(RSA)^2: A Rhetorical-Strategy-Aware Rational Speech Act Framework for Figurative Language Understanding

[*]Cesare Spinoso-Di Piano[1,2] [*]David Austin[1,2]
Pablo Piantanida[2,3,4] Jackie Chi Kit Cheung[1,2,5]
[1]McGill University, [2]Mila - Quebec AI Institute
[3]International Laboratory on Learning Systems (ILLS)
[4]CNRS, CentraleSupélec - Université Paris-Saclay, [5]Canada CIFAR AI Chair, Mila
{cesare.spinoso,david.austin,pablo.piantanida,cheungja}@mila.quebec

Abstract

Figurative language (e.g., irony, hyperbole, understatement) is ubiquitous in human communication, resulting in utterances where the literal and the intended meanings do not match. The Rational Speech Act (RSA) framework, which explicitly models speaker intentions, is the most widespread theory of probabilistic pragmatics, but existing implementations are either unable to account for figurative expressions or require modeling the implicit motivations for using figurative language (e.g., to express joy or annoyance) in a setting-specific way. In this paper, we introduce the **R**hetorical-**S**trategy-**A**ware **RSA** **(RSA)**^{2} framework which models figurative language use by considering a speaker's employed rhetorical strategy. We show that **(RSA)**^{2} enables human-compatible interpretations of non-literal utterances without modeling a speaker's motivations for being non-literal. Combined with LLMs, it achieves state-of-the-art performance on the ironic split of PragMega+, a new irony interpretation dataset introduced in this study.[1]

1 Introduction

Figurative uses of language — where a speaker does not literally say what they mean — are ubiquitous in human communication. For example, when faced with a blizzard with heavy winds, a speaker may say "It's a little chilly out." to indicate that it is very cold outside. Similarly, as illustrated in Fig. 1, a teacher might refer to a student as "really sharp" when they intend to convey that they are in fact not very clever. As a result, it is imperative for language technologies and computational models of language to account for this linguistic phenomenon.

However, recent studies have shown that large language models (LLMs) struggle with figurative interpretations of language. For instance, Hu et al. (2023) show that LLMs often misinterpret utterances which subvert listener expectations such as in irony and humour. In Fig. 1, we show an example where the Mistral-7B-Instruct (V3) LLM overwhelmingly favours an incorrect literal interpretation of the utterance (The student is smart.) over the correct non-literal one (The student is not very clever.). Consequently, this behaviour suggests that LLMs do not correctly or fully model the underlying *communicative goal* (i.e., intention) of utterances needed to interpret them, to the extent that this can be shown through behavioural analysis.

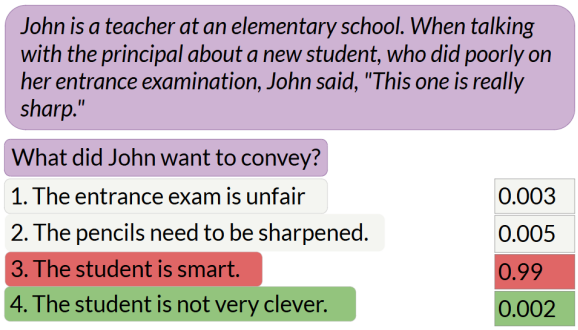

Figure 1: A sample from the PragMega dataset (Hu et al., 2023) showing probabilities of intended meanings for a given scenario. The probabilities are computed using the Mistral-7B-Instruct model (averaged across 10 different orders of presenting the meaning options).

A potential solution to the poor performance of LLMs on figurative understanding is the Rational Speech Act (RSA) framework from probabilistic pragmatics in which the modeling of communicative intents is central. In RSA, an utterance's meaning is interpreted probabilistically by a *pragmatic listener* which reasons about a posited *pragmatic speaker*'s likelihood of generating the utterance under different intended meanings. However, while RSA was designed to interpret utterances based on their communicative goal, its original formulation

[1]Code and data available at https://github.com/cesare-spinoso/rsa2.

[*]Equal contribution.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20898–20938

does not allow non-literal meaning interpretations of utterances. Previous attempts to address this limitation require modeling a speaker's motivations for being non-literal (e.g., to express their affect of annoyance or joy) in a setting-specific way (Kao et al., 2014b; Kao and Goodman, 2015).

The key insight that we leverage is that non-literal language usage follows systematic patterns, which can be grouped into *rhetorical strategies* (Bertin et al., 2016). For example, irony can be characterized by an intended meaning being the opposite of its literal meaning, whereas hyperbole involves overstating the intended meaning. Thus, we argue that pragmatic models should reify the rhetorical strategy as a mediating mechanism between language form and communicative intent. To this end, we introduce a novel formulation of the RSA framework: **R**hetorical-**S**trategy-**A**ware **RSA** **(RSA)**^{2}. In **(RSA)**^{2}, a speaker's rhetorical strategy is explicitly modeled as a latent variable and used by the pragmatic listener to interpret an utterance's potentially non-literal intended meaning.

We show experimentally that **(RSA)**^{2} enables pragmatic listeners to infer non-literal interpretations of utterances without needing to model a speaker's motivations for being non-literal. Using datasets of non-literal number expressions (Kao et al., 2014b) and of ironic weather utterances (Kao and Goodman, 2015), we show that **(RSA)**^{2} provides non-literal interpretations of utterances which closely match those of humans and which often outperform those made by existing *affect-aware RSA* methods. In addition, we couple **(RSA)**^{2} with LLMs and prompt engineering to achieve state-of-the-art performance on the ironic split of an expanded utterance interpretation dataset, PragMega+, which we design for this study. This latter result suggests that probabilistic pragmatics methods may help mitigate LLMs' bias towards literal interpretations and lexical overlap as exhibited in Fig. 1.

To summarize, in this work, we design **(RSA)**^{2}, a communicative-goal-centered procedure of figurative language interpretation. By explicitly considering the rhetorical strategies a speaker might use, this framework provides pragmatic non-literal interpretations of utterances which are aligned with human interpretations. In addition, we show that **(RSA)**^{2} can be used as a tool to help align LLM interpretations of figurative utterances with their intended meaning. Beyond figurative language understanding, our work re-establishes the importance of computational models of language which focus on modeling the communicative goals of speakers and listeners.

2 Related Work

Modeling Human Uses of Figurative Language There has been extensive work to explain why humans use figurative language. Early studies in pragmatics (Grice, 1975; Horn, 1984) attempt to explain such uses of language by positing that language users communicate with each other "cooperatively" by following certain "conversational maxims". In this account, figurative statements can result from reconciling violations of conversational maxims with the cooperative principle. Subsequent studies have shown that speakers use figurative expressions to soften the tone of a critique or to be humorous (Roberts and Kreuz, 1994; Colston, 1997), that figurative uses of language can sometimes be easier to process cognitively than their literal counterparts (Gibbs Jr., 1979) and that listeners understand figurative statements by explicitly interpreting the mental states of speakers through the "Theory of Mind"-network part of the brain (Spotorno et al., 2012).

Kao et al. (2014b,a); Kao and Goodman (2015) adapt the RSA framework originally introduced by Frank and Goodman (2012) to enable non-literal interpretations of figurative utterances (e.g., hyperbolic utterances about prices, ironic utterances about the weather). To enable non-literal interpretations, the authors assume that figurative language use is motivated by affect (e.g., to convey joy, annoyance, etc.) and perform joint inference over both intended meaning and affect to achieve non-literal interpretations. In this work, we seek to computationally model figurative language *without* having to explicitly model its motivation which may be difficult to determine in general. For instance, when John ironically says "This one is really sharp." (Fig. 1), he may be motivated by one (or more) of several reasons: expressing some kind of affect (e.g., humour, frustration), gauging the principal's attention, using a conventionalized way of referring to students, etc. In contrast, **(RSA)**^{2} lifts its dependence on explicitly modeling a speaker's motivation to use figurative language and instead directly accounts for the possible rhetorical strategies being used to produce non-literal interpretations of utterances.

Figurative Language Modeling in NLP In NLP, the ability of language systems to understand figurative language has mainly been evaluated through detection and generation (Li and Sporleder, 2010; Tsvetkov et al., 2014; Van Hee et al., 2018; Balestrucci et al., 2024; Lai and Nissim, 2022; Lai et al., 2023). A common approach to solving these tasks has been to collect figurative-language-specific datasets and to fine-tune language classification and generation models on them (Van Hee et al., 2018). With the shift towards LLMs, there has been growing interest in evaluating pre-trained and instruction-tuned LLMs on their ability to recognize, interpret and generate figurative language such as irony and sarcasm (Gu et al., 2022; Liu et al., 2022; Hu et al., 2023). LLMs have been shown to perform significantly worse than humans in interpreting figurative language, such as humour and irony. In this work, we propose a computational model of figurative language that explicitly models rhetorical strategies and which we show can be coupled with LLMs to better align their interpretations with speaker intentions.

RSA in NLP Andreas and Klein (2016) were the first to use neural listeners and speakers with RSA in the context of image captioning. Fried et al. (2018) extended the use of RSA to instruction following and generation. Shen et al. (2019) investigated RSA for abstractive summarization and Cohn-Gordon and Goodman (2019) applied it to machine translation. In closer connection to our work, Carenini et al. (2024) combined the RSA framework with LLMs for the task of metaphor understanding. Tsvilodub et al. (2025) leverage affect-aware RSA to enable LLMs to provide non-literal meaning interpretations of utterances with number words. In contrast, we develop a novel RSA framework, **(RSA)**^{2}, and apply it to several non-literal utterance interpretation tasks.

3 Rhetorical-Strategy-Aware RSA

In this section, we introduce our novel formulation of the RSA framework — **R**hetorical-**S**trategy-**A**ware **RSA (RSA)**^{2}— which explicitly incorporates the rhetorical strategy into the RSA framework. We first review the inner workings of the RSA framework and then present the novel **(RSA)**^{2} framework. Throughout this presentation, we will use the running example of a speaker ironically saying "The weather is amazing." during a blizzard, inspired by Kao and Goodman (2015).

3.1 The RSA Framework

The RSA framework was introduced to model human communication by accounting for listener and speaker expectations, originally in the context of scalar implicature (Frank and Goodman, 2012). The goal of this framework is to derive a probability distribution over some fixed (finite) set of meanings $\mathcal{M}$ for some observed utterance $u \in \mathcal{U}$ generated in some context[2] $c \in \mathcal{C}$. To do so, this framework posits the existence of multiple speakers and listeners who reason about each other in a recursive fashion. This recursive procedure begins with a *literal listener* L_0, which reasons literally about the interpretation of utterances. Its conditional distribution over $\mathcal{M}$ is defined as

$$P_{L_0}(m|c,u) \propto \mathbb{1}_{m \in [\![u]\!]} \cdot P_{M|C}(m|c), \quad (1)$$

where $P_{M|C}$ is a prior conditional distribution over all meanings in $\mathcal{M}$ given a context c; and $[\![\cdot]\!] : \mathcal{U} \rightarrow \mathcal{P}(\mathcal{M})$ denotes a shared semantic understanding function[3] which, given some utterance $u \in \mathcal{U}$, returns the set of possible intended meanings which are *literally* compatible with u.

For instance, given our running example "The weather is amazing." and a discrete set of weather states which a speaker might want to convey, $\mathcal{M} = \{\texttt{terrible}, \texttt{bad}, \texttt{ok}, \texttt{good}, \texttt{amazing}\}$, the semantic understanding function would return the following set:

$$[\![\text{"The weather is amazing."}]\!] = \{\texttt{amazing}\}.$$

With the base case of the reasoning procedure defined, the RSA framework provides a speaker-aware interpretation of the utterance u by positing the existence of a *pragmatic speaker*, S_1, which selects an utterance u based on its prior probability, $P_{U|C}$, and L_0's expected information gain.[4] That is,

$$P_{S_1}(u|c,m) \propto P_{L_0}(m|c,u)^{\alpha} P_{U|C}(u|c), \quad (2)$$

where α is a rationality parameter that controls how much the pragmatic speaker will favor selecting the most informative utterance according to the literal listener. Note that the pragmatic speaker is

[2]The context c can be linguistic (e.g., conversation history) as well as situational (e.g., the current state of the world).

[3]$\mathcal{P}()$ is the power set function.

[4]While Equation 2 may appear different from its common expected-utility formulation (Goodman and Stuhlmüller, 2013), we show the equivalence of these two formulations in Appendix A.1.

normalized across a set of utterances which are either predefined or generated based on c (Andreas and Klein, 2016).

According to Equation 2, a speaker intending to convey meaning m will distribute its probability mass proportional to the probability that L_0 infers m from u. In our running example, assuming a discrete set of utterances of the form $\mathcal{U} = \{\text{“The weather is } m\text{.”} : m \in \mathcal{M}\}$, a speaker wishing to convey $m = \texttt{amazing}$ will place *all* of its probability mass on the utterance “The weather is amazing.” since no other utterance enables L_0 to recover this intended meaning.

Finally, to interpret a given utterance u pragmatically, the *pragmatic listener*, L_1, updates their prior belief about the distribution of possible intended meanings $P_{M|C}$ by using the likelihood that the utterance was generated by S_1 in the given context. In other words, the pragmatic listener's conditional probability distribution over $\mathcal{M}$ becomes the following posterior Bayesian update:

$$P_{L_1}(m|c,u) \propto P_{S_1}(u|c,m)P_{M|C}(m|c). \quad (3)$$

In the case of our ironically uttered statement “The weather is amazing.”, the pragmatic listener will reason through S_1's generation process and will conclude that they intended to say that the weather is $\texttt{amazing}$ *despite* the blizzard. Thus, the pragmatic listener is unable to escape the literal interpretation of the utterance determined by L_0.

In general, it can be shown that, under the RSA framework, interpretations of utterances which are not compatible with the literal meaning of an utterance will be assigned zero probability by the pragmatic listener (Proof in Appendix A.2). Existing *affect-aware RSA* approaches discussed in Section 2 have attempted to mitigate this limitation by expanding the model of the speaker to include a random variable which accounts for their motivation to be figurative (e.g., for affect-related reasons). In contrast, our **(RSA)**^{2} approach does not require modeling a speaker's implicit motivations for being non-literal. Rather, the pragmatic listener in **(RSA)**^{2} reasons directly and explicitly about the possible rhetorical strategies being employed to achieve non-literal interpretations of utterances.

3.2 The Rhetorical-Strategy-Aware RSA Framework

We introduce a rhetorical-strategy-aware RSA framework, **(RSA)**^{2}, which defines a rhetorical strategy variable $r \in \mathcal{R}$ and a rhetorical function for each value of r, $f_r : \mathcal{C} \times \mathcal{M} \times \mathcal{U} \rightarrow [0,1]$. This function generalizes the literal semantic understanding indicator function $\mathbb{1}_{m \in [\![u]\!]}$ to an arbitrary function over $[0,1]$. The rhetorical function enables non-literal interpretations of utterances to arise based on the employed rhetorical strategy. For instance, using our running blizzard example, the ironic and hyperbolic rhetorical strategies might return the following, where $c = \texttt{blizzard}$ is the context variable:

$$f_{irony}(c, \texttt{terrible}, \text{“The weather is amazing.”}) = 1,$$
$$f_{irony}(c, \texttt{amazing}, \text{“The weather is amazing.”}) = 0,$$
$$f_{hyperbole}(c, \texttt{good}, \text{“The weather is amazing.”}) = 1,$$
$$f_{hyperbole}(c, \texttt{amazing}, \text{“The weather is amazing.”}) = 0.$$

We replace the semantic understanding function with our generalized rhetorical function within the P_{L_0} equation as follows:

$$P_{L_0}(m|c,u,r) \propto f_r(c,m,u)P_{M|C}(m|c). \quad (4)$$

Thus, a meaning which receives zero probability mass when an utterance is interpreted literally by L_0 may still receive non-zero probability mass when that same utterance is interpreted using irony, hyperbole or some other non-literal rhetorical strategy. For example, using f_{irony} and $f_{hyperbole}$ in our running example would enable the probability distribution of L_0 to shift towards non-literal interpretations of the utterance such as $\texttt{terrible}$ and $\texttt{good}$ respectively. Previous extensions of the RSA framework have modified the semantic understanding indicator function to account for lexical uncertainty (Bergen et al., 2016). To the best of our knowledge, we are the first to propose this generalization to model figurative language.

We define the pragmatic speaker and listener like in standard RSA, with our rhetorical strategy variable as an additional conditional factor:

$$P_{S_1}(u|m,c,r) \propto P_{L_0}(m|c,u,r)^{\alpha} P_{U|C}(u|c), \quad (5)$$

$$P_{L_1}(m|c,u,r) \propto P_{S_1}(u|m,c,r)P_{M|C}(m|c). \quad (6)$$

At inference time, the rhetorical strategy being used is an unobserved latent variable which must be marginalized out while accounting for its probability, $P_{R|CU}(r|c,u)$ on $\mathcal{R}$, as follows:

$$P_{L_1}(m|c,u) = \sum_{r'} P_{L_1}(m|c,u,r')P_{R|CU}(r'|c,u). \quad (7)$$

Returning to our running example, let us consider a discrete set of rhetorical strategies $\mathcal{R} = \{literal, irony, hyperbole\}$. We observe that the pragmatic listener conditioned with the *literal* strategy would produce the same distribution as the standard-RSA-derived pragmatic listener. However, the utterance would also be interpreted by $P_{L_1}(m|c, u, irony)$ with most probability mass concentrated around `terrible`, the opposite of the utterance's literal meaning, and by $P_{L_1}(m|c, u, hyperbole)$ with most mass concentrated around good, the utterance's scaled down literal meaning. By marginalizing these listener distributions using $P_{R|CU}$,[5] we would obtain a distribution which assigns mass to meanings beyond the utterance's literal interpretation. In this way, **(RSA)**^{2} allows modeling figurative interpretations of language *without* needing to explicitly model the motivations behind its use. In Appendix A.3, we prove that this latter approach to non-literal language is a special case of **(RSA)**^{2} wherein the rhetorical strategy function is repurposed to model the motivations of non-literal utterances (e.g., affect).

