# Large Language Models Behave (Almost) As Rational Speech Actors: Insights From Metaphor Understanding

**Gaia Carenini, Louis Bodot**
École Normale Supérieure (ENS-PSL)
Paris, France
name.surname@ens.psl.eu

**Walter Schaeken**
KU Leuven
Leuven, Belgium
walter.schaeken@kuleuven.be

**Luca Bischetti and Valentina Bambini**
University School for Advanced Studies IUSS
Pavia, Italy
name.surname@iusspavia.it

## Abstract

What are the inner workings of large language models? Can they perform pragmatic inference? This paper attempts to characterize from a mathematical angle the processes of large language models involved in metaphor understanding. Specifically, we show that GPT2-XL model's reasoning mechanisms can be well predicted within the Rational Speech Act framework for metaphor understanding, which has already been used to grasp the principles of human pragmatic inference in dealing with figurative language. Our research contributes to the field of explainability and interpretability of large language models and highlights the usefulness of adopting a Bayesian model of human cognition to gain insights into the pragmatics of conversational agents.

## 1 Introduction

Large language models (LLMs, e.g., GPT [1], PALM [2], and LLaMA [3]), with their huge number of parameters and enormous computational power, have exhibited unprecedented skills in tasks such as language generation, translation, and understanding. Nevertheless, along with their complexity, comes challenges related to their interpretability, that is, the comprehension of the inner machinery of these models, their strengths, and limitations. Although the skills of large language models have been extensively quantified, our understanding of the underlying cognitive mechanisms that enable these skills remains on the surface.

Gaining insights concerning the cause-and-effect relationship that leads to a particular output in artificial neural networks, along with providing clear and coherent explanations for their decision, is the object of study of the field known as explainability, or interpretability. The motivation for this line of work relates to building trust in AI systems, ensuring ethical use, and enabling humans to assess and potentially correct model behavior when necessary. Investigations in these areas have shown that LLMs can be seen as few-shot-learners [1], fully-zero-shot learners [4], zero-shot-reasoners [5], few-shot table-reasoners [6], or agent models [7]. Nonetheless, none of these insights has helped in structuring a reliable model of pragmatic reasoning for LLMs able to justify their successful performance in linguistic assessments [8].

Information-Theoretic Principles in Cognitive Systems Workshop at 37th Conference on Neural Information Processing Systems (NeurIPS 2023).

The only general-purpose cognitive model for LLMs, called *bounded pragmatic speaker*, has been recently presented in [9]. In this model, the bounded pragmatic speaker embodies a speech agent that attempts to communicate pragmatically but is limited by its computational capacity. Consequently, it develops a base speaker model to opportunely resize the space of utterances under consideration, and a pragmatic model to predict how a listener would interpret each utterance. As a whole, this model captures relevant facets of pragmatic reasoning, yet it has several weaknesses. Firstly, it sees large language models almost-exclusively as speakers bypassing the fact that they are also listeners. Secondly, no experimental evidence is provided to support the model proposed. Lastly, the ways to improve LLMs, suggested by that model, demand either important structural changes or re-training, extremely expensive operations from a computational perspective. Our work partially addresses these weaknesses.

In particular, here we focus on studying how the processes supporting metaphor comprehension may take place in LLMs. In humans, the key mechanism underlying metaphorical understanding is that of pragmatic inference, extensively studied from numerous perspectives such as the pragmatic one within Relevance Theory [10, 11] and, more recently, also from the statistical and computational angle. The Rational Speech Act (RSA) framework [12, 13], is perhaps the most significant formal system able to capture pragmatic reasoning, also in the case of metaphor. RSA is both a game-theoretic and information-theoretic model that looks upon language use as an instance of a signaling game, and upon pragmatic inference as an optimization task [14].

In this work, we contribute to the interpretability and explainability of large language models by showing that the Rational Speech Act framework provides quantitatively accurate explanations for LLMs' behavior in metaphor interpretation. In particular, we exhibit how GPT2-XL model reasons according to a mechanism that can be simulated by the pragmatic listener in the RSA framework. Moreover, we provide an information-theoretic analysis of the similarity among the interpretations computed by an RSA model and by GPT2-XL. Based on the insights gained, we wrap-up providing suggestions on how to improve LLMs without any structural manipulation or re-training.

## 2 Preliminaries

Metaphor is a loose use of language where some concepts are broadened to emphasize some specific features of an element [10, 11]. An example is *love is a rose*: specific features of the rose are magnified and add further characterization to the concept of *love*, such as being *beautiful*, *delicate*.

