Pregnant Questions: The Importance of Pragmatic Awareness in Maternal Health Question Answering

Neha Srikanth[♢*] **Rupak Sarkar**[♢*]
Heran Mane[♠] **Elizabeth M. Aparicio**[♣] **Quynh C. Nguyen**[♠]
Rachel Rudinger[♢] **Jordan Lee Boyd-Graber**[♢]
[♢]Department of Computer Science, University of Maryland
[♠]Department of Epidemiology and Biostatistics, University of Maryland
[♣]Department of Behavioral and Community Health, University of Maryland
{nehasrik,rupak}@umd.edu

Abstract

Questions posed by information-seeking users often contain implicit false or potentially harmful assumptions. In a high-risk domain such as maternal and infant health, a question-answering system must recognize these pragmatic constraints and go beyond simply answering user questions, examining them in context to respond helpfully. To achieve this, we study assumptions and implications, or *pragmatic inferences*, made when mothers ask questions about pregnancy and infant care by collecting a dataset of 2,727 inferences from 500 questions across three diverse sources. We study how health experts naturally address these inferences when writing answers, and illustrate that informing existing QA pipelines with pragmatic inferences produces responses that are more complete, mitigating the propagation of harmful beliefs.

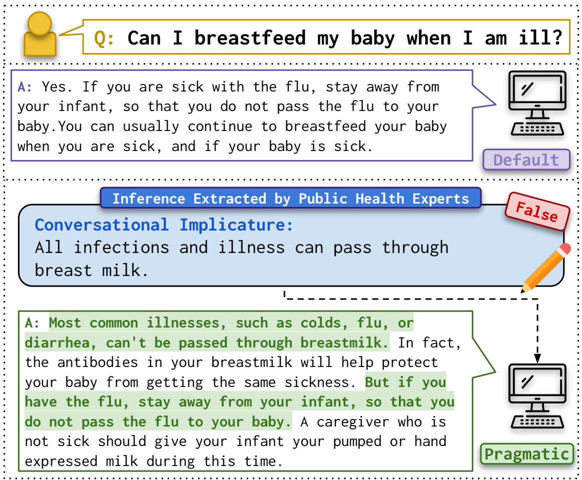

Figure 1: We ask public health experts to identify assumptions and implications from questions and find that incorporating them in a QA pipeline produces a more complete answer.

1 Introduction

Humans have varying information needs when they ask questions (Taylor, 1962). Sometimes these needs are easily inferred from the surface form, such as in factoid questions (e.g. *"Who is the 44th president of the United States"*). However, in a question such as "*Is there a good non-dairy baby milk I can supplement for my newborn?*", addressing the underlying false assumption "*Newborns can safely drink non-dairy milk*" **becomes part of satisfying the unexpressed information need**. Complete answers to these types of questions must not only address the surface question itself, but also "question the question", critically examining its *pragmatic needs*.

These needs become magnified in sensitive domains, such as consumer health or the legal domain. In these settings, addressing pragmatic needs of questions involves proactively addressing false assumptions or implications in questions to ensure that the asker does not continue holding inaccurate beliefs that they may act on. For example, a complete answer to the question about non-dairy milk for newborns should address that while non-dairy milk is viable for older babies, newborns and infants need human breast milk or dairy-based formula because they offer complete nutrition.

Language models have been shown to exhibit sycophancy (Sharma et al., 2023), sometimes adjusting responses to align with a human user's view. However, helpful QA systems should not only *challenge* false or subjective assumptions in questions (Kim et al., 2022) by verifying them against a vetted corpus, but also infer the asker's intent to make sure that its answer satisfactorily addresses their deeper information needs (Taylor, 1962), just as humans do.[1]

We construct a dataset[2] of 2,727 assumptions

[*] Equal contribution.

[1] "Pregnant" in our title also refers to its secondary definition, "full of meaning" (as in "a pregnant pause") alluding to the idea that questions are laden with implicit beliefs.

[2] https://github.com/nehasrikn/pragmatic-inferences-qa

	Question	Expert-Annotated Pragmatic Inference & Veracity
ROSIE (Maternal Health QA System)	`Is it okay for my to color my hair after giving birth?`	Hair dye chemicals can pass through breast milk from mother to child. (`False/Unsure`)
	`What is the advantage for not having an epidural during the labor?`	Avoiding an epidural contributes to a more "natural" and unmedicated birthing experience. (`False/Unsure`)
	`What cough medicine is appropriate for breastfeeding mothers?`	Some cough medicines can be secreted in breast milk. (`True`)
Reddit	`Is it safe to lay on my stomach at 28 weeks of pregnancy?`	Sleeping on the stomach while pregnant may have potential risks (`False/Unsure`)
	`Is it bad to use different bottles/nipples during feedings?`	Using different bottles or nipples for feeding may compromise the baby's latch. (`False/Unsure`)
	`How can I increase the time between feedings for my 3-month-old baby?`	It may be possible to sleep through the night while still ensuring the baby is fed. (`True`)
Natural Questions	`When does the fetus begin to develop memory?`	Fetuses have the ability to form memories. (`True`)
	`What causes a rupture in the amniotic sac?`	There may be ways to prevent early amniotic sac rupture. (`False/Unsure`)
	`When do the clinical manifestations of an ectopic pregnancy appear?`	There may be clinical manifestations of an ectopic pregnancy that do not appear early on. (`True`)

Table 1: Health experts identify pragmatic inferences from questions from three sources: Reddit, Natural Questions (Kwiatkowski et al., 2019), and questions asked to our domain-specific QA system, ROSIE (Mane et al., 2023). They also determine the veracity of each inference and provide supporting evidence from a trusted web document.

and implications in 500 questions (§2) collected from three diverse sources to study (1) how humans embed such assumptions and implications in questions, and (2) the extent to which they are *naturally* addressed in answers written by public health experts. We then ground assumptions and implications, two primary ways humans embed beliefs in questions, in existing linguistic theory of presuppositions and implicatures respectively (§3). We refer to presupposition and implicature collectively as *pragmatic inference*. While recent work has focused on the task of detecting and addressing false *presuppositions* in open-domain QA (Yu et al., 2022), we find that false beliefs of question askers are more likely to present as implicatures than presuppositions (§4). We experiment with *inducing* pragmatic behavior in existing QA pipelines with state-of-the-art retrieval and machine reading models (§5). On questions with at least one highly plausible false pragmatic inference, our expert annotators rated responses from our pragmatic QA system as more helpful and informative.

Thus, QA systems of the future must proactively address assumptions and implications in questions as they are increasingly deployed in sensitive domains.

2 Collecting Assumptions and Implications in the Wild

In contrast with factoid QA, systems deployed in *sensitive* domains such as consumer health must proactively mitigate harm. In these settings, correcting false assumptions is *not optional*: systems must provide contextual answers that balance *information completeness* with brevity.

Access to high quality healthcare in the United States vastly differs across socioeconomic backgrounds (Becker and Newsom, 2003). Such users are often likely to turn to accessible internet resources and—as of late—general purpose chatbots (Palanica et al., 2019). This motivates us to focus on maternal and infant care, a challenging area of consumer health where patients are concerned with *both* their own physical health as well as the health of their child.

