Testing Bayesian measures of relevance in discourse[1]

Alex WARSTADT — *New York University, Department of Linguistics*
Omar AGHA — *New York University, Department of Linguistics*

Abstract. Despite their popularity among linguists, categorical notions of relevance based on partial answerhood have well-known problems. Gradient information-theoretic measures of utility adopted in cognitive modeling, on the other hand, have not been tested as measures of relevance in discourse. We begin to fill this gap by experimentally evaluating two gradient measures of question under discussion (QUD) relevance in question-answer pairs in comparison to the categorical theory: entropy reduction, which measures the degree to which an answer decreases uncertainty about the resolution of the QUD; and KL divergence, which measures the degree to which an answer changes the probability distribution over the alternatives. Our experiments provide decisive evidence against the categorical theory of relevance, but do not give strong support to any one gradient measure. Both KL divergence and entropy reduction have systematic failure modes, and are less predictive of relevance judgments than comparatively unmotivated measures like the difference between prior and posterior. We outline several criteria for an adequate gradient theory of relevance, and identify candidate measures for future investigation.

Keywords: question under discussion (QUD), relevance, Bayesian, entropy, discourse, pragmatics, information gain

1. Introduction

Relevance has played a central role in the formal analysis of discourse structure since Grice (1975) first introduced the Maxim of Relation as a governing principle of pragmatics. The pioneering work of Roberts (2012), alongside others (Groenendijk and Stokhof, 1984; van Dijk, 1995), provided a theory of relevance that is based on informativity relative to the question under discussion (QUD). This paper aims to further this program by experimentally evaluating novel theories of relevance that are informativity-based, like the QUD framework, but are gradient rather than categorical.

In the classical QUD framework, all relevant assertions must be partial answers to the QUD. A response is a partial answer if it is incompatible with at least one alternative in the question's denotation. Because a response either is or is not a partial answer, the partial answer theory delivers only categorical judgments for question-answer pairs. The theory is therefore unable to account for differences between relevant responses. (1) gives an example of three responses, of which all are intuitively relevant, but each differs in the degree to which they address the QUD.

(1) Q: Is it going to rain?
It rained last week. < It's cloudy. < The forecast predicted rain.

A closely related problem is that many intuitively relevant responses are not relevant according

[1] We would like to thank Michael Franke, Vera Klütz, Dan Lassiter, Craige Roberts, and audiences at NYU Semantics Group and Sinn und Bedeutung 26.

to the partial answer theory (Agha and Warstadt, 2020). For example, the response *It's cloudy.* is not a partial answer in (1) because the response is compatible with *all* alternatives in the question: Of the possible worlds in which the weather is cloudy, it rains in some and does not in others. The fact that cloudy weather makes rain *more likely* than sunny weather is not sufficient to make the response relevant under the partial answer theory.

The problems with categorical relevance are well-known. Büring (2003) and Hyska (2015) suggest that relevance be expanded to include responses that shift the probability of alternatives. Van Rooy (2004) first proposed a concrete model of relevance based on entropy reduction, which we evaluate experimentally in this paper. However, while much work has adopted the categorical QUD theory, gradient theories are less well-explored, and rarely put to use in semantics and pragmatics. We hope to reinvigorate research on this topic by performing the first experimental evaluation of gradient theories of relevance.

We compare the categorical theory of relevance to two gradient models based on information-theoretic measures adopted in Bayesian cognitive modeling. These models use these measures to assign a utility value to a response to the extent that it shifts the probabilities of the QUD alternatives in a useful way. However, models differ in how they assign utility values to different updates, and these divergent predictions have not been tested in the context of discourse. Partial answers will still have some degree of relevance in a gradient theory, but (i) many more responses will be relevant, and possibly more or less relevant than partial answers, and (ii) different partial answers will also differ in relevance.

We test two gradient information-theoretic measures: entropy reduction (following Oaksford and Chater, 1994; van Rooy, 2004) measures the degree to which an answer decreases uncertainty about the QUD, and KL divergence (following Hawkins and Goodman, 2017) measures the degree to which an answer shifts the probability distribution over alternatives. In our experiment, we find that a categorical notion of relevance does not reflect participants' behavior. Not only do participants make full use of the relevance scale, but the categories for partially relevant answers suggested by the categorical theory (partial answers and the like) do not correspond to a meaningful class in our data.

We also find that the KL divergence model outperforms entropy reduction, but still misses out on several classes of relevant responses. Finally, while shifting probabilities is an important factor in accounting for relevance judgments, there are other important factors that are not captured by the simple Bayesian model.

2. Limitations of categorical relevance

2.1. Partial answerhood

In the categorical QUD theory of relevance (Roberts, 2012), a response is considered relevant if and only if it is a partial answer to a contextually determined QUD. A partial answer (2b) is any response which eliminates at least one alternative[2] from the QUD (this includes resolving

[2] Here we represent the denotation of a question Q as a partition over possible worlds (Groenendijk and Stokhof, 1984). We refer to each cell in the partition as an **alternative**.

answers as well (2a)). Thus, irrelevant responses are exactly the set of non-eliminating answers, which are consistent with all alternatives in the QUD (2c). Examples of these three answer types are given in (3).

(2) a. **Definition: Resolving answer** (Agha and Warstadt, 2020: 22: (11))
Proposition a is a resolving answer to Q iff $\exists q_{\in Q}[a \subseteq q]$
b. **Definition: Partial answer**
Proposition a is a partial answer to Q iff $\exists q_{\in Q}[a \cap q = \emptyset]$
c. **Definition: Non-eliminating answer**
Proposition a is a non-eliminating answer to Q iff $\forall q_{\in Q}[a \cap q \neq \emptyset]$

(3) Q: *Who (of Jane, Lucy, and Steve) ate the cookies?*
A: *(Only) Jane did.* resolving answer
B: *Jane or Lucy, but not Steve.* partial answer
C: # *Jane ate the cake.* non-eliminating answer

Agha and Warstadt (2020) identified counterexamples to the partial answer theory of relevance. Every example in (4) is intuitively relevant, despite being non-eliminating. Figure 1, makes clear why these are not partial answers: In each case, the response A fails to eliminate either of the alternatives to the polar question Q.

(4) Examples of reductive answers (Agha and Warstadt, 2020: 25: (24-26))
a. Q: *Is John going to Coachella?* A: *Or Lollapalooza.*
b. Q: *Will we cancel the picnic?* A: *If it rains.*
c. Q: *Did Lucy win the race?* A: *She might have.*

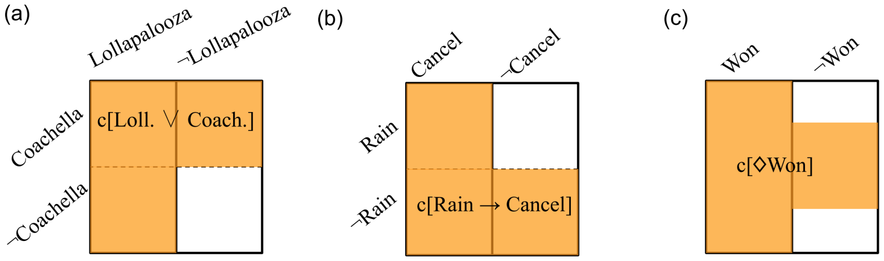


Figure 1: Illustration of the dialogues in (4). The large box represents the context set, divisions correspond to cells in the QUD, and the gold region represents answers. The dashed line represents the divisions corresponding to a different question. (Agha and Warstadt, 2020: 26: Fig. 3)

However, in each case, the eliminated set of worlds is *included within* a single alternative in Q. For this reason, Agha and Warstadt (2020) suggested a new class of relevant responses in addition to partial answers, which we called **reductive answers**. A reductive answer is any proposition a such that a or the negation of a counts as a partial answer (5). For example, in Part (a) of Figure 1, the response *Coachella or Lollapalooza* is consistent with both *Coachella* and *not Coachella*. But the negation of *Coachella or Lollapalooza* is inconsistent with *Coachella*, so the negation is a partial answer.[3]

[3] Agha and Warstadt (2020) also showed how reductive answerhood generalized to *wh*-questions, relevance of questions to other questions, and QUD shifting in discourse.