For present purposes, we focus on nominal predicative metaphors of the form ("$X$ is $Y$ "), with $X$ (topic, that is, the subject of the metaphor), and $Y$ (vehicle, that is, the term used metaphorically to predicate about the topic) being of the category of humans, animals or objects. An example of a nominal predicative metaphor is *workers are ants*, where *workers* is the topic of the metaphor and *ants* is its vehicle, which conveys the strength and industriousness of ants.

Metaphors have been recently studied through mathematically-grounded methods within the Rational Speech Act (RSA) framework. To date, three RSA models for metaphorical understanding of increasing expressiveness are available, whose predictions largely agree with human ones [15, 16, 17]. In this work, we adopt the one introduced by Carenini and colleagues [17], which improves the previous models [i.e., 15 and 16] providing an interpretable and scalable framework based on the notion of *feature typicality*. Informally, we say that a feature is *typical* for a person, object, or animal if it is a property conventionally associated with the latter. A peculiarity of this RSA model lies in the fact that it exploits feature typicality both with respect to the topic and the vehicle.

Formally, let $\mathcal{F}$ be the set of all the features of any topic or vehicle under consideration, and let $f = [f_1, \ldots, f_n]$ with $n = \|\mathcal{F}\|$, be a vector whose entries, $f_i \in [0, 1]$, quantify how much typical $f_i$ is with respect to $X$. The first speech actor in the RSA model is the literal listener $L_0$, which interprets the utterance "$X$ is $Y$" as meaning that $X$ is literally a member of the category $Y$ and has corresponding features. Formally, if $u$ is the uttered category:

$$L_0(c, f|u) = \begin{cases} P(f|c), & \text{if } c = u \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $P(f|c)$ is the prior probability that a member of category $c$ has a feature vector $f$. The second agent of the model is the pragmatic speaker $S_1$. We assume that $S_1$ has the goal of communicating

about a feature $f_i$, formally $g_i(f) = f_i$, and that she chooses what to say according to the following utility function:

$$U(u|g, f) = \log \sum_{c,f'} \delta_{g(f)=g(f')} L_0(c, f'|u)\,. \tag{2}$$

In particular, the pragmatic speaker employs a softmax decision rule to select an utterance:

$$S_1(u|g, f) \propto e^{\lambda U(u|g,f)} \tag{3}$$

where $\lambda$ is the rationality parameter, a quantity that can be learned from data through gradient-based methods.

The last actor of the RSA model is the pragmatic listener $L_1$, who uses Bayesian reasoning to infer the intended meaning based on prior knowledge and her understanding of the speaker. Specifically, $L_1$ considers all possible speaker goals and integrates over them. $L_1$ is defined as:

$$L_1(c, f|u) \propto P(c)P(f|c) \sum_{g} \mathcal{R}(g|t)S_1(u|g, f) \tag{4}$$

where $c$, $f$, $u$, and $t$ denote respectively a category, a feature vector, an utterance, and the topic of the metaphor under discussion, and $\mathcal{R}$ is a function expressing the relevance of the goal $g$ to the topic $t$.

# 3 Methods

## 3.1 Overview

Briefly, we reproduced, replacing the human sample with a large language model, two out of the three behavioral experiments performed in the study by Carenini and colleagues [17]. For this purpose, we interrogated `GPT2-XL`, an open-source large language model developed by OpenAI [1].

We proceeded as follows. In Experiment 1, we measured features' typicality and estimated the value of conditional probabilities involved in the equation encoding for the behavior of pragmatic listeners (Equation 4). From the first experiment, we obtained $P(f|c)$ and $\mathcal{R}(g|t)$, accounting respectively for the probability that a member of category $c$ has a fixed feature $f$ and the estimator for the probability that the communicative goal is $g$ given that the metaphor topic is $t$. In Experiment 2, we retrieved the probability distribution over all the possible alternative interpretations for each of the metaphors in the study. After, we used the values obtained for $P(f|c)$ and $\mathcal{R}(g|t)$ to compute the probability distribution over all the possible interpretations according to $L_1$.

Unlike the pragmatic skills assessment experiments ran in [8], we did not rely upon the responses provided by the standard GPT user interface [18]; instead, we used prompting-based methods and directly accessed the probability distribution generated by the model. This drastically improved the quality of the previous analysis since it allowed us to take into account the entire rank of possible predictions generated by the model rather than exclusively the one outputted.