To effectively study and induce pragmatic behavior in QA systems, the evaluation questions we choose must reflect real-world experiences and situations for which there may not be a straightforward answer explicitly addressed in a single web document. For example, answers to Natural Questions (Kwiatkowski et al., 2019, NQ)—a popular open-domain question-answering dataset—can be

	Maternal Health QA	Reddit	Natural Questions
# questions	200	200	100
ans. length (# sent)	3.9	6.6	5.6
# inferences	1161	1114	452
% false/subjective	22.5	30.8	20.1
% true	77.5	69.2	79.9

Table 2: Dataset statistics stratified by question source.

found directly in short extracted text snippets from Wikipedia (Table 1). In contrast, effectively answering the subjective questions sourced from Reddit requires commonsense reasoning and domain knowledge while identifying the asker's intent.

We carefully construct a dataset of questions from three distributionally distinct sources: a domain-specific QA system we design and deploy to pregnant and postpartum participants we recruit (Mane et al., 2023), Reddit, and NQ. Then, we introduce an annotation scheme to elicit assumptions and implications from these questions, validate their plausibility, and finally collect supporting evidence to determine their veracity. Our final dataset contains 2,727 assumptions from 500 evaluation questions (Table 2). We also include 150 development questions used to train annotators and develop our QA systems.

2.1 Gathering a Diverse Set of Maternal and Infant Health Questions

Maternal Health QA System. We source questions come from a maternal and infant health-specific question answering system that we build (Mane et al., 2023), henceforth referred to as ROSIE. Users ask questions pertaining to pregnancy or infant health and are instructed that the QA system does not have any personalized knowledge of their individual medical history or pregnancy.

This system operates over a corpus of web documents we construct[3] from trusted sources including United States governmental and hospital organizations on maternal and infant health, and spans salient topics such as pregnancy and postpartum symptoms, developmental milestones, and infant safety. Our end-to-end QA system, ROSIE, uses a passage retriever and reranker to provide web passages as answers to study participants via a mobile application. We randomly sample 200 anonymized questions asked to ROSIE for our evaluation set and 50 questions for our development set.

Reddit. While the questions asked to ROSIE do reflect real-world experiences, they are asked *to an automatic system* and thus tend to include less situational detail or implicit content. We turn to Reddit[4] to capture long-tail questions that are about the diverse set of unique situations a new or expecting parent goes through. Table 1 highlights some distributional differences between questions from Reddit and other data sources. Our questions come from four popular subreddits about maternal and infant health: `r/BabyBumps`, `r/breastfeeding`, `r/NewParents`, `r/Mommit`, and `r/beyondthebump` from the `pushshift`[5] dump.

We develop a series of heuristics as a recall-oriented first step to identify questions with false or subjective assumptions. We begin by selecting questions where an upvoted comment shows assumption-correcting behavior or where a user invokes their medical expertise, identified by a select list of discourse markers (Appendix A). Of these, we only retain posts beginning with a "wh" word, filtering a few hundred thousand posts down to 2,858 questions.

As Reddit encourages community participation, many questions are "community seeking" as opposed to *information*-seeking. To identify information-seeking questions, we use GPT-3.5 (Ouyang et al., 2022) to filter medical questions from non-medical questions (Prompt A.1) then manually vet the final set of 297 questions. We randomly sample 200 questions for our evaluation dataset and 50 questions for our development set, discarding the rest.

Titles of Reddit posts are often a hook or a summary of the entire post. Using the 50 development questions, we use GPT-3.5 to minimally edit the titular question to include crucial details from the post description, providing a series of exemplars (Prompt A.2). These rewrites mainly include the age of a newborn or the stage of pregnancy from the description, but sometimes include small situational details that contextualize the question. Two authors validate all rewrites, keeping the original title wherever both authors agree that the rewrite changed the communicative goal of the asker.

Natural Questions. Lastly, we include maternal and infant health questions from NQ to study pragmatic aspects of factoid-style questions. We embed all questions in the train set of NQ using the `sentence-transformers` (Reimers and Gurevych, 2019) implementation of `all-mpnet-base-v2` (Song et al., 2020), including unanswerable questions (Asai and Choi, 2021). We identify 2500 answerable questions and 2500 unanswerable questions as maternal health-related by identifying the top 100 nearest

[3]Corpus available upon request. We use Barbaresi (2021) to scrape 408,000 web documents which we split into passages of 100 tokens following Karpukhin et al. (2020).

[4]https://www.reddit.com/

[5]https://github.com/Watchful1/PushshiftDumps

neighbors of 50 randomly sampled questions from the development sets of Reddit and ROSIE.[6] From this set, we randomly sample 100[7] questions for our evaluation set and 50 for our development set. Though obtained with a nearest neighbors approach, these questions greatly differ from those obtained from our previous sources, as they reflect the factoid QA-oriented tasks and goals of the original dataset creators (Table 2).

Collecting Human Answers from Health Experts. We recruit a team of twenty health experts using Upwork[8] to annotate our data including obstetricians and gynecologists (OB/GYNs), nurses, lactation consultants, and public health experts, many of whom have experience with patients. In addition, many of these expert annotators are either currently pregnant or postpartum or have been in the past. We ask a subset of six experts to write helpful and informative long-form answers to all 500 questions in our dataset (Figure 6, bottom panel). While annotators write answers from scratch, they must provide supporting web documents from the same list of verified sources we use to build the corpus for ROSIE.

2.2 Identifying Assumptions and Implications

Inferring *possible* assumptions, implications, and asker beliefs from patient questions in our domain are challenging. In the past, others have extracted assumptions using shallow signals from the surface form of a question (Kim et al., 2021; Parrish et al., 2021). While some assumptions or implications in our dataset can be inferred directly from the question expression, others require deeper domain or experiential knowledge (Table 1).

Eliciting these assumptions and implications from non-linguists is challenging as existing linguistic frameworks (§3.1) are inaccessible or cumbersome for those unfamiliar with the theoretical concepts behind them. As such, we operationalize large-scale data collection by asking five annotators from a *different* subset of our expert annotator pool to first write a list of subquestions that an answer to the original question would address (Figure 6, top panel). Doing so primes annotators to reason about the intent behind a question as well as the information needs of an asker. Then, we ask them to write a set of sentences reflecting possible beliefs or assumptions that the patient may hold (or, alternatively, beliefs that any complete answer to the question must address). We emphasize that the assumptions they write can be either medically or scientifically true or false.

Then, we *consolidate* the set of subquestions and human-written assumptions and beliefs into a single set of assumptions and implications using GPT-3.5 (Prompt B.1).

2.3 Annotating Inference Veracity

Lastly, we ask a new subset of eight expert annotators to annotate whether each assumption and implication in our dataset is medically or scientifically true, false, or subjective and provide a supporting web document from our list of verified sources along with a passage from the document (Figure 6, middle panel).

Validation. To verify that the assumptions and implications we extract are plausibly inferrable from the question, we recruit an additional pair of health experts, which we refer to as expert *validators*, to rate inferences.[9] We sample 100 assumptions and implications judged as false or subjective, and 100 true inferences and ask our expert validators to rate the plausibility of an inference on a 1–5 Likert scale based on how likely the question asker is to believe the assumption or implication. Henceforth, we refer to this sample of 200 inferences coming from 152 unique questions as INFERENCESAMPLE. Both annotators judge the majority of our inferences as plausible, with 80% and 95% rated with a score of at least 3 (see Figure 4 for the rating scale). Spearman's correlation between the two annotators is 0.69. See Appendix C for more detail.

