Interpreting Answers to Yes-No Questions in Dialogues from Multiple Domains

Zijie Wang[1] **Farzana Rashid**[2] **Eduardo Blanco**[1]
[1]University of Arizona [2]University of North Carolina Asheville
{zijiewang, eduardoblanco}@arizona.edu frashid@unca.edu

Abstract

People often answer yes-no questions without explicitly saying *yes*, *no*, or similar polar keywords. Figuring out the meaning of indirect answers is challenging, even for large language models. In this paper, we investigate this problem working with dialogues from multiple domains. We present new benchmarks in three diverse domains: movie scripts, tennis interviews, and airline customer service. We present an approach grounded on distant supervision and blended training to quickly adapt to a new dialogue domain. Experimental results show that our approach is never detrimental and yields F1 improvements as high as 11-34%.

1 Introduction

While state-of-the-art models obtain results as high as 93% F1 (Zhang et al., 2021) with question-answering benchmarks such as SQuAD (Rajpurkar et al., 2018), challenges remain. For example, natural questions submitted to search engines remain challenging (Kwiatkowski et al., 2019). Similarly, existing models face challenges with open-ended questions checking for comprehension (Xu et al., 2022), yes-no questions that require deriving an answer (yes or no) from text (Clark et al., 2019), false presuppositions and assumptions in the question (Yu et al., 2023; Kim et al., 2023), and negation (Ravichander et al., 2022).

Many questions in dialogues expect a *yes* or *no* for an answer. Yet many follow-up turns answer this kind of questions without explicitly saying *yes*, *no*, or similar polar keywords. Hockey et al. (1997) analyze 18 hours of speech (Rossen-Knill et al., 1997) and report that 27% of questions fall in this category. Indirect answers to yes-no questions are used to ask follow-up questions or provide explanations for negative answers (Stenström, 1984), prevent incorrect interpretations of direct answers (Hirschberg, 1985), or show politeness (Brown and Levinson, 1978).

A_1: I understand, but I noticed that the Fire Marshall is here with you. Is this somehow related to the fire department?
B_1: I really can't give out any information right now at this point.
A_2: Okay. But I do understand that your partner, Leon Jackson's been injured. Is that correct?
B_2: He was hurt, but not seriously. He'll be fine.
A_3: Do you have the suspect in custody?
B_3: [...] is not a good time, okay. Detective Jackson's hurt. He's fine. I've got a Fire Marshall shot, Detective Jackson is hurt but not seriously.

Figure 1: Movie dialogue with three yes-no questions (A_1, A_2, and A_3). Answers are indirect as they do not include polar keywords (*yes*, *no*, etc.). In B_1 and B_3, the author declines to answer, whereas in B_2 the author indirectly answers *no* by minimizing the incident (*injured* requires loss of function, while *hurt* does not).

Consider the dialogue from the Movie *15 Minutes* in Figure 1. None of the questions are answered explicitly. Speaker B declines to answer the first and third questions. The answer to the first question states that B is not allowed to answer. The answer to the third question restates information known from previous utterances but provides no answer. On the other hand, the answer to the second question implicitly denies that Leon was injured by stating that he was (only) not seriously hurt.

This paper tackles the problem of interpreting indirect answers to yes-no questions (i.e., answers that do not contain *yes*, *no*, or other polar keywords). Our contributions are:[1]

1. Demonstrating that the problem of identifying yes-no questions in dialogues can be automated with high precision, even in out-of-domain dialogues;
2. Creating three new benchmarks (300 instances each) to evaluate models to interpret answers to yes-no questions in three domains;

[1]New benchmarks and code available at https://github.com/wang-zijie/yn-question-multi-domains

Findings of the Association for Computational Linguistics: NAACL 2024, pages 2111–2128

3. A methodology using distant supervision to obtain additional (noisy) training data in a new domain with minimal human intervention;
4. Experimental results showing that blended training with the additional (noisy) data is always beneficial across domains; and
5. Error analysis providing insights into the most difficult indirect answers to interpret correctly.

As our experimental results show, interpreting indirect answers to yes-no questions is a challenging problem. In addition, this problem opens the door to several applications. For example, the work presented here could help dialogue systems avoid conflicts in follow-up turns (Qin et al., 2021) and alleviate the need for clarification questions (Rao and Daumé III, 2018). Further, knowing the interpretations of an indirect answer could help reveal the intention behind questions (Mirzaei et al., 2023).

