The lady doth protest too much!
Informativity expectations affect judgments of argument strength

S.R. Pulapura* **Hannah Rohde** **Daniel Lassiter** **Catherine Lai**

School of Philosophy, Psychology, and Language Sciences
University of Edinburgh

Abstract

Bayesian models of argumentation predict that a rational agent's credence should increase monotonically as new evidence supporting a claim becomes available. This paper investigates a competing prediction: in natural language dialogues, listeners may instead become suspicious when speakers justify their beliefs with *excess* reasons. We analyze this phenomenon as a rational social inference in the tradition of Grice, whereby listeners may "rationalize" seemingly unnecessary argumentation by inferring that the speaker expected a disagreement. We introduce a pragmatic model of argumentation, adapting the Rational Speech Act framework, and report a behavioral study that tests its predictions against a first-order Bayesian model. We find that information-theoretic expectations can modulate belief revision when listeners evaluate evidence in a dialogue setting, an effect which standard models of rationality fail to explain.

1 Introduction

Language is built upon a foundation of trust, and there is a case to be made that this trust is rational: our modern lives depend substantially on testimonial knowledge (Levy, 2021; Mercier and Sperber, 2017), and communication would be of little use if it were not generally reliable (McCready, 2015; Sperber et al., 2010). However, listeners obviously do not always accept assertions on the basis of conversational norms (Sperber et al., 2010), and speakers do not expect addressees to take implausible testimony at face value (Oey et al., 2023). Even when the Quality maxim (Grice, 1989) is presumably in effect, interlocutors might have access to different bodies of information against which they evaluate facts, or else might differ subjectively on matters of taste or opinion (Stephenson, 2007). This study concerns the use of *reasons* to negotiate controversial beliefs.

When some belief C is difficult to accept on the basis of trust alone, a speaker may provide reasons to make it available by inference instead (Mercier, 2020; van Eemeren et al., 2014):

(1) We should take the bus to New York today, ***because the train is delayed by 2 hours.***

A reason R is said to "count in favor" of C when the truth of R raises the probability of C in the context of an agent's other beliefs. Conclusions justified by many coherent reasons should then be more credible– or at least, no less credible– than conclusions justified by few or none. This sensible analysis is predicted by standard Bayesian models of argument strength (Oaksford and Hahn, 2004; Godden and Zenker, 2018), but seems contra the intuition that superfluous reasons are somehow suspicious (*The lady doth protest too much!*).

We propose that the latter intuition arises as an effect of rational *language use*. Rational speakers tend to be more redundant when expressing surprising or low-probability content, and less redundant when expressing predictable content; this tendency helps maximize the amount of information a speaker can convey while keeping production cost to a minimum (Aylett and Turk, 2004; Levy and Jaeger, 2007; Piantadosi et al., 2011, a.m.o.). Following Grice (1989), recent experiments by Kravtchenko and Demberg (2022a,b) suggest that listeners expect this production preference, and can "rationalize" assertions that seem redundant (or trivially predictable) by inferring that the content might actually be "surprising" in a given context:

(2) "Alice went to the grocery store, and ***paid the cashier!***" $\rightsquigarrow$ *Alice doesn't always pay.*

We investigate whether interlocutors are similarly economical when coordinating beliefs via argumentation. If so, a rational speaker should avoid giving reasons when her[1] interlocutor is likely to

*Correspondence: s.pulapura@sms.ed.ac.uk

[1]We refer to the speaker as "she" and the listener as "he".

agree with her anyways. If listeners expect this production preference, effortful argumentation for a seemingly innocuous conclusion might signal to a listener that there could be relevant grounds for disagreement which he hadn't considered before.

We develop a Rational Speech Act model (Frank and Goodman, 2012) that predicts this inference, taking reasons that justify actions as our case study. A behavioral experiment confirms that seemingly unnecessary reasons can signal to a listener that the speaker expected a disagreement. The inference seems to mitigate the effectiveness of the evidence, too, though further experimentation is required for a more conclusive result. We also discuss some speaker-specific and sociolinguistic factors which could modify or suppress the observed inference. Ultimately, our contributions are twofold: (1) we provide empirical support for our prediction that verbal reasons are evaluated in light of a speaker's rational choice to provide them, and therefore (2) we explain why an intuition that seems at odds with Bayesian rationality might yet be rational after all.

2 Theoretical Background

Rational communication. Going back at least to Zipf (1935), it has been of great interest to linguists and psychologists whether language users construct their utterances "rationally"– that is, by maximizing information exchange while minimizing resource expenditure. Most modern psycholinguistic accounts treat informativity as a measure of *surprise*: all linguistic units, from phonemes to full utterances, are more informative when surprising, and less informative when predictable (Aylett and Turk, 2004; Levy, 2008; Levy and Jaeger, 2007). Following Shannon (1948), the information content (*surprisal*) of an utterance u is given as its negative log-probability in context.

$$I(u_i) = -\log P(u_i|u_0...u_{i-1}) \qquad (1)$$

On such accounts, rational communication is modeled as an efficient flow of information through a noisy channel with limited bandwidth. To maximize the amount of information that can be transmitted through the channel, a constant rate of information that approaches, but does not exceed, its capacity must be maintained. Thus, rational speakers spread surprising content across longer signals, and compress predictable content into shorter signals, in order to maintain the optimal information rate (Aylett and Turk, 2004; Levy and Jaeger, 2007).

Reduction in low-surprisal contexts. Natural languages are not maximally efficient, of course. However, over the past 20 years, numerous adaptations of the noisy-channel model have captured patterns of signal reduction in low-surprisal contexts across all levels of linguistic structure. These include phonetic reductions (Aylett and Turk, 2004; Bell et al., 2003; Pluymaekers et al., 2005), contractions (Piantadosi et al., 2011), lexical reductions (e.g. chimp/chimpanzee; Mahowald et al., 2013), complementizer omission (Jaeger, 2010), fragments (Bergen and Goodman, 2015; Lemke et al., 2021), and omission of discourse connectives (Torabi Asr and Demberg, 2012). All of these phenomena occur more frequently when the reduced or omitted content is contextually predictable.

Redundancy in high-surprisal contexts. Despite these findings, rational theories (including the submaxim of Quantity *"Do not make your contribution more informative than is required"*; Grice, 1989) have come under fire because of their apparent inability to account for the logical redundancies that abound in naturalistic communication (Engelhardt et al., 2006; Heller et al., 2012). For example, speakers often provide logically stronger references than strictly required to identify an object:

(3) *[Context: the display has just one square.]* "Look at **the blue square**!"

However, more recent work suggests that speakers are not simply redundant at random. Often times, they "overspecify" to ensure successful communication of *high*-surprisal content. For example, optional modification occurs more frequently with unpredictable attributes ("yellow strawberry" rather than "yellow banana"; Degen et al., 2020), in complex visual displays (Rubio-Fernandez, 2016), and in learner-directed speech (Bergey et al., 2020).

Rationalizing apparent redundancies. If rational speakers tend to be economical in this way, rational *listeners* might then be expected to interpret utterances under the assumption that they are sufficiently informative or "mention-worthy". A growing body of literature suggests that this is often the case (Hao et al., 2024; Rohde et al., 2021; Rohde and Rubio-Fernandez, 2022; Reksnes et al., 2024). Moreover, utterances that seem redundant at first may generate inferences that change the listener's model of the context into a state where the content would be informative after all. One such inference is shown by Kravtchenko and Demberg (2022a,b):

participants in their studies rationalized assertions of seemingly trivial script continuations by inferring that the content was actually *un*predictable in the current context (repeated from (2)):

(4) "Alice went to the grocery store, and ***paid the cashier!***" $\rightsquigarrow$ *Alice doesn't always pay.*

Beyond cooperativity. These works are situated within the broader Gricean program, wherein listeners are expected to update their beliefs to align with a cooperative speaker's intended meaning. Therefore, redundancy inferences that reduce the contextual predictability (or *prior* probability) of a claim do not adversely impact the listener's *posterior* belief. Upon hearing (4), the listener is predicted to accept that "Alice paid the cashier" is *true*, despite also inferring that it was less contextually *likely* than he thought. However, when listeners are not expected to update their beliefs on the basis of trust alone, redundancy inferences may also impact their posteriors, we will show.