4 Non-Literal Interpretations of Figurative Language with (RSA)^2

In this first experimental study, we aim to show that **(RSA)**^{2} can be used to derive non-literal meaning interpretations which match human interpretations as well as by existing affect-aware RSA methods. We implement **(RSA)**^{2} in two settings: non-literal number price expressions (e.g., "This kettle costs 10000$.") and ironic weather utterances (e.g., "The weather is amazing." during a winter blizzard). We show that **(RSA)**^{2} produces utterance interpretations that are on par with or better than affect-aware RSA interpretations.

4.1 Datasets

Non-literal number expressions. We use the dataset from Kao et al. (2014b) which contains a collection of literal and non-literal number expressions related to the price of objects. The context space $\mathcal{C}$ is the set of objects being described, the meaning space $\mathcal{M}$ is the price of the object being described and the utterance space $\mathcal{U}$ is a verbalization of the object's price. These three sets are listed below:

$$\mathcal{C} = \{\text{electric kettle}, \text{laptop}, \text{watch}\},$$
$$\mathcal{M} = \{50, 51, 500, 501, 1000, 1001, 5000, 5001, 10000, 10001\},$$
$$\mathcal{U} = \{\text{"The } c \text{ costs } m \text{ dollars."} : c \in \mathcal{C}, m \in \mathcal{M}\}.$$

The authors collected meaning and affect priors $P(m|c), P(a|c)$ from 30 human participants as well as meaning and affect posteriors $P(m|c, u), P(a|c, u)$ from 120 human participants via a Likert-scale probability elicitation technique.

To enable **(RSA)**^{2}-based non-literal interpretations, we define the space of rhetorical strategies $\mathcal{R}$ as consisting of four types: *literal* (describing the price exactly), *hyperbole* (overstating the price), *understatement* (understating the price) and *halo* (providing a round figure rather than the exact one).

Ironic weather utterances. We use the dataset from Kao and Goodman (2015) which contains a collection of utterances about the weather similar to the one from our running example in Section 3. In this case, the context space $\mathcal{C}$ is represented visually through images depicting different weather conditions (see Appendix B.1.1 for the images), the utterance space $\mathcal{U}$ consists of statements about the weather (e.g., "The weather is amazing."), and the meaning space $\mathcal{U}$ corresponds to the true weather state being communicated by the speaker. These three sets are listed below:

$$\mathcal{C} = \{c_i : i \in [1 \dots 9]\},$$
$$\mathcal{M} = \{\texttt{terrible}, \texttt{bad}, \texttt{ok}, \texttt{good}, \texttt{amazing}\},$$
$$\mathcal{U} = \{\text{"The weather is } m\text{"} : m \in \mathcal{M}\}.$$

The context space contains three types of contexts: those where the weather is visibly good ($\{c_1, c_2, c_3\}$), those where the weather is visibly bad ($\{c_7, c_8, c_9\}$) and those where the weather is neither visibly good nor bad ($\{c_4, c_5, c_6\}$). The original authors of the paper collected meaning and affect priors $P(m|c), P(a|c)$ from 49 human participants as well as meaning, affect and irony posteriors $P(m|c, u), P(a|c, u), P(r|c, u)$ from 120 human participants using both normalized counts and a Likert-scale probability elicitation technique similar to Kao et al. (2014b).

To enable **(RSA)**^{2}-based non-literal interpretations, the space of rhetorical strategies $\mathcal{R}$ is defined using both the *literal* (describing the weather as it is perceived) and *irony* (describing the weather

[5] $P_{R|CU}$ may be uniform, human-derived or computed experimentally.

as the opposite of how it is perceived) rhetorical strategies.

4.2 RSA Model Experiments

4.2.1 Baseline

Affect-aware RSA. For the non-literal number expressions dataset, we use the pragmatic listener as computed by Kao et al. (2014b). For the weather utterance dataset, we re-implement both the affect-aware literal and pragmatic listeners (see Appendix B.1.2 for implementation details).

4.2.2 (RSA)^2 Implementations

Non-literal number expressions. We manually define f_r based on the intuitive definitions of each of the rhetorical strategies in $\mathcal{R}$. To simplify notation, we only consider the integer parts of the utterance i.e., $u =$ "The c costs y dollars." becomes $u = y$. We define f_r as:

$$f_r(c, m, u) = \begin{cases} 1 & \text{if } r = \text{literal}, u = m \\ 1 & \text{if } r = \text{hyperbole}, \\ & \quad u - m > 10 \\ 1 & \text{if } r = \text{understatement}, \\ & \quad m - u > 10 \\ 1 & \text{if } r = \text{halo}, |u - m| = 1, \\ & \quad u \text{ ends in } 0 \\ 0.001 & \text{otherwise} \end{cases}$$

Thus, for instance, if a kettle's price is $m = 50$, but the utterance used is $u = 500$, then this will trigger the *hyperbole* rhetorical strategy. Similarly, if the kettle's price is $m = 501$ and the utterance is $u = 500$, then this will activate the *halo* rhetorical strategy. We ignore the context c in f_r for simplicity, although in principle the effect of a rhetorical strategy like hyperbole could vary across objects. We substitute f_r in L_0's Equation 4 and use the meaning prior $P(m|c)$ derived from Kao et al. (2014b) and a uniform rhetorical strategy prior, $P(r|c, u) = \frac{1}{4}$, for the marginalization of Equation 7. We set the rationality parameter $\alpha = 1$.

Ironic weather utterances. To highlight the flexibility of (RSA)^2, we approximate the rhetorical function f_r across all 450 (c, m, u, r) quadruples[6] using a neural network. Specifically, we split the dataset into training, validation and test sets (60%/20%/20%) and train a 2-layer neural network ($16 \times 16 \times 5$) with sigmoid activations. We encode the contexts, meanings, utterances, and rhetorical strategies using a one-hot encoding and use a cross-entropy loss between $P_{L_1}(m|c, u)$ and the meaning selected by human participants given the context c and the utterance u. The rationality parameter is set to $\alpha = 1$. We use the human-elicited probability $P(r|c, u)$ for the marginalization across the two rhetorical strategies in Equation 7. Additional details can be found in Appendix B.1.3.

[6] 9 contexts × 5 meanings × 5 utterances × 2 rhetorical strategies

4.3 Results

We plot the meaning distribution of human posteriors and of affect-aware and (RSA)^2 listeners for the hyperbolic number expression "The electric kettle costs 1001 dollars." in Fig. 2 and for the ironic expression "The weather is amazing." uttered in the context of a blizzard in Fig. 3. In both cases, we observe that the (RSA)^2 listeners, especially the pragmatic listeners, induce human-like meaning distributions. Figures for all other context–meaning–utterance triples can be found in Appendices B.2 and B.3.

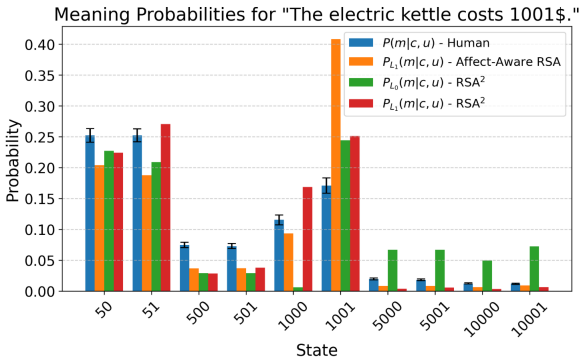

Figure 2: Meaning distribution of human posteriors and of affect-aware and (RSA)^2 listeners for the utterance "The electric kettle costs 1001 dollars".

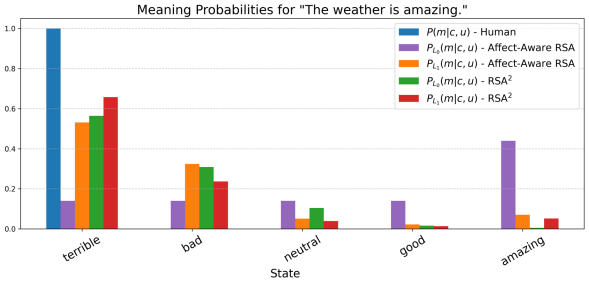

Figure 3: Meaning distribution of human posteriors and of affect-aware and (RSA)^2 listeners for the utterance "The weather is amazing." in the context of a blizzard (Image c_8 in Appendix B.1.1).

We evaluate our model's predictive power by

computing the mean absolute difference (MAD) between the human-derived meaning probability distributions and those generated by the listeners in both affect-aware RSA and **(RSA)**^{2} (Table 1). We note that **(RSA)**^{2} is competitive with affect-aware RSA, surpassing it on the ironic weather utterances dataset. While the affect-aware listener outperforms our **(RSA)**^{2} listeners on non-literal number expressions, we believe this is due to the poorly calibrated uniform rhetorical strategy posterior we use, which sometimes causes probability mass to be distributed to the tail ends of the meaning space. For instance, while the expression "The kettle costs 51 dollars." is most likely to be uttered with the *literal* rhetorical strategy, a significant portion of the meaning mass is distributed across the rest of meaning space due to the non-zero posterior probability of the *understatement* rhetorical strategy. Overall, we believe this initial study demonstrates that **(RSA)**^{2} can induce listener meaning distributions for figurative language that are at least as compatible with human interpretations as those produced by existing affect-aware RSA models.

Model	L_i	Non-literal Numbers ↓	Weather Utterances ↓
Affect-Aware RSA	L_0	-	0.2377
	L_1	**0.0436**	0.1278
(RSA)^2	L_0	0.0438	0.1647
	L_1	0.0467	**0.1229**

Table 1: Mean absolute differences between listener meaning distributions, $P_{L_i}(m|c,u), i = 0, 1$, and the human posterior, $P(m|c,u)$, for both affect-aware and **(RSA)**^{2} on both the non-literal number expressions and ironic weather utterances datasets.

5 LLM Irony Interpretation with (RSA)^2

In this section, we apply **(RSA)**^{2} to LLMs to generate non-literal interpretations of utterances that better align with human interpretations of figurative language. In particular, we focus on the task of interpreting ironic utterances in situational contexts beyond the weather examples from Section 4. Our results demonstrate that integrating LLMs into **(RSA)**^{2} can, to a certain extent, mitigate their biases toward lexical-overlap and literal interpretations.

5.1 The PragMega+ Dataset

We expand the PragMega dataset from Hu et al. (2023) to evaluate the **(RSA)**^{2} framework on ironic utterance interpretation. The original PragMega dataset contains 25 ironic scenarios where the intended meaning of an utterance is non-literal, such as in Fig. 1. Each scenario includes a background context c and an utterance u produced ironically by a speaker. In addition, each scenario is accompanied by four candidate intended meanings, which act as our meaning space$\mathcal{M}$: the literal meaning of u (`Literal Meaning`), the intended meaning of u (`Non-Literal Meaning`), a lexical overlap distractor meaning (`Overlap Meaning`), and a non-sequitur distractor meaning (`Non-Sequitur Meaning`).

Our expanded dataset, PragMega+, adds 25 ironic scenarios in the same format as part of our test set. In addition, we manually modify each of the 50 scenarios to create 50 *literal* scenarios where the intended meaning is the `Literal Meaning`. Examples of scenarios with `Non-Literal` and `Literal` intended meanings from PragMega+ can be found in Appendix C.1.

5.2 Experimental Setup

5.2.1 LLM Probability Estimation

To integrate LLMs within the **(RSA)**^{2} framework, we use an LLM N to estimate all conditional probabilities of the form $P_N(y|x)$. To do so, we use a prompt template similar to Hu et al. (2023) which, given x, lists all possible values of y in a multiple-choice question (MCQ) format. After passing the prompt to the LLM N, we extract the next-token logit of the number corresponding to each option. These logits are re-normalized so that the sum of their corresponding probabilities equals one. In addition, to avoid positional bias, we randomly shuffle the order of the options in the prompt 10 times[7] and average across shuffles to obtain $P_N(y|x)$. We use this approach to estimate the meaning prior $P_N(m|c)$, the rhetorical strategy posterior $P_N(r|c,u)$, the meaning posterior $P_N(m|c,u)$, and the meaning posterior *conditioned on* the rhetorical strategies, $P_N(m|c,u,r)$, where $\mathcal{R} = \{\textit{literal}, \textit{irony}\}$. We experiment with two instruction-tuned models: Mistral-7B-Instruct (V3) LLM (Jiang et al., 2023) and Llama-8B-Instruct (V3.1) AI@Meta (2024) via HuggingFace (Wolf et al., 2020). We performed prompt engineering on the validation set. All prompts used on the test set are presented in Appendix C.2.

[7] If the total number of permutations is less than 10, we use that total instead.

5.2.2 Alternative Utterance Generation

We use an LLM G conditioned on a scenario's context c to generate the alternative utterances needed to compute the pragmatic speaker's distribution. To do so, we condition G's next-token generation on the full PragMega+ scenarios up to the original utterance's first quotation mark (e.g., *... John said,"* in Fig. 1), and continue generation until another quotation mark is produced. We use a decoding temperature of 1.0 and generate 50 alternative utterances for each context. We also use the LLM's generation likelihood, $P_G(u \mid c)$, as the utterance prior in our model. We experiment with the base versions of Mistral-7B (V3) (Jiang et al., 2023) and Llama-8B (V3.1) (AI@Meta, 2024).

5.2.3 LLM Listeners

LLM RSA We implement an LLM RSA baseline by setting $P_{L_0}(m|c,u)$ equal to the raw LLM-derived probabilities, $P_N(m|c,u)$. This baseline follows prior neural listener work (Andreas and Klein, 2016; Monroe et al., 2017) which approximate $P_{L_0}(m|c,u)$ using trained neural networks. We set $\alpha = 1$.

LLM (RSA)^2 We implement our LLM (RSA)^2 listeners by setting $P_{L_0}(m|c,u,r)$ equal to the LLM-derived probabilities $P_N(m|c,u,r)$. For the rhetorical strategy posterior we use both the raw LLM-derived probabilities, $P_N(r|c,u)$, and an indicator posterior, $I(r|c,u) = \mathbb{1}_{r=\arg\max_{r'} P_N(r'|c,u)}$. We set $\alpha = 1$.

Where is $f_r(c,m,u)$ in LLM (RSA)^2? Based on the LLM (RSA)^2 literal listener, the rhetorical strategy function is implicitly set as:

$$f_r(c,m,u) = \frac{P_N(m|c,u,r)}{k \cdot P_N(m|c)}, \tag{8}$$

where $k = \max_{m'} f_r(c,m',u)$ ensures that $f_r(c,m,u) \in [0,1]$.

Why is affect-aware RSA not included as a baseline? While it is in principle possible to implement an LLM-based Affect-Aware RSA baseline (Tsvilodub et al., 2025), we do not pursue this direction, as it would require defining a suitable affect space for each situational context.

5.3 Results

Table 2 presents the average listener meaning probabilities for the correct, incorrect and distractor meanings (aggregated) across all 50 scenarios in the test set. We use the LLM pair G = Llama-8B (V3.1) and N = Mistral-7B-Instruct (V3), since this pair performed best on the validation set. Listener probabilities conditioned on each rhetorical strategy are shown in Table 6. Overall, we see that the LLM (RSA)^2 listeners outperform the baseline LLM RSA listeners. In particular, the LLM (RSA)^2 literal listener marginalized using the indicator rhetorical strategy posterior, $I(r|c,u)$, performs the best while the baseline LLM RSA literal listener performs the worst.

Model	L_i	Correct ↑	Incorrect ↓	Distractor ↓
LLM RSA	L_0	0.73	0.24	0.02
	L_1	0.76	0.22	0.01
LLM (RSA)^2 with $P_N(r|c,u)$	L_0	0.74	0.23	0.02
	L_1	0.80	0.16	0.02
LLM (RSA)^2 with $I(r|c,u)$	L_0	**0.85**	**0.13**	0.01
	L_1	0.84	0.13	**0.01**

Table 2: Average listener probabilities, $P_{L_i}(m|c,u), i = 0,1$ for the correct, incorrect and distractor intended meanings on the test set averaged across all 50 scenarios.

We study these results more carefully by plotting the listener probability distributions split between ironic and literal scenarios in Fig. 4. On ironic scenarios, the LLM (RSA)^2 listener probabilities for the correct non-literal interpretation are all above 0.8. In contrast, the LLM RSA literal listener assigns a probability of less than 0.5 (0.48) to the correct non-literal intended meaning. However, on literal scenarios, the trend is reversed. The LLM RSA literal listener assigns most of its probability mass (0.95) to the correct literal meaning while the best performing of the LLM (RSA)^2 listeners – L_0 marginalized with $I(r|c,u)$ – assigns 0.77 of its mass to the correct literal intended meaning.