(5) **Definition: Reductive Answer** (Agha and Warstadt, 2020: 24: (21))
Proposition a is a reductive answer to Q iff $\exists_{q \in Q} \exists_{b \in ?a}[b \cap q = \emptyset]$, *where* $?a = \{a, \neg a\}$.

Our revision to Roberts's account inherits many of its virtues and shortcomings. As far as categorical theories of relevance go, a notion at least as inclusive as reductive answerhood appears to be necessary. Still, it remains too restrictive. We find that it excludes all the (relevant) responses in (1), to say nothing of capturing gradient distinctions. Our goal in the present work is to address some of the shortcomings of the categorical theory, by considering candidates for a gradient measure of relevance.

3. Bayesian approaches to relevance

The need for a gradient measure of relevance suggests that some probabilistic reasoning is at play. Work in Bayesian pragmatics has already made probabilistic extensions to the QUD framework, and offers several alternative operationalizations of relevance, usually under the guise of *utility*. The most notable alternatives are **entropy reduction** (Oaksford and Chater, 1994; Nelson et al., 2010; van Rooy, 2004; Rothe et al., 2018) and **KL divergence** (Nelson et al., 2010; Hawkins and Goodman, 2017).

These frameworks add the assumption that conversational agents are able to reason not just about the entailment relations between a proposition a and the context C, but also the probability of a proposition being true in a given context $P(a|C)$. Each alternative in the QUD Q is now associated with a probability in context, and since Q is a partition over the context set, the probability of its alternatives sums to 1. In other words, Q is a random variable with its domain equal to the set of QUD alternatives, distributed according to $P(Q|C)$ in context C.

3.1. Entropy reduction

Information that reduces uncertainty about the QUD is relevant: this is the intuition behind hypothesizing entropy reduction as a measure of relevance. Entropy reduction (7) measures the change in Shannon entropy (Shannon, 1948) of the probability distribution over the QUD after gaining a new piece of information.

(6) **Definition: Entropy of a question**

$$H(Q) := \sum_{q \in Q} P(q)(-\log P(q))$$

(7) **Definition: Utility of an Answer – Entropy reduction**

$$\begin{aligned} U_{ER}(a;Q) : &= H(Q) - H(Q|a) \\ &= \sum_{q \in Q} P(q)(-\log P(q)) - \sum_{q \in Q} P(q|a)(-\log P(q|a)) \end{aligned}$$

Defining entropy. Entropy itself is defined in (6). To give the unfamiliar reader an intuition about this definition, it makes sense to start with the **surprisal** of a proposition: $S(a) := -logP(a)$. The negative log probability of a is a quantity that matches our intuitions about what it means for a to be surprising, or provide information: $S(a) = 0$ when $P(a) = 1$, and $S(a)$ approaches positive infinity as $P(a)$ approaches 0. The less probable a proposition is, the higher its surprisal.[4]

The Shannon entropy of a random variable Q (in our case, a QUD), is the expected value of the surprisal of Q. This is shown in definition (6). Intuitively, the entropy of Q tells you how much information you can expect to gain upon learning the answer to a question. Entropy is always highest when each of the alternatives are equally likely: For example if there is a 50% chance of rain, the entropy of the question *Q=Is it raining* is:

$$\begin{aligned} H(Q) &= P(yes)(-\log_2 P(yes)) + P(no)(-\log_2 P(no)) \\ &= 0.5(-\log_2 0.5) + 0.5(-\log_2 0.5) \\ &= 1 \end{aligned}$$

If instead there is a 10% chance of rain, the entropy is:

$$\begin{aligned} H(Q) &= P(yes)(-\log_2 P(yes)) + P(no)(-\log_2 P(no)) \\ &= 0.1(-\log_2 0.1) + 0.9(-\log_2 0.9) \\ &\approx 0.332 + 0.137 = 0.469 \end{aligned}$$

Although the information that it is raining has high surprisal in this context, there is a much greater chance that one will end up in the less surprising outcome.[5]

Examples of entropy reduction. This brings us to entropy reduction (7). This quantity, in the context of discourse, measures how much a new piece of information a decreases the entropy of the QUD Q. The graphical representation of entropy reduction in Figure 2 below shows how entropy reduction for a polar QUD changes as a function of the prior and posterior probability of the *yes* alternative:

(I) When new information increases one's certainty about which answer to the QUD is true, entropy reduction takes a positive value. This matches intuitions about relevance: in Table 1, row (a), entropy reduction (ER) is 0.92 bits. **(II)** When new information is irrelevant, the prior and posterior are equal and entropy reduction is 0.[6] This accounts for irrelevance: in Table 1, row (b), entropy reduction is 0 bits. **(III)** When new information *decreases* one's certainty about which answer is true, entropy reduction is negative. This violates intuitions: in Table 1, row (c), entropy reduction is -0.92 bits, yet the information is relevant.[7] **(IV)** When new information alters the prior in such a way that the affirmative and negative alternatives swap

[4] To give a simple example, suppose the weather forecast says there is a 100% chance that it is raining in my area. Then, the surprisal of the proposition *a=it is raining* in bits is $S(a) = -(\log_2 1) = 1(0) = 0$, and analogously I will not gain any information if I step outside and see that it is raining. On the other hand, if the forecast says there is a 50% chance of rain, then the surprisal of a in bits is $S(a) = -\log_2 0.5 = 1$, i.e. a provides 1 bit of information. If the chance of rain is even smaller, say 1%, then the surprisal is $S(a) = -\log_2 0.01 \approx 6.64$, and indeed I would gain even more information, and be even more surprised, to step outside and see that it is raining.

[5] Note: It is conventional to define the surprisal of a proposition with probability 0 to be $+\infty$, and in calculations of entropy, $0 * +\infty$ is treated as 0. Hence, if there is 100% chance of rain, $H(Q) = 0$.

[6] The converse is not true: there is some relevant information that does not alter the prior (see Section 6).

[7] We could avoid negative ER values by taking the absolute value of entropy reduction. However, this is not typically done in work that uses entropy reduction, and this would not fix other bad predictions, such as the one corresponding to Table 1, row (d).

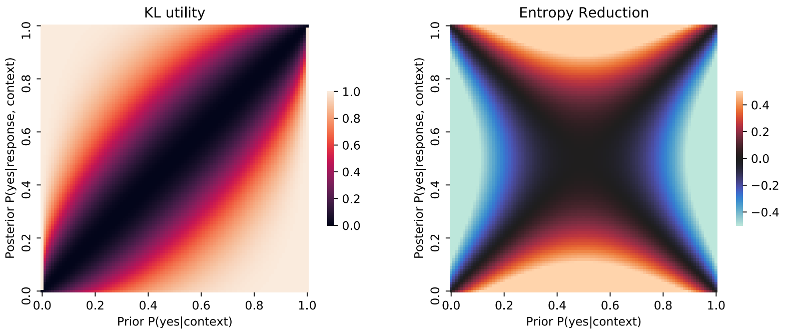
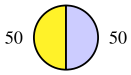
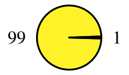
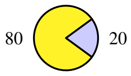
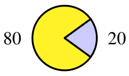
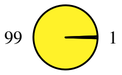
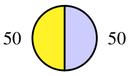
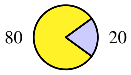
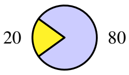
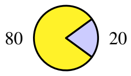
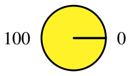


Figure 2: Graphical representation of KL utility and entropy reduction as a function of the prior and posterior, for a polar question. The prior and posterior probabilities represent the probability of the affirmative answer.

probability, entropy reduction is 0. This is rather bizarre: in Table 1, row (d), entropy reduction is 0 bits, yet the information is relevant and shifts one's prior substantially. **(V)** Exhaustive answers are not always maximally relevant. As Table 1, row (e) shows, entropy reduction of an exhaustive answer is equal to the entropy of the prior distribution.