## 3.2 Experimental Pipeline

We based our study on the dataset of 84 nominal predicative metaphors (i.e., in the form "$X$ *is* $Y$") previously investigated by Roncero & De Almeida [19]. We started by isolating metaphors' topics and vehicles. Next, we grouped the properties associated with at least one of the topics or vehicles – 574 in total – into the set $\mathcal{F}$, the *features set*. For the sake of clarity, we provide explicit explanations of the entire procedure applied to the example: *workers are ants*.

**Experiment 1: Feature Typicality** *Materials:* 168 nouns (topics and vehicles of the 84 metaphors), $\mathcal{F}$. *Methods:* We run a fragment of code that for each noun, $z$, e.g., *workers* or *ants* (a) asks `GPT2-XL` to complete the prompt "$z$ *is*" (or alternatively "$z$ *are*", coherently with the arity of $z$), e.g., "*workers are*"; (b) extracts the probability distribution over all the possible completion, e.g., the probability that the model assigns to the completed sentence *workers are strong*; (c) restricts the latter to the subset of 574 features considered in the study; and (d) re-normalizes the restriction to re-obtain a probability distribution. *Results:* Estimation for $P(f|c)$ (resp. $\mathcal{R}(g|t)$) when the noun is a metaphor vehicle (resp. a metaphor topic).

[1] The model is available at `https://github.com/openai/gpt-2`

**Experiment 2: Metaphor Interpretation** *Materials:* 84 metaphors, 574 features. *Methods:* We run a fragment of code that, for each selected metaphor, $m$: (a) asks `GPT2-XL` to complete the sentence " $m$. *This means that* $t(m)$ *is*"(or alternatively "$t(m)$ are", coherently with the arity of $t(m)$) where $t(m)$ is the topic of $m$, e.g., "*Workers are ants. This means that workers are*; (b) extracts the probability distribution over all the possible completion, e.g., the probability that the model assigns to the completed sentence *Workers are ants. This means that workers are organised.*; (c) restricts the latter to the subset of 574 features considered in the study, and (d) re-normalizes the restriction to re-obtain a probability distribution. *Results:* Distribution over the interpretations generated by `GPT2-XL`.

**Implementation** We implemented in Python all the fragments of code needed to perform Experiments 1 and 2, which for the sake of reproducibility will be made available as part of the supplementary materials in the final version of this work. The implementation relies on the code in [20].

## 4 Results

We compared the distributions over all possible interpretations computed through the RSA model from the data of Experiment 1 with the ones generated by the model, identified in Experiment 2.

To carry out this comparison, we first defined the notion of $k$-*agreement* among two probability distributions to be the number of common features among the $k$ most probable ones, and remarked that the standard notion of *accuracy* for a predictive model corresponds to $1-$agreement. We obtained that in 30 out of 84 metaphors ($\sim$35.7%), the most probable feature (metaphor interpretation) was the same for both the distributions under discussion, resulting in a $1-$agreement of .36. Since a metaphor may admit more than one interpretation, we considered the agreement for increasing values of $k$. If $k = 3$, the $k$-agreement was $\sim 1.2$. In particular, 65 out of 84 metaphors (77.4%) admitted at least a common feature among the 3 most probable. When $k = 5$, the $k$-agreement was $\sim 2.4$, in particular, more than 90.4% of the metaphors admit at least a common feature among the five most probable.

Next, we estimated the global dissimilarity among the two distributions through the measure of the mean Pearson correlation coefficient, which in $r$=.47, (SD=.32). We computed the Jensen-Shannon Divergence (JSD), an information-theoretic measure of dissimilarity. The average JSD .33, (SD=.01). This index suggests that more than 67% of the information is shared among the two distributions.

## 5 Discussion

In this work, we showed the usefulness of fostering interdisciplinary connections between cognitive science – in particular, linguistics and pragmatics – and prompt-based methods to improve the interpretability of large language models. In particular, we show that GPT2-XL behaves in a way that can be represented by the RSA pragmatic listener.

More precisely, our results indicated a strong positive correlation between the probability distributions over the predictions of the RSA model (computed on typicality values extracted from GPT) and the one obtained directly questioning the model. Similarly, the value of Jensen-Shannon Divergence was surprisingly positive, and it points out that GPT2-XL might be better approximated by RSA than humans [17]. Globally, these empirical results exhibit that GPT2-XL is able to perform some form of pragmatic inference that tightly approximates the one in the Rational Speech Act framework, consistently with the formalization proposed in [9].