While the strong performance of the LLM RSA listeners on the literal scenarios is expected – the correct intended meaning is consistent with the utterance and the context – the drop in LLM (RSA)^2 performance in these cases is surprising. Further analysis reveals that this result stems from the listener marginalization, specifically the rhetorical strategy posterior $P_N(r \mid c,u)$. This posterior performs poorly in the literal scenarios where the utterance is intended to be interpreted literally (Table 6). While for ironic scenarios the average probability of $P_N(r = \textit{irony}|c,u) = 0.88$ is rea-

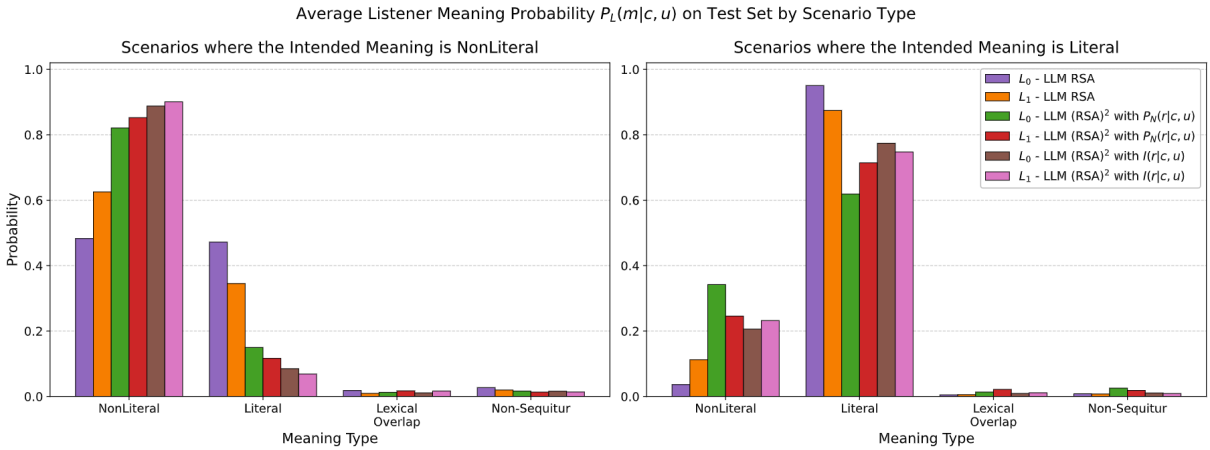

Figure 4: Average listener probabilities on the test set, split between the first 25 scenarios where the intended meaning is non-literal and the last 25 scenarios where the intended meaning is literal.

sonable, it is far worse on the sincere scenarios with $P_N(r = \textit{literal}|c, u) = 0.55$. This asymmetry in rhetorical strategy posteriors also explains why using $I(r|c, u)$ in the marginalization helps the LLM **(RSA)**^{2} listeners. The indicator function blocks more of the probability mass associated with the incorrect meaning from being incorporated into the marginalization.

5.4 Probability Ablation Study

To study the impact of the RSA reasoning process on LLMs, we run an ablation study in which we set the probabilities $P(m \mid c)$ and $P(u \mid c)$ to uniform in both LLM RSA and LLM **(RSA)**^{2}. Our analysis reveals that ablating $P(u|c)$ does not hinder performance and may slightly improve it. Surprisingly, ablating $P(m|c)$ reveals that most of the performance gain obtained by L_1 comes from the meaning prior *and not* from the RSA reasoning process itself. We attribute this result to the way in which we generate alternative utterances (Section 5.2.2) which tend to produce literal paraphrases of the intended meaning. For instance, Llama-8B (V3.1) generates alternatives to "This one is really sharp." from Fig. 1 such as "Her scores were well below average." Thus, the pragmatic speaker conditioned on the correct intended meaning distributes probability mass across many compatible alternative utterances, leading to a Bayesian update that *spreads*, rather than *narrows*, the distribution over meanings. Future work on coupling LLMs with RSA may want to ensure diversity in the alternatives generated, for instance by conditioning the generation on individual meanings rather than the context.

Overall, despite RSA reasoning offering limited improvements, the results obtained with L_0 still suggest that explicitly modeling rhetorical strategies may help mitigate lexical-overlap and literal interpretation biases in LLMs.

Model	L_i	w/o $P(m\|c)$	w/o $P(u\|c)$
LLM RSA	L_1	0.44 (-42.7%)	0.78 (+1.8%)
LLM **(RSA)**^{2} with $P_N(r\|c,u)$	L_1	0.44 (-44.8%)	0.80 (+0.3%)
LLM **(RSA)**^{2} with $I(r\|c,u)$	L_1	0.51 (-39.4%)	0.84 (+0.2%)

Table 3: Pragmatic listener probability ablations. We report the average listener posterior probabilities of the correct meaning on the test set (across all 50 scenarios) and the relative change with respect to the unablated model.

6 Conclusion

In conclusion, we presented **(RSA)**^{2}, a computational model of figurative language based on the RSA framework that explicitly incorporates rhetorical strategy to support non-literal interpretation. We show that **(RSA)**^{2} enables human-compatible interpretations of figurative utterances—including non-literal number expressions and ironic weather statements—*without* explicitly modeling the speaker's motivations behind using figurative language. We further demonstrated that combining **(RSA)**^{2} with LLMs yields state-of-the-art performance on the ironic split of the PragMega+ dataset. We hope this work inspires future research to incorporate pragmatic reasoning techniques within language technologies.

Limitations

Limited dataset size and diversity. While our experimental and theoretical results (Appendix A.3) support the usefulness and applicability of **(RSA)**^{2}, we believe that future work in computational pragmatics should include broader examples across languages and cultural contexts. Work such as Park et al. (2024); Sravanthi et al. (2024) has begun to address these issues. Regarding **(RSA)**^{2} specifically, while we have demonstrated its applicability to irony, hyperbole, understatement, and pragmatic halo, evaluating our framework on additional pragmatic phenomena such as metaphor and politeness are promising directions for future research.

Alternative utterance generation. As discussed in Section 5.4, pragmatic reasoning does not yield the typical probability mass narrowing associated with RSA due to the distribution of mass across multiple alternative utterances that effectively "mean" the same thing within this framework. We believe that this poses an interesting research problem for the field of computational pragmatics in particular (e.g. How do we generate relevant and non-overlapping alternative utterances?), and for natural language processing (e.g., LLM decoding) more broadly.

What if the rhetorical strategies are not known in advance? While **(RSA)**^{2} addresses the challenge of explicitly modeling affect in affect-aware RSA, it introduces a new question: how to determine which rhetorical strategies are appropriate in a given context? We take a first step toward this in Appendix D, where we propose and test a clustering-based algorithm that automatically induces *rhetorical strategy clusters*. While our algorithm underperforms LLM **(RSA)**^{2}, we believe that developing a more principled approach to inducing rhetorical strategies is an important direction for future work.

Ethics Statement

We propose **(RSA)**^{2} as a computational model of figurative language that aims to better capture human interpretations of non-literal utterances and to help LLMs interpret such utterances in a way that aligns more closely with their intended meaning. We recognize, however, that figurative language differs significantly across languages and cultures. Our model has not yet been validated on languages beyond English, or on figurative phenomena that may be less common in English. We encourage future work—similar to Park et al. (2024)—to explore figurative language in a range of linguistic and cultural contexts beyond those commonly associated with English.

Acknowledgments

The authors would like to thank the reviewers for their valuable comments. We would also like to thank Gaurav Kamath, Arkil Patel and Siva Reddy for their insightful comments on an earlier version of this paper. We are also extremely grateful to Justine Kao and Polina Tsvilodub for sharing the non-literal number expressions and ironic weather utterances datasets and for their helpful feedback. This work was supported by the Fonds de Recherche du Québec – Nature et Technologies (FRQNT) and the Natural Sciences and Engineering Research Council of Canada (NSERC). Jackie Chi Kit Cheung is supported by Canada CIFAR AI Chair program. We acknowledge material support from NVIDIA Corporation in the form of computational resources provided to Mila.

References

AI@Meta. 2024. Llama 3 model card.

Jacob Andreas and Dan Klein. 2016. Reasoning about Pragmatics with Neural Listeners and Speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Pier Felice Balestrucci, Silvia Casola, Soda Marem Lo, Valerio Basile, and Alessandro Mazzei. 2024. I'm sure you're a real scholar yourself: Exploring ironic content generation by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14480–14494, Miami, Florida, USA. Association for Computational Linguistics.

Leon Bergen, Roger Levy, and Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9:20–1.

Marc Bertin, Iana Atanassova, Cassidy R Sugimoto, and Vincent Lariviere. 2016. The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. *Scientometrics*, 109:1417–1434.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort,

Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122.

Gaia Carenini, Luca Bischetti, Walter Schaeken, and Valentina Bambini. 2024. Towards a Fully Interpretable and More Scalable RSA Model for Metaphor Understanding. Publisher: arXiv Version Number: 1.

Reuben Cohn-Gordon and Noah Goodman. 2019. Lost in Machine Translation: A Method to Reduce Meaning Loss. ArXiv:1902.09514 [cs].

Herbert L. Colston. 1997. Salting a wound or sugaring a pill: The pragmatic functions of ironic criticism. *Discourse Processes*, 23(1):25–45. Publisher: Routledge _eprint: https://doi.org/10.1080/01638539709544980.

Michael C. Frank and Noah D. Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science*, 336(6084):998–998. Publisher: American Association for the Advancement of Science.

Daniel Fried, Jacob Andreas, and Dan Klein. 2018. Unified Pragmatic Models for Generating and Following Instructions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1951–1963, New Orleans, Louisiana. Association for Computational Linguistics.

Raymond W. Gibbs Jr. 1979. Contextual effects in understanding indirect requests. *Discourse Processes*, 2(1):1–10. Publisher: Routledge _eprint: https://doi.org/10.1080/01638537909544450.

Noah D Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

HP Grice. 1975. Logic and conversation. *Syntax and Semantics*, 3:43–58.

Yuling Gu, Yao Fu, Valentina Pyatkin, Ian Magnusson, Bhavana Dalvi Mishra, and Peter Clark. 2022. Just-DREAM-about-it: Figurative language understanding with DREAM-FLUTE. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 84–93, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Laurence Horn. 1984. Towards a new taxonomy for pragmatic inference: Q-and r-based implicature. *Meaning, form and use in context*.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

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

Justine Kao, Leon Bergen, and Noah Goodman. 2014a. Formalizing the pragmatics of metaphor understanding. In *Proceedings of the annual meeting of the Cognitive Science Society*, volume 36.

Justine T. Kao and Noah D. Goodman. 2015. Let's talk (ironically) about the weather: Modeling verbal irony. *Cognitive Science*.

Justine T. Kao, Jean Y. Wu, Leon Bergen, and Noah D. Goodman. 2014b. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, 111(33):12002–12007.

Huiyuan Lai and Malvina Nissim. 2022. Multi-figurative language generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5939–5954, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267, Toronto, Canada. Association for Computational Linguistics.

Linlin Li and Caroline Sporleder. 2010. Using Gaussian mixture models to detect figurative language in context. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 297–300, Los Angeles, California. Association for Computational Linguistics.

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

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Dojun Park, Jiwoo Lee, Seohyun Park, Hyeyun Jeong, Youngeun Koo, Soonha Hwang, Seonwoo Park, and Sungeun Lee. 2024. MultiPragEval: Multilingual pragmatic evaluation of large language models. In *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 96–119, Miami, Florida, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Richard M. Roberts and Roger J. Kreuz. 1994. Why Do People Use Figurative Language? *Psychological Science*, 5(3):159–163. Publisher: [Association for Psychological Science, Sage Publications, Inc.].

Gregory Scontras, Michael Henry Tessler, and Michael Franke. 2025. Probabilistic language understanding: An introduction to the rational speech act framework. https://www.problang.org.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. 2019. Pragmatically Informative Text Generation. ArXiv:1904.01301 [cs].

Nicola Spotorno, Eric Koun, Jérôme Prado, Jean-Baptiste Van Der Henst, and Ira A. Noveck. 2012. Neural evidence that utterance-processing entails mentalizing: The case of irony. *NeuroImage*, 63(1):25–39.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12075–12097, Bangkok, Thailand. Association for Computational Linguistics.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Polina Tsvilodub, Kanishk Gandhi, Haoran Zhao, Jan-Philipp Fränken, Michael Franke, and Noah D. Goodman. 2025. Non-literal understanding of number words by language models.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

A Additional Theoretical Results and Discussion

A.1 Equivalence of RSA Formulations

We demonstrate that the pragmatic speaker formula introduced in Section 3.1 is equivalent to the formulation originally presented in Frank and Goodman (2012), which is given by

$$P_{S_1}(u|m,c) \propto \exp(\alpha \cdot U(m,u,c)), \tag{9}$$

where $U(m,u,c)$ is the utility function defined via information theory as

$$U(m,u,c) = \log P_{L_0}(m|c,u) - \kappa(u), \tag{10}$$

where $\kappa : \mathcal{U} \to \mathbb{R}_{\geq 0}$ is some cost function.

Replacing $U(m,u,c)$ in Equation 9, this is equivalent to

$$P_{S_1}(u|m,c) \propto \exp(-\alpha \cdot \kappa(u)) \cdot P_{L_0}(m|c,u)^{\alpha}. \tag{11}$$

Firstly, we show that there exists a particular setting of $\kappa(u)$ in Equation 11 for which we can recover Equation 2 presented in Section 3.1:

$$\begin{aligned} P_{S_1}(u|m,c) &\propto \exp(-\alpha \cdot \kappa(u)) \cdot P_{L_0}(m|c,u)^{\alpha} \\ &= \exp(\log(P(u|c)) \cdot P_{L_0}(m|c,u)^{\alpha} && \text{Setting } \kappa(u) = \frac{\log P(u|c)}{-\alpha} \\ &= P_{L_0}(m|c,u)^{\alpha} \cdot P(u|c) \end{aligned}$$

Secondly, we show that there exists a particular setting of $P(u|c)$ for which we can recover Equation 11:

$$\begin{aligned} P_{S_1}(u|m,c) &\propto P_{L_0}(m|c,u)^{\alpha} \cdot P(u|c) \\ &= P_{L_0}(m|c,u)^{\alpha} \cdot \texttt{softmax}(-\alpha \cdot \kappa(u)) && \text{Setting } P(u|c) = \texttt{softmax}(-\alpha \cdot \kappa(u)) \\ &\propto P_{L_0}(m|c,u)^{\alpha} \cdot \exp(-\alpha \cdot \kappa(u)) \end{aligned}$$

This proves the equivalence.

A.2 Standard RSA Provides Zero Probability Mass for Non-Literal Interpretations

We show that in the standard RSA framework, if $\alpha > 0$, then for all $m \in \mathcal{M}$ and $u \in \mathcal{U}$, the condition $m \notin [\![u]\!]$ implies that $P_{L_1}(m|u) = 0$.

Let $\alpha > 0$ and consider $m \in \mathcal{M}$ and $u \in \mathcal{U}$ such that $m \notin [\![u]\!]$. Then, we have:

$$\begin{aligned} P_{L_0}(m|c,u) &\propto \mathbb{1}_{m \in [\![u]\!]} \cdot P(m|c) \\ &= 0 \cdot P(m|c) \\ &= 0, \\ P_{S_1}(u|c,m) &\propto P_{L_0}(m|c,u)^{\alpha} \cdot P(u|c) \\ &= 0 \cdot P(u|c) \\ &= 0, \\ P_{L_1}(m|c,u) &\propto P_{S_1}(u|c,m) P(m|c) \\ &= 0 \cdot P(m|c) \\ &= 0. \end{aligned}$$

This proves the above statement.

A.3 QUD-RSA is a Special Case of (RSA)^2

We show that affect-aware RSA is a special case of **(RSA)**^{2}. To do so, we use the more general formulation of affect-aware RSA, Question Under Discussion RSA (QUD-RSA) (Scontras et al., 2025).

In QUD-RSA, the meaning space $\mathcal{M}$ is projected to meaning subspaces $\mathcal{X}$ using projection functions $q : \mathcal{M} \to \mathcal{X}$. Typically, this is done when the meaning space has been augmented to a vector space in order to include additional meaning dimensions. For instance, in affect-aware RSA, the meaning vector space $\mathcal{M} = \mathcal{S} \times \mathcal{A}$ is broken into the state of the world being conveyed, $\mathcal{S}$, and the affect $\mathcal{A}$. In this setting, an affect projection $q_{\texttt{affect}} : \mathcal{S} \times \mathcal{A} \to \mathcal{A}$ might look like the following:

$$q_{\texttt{affect}}(s, a) = a.$$

We first define the equations of QUD-RSA and of **(RSA)**^{2} and then show that any instance of the QUD-RSA equations can be re-written with the **(RSA)**^{2} equations, but not vice-versa.