Table 1: Examples of entropy reduction and KL utility in different scenarios with the QUD Q *Will admiral win?*. Pie charts show the probability that Admiral will win on the left in yellow, and the probability that Admiral will not win on the right in blue. 'Ent.' is the entropy of Q in the current information state, 'ER' is the entropy reduction of the announcer's statement a ($U_{ER}(a;Q)$), and 'KL' is the KL utility of the announcer's statement ($U_{KL}(a;Q)$).

	Prior			Posterior			Utility	
	Context	**QUD**	**Ent.**	**Announcer**	**QUD**	**Ent.**	**ER**	**KL**
(a)	Admiral and Barney are neck and neck.		1	Admiral takes a huge lead in the last lap!		0.08	0.92	0.88
(b)	Admiral is a strong favorite.		0.72	Today's sponsor is Nike!		0.72	0	0
(c)	Admiral is about to win.		0.08	Barney catches up, it's a photo finish!		1	-0.92	0.99
(d)	Admiral is a strong favorite.		0.72	Barney takes a 50m lead!		0.72	0	0.94
(e)	Admiral is a strong favorite.		0.72	Admiral wins!		0	0.72	0.52

3.2. KL divergence

KL divergence (8) is a measure of the difference between two probability distributions P and P'. Unlike entropy reduction, it does not directly track how different P and P' are from uniform distributions. It measures the expected excess code length if the code is optimized not for the true distribution (or posterior) P, but for an approximation (or prior) P'. Essentially, the KL divergence is the extra encoding cost of using the approximation (prior) rather than the true distribution (posterior).

(8) **Definition: KL divergence** of two distributions (over the same alternative set Q)

$$KL_Q(P||P') := \sum_{q \in Q} P(q) \cdot \log \frac{P(q)}{P'(q)}$$

Using the definition of KL divergence, we give a formula in (9) for the utility of an answer a in the context of a QUD Q. Formula (9) contains an instance of definition (8) in which the prior P is the approximation, and the posterior upon learning a, $P(\bullet \big| a)$, acts as the true distribution.

(9) **Definition: KL utility**

$$\begin{aligned} U_{KL}(a;Q) :&= g\Big(KL\big(P(\bullet|a)\big|\big|P(\bullet)\big)\Big) \\ &= g\left(\sum_{q \in Q} P(q|a) \cdot \log \frac{P(q|a)}{P(q)}\right) \qquad \text{where } g(x) = 1 - a^{-x} \end{aligned}$$

We apply a squashing function g to map KL divergence, which is unbounded above, onto the interval $[0,1)$. We set the free parameter a to 10. We use **KL divergence** to refer to the general definition in (8), and **KL utility** to refer to the squashed utility function in (9).

3.3. Deriving KL divergence

For the unfamiliar reader, the intuition behind KL divergence begins with the concept of **cross entropy** (10). Let P be the true probability distribution over the alternatives of a QUD Q, and let P' be an approximation of P. The cross entropy of P and P' tells you how much information you will gain on average if you falsely believe that Q is distributed according to P' rather than P. Like entropy, the cross entropy is the expected surprisal of the answer to Q. The difference is that the surprisal for the question is measured using an approximation of the true distribution P', whereas the expectation is taken with respect to the true distribution P.

(10) *Cross entropy of two distributions*

$$H(P,P') := \sum_{q \in Q} P(q)(-\log P'(q))$$

(11) *KL divergence in terms of cross entropy*[8]

$$KL(P\big|\big|P') = H(P,P') - H(P)$$

[8] This is a slight abuse of notation: We write $H(P)$ to denote the entropy of Q distributed by P.

The KL divergence of P and P' is simply the difference of the cross entropy of P and P' with respect to Q and the entropy of Q as distributed by P. In other words, KL divergence tells you how much *more* information the answer to Q will convey on average, if you falsely believe Q to be distributed according to P' rather than P. The identity in (11) is easily shown to be equivalent to the definition in (8).

Examples of KL divergence. KL utility, defined in (9), is the KL divergence of the posterior from the prior, squashed into the interval [0,1). A graphical representation of KL utility is given in Figure 2. The following points illustrate key behaviors of KL utility:

(I) When new information shifts one's probability distribution over the QUD, the KL utility is positive. This accounts for the relevance of rows (a), (c), (d), and (e) in Table 1. **(II)** When new information is irrelevant, the prior and posterior are equal and KL utility is 0 (see Table 1, row (b)). **(III)** KL utility approaches 1 as the prior probability gets arbitrarily close to 0 for some alternative with non-zero posterior probability. As Table 1, row (c) shows, the posterior does not need to be biased for the KL utility to be near-maximal. When some alternative with 0 prior probability has non-zero posterior probability (corresponding to a destructive update), KL utility is undefined. **(IV)** KL utility is asymmetrical, i.e. $KL(P||P')$ does not always equal $KL(P'||P)$ (compare Table 1, rows (a) and (c)). **(V)** Exhaustive answers are not maximally relevant. This defies intuition: as Table 1 shows, an exhaustive answer that confirms expectations (row (e)) is predicted to have lower utility than some non-exhaustive answers that challenge prior expectations more strongly (rows (a) and (c)).

4. Method

Our experiment directly tests participants' introspective judgments of response relevance in dialogues against gradient measures of relevance. entropy reduction and KL utility are computed from participants' estimates of prior and posterior probabilities with respect to the QUD.

4.1. Materials

We constructed 150 dialogues, where each dialogue consists of a polar question, a context, and an answer. There were ten distinct polar questions, and for each question, there were three contexts and five answers, for a total of fifteen conditions per question. Table 2 below shows the full set of conditions for one question.

The three context conditions correspond to which answer to the question we expect to be favored by the context: A *negative bias* context favors a high prior probability for the negative answer, a *positive bias* context favors the affirmative answer, and a *neutral bias* context favors neither answer strongly.

The five response conditions vary in the degree to which they favor the affirmative answer (all answers were written to *increase* or *not alter* the probability of the *yes* answer). *Exhaustive* answers entail the affirmative answer, *high certainty* answers strongly favor the affirmative answer, *low certainty* answers weakly favor the affirmative answer, and *non-answers* favor neither alternative. We also include a fifth condition for *reductive answers* (see Section 2).

Table 2: This table provides examples of different contexts and response types within a single vignette. The positive bias context suggests that the family is inside the house (the *yes*-answer to the question), while the negative bias context gives reasons to think they have left. The response types favor either the *yes*-answer or neither answer, and they differ in informative strength.

Context		
Negative bias	Neutral	Positive bias
You're a burglar, trying to rob a house with your accomplice Chris. You've been staking out a good looking house for the last week, ***and noticed that the family seems to have left for vacation.***	*You're a burglar trying to rob a house with your accomplice Chris. You've been staking out a good looking house for the last week,* ***and noticed that the family often goes out to dinner around this time.***	*You're a burglar trying to rob a house with your accomplice Chris. You've been staking out a good looking house for the last week,* ***and noticed that the family usually watches TV together around this time.***

Question
Is there anyone inside the house?

Response				
Exhaustive	High Certainty	Low Certainty	Non-answer	Reductive
The whole family is inside the house.	*I heard voices coming from inside.*	*There's a light on.*	*The house has two entrances.*	*If there are cars in the driveway.*

4.2. Experiment

We collected three kinds of judgments from participants: priors, posteriors and helpfulness judgments, described in Figure 3. We collected each judgment from three different participants, and each rating was given on a slider from 0 to 1, see Figure 4 further below for an example. We used the prior and posterior judgments to compute $U_{KL}(a;Q)$ and $U_{ER}(a;Q)$.

