gargett and Barnden, 2015).

author can give it a non-neutral connotation. This connotation is not motivated by explicit arguments. The intention is for the audience to adopt the utterer’s perspective or stance towards the TARGET.

(4) a. The islamic wave.

b. This slender and anaemic first novel by a notable poet.

As already stressed by Lakoff and Johnson (1980), metaphors generally highlight some aspects of the TARGET, while at the same time hiding others. This process of highlighting and hiding causes a *framing effect* on the receiver, whereby the TARGET is seen, as it were, through the distorting lens of the SOURCE. The availability of several experiments and of meta-studies (Sopory and Dillard, 2002; van Stee, 2018) makes the Persuasiveness category one that is most supported empirically.

Explanation. This type of metaphors are used for didactic purposes. The intention is to explain a new or already familiar concept to the addressee. There is some knowledge asymmetry in the discourse from specialists to non-specialists, *e.g.* from teacher to students.

(5) a. The atmosphere is the blanket of gases that surrounds the earth.

b. When the neutron falls apart, spits out an electron, it becomes a proton.

The clarifying effect of metaphor has been recognized in the existing study of intentions behind it by Roberts and Kreuz (1994). The role metaphors play in educational settings–*viz.* in primary education–has been analysed in detail by Cameron (2003). Moreover, there is some empirical evidence for the usefulness of certain (deliberate) metaphors in undergraduate lectures (Beger and Jäkel, 2015). However, the use of metaphors in education does not go without risks of blocking further understanding, as highlighted by Spiro et al. (1989).

Argumentative metaphor. These metaphors are part of explicit arguments intended by the author to convince the audience of a certain claim. The intention is to make the argument more compelling.

(6) a. The EU is like a big family.

b. Human beings are responsible for their actions since they are not machines.

As pointed out, among others, by van Poppel (2021), argumentative metaphors can be used to make an effective statement, either as a standpoint or as a starting point (premise) for an argument (6a). Moreover, they can also actively contribute to the flow of argumentation (6b).

Social interaction. These metaphors focus on interpersonal relations, group or cultural conventions and the like. The intention is to create or strengthen some bond between producer and receiver.

(7) a. Sleepy Joe, Crooked Hillary.

b. She passed away.

A metaphor can bring closer its maker and appreciators in a number of different ways. First, it can exploit the fact that they belong to the same group–*e.g.*, Trump’s supporters (7a). In such cases, a social metaphor is used to isolate the desired receiver from the general public (Cohen, 1978), thus reinforcing the in-group/out-group dynamic. Second, metaphor can be used to conceal a TARGET that is experienced as negative. If they understand this, the receiver becomes aware of the additional care put by the producer in their utterance. The shared awareness fosters intimacy building between the pair and stimulates empathetic effects (7b).

Humour. The intention is to entertain the addressee, to be funny. Metaphoric language is exploited for its divertive effects, which would fade in literal paraphrases.

(8) a. I’m a doormat in the world of boots.

b. You walked into what I would call a cupboard but they classed it as the bathroom.

Language is not only used to communicate. Among the many and varied uses of language, there is also the one of entertaining others, and being entertained in return. Steen (2008, 2014) cites typical cases of humorous metaphors: sports newspaper headers, jokes, riddles and so on. In fact, the expression ”humorous metaphor” could stand for an umbrella concept grouping different phenomena, as suggested by Attardo (2015). The *Resolvable Incongruity* view offers a possible explanation for the divertive potential of certain metaphors (Oring, 2003; Dynel, 2009).

Heuristic reasoning. The intention is to provide an interpretative model for a theory, an artwork, etc., typically an abstract domain which is otherwise difficult to structure and conceive of. The metaphoric expression is used to organize the addressee’s conceptualization of the TARGET, based on their prior knowledge about the SOURCE. The discourse generally remains among specialists.

(9) a. A gas is like a collection of billiard balls in random motion.

b. It is her body as the canvas, her appearance as art.

Metaphor is a matter of seeing something *as* something else, that is, of interpreting things from a certain perspective. In cognitive terms, we map the SOURCE to the TARGET in order to better understand it. Thus, a primary intention of metaphor, especially within academic contexts, is to provide an interpretation for the products of science (9a), as illustrated by Hesse in her seminal book (Hesse, 1966), or of art (9b) and literature (Ricœur, 1975).

4 Data collection and annotation

4.1 Collecting the data

In order to empirically test the proposed taxonomy, we collected and annotated data ($\sim$ 1.2k metaphors) from the VU Amsterdam Metaphor Corpus (VUAMC; Steen et al., 2010a)[3]. This freely-accessible corpus was chosen since it contains fine-grained metaphoricity annotations at word level; it includes different genres; it contains metaphors in different grammatical constructions; and it has been extended in subsequent work with other relevant annotations, such as metaphor novelty scores (Do Dinh et al., 2018). Metaphor-related words (MRWs) in the VUAMC are identified following the MIPVU identification procedure (Steen et al., 2010b). The core idea behind the procedure is the distinction between *contextual* and *basic* meaning of words. Text fragments are collected from the British National Corpus (BNC) Baby (The BNC Baby, 2005), a 4-million-words corpus of English language covering 4 registers (Academic, News, Fiction, Conversation). The VUAMC encodes multiple information at word level, including information on metaphor type, distinguishing among *direct* and *indirect* metaphors.

Direct metaphors are expressions whose dictionary meaning coincides with the contextual meaning. For example, the word *ferret* in the phrase *he's like a ferret* is a direct metaphor. Indirect metaphors, instead, are defined as expressions having a more basic dictionary meaning that differs from the contextual meaning. Consider, for instance, the use of *valuable* in the sentence *teachers do a valuable work*.