A.3.1 Definitions

Definition 1 (QUD-RSA). Given a meaning space $\mathcal{M}$, an utterance space $\mathcal{U}$, a context space $\mathcal{C}$ and a set of n projection functions $Q = \{q_i : i \in [1 \dots n] \;\; \& \;\; q_i \text{ is a projection function from } \mathcal{M} \text{ to some subspace } \mathcal{X}\}$, the QUD-RSA equations $\mathcal{E}_{\texttt{QUD-RSA}}$, are as follows:[8]

$$P_{L_0}(m \mid u, c) \propto \mathbb{1}_{m \in [\![u]\!]} \cdot P(m \mid c), \tag{12}$$

$$P_{L_0}(m \mid c, u, q) \propto \sum_{m' \in \mathcal{M}} \mathbb{1}_{q(m')=q(m)} \cdot P_{L_0}(m' \mid u, c), \tag{13}$$

$$P_{S_1}(u \mid c, m, q) \propto P_{L_0}(m \mid c, u, q) \cdot P(u \mid c), \tag{14}$$

$$P_{L_1}(m \mid c, u, q) \propto P_{S_1}(u \mid c, m, q) \cdot P(m \mid c). \tag{15}$$

Definition 2 ((RSA)^2). Given a meaning space $\mathcal{M}$, an utterance space $\mathcal{U}$, a context space $\mathcal{C}$, a rhetorical strategy space $\mathcal{R}$ and a set of rhetorical strategy functions $F_r = \{f_r : r \in \mathcal{R} \;\; \& \;\; f_r : \mathcal{M} \times \mathcal{C} \times \mathcal{U} \to [0, 1]\}$, the **(RSA)**^{2} equations, $\mathcal{E}_{\textbf{(RSA)}^2}$ are as follows:

$$P_{L_0}(m \mid c, u, r) \propto f_r(c, m, u) \cdot P(m \mid c), \tag{16}$$

$$P_{S_1}(u \mid c, m, r) \propto P_{L_0}(m \mid c, u, r) \cdot P(u \mid c), \tag{17}$$

$$P_{L_1}(m \mid c, u, r) \propto P_{S_1}(u \mid c, m, r) \cdot P(m \mid c). \tag{18}$$

A.3.2 QUD-RSA Can Be Simulated by (RSA)^2

Lemma 1. If $P(m \mid c) > 0$ for all $m \in \mathcal{M}, c \in \mathcal{C}$, any instance of the QUD-RSA equations can be represented as an instance of the **(RSA)**^{2} equations.

Proof. Given an instance of the QUD-RSA equations, $\mathcal{E}_{\texttt{QUD-RSA}}$, we can build an equivalent instance of the **(RSA)**^{2} equations, $\mathcal{E}_{\textbf{(RSA)}^2}$.

We do this by setting the spaces $\mathcal{C}, \mathcal{M} \;\; \& \;\; \mathcal{U}$ in $\mathcal{E}_{\textbf{(RSA)}^2}$ to be the same as the ones from $\mathcal{E}_{\texttt{QUD-RSA}}$. In addition, we define a rhetorical strategy variable r and a corresponding rhetorical function f_r for each projection variable and corresponding projection function q of $\mathcal{E}_{\texttt{QUD-RSA}}$ as follows:

$$f_r(m, c, u) = \frac{\sum_{m' \in \mathcal{M}} \mathbb{1}_{q(m')=q(m)} \cdot P_{L_0}(m' \mid c, u)}{k \cdot P(m \mid c)}, \tag{19}$$

where the division by $k = \max_{m'} f_r(c, m', u)$ is needed such that $f_r(c, m, u) \in [0, 1]$.

Replacing f_r in $\mathcal{E}_{\textbf{(RSA)}^2}$'s Equation 16 enables us to recover the instance of the QUD-RSA equations, $\mathcal{E}_{\texttt{QUD-RSA}}$.

[8] We overload notation here and use q to represent both the latent variable which indexes its corresponding projection *as well as* the projection function itself.

A.3.3 Not All Instances of (RSA)^2 Can Be Simulated by QUD-RSA

Lemma 2. Given context space $\mathcal{C}$, meaning space $\mathcal{M}$, and utterance space $\mathcal{U}$, there exists an instance of the **(RSA)**^{2} equations which cannot be simulated by any instance of the QUD-RSA equations.

Proof Idea. The crux of this proof lies in the observation that the QUD-RSA equations can only induce literal listener distributions which are *binary combinations* of meaning priors $P(m|c)$. Thus, one can pick some probability vector which is not such a combination and compute it using the rhetorical strategy function f_r.

Proof. We first demonstrate that the literal listener of any instance of the QUD-RSA equations is a binary combination of meaning priors.

Consider the normalized literal listener distribution $P_{L_0}(m \mid c, u)$ from the QUD-RSA equations:

$$P_{L_0}(m \mid c, u) = \frac{\mathbb{1}_{m \in [\![u]\!]} \cdot P(m \mid c)}{\sum_{m''} \mathbb{1}_{m'' \in [\![u]\!]} \cdot P(m'' \mid c)}. \tag{20}$$

Replacing it in Equation 13, we get the following:

$$P_{L_0}(m \mid c, u, q) \propto \sum_{m'} \mathbb{1}_{q(m')=q(m)} \cdot \Big(\frac{\mathbb{1}_{m \in [\![u]\!]} \cdot P(m \mid c)}{\sum_{m''} \mathbb{1}_{m'' \in [\![u]\!]} \cdot P(m'' \mid c)} \Big) \tag{21}$$

$$\propto \sum_{m'} \mathbb{1}_{q(m')=q(m)} \cdot \mathbb{1}_{m \in [\![u]\!]} \cdot P(m \mid c) \tag{22}$$

$$\propto \sum_{m'} \mathbb{1}_{q(m')=q(m) \wedge m \in [\![u]\!]} \cdot P(m \mid c) \tag{23}$$

Thus, the QUD-RSA literal listeners are probability distributions induced by taking binary combinations of the meaning priors, $P(m|c)$. Since the meaning space $\mathcal{M}$ is finite, this implies that there is a finite set of distributions which the QUD-RSA literal listener can be set to. We define this set as follows:

$$\mathcal{P}_{\mathcal{L}0} = \{p_{l_0} \in \mathbb{R}^{|\mathcal{M}|} : p_{l_0} \text{ is a QUD-RSA literal listener probability vector}\}. \tag{24}$$

Then, to complete this proof, we need only define a function f_r such that $P_{L_0}(m \mid c, u, r) \notin \mathcal{P}_{\mathcal{L}0}$.

We create an f_r which induces $P_{L_0}(m \mid c, u, r)$ outside of $\mathcal{P}_{\mathcal{L}0}$ by picking a real number k such that it lies between 0 and the minimum non-zero probability value in $\mathcal{P}_{\mathcal{L}0}$. That is,

$$k = \frac{p_{\text{min}}}{2} \text{ such that } p_{\text{min}} = \min_{\vec{p} \in \mathcal{P}_{\mathcal{L}0},\, i \in 1 \ldots |\mathcal{M}|,\, \vec{p}[i] > 0} \vec{p}[i] \tag{25}$$

We define f_r such that $P_{L_0}(m \mid c, u, r) = k$ for some $m \in \mathcal{M}$. Let $m_1, m_2 \in \mathcal{M}, m_1 \neq m_2$ and f_r be defined as follows:

$$f_r(m, c, u) = \begin{cases} \frac{k}{P(m_1|c)} & \text{if } m = m_1 \\ \frac{1-k}{P(m_2|c)} & \text{if } m = m_2 \\ 0 & \text{otherwise} \end{cases}$$

With this function f_r[9], we can show that $P_{L_0}(m_1 \mid c, u, r) = k$ which we have just shown is not a probability value in any of the probability vectors found in $\mathcal{P}_{\mathcal{L}0}$. This can be seen as follows:

$$P_{L_0}(m_1 \mid c, u, r) = \frac{f_r(m_1, c, u) \cdot P(m_1 \mid c)}{\sum_{m'} f_r(m', c, u) \cdot P(m' \mid c)}, \tag{26}$$

$$= \frac{\frac{k}{P(m_1|c)} \cdot P(m_1 \mid c)}{\frac{k}{P(m_1|c)} \cdot P(m_1 \mid c) + \frac{1-k}{P(m_2|c)} \cdot P(m_2 \mid c)}, \tag{27}$$

$$= \frac{k}{k + 1 - k} = k \quad \text{as desired.} \tag{28}$$

This shows that there exists an instance of the **(RSA)**^{2} equations which cannot be represented with QUD-RSA.

[9] f_r should be divided by its maximum value to ensure it respects its $[0, 1]$ co-domain constraint. We omit this step for compactness.

B Deriving Non-Literal Interpretations of Figurative Language with (RSA)^2

We provide additional details regarding the ironic weather utterances experiment along with additional figures for both our non-literal number expression and ironic weather utterances experiments.

B.1 Ironic Weather Utterances Experimental Details

B.1.1 Weather Contexts Images

We include the images associated with each of the 9 weather contexts in the ironic weather utterances dataset. These images were shown to human participants to elicit both prior and posterior probabilities. These can also be found in the original work by Kao and Goodman (2015).

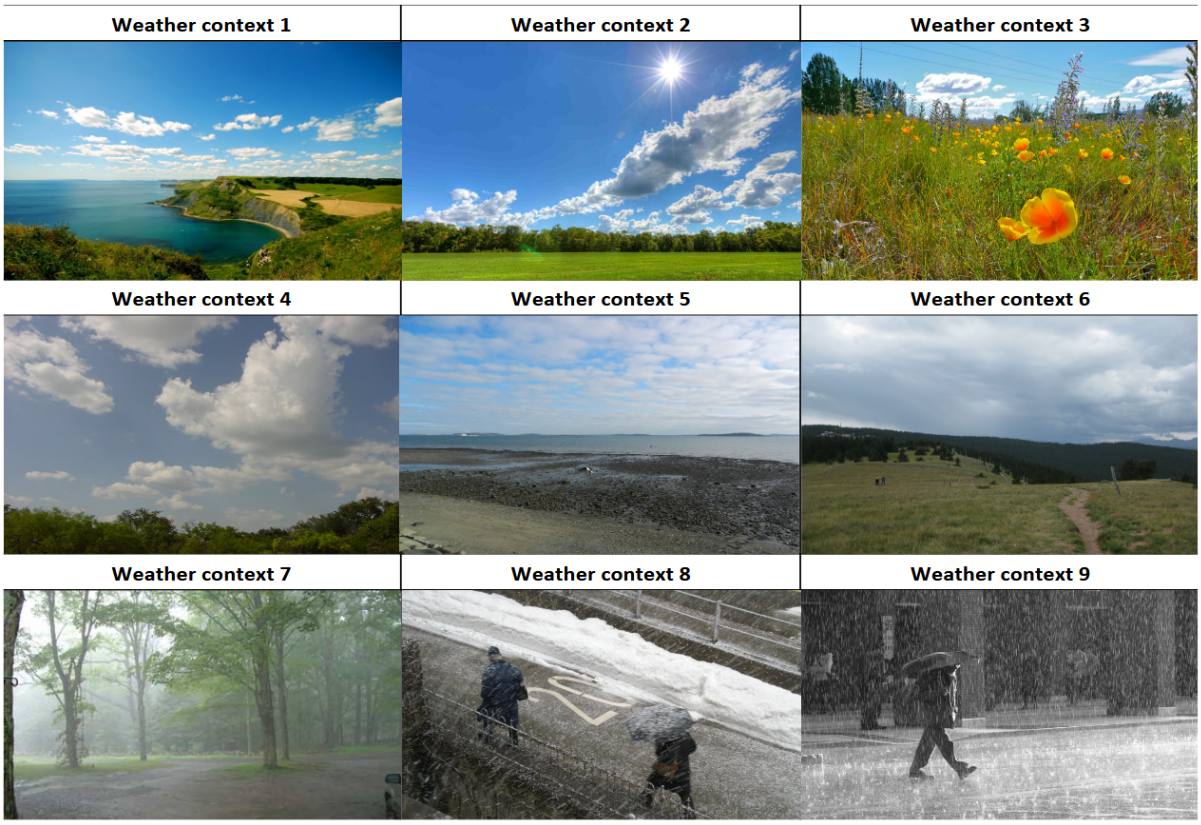

Figure 5: Images associated with the 9 weather contexts in the ironic weather utterances dataset.

B.1.2 Re-Implementation Details of Affect-Aware RSA

Using the ironic weather utterances dataset from Kao and Goodman (2015), we re-implement their affect-aware RSA model using the following equations where $s \in \mathcal{S}$ represents the conveyed state of the world (which we typically call m in our notation), $a, v \in \mathcal{A}, \mathcal{V}$ represent the arousal and valence dimensions of affect respectively and where the three QUD projections used are $q_{\texttt{literal}}(s,a,v) = s, q_{\texttt{arousal}}(s,a,v) = a, q_{\texttt{valence}}(s,a,v) = v$:

$$P(s,a,v|c) = P(s|c) \cdot P(a|c) \cdot P(v|c) \tag{29}$$

$$P_{L_0}(s,a,v|c,u) \propto P(s,a,v|c) \cdot \mathbb{1}_{s \in [\![u]\!]}, \tag{30}$$

$$P_{L_0}(s,a,v|c,u,q) \propto \sum_{(s',a',v') \in \mathcal{S} \times \mathcal{A} \times \mathcal{V}} \mathbb{1}_{q(s',a',v')=q(s,a,v)} \cdot P_{L_0}(s',a',v'|c,u), \tag{31}$$

$$P_{S_1}(u|c,s,a,v,q) \propto P_{L_0}(s,a,v|c,u,q)^{\alpha} \tag{32}$$

$$L_1(s,a,v|c,u,q) = P(s,a,v|c) \cdot P_{S_1}(u|c,s,a,v,q) \tag{33}$$

The QUD variables can be marginalized out from either listener with the following equations:

$$P_{L_0}(s, a, v|c, u) = \sum_{q \in Q} P_{L_0}(s, a, v|c, u, q) \cdot P(q|c), \tag{34}$$

$$P_{L_1}(s, a, v|c, u) \propto \sum_{q \in Q} L_1(s, a, v|c, u, q) \cdot P(q|c), \tag{35}$$

where, as described in Kao and Goodman (2015), we use $P(q_{\texttt{literal}}) = 0.3, P(q_{\texttt{arousal}}) = 0.4, P(q_{\texttt{valence}}) = 0.3$.

Finally, we can marginalize out the affect variables a and v from the $P_{L_i}(s, a, v|c, u)$ to get $P_{L_i}(s|c, u)$:

$$P_{L_i}(s|c, u) = \sum_{a \in \mathcal{A}} \sum_{v \in \mathcal{V}} P_{L_i}(s, a, v|c, u) \tag{36}$$

To verify the correctness of our re-implementation, we reproduce Fig. 5 from the Kao and Goodman (2015) paper and verify qualitatively that each context-utterance-meaning triple matches. Our reproduction of this figure can be found in Fig. 6.

B.1.3 (RSA)^2 Training Details

To learn the rhetorical strategy function f_r, we trained a neural network with an architecture of $16 \times 16 \times 5$, employing sigmoid activation functions throughout. The input to the network was a 16×1 one-hot encoding vector, where the first 9 entries were reserved for the context indicator (c_1 to c_9), the next 5 entries were reserved for the utterance indicator, and the final 2 entries were reserved for the rhetorical strategy indicator. The network's output was a 5×1 vector, representing the values of f_r for each of the five meanings. The use of sigmoid activations ensures that the output values are constrained within the interval $[0, 1]$, thereby respecting the defined image of f_r. For the training process, the entire dataset was utilized as a single batch, and training proceeded for 500 epochs. We employed a strategy of saving the model that achieved the best performance, as evaluated by the validation loss. The Adam optimizer was chosen for optimization, configured with a learning rate of 0.001 and a weight decay of 0.001. The network was trained using a cross-entropy loss function with $P_{L_1}(m|c, u)$.

B.2 Non-Literal Number Expressions

We show the listener meaning distributions for all context-meaning-utterances triples from the non-literal number expressions experiment. Figures 7 and 8 present these distributions for the electric kettle, Figures 9 and 10 present the distributions for the laptop and Figures 11 and 12 do the same for the watch.

B.3 Ironic Weather Utterances Additional Results

Meaning probability distributions by humans, alongside the listener and pragmatic listeners of both affect-aware RSA and **(RSA)**^{2} models, are presented for each of the nine weather contexts (which are associated with Fig. 5). Figures 13, 14, and 15 illustrate these distributions for weather contexts 1, 2, and 3 where the weather is visibly good. Figures 16, 17, and 18 show the distributions for weather contexts 4, 5, and 6 where the weather is neither visibly good nor bad. Finally, distributions for weather contexts 7, 8, and 9 where the weather is visibly bad are shown in Figures 19, 20, and 21.