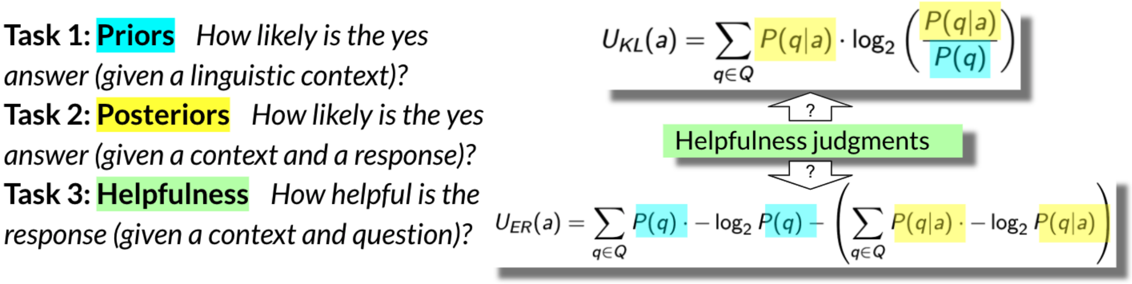
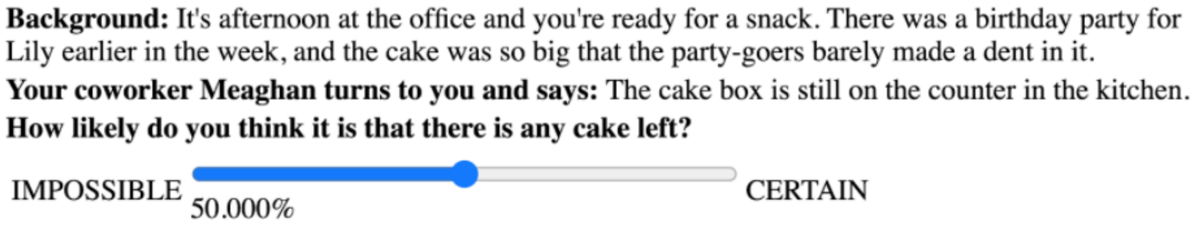

Task 1: Priors How likely is the yes answer (given a linguistic context)?
Task 2: Posteriors How likely is the yes answer (given a context and a response)?
Task 3: Helpfulness How helpful is the response (given a context and question)?
$U_{KL}(a) = \sum_{q \in Q} P(q|a) \cdot \log_2 \left(\frac{P(q|a)}{P(q)} \right)$
?
Helpfulness judgments
?
$U_{ER}(a) = \sum_{q \in Q} P(q) \cdot -\log_2 P(q) - \left(\sum_{q \in Q} P(q|a) \cdot -\log_2 P(q|a) \right)$

Figure 3: The three tasks in the study. We collect measurements of prior and posterior beliefs to calculate utility values based on two different metrics used to predict helpfulness judgments.

Each participant completed 10 experimental trials – one per question in our stimuli – and provided the same type of judgment in all trials (i.e. individual participants were to judge only priors, only posteriors, or only helpfulness). In addition, participants completed five filler trials, three of which served as attention checks, and two of which served as comprehension checks.



Background: It's afternoon at the office and you're ready for a snack. There was a birthday party for Lily earlier in the week, and the cake was so big that the party-goers barely made a dent in it.

Your coworker Meaghan turns to you and says: The cake box is still on the counter in the kitchen.

How likely do you think it is that there is any cake left?

IMPOSSIBLE 50.000% CERTAIN

Figure 4: One trial for the posterior judgment task.

4.3. Slider

Previous work that uses sliders for belief estimation has found that participants are resistant to using the endpoints of the slider, which causes an artificial lack of extreme probability judgments (Chen et al., 2020; von der Malsburg et al., 2020). To counteract this bias, we transform the values on the slider using the inverse of the Linear Log Odds (LLO) transformation.[9] The LLO transformation has been proposed in psychometric literature to correct for probability distortion, the tendency for humans to perceive differences in probability differently for extreme probabilities (Zhang and Maloney, 2012). We use the inverse of the LLO transformation, to stretch the probability scale at the endpoints of the slider and compress the probability scale near the middle. This gives participants more flexibility when choosing values near the endpoints of the scale.[10]

4.4. Participants

We had 147 HITs, of which 105 passed at least four out of five quality checks. Two HITs were excluded due to duplicate workers. Participants whose responses passed four or more quality checks were paid \$2.50, or approximately \$15/h, assuming a ten minute completion time for fifteen trials. Participants whose responses passed fewer than four quality checks were compensated \$1.00. The different pay rates were accomplished by compensating all participants a base rate of \$1.00 and issuing a bonus of \$1.50 for work that passes quality checks.

4.5. Filtering

32 items out of 150 items were filtered due to low crowdworker agreement. We filtered items based on the range of judgments: If the range of the prior, posterior, or helpfulness judgments was greater than 0.75, we removed the whole item. For example, if the prior judgments for

[9] For probability p, the LLO transformation $\pi(p)$ is defined implicitly as:

$$\log \frac{\pi(p)}{1-\pi(p)} = \gamma \left(\log \frac{p}{1-p} \right) + (1-\gamma) \left(\log \frac{p_0}{1-p_0} \right).$$

The LLO transformation can be visualized as an S-shaped curve, where the center of the S is at the fixed point $p_0 = 0.5$, and the slope is determined by γ. In our application, the fixed point parameter p_0 is set to 0.5, and the slope parameter γ is set to 1.5.

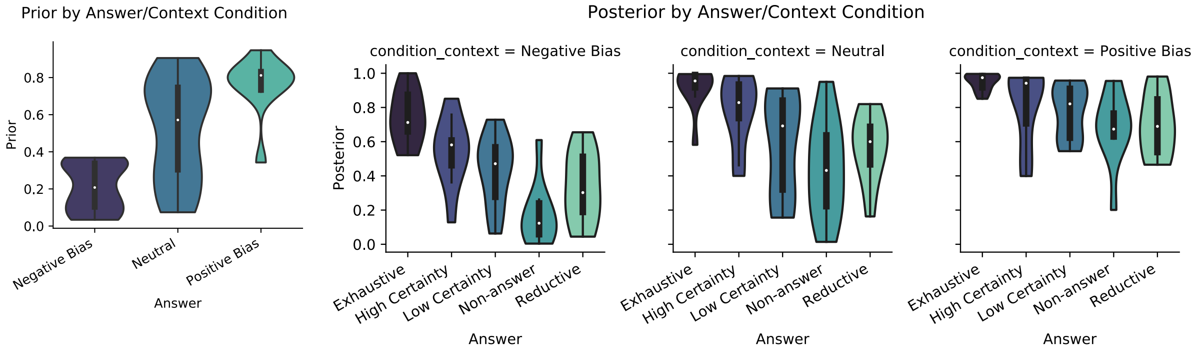
[10] For example, the point which would have the value 0.9 on a non-transformed slider instead shows 0.964.

a given context were 0.1, 0.2, and 0.9, we would remove all prior, posterior, and helpfulness judgments including that context.

4.6. Sanity checks

We find that participants' probability judgments are highly predictable from context and answer condition, as shown in Figure 5. This serves as a validation of our stimuli: Since we used a 3×5 design in our stimuli simply to ensure that items take on a wide range of prior and posterior values, these findings do not directly bear on our hypothesis.

The distribution of prior judgments by context condition (the leftmost plot in Figure 5) increases precipitously as we go from negatively biased, to neutral, to positively biased contexts, with medians increasing from 0.12 to 0.57 to 0.80. The distribution of posterior judgments by both context and answer condition (the right three plots in Figure 5) also shows an expected pattern. For a given context condition, posterior judgments decrease consistently as the answer condition goes from exhaustive to high certainty to low certainty to non-answer. Reductive answers generally lead to posteriors slightly lower than a low certainty answer. We find this same decreasing pattern in posterior judgments for all context conditions, but unsurprisingly, posterior judgments are higher for all answer types when context bias is more positive.