Learning from reasons. While the "cooperative information exchange" model is a highly theoretically useful idealization, cooperative assertion (with corresponding pragmatic enrichment) is not the only strategy available for changing another agent's beliefs. Indeed, a "vigilant" listener will probably be unmoved by a simple assertion of implausible content (Mercier, 2020; Sperber et al., 2010). However, since new information is easier to accept when it coheres well with the rest of one's beliefs, a speaker may guide a skeptical listener to accept a challenging conclusion nonetheless by increasing its contextual plausibility instead (Mercier, 2020; Sperber et al., 2010; van Eemeren et al., 2014). The purpose of *argumentation* is, therefore, to reduce or eliminate the need for trust among interlocutors by making the target content inferrable from other propositions (*reasons*) which the listener either already believes, or will accept more easily.

As noted by Sperber et al. (2010), Grice himself acknowledges a difference between belief changes induced by trust vs. by reasons:

> *"Conclusion of argument: p, q, therefore r: While U[tterer] intends that A[ddressee] should think that r, he does not expect (and so intend) A to reach a belief that r on the basis of U's intention that he should reach it.* ***The premises, not trust in U, are supposed to do the work****"* (Grice, 1989).

An arguer does not attempt to update the context directly with a claim C (e.g. Stalnaker, 1978), or else to set the probability of C to 1 regardless of its predictability in the prior context. Instead, she demonstrates that C must be so in light of some uncontroversial upstream premises R. In other words, C is not simply true, but rather *coherent with* (and so *predictable from*) the listener's other beliefs.

Bayesian models of argument strength. Linguists studying argumentation have endeavored to formalize this sort of "evidential" meaning (particularly those theorists who take it to be a primary component of the semantics, e.g. Anscombre and Ducrot, 1983). Bayesian models have increasingly been recruited for this purpose (Merin, 1999; Wintersteín, 2018). Inspired by work in the psychology of reasoning (Hahn and Oaksford, 2007; Godden and Zenker, 2018; Oaksford and Hahn, 2004), Bayesian models quantify the evidential contribution[2] of a set of reasons R as the extent to which the truth of R raises the probability of C by application of Bayes' Rule (or related measures, e.g. Good, 1950):

$$P(C|R) \propto P(R|C)P(C) \tag{2}$$

A tacit assumption of this model is that, to a purely rational addressee, the most effective argument should be the one that *maximizes* the contextual probability of the conclusion. Indeed, in explicitly adversarial contexts, listeners may assume that any "better" reasons that the speaker has declined to mention are either false or unknown (analogous to a simple scalar implicature; Barnett et al., 2022; Cummins and Franke, 2021). While these models do predict "diminishing returns" when the listener's prior is already high, they do not predict that excess reasons might actually count *against* the conclusion. Divergent behavior may be attributed to a reasoning error (Hahn and Oaksford, 2007).

Limitations of the standard model. In ordinary conversation, though, it is unclear whether we actually stand to benefit from having the most conclusive evidence for all beliefs in all cases. Indeed, as observed by Grice (1989) and numerous subsequent works (Barker and Taranto, 2003; Fishman and Degen, 2023; Lai, 2012; Romero and Han, 2004; von Fintel and Gillies, 2010), affirming the truth, certainty, evidential justification etc. of one's conversational contribution is always somewhat redundant when Quality is guaranteed. From a prag-

[2]Note that a reason may be mentioned in order to evidentially support a conclusion even if the reason itself is fully redundant in the traditional sense (i.e. commonly believed by all interlocutors), particularly when it may not be readily accessible otherwise due to attention or memory limitations. See Walker (1993) and Crone (2019) for related discussion.

matic perspective, then, the optimal argumentative strategy for *maximizing total information transmission* would be to reserve resource intensive justifications for high-surprisal/low-probability content, where efficient methods (assertion, abstention) would not suffice to secure agreement. Therefore, rational speakers should provide reasons more often when the listener is predicted to disagree, and omit their reasons more often when the listener would probably believe the conclusion in any case.

Following Kravtchenko and Demberg (2022a,b), we propose that rational listeners who expect this production preference could interpret seemingly excessive reasoning by revising their model of the context, such that the prior probability of C would be low enough to necessitate the argument. But unlike the assertoric case described by Kravtchenko and Demberg (2022a,b), in the argumentative case we describe, the listener's posterior belief is derived by conditioning his prior on new evidence, not just by agreeing with the speaker's belief. Therefore, any inference that reduces the contextual prior may also adversely impact the listener's posterior credence. In sum, when the speaker "protests too much", the listener may infer that something is amiss that warrants such protestations.

3 The Present Study

The objective of our study is to determine whether listeners make inferences to rationalize *evidentially favorable* but *pragmatically unnecessary* reasons. We compare the predictions of two probabilistic models: a first-order Bayesian **Evidential Model**, and our novel **Pragmatic Model**, which embeds a simple Bayesian inference within a higher-order social inference about the speaker's choice to provide evidence. We compare the models' predictions with respect to two research questions:

(Q1) Do listeners rationalize seemingly unnecessary reasons by inferring that the speaker expected a disagreement (i.e. unfavorable prior)?

(Q2) If so, does the inference described in (Q1) impact the listener's posterior credence?

Reasons for action. Reasons are used in dialogue to justify any type of conclusion on which interlocutors wish to coordinate: facts, preferences, actions, emotive attitudes, etc. (Mercier and Sperber, 2017; Winterstein, 2018). We focus on a case study wherein speakers provide reasons to justify their actions (*"I did A because* $R_0 ... R_n$*"*). In this case, reasons constitute evidence for the action's normative goodness (deliberative or deontic status, Alvarez and Way, 2024; Kearns and Star, 2009). Thus, we will consider the reasons to be effective if they cause the listener to believe that the action defended by the speaker is *better* or *more favorable*[3].

Notation. Let $\mathcal{V} : A \times W \rightarrow \mathbb{R}$ be a decision-theoretic value function, which accepts an action $a \in A$ and a possible world $w \in W$, and returns a scalar value $v \in V$, representing how good or favorable action a would be if the true state of the world were w. The real numbers $\mathbb{R}$ comprise the scale of values an action may take at any given world, with the best actions approaching a value v of ∞, and the worst actions approaching a value v of $-\infty$ (e.g. Jeffrey, 1983).

Let π be a probability density function representing an agent's prior over v, the goodness of the action described by the speaker. When π places more probability density on low values of v, the agent has an unfavorable opinion. He believes that the actual world is most probably one in which $\mathcal{V}$ assigns a low value v to the action a. Conversely, when π places more probability density on *high* values of v, the agent has a favorable opinion (Fig. 1). See also Achimova et al. (2025) for further discussion on modeling opinion states in this way.

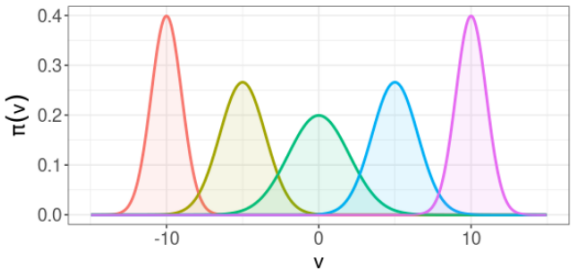

Figure 1: Some possible priors π, ranging from very unfavorable (red) to very favorable (purple).