3 Grounding Assumptions and Implications in Linguistic Theory

Assumptions and implications in our dataset map to two well-studied phenomena in linguistic pragmatics: presupposition and implicature (Grice, 1975; Stalnaker et al., 1977). We begin with a short primer of both types of *pragmatic inference* (§ 3.1) and then discuss the implications of both types in a QA setting (§ 3.2).

[6] We tried several different filtering heuristics, including keyword-based detection, but the nearest neighbors approach yielded the most topical questions.

[7] NQ questions make up a smaller proportion of our evaluation dataset as we avoid diverting large amounts of annotation resources to factoid-style questions.

[8] https://www.upwork.com/

[9] These expert validators were not a part of our dataset construction.

3.1 Two Types of Pragmatic Inference: Presupposition and Implicature

A sentence S is a *pragmatic inference* of a question Q if, depending on the *context* and *conversational goals* of discourse participants (Jeretic et al., 2020), a human would believe that the asker of Q believes or assumes S to be true. Henceforth, we refer to the assumptions and implications that we collect in our dataset as *pragmatic inferences*. We review the two most relevant types of pragmatic inferences: *presupposition* and *implicature*.

Presupposition. Presuppositions are implicit assumptions in utterances taken for granted by discourse participants (Beaver, 1997). The question *"What vitamins should I stop taking after becoming pregnant?"* presupposes *"I was taking vitamins before becoming pregnant."* Presuppositions can often be detected solely by the presence of a lexical or syntactic trigger (Levinson et al., 1983). In the example above, the word *stop* presupposes that an activity was already in motion. We refer to these presuppositions as "trigger-based".

As we observe during the collection of our dataset, domain or world knowledge is often needed to capture presuppositions in real-world data that are not apparent from lexical or syntactic cues (Abusch, 2002). For example, the question *"Are multiple ultrasounds dangerous for my baby?"* does not directly result in non-trivial trigger-based presuppositions. However, the asker of the question presupposes that the effects of an ultrasound are additive and hence asks about whether that additive effective is *dangerous*.

Implicature. Implicature is a type of pragmatic inference that is *suggested* by an utterance as opposed to part of its literal meaning (Grice, 1975). Consider the question *"Do most babies fit in newborn clothes?"* While the speaker understands that newborn clothes fit *some* babies, their question implies that not all babies fit in newborn clothes. As we discover, a significant portion of inferences in our dataset are *implied* from questions rather than presupposed, but detecting and generating implicatures remain understudied in NLP.

Some implicatures are related to lexical items or syntactic structure of utterances. For example, the statement *"These prenatal vitamins are in gumdrop form, but are healthy"* implies that gumdrops are usually *not* healthy. Others are a function of a speaker's intent, beliefs, and other contextual elements (Zheng et al., 2021). While they are a part of the content of an utterance, these implicatures are not *at-issue* (e.g. the main point under discussion (Potts, 2004; Koev, 2018)) and are not encoded by the linguistic properties of a sentence (Allott, 2018). Consider the question *"How can you tell the difference between postpartum depression and exhaustion?"*. Reasoning about asker belief, we may conclude that they are implying that the two conditions should be treated differently, as one is more serious than the other.

3.2 Presupposition and Implicature in QA

In a natural setting, as we discover, humans embed both presuppositions and implicatures nearly equally in questions (§4). However, from a linguistic perspective, they represent different levels of an asker's commitment to the propositional content of the inference (Peters, 2016). Presuppositions are already a part of an asker's world model. In contrast, implicatures are *likely* beliefs that may be negated in an asker's subsequent utterances. Consider the question *"Is it normal for my baby to move more than usual when closer to due date?"* with both the presupposition *"There are factors that contribute to changes in fetal movement as the due date approaches"* and the implicature *"It may not be necessary to be concerned if there is a significant increase in fetal movement close to the due date."* While both are false, the presupposition is stronger, and is clearly in need of addressing in a potential answer. As illustrated, these distinct phenomena must be dealt with differently when answering a question.

Related Work. Existing work in pragmatics in QA focuses on open-domain question answering. Kim et al. (2021) present the first study of presuppositions in Google search queries using the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019) that are unanswerable due to false presuppositions. However, their system only addresses trigger-based presuppositions, overlooking the type of deeper presuppositions present in our dataset derived from world or domain knowledge. Other work has looked at Google queries with *questionable* assumptions (Kim et al., 2022) and false presuppositions in open-domain Reddit questions (Yu et al., 2022). Computational studies of implicature have only focused on specific types, such as scalar implicature (e.g., some $X \rightarrow$ not all X) (Schuster et al., 2020; Zheng et al., 2021; Kabbara and Che-

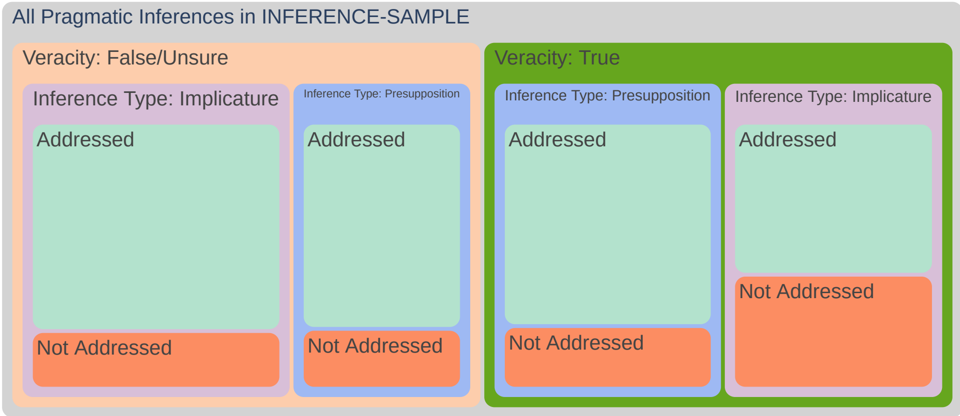

Figure 2: Tree-map (Shneiderman, 1992) visualizing the distribution of expert-annotated pragmatic inferences in INFERENCE-SAMPLE with their veracity, inference type, and whether or not they were addressed or not in the expert-written answer to the question from which they came from. When users make false or subjective inferences, they are more likely to do so as an implicature. Moreover, when an inference is false, it is more likely to be naturally addressed in an answer by public health experts.

ung, 2022; Jeretic et al., 2020). As a result of the context induced by our domain, implicatures in our dataset extend beyond scalar implicature.

4 How to ask and answer questions

Before we investigate the behavior of QA systems, we first study how humans embed pragmatic inferences in their questions (§4.1) as well as to what extent they are *naturally* addressed by human public health experts (§4.2).

4.1 Pragmatic Inference Type: Understanding Speaker Commitment

When users ask questions, how strongly are they committed to the inferences that experts identify in their questions? Presupposition is a phenomenon based on *mutual* acknowledgment of facts: when a human makes a presupposition, not only are they presuming the content of the inference, they are also signaling the belief that their *interlocutor* (here, a QA system) should believe it too.