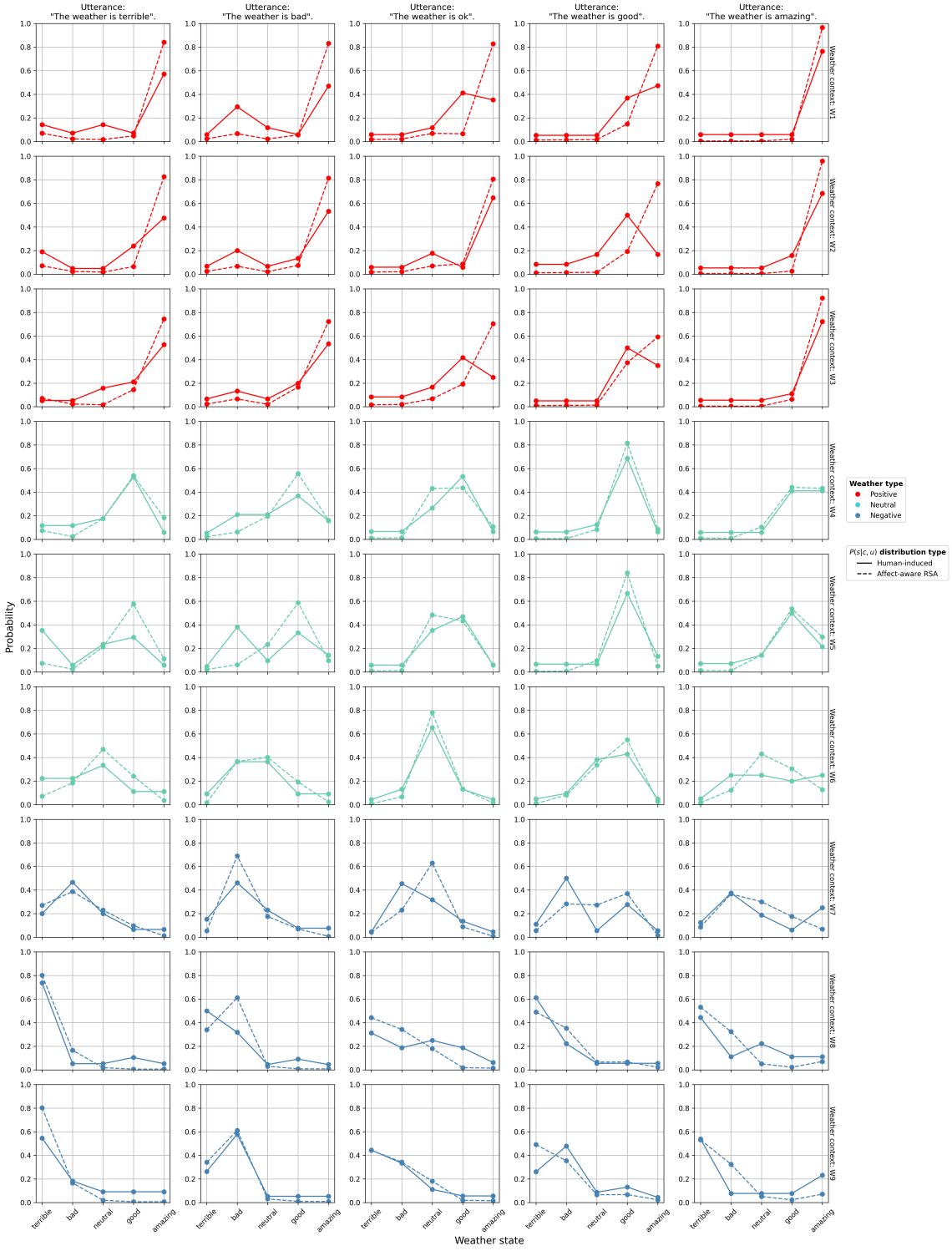

Figure 6: Meaning distributions, $P(m|c, u)$, for humans and for the affect-aware RSA method. This figure was originally generated by Kao and Goodman (2015). We reproduce it here with our re-implementation to verify the correctness of the re-implementation.

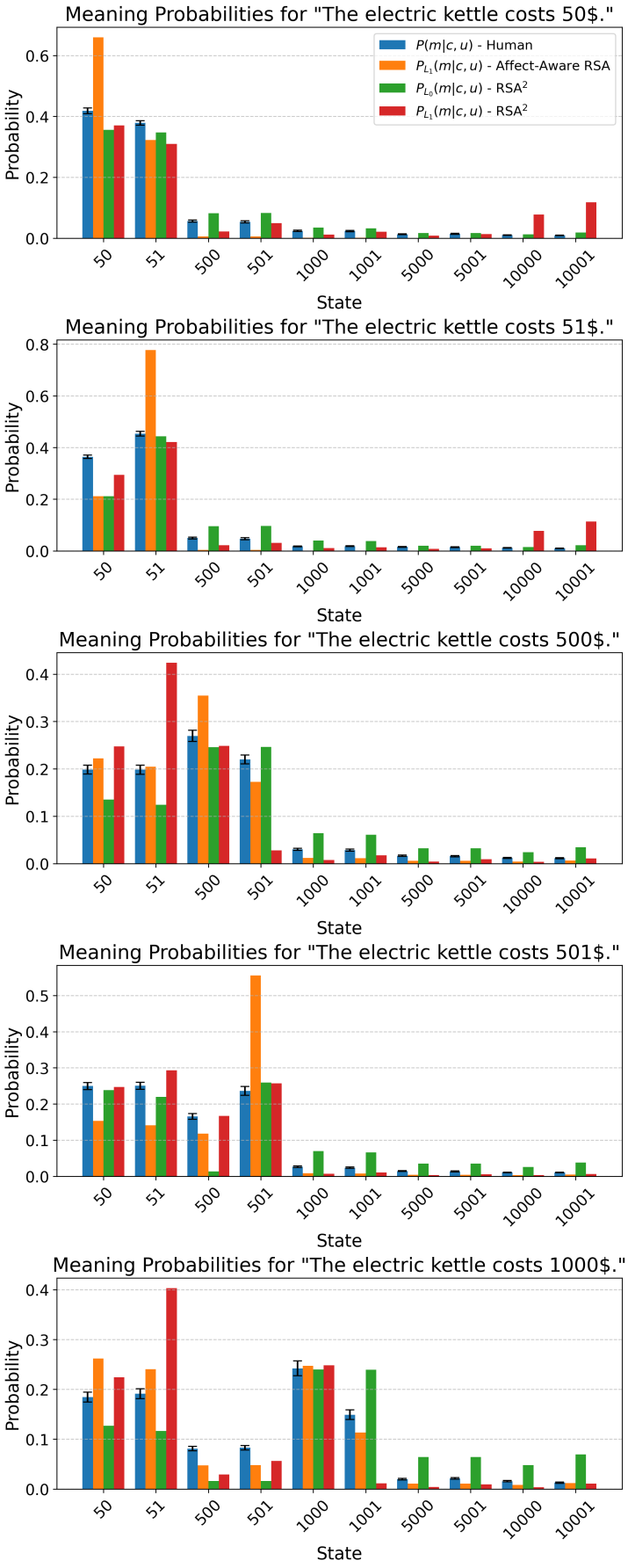

Figure 7: Meaning probability distributions by humans along with the affect-aware pragmatic listener and the **(RSA)**^{2} literal and pragmatic listeners on utterances about the price of an electric kettle.

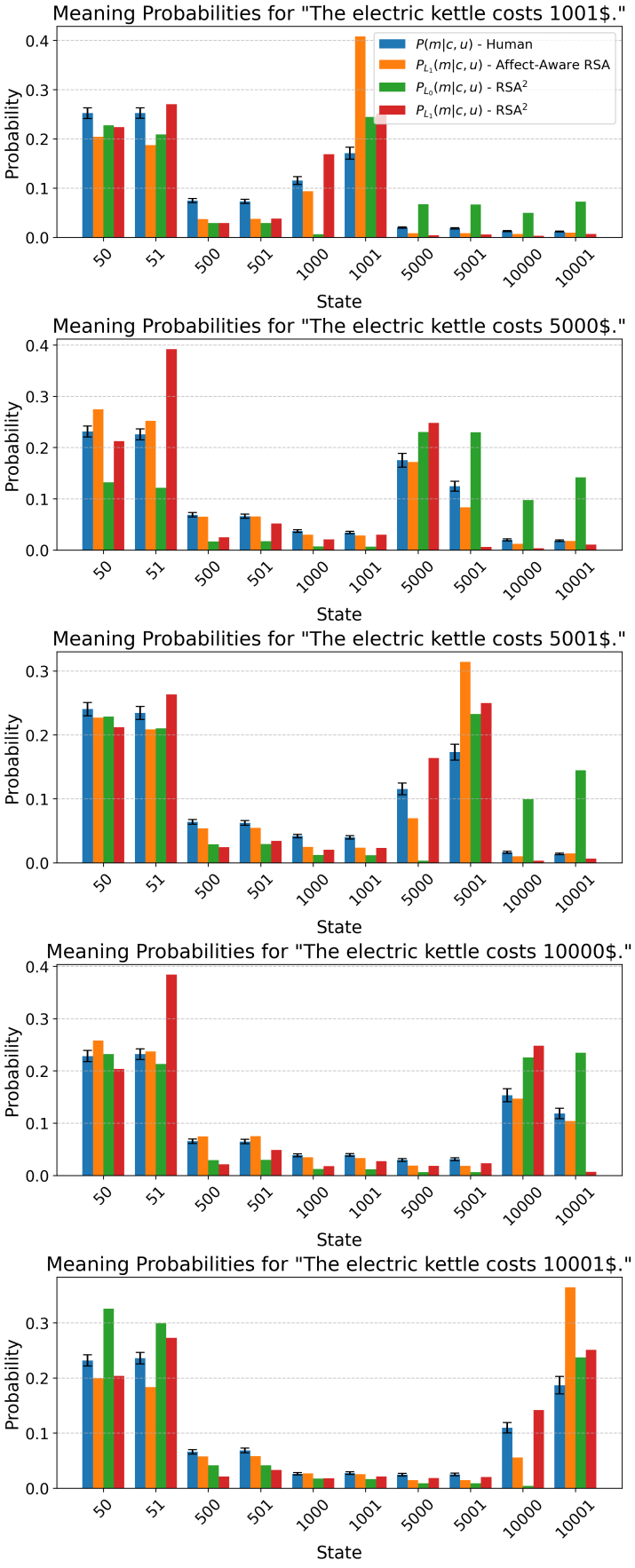

Figure 8: Meaning probability distributions by humans along with the affect-aware pragmatic listener and the **(RSA)**^{2} literal and pragmatic listeners on utterances about the price of an electric kettle (Continued).

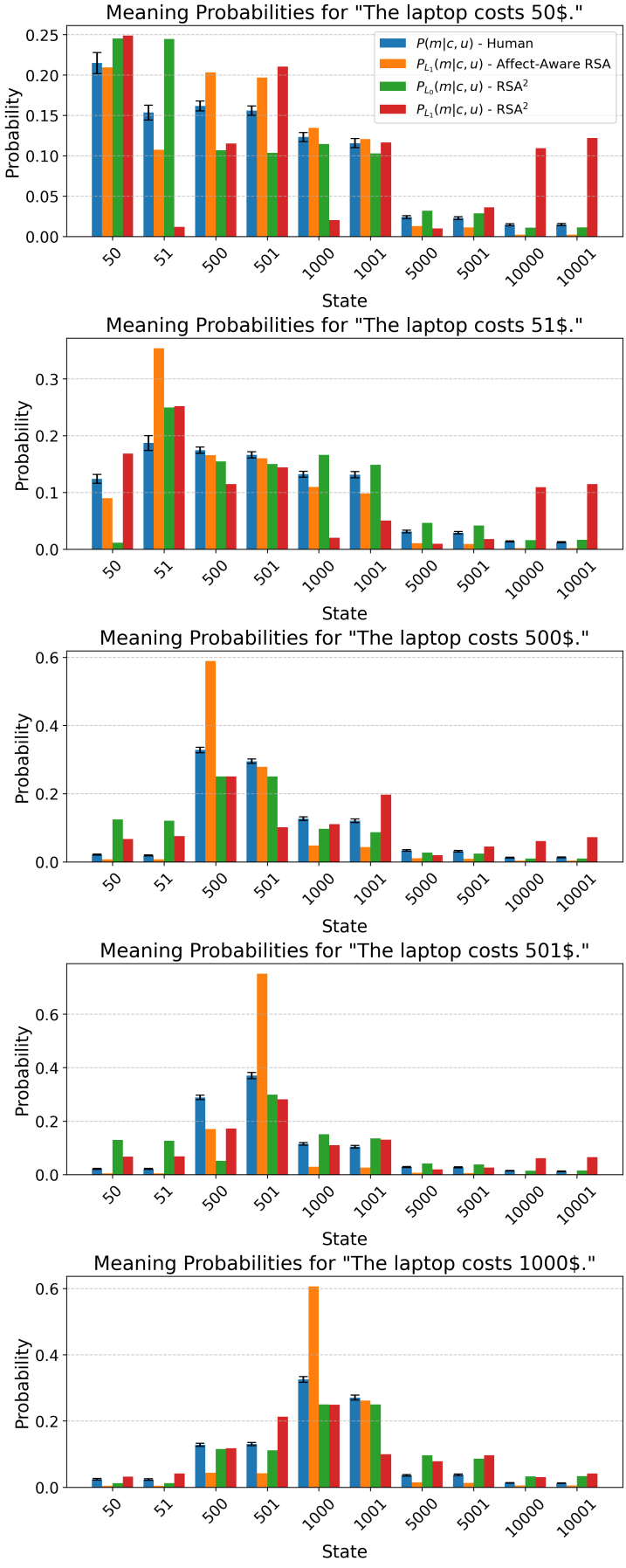

Figure 9: Meaning probability distributions by humans along with the affect-aware pragmatic listener and the **(RSA)**^{2} literal and pragmatic listeners on utterances about the price of a laptop.

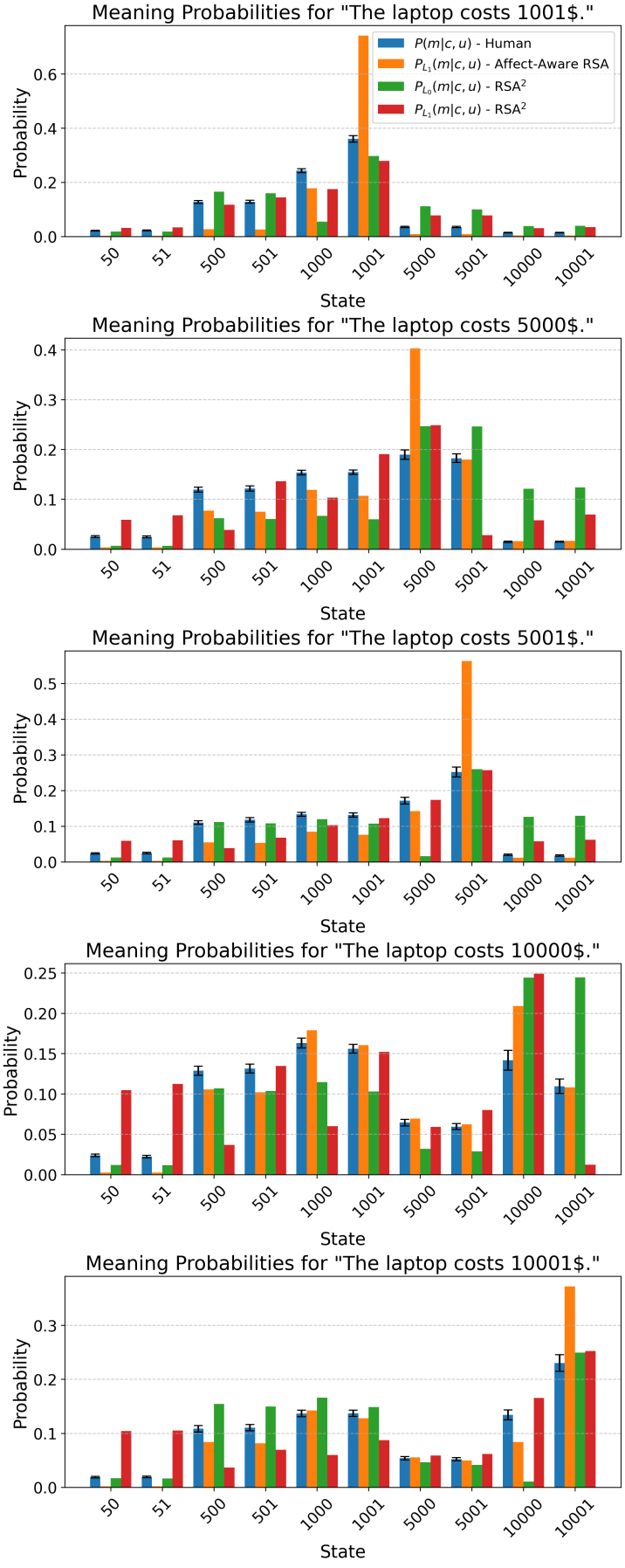

Figure 10: Meaning probability distributions by humans along with the affect-aware pragmatic listener and the **(RSA)**^{2} literal and pragmatic listeners on utterances about the price of a laptop (Continued).

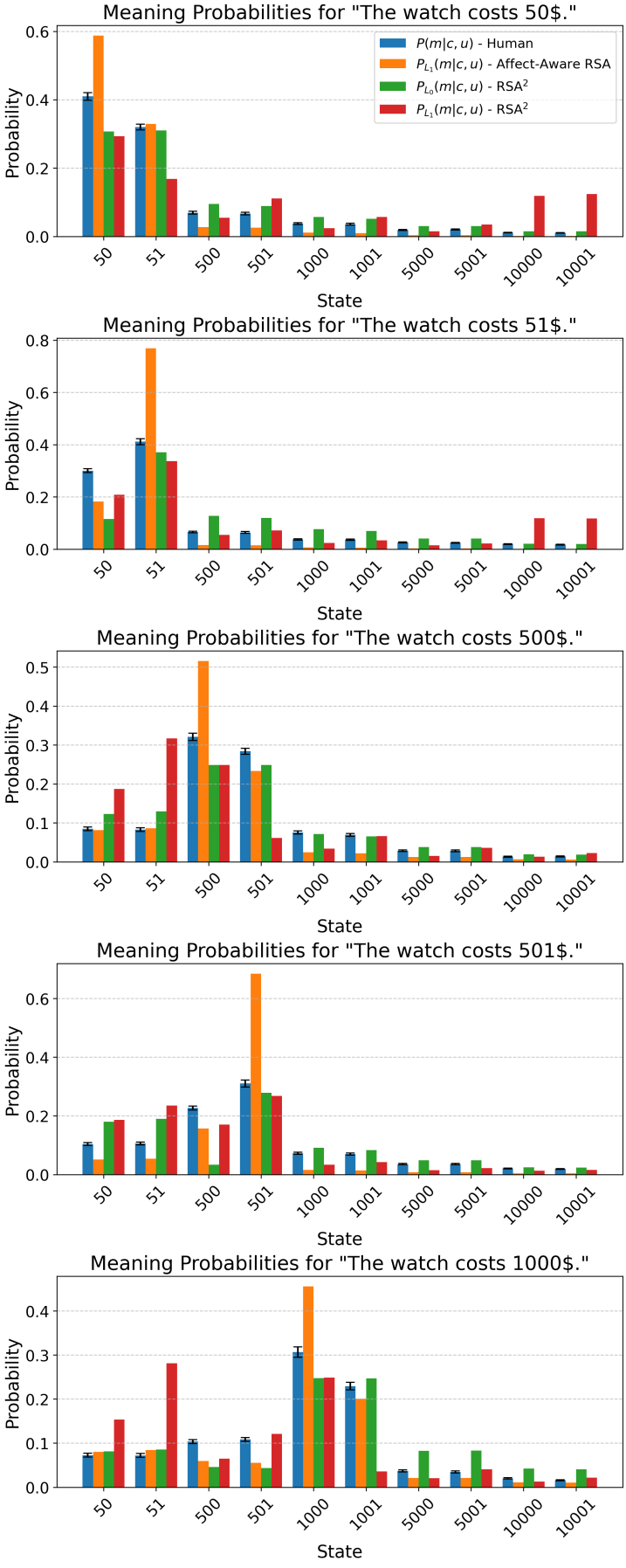

Figure 11: Meaning probability distributions by humans along with the affect-aware pragmatic listener and the **(RSA)**^{2} literal and pragmatic listeners on utterances about the price of a watch.