Prior by Answer/Context Condition

Posterior by Answer/Context Condition

condition_context = Negative Bias

condition_context = Neutral

condition_context = Positive Bias

Figure 5: Distribution of participants' prior judgments by context condition (left) and posterior judgments by context and response condition (right). Prior judgments vary with context condition as expected. Posterior judgments vary predictably depending on the response condition, from exhaustive answer (highest posterior) to non-answer (lowest posterior). Reductive answers are more spread out. In positive bias contexts, differences between answer types are compressed due to ceiling effects.

5. Results

We had two hypotheses: first, that the categorical theory of relevance could not meaningfully predict helpfulness judgments, and second, that KL utility would be more predictive of helpfulness judgments than entropy reduction. As we discuss below, our results strongly support the first of these hypotheses, but not the second.

5.1. Testing the categorical theory

Figure 6 shows the distribution of helpfulness judgments for different answer types. helpfulness varies starkly by answer type, but context condition appears to have little effect. To test this intuition, we fit linear models to predict helpfulness judgments from context and answer conditions. We find that answer type is strongly predictive of helpfulness ($R^2 = 0.78, p < 10E-5$), while context condition is not predictive at all ($R^2 = 0.002, p > 0.8$). A Bayes factor comparison (Figure 10 in Section 5.4) confirms that the interaction between answer type and context condition is not significant.

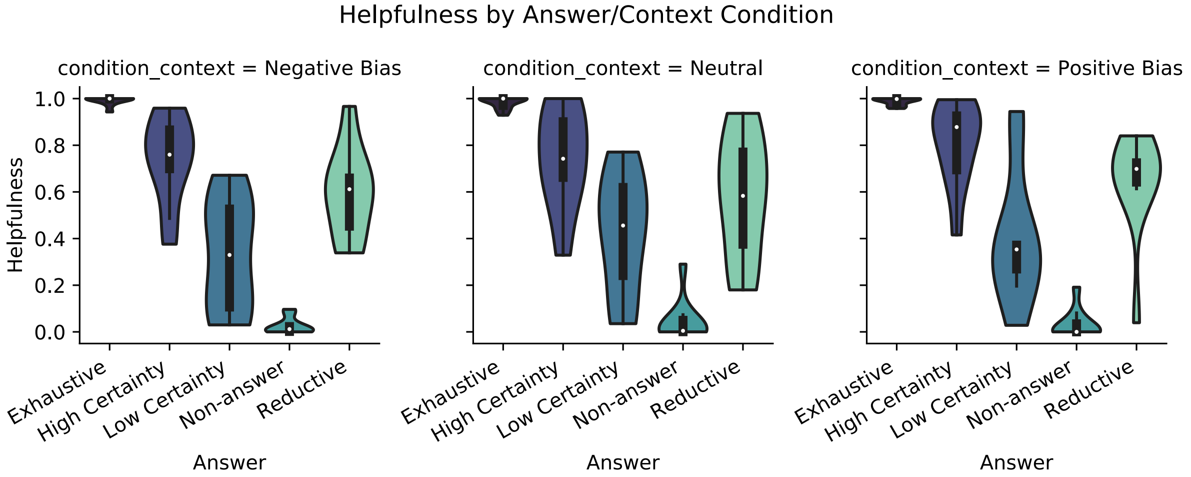


Figure 6: Distribution of participants' helpfulness judgments by response type (Answer) and context condition. Response types are strongly predictive of helpfulness judgments across all context conditions.

In order to assess the meaning of these results for the categorical partial answer theory of relevance, we have to imagine how participants would respond on a scale given categorical judgments. On one version of the categorical theory, only partial answers (including exhaustive answers) would be judged relevant, and all other responses would be irrelevant. Call this the **strong version** of the partial answer theory. On the strong theory, we would expect to see a strongly bimodal pattern to helpfulness judgments. This is clearly not what we observe in Figure 6. While exhaustive answers and non-eliminating answers that we classified as non-answers show clear categorical behavior, participants' helpfulness judgments cover the full scale for other answer types.

On the **weak version** of the partial answer theory, we might expect to see three categories of relevance: full relevance, intermediate relevance, and irrelevance. This interpretation of the partial answer theory is more plausible, but it is still inconsistent with our results. It does not explain the observed differences between low certainty and high certainty responses. Furthermore, it does not capture the difference between answers that we classified as non-answers on the one hand, and answers that we classified as low- or high-certainty on the other hand. As non-eliminating answers, all three answer types are predicted in the categorical theory to be minimally relevant.

The strong and weak categorical theories differ in their predictions for partial answers. On the strong theory, partial answers must be maximally relevant. However, for at least some *wh*-questions, it is possible to find examples of non-exhaustive partial answers that are intuitively less relevant than non-partial answers. If we compare two possible responses to the question (12), our judgment that the high-certainty answer (12a) is more relevant than the partial answer (12b).[11]

(12) What are we having for dinner?
 a. There are hamburger buns on the counter.
 b. Not lasagna.

On the weak categorical theory, only exhaustive answers are maximally relevant, and non-exhaustive partial answers fall into the intermediate category. However, in the case of polar questions, all partial answers are exhaustive answers. So we are back to the strongly bimodal predictions of the strong partial answer theory addressed above. Even if the categorical theorist adopts Agha and Warstadt's (2020) reductive answerhood revision to the categorical theory, they predict that all reductive answers should be (at least) more relevant than non-reductive answers. However, our results show that on average reductive answers are less relevant than high certainty answers, which are not reductive answers.

In summary, there is no straightforward path for the categorical partial answer theory. If there is a correct categorical theory, it likely involves more than two categories of relevance, and these categories do not seem to map cleanly onto either partial answers or reductive answers.

5.2. Testing the gradient theory

The distribution of helpfulness judgments in Figure 6 clearly supports a gradient theory of relevance. To test which measure provides a better fit, we measure Spearman's rank order correlation between helpfulness judgments and several gradient measures. In addition to entropy reduction and KL utility, we evaluate two baseline models of relevance: the posterior alone, and the (positive) distance between the prior and posterior. Neither baseline was hypothesized to be a strong model of relevance.

Figure 7, top row, shows the correlations. Contrary to our predictions, neither KL utility nor entropy reduction is strongly correlated with helpfulness judgments, and KL utility is not a substantially stronger predictor of helpfulness. We therefore ask whether these information-theoretic measures provide a better fit than alternative, less well-motivated measures. Surprisingly, in Figure 7 we find that helpfulness correlates *more strongly* with posterior on its own, and the (positive) distance between the posterior and prior probabilities both correlate *more strongly* than with KL divergence or entropy reduction.

[11] Our data does not include *wh*-questions, but this example suggests that a followup study on *wh*-questions might be useful.

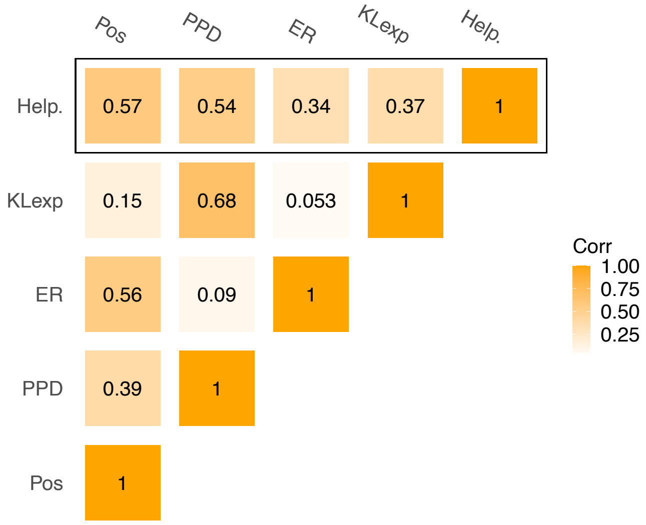


Figure 7: Spearman correlation plot showing the relationship between helpfulness and several predictor variables. The posterior (Pos) and the posterior-prior difference (PPD) show the strongest correlation with helpfulness, followed by KL utility. ER is the weakest.