Our corpus consists of 1214 MRWs collected from the VUAMC. We annotated all unique instances of direct metaphors found in the corpus (301 MRWs) and a subset of indirect metaphors (913 MRWs). The VUAMC contains redundant instances of the same direct metaphor–several MRWs correspond, *e.g.*, to the phrase *like a piñata above the teeming streets of the city*. However, for the purpose of annotating intentions the most natural unit of analysis is the phrase since the same intention is typically attributed to all MRWs in it. Thus, for each direct metaphor we assigned an intention only to one MRW. Annotators manually selected which word to annotate, based on their intuition of which lexical unit contributes the most to the metaphoricity of the phrase.

To select a subset of indirect metaphors to annotate we used Do Dinh et al. (2018)'s novelty scores. We divided all indirect metaphors into 5 bins according to their novelty scores. We opted to focus only on the top two bins—MRWs with novelty scores in [1,0.6] or (0.6,0.2]—which correspond to the most novel metaphors. Our rationale was that more creative uses of metaphor would yield more interesting material for investigating intentions. Within these indirect metaphors, we annotated 913 MRWs.

Some further cases were excluded from the annotation of intentions:

- Cases where there was *not sufficient context* to fully interpret the metaphor and assign an intention. A wider context could in general facilitate annotation since the attribution of intentions is likely informed by the surrounding discourse. Example: ”contraption!”

- Cases of *idiomatic* use. Idiom is a kind of figurative language use that should be distinguished from metaphor. While idioms represent relatively fixed and stable expressions within a linguistic community, metaphors are more productive and can show variation. Example: ”Even so, no room to swing a cat.”

- Some highly conventionalized *interjections* were also excluded since, just like idioms, they do not seem to require any active metaphorical interpretation in terms of meaning transfer. Example: ”Bloody hell!”

[3] `http://www.vismet.org/metcor/about.html`

Instances marked as cases to be excluded were not considered in subsequent study phases. The final dataset comprises 988 MRWs, each annotated with at least one intention from the taxonomy.

4.2 Annotation procedure and guidelines

The procedure for the annotation of direct and indirect metaphors consists of two key steps:

1. The annotator should distinguish lexicalized metaphors from other types of metaphors. If they perceive some intention behind the metaphor other than pure communication of information to the receiver, then they shall move on to step 2.

2. The annotator is asked to assign up to three intentions to the metaphor under analysis. In order to complete the task, they are provided with a table listing the taxonomic categories, each with its description and some examples.

The full guidelines can be found in Appendix A. In the guidelines, we provide a detailed description of the sequential steps to be followed during annotation. We also work out at length an example of annotation performed following the guidelines.

4.3 Corpus annotation

The annotation was carried out by an author of this paper, who was a Master's student in logic and philosophy of language. In addition to the 9 intention categories in the taxonomy, we also include a "dummy category" to keep track of cases where an intention could not be attributed.

Inter-annotator reliability. Another author, a metaphor researcher, annotated a subset of the data (360 MRWs). This subset is representative of the whole annotated corpus and replicates its proportions between different metaphor types: direct metaphors, indirect metaphors with novelty score in [1-0.6] and in (0.6-0.2].

We calculate inter-annotator reliability for 301 of the 360 items, to which both annotators assign at least one intention category. Their agreement in terms of Krippendorff's α (Artstein and Poesio, 2008) is 0.77, indicating moderate-to-fair agreement. More details about the metric used can be found in Appendix C.

5 Corpus analysis

We analysed our corpus to shed some light on the relationship between intentions and metaphor type, genre and novelty. Only the first attributed intention was considered for data analysis since no other intention was selected in most cases (827/988). Distribution of intention categories in the whole corpus and per metaphor type is shown in Table 1, and further analysis of metaphor type can be found in Appendix B.

	Direct	Indirect [1,.6]	Indirect (.6,.2]	**Total**
Lexicalized metaphor	9	19	379	407
Artistic metaphor	19	13	43	75
Visualization	53	11	132	196
Persuasiveness	2	15	51	68
Explanation	9	3	30	42
Argumentative metaphor	4	7	48	59
Social interaction	5	2	26	33
Humour	12	10	28	50
Heuristic reasoning	16	3	39	58

Table 1: Distribution of intentions by metaphor type.

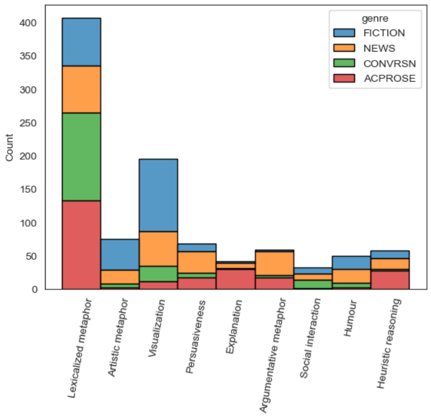
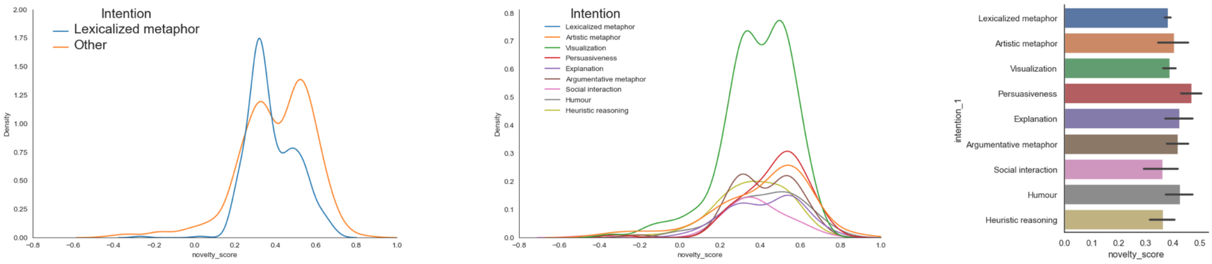


Figure 1: Distribution of intention categories per genre.