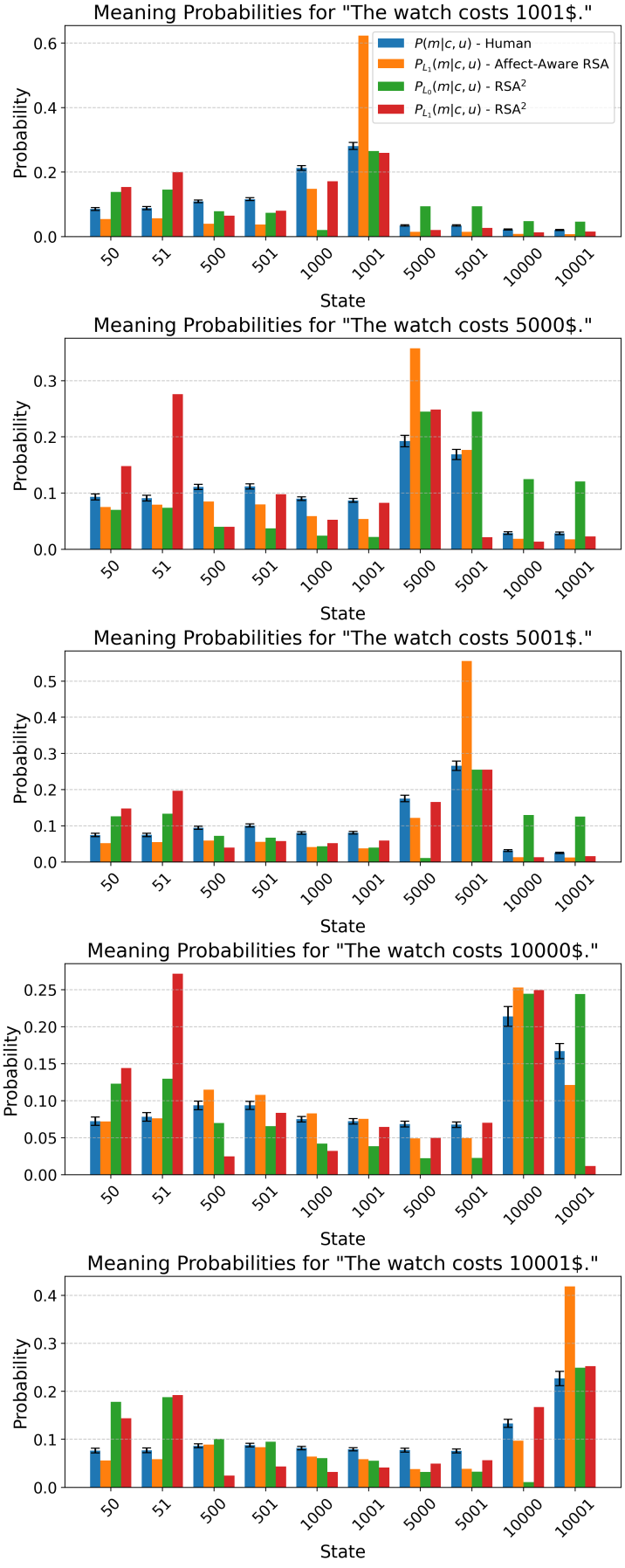

Figure 12: Meaning probability distributions by humans along with the affect-aware pragmatic listener and the **(RSA)**^{2} literal and pragmatic listeners on utterances about the price of a watch (Continued).

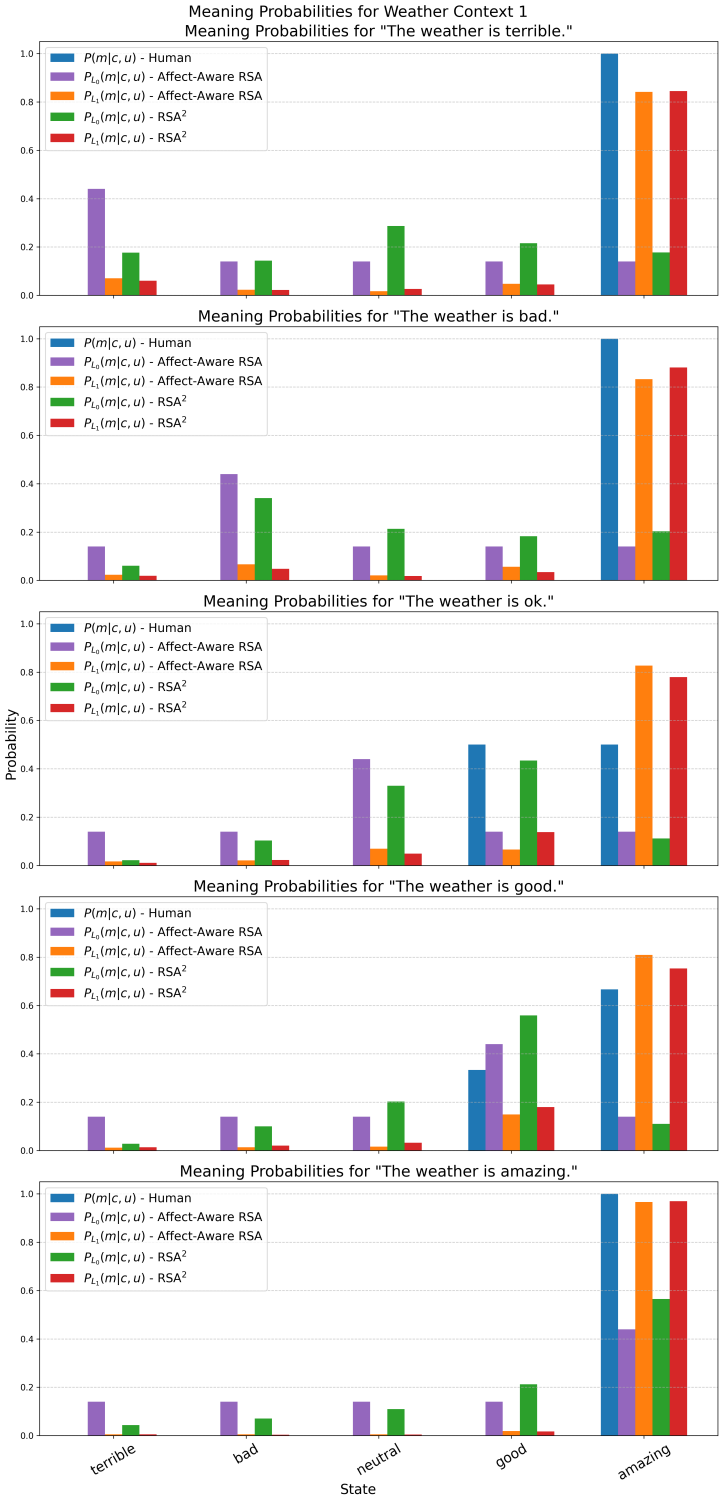

Figure 13: Meaning probability distributions by humans along with the listener and pragmatic listeners of both affect-aware RSA and **(RSA)**^{2} for weather context 1 from Fig. 5.

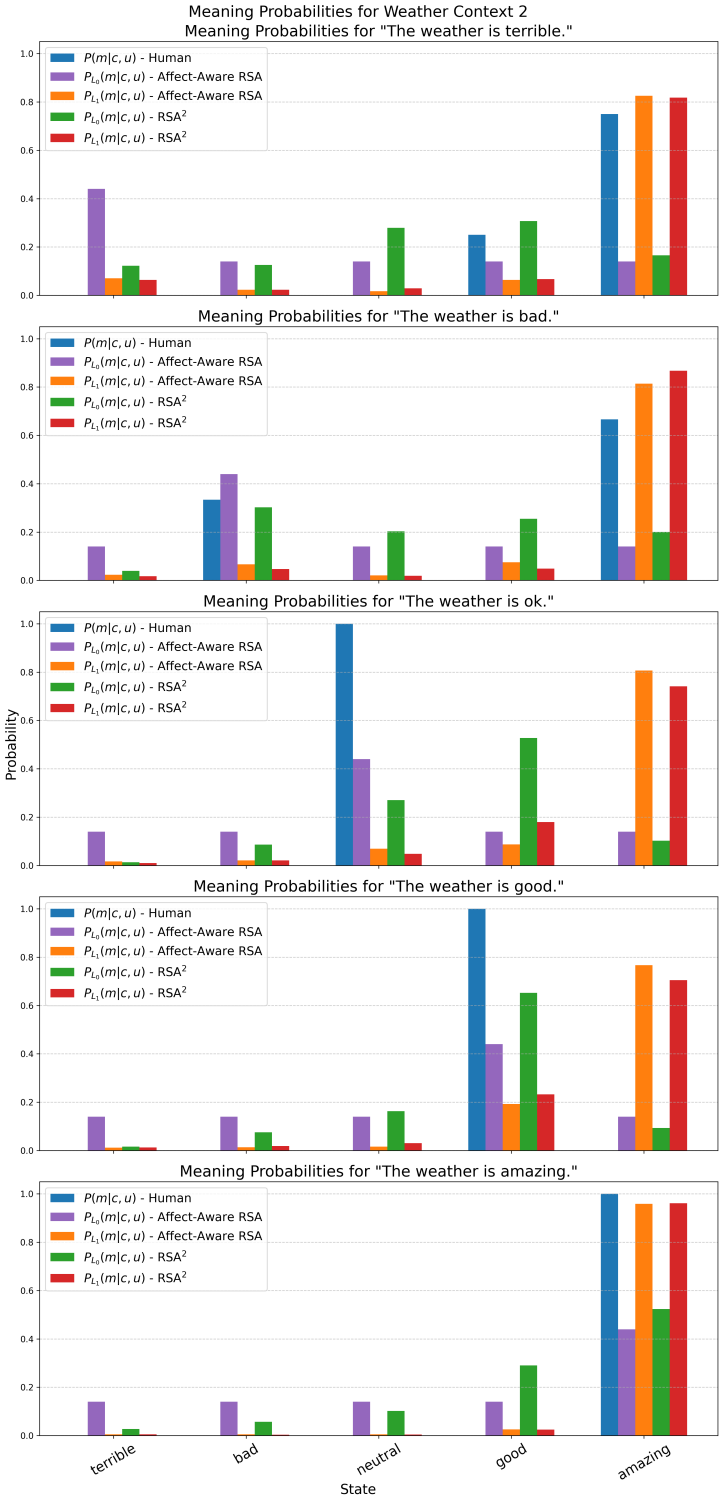

Figure 14: Meaning probability distributions by humans along with the listener and pragmatic listeners of both affect-aware RSA and **(RSA)**^{2} for weather context 2 from Fig. 5.

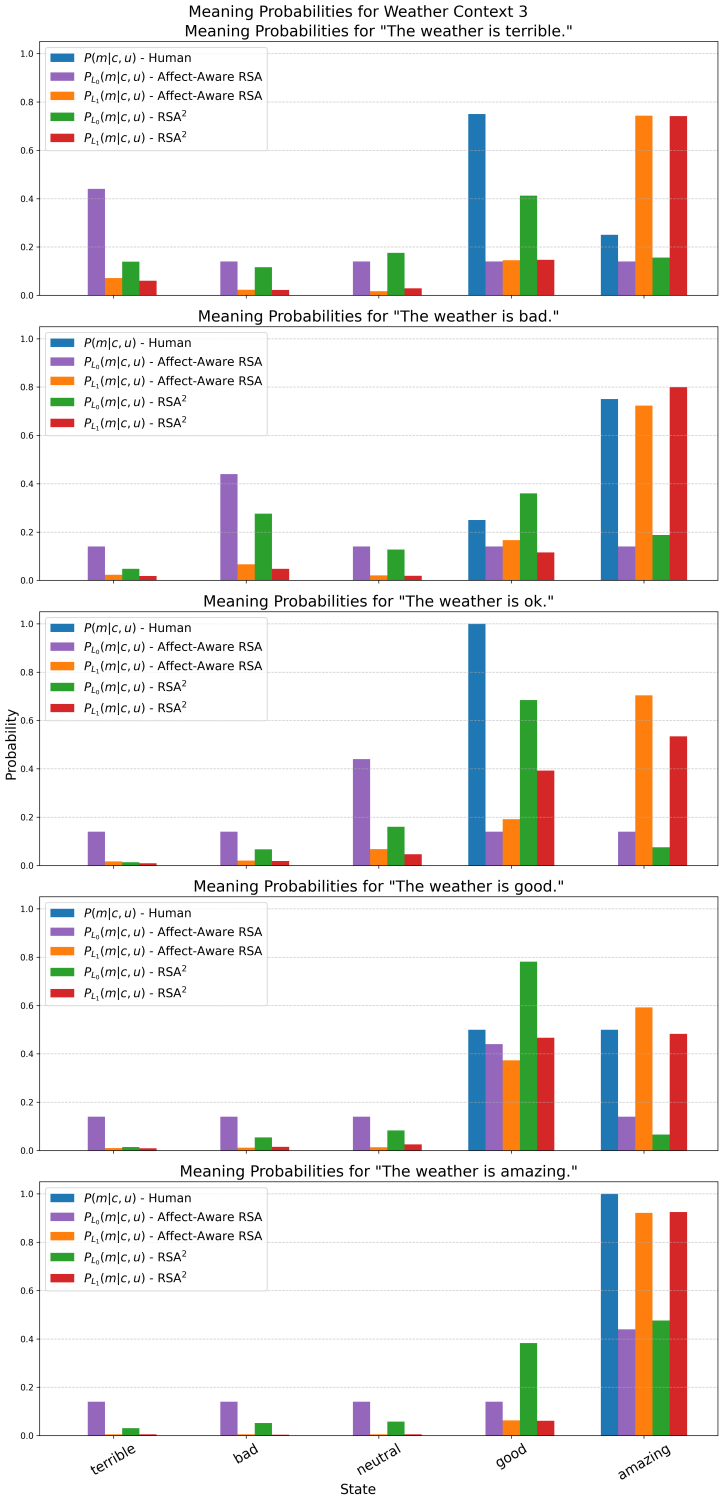

Figure 15: Meaning probability distributions by humans along with the listener and pragmatic listeners of both affect-aware RSA and **(RSA)**^{2} for weather context 3 from Fig. 5.

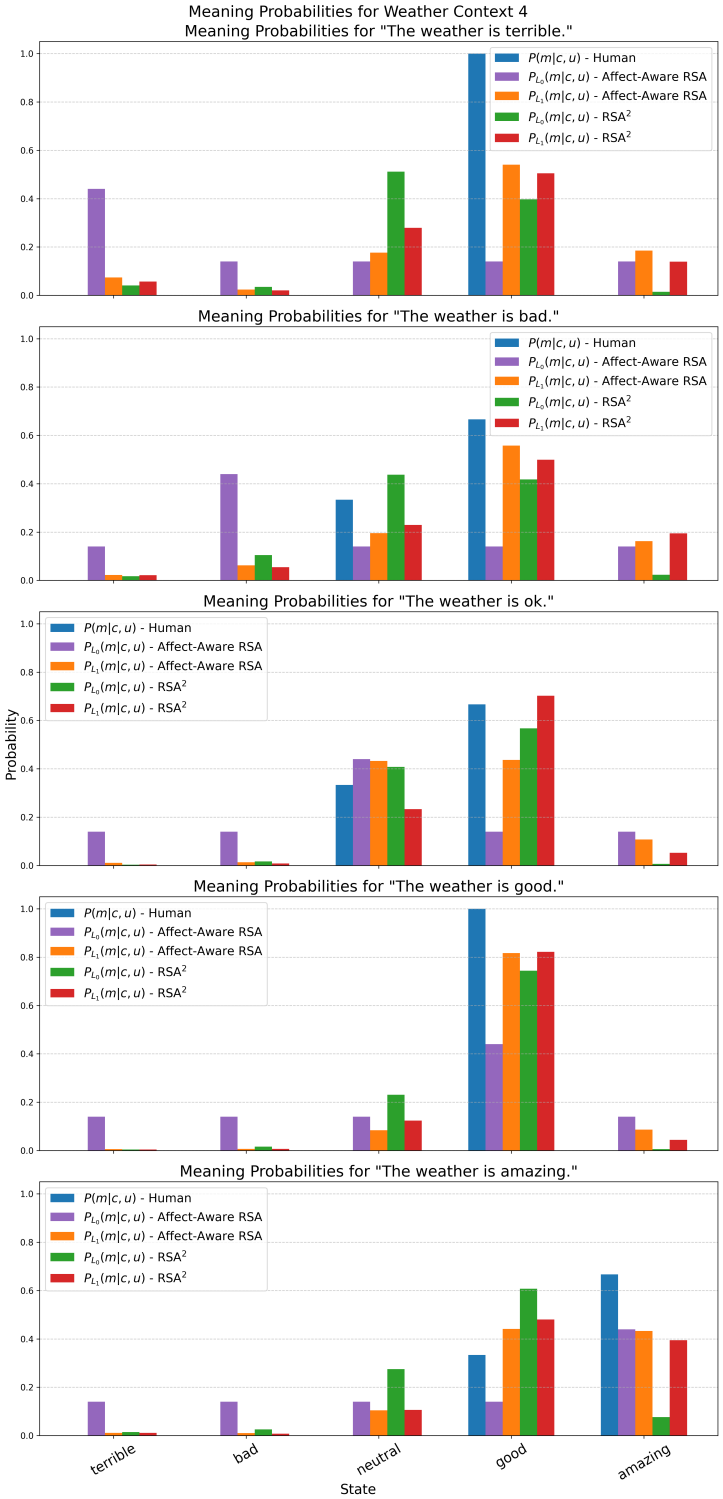

Figure 16: Meaning probability distributions by humans along with the listener and pragmatic listeners of both affect-aware RSA and **(RSA)**^{2} for weather context 4 from Fig. 5.