2 Terminology and Existing Corpora

We use the term *yes-no question* to refer to a question that expects a *yes* or *no* for an answer. Answers to yes-no questions may not include *yes*, *no*, or other polar keywords (e.g., positive: *sure*, *of course*, etc.; negative: *not at all*, *no way*, etc.). We refer to answers with and without polar keywords as *direct* and *indirect* answers.

We make a distinction between the source of dialogues—who the speakers are and why they communicate. We use the term *synthetic* dialogue to refer to dialogues between people who are instructed (and usually paid) to talk about a given topic. The speakers in synthetic dialogues include crowdworkers. We use *genuine dialogue* to refer to naturally-occurring dialogues between people.

Finally, in this paper we work on two problems related to yes-no questions. Given a dialogue, *identifying* yes-no questions pinpoints where the yes-no questions are. On the other hand, *interpreting* answers to yes-no questions figures out the underlying meaning of the answer (*yes*, *no*, or *middle*). Unlike traditional question answering, answers are readily available—the problem is to figure out what the answer means.

Existing Corpora We work with several existing dialogue corpora in multiple domains. In order to identify yes-no questions, we work with the following as training corpora (in-domain): SWDA (Stolcke et al., 2000), telephone conversations with dialogue act annotations (122k turns); MRDA (Shriberg et al., 2004), meeting transcripts with dialogue act annotations (43k turns); DailyDialog (Li et al., 2017), multi-turn conversations written by crowdworkers to simulate human daily conversations (87k turns); Friends (Chen and Choi, 2016), scripts of the TV show (58k turns); and MWOZ (Zhu et al., 2020), task-oriented dialogues written by crowd workers (105k turns). For evaluation purposes (out-of-domain), we use the following: Tennis (Liye et al., 2016), transcripts of post-match interviews of tennis players (164k turns); Movie (Danescu-Niculescu-Mizil and Lee, 2011), movie transcripts (304k turns); and Air (Wei et al., 2018), task-oriented dialogues with topics limited to travel and flights (3,805k turns).

In order to interpret answers to yes-no questions, we work with the following as training corpora: Circa (Louis et al., 2020), 34k yes-no questions and indirect answers written by crowdworkers; and SWDA-IA (Sanagavarapu et al., 2022), 2.5k yes-no questions and indirect answers from the SWDA. Both corpora include manual annotations of the interpretations of answers. For evaluation purposes, we use the same corpora than for identifying yes-no questions: Tennis, Movie, and Air. Specifically, we create new benchmarks (300 questions and indirect answers from each corpus) and use the rest (questions and direct answers) for training purposes via distant supervision.

3 Related Works

Yes-no questions have received considerable attention recently. BoolQ (Clark et al., 2019) is a collection of 16,000 yes-no questions and Wikipedia articles from which answers (*Yes* or *No*) can be derived. Sulem et al. (2022) enhance BoolQ with questions that cannot be answered. Unlike them, in this paper we target yes-no questions in dialogues, which are more open-ended (Figure 1) than the fact-seeking questions (e.g., Has the UK been hit by a hurricane?). Two recent works target yes-no questions in dialogues (Choi et al., 2018; Reddy et al., 2019). Unlike us, both of them work with synthetic dialogues written by crowdworkers and are constrained to a handful of scenarios.

Yes-no questions in genuine dialogues have been studied before. de Marneffe et al. (2010) study 224 yes-no questions including gradable adjectives, and de Marneffe et al. (2009) present a typology for 623 yes-no questions from SWDA. We work with an order of magnitude more data, several dialogue domains, and modern learning strategies.

	SWDA	MRDA	DailyDialog	Friends	MWOZ	All
genuine dialogue?	Yes	Yes	No	Yes	No	n/a
# turns	122k	43k	87k	58k	105k	415k
# yes-no questions						
using strict rules	1.8k	1.3k	8.0k	2.4k	16.2k	29.8k
# with indirect answers	0.0k	0.0k	0.0k	0.0k	0.0k	0.0k
precision (in 200 samples)	1.00	1.00	1.00	1.00	1.00	1.00
using relaxed rules	3.7k	2.8k	13.5k	4.9k	34.5k	59.4k
# with indirect answers	1.9k	1.5k	5.5k	2.5k	18.3k	29.6k
precision (in 200 samples)	0.99	0.99	1.00	0.99	1.00	0.99

Table 1: Evaluation of rules to collect yes-no questions. Precision is calculated with a random sample of size 200 for each corpus. The relaxed rules yield twice as many yes-no questions (i.e., twice the relative recall) without lowering precision. Note that many answers to yes-no questions are *indirect*.