Let θ be a minimum "goodness" threshold, above which the action is considered "justified" in the given context. The speaker produces an utterance u aiming to convince the listener that $v \geq \theta$. Her utterance u is either a non-empty set of reasons R, or $\emptyset$ (when she forgoes the reasons entirely).

We now introduce the logic of each model from a theoretical perspective; please refer to Appendix A for more implementation details.

[3]Our basic logic is easily adapted to suit reasons for other types of conclusion, but our experiment focuses on actions.

4 Evidential Model: "More is more"

We begin with the Evidential Model, which follows a standard Bayesian approach to argument strength. An evidential listener updates his prior over v (goodness of the action) in response to utterance u (either a non-empty set of positive reasons R, or $\emptyset$) via simple Bayesian conditionalization on the available evidence. We use the notation L_0^π for an evidential listener whose prior is π.

$$p_{L_0^\pi}(v|u) \propto p(u|v)\pi(v) \tag{3}$$

Since an evidential listener does not reason pragmatically about *why* the speaker has chosen to provide or omit reasons, his model of the prior is unaffected by her utterance choice (Q1). And as long as the speaker's reasons R raise the probability density function via Bayes' theorem, the listener will rate the action *more* favorably after hearing R (Q2).

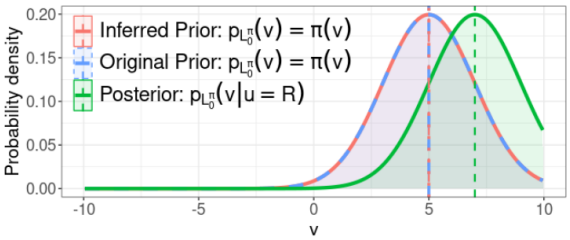

Figure 2: The **Evidential Model** predicts that an utterance of favorable reasons will not affect the listener's model of the prior, and will thus increase his posterior.

5 Pragmatic Model: "Less is more"

Our novel model predicts that reasons which *evidentially* count in favor of a conclusion might still *pragmatically* count against. We adapt the Rational Speech Act framework (Frank and Goodman, 2012) to formalize the pragmatic reasoning.

The RSA framework. We will briefly introduce the three main components of the RSA framework before discussing how they can be adjusted to suit our argumentative scenario. See Degen (2023) for a detailed introduction.

The core intuition of the model is that pragmatic speakers take into consideration the interpretive process of a listener when selecting the optimal utterance u to convey meaning m, and pragmatic listeners likewise consider the production preferences of the speaker when inferring which meaning m she intended by utterance u.

However, to prevent this "back-and-forth" reasoning between speaker and listener from running unchecked, RSA models are grounded in a third agent, a hypothetical "literal" listener L_0, who performs no pragmatic reasoning. When he hears an utterance u, he simply eliminates all possibilities that fail to satisfy the truth conditions of u (where $\delta_{m \in [\![u]\!]} = 0$) and weighs all the remaining options (where $\delta_{m \in [\![u]\!]} = 1$) according to his prior:

$$P_{L_0}(m|u) \propto \delta_{m \in [\![u]\!]} \cdot P_{L_0}(m) \tag{4}$$

The pragmatic speaker S_1 selects her utterance u to minimize the literal listener's surprisal at her intended meaning m, while also minimizing her production cost (i.e. conveying as much Shannon information as possible with minimal resources, as per section 2). The *utility* of an utterance, therefore, is a tradeoff between its informativity and cost:

$$\mathbb{U}_{S_1}(u; m) = \ln P_{L_0}(m|u) - \text{cost}(u) \tag{5}$$

Rather than deterministically selecting the highest utility utterance every time, though, the pragmatic speaker employs a softmax choice instead. The temperature parameter α represents the extent to which she maximizes her utility (if $\alpha = 0$, she chooses randomly, and as $\alpha \to \infty$, she always chooses u with the highest utility).

$$P_{S_1}(u|m) = \text{softmax}(\alpha \cdot \mathbb{U}_{S_1}(u; m)) \tag{6}$$

Ultimately, the pragmatic listener L_1 infers which m the speaker intended (out of a set of relevant alternatives), taking into consideration both his independent prior beliefs and the likelihood of S_1 selecting u to convey m to L_0:

$$P_{L_1}(m|u) \propto P_{S_1}(u|m) \cdot P_{L_1}(m) \tag{7}$$

We will now specify each of the three RSA agents for our Pragmatic Model of argumentation.

Literal listener. First, we adjust the model to represent belief changes in response to evidence, rather than truth-conditional meaning. Rather than updating his prior with the literal semantics of the speaker's utterance, our literal listener should behave exactly like the Evidential Listener described in (3), by simply conditioning on the proffered evidence with no further pragmatic reasoning.

The posterior probability assigned by L_0^π to $v \geq \theta$ (i.e. that the action is at least as good as some threshold) can thus be computed as:

$$p_{L_0^\pi}(v \geq \theta|u) = \int_\theta^\infty p_{L_0^\pi}(v|u)\, dv \tag{8}$$

Pragmatic speaker. Following the usual RSA setup, our pragmatic speaker S_1 selects her utterances according to the economy principle outlined

in section 2. She expends more resources on content that she thinks will be informative or unpredictable to the listener, and less resources on content she thinks he will already believe. In our case, the belief she aims to induce is is $v \geq \theta$.

If L_0^π's prior assigns low probability to $v \geq \theta$ as per (8), his belief in $v \geq \theta$ can substantially *increase* (his surprisal can substantially *decrease*) after hearing the speaker's reasoning. Thus, the informativity of the reasons suffices to offset their cost. However, if L_0^π is *already* confident that $v \geq \theta$, the speaker has little to gain from providing her reasoning, and can conserve resources by omitting it. The utility of u is given as:

$$\mathbb{U}_{S_1}(u;\theta,\pi) = \ln p_{L_0^\pi}(v \geq \theta|u) - \text{cost}(u) \quad (9)$$

Note that providing reasons will still have high utility, even when the listener has a generally favorable prior opinion, if the relevant threshold θ is also high (for example, an item must be highly desirable to justify a large expense, but need only be moderately desirable to justify a small expense).

S_1 selects an utterance by soft-maximizing her utility, given her desired threshold θ and the prior belief π she *expects* the listener to have.

$$P_{S_1}(u|\theta,\pi) = \text{softmax}[\alpha \cdot \mathbb{U}_{S_1}(u;\theta,\pi)] \quad (10)$$

Pragmatic listener. By assuming that the speaker has chosen her utterance "rationally" (as per Eq. 10), a pragmatic listener L_1 can infer which distribution π she is treating as his prior. In other words, he can infer whether the speaker expected him to agree or disagree with her. Applying Bayes' theorem, L_1 jointly infers π and θ based on his priors over each of those, and the likelihood $P_{S_1}(u|\theta,\pi)$ of S_1 producing utterance u for a given π and θ.

$$p_{L_1}(\theta,\pi|u) \propto P_{S_1}(u|\theta,\pi)\, p(\theta)\, p(\pi) \quad (11)$$

L_1 can isolate the speaker's model of his prior by marginalizing over all possible thresholds θ:

$$p_{L_1}(\pi|u) \propto \int_{-\infty}^{\infty} p_{L_1}(\theta,\pi|u)\, d\theta \quad (12)$$

Since the speaker tends to give more reasons when she expects L_0^π to have an unfavorable prior, L_1 tends to infer lower π's upon hearing more reasons. The speaker may have anticipated a disagreement, he thinks, since she found it worthwhile to defend herself despite the added cost (Q1).