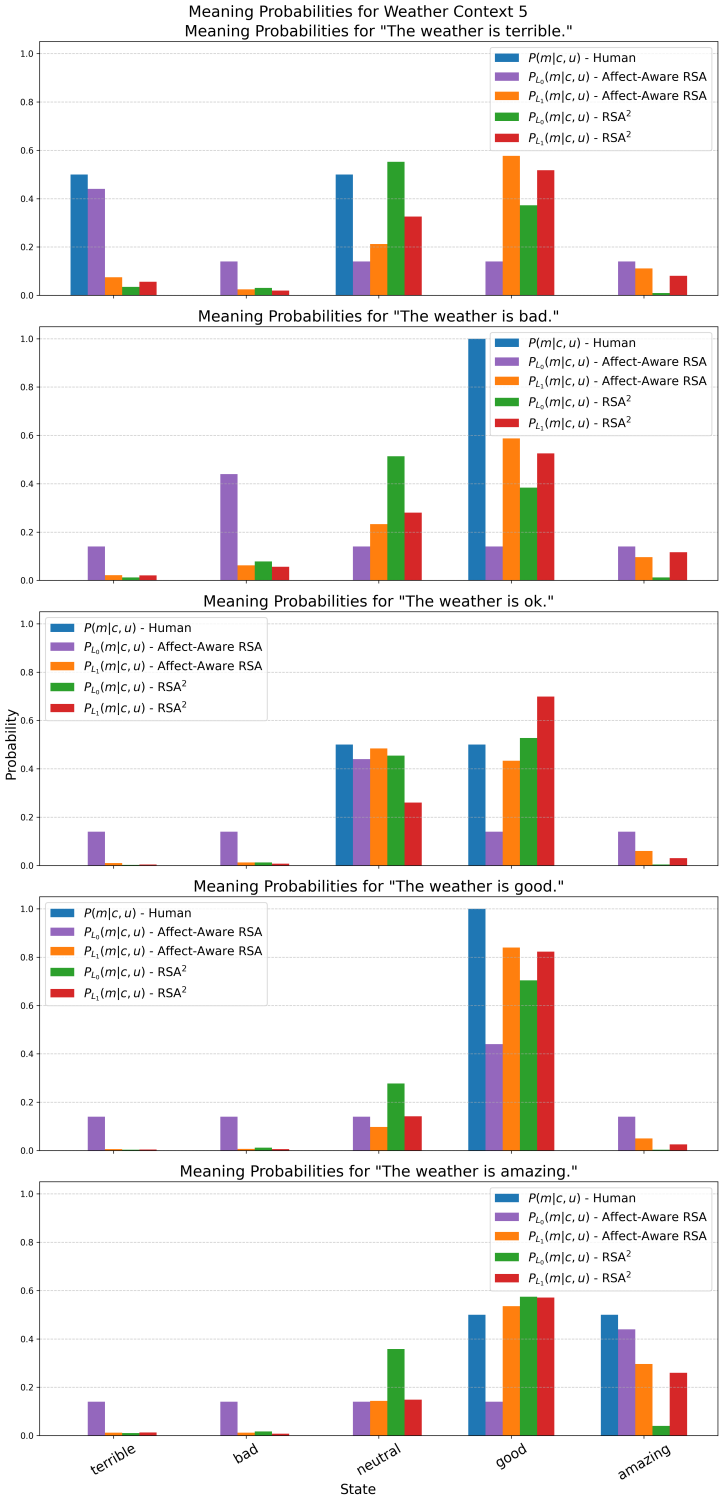

Figure 17: Meaning probability distributions by humans along with the listener and pragmatic listeners of both affect-aware RSA and **(RSA)**^{2} for weather context 5 from Fig. 5.

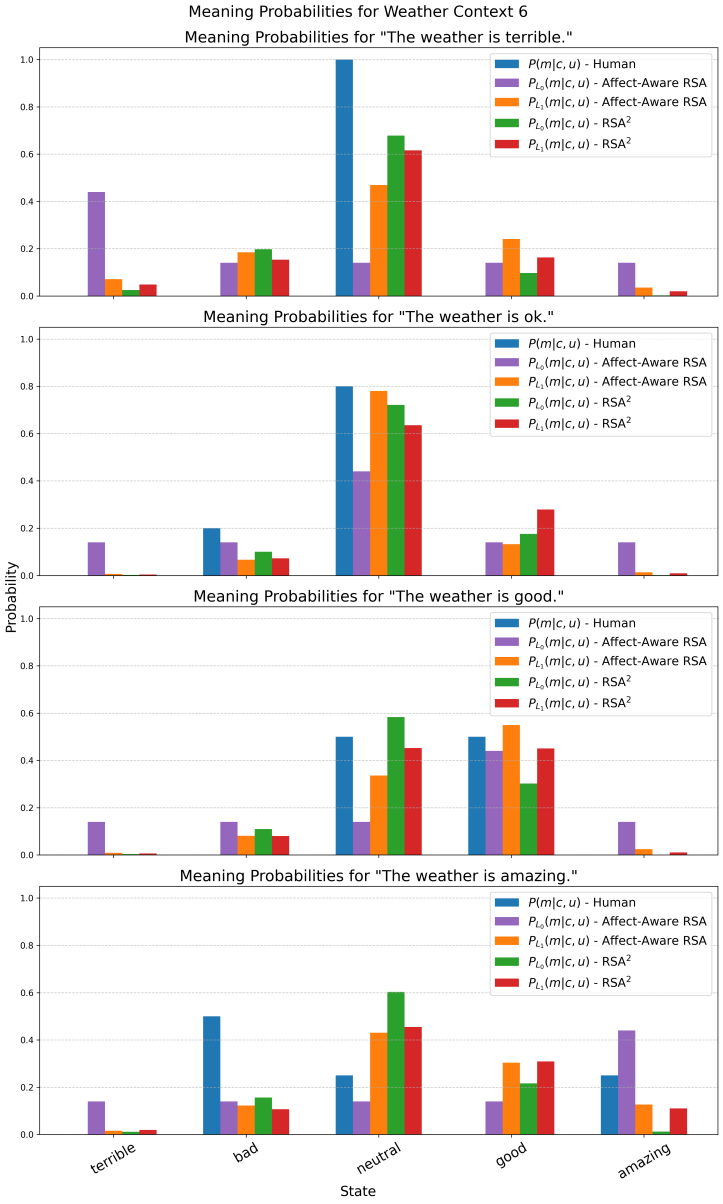

Figure 18: Meaning probability distributions by humans along with the listener and pragmatic listeners of both affect-aware RSA and **(RSA)**^{2} for weather context 6 from Fig. 5.

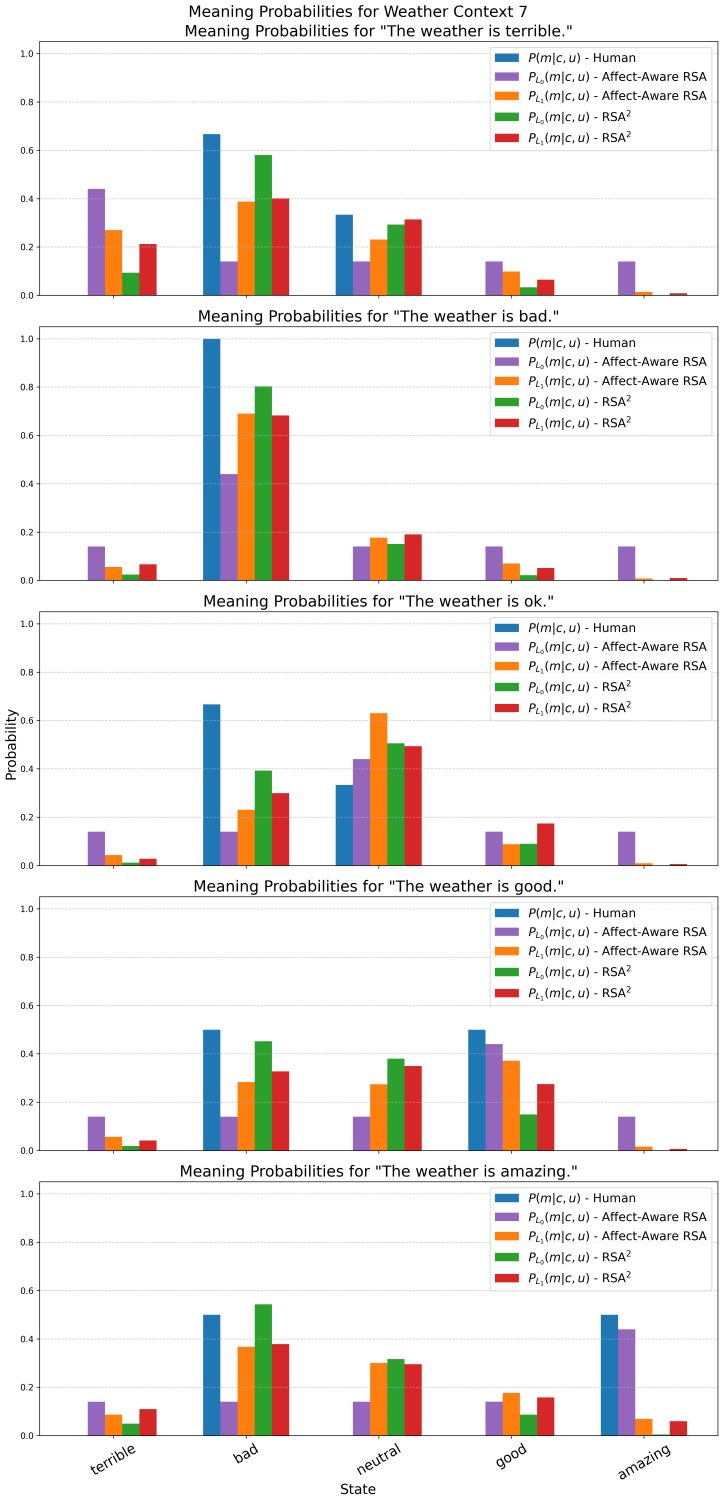

Figure 19: Meaning probability distributions by humans along with the listener and pragmatic listeners of both affect-aware RSA and **(RSA)**^{2} for weather context 7 from Fig. 5.

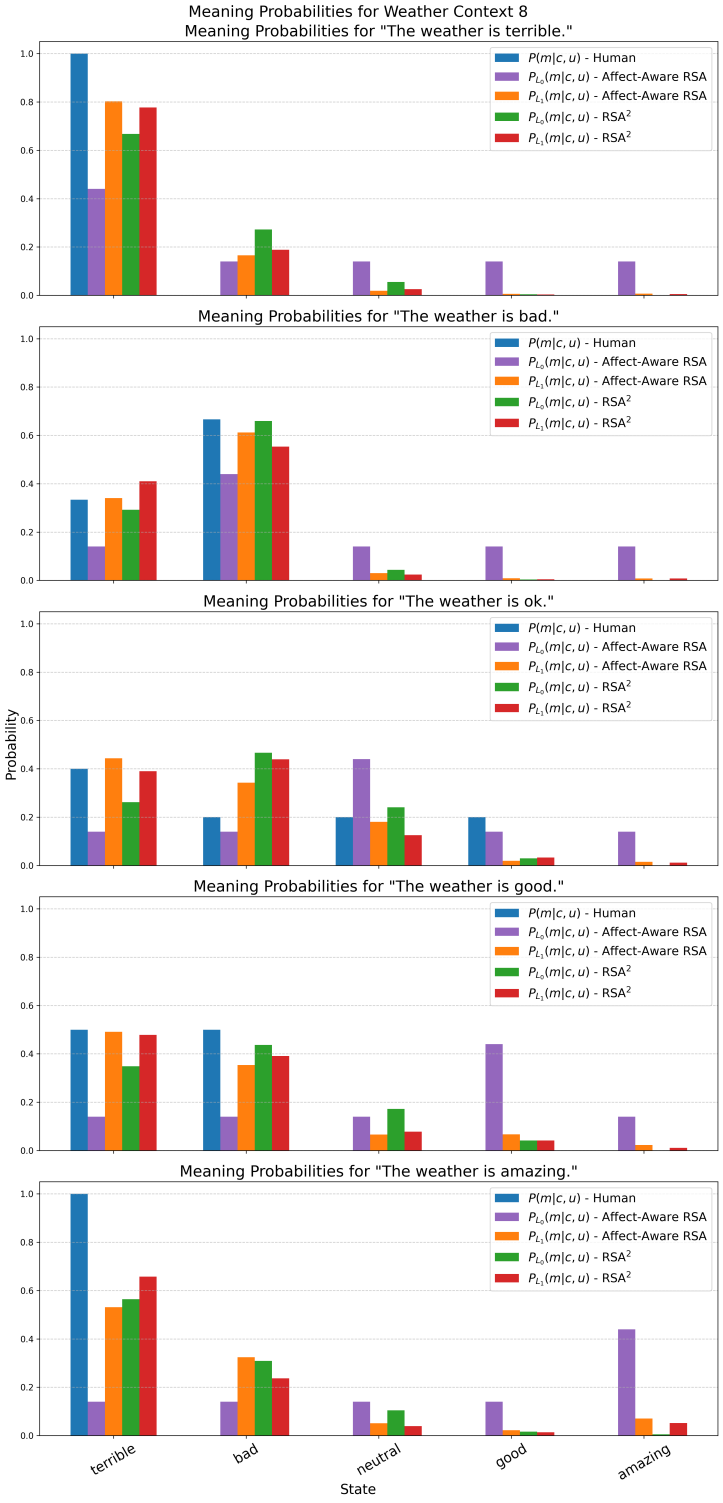

Figure 20: Meaning probability distributions by humans along with the listener and pragmatic listeners of both affect-aware RSA and **(RSA)**^{2} for weather context 8 from Fig. 5.

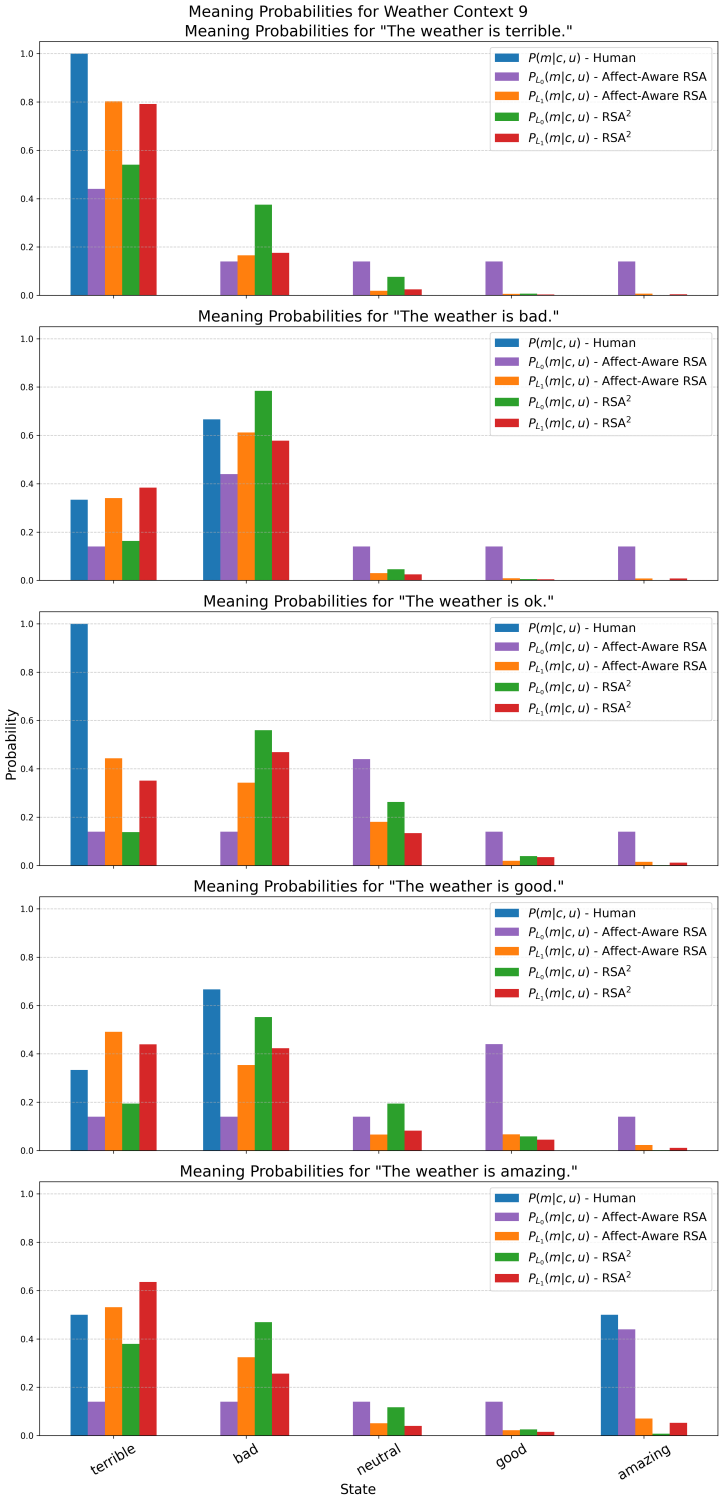

Figure 21: Meaning probability distributions by humans along with the listener and pragmatic listeners of both affect-aware RSA and **(RSA)**^{2} for weather context 9 from Fig. 5.