Genre. The genre of the discourse in which a metaphor appears should intuitively tell us something about its presumed intention. In particular, one would expect to find relatively more metaphors with a specific intention in extracts from certain genres, and not from others. This is suggested also by Steen (2008), who claims in passing that the function of a deliberate metaphor depends on the function of the discourse in which it is found. In the VUAMC, the information on the genre of each fragment is directly available under four intuitive tags: FICTION, NEWS, CONVRSN, ACPROSE.

In Figure 1, we report how individual intention categories (the vertical bars) are distributed over the four genres (the coloured parts in each bar). Our findings support the assertion that intentions behind metaphor use seem to correlate with the discourse genre in which the metaphor is found. For instance, Artistic metaphor and Visualization are found mostly in Fiction; Persuasiveness and



(a) (b) (c)

Figure 2: Distribution of novelty scores: (a) comparison between Lexicalized *vs* other metaphors, (b) individual distributions, and (c) mean novelty scores. Figures (a), (b) show probability densities.

Argumentative metaphor in News; Explanation and Heuristic reasoning in Academic texts; Social interaction in Conversation. All of these results agree with what one would intuitively expect. However, reality is complex and suggests that drawing one-to-one correspondences would be too simplistic. In most cases, instances of the same intention are found in all four registers. Genre can thus help to track most common uses but not all uses.

Novelty score. Information on the novelty *vs* conventionality of metaphors is crucial for understanding how different intentions are reflected in different language choices. In particular, certain intentions seem to correlate with highly conventional metaphors, while others result in more original ones.

Figure 2(a) contrasts the distribution over Lexicalized metaphor (the blue line) *vs* all other intentions merged together (the orange line). In Figure 2(b), we zoom in and plot individual distributions. Each coloured line corresponds to the distribution of a single intention category. Finally, we have computed mean novelty scores per intention with confidence intervals, as shown in Figure 2(c).

In terms of novelty, metaphors with different perceived intentions show different degrees of conventionality. Taking into account average novelty scores and estimated distributions, categories such as Persuasiveness, Explanation, Humour and Artistic metaphor are generally more original, while Lexicalized metaphor, Social interaction and Heuristic reasoning are more conventional.

6 Evaluation of LLMs

We use our dataset to test GPT-4 Turbo (`gpt-4-0125-preview`; OpenAI et al., 2023), Llama2-13B-Chat, and Llama2-70B-Chat (Touvron et al., 2023) in terms of their ability to predict the intentions behind metaphor use (for details regarding model access, parameters, and computational budget, see Appendix D). The task requires the models to choose a single intention category from our taxonomy, given a highlighted metaphorical expression in a sentence. We test the models in zero-shot and five-shot in-context learning settings. In the zero-shot setting, a short explanation for each intention category is provided.

In the five-shot settings, we randomly sample five in-context examples for each test item, and at least one of the examples is from the same intention category as the test item. Since the in-context examples implicitly explain the intention categories, we conduct two five-shot experiments: One experiment provides the same explanations for the intention categories that are used in zero-shot experiments (5-shot); the other removes those explanations from the prompt (5-shot-short). The latter setup tests whether the models are able to correctly infer what each intention category means from in-context examples. For each setting, we compute the average performance of the models across 3 different prompts, as shown in Appendix E.

Results. Table 2 shows the models' performance in these tasks in terms of accuracy. All three models reach accuracies that are above the random baseline in the zero-shot experiments, although the accuracies are still relatively low, demonstrating that this is a challenging task for the LLMs. GPT-4 is the best-performing model among the three, and the 70B Llama2-Chat model slightly outperforms the 13B one.

In the 5-shot experiments, the accuracy of GPT-4 and Llama2-13B-Chat increases when the explanations for intention categories are retained in the instructions. When the explanations are removed, the performance of all three models decreases, indicating that the models are not able to infer a correct

	0-shot	5-shot	5-shot-short
Llama2			
`-13b-chat`	24.79 (2.42)	26.75	22.90
`-70b-chat`	27.29 (5.45)	22.49	14.39
GPT-4	42.99 (1.64)	44.68	41.44
Random			13.01

Table 2: Model accuracy (%) in zero- and few-shot settings, compared to random baseline. Zero-shot accuracy is averaged over 3 runs that use different prompts; standard deviation is given in parentheses. The 5-shot-short setting removes explanations for intention categories from the prompt.

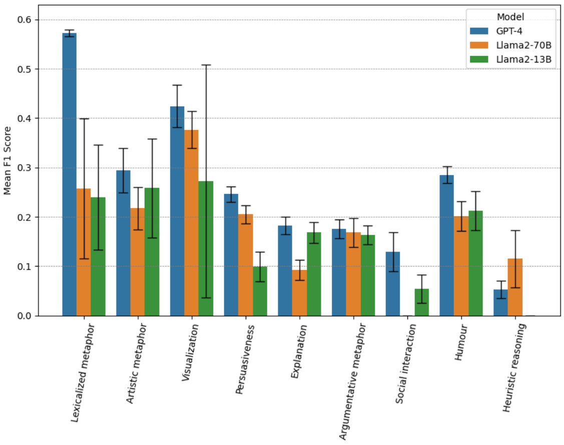


Figure 3: Model F_1 score in the zero-shot experiment, averaged across three prompts. Confidence intervals are computed with standard deviation across the prompts.

characterization of the intention from the examples. The accuracy of Llama2-70B-Chat is very close to the random baseline in the 5-shot-short setting, indicating that the model is particularly dependent on the explanations in making correct predictions.