The work presented here is closest to Circa (Louis et al., 2020), DIRECT (Takayama et al., 2021), and SWDA-IA (Sanagavarapu et al., 2022). Unlike us, Circa works with synthetic yes-no questions and answers without any conversational context. DIRECT also works with synthetic dialogues. SWDA-IA works with telephone conversations from SWDA. To our knowledge, we are the first to explore yes-no questions in multiple dialogue domains. We show that existing corpora are beneficial, and more importantly, that combining additional training data obtained via distant supervision in the new dialogue domains brings additional improvements across all domains.

4 Identifying Yes-No Questions

We first tackle the problem of identifying yes-no questions in dialogue. To our knowledge, previous work on yes-no questions is limited to interpreting the answers. We first present our rule-based approach to collect yes-no questions. Then, we describe how to leverage these rules to build a classifier to automate the task.

4.1 Collecting Yes-No Questions

We define rules to identify yes-no questions in dialogues based on (a) dialogue acts if gold annotations are available or (b) lexical matching. Our rules look at (a) all turns within a dialogue for turns that may contain a yes-no question and (b) the next turn to check for direct answers. We refer to the set of rules that only look at the question as *relaxed rules*, and to the combination of rules that look at the question and answer as *strict rules*.

Corpora with Dialogue Acts Annotations For SWDA and MRDA, the only two corpora with dialogue act annotations, we use these annotations as they indicate yes-no question presence. Specifically, we refine the list of dialogue acts by Sanagavarapu et al. (2022), as SWDA and MRDA use different label sets (see Appendix A).

Corpora without Dialogue Acts Annotations For the other corpora we work with (DailyDialog, Friends, and MWOZ), we define simple rules that identify yes-no questions with high precision:

The conversation turn:

1. includes common auxiliary verbs in yes-no questions (*do*, *does*, *did*, *don't*, *doesn't*, *didn't* *is*, *isn't*, *are*, *aren't*, *was*, *wasn't*, *were*, *weren't*, *have*, *haven't*, *has*, *hasn't*, *can*, *can't*, *could*, *couldn't*, *will*, *won't*, *would*, *wouldn't*, *may*, and *might*) and does not include wh-question words (*what*, *when*, *where*, *which*, *who*, *whom*, *whose*, *why*, and *how*); and
2. has more than three tokens and ends in '?'.

Regardless of how questions are identified, we experiment with an extra rule to check if the next turn is a direct answer. Here we are not concerned with the interpretation of the answer. Rather, we consider the subset of yes-no questions that are followed by a direct answer regardless of its interpretation. We identify direct answers by checking whether the first two sentences in the next turn contain *yes*, *yea*, *yup*, *yep*, *yeah*, *sure*, *no*, or *nope*.

Table 1 analyzes the outcome of the rules. We estimate precision using a sample of 200 matches per dialogue domain (total: 1,000), and use the number of matches (in our case, the number of yes-no questions) to approximate relative recall (Pantel and Pennacchiotti, 2006). Overall, The relaxed rules yield twice as many yes-no questions than the strict rules (twice relative recall: 29.8k vs. 59.4k matches) while being equally precise.

	In-Domain		Out-of-Domain							
			Tennis		Movie		Air		All	
	# k	P	# k	P	# k	P	# k	P	# k	P
Rule-based classifier										
strict rules	29.8	1.00	23	1.00	9	1.00	364	1.00	396	1.00
relaxed rules	59.4	0.99	34	1.00	18	0.99	808	1.00	860	1.00
BERT, distant supervision with										
strict rules	n/a	n/a	40	0.99	25	0.98	826	1.00	891	0.99
relaxed rules	n/a	n/a	42	0.99	24	0.98	825	1.00	891	0.99

Table 2: Evaluation of the rule-based and BERT classifiers to identify yes-no questions. *In-domain* refers to the corpora used to define the rules (Table 1) and train the BERT classifier using distant supervision. '# k' stands for *number of yes-no questions identified in thousands*. The classifiers with strict and relaxed rules (top block) are equally precise, but the latter doubles recall (twice # k). The BERT classifiers are equally precise.

4.2 Classifiers to Identify Yes-No Questions

The rules to identify yes-no questions were iteratively defined using the five corpora in Table 1: SWDA, MRDA, DailyDialog, Friends and MWOZ. While the precision is high in these corpora (in-domain), our goal is to identify yes-no questions in any dialogue (out-of-domain). To do so, we evaluate with out-of-domain corpora (Movie, Tennis, and Air) with (a) a rule-based classifier and (b) a classifier trained with the output of the rules using distant supervision.