On the other hand, implicatures are a softer way for humans to express uncertainty. For example, *"Which immunity injections can I skip for my baby?"* and *"Is it sufficient if my baby takes most immunity injections"* have the same underlying inference (*"It is okay to pick and choose vaccines"*), but is taken for granted in the first (presupposition), whereas loosely suggested in the second (implicature). We want to distinguish inferences—separating implicatures from presuppositions—in our questions to better characterize *so that we can prioritize addressing stronger false inferences.*

Annotation Framework. A pair of authors independently annotate all inferences in INFERENCE-SAMPLE as a presupposition or implicature by first determining whether it is a proposition about the world that the asker believes to be true, without which the question would not be felicitous (presupposition) or whether it involves deriving asker belief through communicative principles (implicature). Between authors, Cohen's kappa is $\kappa = 0.85$, indicating strong agreement. Author annotators adjudicated the final inference type (see Figure 2 for overall distribution), but individual annotator labels and adjudication rationales are preserved as a part of our dataset.

Findings. Presuppositions and implicatures are balanced in INFERENCE-SAMPLE (Figure 2), with a slight majority of inferences as implicatures, indicating that many inferences that health expert annotators identify are more subtle. When an inference is true, it is almost equally likely to be a presupposition or an implicature. However, when users make false or subjective (veracity marked "Unsure") inferences, **they are more likely to do so via implicature** (Figure 2). Past work has looked into generating and verifying *presuppositions* in open-domain QA, but identifying and addressing *implicatures* in an effort to make answers information-complete remains heavily underexplored. This finding highlights a key strength of our work: the ensuing context from our specific domain tests the usefulness of pragmatic inference in QA by allowing us to extract a greater range of inferences.

In settings that lack such context (e.g. single-turn

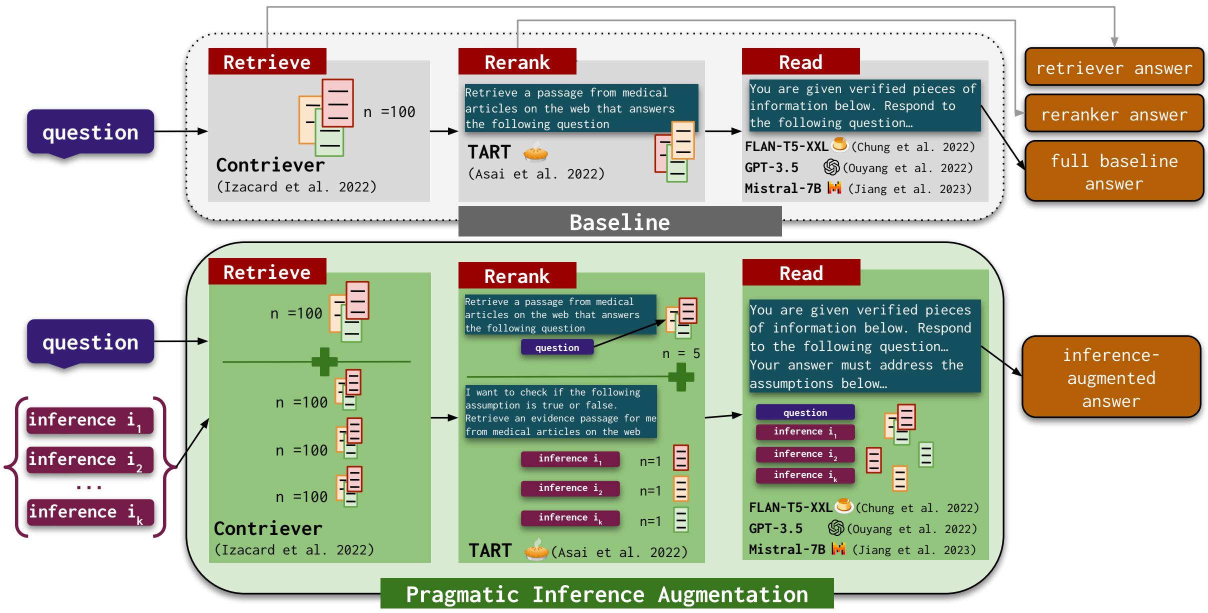

Figure 3: Our baseline and pragmatic inference-augmented QA pipelines. We experiment with retrieval, reranking, and reading stages and a variety of instruction-tuned and prompt-based models.

open-domain QA), we are restricted to leveraging lexical or syntactic signals from the surface form of the question (Kim et al., 2021) since reasoning about asker belief is not possible without other contextual signals. For example, in the absence of additional context, the question *"Should I push grandparents for flu shot and tdap?"* may give rise to inferences involving the safety or effectiveness of these vaccines for the elderly. However, upon learning that this was asked in a web forum by a postpartum mother, we may reason that she believes her infant may be at risk for contracting the flu or other diseases if their grandparents handle them unvaccinated.

4.2 Addressing Inferences in Expert Answers

When health experts are tasked with *answering* questions, how likely are they to naturally address inferences that users implicitly make? Studying whether or not answers *naturally* address pragmatic inferences (§ 4.2) gives us better insight into the types of inference health experts, and in turn models, should prioritize when answering questions.

Annotation. We ask two annotators from our expert annotator pool to determine whether each inference in INFERENCE-SAMPLE is addressed, either implicitly or explicitly, by the human-written answer to its source question.

Findings. The majority of inferences in INFERENCE-SAMPLE are addressed by the human-written answer *naturally* (Figure 2). Importantly, **when an inference is false, it is more likely to be naturally addressed.** Moreover, a significant number of *true* inferences are also addressed by an answer, indicating that health experts not only aim to correct false or subject inferences but also prioritize completeness. This key finding supports one of the main arguments of this work: QA systems *must* address pragmatic inferences in their answers, just as humans do.

5 Inducing Pragmatic Behavior in QA

Inducing pragmatic behavior in QA systems is not straightforward. Existing systems are not trained to proactively reason about asker beliefs, since many popular QA datasets do not necessitate this type of behavior (e.g. factoid QA).

We experiment with eliciting model answers that address the pragmatic needs of questions, such as refuting false inferences, using the pragmatic inferences in our dataset. We inject inferences at each stage of the classic QA pipeline: passage retrieval, reranking, and machine reading (§5.1) and evaluate outputs against expert-written answers with both automatic and human evaluation (§5.2).

5.1 Experimental Setup

Corpus. We use the corpus from Mane et al. (2023) of 408,000 documents from verified web sources on maternal health and infant care and augment the corpus with the sources that our expert annotators found while both writing answers and determining the veracity of inferences.

Baseline Models. As a baseline system, we use a retrieval, reranking, and reading-based QA pipeline. Contriever (TIzacard et al., 2022), an unsupervised dense passage retriever, identifies top relevant documents ($n = 100$) in our corpus given a question. Those documents are reranked using TART-full (Asai et al., 2022), a multi-task retrieval system with a cross-encoder architecture (Instruction E.1). TART is instruction-tuned, equipping it with the flexibility to redefine passage relevance for different tasks. We feed the top five reranked documents to three different reader models: FLAN-T5-XXL (Chung et al., 2022, 11 billion parameters), an instruction-tuned, *prompt*-based encoder-decoder model jointly trained on a multiple tasks with a standard answer extraction prompt from Mishra et al. (2022) (Instruction E.3), MISTRAL-7B (Jiang et al., 2023) (an open source large language model, Prompt E.4), and GPT-3.5 (Prompt E.4).