Error analysis. Figure 3 shows the mean F_1 score for each intention category in the zero-shot experiment, averaged across the three prompts. GPT-4 reaches the highest F_1 scores when it comes to Lexicalized metaphor and Visualization, closely followed by Llama2-70B-Chat with regard to Visualization. On the other hand, GPT-4 unsurprisingly gets the worst results in the Heuristic reasoning and Social interaction categories–recall from Section 5 that these are the least represented categories in our dataset. F_1 scores for the 5-shot experiments are provided in Appendix F.

Both GPT-4 and Llama2-70B-Chat mistake Lexicalized metaphor as Visualization. These concerns conventional metaphors whose TARGET domains pertain to visible objects or the action of seeing, e.g., *a glimpse of the impact of the 1980–1 riots*, and *channels of communication*.

These two models also mistake Visualization for other intention categories, such as Artistic metaphor or Persuasiveness, e.g., *as enjoyable as feeling gently hungry or amorous*; *the wide sleeves of limp cotton hung from her freckled arms like rags thrown over a stick*. These errors can be attributed to the absence of embodied experience in LLMs. These metaphors naturally evoke a mental image or sensory experience in humans. LLMs, on the other hand, do not automatically form representations of the meaning of a text in another modality.

7 Conclusion

Contributions. In our analysis, we have gathered evidence from the existing literature and incorporated it into a novel taxonomy of intentions commonly attributed to metaphor. The taxonomy can be used to annotate metaphors in unrestricted text, as demonstrated by our corpus annotation effort. Data collected from the VUAMC helped to better understand the nature of the different intentions and how these are realized in linguistic metaphors varying in their type, genre, and novelty score. Our taxonomy should be considered as a first step towards the systematization of the various findings on discourse goals that metaphor can accomplish. Individual categories can now be investigated further, and perhaps refined in future studies.

We have also created and released a first dataset with metaphors annotated according to the taxonomy. Lastly, our experiments with GPT-4 and Llama2-Chat models show that inferring intentions behind metaphor use is still a challenging task for current LLMs.

Future directions. Since the VUAMC contains word-based annotations, we have currently adopted MRW as the basic unit for the annotation of intentions. A more natural choice, however, would be to analyse metaphoric phrases. Alternatively, the process of choosing a single MRW for each metaphoric expression could be automatized, for instance exploiting semantic information available at word level such as novelty, imageability or concreteness scores (Wilson and Division, 1997). We intend to explore both routes in future work.

Limitations

This study inevitably has some limitations; we discuss three of them here. First, the corpus used for the annotation, the VUAMC, contains mostly indirect metaphors, which are generally quite conventional. Adopting a corpus with more direct and novel metaphors would probably yield interesting results in terms of attributed intentions. However, such a corpus, comparable in size and range to the VUAMC, is missing. Second, while the output of the reliability study is encouraging, there is the need for extra annotators to make our results more trustworthy. Third, the current experimental setup asks LLMs to select only one intention category per metaphor, in contrast with our annotation guidelines. While making the task more straightforward, this choice does not reflect the complexity inherit to the analysis of metaphors in language use.

Ethical considerations

Our dataset is created from metaphors sampled from the VUAMC, which is freely and publicly accessible and suitable for research purposes. The two annotators are authors of this paper and volunteered to annotate the dataset.

References

G. E. M. Anscombe. 1957. *Intention*. Cambridge, Mass.: Harvard University Press.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Salvatore Attardo. 2015. Humorous metaphors. In Geert Brône, Kurt Feyaerts, and Tony Veale, editors, *Cognitive Linguistics and Humor Research*, pages 91–110. De Gruyter Mouton, Berlin, München, Boston.

Anke Beger and Olaf Jäkel. 2015. The cognitive role of metaphor in teaching science: Examples from physics, chemistry, biology, psychology and philosophy. *Philosophical Inquiries*, 3:89–112.

Beata Beigman Klebanov and Michael Flor. 2013. Argumentation-relevant metaphors in test-taker essays. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20, Atlanta, Georgia. Association for Computational Linguistics.

George Aaron Broadwell, Umit Boz, Ignacio Cases, Tomek Strzalkowski, Laurie Feldman, Sarah Taylor, Samira Shaikh, Ting Liu, Kit Cho, and Nick Webb. 2013. Using imageability and topic chaining to locate metaphors in linguistic corpora. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 102–110, Berlin, Heidelberg. Springer Berlin Heidelberg.

Lynne Cameron. 2003. *Metaphor in educational discourse*. Advances in Applied Linguistics. Continuum, London, UK.

A. Ciaramidaro, M. Adenzato, I. Enrici, S. Erk, L. Pia, B.G. Bara, and H. Walter. 2007. The intentional network: How the brain reads varieties of intentions. *Neuropsychologia*, 45(13):3105–3113.

Ted Cohen. 1978. Metaphor and the cultivation of intimacy. *Critical Inquiry*, 5(1):3–12.

Sara Dellantonio, Claudio Mulatti, Luigi Pastore, and Remo Job. 2014. Measuring inconsistencies can lead you forward: Imageability and the x-ception theory. *Frontiers in Psychology*, Front. Psychol.(708):1–9.

Erik-Lân Do Dinh, Hannah Wieland, and Iryna Gurevych. 2018. Weeding out conventionalized metaphors: A corpus of novel metaphor annotations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1424, Brussels, Belgium. Association for Computational Linguistics.

Marta Dynel. 2009. Creative metaphor is a birthday cake: Metaphor as the source of humour. *Metaphorik.de*, 17.

Susan R Fussell and Mallie M Moss. 2014. Figurative language in emotional communication. In *Social and cognitive approaches to interpersonal communication*, pages 113–141. Psychology Press.

Andrew Gargett and John Barnden. 2015. Modeling the interaction between sensory and affective meanings for detecting metaphor. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 21–30, Denver, Colorado. Association for Computational Linguistics.

Raymond W. Jr Gibbs. 1999. *Intentions in the Experience of Meaning*. Cambridge University Press.

