QUD variability in naturalistic discourse predicts scalar inference variability

Background: Question under Discussion (QUD) is a key linguistic concept for modeling the structure and dynamics of discourse. Most previous research on QUDs, however, has focused on highly controlled contexts with carefully curated examples, and researchers heavily rely on their own intuitions to reconstruct QUDs. A major challenge remains as to how to identify QUDs in naturalistic settings and in more principled ways, particularly given that different comprehenders may have divergent intuitions about which QUD is most salient. Using naturalistic discourse and scalar implicature as a case study, we develop experimental and computational methods to examine effect of QUD variability on pragmatic inferences. Recent literature has shown that interlocutors do not draw scalar implicatures (SIs) at the same rate across different lexical scales, and such inter-scale variation can be reduced when the *explicit* QUD is manipulated to make the relevant alternatives highly salient to comprehenders [2]. The study has two goals. First, we examine the hypothesis with *implicit* QUDs using a QUD elicitation task and an inference task in naturalistic discourse. If QUD plays a role in calculating SIs, then the rates of SI calculation across all scales should correlate with the proportion of elicited SI-relevant QUDs. Second, we evaluate if computational methods can help to predict whether QUDs are SI-relevant.

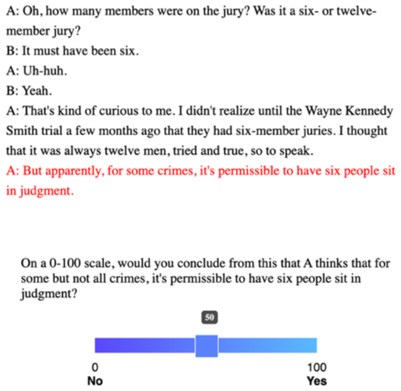
Human experiments: 37 <weak, strong> (e.g., <some, all>) lexical scales were selected from two existing datasets [2,3]. 37 short naturalistic conversations were then selected from [1], with each conversation containing a target “weak” scalar term in the last utterance of the conversation. In **Experiment 1**, participants (N=38) were presented with the 37 conversations. For each trial, participants read the entire conversation and were probed for the SI inference using a slider bar on a 0-100 scale (see an example in Figure 1). Robust cross-item SI variabilities were found (Figure 3). In **Experiment 2** the QUD elicitation task (N=120), we adapted the web-based data collection paradigm for crowdsourcing QUDs [4]: conversations were presented to participants and QUDs were probed at two different probing points (before and after the last utterance), see Figure 2. Since results from both probing points were similar, here we only report the results relevant for responses elicited at probing point 2.

Human annotation benchmark: Experiment 2 collected 1126 questions for each probing point. For each item, we manually clustered elicited questions into bins that represent only the unique questions based on semantic similarities. Next, we manually identified which bin(s) of the unique questions was the most relevant for the SI calculation. The item-specific proportion of the SI-relevant questions was then calculated.

SI-relevance prediction: Two computational tasks were developed for more efficiently quantifying SI-relevance in questions. In the **Question Clustering** task, we embedded each question together with its corresponding context using OpenAI text-embedding-3-large. KMeans and hierarchical clustering were performed with an optimal average k=5 for each item based on Elbow and Silhouette scores. Each cluster was then calculated a purity score, (i.e. the proportion of human-identified SI-relevant questions in each cluster). For each item, clusters with purity ≥ 0.5 were identified as SI-relevant. In the **Fine-tuning** task, we fine-tuned GPT-4o to examine if SI-relevant questions can be identified via a “question-answer” relation. The training dataset consists of 1029 examples from probing point 1. The test set contains 258 examples from probing point 1 and 1126 examples from probing point 2. We upsampled the training set to balance the SI/non-SI-relevant question distribution. Additionally, we established **LLM Zero-/Few-shot Baselines** with GPT-4o, GPT-4-turbo, and Llama 3.1 instruct 8B, with a prompt for zero-shot, few-shot, or zero-shot Chain-of-Thought (CoT) conditions. A sample prompt is shown in (2).

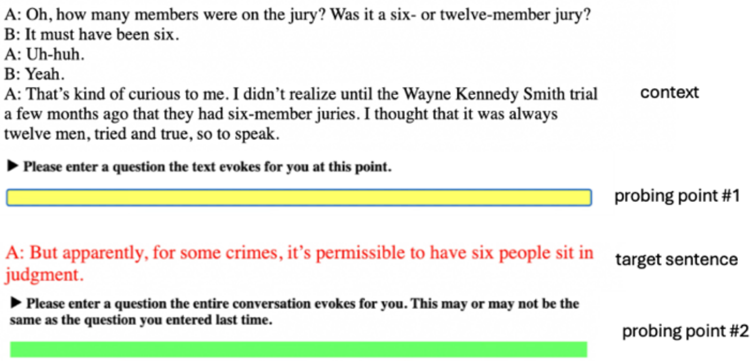
Results: To quantify the relation between QUD-variability and SI-variability, we constructed a mixed effects model predicting the SI rate from Experiment 1, with the proportion of SI-relevant-QUDs as the only predictor. In human annotation results, the QUD predictor had a significant effect (b=.39, t.= 3.76, p<0.001), see Figure 4. For the LLM methods, we measure their performance using three standard metrics: accuracy, macro F1, and Cohen’s Kappa. Additionally, each method was examined if the predicted SI-relevant question proportion is correlated with SI-rates. Table 1 shows that clustering and fine-tuning methods clearly outperform zero- or few-shot LLMs.