Augmenting Systems with Pragmatic Inferences. In addition to retrieving the top 100 passages using the question as input, we retrieve the top 100 passages for each pragmatic inference of the question ($i_1 ... i_k$) as well. Then, for each pragmatic inference i, we rerank the top 100 passages using a new inference-informed instruction (Instruction E.2) and select the top passage post-reranking. We augment the top five reranked passages from the question with these k top passages from each pragmatic inference to feed to each reader (Prompt E.5). During reading, we prompt MISTRAL-7B and GPT-3.5 to address all k assumptions when generating an answer.[10] To keep the same *number* of passages fed to readers in the baseline pipelines as in the inference-augmented pipeline, we add k extra passages to the top five existing ones. This ensures that while the volume of information presented to machine readers is the same in both pipelines, the nature of the content differs, allowing us to measure the utility of inference augmentation during retrieval and reranking. Figure 3 visualizes our baseline and inference-augmented QA pipelines.

5.2 Evaluation

We evaluate answers from seven pipeline variations against expert answers (Table 3): (1) The **No Reader** baselines consist of the top retrieved passage based on the input question (RETRIEVE-ONLY) and the top reranked passage from the input question (RERANK-ONLY), (2) three **Baseline Readers** (BASELINE-FLAN-T5-XXL, BASELINE-MISTRAL-7B, and BASELINE-GPT-3.5), and (3) two **Inference-Augmented** pipelines INFERENCE-MISTRAL-7B, and INFERENCE-GPT-3.5).

Automatic Evaluation Metrics. Three automatic evaluation metrics measure the quality of generated answers: ROUGE (Lin, 2004) (both F1 and recall),BLEURT (Sellam et al., 2020), and QAFACTEVAL (Fabbri et al., 2022), a more recent QA evaluation metric originally designed to measure the faithfulness of summaries. GPT-3.5 scores the strongest according to QAFACTEVAL, our main evaluation metric because it—of the three metrics—most closely captures information content. However, automatic evaluation of generated answers does *not* capture several higher-level semantic and pragmatic aspects of the question. Thus, we still need experts to validate the answers.

Human Judgments. We ask our expert *validators* to score answers from the top-performing baseline and inference-augmented pipeline (BASELINE-GPT-3.5 and INFERENCE-GPT-3.5, on QAFACTEVAL respectively). For each of the 152 questions in INFERENCE-SAMPLE, expert validators score both model outputs simultaneously from 1–5 based on completeness (instructions in Figure 5). Answers typically received a score of 1 when they were off-topic and missing crucial information, a score of 2 when they were topical but still missing crucial information, 3 when containing all essential information to the question, 4 when most information was present for completeness, and a score of 5 when the answer was information complete. Judging the information completeness of an answer is a subjective task, as reflected by the Spearman rank correlation between their annotations ($\rho = 0.34$). While the mean score of inference-augmented examples is comparable to baseline answers (4.43 vs. 4.45), annotators rated the inference-augmented answer as equivalent or better than its baseline counterpart in 75% of questions in INFERENCE-SAMPLE (see Table 5 for examples).

We further focus on annotator preferences on our original motivating population of questions—those with highly plausible, false assumptions. Both annotators rate inference-augmented answers higher

[10] We do not use FLAN-T5-XXL here because it struggled with reading in the baseline setting.

	No Reader		Baseline Reader			Inference-Augmented	
	RETRIEVE	RERANK	FLAN-T5-XXL	MISTRAL-7B	GPT-3.5	MISTRAL-7B	GPT-3.5
ROUGE-L (F1)	$15.6_{(10.3)}$	$17.5_{(10.3)}$	$15.89_{(13.7)}$	$17.4_{(5.1)}$	$18.7_{(6.0)}$	$16.2_{(4.8)}$	$18.6_{(5.9)}$
ROUGE-L (Recall)	$23.0_{(15.7)}$	$19.5_{(10.9)}$	$36.9_{(21.9)}$	$15.2_{(6.8)}$	$17.4_{(8.1)}$	$13.2_{(5.7)}$	$16.6_{(7.6)}$
BLEURT	$-0.72_{(0.34)}$	$-0.61_{(0.29)}$	$-0.79_{(0.45)}$	$-0.38_{(0.22)}$	$-0.37_{(0.22)}$	$-0.36_{(0.21)}$	$-0.38_{(0.22)}$
QAFACTEVAL	$0.69_{(1.05)}$	$0.76_{(0.80)}$	$1.16_{(1.5)}$	$1.02_{(0.65)}$	$1.15_{(0.73)}$	$0.96_{(0.62)}$	$1.17_{(0.75)}$
Human (/5)	-	-	-	-	4.43	-	4.45

Table 3: Mean and standard deviations of automatic (ROUGE, BLEURT, QAFACTEVAL) and human evaluation metrics per question. We report results for the top retrieved passage and the top reranked passage, and two modes with and without access to human-written assumptions. Inference-augmented models perform competitively with baselines, indicating the promise of inducing pragmatic behavior in QA models to mitigate harm.

than the default answers in the subset of questions with at least one *highly plausible*, (plausibility=5) *false* pragmatic inference (Table 4). We hypothesize that the similar ratings received by the two systems across *all* questions is due to shortcomings in the instruction-following capabilities of LLMs. Forcing the reader model to address pragmatic inferences distracts it from answering the question more completely, and does not always result in more helpful answers when the pragmatic inferences are true. This illustrates the promise of inducing pragmatic behavior in QA models and represents a lower bound: none of the models we experiment with were trained to optimize for addressing assumptions in questions.

6 Can inference extraction be automated?

While pragmatic inferences elicited from health experts are informed by their expertise, they are slow and costly to collect. Our QA experiments use *human-written inferences* to establish an upper bound of answer quality with existing models. However, a fully automatic pragmatic QA pipeline must first generate pragmatic inferences relevant to a question and *then* generate an answer that addresses the subset of false inferences. As such, we experiment with generating pragmatic inferences with GPT-3.5 (Ouyang et al., 2022) to understand to what extent automating the process is feasible with existing prompting and in-context learning.

Experimental Setup. We generate inferences with GPT-3.5 for all questions in INFERENCE-SAMPLE using seven in-context examples corresponding to 37 different pragmatic inferences, as more in-context examples yields diminishing returns (Liu et al., 2022). We select pragmatic inferences written by multiple expert annotators from diverse user questions and randomly shuffle them to prevent unwanted effects emerging from example order (Si et al., 2022), including exemplars from all three sources (ROSIE, Reddit, and NQ) to capture distributional differences in their pragmatic inferences. As humans naturally did, we let GPT-3.5 generate *varied* numbers of inferences per question.

Evaluation: Can GPT-3.5 generate human-like pragmatic inferences? For each inference in INFERENCE-SAMPLE, a pair of authors annotate whether or not each human-written assumption is semantically equivalent to at least one inference generated by GPT-3.5 (Prompt F.1) with a Cohen's kappa of $\kappa = 0.88$. Post-adjudication, 63% of inferences were not present among model generations. When stratifying by inference type, 53% of presuppositions and 71% of implicatures were not present. This illustrates that just as they are more difficult to detect, implicatures grounded in domain knowledge are more difficult for language models to generate.

7 Conclusion

We show that it is possible to induce pragmatic behavior in QA systems to correct latent false assumptions in the sensitive domain of maternal and infant health. Next-generation QA systems deployed in real-world settings *must* learn to address the pragmatics of user questions. Though we have shown the viability of *explicitly* inducing pragmatic behavior in models in this work, directions for future work include training retrievers to inherently search for evidence to address pragmatic inferences and readers to reason on top of such evidence to tactfully and effectively challenge user beliefs.