Raymond W. Jr Gibbs. 2011. Are 'deliberate' metaphors really deliberate? a question of human consciousness and action. *Metaphor and the Social World*, 1(1):26–52.

A. Goatly. 1997. *The Language of Metaphors*, 1 edition. Routledge.

H. Paul Grice. 1957. Meaning. *Philosophical Review*, 66(3):377–388.

Michael Haugh and Kasia M. Jaszczolt. 2012. Speaker intentions and intentionality. In Keith Allan and Kasia Jaszczolt, editors, *Cambridge Handbook of Pragmatics*, pages 87–112. Cambridge University Press.

M. B. Hesse. 1966. *Models and Analogies in Science*. Newman history and philosophy of science series. Ind.

Arthur M. Jacobs and Annette Kinder. 2017. "the brain is the prisoner of thought": A machine-learning assisted quantitative narrative analysis of literary metaphors for use in neurocognitive poetics. *Metaphor and Symbol*, 32(3):139–160.

Lynn Katz Fainsilber and Andrew Ortony. 1987. Metaphorical uses of language in the expression of emotions. *Metaphor and Symbol - METAPHOR SYMB*, 2:239–250.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Methodology*. Sage Publications, Inc., Beverly Hills, CA.

George Lakoff and Mark Johnson. 1980. *Metaphors we Live by*. University of Chicago Press, Chicago.

Andreas Musolff. 2004. *Metaphor and political discourse: Analogical reasoning in debates about Europe*. Palgrave Macmillan London.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

E. Oring. 2003. *Engaging Humor*. University of Illinois Press.

Andrew Ortony. 1975. Why metaphors are necessary and not just nice1. *Educational Theory*, 25:45 – 53.

Allan Paivio, John C. Yuille, and Stephen A. Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, 76:1–25.

Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Rebecca J. Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of the*

Fourth International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal. European Language Resources Association (ELRA).

Prisca Piccirilli and Sabine Schulte Im Walde. 2022. What drives the use of metaphorical language? negative insights from abstractness, affect, discourse coherence and contextualized word representations. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 299–310, Seattle, Washington. Association for Computational Linguistics.

Carina Rasse, Alexander Onysko, and Francesca M. M. Citron. 2020. Conceptual metaphors in poetry interpretation: a psycholinguistic approach. *Language and Cognition*, 12(2):310–342.

François Recanati. 1986. On defining communicative intentions. *Mind and Language*, 1(3):213–41.

Gudrun Reijnierse, Christian Burgers, Tina Krennmayr, and Gerard Steen. 2018. Dmip: A method for identifying potentially deliberate metaphor in language use. *Corpus Pragmatics*, 2:129–147.

Paul Ricœur. 1975. *La métaphore vive*. Éditions du seuil, 27, rue Jacob, Paris VIe.

Richard M. Roberts and Roger J. Kreuz. 1994. Why do people use figurative language? *Psychological Science*, 5(3):159–163.

John R. Searle. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

P. Sopory and J. P. Dillard. 2002. The persuasive effects of metaphor: A meta-analysis. *Human Communication Research*, 28(3):382–419.

Rand Spiro, Paul J. Feltovich, Richard Coulson, and Daniel Anderson. 1989. Multiple analogies for complex concepts: Antidotes for analogy-induced misconception in advanced knowledge acquisition. In S. Vosniadou and A. Ortony, editors, *Similarity and analogical reasoning*, pages 498–530. Cambridge University Press.

G. J. Steen, A. G. Dorst, J. B. Herrmann, A. A. Kaal, and T. Krennmayr. 2010a. VU Amsterdam Metaphor Corpus.

Gerard Steen. 2008. The paradox of metaphor: Why we need a three-dimensional model of metaphor. *Metaphor and Symbol*, 23(1):213–241.

Gerard Steen. 2011. The contemporary theory of metaphor - now new and improved! *Review of Cognitive Linguistics*, 9(1):26–64.

Gerard Steen. 2014. Deliberate metaphor affords conscious metaphorical cognition. *Cognitive Semiotics*, 5(1-2):179–197.

Gerard Steen. 2017. Deliberate metaphor theory: Basic assumptions, main tenets, urgent issues. *Intercultural Pragmatics*, 14(1):1–24.

Gerard Steen. 2023. Thinking by metaphor, fast and slow: Deliberate metaphor theory offers a new model for metaphor and its comprehension. *Frontiers in Psychology*, 14.

Gerard Steen, Lettie Dorst, J. Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010b. *A method for linguistic metaphor identification: From MIP to MIPVU*. John Benjamins.

The BNC Baby. 2005. The BNC Baby, version 2. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

L. van Poppel. 2021. The study of metaphor in argumentation theory. *Argumentation*, 35:177–208.

Stephanie van Stee. 2018. Meta-analysis of the persuasive effects of metaphorical vs. literal messages. *Communication Studies*, 69:545–566.

J. H. M. Wagemans. 2016. Analyzing metaphor in argumentative discourse. *Rivista Italiana Di Filosofia Del Linguaggio*.

Michael Wilson and Informatics Division. 1997. Mrc psycholinguistic database: Machine usable dictionary, version 2.00. *Behav Res Methods*, 20:1–12.

A The annotation guidelines

In this task, you are asked to annotate the intentions behind direct and indirect metaphors. For each sentence you are presented with, please annotate the text delimited by <b> and < /b>. For instance, in the sentence "Usually the slightest whisper travelled like jungle <b>drums< /b> through the world of fashion" you should annotate the word "drums", following the steps that are detailed below.

- **Step 1**: decide if the metaphoric expression could be avoided.
  If there are (literal) paraphrases that would convey roughly the same message in the given context, please continue the annotation and proceed with Step 2. If you cannot think of any paraphrase that avoids the metaphor and would work just fine, then mark the metaphor as a *Lexicalized metaphor* and skip Step 2.