As a consequence, we speculate that, in order to improve the capabilities of LLMs in dealing with figurative language, it may be helpful to enforce values of typicality closer to the ones measured in humans, rather than reinforcing pragmatic inference rules, which our results suggest being already inherent in LLMs, at least in the case of deriving metaphorical meanings. Providing these models with better estimates for typicality may help them capture finer nuances of non-literal language, opening the door to new broad-spectrum applications.

## References


[1] Tom Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[2] Aakanksha Chowdhery et al. "PaLM: Scaling Language Modeling with Pathways". In: *arXiv e-prints* (Apr. 2022). DOI: 10.48550/arXiv.2204.02311.

[3] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL].

[4] Xuandong Zhao et al. "Pre-trained Language Models Can be Fully Zero-Shot Learners". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 15590–15606. DOI: 10.18653/v1/2023.acl-long.869. URL: https://aclanthology.org/2023.acl-long.869.

[5] Takeshi Kojima et al. "Large Language Models are Zero-Shot Reasoners". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 22199–22213. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

[6] Wenhu Chen. "Large Language Models are few(1)-shot Table Reasoners". In: *Findings of the Association for Computational Linguistics: EACL 2023*. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1120–1130. URL: https://aclanthology.org/2023.findings-eacl.83.

[7] Jacob Andreas. "Language Models as Agent Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 5769–5779. DOI: 10.18653/v1/2022.findings-emnlp.423. URL: https://aclanthology.org/2022.findings-emnlp.423.

[8] Chiara Barattieri di San Pietro et al. "The pragmatic profile of ChatGPT: Assessing the communicative skills of a conversational agent". it. In: *Sistemi intelligenti, Rivista quadrimestrale di scienze cognitive e di intelligenza artificiale* 2/2023 (2023), pp. 379–400. ISSN: 1120-9550. DOI: 10.1422/108136.

[9] Khanh Nguyen. "Language Models are Bounded Pragmatic Speakers: Understanding RLHF from a Bayesian Cognitive Modeling Perspective". In: *Proceedings of the First Workshop on Theory of Mind in Communicating Agents at ICML* (2023). DOI: 10.48550/arXiv.2305.17760.

[10] Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition (2nd edition)*. Oxford: Blackwell, 1995.

[11] Robyn Carston. "XIII—Metaphor: Ad hoc concepts, literal meaning and mental images". In: *Proceedings of the Aristotelian society*. Vol. 110. 2010, pp. 295–321.

[12] Michael C. Frank and Noah D. Goodman. "Predicting Pragmatic Reasoning in Language Games". In: *Science* 336.6084 (2012), pp. 998–998. DOI: 10.1126/science.1218633.

[13] Noah D. Goodman and Michael C. Frank. "Pragmatic Language Interpretation as Probabilistic Inference". In: *Trends in Cognitive Sciences* 20.11 (2016), pp. 818–829. DOI: 10.1016/j.tics.2016.08.005.

[14] Judith Degen. "The Rational Speech Act Framework". In: *Annual Review of Linguistics* 9.1 (2023), pp. 519–540. DOI: 10.1146/annurev-linguistics-031220-010811.

[15] Justine T. Kao, Leon Bergen, and Noah D. Goodman. "Formalizing the Pragmatics of Metaphor Understanding". In: *Proceedings of the 36th Annual Meeting of the Cognitive Science Society, CogSci 2014* (2014), pp. 719–724.

[16] Alexandra Mayn and Vera Demberg. "Pragmatics of Metaphor Revisited: Formalizing the Role of Typicality and Alternative Utterances in Metaphor Understanding." In: *Proceedings of the 44th Annual Meeting of the Cognitive Science Society: Cognitive Diversity, CogSci 2022*. 2022, pp. 3154–3160.

[17] Gaia Carenini et al. "Towards a Better Rational Speech Act Framework for Context-Aware Modeling of Metaphor Understanding". In: *Proceedings of the First Workshop on Theory of Mind in Communicating Agents at ICML* (2023).

[18] Jennifer Hu and Roger Levy. *Prompting is not a substitute for probability measurements in large language models*. 2023. arXiv: 2305.13264 [cs.CL].

[19] Carlos Roncero and Roberto De Almeida. "Semantic properties, aptness, familiarity, conventionality, and interpretive diversity scores for 84 metaphors and similes". In: *Behavior research methods* 47 (July 2014). DOI: 10.3758/s13428-014-0502-y.

[20] Adnan Ben Mansour, Gaia Carenini, and Alexandre Duplessis. *An intuitive logic for understanding autoregressive language models*. https://github.com/Adnan-Ben-Mansour/hackathon2022. 2022.