5.3. Breaking down the fit by prior and posterior

To gain more insight into the success- and failure-modes of gradient relevance metrics, Figure 8 shows how helpfulness judgments *as a function of prior and posterior* compare the overall plots of entropy reduction and KL utility (repeated from Figure 2), for all combinations of prior and posterior. The first thing to notice is that our stimuli were not designed to elicit data points in the lower right-hand triangle of the scatter plot, the region where the posterior is lower than the prior. We only consider responses that favor the positive resolution of the question.[12] However, participants did sometimes report a posterior lower than their prior. Some of these points are likely noise, but some might be due to systematic misconstruals of the stimuli. For our analysis, we restrict our attention to the data close to and above the $x = y$ diagonal. Both models predict zero utility for data along this diagonal, where the prior and posterior are equal.

In other regions of the space, the two models make very different predictions. First, the entropy reduction model predicts zero utility for the $x = -y$ diagonal, the region occupied by examples

[12] Due to this asymmetry, we cannot use our data to compare correlations between helpfulness judgments and measurements in subsets of the data determined by prior probability alone, or posterior probability alone. This is because these slices of data will systematically exclude data points that would affect the outcome of the analysis. For instance, we observe that ER correlates more strongly with helpfulness than KL divergence when the prior is greater than 0.5. This is merely an artifact of the fact that, when the prior is greater than 0.5, problematic cases of negative ER will only be observed when the posterior is *lower* than the prior, and our stimuli avoid such examples. We leave such an analysis to future work.

like (d) in Table 1. The KL utility model, on the other hand, predicts high utility when the prior is low and the posterior is high, and when the prior is high and the posterior is low. In other words, KL utility predicts that large posterior-prior differences always lead to high relevance.

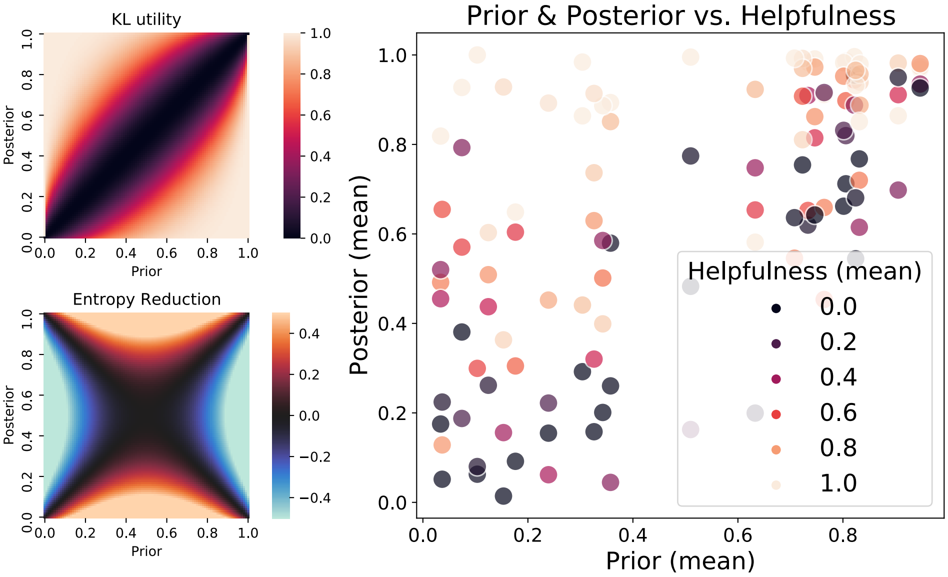


Figure 8: The scatter plot on the right shows the mean judgments for each item (after filtering to remove low-agreement items). The item's mean helpfulness rating (color) is plotted against the mean prior (x-axis) and posterior (y-axis) ratings. It can be compared to the heatmaps on the left that show the predictions of each KL utility and entropy reduction.

Second, as we saw in Section 3, entropy reduction penalizes responses that increase uncertainty, assigning them negative utility (colored in blue in Figure 8). Contradicting entropy reduction, the scatterplot shows that responses close to the point (0, 0.5) are judged *more* relevant than responses along the $x = y$ diagonal. The ER model can be altered by taking the absolute value of ER utility–as it were, turning the blue regions light orange. But the altered ER model still has a problem in the upper left hand corner, where the prior is low and the posterior is high. In that region, participants' judgments resemble the KL utility plot, assigning high relevance to large posterior-prior differences.

Both measures share a flaw when it comes to high-prior, high-posterior cases. Conceptually speaking, they take for granted that small shifts in probability (i.e., points near the $x = y$ diagonal) indicate that the response provided little new information, and have low utility on those grounds. This is even the case for exhaustive answers, which both measures assign low utility to when the context is positively biased. However, we observe that the upper right hand corner of the scatter plot contains many points with high helpfulness, telling us that a relevant response does not need to shift one's prior substantially. More generally, this is unexpected for any model that assumes that helpfulness tracks informativity while reducing informativity to updates in point-estimates of probabilities.

Although the baseline measures – the posterior alone, and the prior-posterior distance – are more strongly correlated with helpfulness, obvious counterexamples to these measures exist. First of all, low posteriors are only rated less helpful because our stimuli lack examples where the prior is high and the posterior low. Even if we attempt to account for such cases by considering the distance of the posterior from 0.5, the posterior will fail to predict low helpfulness in cases where the response is largely irrelevant, but the prior and posterior are both high or both low, as in (13).

(13) *Tony got Ethan a book-shaped birthday gift. The book looks big and is probably a hardcover. You know that Ethan is really into cooking, so the first possibility that comes to your mind is that Tony got him a cookbook.* prior: .75
Ethan's sister gave him fuzzy socks. posterior: 0.72
helpfulness: 0 KL: 0.08

The posterior-prior distance (PPD) does not run into the same problem with non-answers, since the PPD (like KL) is close to zero in these cases. This makes it a more interesting alternative to the KL divergence model. PPD and KL divergence are highly correlated, and the relationship is nonlinear.

To make this comparison easier, Figure 9 shows the difference between KL divergence and PPD. Notice that KL divergence is higher than PPD when the prior is greater than 0.5 and the posterior is extremely close to 1 (or when the prior is less than 0.5 and the posterior is extremely close to 0). This is because PPD is bounded above by 1 minus the prior. As a result, PPD predicts much lower helpfulness for responses where the posterior is close to 1. This can be seen in the center heatmap in Figure 9 by looking at the high-value regions near the bottom (posterior = 0) and the top (posterior = 1).

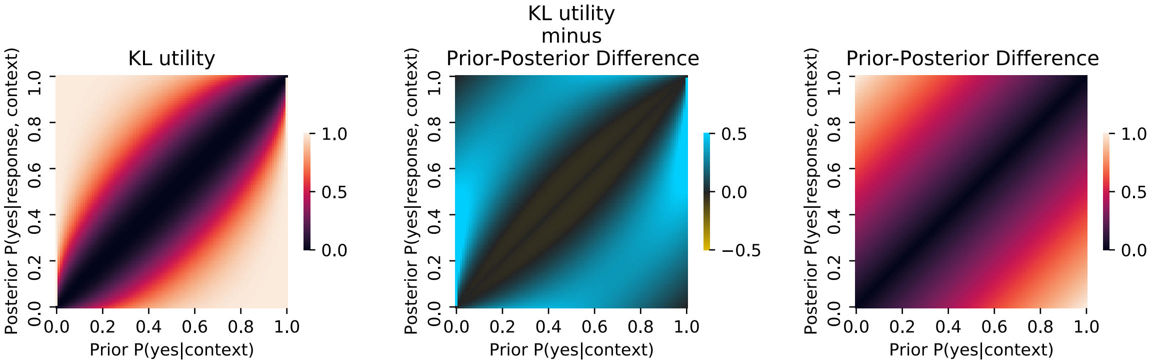


Figure 9: KL divergence (left) compared to the PPD (right). The middle heatmap shows the regions where KL is higher than the PPD.