- **Step 2**: select categories from the taxonomy of intentions.
  In this step, you are asked to select a possible intention behind the metaphor you are analysing. The list of categories that you should use is the following one: *Artistic metaphor*, *Visualization*, *Persuasiveness*, *Explanation*, *Argumentative metaphor*, *Social interaction*, *Humour*, *Heuristic reasoning*. If you think that more intentions might play a role, feel free to select multiple categories–up to a maximum of 3.

A.1 Explanation

Lexicalized metaphors. To discriminate between lexicalized metaphors and other metaphors, try to think about the subject matter (the Topic) of the metaphor. If the metaphor is just the most common way to talk about the Topic, then mark it as *Lexicalized*. On the other hand, if the metaphor could be avoided, and the intended message could be expressed in a different way, then the metaphor is not lexicalized. Consider the following examples:

(10) a. Do you <b>follow< /b>?

b. Usually the slightest whisper travelled like jungle <b>drums< /b> through the world of fashion.

(18a) is an example of a Lexicalized metaphor. The speaker is asking the hearer if they are "following" (most likely) their words. This simply reflects the way in which we generally conceptualize discourse, namely in spatial terms (*e.g.* as a path).

On the other hand, the metaphor in (18b) is not lexicalized. The noun "drum" is not commonly used to talk about fashion. One could express the intended message through the following paraphrase "Usually the slightest whisper spread very fast and loud though the world of fashion".

Intention categories. For Step 2, try to think of which communicative goals the metaphor might accomplish better than its paraphrases. To decide which intention(s) to select, refer to the following overview of the taxonomic categories. Each item is provided with its description and some paradigmatic examples.

Intention	**Description**	**Examples**
Artistic metaphor	These metaphors are used to predicate at once a whole set of features of the Topic. These features need not to be all clearly determined in advance. Ultimately, the intention is to stimulate the receiver's creative interpretation.	• To her, the long summer days had stretched ahead, <b>world< /b> without end. • Amaldi dodged the American invitation, perhaps because (with Rome liberated) Fermi's <b>mantle< /b> in physics had fallen on his young shoulders and there were younger minds to teach. • The summer's <b>sprawl< /b> begins to be oppressive at this stage in the year and trigger fingers are itching to snip back overgrown mallows, clear out the mildewing foliage of golden rod and reduce the overpowering bulk of bullyboy ground cover.
Visualization	The utterer might resort to a metaphor whose Vehicle (*i.e.* the conventional referent) is easier to visualize than the Topic (the contextual referent). Typically, this happens when the latter belongs to an abstract domain or when the audience is not familiar with it. The intention is to help the receiver to form an intuitive representation of the Topic.	• Relief surged through her like a physical <b>infusion< /b> of new blood. • And beyond, green grass and geraniums like <b>splashes< /b> of blood. • The results are terse and sharply <b>etched< /b>, like the best line drawings.
Persuasiveness	Using the metaphor to refer to the Topic, the author gives it a non-neutral connotation, which is not motivated on explicit grounds. The intention is for the audience to adopt the utterer's positive or negative attitude towards the Topic.	• The <b>ramshackle< /b> Whitley Council negotiating machinery is the other reason why the ambulance workers have lost out. • America may have changed Presidents a year ago, but the fiscal ticket remains as <b>inpenetrable< /b> as ever. • An atmosphere <b>poisoned< /b> by mistrust.
Explanation	These metaphors are used for didactic purposes. The intention is to explain a new or already familiar concept to the addressee.	• Canals within the algae stand out as <b>rods< /b> in this kind of preservation, which is common in Ordovician rocks. • Thus one can and must say, that each fight is the singularisation of all the circumstances of the social whole in movement and that by this singularisation, it <b>incarnates< /b> the enveloping totalization which the historical process is. • The ego-identity of that person is <b>shaped< /b> by these choices.

Argumentative metaphor	These metaphors are part of explicit arguments intended by the author to convince the audience of a certain claim. The intention is to support the argument, to make it more compelling for the addressee.	• The effect is rather like an extended <b>advertisement< /b> for Marlboro Lights. • There was already a rather perfunctory air to the Queen's visit three years ago, as if it were just a required <b>coda< /b> to her tour of China. • But the villages are dying, becoming suburbs or <b>dormitories< /b> where few people work but many sleep.
Social interaction	These metaphors focus on interpersonal relations, group or cultural conventions and the like. The intention is to create or strengthen some bond between producer and receiver.	• But I'm starting to think that everything's a turn-off for you, <b>doll< /b>. • Smoking heroin ("<b>chasing< /b> the dragon") was one feature of the upsurge. • Political correctness, just as we suspected, will be perfectly <b>grey< /b>.
Humour	The intention is to entertain the addressee, to be funny. Metaphoric language is exploited for its divertive effects, which would go missing in literal paraphrases.	• Not sure of the music policy, but the name sounds like the <b>ingredients< /b> of a takeaway from a less salubrious Chinese. • From there, like a <b>buzzard< /b> in its eyrie, he would make forays round the US and abroad in spite of his advanced age. • It 's my life which is about to go down the <b>plughole< /b>.
Heuristic reasoning	The intention is to provide an interpretative model for a scientific theory, a work of art, etc. The metaphoric expression is used to organize the addressee's conceptualization of the Topic, based on their prior knowledge about another domain.	• It is her body as the <b>canvas< /b> her appearance as art. • It is as if it is walking through a <b>minefield< /b>. • At the moment, history is made without being known (l'histoire se fait sans se connaître); history constitutes, we might say today, a political <b>unconscious< /b>.

A.2 Example

Here below is one example annotated following the guidelines.