Finally, L_1 updates his posterior belief over v, the goodness of the action, in response to the evidence provided (or indeed, not provided) by the speaker. Here, he factors in his actual prior *and* the prior he just inferred based on the speaker's utterance (i.e. π). Conditionalization from his *original* prior is another simple Bayesian inference:

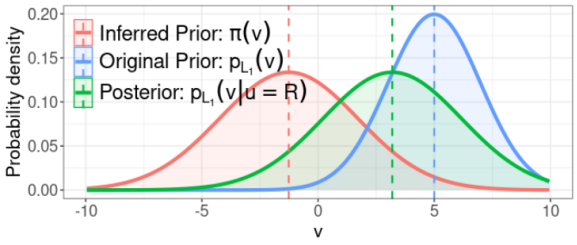

Figure 3: The **Pragmatic Model** predicts that an utterance of favorable reasons may reduce listener's model of the prior, and thus reduce his posterior too, depending on parameter selection (cost, λ, etc). See Appendix A for further detail.

$$p_{L_1^{\text{orig}}}(v|u) \propto p(u|v)\, p_{L_1}(v) \quad (13)$$

Conditionalization from the *inferred* prior is calculated by applying Bayes' theorem from each possible π , weighted by $p_{L_1}(\pi|u)$, the probability the listener assigns to that particular π (e.g. Eq. 12):

$$p_{L_1^\pi}(v|u) \propto \int_\pi p(u|v)\, \pi(v) \cdot p_{L_1}(\pi|u)\, d\pi \quad (14)$$

If the pragmatic listener makes a weighted tradeoff between (13) and (14), his overall posterior is:

$$p_{L_1}(v|u,\lambda) \propto (1-\lambda)\, p_{L_1^{\text{orig}}}(v|u) + \lambda\, p_{L_1^\pi}(v|u) \quad (15)$$

Here, the weight parameter λ determines how much L_1 weighs his original opinion vs. the opinion he thinks the speaker *expected* him to have, when evaluating the evidence. If the listener hears a costly justification for an action he would otherwise have found perfectly acceptable, as long as he places enough consideration on the inferred prior (which would be much lower than his original prior), the evidential strength of the reasons may not be enough to "recoup the loss." Or, in plain terms: once the listener becomes aware that there may be serious grounds for disagreement which he hadn't previously considered, the available evidence may not be enough to convince him anymore. In that case, a verbose justification may result in a lower posterior judgment than no justification at all (Q2, see Fig. 3).

6 Materials & Methods

Study design. We conducted a behavioral study to adjudicate between the models. The experiment utilized a 2x2 Latin square design, crossing utterance type $\in$ {NO-REASONS, REASONS} with stakes $\in$ {LOW-STAKES, HIGH-STAKES}. We included

the stakes manipulation to test whether an inference arises primarily when the reasons seem *unnecessary*. Utterance type was manipulated within items; stakes was manipulated between items.

Materials. Target stimuli consisted of text message conversations between two named characters, since prior studies show that informativity inferences arise more readily when the speaker is salient (Reksnes et al., 2024). For consistency, the speaker character's name started with A, and the listener character's name started with B. The speaker character was always the "incoming"/grey-bubble character, and the listener was the "outgoing"/blue-bubble character, allowing the participant to assume the perspective of the listener (Fig. 4).

In the NO-REASONS condition, the speaker simply mentions an action. In the REASONS condition, she offers four different reasons. We opted for these verbose justifications to ensure that their production cost would be high enough to trigger an inference. Each set of reasons was designed to be internally consistent, and to strongly support the action (to avoid a confound with the "weak evidence effect", e.g. Barnett et al., 2022). Items were approximately the same length overall, with minor variability to maintain naturalism.

In the LOW-STAKES condition, the speaker discusses an innocuous or day-to-day action (e.g. ordering takeout), and provides the reasons out of the blue. In the HIGH-STAKES condition, the action involves a more complex or significant decision (e.g. moving to a new apartment), where the requisite threshold value θ to justify the action would be comparatively higher, and the reasons are provided in response to indirect probing from character B. Our goal was to distinguish between contexts where multiple reasons feels excessive, and contexts where multiple reasons feels necessary.

On each of the 16 target trials, participants rated the following on a 7-point Likert scale:

RATING1: A expected B to feel ["very unfavorably" - "very favorably] about [action].

RATING2: Now, B will feel ["very unfavorably" - "very favorably"] about [action].

The first rating probes Q1: do seemingly unnecessary reasons generate an inference that speaker expected the listener to have an unfavorable prior? The second rating probes Q2: does recognition of that expectation undercut persuasiveness?

Participants also saw 4 attention checks, where they were asked to responded to rating scale questions about explicitly stated opinions. Refer to Appendix B for sample trials and attention checks.

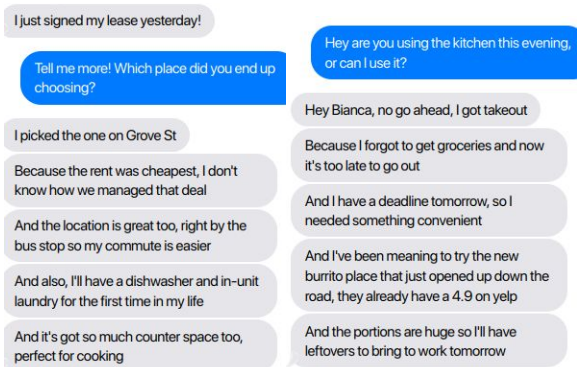

Figure 4: Sample dialogue in the REASONS condition, for HIGH-STAKES (left) and LOW-STAKES (right) actions

Participants. 60 native English speakers from the US or UK were recruited on Prolific for the study. All participants passed the attention checks, so they were all included in the analysis. The survey took a median of 17 minutes, 52 seconds to complete, and participants were compensated £5.00.

Procedure. The survey was hosted online on Qualtrics. Participants first read an information sheet and consented to participation, and confirmed their Prolific ID and native-speaker status. They were then instructed to read each set of messages closely and reason about the characters' beliefs.

Participants were randomly assigned to one of two stimulus lists, with 4 trials in each of the 4 conditions. Each item appeared only once per list. With 4 attention checks, participants saw a total of 20 trials in random order. Each trial appeared on a separate page, and participants were required to complete both ratings before advancing. At the end of the survey, participants were asked to provide feedback and guess what the experiment was about.

7 Results

Model selection. Analysis for the Likert rating questions was conducted with a Bayesian cumulative logit model, as is best practice for ordinal data (Bürkner and Vuorre, 2019). We used the R (R Core Team, 2021) package `brms` (Bürkner, 2017) with default priors. The fixed effects were utterance type, stakes, and their interaction, treating NO-REASONS and LOW-STAKES as the baselines, with random effects for PARTICIPANT and ITEM.

Q1: Reduction of the prior. The first Likert rating (*A expects B to feel...*) targets Q1: whether participants would rationalize the speaker's choice

to provide detailed justification by inferring that she expected the the listener to disapprove otherwise. Responses are plotted in Figure 5. The model identified a credible main effect of REASONS (β = -0.77, 95% CI -1.09, -0.43), indicating that the inferred prior was *lower* in the REASONS condition for LOW-STAKES actions. A credible interaction of REASONS with HIGH-STAKES (β = 0.83, 95% CI 0.35, 1.30) reversed the main effect entirely.

These results support the Pragmatic Model, which predicts that a speaker's production of seemingly unnecessary reasons signals that she expected the listener to have an unfavorable prior (i.e. he would need to be *convinced* to share her opinion). No inference is required when reasons are necessary to obtain the relevant threshold. The Evidential Model instead predicts no effect for either type of action, since the listener does not reason about the speaker's choice to justify in either case.