Acknowledgments

We thank Philip Resnik, Sweta Agarwal, Sathvik Nair, and other members of the University of Maryland CLIP lab for their helpful feedback. We also thank Ximena Marin Gutierrez, Marina Yue, Michelle Jasczynski, and Amara Channell Doig for their help in the development of ROSIE. We are grateful for our pool of expert annotators and for users of ROSIE participating in our clinical trial.

The study reported in this paper was supported by research grants from the National Institute on Minority Health and Health Disparities (grant number R01MD016037) and by the National Library of Medicine (grant number R01LM012849). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the paper.

Limitations

Data are not multilingual. Although our participants who provide questions come from diverse socioeconomic and racial backgrounds, all of the data we collect are in English. In addition, since we require participants to be located in the United States, the questions provided by participants are only reflective of the healthcare needs of English-speaking residents in the United States.

Choice of a single domain. While our approach can be generalized to any other domain, all of our data and experiments are confined to a single domain (maternal and infant health). We have not validated that our conclusions generalize beyond this particular important domain.

Pragmatics can be annotator-dependent. Finally, some degree of pragmatic inference is always dependent on the annotator, and we have not validated that this is consistent across different annotator backgrounds.

Ethical Considerations

NLP systems are never a replacement for doctors or clinical expertise, especially in high-stakes settings. This work has grown out of collaboration with public health experts to help disseminate medically accurate but *contextual* information to new or expectant mothers with limited access to healthcare. Upon detection of false or potentially problematic assumptions, patients can then be referred to healthcare providers better able to provide information than current QA systems. All of our data was collected with IRB approval in consultation with public health professionals.

References

Dorit Abusch. 2002. Lexical alternatives as a source of pragmatic presuppositions. In *Semantics and linguistic theory*, volume 12, pages 1–19.

Nicholas Allott. 2018. Conversational implicature.

Akari Asai and Eunsol Choi. 2021. Challenges in information-seeking QA: Unanswerable questions and paragraph retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1492–1504, Online. Association for Computational Linguistics.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2022. Task-aware retrieval with instructions. *arXiv preprint arXiv:2211.09260*.

Adrien Barbaresi. 2021. Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 122–131. Association for Computational Linguistics.

David Ian Beaver. 1997. Presupposition. In *Handbook of logic and language*, pages 939–1008. Elsevier.

Gay Becker and Edwina Newsom. 2003. Socioeconomic status and dissatisfaction with health care among chronically ill african americans. *American journal of public health*, 93(5):742–748.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

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

Jad Kabbara and Jackie Chi Kit Cheung. 2022. Investigating the performance of transformer-based NLI models on presuppositional inferences. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 779–785, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Najoung Kim, Phu Mon Htut, Sam Bowman, and Jackson Owen Petty. 2022. (qa)2: Question answering with questionable assumptions. *ArXiv*, abs/2212.10003.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3932–3945, Online. Association for Computational Linguistics.

Todor Koev. 2018. Notions of at-issueness. *Language and Linguistics Compass*, 12(12):e12306.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Stephen C Levinson, Stephen C Levinson, and S Levinson. 1983. *Pragmatics*. Cambridge university press.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Heran Y Mane, Amara Channell Doig, Francia Ximena Marin Gutierrez, Michelle Jasczynski, Xiaohe Yue, Neha Pundlik Srikanth, Sourabh Mane, Abby Sun, Rachel Ann Moats, Pragat Patel, et al. 2023. Practical guidance for the development of rosie, a health education question-and-answer chatbot for new mothers. *Journal of Public Health Management and Practice*, 29(5):663–670.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *ACL*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Adam Palanica, Peter Flaschner, Anirudh Thommandram, Michael Li, and Yan Fossat. 2019. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. *Journal of medical Internet research*, 21(4):e12887.

Alicia Parrish, Sebastian Schuster, Alex Warstadt, Omar Agha, Soo-Hwan Lee, Zhuoye Zhao, Samuel R. Bowman, and Tal Linzen. 2021. NOPE: A corpus of naturally-occurring presuppositions in English. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 349–366, Online. Association for Computational Linguistics.

Stanley Peters. 2016. Speaker commitments: presupposition. In *Semantics and Linguistic Theory*, pages 1083–1098.

Christopher Potts. 2004. *The logic of conventional implicatures*, volume 7. OUP Oxford.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Ben Shneiderman. 1992. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on graphics (TOG)*, 11(1):92–99.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *CoRR*, abs/2004.09297.

Robert Stalnaker, Milton K Munitz, and Peter Unger. 1977. Pragmatic presuppositions. In *Proceedings of the Texas conference on per˜formatives, presuppositions, and implicatures. Arlington, VA: Center for Applied Linguistics*, pages 135–148. ERIC.

Robert S Taylor. 1962. The process of asking questions. *American documentation*, 13(4):391–396.

Gautier TIzacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. In *Transactions on Machine Learning Research*.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Xinyan Velocity Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. Crepe: Open-domain question answering with false presuppositions. *ArXiv*, abs/2211.17257.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. 2021. GRICE: A grammar-based dataset for recovering implicature and conversational rEasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2074–2085, Online. Association for Computational Linguistics.

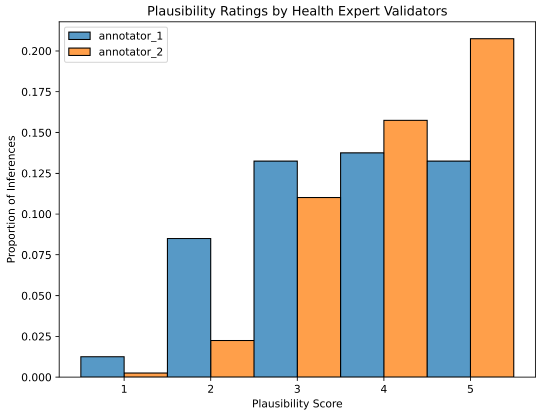

Figure 4: Ratings of expert validators of the plausibility of inferences written by health experts in our dataset. The majority of inferences are plausible.

A Reddit Question Filtering

A.1 Discourse Markers

Assumption Correcting Markers: "however,", "actually,", "as a matter of fact", "in fact", "not true", "despite what you", "on the contrary", "common misconception", "not exactly", "just to clarify", "you're confusing", "correct me if i'm wrong", "correct me if i am wrong", "you're wrong", "we have to remember that", "while that's true", "could be dangerous", "might not be the best thing"

Expertise Invoking Markers: "as a doctor", "as a medical professional", "i'm a doctor", "being a doctor", "as a nurse", "i'm a nurse", "i'm a medical professional", "being a nurse"

Prompt A.1: Medical vs. Non-Medical Question Identification

`Prompt:` You are an expert in maternal and infant health who specializes in finding out whether a question posed by a new or expecting mother is seeking opinion/community participation, or whether it is a medical question. Given a question, you must answer whether it is question seeking medical advice or if it is seeking personal anecdotes and sharing experience. If it's seeking medical advice, answer with "medical". Otherwise, answer "non-medical". If a question is under-specified, answer with "non-medical".