Allan Ahlberg says: "In the past, a lot of children's books seemed to be the work of talented illustrators whose pictures looked brilliant framed in a gallery, but when you tried to read the book, there was nothing there, because the words started as a <b>coat-hanger< /b> to hang pictures on."[4]

Step 1. This sentence from a news fragment is about old children's books. The author highlights the characteristic of these books of focusing more on the quality of the illustrations, rather than on the narration. The words that make up the story are metaphorically compared to coat-hangers. The utterer invites us to think of the relation between the illustrations and the words as the one existing between a coat and a coat-hanger. The latter is just instrumental, it has no purpose or value in itself which is independent of the former. Through the metaphor, the author predicates these features of the words in the children's books. The same message could have been conveyed in a literal way, along the following lines: "the words had no value in themselves, they were just instrumental for the illustrations". Thus, the output of Step 1 is that the metaphor is *not Lexicalized* and we may move on to Step 2.

Step 2. The metaphoric expression is used in this case to explain the way in which illustrations and words are related in old children's books. The author invites the addressees to understand this relation in terms of the more familiar and concrete relation between coats and hangers. For this reason, the metaphor can be annotated as *Explanation*. It should be noted, however, that also other intentions seem to play a role. For instance, one might read a negative judgment of value in the author's remark. Thus, the annotation could also be *Persuasiveness* or *Argumentative metaphor*, depending on whether some rational justification is given by the utterer to support their judgment.

B Corpus analysis: Type

Proponents of DMT maintain that direct metaphors constitute principled examples of deliberate metaphors. Since direct metaphors overtly introduce a referent from a SOURCE domain from which a conceptual mapping has to be made (Steen, 2011), they would require the intentional use of metaphor *as* metaphor. On the contrary, given the availability of a contextually relevant non-basic meaning, indirect metaphors would be non-deliberate–though ambiguous cases are possible (Steen, 2023). Thus, information on the type of linguistic metaphor would help to identify deliberate uses in communication. In Table 1, we outline the distribution of metaphors in our dataset across the intention categories for all metaphor types.

The results partially align with the claim that direct and indirect metaphors show different tendencies when it comes to their perceived intentions. While all meaningful metaphors are uttered with the minimal intention to communicate, direct metaphors generally correlate with other discourse goals, too. The categories mostly associated to direct metaphors are Visualization, Artistic metaphor, Heuristic reasoning. Indirect metaphors, especially the most conventional ones, are instead judged as lexicalized metaphors.

C Inter-annotator agreement

Our annotation task consists of a multi-label classification with multiple annotators–individual instances can be associated with multiple, non-exclusive intentions. After a brief survey of the available options (Artstein and Poesio, 2008), we opted for a variant of Krippendorff's α as an indicator of the inter-annotator agreement. In particular, we adopted the MASI distance, which is suitable for set-valued labelling tasks such as ours[5].

Out of the 360 MRWs included in the reliability study, 59 distinct items were judged as cases to be excluded by either or both of the two coders. Inter-annotator agreement was computed on the remaining 301 metaphors, where at least one intention was assigned by each annotator. The inter-annotator agreement score was 0.77, which indicates moderate-to-fair agreement. While in his seminal work Krippendorff (1980) sets 0.8 as the minimal requirement for reliable annotation schemes, we agree with Artstein and Poesio (2008), who "doubt that a single cutoff point is appropriate for all purposes" and indicate 0.7 as a more reasonable

[4] The example is taken from a News text in the VUAMC (document id: a1l-fragment01; sentence id: 29).

[5] The metric has been applied by Passonneau and colleagues to the annotation of co-reference chains (Passonneau, 2004) and Summary Content Units (Passonneau, 2006).

goal, especially for complex semantic tasks.

D Model details

The GPT-4 model is accessed through the OpenAI API, and the two Llama2-Chat models Hugging Face. We employ greedy search for all 3 models. For the two Llama2-Chat models, this is done by setting `do_sample=False` and `num_beams=1`; for the GPT-4 model, `temperature` is set to 0.

Our GPT-4 queries cost ~ 60 USD. Our Llama2-Chat queries used ~ 460 GPU hours (58946:35 SBU).

E Prompts

The prompts for zero-shot and five-shot experiments are presented in Figure 4 and 5 respectively. In the zero-shot experiments, the GPT-4 model always starts its answer with the intention category it predicts for the given metaphor. The Llama2-Chat models, on the other hand, need to generate some text (for example, *Based on the provided sentence, I would select the category of ...*) before providing its prediction. We thus provide the Llama2-Chat models the text they tend to generate at the start of their assistant messages (as part of the prompts), so that the first few new tokens they generate will be the intention category they predict.

Such assistant prompts are determined in the following way: We first take a prompt (system message and user message) that works for GPT-4 and apply it directly to a Llama2-Chat model (the 13B model for the first 2 prompts, and the 70B model for the last one). We do this for 3 different input sentences to obtain the text the model is most likely to produce before providing its prediction. This text is then used as the assistant prompt for both Llama2-Chat models. As shown in Figure 4, the 3 prompts contain different assistant messages, as we follow the messages that the Llama2-Chat models naturally produce when provided with different system prompts.

