The Question Under Discussion (QUD) model has been an influential theoretical device in pragmatics, but efforts to derive QUDs from naturalistic data are few. In this work, we crowdsource QUD annotations of radio interviews. We address a fundamental issue at the center of QUD theory: can discourse agents reliably infer an implicit question or questions being addressed in naturalistic discourse? The secondary question we address is whether, as most QUD theories presuppose, there are multiple salient QUDs. We compare several similarity metrics for questions and answers, demonstrating that our user interface encourages annotators to obey useful theoretical constraints like Q-A Congruence. Overall, we find moderate annotator agreement forming qualitatively identifiable clusters, consistent with the existence of multiple contextually-restricted immediate QUDs. We further find, unexpectedly, that annotators are unreliable at reconstructing masked overt questions, suggesting that explicit questions may correspond to QUD/topic shifts.

Background Roberts 2012/1996 characterizes discourse as a game in which possible moves (utterances) are guided by whether they help answer the immediate QUD, usually a single implicit question. QUDs help formalize Gricean Relevance, important also in theories of focus, exhaustivity, coherence, etc. Existing resources fall broadly into two camps: rigorous, theory-grounded approaches, such as the hierarchical annotations in De Kuthy et al. 2018 and Hesse et al. 2020, albeit limited in scope by ontological complexity; or large, crowdsourcing approaches working with various kinds of implicit question, such as evoked questions (Westera et al., 2020) or elaborations (Wu et al., 2023), albeit not necessarily targeting theoretical properties of immediate QUDs.

Procedure We selected 10 complete two-party dialogue transcripts from INTERVIEW (Majumder et al., 2020), a corpus of NPR interviews in American English, split by sentence and annotated with turn information. Episodes were chosen to have between 29 and 32 sentences, of which at least 5 were overt questions (μ=5.5). 10 native English speakers per episode were recruited on Prolific, resulting in 100 unique sets of annotations. For each episode, annotators read the dialogue one sentence at a time, in a moving two-sentence window to simulate linear processing, as inspired by Westera et al. 2020. For each new sentence, annotators were prompted to (i) write a question that can be answered by that sentence, and (ii) select a contiguous span from that sentence best representing the answer to their question (Fig. 1). Annotators could opt to mark "no clear question" (e.g., for moves like *Good morning.*) While participants were free to write any question that the sentence addresses, we assume that discourse context makes certain potential QUDs more likely.

Evaluation We consider several similarity metrics for measuring QUD agreement. The first is *token edit distance* (ED), which counts the minimum number of words that must be inserted, deleted, or substituted to transform one array of tokens into another. This metric is useful for measuring answer similarity (μ=6.6), since all answers are forced by our interface to be subsets of the target sentence. Measuring similarity among questions is more challenging. Assuming Q-A Congruence, we hypothesize that annotators who select similar (low ED) answer spans are more likely to be writing similar QUDs, since they place focus on the same information. To test this hypothesis, we look at how answer ED correlates with three question similarity metrics: *question edit distance* (μ=8.0); rescaled *BERTScore* (Zhang et al., 2020) (μ=0.37), which encodes two sentences using a large transformer language model and measures the cosine similarity of their embeddings (between –1 and 1, where higher values are more similar); and *Wh-word agreement* (μ=0.39). Examples of these metrics applied to the collected data below in (1) are given in Table 1.

(1) a. Who else had been watching the radar? [**One of my graduate students**]
b. Who saw the occurrence and effects on the radar? [**my graduate student**]
c. Where are the clouds coming from? [**southwest about five miles**]

Results We find a moderate correlation for answer ED and question ED (Spearman's ρ=0.41), as well as for answer ED and BERTScore using DeBERTa (ρ=−0.37), the model recommended