Q2: Reduction of the posterior. The second rating (*Now, B will feel...*) targets Q2: whether higher-order inferences about *why the speaker felt that reasons were necessary* would mitigate their effectiveness in persuading the listener. Responses are plotted in Figure 6. There was no credible effect of REASONS for LOW-STAKES items. However, there was a credible interaction between REASONS and HIGH-STAKES (β = 0.75, 95% CI 0.28, 1.22), meaning that the reasons improved the posterior favorability ratings *only for high-stakes items*. Further, there was no credible main effect of stakes, indicating that LOW- and HIGH-STAKES items were rated about the same when reasons were absent.

Listeners' posterior favorability ratings were never credibly *lower* in the presence of reasons. Since there is no clear evidence that verbose argumentation can be *less* persuasive than no argumentation whatsoever (which only the Pragmatic Model can predict qualitatively), we cannot conclusively rule out the Evidential Model for Q2. Nevertheless, the results seem amenable to a pragmatic analysis. While HIGH and LOW-STAKES items received similar ratings with no reasons, only the HIGH-STAKES decisions were judged more favorably with reasons. Conversely, participants judged the LOW-STAKES actions about the same when given multiple positive reasons, and when given no reasons at all.

Since no reduction of the prior was observed for HIGH-STAKES items, the evidential strength of the reasons yielded a net increase in the posterior. By contrast, since a reduction of the prior *was* observed for LOW-STAKES items, the evidential strength of

Figure 5: Responses to Q1. Prior favorability ratings reduced when reasons were provided for low-stakes actions. The effect was entirely reversed in the interaction with high-stakes.

Figure 6: Responses to Q2. Posterior favorability ratings were unchanged when reasons were provided for low-stakes actions. However, ratings credibly increased in the interaction with high-stakes.

the reasons might have been sufficient to "recoup the loss", but insufficient to further improve the posterior. In other words, the evidential and pragmatic reasoning may have effectively canceled out, resulting in no net gain to the posterior for LOW-STAKES items. See Appendix C for detailed results.

8 Discussion

In this study, we set out to explain why unnecessary argumentation sometimes seems suspicious. This intuition seems contrary to basic principles of rational belief revision, which predict that agents should only become *more* confident of a conclusion, when they have access to more evidence in favor of it. We proposed that this effect might be produced by social reasoning about the expected behavior of an efficient communicator. Adapting the RSA framework, we provided a pragmatic account of the "protest too much" effect, and tested the predictions of our model against a first-order Bayesian model. The study confirmed our prediction that seemingly unnecessary reasoning can signal to a listener that the speaker expected a disagreement, and suggests that this inference might diminish the

persuasiveness of the reasons as well.

However, the inference we observed from redundant reasons was not strong enough to result in a net *reduction* of the posterior, which our RSA model predicts is possible under certain circumstances. Furthermore, there are surely many other factors which could prevent a protest-too-much inference from arising, and many other inferences besides the one we investigated which might serve to "rationalize" superfluous arguments. Here we will outline some avenues for further study, which could address these open issues.

Binary vs. scalar utterance choice. In this initial study, we simplified the set of utterances a speaker could provide to a binary choice (i.e. either an argument, or no argument). This simplification may have prevented us from observing a robust "non-monotonicity" effect in the listener's posterior credence. As suggested by some participants in the free-text response, it may be difficult to make an explicit value judgment of an action in an experimental context with no additional information (even if that action would not require justification in a casual conversation). If the NO-REASONS condition was actually perceived as *under*informative, the "optimal" number of reasons might be somewhere between 1 and 4, and not 0. A larger follow-on study testing a range of argument lengths might elicit the "non-monotonicity" effect that did not conclusively arise in this smaller experiment.

Speaker-specific adaptations. The stakes manipulation that we included in this experiment, which suppressed the informativity inference as predicted, is only one of many manipulations which might prevent listeners from making an inference in response to extensive reasoning from a speaker. Another possibility, of course, is that some people are just more talkative than others. Listeners are frequently sensitive to the production preferences of individual speakers (Schuster and Degen, 2020), not just general expectations for "rationality." In sentence completion tasks conducted by Reksnes et al. (2024), participants anticipated "over-informative" continuations more frequently from speakers whose overall style was more redundant. So for a particularly loquacious speaker, an inference may not be required to make sense of a verbose justification.

In addition to the speaker's individual attributes, we expect that listeners may also be sensitive to social factors. For example, Beltrama and Schwarz (2025) have shown that sociolinguistic *personae* such as "Nerdy" (associated with precise speech) and "Chill" (associated with imprecise speech) can affect pragmatic interpretation of round numbers; it seems plausible that similar factors may be at play when interpreting discourses with meticulously granular reasoning. Future studies should also consider the effects of *testimonial injustice* (Fricker, 2007), the systematic discounting of evidence produced by certain social groups. Indeed, McCready and Winterstein (2019) show experimentally that evidence produced by women may be perceived as less reliable than evidence produced by men.

"Virtue-signaling" and other inferences. In their responses to the free-text question, two participants mentioned that the speaker sometimes seemed to be "virtue-signaling"; these comments draw attention to another inference which might serve to rationalize superfluous argumentation. The speaker might have chosen to justify her action because she hopes to be perceived as *extraordinarily* virtuous, diligent, etc., and not (just) because she fears she may be perceived as *insufficiently* reasonable otherwise. There are surely many more inferences that might arise in response to fluctuations in information rate, which may be identified in future studies.

9 Conclusion

The cooperative principle has long been essential to the study of pragmatics, but in recent years, there has been a surge of interest in expanding Grice's insights to mixed-motive and noncooperative communication (Cummins, 2025). Much of this work investigates whether and why "classic" inferences arise in low-trust contexts (where the status of the cooperative principle is unclear; Asher and Lascarides, 2013), and how speakers utilize implicature, vagueness, presupposition, etc. to pursue adversarial goals (Cummins and Franke, 2021; Franke et al., 2020; Meibauer, 2014). Our work contributes a new perspective to this growing literature, demonstrating how pragmatic models might be harnessed to explain intuitive judgments that cannot be accounted for by other theories of trust and rationality.

Limitations

Our study faces a number of limitations in terms of its theoretical claims, empirical methods, and modeling choices. As discussed in section 8, the inference we describe is surely defeasible by numerous contextual factors, and our findings are sig-

nificantly constrained by the "verbose argument" vs "no argument" experimental setup. A larger study with additional conditions and participants is required to support our hypothesis more conclusively. Finally, a more robust quantitative fit of our model to the empirical data will require further norming studies to estimate, e.g., the likelihood of each individual reason used in each stimulus, and $p(\theta)$, the listener's prior over the justification threshold for each action.

Acknowledgments

Many thanks to the three anonymous reviewers and Kajsa Djärv for helpful comments and suggestions.

References

Asya Achimova, Michael Franke, and Martin V. Butz. 2025. The alignment model of indirect communication. *PLOS ONE*, 20(5):1–33.

Maria Alvarez and Jonathan Way. 2024. Reasons for Action: Justification, Motivation, Explanation. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2024 edition. Metaphysics Research Lab, Stanford University.

Jean-Claude Anscombre and Oswald Ducrot. 1983. *L'argumentation dans la langue*. Pierre Mardaga, Bruxelles.

Nicholas Asher and Alex Lascarides. 2013. Strategic conversation. *Semantics and Pragmatics*, 6(2):1–62.

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31–56.

Chris Barker and Gina Taranto. 2003. The paradox of asserting clarity. In *Proceedings of the Western Conference on Linguistics (WECOL) 2002*, volume 14, pages 10–21, Fresno, CA. Department of Linguistics, California State University.

Samuel A. Barnett, Thomas L. Griffiths, and Robert D. Hawkins. 2022. A pragmatic account of the weak evidence effect. *Open Mind*, 6:169–182.

Alan Bell, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory, and Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, 113(2):1001–1024.