Prompt A.2: Reddit Question Rewriting

`Prompt:` You will be shown questions about maternal and infant health asked by users. Each question contains a TITLE and DESCRIPTION that elaborates on it, containing details that are both relevant and irrelevant to answering the question. Given a question TITLE and a DESCRIPTION, your task is to incorporate only the relevant details from the DESCRIPTION and rewrite the TITLE into a REWRITE. If there are no relevant details, return the TITLE. As a general rule, keep the rewrite as similar to the original question as possible. The rewrite should be a question in a single sentence.

`Title:` How to Stop Co-Sleeping

`Description:` ...

`Rewrite:` How to wean my 11-month-old out of Co-Sleeping?

B Consolidating Subquestions and Assumptions and Implications into Pragmatic Inferences

Prompt B.1: Question Consolidation

`Prompt:` Given questions asked by new or expecting mothers, your task is to identify the assumptions in them. For this task, you will be given a QUESTION asked by a new or expecting mother, some ASSUMPTIONS (as a list of beliefs or assumptions) in those questions identified by health experts, and some possible SUBQUESTIONS (as a list) that public health experts have identified to have the same information goals as the original question. Given all three of these, your task is to consolidate the SUBQUESTIONS and ASSUMPTIONS into a single, exhaustive list, called INFERENCES. Turning a SUBQUESTION into an inference may involve just turning it into a declarative sentence, or identifying the assumptions made in the SUBQUESTION. Finally, add the INFERENCES to the list of ASSUMPTIONS and remove any duplicates.

C Further Details on Inference Validation

Plausibility scores are the outcome of a three-stage process: (1) a pregnant or new mother holding a belief that is latent while asking a question, (2) a maternal health expert reasoning about these latent beliefs of the mother from the question text, and finally (3) a different expert estimating the likelihood of the beliefs extracted in Step 2 of this process.

The plausibility distribution in Figure 4 represents the results of Step 3. It is important to note that the humans involved in each step of the process are completely disjoint, and have little to no information about each other. In 90% of inferences, the validators agree that the extracted inferences in Step 2 are indeed plausible. The 10% of inferences

Your Task

The question below is asked by a pregnant or postpartum mother on a topic related to pregnancy and postpartum symptoms, developmental milestones, infant care, etc. We will present you with two answers to this question. **Please read the two answers and provide a score for each on a discrete scale from 1 to 5 evaluating the quality of answer completeness.** If both answers are similarly poor, or similarly strong, you can give them similar scores. A high quality answer should identify the information needs of the question and make sure any assumptions are addressed without excluding any major content you feel is necessary in a response to this question. Higher quality answers may provide more comprehensive information or include cautionary caveats.

Figure 5: Human evaluation instructions provided to two expert annotators.

	Mean Likert Score: Default Answer	Inference-Augmented
Expert Annotator 1	4.45	4.6
Expert Annotator 2	4.37	4.45

Table 4: Human preferences of answers on questions with at least one high-plausibility false assumption.

that validators found less plausible is more a reflection of the subjectivity of pragmatic inference in QA in general (see lines 686-690) than our usage of GPT-3.5 a textual transformation tool.

In fact, experts themselves can disagree about answers to questions or background inferences (a finding echoed in Xu et al. (2023)), so we include the plausibility study to reassure audiences that experts do find the inferences in our dataset plausible.

D Human Evaluation Details

Human Evaluation Annotation Instructions. See Figure 5 for annotator instructions.

Annotator preferences on questions with a highly plausible, false assumption. Human annotators prefer inference-augmented answers for questions with a highly plausible, false assumption. See Table 4.

E QA Model Prompts

E.1 Reranker Instructions

Prompt E.1: TART Baseline Instruction

```
Retrieve a passage from medical articles on the web that answers the following question.
```

Prompt E.2: TART Pragmatic Inference Instruction

```
I want to check if the following assumption is true or false. Retrieve an evidence passage for me from medical articles on the web.
```

E.2 Reader Prompts

Prompt E.3: Flan-T5 Reader Prompt

```
You are an expert in maternal and infant health. You are given a few passages below:

Context: {context}

Using information only from the passages above, respond to the following question with a helpful and complete answer. Use information from multiple passages if necessary to answer the question. Again, you are allowed to only use information from the passages above.

Question: {question}
```

Prompt E.4: Mistral/GPT Reader Prompt

```
You are an expert in maternal and infant health. You are given a few verified pieces of information below:

Context: {context}

Using evidence only from verified pieces of information above, respond to the following question with a helpful and complete answer. Use information from multiple sources above if necessary to answer the question. In your answer, do not reveal that you are fetching information from these pieces of evidence. Use information only from the verified sources above, and not from any other sources.

Question: {question}
Answer:
```

Prompt E.5: Inference-Augmented Reader Prompt

```
You are an expert in maternal and infant health. You are given a few verified pieces of information below:

Context: {context}

Using information only from verified pieces of information above, respond to the following question with a helpful and complete answer. As humans often make assumptions while asking questions, your answer must address assumptions made by the asker listed below.

Assumptions: {assumptions}

Use information from multiple sources above if necessary to answer the question below and address the ASSUMPTIONS. In your answer, do not reveal that you are fetching information from these pieces of evidence. Use information only from the verified sources above, and not from any other sources.
Question: {question}
Answer:
```

Question (Q)	Default Answer (A_D)	Pragmatic Inference-Augmented Answer (A_{IA})
When should I give my baby fever reducing meds after shots?	You should give your baby fever reducing medication after shots if they are experiencing discomfort or extreme discomfort due to the fever. However, it is important to note that in most cases, a child does not need to be medicated for a fever unless they are extremely uncomfortable. The most important thing to do when a child has a fever is to make sure they stay hydrated by drinking plenty of fluids. If you are unsure whether your child should receive medication to address the fever or other symptoms, it is recommended to consult with your child's healthcare provider.	**It is not recommended to give your baby fever reducing medication before or immediately after receiving shots. Studies have shown that giving medication prior to a vaccine visit may decrease the child's immune response to the vaccine. Fevers after vaccination are actually a sign that the immune system is responding to the vaccine.** If your baby develops a fever after receiving shots, it is generally not necessary to give them medication unless they are extremely uncomfortable. The most important thing to do when a child has a fever is to make sure they stay hydrated by drinking plenty of fluids. If you are unsure whether your baby should receive medication to address the fever or other symptoms, it is best to consult with your child's healthcare provider.
How warm can my bath be while staying safe for my 4-month-old baby?	To ensure the safety of your 4-month-old baby during bath time, it is recommended to fill the basin with water that feels warm, but not hot, to the inside of your wrist or elbow. The water temperature should be comfortable for your baby and should not pose a risk of burns or scalds. It is important to check the water temperature before placing your baby in the bath to prevent any potential harm.	The safe temperature for a baby may be different from that for an adult. It is important to ensure that the bathwater is warm, but not hot, for a 4-month-old baby. According to the provided information, the water should feel warm to the inside of your wrist or elbow. It is recommended to fill the basin with 2 inches of water at this warm temperature. This is to prevent the baby from getting burned or scalded by hot water. **It is important to note that babies are not as adaptable as adults to temperature change, and they can lose heat rapidly. Therefore, it is crucial to maintain a safe and comfortable water temperature for the baby during bath time.**
Can pregnancy tests be trusted?	According to pregnancy kit manufacturers, most at-home pregnancy tests are 98% to 99% accurate when used exactly as instructed. This means that positive results can generally be trusted. However, it is important to note that false negative results can occur if the test is taken too soon. Therefore, it is recommended to wait until a sufficient amount of time has passed since the possible conception before taking a pregnancy test to ensure more accurate results.	According to the information provided by pregnancy kit manufacturers, most at-home pregnancy tests are 98% to 99% accurate when used exactly as instructed. Positive results from these tests can be trusted. However, it is important to note that false negative results can occur if the test is taken too soon. **Blood tests, on the other hand, can detect pregnancy sooner and are the only tests that can show specific amounts of HCG in the body...**

Table 5: Outputs generated by a standard QA system and a QA systems informed by pragmatic inferences. Answers from the pragmatic system better explicate their underlying inferences.