(14) *Your friend Adam grew up listening to the Les Miserables soundtrack. This year, the school's drama club is putting on the show, and (naturally) Adam auditioned for the lead role. As a loyal friend, you are trying to find some information about how he did.*
prior: 0.71
Adam was very happy to learn that he got the part. posterior: 0.99
helpfulness: 1 KL: 0.65 PPD: 0.28

5.4. Model comparison using Bayes Factors

So far, the only measure we have used to compare our two information-theoretic metrics is the Spearman rank-order correlation. However, it is also useful to consider the relative performance of various linear models that we fit to our data. To get a sense of how well various choices of factors might explain the data, we performed a Bayesian model comparison (Morey et al., 2016). The Bayes Factor is the likelihood ratio between the marginal likelihood of two models H_1 and H_2, i.e. $P(X|H_1)/P(X|H_2)$, where X is the data. In Figure 10, we take the ratio of the model likelihood relative to the intercept. Bayes Factor comparison does not offer a categorical notion of significance analogous to that of classical hypothesis testing. Rather, Bayes Factors should be interpreted as a measure of the weight of evidence in favor of one hypothesis over another.[13]

In Figure 10, notice that context type is the worst predictor overall (plot (c)), and answer type is the best (plots (c-d)). We do not interpret this as supporting the categorical theory. Rather, this is a gap that future gradient theories should aim to fill. Among the continuous predictors, the single factor model for the posterior (Pos) is significantly more likely than the intercept compared to the PPD, KL utility, and entropy reduction.[14]

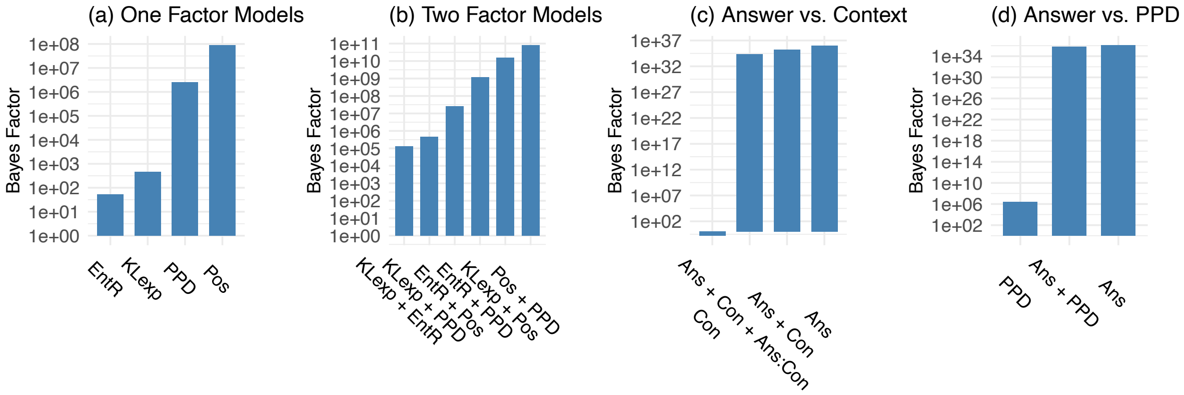


Figure 10: These plots show the Bayes Factor (relative to the intercept) for a variety of linear models. Plots (a) and (b) show models with factors chosen from {entropy reduction, KL exponential, posterior-prior difference, posterior}. Plots (c) and (d) compare the answer type (Ans) and context type (Con) to other predictors of helpfulness.

[13] As a rule of thumb, we consider one Bayes Factor to be significantly higher than another if it is at least ten times greater.

[14] More precisely, the posterior model's BF is over ten times higher than the next best model, the PPD.

6. Discussion

6.1. KL divergence: Findings that support the model

Shifting probabilities is helpful. The key assumption behind Bayesian models of relevance is that the relevance of a response can be predicted by changes in the probabilities of the possible answers. This assumption is shared by both the entropy reduction and KL divergence models. (15) gives a typical example of a low prior and a high posterior, leading to a high KL score.

(15) Context: *It's afternoon at the office and you're ready for a snack. There was a birthday party for Lily earlier in the week, and the cake was delicious.*
Q: *Is there any cake left?* prior: 0.15
A: *There are three slices left.* posterior: 0.93
helpfulness: 0.94 KL: 0.99

Increasing uncertainty can be helpful. The major conceptual difference between entropy reduction and KL divergence is that under the entropy reduction model, only responses that decrease uncertainty have positive ER values.[15] If we interpret responses with negative ER values as less relevant than responses with positive ER values, then the model predicts that responses that decrease uncertainty will always be more relevant than responses that increase uncertainty.[16]

KL divergence does not have this property. Under the KL divergence model, updating from a peaked distribution to a uniform distribution can be just as helpful as updating from a uniform distribution to a peaked distribution.[17] Unhelpful responses are unhelpful because they only create small changes in the probability distribution, and the direction of the change does not matter.

This is easy to see in cases where the prior is low. In (16), the context mentions that the family seems to have left for vacation. The prior probability that someone is in the house is therefore low (12%). After learning that there are voices coming from inside the house, the posterior is 60%, which is much closer to a uniform distribution than the prior. Thus, the entropy reduction score is low (in fact, negative) but the KL score is moderate, and participants judged this example very helpful.

(16) Context: *You're a burglar, trying to rob a house with your accomplice Chris. You've been staking out a good looking house for the last week,* ***and noticed that the family seems to have left for vacation.***
Q: *Is there anyone in the house?* prior: 0.12
A: *I heard voices coming from inside.* posterior: 0.60
helpfulness: 0.96 KL: 0.88 ER: -0.43

[15] For an intuitive example, see Table 1, row (c).

[16] We could avoid negative ER values by taking the absolute value of entropy reduction. However, this method would still assign 0 ER to cases like (d) in Table 1.

[17] However, KL divergence is not symmetric, so in general, updating from a uniform distribution U to a peaked one P is not as helpful as updating from the *same* peaked distribution P to U.

6.2. KL divergence: Findings that are incompatible with the model

Confirming suspicions is helpful. Under the KL divergence model, the degree of change from prior to posterior determines the helpfulness of the response. This entails that small changes in probability are unhelpful. However, our data shows that this is not always the case. When participants assign a high prior probability to the *yes* answer, then even if the response is an exhaustive answer, the posterior will not be significantly higher than the prior. This is simply because there is not much room to move.[18]

In (17), the context sets up a fairly high prior probability for the *yes* answer (that Tony got Ethan a cookbook). The high-certainty response offers further evidence for the *yes* answer, which participants judged very helpful. But the difference between prior and posterior is low, and the KL score is small.

(17) Context: *Tony got Ethan a book-shaped birthday gift. The book looks big and is probably a hardcover. You know that Ethan is really into cooking, so the first possibility that comes to your mind is that Tony got him a cookbook.*
Q: *Did Tony get Ethan a cookbook?* prior: 0.72
A: *Tony asked about cookbook recommendations the other day.* posterior: 0.81
helpfulness: 0.94 KL: 0.07

In general, exhaustive and high certainty responses are judged very helpful, regardless of whether the prior is low or high. This suggests that in certain cases, the prior matters less than the posterior.

Strategizing about QUD resolution is helpful. In the original QUD framework, Roberts (2012) introduces the useful concept of a **strategy of inquiry** to resolve the QUD. A strategy of inquiry is a sequence of discourse moves aimed at resolving a particular question. These discourse moves can be other questions (subinquiries) that are considered easier to answer, and whose answers provide information about how to resolve the main question. This dimension of discourse structure is not captured by our simple Bayesian models, though such strategies could be included in a more complete model.