C LLM Irony Interpretation with (RSA)^2

C.1 PragMega+ Dataset Examples

Non-literal Intended Meaning	Literal Intended Meaning	Intended Meanings
In a shop, Lara tries on a dress. **The dress is far too long for her.** Lara asks Simon: "Does this dress fit me?" Simon answers: "Wow! That must be custom made! It's clearly the perfect size and length for you."	In a shop, Lara tries on a dress. **The dress fits her perfectly.** Lara asks Simon: "Does this dress fit me?" Simon answers: "Wow! That must be custom made! It's clearly the perfect size and length for you."	1. The dress is fitting well. (LM) 2. Lara needs to get a dress of a shorter length. (NLM) 3. Simon does not like the color of this dress. (OM) 4. Simon has to get back to work. (NSM)
While Tom and a new acquaintance from work, Sara, were chatting at a party, they noticed a colleague across the room. **She was standing alone holding a drink and a CD.** Tom points at the girl and comments: "The life of the party, right there."	While Tom and a new acquaintance from work, Sara, were chatting at a party, they noticed a colleague across the room. **She was standing at the center of a group of colleagues, holding a drink and and telling another one of her classic stories.** Tom points at the girl and comments: "The life of the party, right there."	1. Their colleague is very sociable. (LM) 2. Their colleague is quite unsociable. (NLM) 3. Their colleague has good taste in music. (OM) 4. Their colleague has good taste in fashion. (NSM)
After a long day, Bruce returns home and notices **that his kids have not cleaned their room.** Bruce says "I love how clean your room is." What did Bruce want to convey?	After a long day, Bruce returns home and notices **that his kids have cleaned their room.** Bruce says "I love how clean your room is." What did Bruce want to convey?	1. Bruce is happy with how clean his kids' room is. (LM) 2. Bruce is annoyed that his kids have not cleaned their room. (NLM) 3. It's important to keep one's room clean. (OM) 4. Bruce forgot to make dinner. (NSM)
After asking his parents several times, Edward has finally received a new gaming console for his birthday. **However, even though Edward promised that he would not spend more than two hours a day gaming, his parents quickly realize that Edward has no intention of keeping that promise.** Edward's mom tells he, "I see you are quite good at keeping your promises."	After asking his parents several times, Edward has finally received a new gaming console for his birthday. **Edward promised that he would not spend more than two hours a day gaming.** To their surprise, Edward's parents realize that Edward has every intention of keeping that promise. Edward's mom tells he, "I see you are quite good at keeping your promises."	1. She is disappointed that Edward has not kept he promise. (LM) 2. Edward has kept his promise. (NLM) 3. Spending too much time gaming is bad for one's health. (OM) 4. She needs to buy a new remote control. (NSM)

Table 4: Examples of a scenario with a non-literal intended meaning as well as its corresponding modified scenario with a literal intended meaning in the PragMega+ dataset as well as the shared intended meanings: 1. `Literal Meaning` (LM) 2. `Non-Literal Meaning` (NLM) 3. `Overlap Meaning` (OM) 4. `Non-Sequitur Meaning` (NSM). The first two scenarios are from the original PragMega dataset while the last two were manually generated by the authors as part of the test set.

C.2 Prompt Templates

We designed four prompt templates to elicit different conditional meaning distributions from language models. Each prompt targets a specific probability of interest within our framework. Figure 22 elicits $P(m \mid c, u)$, Figure 23 elicits $P(r \mid c, u)$, Figure 24 elicits for $P(m \mid c, u, r)$, and Figure 25 elicits $P(m \mid c)$.

Prompt Template for $P(m|c, u)$

Template:

```
Task: You will read short stories that describe everyday situations.
Each story will be followed by a multiple-choice question.
Read each story and choose the best answer.
Your task is to decide what the character in the story is trying to convey.
The answer options are 1, 2, 3, or 4.


[scenario]
What meaning is X likely conveying?

[options]

Answer:
```

Sample:

```
Task: You will read short stories that describe everyday situations.
Each story will be followed by a multiple-choice question.
Read each story and choose the best answer.
Your task is to decide what the character in the story is trying to convey.
The answer options are 1, 2, 3, or 4.


John is a teacher at an elementary school. When talking with the
principal about a new student, who did poorly on her entrance
examination, John said, ``This one is really sharp.''
What meaning is John trying to convey?

1. The pencils need to be sharpened.
2. The student is smart.
3. The student is not very clever.
4. The entrance exam is unfair.

Answer:
```

Figure 22: Prompt template for $P(m|c, u)$ where [scenario] is replaced with one of the (ironic or literal) scenarios from the PragMega dataset with the utterance and [options] are the 4 possible intended meanings for this particular scenario. X is replaced with the name of the speaker in the scenario which is extracted using a regular expression.

Prompt Template for $P(r|c, u)$

Template:

```
Task: You will read short stories that describe everyday situations.
Each story will be followed by a multiple-choice question.
Read each story and choose the best answer.
Your task is to decide, given the situation and what the character has said,
whether the character in the story is being "Sincere" or "Not Sincere".
A character is "Sincere" if what they are saying is consistent with the
context of the story. For instance, if the weather outside is
sunny then a "Sincere" utterance might be "The weather
outside is amazing." Likewise, if the weather outside is rainy then
a "Sincere" utterance might be "The weather is terrible."
In contrast, a character is "Not Sincere" if they are saying something
that contradicts the context of the story. For instance, if the weather
outside is rainy then a "Not Sincere" utterance might be "The weather
outside is amazing." Similarly, if the weather outside is sunny then
a "Not Sincere" utterance might be "The weather outside is terrible."
The answer options are 1 or 2.

[scenario]
Was X being sincere or not sincere?

[options]

Answer:
```

Sample:

```
Task: You will read short stories that describe everyday situations.
Each story will be followed by a multiple-choice question.
Read each story and choose the best answer.
Your task is to ... [see above].
The answer options are 1 or 2.


John is a teacher at an elementary school. When talking with the
principal about a new student, who did poorly on her entrance
examination, John said, ``This one is really sharp.''
Was John being sincere or not sincere?

1. Sincere
2. Not Sincere

Answer:
```

Figure 23: Prompt template for $P(r|c, u)$ where [scenario] is replaced with one of the (ironic or literal) scenarios from the PragMega dataset with the utterance and [options] are either sincere or not sincere. X is replaced with the name of the speaker in the scenario which is extracted using a regular expression. We used the "Sincere"/"Not sincere" terminology because we found that the term "Irony" would hurt performance.

Prompt Template for $P(m|c, u, r)$

Template:

```
Task: You will read short stories that describe everyday situations and which
finish with a character saying something. Your task is to decide, given the
situation and what the character has said, what meaning the character is trying
to convey. Each story will be followed by 4 possible meaning interpretations
listed from 1 to 4. Read each story and choose the number corresponding to the
best meaning interpretation. You can only answer with 1, 2, 3, or 4.
```

If sincere:

```
Assume that the character is saying exactly what they want to convey literally
when choosing the character's true intended meaning **even if** it contradicts
the context. For instance, if it's a sunny day and the character says "The
weather is terrible." then they do actually mean that the weather is terrible.
Similarly, if it's a rainy day and the character says "The weather is amazing."
then they do actually mean that the weather is amazing.
```

If not sincere:

```
In addition, when choosing the intended meaning, assume that the character is
saying the opposite of what they want to convey. For instance, if they say "The
weather is terrible." then they actually mean that the weather is amazing.
Similarly, if they say "The weather is amazing." then they actually mean that
the weather is terrible.
```

```
[scenario]
What meaning is X trying to convey?

[options]

Answer:
```

Figure 24: Prompt template for $P(m|c, u, r)$ where [scenario] is replaced with one of the (ironic or literal) scenarios from the PragMega dataset with the utterance and [options] are the 4 possible intended meanings for this particular scenario. X is replaced with the name of the speaker in the scenario which is extracted using a regular expression.

Prompt Template for $P(m|c)$

Template:

```
Task: You will read short stories that describe everyday situations
and which finish with a character saying something. Your task is to
decide, given the situation, what meaning the character is most
likely conveying. Each story will be followed by 4 possible meaning
interpretations listed from 1 to 4. Read each story and choose the
number corresponding to the most likely meaning. You can only answer
with 1, 2, 3, or 4.

[scenario]
What meaning is X likely conveying?

[options]

Answer:
```

Sample:

```
Task: You will read short stories that describe everyday situations
and which finish with a character saying something. Your task is ... [see above]


John is a teacher at an elementary school. When talking with the
principal about a new student, who did poorly on her entrance
examination, John said something.
What meaning is John likely conveying?

1. The pencils need to be sharpened.
2. The student is smart.
3. The student is not very clever.
4. The entrance exam is unfair.

Answer:
```

Figure 25: Prompt template for $P(m|c)$ where [scenario] is replaced with one of the (ironic or literal) scenarios from the PragMega dataset without the utterance and [options] are the 4 possible intended meanings for this particular scenario. X is replaced with the name of the speaker in the scenario which is extracted using a regular expression.

C.3 Additional Results

	$P_{L_0}(m = \texttt{NLM} \mid c, u, r = \textit{irony})$	$P_{L_1}(m = \texttt{NLM} \mid c, u, r = \textit{irony})$	$P_{L_0}(m = \texttt{LM} \mid c, u, r = \textit{literal})$	$P_{L_1}(m = \texttt{LM} \mid c, u, r = \textit{literal})$
Scenarios where the intended meaning is non-literal	0.92	0.94	0.47	0.62
Scenarios where the intended meaning is literal	0.16	0.59	0.95	0.88

Table 5: Average listener probability distributions *conditioned on* the *irony* and *literal* rhetorical strategies on both the ironic split (i.e., the split in which the intended meaning is non-literal) and the literal split (i.e., the split in which the intended meaning is literal).

	$P_N(r = \textit{irony} \mid c, u)$	$P_N(r = \textit{literal} \mid c, u)$
Scenarios where the intended meaning is non-literal	0.88	0.12
Scenarios where the intended meaning is literal	0.45	0.55

Table 6: Rhetorical strategy posteriors $P_N(r|c, u)$ on both the ironic split (i.e., the split in which the intended meaning is non-literal) and the literal split (i.e., the split in which the intended meaning is literal).

D RSC–RSA : A Rhetorical Strategy Clustering Algorithm

Algorithm 1 RSC–RSA

Ensure: Context c, Meaning m, Utterance u

Require: Set of induced rhetorical function values $F(c, m, u) = \{f_1(c, m, u), \ldots, f_n(c, m, u)\}$

▷ Initialize the following objects.

LLMs M and G

Embedding function $\texttt{embed} : \mathcal{U} \to \mathbb{R}^n$

K-means clustering function $\texttt{k-means} : \mathcal{P}(\mathbb{R}^n) \to \mathcal{P}(\mathbb{R}^n)$

Cosine similarity $\texttt{cosine-sim} : \mathbb{R}^n \times \mathbb{R}^n \to [0, 1]$

▷ Generate, embed and cluster alternative utterances.

$\mathcal{U}_{\text{alt}} \leftarrow \{u\} \cup \{u_i \sim P_G(u|c)\}$

$E \leftarrow \{\texttt{embed}(u') : u' \in \mathcal{U}_{\text{alt}}\}$

$\mathcal{X} \leftarrow \texttt{k-means}(E)$

▷ Compute rhetorical function value for each cluster.

$F(c, m, u) \leftarrow \emptyset$

for $r \in \mathcal{X}$ **do**

  $\mathcal{U}_r \leftarrow \{u' \in \mathcal{X}_{\text{alt}} : u' \in \text{ cluster } r\}$

  ▷ Compute cluster meaning probability and weight it by the utterance distance to the centroid.

  $p_{mc} \leftarrow \frac{\sum_{u' \in \mathcal{U}_r} P_M(m|c,u')}{\sum_{m' \in \mathcal{M}} \sum_{u' \in \mathcal{U}_r} P_M(m'|c,u')}$

  $f_r(c, m, u) \leftarrow \frac{p_{mc}}{P_M(m|c,u)}$

  Append $f_r(c, m, u)$ to $F(c, m, u)$

end for

return $F(c, m, u)$

We present our clustering algorithm, **R**hetorical **S**trategy **C**luster RSA **RSC–RSA**, which attempts to automatically *induce* the most salient rhetorical strategies and their corresponding rhetorical functions. To do so, we rely on the intuition that utterances which are generated with the same intended meaning and rhetorical strategy are likely to be semantically similar. For example, the utterances "The weather is amazing." and "Gosh, this weather is so great!", uttered in the context of bad weather, are likely to both employ the same rhetorical strategy (irony) to convey the same intended meaning (that the weather is in fact terrible). The algorithm which computes the set of induced rhetorical function values, $F(c, m, u) = \{f_1(c, m, u), \ldots, f_n(c, m, u)\}$, generates alternative utterances using a base LLM, embeds them and clusters them with k-means to induce *rhetorical strategy clusters* which act as proxies of prototypical rhetorical strategies. The full procedure is presented in Algorithm 1.

In brief, the **RSC–RSA** algorithm generates, embeds and clusters a set of alternative utterances which could have occurred in context c to create the rhetorical strategy clusters $\mathcal{X}$. Using the generated clusters $\mathcal{X}$, **RSC–RSA** computes the corresponding rhetorical function values, $f_r(c, m, u)$, for each cluster r for a given c, m, u triple. To do so, **RSC–RSA** uses a formula which averages the meaning probabilities of all the utterances in that cluster and divides them by $P_M(m|c, u)$. In this way, the induced rhetorical strategy values parallel those from Equation 8 in that they return a ratio of two probabilities. The equations for computing f_r are as follows:

$$p_{mc} = \frac{\sum_{u' \in \mathcal{U}_r} P_M(m|c, u')}{\sum_{m' \in \mathcal{M}} \sum_{u' \in \mathcal{U}_r} P_M(m'|c, u')}, \tag{37}$$

$$f_r(c, m, u) = \frac{p_{mc}}{P_M(m|c, u)}. \tag{38}$$

These values are then used to compute $P_{L_0}(m|c, u, r)$ as defined in Equation 4. To marginalize across rhetorical strategy clusters, we use the relative cluster size, i.e. $P_{R|CU}(r|c, u) = \frac{|\mathcal{U}_r|}{|\mathcal{U}|}$.

If the alternative utterances for u = "The weather is amazing." include u_1 = "Gosh, this weather is so great!", u_2 = "The weather is terrible." and u_3 = "The weather is so bad.", then we expect for the embeddings of u and u_1 to be in one cluster and for the embeddings of u_2 and u_3 to be in another cluster. These two rhetorical strategy clusters would then resemble the *ironic* and *literal* rhetorical strategies from Section 3.2 and their **RSC–RSA**-derived rhetorical function values would approximate those of $f_{\textit{irony}}$ for u and u_1 and of $f_{\textit{literal}}$ for u_2 and u_3.

D.1 RSC–RSA Implementation Details

To implement the embedding and clustering procedures of **RSC–RSA**, we utilized the Sentence-BERT architecture (Reimers and Gurevych, 2019) and the k-means clustering algorithm from scikit-learn (Buitinck et al., 2013). The k-means algorithm was executed with 10 initializations, using default settings from scikit-learn.

Listener	Average $P(m\|c,u)$ across all scenarios	Average $P(m\|c,u)$ across ironic scenarios	Average $P(m\|c,u)$ across literal scenarios
RSA-RSC – L_0 with $P(m\|c)$	0.59	0.91	0.28
RSA-RSC – L_1 with $P(m\|c)$	0.66	0.91	0.41
RSA-RSC – L_0 without $P(m\|c)$	0.27	0.50	0.034
RSA-RSC – L_1 without $P(m\|c)$	0.28	0.52	0.036

Table 7: Average $P(m|c,u)$ for different RSA-RSC listener models.

D.2 Clustering Results

We experimented with 2, 4, and 8 clusters yielding no significant difference. As a result, we report results of the **RSC–RSA** clustering method with 4 clusters with and without the meaning prior, $P(m|c)$, in Table 7. We see that across all listeners and scenario types, the performance is largely driven by the meaning prior $P(m|c)$ with the performance on the literal scenarios being worse than random. We encourage future work to improve upon our initial attempt at automatically uncovering the most salient rhetorical strategies in a given context and leveraging them within the **(RSA)**^{2} framework.