Andrea Beltrama and Florian Schwarz. 2025. (im)precise personae: The effect of socio-indexical information on semantic interpretation. *Language in Society*, 54(4):787–814.

Leon Bergen and Noah D. Goodman. 2015. The strategic use of noise in pragmatic reasoning. *Topics in Cognitive Science*, 7(2):336–350.

Claire Augusta Bergey, Benjamin Morris, and Daniel Yurovsky. 2020. Children hear more about what is atypical than what is typical. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Paul-Christian Bürkner. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28.

Paul-Christian Bürkner and Matti Vuorre. 2019. Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, 2(1):77–101.

Phil Crone. 2019. Assertions of clarity and raising awareness. *Journal of Semantics*, 36(1):53–97.

Chris Cummins. 2025. Noncooperative communication. *Annual Review of Linguistics*, 11(Volume 11, 2025):35–52.

Chris Cummins and Michael Franke. 2021. Rational interpretation of numerical quantity in argumentative contexts. *Frontiers in Communication*, Volume 6 - 2021.

Judith Degen. 2023. The rational speech act framework. *Annual Review of Linguistics*, 9(Volume 9, 2023):519–540.

Judith Degen, Robert D. Hawkins, Caroline Graf, Elisa Kreiss, and Noah D. Goodman. 2020. When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review*, 127(4):591–621.

Paul E. Engelhardt, Karl G.D. Bailey, and Fernanda Ferreira. 2006. Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, 54(4):554–573.

Alon Fishman and Judith Degen. 2023. Evidential uncertainty involves both pragmatic and extralinguistic reasoning: a computational account. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Michael Franke, Giulio Dulcinati, and Nausicaa Pouscoulous. 2020. Strategies of deception: Underinformativity, uninformativity, and lies—misleading with different kinds of implicature. *Topics in Cognitive Science*, 12(2):583–607.

Miranda Fricker. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

David Godden and Frank Zenker. 2018. A probabilistic analysis of argument cogency. *Synthese*, 195(4):1715–1740.

I. J. Good. 1950. *Probability and the Weighing of Evidence*. Griffin, London.

H. Paul Grice. 1989. Logic and conversation. In *Studies in the Way of Words*, pages 22–40. Harvard University Press, Cambridge, MA.

Ulrike Hahn and Mike Oaksford. 2007. The rationality of informal argumentation: A bayesian approach to reasoning fallacies. *Psychological Review*, 114(3):704–732.

Hailin Hao, Muxuan He, and Zuzanna Fuchs. 2024. Greta is a female director: When gender stereotypes interact with informativity expectations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Daphna Heller, Kristen S. Gorman, and Michael K. Tanenhaus. 2012. To name or to describe: Shared knowledge affects referential form. *Topics in Cognitive Science*, 4(2):290–305.

T. Florian Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23–62.

Richard C. Jeffrey. 1983. *The Logic of Decision*, 2 edition. University of Chicago Press, Chicago.

Stephen Kearns and Daniel Star. 2009. Reasons as evidence. In Russ Shafer-Landau, editor, *Oxford Studies in Metaethics*, volume 4, pages 215–242. Oxford University Press.

Ekaterina Kravtchenko and Vera Demberg. 2022a. Informationally redundant utterances elicit pragmatic inferences. *Cognition*, 225:105159.

Ekaterina Kravtchenko and Vera Demberg. 2022b. Modeling atypicality inferences in pragmatic reasoning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 44(44).

Catherine Lai. 2012. *Rises All the Way Up: The Interpretation of Prosody, Discourse Attitudes and Dialogue Structure*. Phd dissertation, University of Pennsylvania, Philadelphia, PA.

Robin Lemke, Ingo Reich, Lisa Schäfer, and Heiner Drenhaus. 2021. Predictable words are more likely to be omitted in fragments–evidence from production data. *Frontiers in Psychology*, 12:662125.

Neil Levy. 2021. *Bad Beliefs: Why They Happen to Good People*. Oxford University Press.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Roger Levy and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Kyle Mahowald, Evelina Fedorenko, Steven T. Piantadosi, and Edward Gibson. 2013. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318.

Elin McCready. 2015. *Reliability in pragmatics*. Oxford studies in semantics and pragmatics ; 4. Oxford University Press, Oxford.

Elin McCready and Grégoire Winterstein. 2019. Testing epistemic injustice. *Investigationes Linguisticae*, 41:86–104.

Jörg Meibauer. 2014. *Lying at the Semantics-Pragmatics Interface*, volume 14 of *Mouton Series in Pragmatics*. De Gruyter Mouton, Berlin and Boston.

Hugo Mercier. 2020. *Not Born Yesterday*. Princeton University Press, Princeton.

Hugo Mercier and Dan Sperber. 2017. *The Enigma of Reason*. Harvard University Press, Cambridge, MA and London, England.

Arthur Merin. 1999. *Information, relevance, and social decisionmaking: some principles and results of decision-theoretic semantics*, page 179–221. Center for the Study of Language and Information, USA.

Mike Oaksford and Ulrike Hahn. 2004. A bayesian approach to the argument from ignorance. *Canadian Journal of Experimental Psychology*, 58(2):75–85.

Lauren A. Oey, Adena Schachner, and Edward Vul. 2023. Designing and detecting lies by reasoning about other agents. *Journal of Experimental Psychology: General*, 152(2):346–362.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2011. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529.

Mark Pluymaekers, Mirjam Ernestus, and R. Harald Baayen. 2005. Lexical frequency and acoustic reduction in spoken dutch. *The Journal of the Acoustical Society of America*, 118(4):2561–2569.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Vilde R. S. Reksnes, Alice Rees, Chris Cummins, and Hannah Rohde. 2024. *Tell me something I don't know: Speaker salience and style affect comprehenders' expectations for informativity*, pages 177–214. Topics at the Grammar-Discourse Interface. Language Science Press.

Hannah Rohde, Richard Futrell, and Christopher G. Lucas. 2021. What's new? a comprehension bias in favor of informativity. *Cognition*, 209:104491.

Hannah Rohde and Paula Rubio-Fernandez. 2022. Color interpretation is guided by informativity expectations, not by world knowledge about colors. *Journal of Memory and Language*, 127:104371.

Maribel Romero and Chung-Hye Han. 2004. On negative yes/no questions. *Linguistics and Philosophy*, 27(5):609–658.

Paula Rubio-Fernandez. 2016. How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in Psychology*, 7:153.

Sebastian Schuster and Judith Degen. 2020. I know what you're probably going to say: Listener adaptation to variable use of uncertainty expressions. *Cognition*, 203:104346.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Dan Sperber, Fabrice Clement, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. 2010. Epistemic vigilance. *Mind & Language*, 25(4):359–393.

Robert C. Stalnaker. 1978. Assertion. In Peter Cole, editor, *Syntax and Semantics 9: Pragmatics*, pages 315–332. Academic Press.

Tamina Stephenson. 2007. Judge dependence, epistemic modals, and predicates of personal taste. *Linguistics and Philosophy*, 30(4):487–525.

Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of discourse relations. In *Proceedings of COLING 2012*, pages 2669–2684, Mumbai, India. The COLING 2012 Organizing Committee.

Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. 2014. *Handbook of Argumentation Theory*. Springer, Dordrecht.

Kai von Fintel and Anthony S. Gillies. 2010. Must . . . stay . . . strong! *Natural Language Semantics*, 18(4):351–383.

Marilyn A. Walker. 1993. Informational redundancy and resource bounds in dialogue. In *Working Notes of the AAAI Spring Symposium on Reasoning about Mental States: Formal Theories and Applications (SS-93-05)*, pages 160–169, Palo Alto, CA. AAAI.

Grégoire Winterstein. 2018. A bayesian approach to argumentation within language. Manuscript.