F Model performance

Figures 6 and 7 show the three models' performance (F_1 scores) in the 5-shot settings with regard to each intention category.

```
[SYSTEM]
You are a linguist. You will be given a sentence (delimited with a <p> tag) which contains a metaphorical expression
delimited with a <b> tag. Please annotate the intention behind the metaphor in <b> tag. Your answer should be one of the
following intention categories:

- Argumentative metaphor: The intention is to support an explicit argument, to make it more compelling.
- Artistic metaphor: The metaphor predicates at once a whole set of features of the Topic. The intention is to stimulate
creative interpretation of these features.
- Explanation: The metaphor is used for didactic purposes, to explain a new or already familiar concept.
- Heuristic reasoning: The intention is to provide an interpretative model for a scientific theory, a work of art, etc..
The metaphor organizes the receiver's conceptualization of the Topic.
- Humour: The intention is to entertain, to be funny.
- Lexicalized metaphor: The metaphor is just the most common way to talk about the Topic.
- Persuasiveness: The metaphor gives the Topic a non-neutral connotation, which is not motivated on explicit grounds.
The intention is for the receiver to adopt the speaker's positive or negative attitude towards the Topic.
- Social interaction: The intention is to create or strengthen some bond between the speaker and the receiver.
- Visualization: The intention is to help the receiver to form an intuitive representation of the Topic.

[USER]
<p>But there is a <b>puff</b> of dust on the horizon .</p>

[ASSISTANT]
Based on the sentence you provided, the intention behind the metaphorical expression in the <b> tag is:

* Visualization
```

```
[SYSTEM]
You are an expert at metaphor analysis, and you are asked to annotate the intentions behind metaphors. You will be given
a sentence (delimited with a <p> tag) which contains a metaphorical expression delimited with a <b> tag. Please annotate
the intention behind the metaphor in <b> tag by selecting from the following intention categories:

[...]

Please select one and only one category.

[USER]
[...]

[ASSISTANT]
Based on the provided sentence, I would select the category of "[...]
```

```
[SYSTEM]
You are a linguist. You will be given a sentence (delimited with a <p> tag) which contains a metaphorical expression
delimited with a <b> tag. Please annotate the intention behind the metaphor in <b> tag. Your answer should be one of the
following intention categories:

[...]

Please always start your answer with the intention category you choose.

[USER]
[...]

[ASSISTANT]
Category: [...]
```

Figure 4: Prompts for zero-shot experiments and example model output. The same explanations for the intention categories are used in all 3 prompts. Assistant messages in violet are provided to the Llama2-Chat models, so that model outputs always start with the predicted intention category.

```
[SYSTEM]
You are a linguist. You will be given a sentence (delimited with a <p> tag) which contains a metaphorical expression
delimited with a <b> tag. Please annotate the intention behind the metaphor in <b> tag. Your answer should be one of the
following intention categories:

- Argumentative metaphor: The intention is to support an explicit argument, to make it more compelling.
- Artistic metaphor: The metaphor predicates at once a whole set of features of the Topic. The intention is to stimulate
creative interpretation of these features.
- Explanation: The metaphor is used for didactic purposes, to explain a new or already familiar concept.
- Heuristic reasoning: The intention is to provide an interpretative model for a scientific theory, a work of art, etc..
The metaphor organizes the receiver's conceptualization of the Topic.
- Humour: The intention is to entertain, to be funny.
- Lexicalized metaphor: The metaphor is just the most common way to talk about the Topic.
- Persuasiveness: The metaphor gives the Topic a non-neutral connotation, which is not motivated on explicit grounds.
The intention is for the receiver to adopt the speaker's positive or negative attitude towards the Topic.
- Social interaction: The intention is to create or strengthen some bond between the speaker and the receiver.
- Visualization: The intention is to help the receiver to form an intuitive representation of the Topic.
[USER]
<p>Lewis got up and pulled the curtains across with a <b>pettish</b> jerk of his hands .</p>
[ASSISTANT]
Persuasiveness
[USER]
<p>‘ He told me you two were pakiteng , little <b>heifers</b> . ’</p>
[ASSISTANT]
Social interaction
[USER]
<p>This time the front door was open and a <b>swathe</b> of sunlight lay across the red-tiled floor .</p>
[ASSISTANT]
Visualization
[USER]
<p>Ruth could think of few worse fates , but her stomach <b>lurched</b> at the thought of leaving Millfield .</p>
[ASSISTANT]
Visualization
[USER]
<p>The storm was abating rapidly , the evening sky clearing in the west with the golden rays of the setting sun adding a
dying colour to the <b>sullen</b> slate blue water .</p>
[ASSISTANT]
Visualization
[USER]
<p>But there is a <b>puff</b> of dust on the horizon .</p>
[ASSISTANT]
Visualization
```

Figure 5: Prompts for five-shot experiments and example model output. The explanations for intention categories are removed in the 5-shot-short setting.

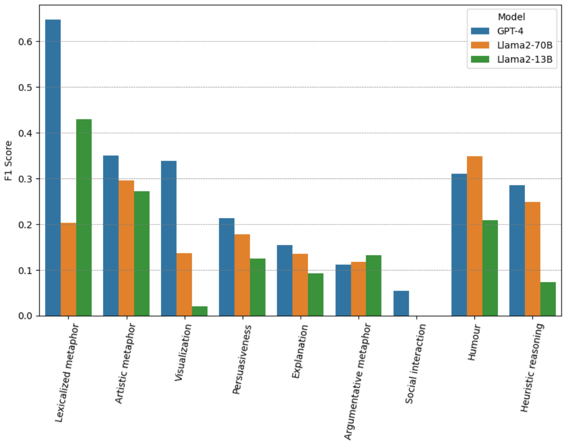


Figure 6: Model F_1 score in the five-shot experiment with explanations for the intention categories.

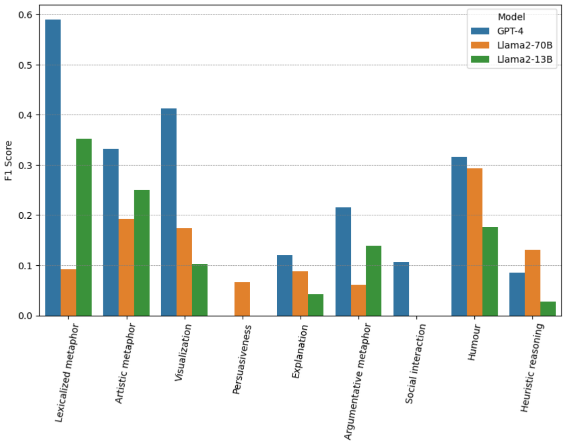


Figure 7: Model F_1 score in the five-shot experiment without explanations for the intention categories.