Discussion: The results showed that SI variability is predicted by the variability of contextually relevant QUDs. More importantly, experimental and computational methods can be leveraged to more precisely track the variability of QUDs in naturalistic contexts. This methodological advance opens up future directions for exploring a wider range of empirical domains at scale.



A: Oh, how many members were on the jury? Was it a six- or twelve-member jury?
B: It must have been six.
A: Uh-huh.
B: Yeah.
A: That's kind of curious to me. I didn't realize until the Wayne Kennedy Smith trial a few months ago that they had six-member juries. I thought that it was always twelve men, tried and true, so to speak.
A: But apparently, for some crimes, it's permissible to have six people sit in judgment.

On a 0-100 scale, would you conclude from this that A thinks that for some but not all crimes, it's permissible to have six people sit in judgment?



A: Oh, how many members were on the jury? Was it a six- or twelve-member jury?
B: It must have been six.
A: Uh-huh.
B: Yeah.
A: That's kind of curious to me. I didn't realize until the Wayne Kennedy Smith trial a few months ago that they had six-member juries. I thought that it was always twelve men, tried and true, so to speak.

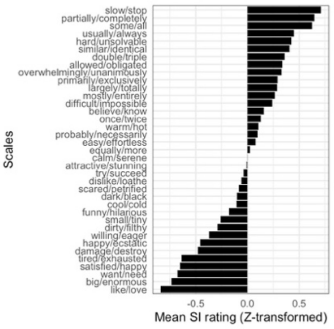
▶ Please enter a question the text evokes for you at this point.

A: But apparently, for some crimes, it's permissible to have six people sit in judgment.

▶ Please enter a question the entire conversation evokes for you. This may or may not be the same as the question you entered last time.

Figure 1: Example trial for Experiment 1

Figure 2: Example trial for Experiment 2



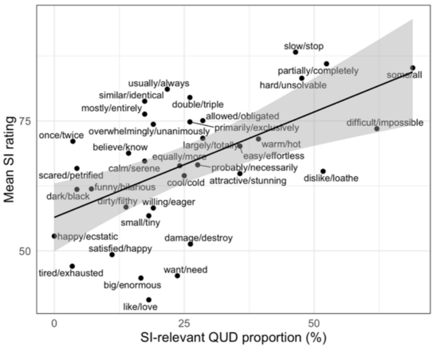


Figure3: SI rates for 37 different scales

Figure 4: Correlation between SI rate and QUD

(1) Instruction for fine-tuning GPT-4o: In the following context, judge whether this continuation answers the question. Provide an answer with 'Yes' or 'No'.
Input: Context: {context}
Question: {question}
Continuation: {continuation}

(2) LLM zero-shot: In the following context, judge whether this continuation answers the question. Answer ONLY with 'Yes' or 'No'.

Method/Model	Accuracy (%) ↑	macro-avg F1 ↑	Cohen's κ ↑	Estimate	t-value	p-value
kmeans	**85.17**	**0.800**	**0.601**	**0.294**	**4.307**	**<0.001**
hierarchical	**85.97**	**0.805**	**0.611**	**0.287**	**4.082**	**<0.001**
GPT-4o fine-tuned	**82.59**	**0.791**	**0.586**	**0.254**	**3.247**	**0.003**
GPT-4o zero-shot	68.29	0.542	0.088	-0.087	-0.388	0.700
GPT-4o few-shot	65.19	0.590	0.190	0.045	0.337	0.738
GPT-4o zero-shot CoT	68.29	0.538	0.082	-0.296	-1.492	0.145
GPT-4 zero-shot	72.11	0.537	0.102	-0.001	-0.006	0.995
GPT-4 few-shot	71.31	0.590	0.184	0.214	0.900	0.374
GPT-4 zero-shot CoT	67.94	0.543	0.090	-0.041	-0.210	0.835
Llama-3 zero-shot	60.75	0.543	0.101	0.174	1.236	0.225
Llama-3 few-shot	51.15	0.503	0.123	0.218	1.622	0.114
Llama-3 zero-shot CoT	63.77	0.364	0.102	0.322	1.480	0.148

Table 1: Model performance on the SI-relevance prediction task

References: [1] Godfrey, J., & Holliman, E. (1993). Switchboard-1 Release 2 LDC97S62. Linguistic Data Consortium, 34. [2] Ronai, E., & Xiang, M. (2024). What could have been said? Alternatives and variability in pragmatic inferences. [3] Shivade, C. et.al. (2015). Corpus-based discovery of semantic intensity scales. [4] Westera, M.et al. (2020). TED-Q: TED talks and the questions they evoke.