George Kingsley Zipf. 1935. *The Psychobiology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin, Boston.

A Appendix: Model implementation

In section 5, we introduced the logic of our Rational Speech Act model from a theoretical perspective. Here, we will provide some further details on how we generated the simulations.

Speaker utterance choice. While the basic logic of our model places no restrictions on the set of reasons the speaker might produce, we ran our simulations to match the constraints of our experimental setup, where only two utterances are possible: NO-REASONS ($u = \emptyset$), and REASONS ($u = R$), where R represents the set of four reasons. In realistic dialogues, intermediate arguments are also possible. Our model is equipped to handle those cases too, once we have the empirical data to match.

Priors. The literal listener's prior π over v may, conceivably, be any probability density function. To make our simulations tractable, though, we restricted the set of possible priors to Gaussian distributions over v, as parameterized by mean μ and standard deviation σ, i.e. $\pi(v) = \mathcal{N}(v; \mu, \sigma^2)$. Gaussian distributions can reasonably approximate agents' opinion states: the mean represents the expected value v assigned by the agent to an action (i.e. his overall judgment), and the standard deviation represents how confidently he holds that opinion. Figure 1 shows some possible Gaussian priors, representing a range of opinions from (approximately) "very unfavorable" to "very favorable". We modeled the continuous and unbounded Gaussian distributions using a grid approximation.

Likelihood. Since the purpose of our study is to show that *evidentially* favorable reasons might still backfire due to a pragmatic inference, we simulated the Bayesian likelihood $p(R|v)$ using an exponential function $e^{v/2}$. This likelihood function always raises the posterior from a Gaussian prior, thus representing the effect of a "positive" or "favorable" reason. While this approach delivers the correct qualitative behavior, a norming study that empirically estimates the likelihood for each individual reason is required for a true quantitative fit of our model to the data.

Evidential/literal listener. Figure 2 shows a simple Bayesian update for a literal listener whose prior is a Gaussian distribution with $\mu = 5$ and $\sigma = 2$. His opinion becomes *more* favorable after hearing the reasons, even though his prior was already fairly favorable.

Pragmatic speaker. Figure 7 compares the speaker utility $\mathbb{U}_{S_1}$ of each utterance, R and $\emptyset$. The utility of each utterance depends crucially on what the speaker takes to be the listener's prior $\pi(v) = \mathcal{N}(v; \mu, \sigma^2)$. When the listener presumably has an unfavorable prior, R has high utility, since the reasons' high informativity offsets their cost. When the listener presumably *already* has a favorable

prior, the reasons are not informative enough to be worth the added cost, and $\emptyset$ has higher utility.

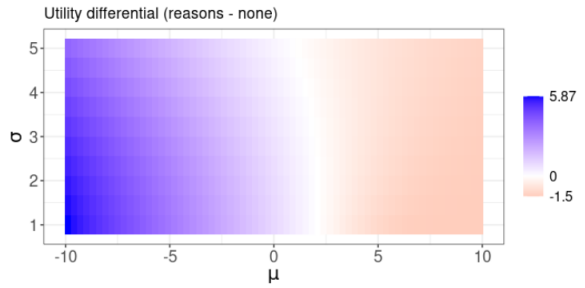

Figure 7: Comparing the utility of R and $\emptyset$, for each possible prior the listener might have (with each π modeled as a Gaussian parameterized by μ and σ). The reasons utterance has a higher utility when the listener has a lower prior (blue). The null utterance has a higher utility when the listener has a higher prior (red).

Figure 7 plots the utilities given a threshold value $\theta = 3$, an utterance cost of 1.5 for the reasons utterance, and a softmax rationality parameter $\alpha = 1$. Bear in mind that $u = R$ will "survive" at higher priors when the cost is lower, or when the threshold is higher. Likewise, $u = \emptyset$ will overtake the reasons at lower priors when the cost is higher, or when the threshold is lower.

Pragmatic listener. Here, the pragmatic listener infers which prior $\pi(v) = \mathcal{N}(v; \mu, \sigma^2)$ the speaker *expected* him to have. Figure 8 shows the probability assigned by the pragmatic listener to each possible π, as computed by Eq. 11 and 12. For this computation, we selected a uniform distribution for $p(\theta)$, the prior over θ. We decomposed the hyperprior over π into $p(\mu) \cdot p(\sigma)$. Since we used $\mu = 5$ and $\sigma = 2$ for the listener's "original" prior, we set $p(\mu) = \mathcal{N}(5, 10^2)$ and $p(\sigma) = \mathcal{N}(2, 2^2)$ for the hyperprior $p(\pi)$.

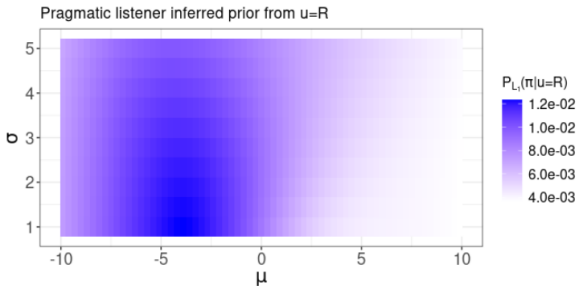

Figure 8: Assuming that the speaker has selected her utterance "rationally", the pragmatic listener is more likely to infer a lower π than a higher π upon hearing verbose reasoning. This inference is constrained by his priors over θ and π.

Recall that the pragmatic listener may take into consideration some combination of his original prior and the pragmatically reduced prior when conditioning on the evidence. In Figure 3, we chose $\lambda = 1$, thus placing maximal weight on the inferred prior. Thus, the reasons fail to recover from the pragmatic inference. However, for lower values of λ, the reasons may still be enough to recover the original prior; indeed, when $\lambda = 0$, the Pragmatic Model predicts the same posterior as the Evidential Model.

B Additional materials

First, before sending the texts, did **ABBY** expect **BRITTANY** to view her decision unfavorably, or favorably?

	Very unfavorable	Unfavorable	Somewhat unfavorable	Neither unfavorable nor favorable	Somewhat favorable	Favorable	Very favorable
Abby expected Brittany to feel...	○	○	○	○	○	○	○

Now, will **BRITTANY** think **ABBY** made a good choice?

	Very unfavorable	Unfavorable	Somewhat unfavorable	Neither unfavorable nor favorable	Somewhat favorable	Favorable	Very favorable
Brittany will feel...	○	○	○	○	○	○	○

Figure 9: Sample Likert rating scale questions.