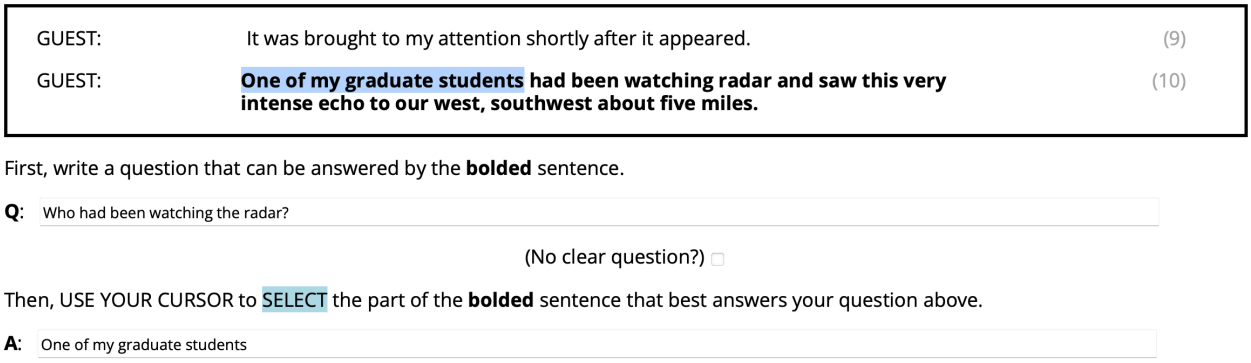

Fig. 1: Annotation interface. The answer box is auto-filled only by selecting from the bold sentence.

Metric	(a,b)	(b,c)	(a,c)
Token ED (A)	3	4	5
Token ED (Q)	6	8	7
BERTScore (Q)	0.41	0.14	0.12
Wh-word (Q)	1	0	0

Table 1: Similarity metrics on (1).

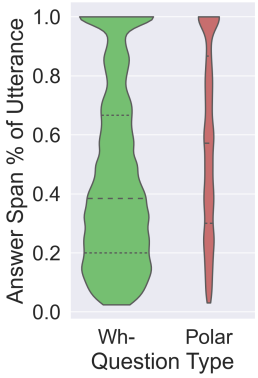

Fig. 2: Ans. spans.

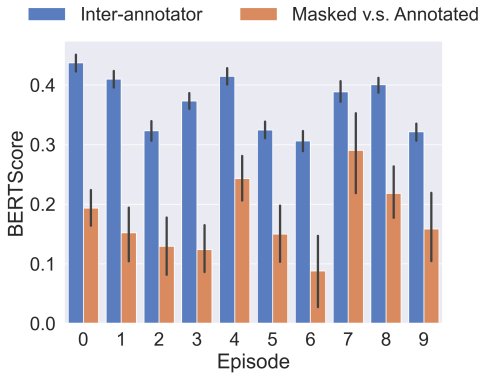

Fig. 3: Mean question similarity.

by the BERTScore authors. We also find correlations for Wh-word agreement with answer ED (ρ=−0.32) and BERTScore (ρ=0.49). The vast majority of QUDs written are Wh-questions, though polar questions exhibit an interesting pattern. With no explicit instruction to do so, for polar QUDs, annotators often select the entire sentence as their answer span (a response consistent with theoretical predictions about focus), while Wh-QUDs have short, constituent-sized spans (Fig. 2).

Masking questions Under most theories of QUDs, in normal circumstances, explicitly asked questions become the new QUD. To see whether annotator-written QUDs match actual questions, we masked all explicitly asked questions, keeping the sentence preceding it intact for context. We found that across episodes, annotators write QUDs consistently less similar to the masked question than to one another (Fig. 3), yet the mean inter-annotator BERTScore for QUDs on post-masked trials is not significantly different from inter-annotator agreement on normal trials. One possibility is that discourse participants may opt to ask explicit questions precisely in contexts with unpredictable topic shifts, making recovery difficult. Another is that our question similarity metrics fail to account for more general superquestions, a limitation of our linear, non-hierarchical method.

Conclusion Our results suggest that naturalistic discourse involves multiple compatible QUDs, but annotators are able to robustly extract these QUDs. The next step is extracting annotations about QUD hierarchy and relations among questions — a challenge we leave to future work.

References **De Kuthy et al. (2018)**. “QUD-based annotation of discourse structure and information structure”. *LREC 2018*. **Hesse et al. (2020)**. “Annotating QUDs for generating pragmatically rich texts”. *Workshop on discourse theories for text planning 2020*. **Majumder et al. (2020)**. “INTERVIEW”. *EMNLP 2020*. **Roberts (2012)**. “Information structure in discourse”. *Semantics and Pragmatics 5*. **Westera et al. (2020)**. “TED-Q”. *LREC 2020*. **Wu et al. (2023)**. “Elaborative simplification as implicit questions under discussion”. arXiv:2305.10387.