F Inference Generation Prompt

Prompt F.1: Inference Generation Prompt

```
When humans ask questions, they often have certain assumptions or implications that are embedded in the questions. These assumptions and implications may be true or false, and they may or may not be present in the surface form of the question. Given a question asked by a new or expecting mother, your task is to identify all relevant assumptions and implications in these questions and write them in a list titled INFERENCES. Each inference under INFERENCES should be an independent and declarative assertion that represents an assumption or an implication that the speaker makes while asking the question.
```

Consider the question asked by an expectant mother to their doctor: ***"What kind of music should I play my baby in the womb?"*** This question implicitly makes some assumptions that may or may not be true:

1. Babies can hear sound in the womb.
2. Babies can differentiate music from other sounds.
3. Hearing music positively influences fetal development.
4. (Assuming #3) Certain genres of music are more beneficial than others.

5. There is a baby in my womb.
6. There are different kinds of music.
7. Music is something that can be played.

Unlike the assumptions in the blue box, the assumptions in the red box are linguistically *valid*, but are trivial. Identifying them doesn't help us write more helpful answers that address them. A *complete* answer to this question might address several sub-questions:

1. Does my baby hear sounds in the womb?
2. Can they differentiate music from other sounds?
3. Will listening to music somehow influence their fetal development?
4. Do particular genres of music influence fetal development more than others?

When patients ask questions to their doctors, they may implicitly make problematic, subjective, or false assumptions that are then explicitly corrected by their doctor. This process is often seamlessly carried out in human conversation:

Patient: "*How many glasses of wine a day can I drink while pregnant?*"
Doctor: "*Actually, there's no safe amount of alcohol to drink during pregnancy! Growing babies are exposed to the same amount of alcohol as you are, and the alcohol will pass through your placenta to your baby.*"

Because this question starts with "how many", the doctor inferred that the patient believed that there was an alcohol amount that was safe to consume. Rather than answering "0 glasses are safe", the doctor corrected this assumption, and explained the consequences of drinking any alcohol while pregnant. On the other hand, *chatbots* may take questions at face value, failing to correct potentially harmful assumptions in questions. As a first step, we want to ensure that *chatbots* can identify assumptions or implications inherent in user questions.

Your Task

1 **Write a list of sub questions that an answer to the original question would answer.** Answers to these sub questions should reflect a complete and exhaustive answer to the original question.

2 **Write a set of sentences reflecting possible beliefs or assumptions that the patient may hold** (or, alternatively, beliefs that a complete answer to the question must address). **The assumptions you write can be either medically or scientifically true or false. Please write all assumptions of both types you can identify.** There is no set number of assumptions you can extract: many questions may contain more than one assumption.

Consider the question asked by an expectant mother to their doctor: ***"What should I consume to increase breast milk?"*** This question implicitly makes some assumptions that may or may not be true:

1. Breast milk production is affected by certain foods.
2. Increasing milk production is desirable.
3. Consumption of particular food or beverages is the optimal way to increase milk production.

When patients ask questions to their doctors, they may implicitly make these sorts of problematic, subjective, or false assumptions that are then explicitly corrected by their doctor. In this task you will be shown:

- **a patient question:** this question was asked by a pregnant or postpartum mother and is related to experiences around pregnancy, postpartum, or caring for an infant.
- **a set of possible assumptions:** these sentences reflect possible beliefs that the question asker may hold, or, alternatively, beliefs that any complete answer to the question must address.

Your Task

You will be given a question and a set of possible assumptions. You must first verify whether or not the assumption is medically or scientifically true (or sound). Some assumptions may be true, and others may be false. In order to mark an assumption as true or false, you must find accompanying evidence from a web article from a domain listed below.

For each assumption associated with a given patient question, first, label whether the assumption is medically or scientifically true by finding a web document. Then, find one or more pieces of text from that web document that explicitly provides evidence supporting or refuting the assumption. This span of text can be a phrase, a sentence, or an entire paragraph.

Your evidence to support or refute a particular assumption must come **from a trusted source document. Do not use sources like Wikipedia, personal blogs, or popular consumer health websites like WebMD or Healthline.** Viable and reputable sources include US government websites (left) or US Hospitals or University Clinics (right):

Center for Disease Control (CDC): https://www.cdc.gov/index.htm
US Department of Health and Human Services: https://health.gov/
Office of Women's Health (from Health and Human Services): https://www.womenshealth.gov/
usa.gov
National Library of Medicine: https://www.nlm.nih.gov/
National Institute of Health: https://www.nih.gov/
National Institute of Child Health and Human Development: https://www.nichd.nih.gov/

Mayo Clinic: https://www.mayoclinic.org/
Seattle Children's: https://www.seattlechildrens.org/
Hopkins Children's: https://www.hopkinsmedicine.org/johns-hopkins-childrens-center
Yale Medicine: https://www.yalemedicine.org
Children's Hospital of Philadelphia: https://www.chop.edu/

Given a question asked by a new or expecting mother, write a **helpful** and **informative** answer to the question. You can keep your answers to a couple short paragraphs. To write your answer, please use and cite reputable sources such as:

Center for Disease Control (CDC): https://www.cdc.gov/index.htm
US Department of Health and Human Services: https://health.gov/
Office of Women's Health (from Health and Human Services): https://www.womenshealth.gov/
usa.gov
National Library of Medicine: https://www.nlm.nih.gov/
National Institute of Health: https://www.nih.gov/
National Institute of Child Health and Human Development: https://www.nichd.nih.gov/

Mayo Clinic: https://www.mayoclinic.org/
Seattle Children's: https://www.seattlechildrens.org/
Hopkins Children's: https://www.hopkinsmedicine.org/johns-hopkins-childrens-center
Yale Medicine: https://www.yalemedicine.org
Children's Hospital of Philadelphia: https://www.chop.edu/

A larger list of trusted knowledge sources websites can be found here. Note that you may use sources outside the above-mentioned sources as long as they are similarly **reputable**. **Please abstain from using blogs or for-profit websites such as WebMD, Healthline, or other consumer health websites. Your answer may contain text verbatim from the sources you visit to write a complete answer to the question, but you can also paraphrase or summarize text across multiple sources.**

Please do not use ChatGPT or other generative AI systems to write your answers, extract evidence, or summarize long documents.

Figure 6: Instructions given to each annotator for each phase of annotation. First, we show questions to annotators and ask them to write sub questions and the assumptions present (top panel). Then, after passing these outputs to a prompt-based model to extract consolidated inferences, we ask a different set of annotators to *verify* the veracity of the inferences along with supporting evidence (middle panel). Simultaneously, we ask a third set of annotators to write answers to questions without any inference supervision (bottom panel).