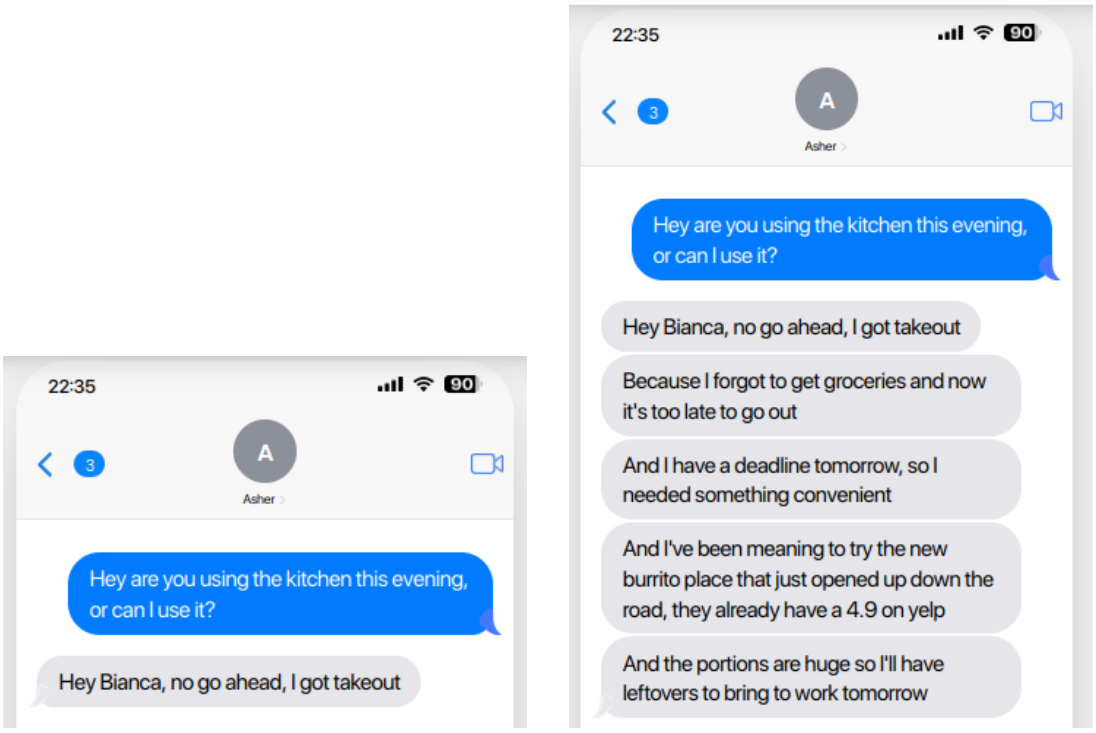

Figure 10: Sample LOW-STAKES context stimuli in NO-REASONS and REASONS conditions.

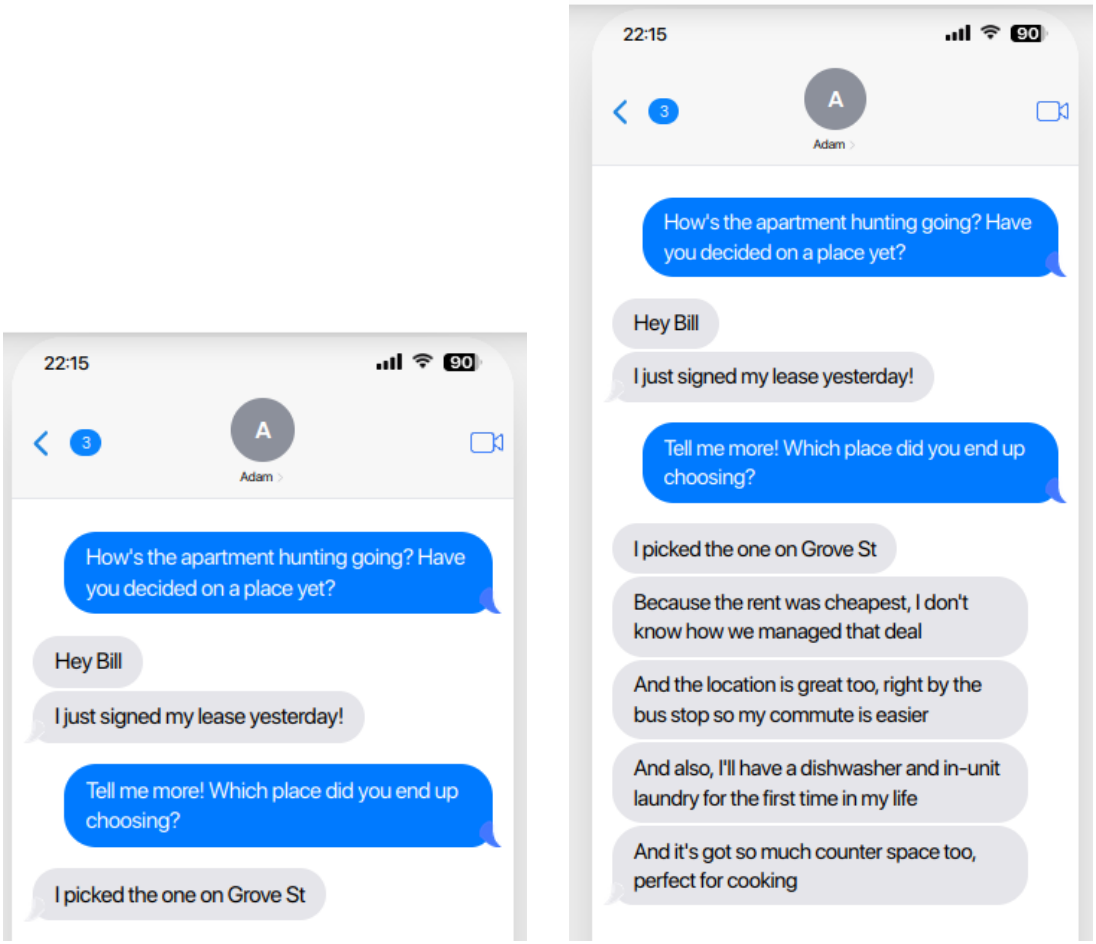

Figure 11: Sample HIGH-STAKES context stimuli in NO-REASONS and REASONS conditions.

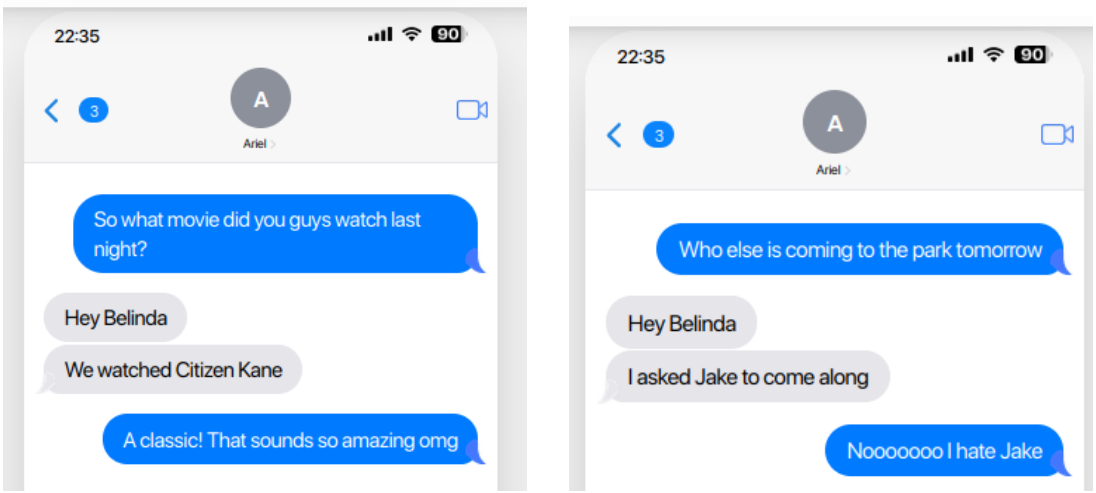

Figure 12: Sample attention checks.

C Full results

Predictor	Estimate	Std. Error	95% CI Lower	95% CI Upper
REASONS	**-0.77**	**0.17**	**-1.09**	**-0.43**
HIGHSTAKES	0.19	0.66	-1.12	1.52
REASONS:HIGHSTAKES	**0.83**	**0.24**	**0.35**	**1.30**

Table 1: Log-odds coefficients from the Bayesian ordinal model of Q1. A **credible** main effect of REASONS, moderated by a **credible** interaction with HIGH-STAKES, is observed.

Predictor	Estimate	Std. Error	95% CI Lower	95% CI Upper
REASONS	0.15	0.17	-0.20	0.50
HIGHSTAKES	0.25	0.76	-1.25	1.79
REASONS:HIGHSTAKES	**0.75**	**0.24**	**0.28**	**1.22**

Table 2: Log-odds coefficients from the Bayesian ordinal model of Q2. A **credible** interaction of REASONS and HIGH-STAKES is observed.