For example, in (18), the response does not lead to a large difference between prior and posterior. But it does provide useful information to the asker: To discover the answer to the question, the asker should go check the driveway. The response sets up a strategy of inquiry, which is conversationally useful even when the response does not significantly shift the probabilities of the possible answers.[19]

[18] However, as the posterior approaches 1 or 0, KL divergence tends to infinity, so in principal a response that alters the prior by a very small amount could have arbitrarily large KL utility. However, participants almost never use very extreme values on the scale, so in practice we observe low KL values when priors are high.

[19] In principal, a response like (18A) could shift the probabilities of possible answers, for example, you had a prior belief that there are two cars in the driveway, independent of your degree of belief in the family's being home. The point here is that such responses are judged relevant even when they do not shift probabilities.

(18) Context: *You're a burglar trying to rob a house with your accomplice Chris. You've been staking out a good looking house for the last week, and noticed that the family often goes out to dinner around this time.*
Q: *Is there anyone inside the house?* prior: 0.30
A: *If there are two cars in the driveway, then someone is home.* posterior 0.44
helpfulness: 0.89 KL: 0.13

Reducing higher order uncertainty is helpful. When participants give probability judgments, two participants might give the same number, but differ in how certain they are about their judgment.[20] A subject's uncertainty about their subjective probability judgments is called **higher-order uncertainty**.[21] The basic model we have presented does not capture differences in higher-order uncertainty, but there is some evidence that participants' helpfulness judgments are sensitive to them.

In (19), the change between prior and posterior is small, and the KL score is correspondingly low. However, participants judge the response to be fairly helpful (70%).

(19) Context: *Maria is the new foreign exchange student at school. You want to introduce yourself, but you forgot where she is from. You know it's a big city in Europe.*
Q: *Is Maria from Madrid?* prior: 0.34
A: *She's either from Madrid or Moscow.* posterior: 0.50
helpfulness: 0.70 KL: 0.16

This suggests that there is some dimension of the data that is not captured by point estimates of probability. One possibility is that the response reduces participants' higher-order uncertainty about their probability judgments, even though it does not significantly shift their first-order probability distribution.

In ongoing work, we elicit judgments of higher-order uncertainty by prompting participants to give a judgment of their certainty level, along with first-order probability judgments. We then explore different Bayesian models that use the additional information.

Relevance of questions. In this paper, we have focused on the relevance of assertions as responses in short question-answer dialogues. However, there is a coherent notion of question relevance that can be found in classic work on the subject (Groenendijk and Stokhof, 1984; Roberts, 2012; Ginzburg, 1995). These notions of question relevance are categorical.

As van Rooy (2004) shows, question relevance can be naturally lifted to our probabilistic setting using **mutual information**. The mutual information between two random variables X and Y is given in (20). It is defined as the KL divergence between their joint distribution $p(X,Y)$ and their product distribution $p(X)p(Y)$.

(20) Mutual Information (Cover and Thomas, 1991: 18: (2.28))
$I(X;Y) = \sum_{x\in X} \sum y \in Y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$

[20] This might also happen to the same participant at different times.

[21] For previous work on higher-order uncertainty in linguistics, focusing on vague expressions, see Herbstritt and Franke (2019) and references therein.

The mutual information provides a measure of the informational value of a question X asked in the context of some QUD Y.[22] Let us provide some intuition for this measure. Recall that the KL divergence represents the new information gained by updating one's beliefs about the QUD from the prior to the posterior (see Section 3.3). The mutual information is the information gained by updating from the joint distribution to the product distribution, which are equal if and only if X and Y are independent.

Thus, the mutual information represents the value of the information that would be added upon learning that the background QUD and the new question are independent. If they are already independent, the mutual information is 0. This means that asking a question that is independent of the prior over the QUD alternatives should count as an irrelevant discourse move. A question is relevant to the extent that its joint distribution with the prior is different from what it would be if the question and the prior were independent.

So far, this gradient notion of question relevance is untested, but future work studying question-question pairs could use mutual information in exactly the way that we use KL divergence to look at assertions.

7. Conclusion

Relevance is a complicated topic that can be approached in many ways. Our focus, which we share with Roberts (2012), van Rooy (2004), and others, is on relevance as informativity with respect to the QUD. To this end, we have performed an exploratory experiment to test two gradient implementations of this basic idea. Though KL divergence has clear advantages over entropy reduction, neither metric does exactly what we want.

In ongoing and future work, we expand on this methodology in several ways. First, we elicit judgments on negative responses as well as positive ones, in order to eliminate the asymmetry in the data. Second, we ask participants for both their point probability estimates *and* their and confidence in those point estimates. We use these confidence judgments to quantify the participants' higher-order uncertainty, in case the effect of a response on higher-order uncertainty plays a role in judgments of relevance. Third, we test additional metrics, such as the Bayes Factor, which may have some advantages over entropy reduction and KL divergence.

In linguistic theory, relevance plays a major role in predicting acceptability judgments for discourses, and these judgments are not always clear cut. Our hope is that a successful gradient theory of relevance can be applied to discourse semantics and pragmatics to enrich the toolset of the discipline, and strengthen the connections between the study of meaning and the study of cognition.

[22] Notice that this definition is symmetric in X and Y, so the choice to use X as the new question and Y as the QUD is arbitrary.

References

Agha, O. and A. Warstadt (2020). Non-resolving responses to polar questions: A revision to the QUD theory of relevance. In M. Franke, N. Kompa, M. Liu, J. L. Mueller, and J. Schwab (Eds.), *Proceedings of Sinn und Bedeutung*, Volume 24, pp. 17–34. Osnabrück University and Humboldt University Berlin.

Büring, D. (2003). On D-trees, beans, and B-accents. *Linguistics and Philosophy 26*(5), 511–545.

Chen, T., Z. Jiang, A. Poliak, K. Sakaguchi, and B. Van Durme (2020). Uncertain natural language inference. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, pp. 8772–8779. Association for Computational Linguistics.

Cover, T. M. and J. A. Thomas (1991). *Elements of Information Theory*. New York: Wiley.

Ginzburg, J. (1995). Resolving questions, I. *Linguistics and Philosophy 18*(5), 459–527.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. L. Morgan (Eds.), *Speech Acts*, pp. 41–58. Brill.

Groenendijk, J. and M. J. B. Stokhof (1984). *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph. D. thesis, University of Amsterdam.

Hawkins, R. X. D. and N. D. Goodman (2017). Why do you ask? The informational dynamics of questions and answers. Technical report.

Herbstritt, M. and M. Franke (2019). Complex probability expressions & higher-order uncertainty: Compositional semantics, probabilistic pragmatics & experimental data. *Cognition 186*, 50–71.

Hyska, M. A. (2015). Discourse-level information structure and the challenge of metadiscursives. Master's thesis.

Morey, R. D., J.-W. Romeijn, and J. N. Rouder (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology 72*, 6–18.

Nelson, J. D., C. R. McKenzie, G. W. Cottrell, and T. J. Sejnowski (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science 21*(7), 960–969.

Oaksford, M. and N. Chater (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review 101*(4), 608–631.

Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics 5*(6), 1–69.

Rothe, A., B. M. Lake, and T. M. Gureckis (2018). Do people ask good questions? *Computational Brain & Behavior 1*(1), 69–89.

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

van Dijk, T. A. (1995). Discourse semantics and ideology. *Discourse & Society 6*(2), 243–289.

van Rooy, R. (2004). Utility, informativity and protocols. *Journal of Philosophical Logic 33*(4), 389–419.

von der Malsburg, T., T. Poppels, and R. P. Levy (2020). Implicit gender bias in linguistic descriptions for expected events: The cases of the 2016 United States and 2017 United Kingdom elections. *Psychological Science 31*(2), 115–128.

Zhang, H. and L. Maloney (2012). Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience 6*.