Rule-based Classifier Our first classifier to identify yes-no questions is simple: run the rules previously defined to identify yes-no questions. Doing so has the advantage of simplicity. However, like any other rule-based system, doing so may suffer from low recall as the rules may not generalize.

BERT Classifier Our second classifier uses distant supervision. We use a BERT classifier (Devlin et al., 2019) trained as follows. We use as positive examples the 59.4k yes-no questions identified with our rules in the training corpora (Table 1). As negative examples, we randomly choose 59.4k turns not identified as yes-no questions with the rules. We use the implementation by Hugging Face (Wolf et al., 2020). While any other models could be used, we chose BERT because it demands less computational resources and obtains hard-to-beat results. Appendix A provides additional details.

4.3 Results and Analysis

Table 2 presents results with the rule-based and BERT-based classifiers. In-domain refers to the corpora used to define the rules (Table 1, same as *All*). Out-of-domain includes three additional dialogue corpora that we will also use to interpret answers to yes-no questions. We approximate precision using a sample of 200 examples per corpora.

The rule-based classifier obtains almost perfect precision with both in-domain corpora and the three out-of-domain corpora, regardless of whether we use strict or relaxed rules. Using relaxed rules, however, obtains twice the amount of yes-no questions in both in-domain and out-of-domain corpora.

Despite it is trained with yes-no questions matching a handful of rules, the BERT-based classifier identifies many more yes-no questions than the rules themselves. While the overall benefit looks somewhat low (860k vs. 891k; 3.6%), this is mostly due to the small improvement with Air (825k vs. 808k; 2.2%). Indeed, in Tennis and Movie the BERT-based classifier identifies 23.5% and 33.3% more yes-no questions (42k vs. 34k and 24k vs. 18k). Note that unlike Tennis and Movie, Air consists exclusively of synthetic dialogues to make travel reservations. These dialogues are very restrictive; most of the yes-no questions are asked by the speaker acting as the travel agent. Further, dialogues are scripted and yes-no questions follow very few patterns that can be easily caught with our rules (e.g., *Do you mean [. . .]?*, *Would you like a late flight?*). Surprisingly, there is no difference in training with the output of strict or relaxed rules.

5 Interpreting Answers to Yes-No Questions

Armed with the highly precise classifier to identify yes-no questions, we move to interpreting answers to yes-no questions in dialogues from multiple domains. To our knowledge, there are two publicly available corpora: Circa and SWDA-IA (Section 2). We aim to explore multiple dialogue domains, so we first create new benchmarks.

	Existing Benchmarks		Our Benchmarks		
	Circa	SWDA-IA	Tennis	Movie	Air
genuine?	No	Yes	Yes	Yes	No
context?	No	Yes	Yes	Yes	Yes
# yes-no questions (train+dev / test)	27.4k / 6.8k	2.0k / 0.5k	0 / 300	0 / 300	0 / 300
% answers with interpretation *Yes*	57.1	61.9	47.0	26.7	90.0
% answers with interpretation *No*	40.1	23.2	18.3	18.3	3.0
% answers with interpretation *Middle*	2.8	14.9	34.7	55.0	7.0

Table 3: Analysis of corpora to interpret indirect answers to yes-no questions. Note that the label distribution (% of *Yes*, *No*, and *Middle*) is very different in each benchmark.

Three New Benchmarks We create three new benchmarks for evaluation purposes in new domains. Specifically, we randomly select 300 yes-no questions followed by an indirect answer from each corpus (Tennis, Movie, and Air; 900 total). Then, we manually annotate the interpretation of the answer using three labels: *Yes*, *No*, or *Middle*. Our definition of *Yes* includes what Circa and SWDA-IA define as *Probably yes*, which include *sometimes yes* and *yes under certain conditions*. For example, we annotate *Q: Do you like Mexican food? A: I am fine with tacos if my friends suggest Mexican* with *Yes*. Similarly, our definition of *No* includes what others define as *Probably no*. For example, we annotate *Q: Do you want to go out for dinner? A: I have a deadline and I may skip dinner* with *No*. On the other hand, our definition of *Middle* follows previous work: we choose it when the answer does not lean toward *yes* or *no*.

Table 3 summarizes all the benchmarks available. Training data is only available for the two existing benchmarks. Note that the label frequency is very different between existing corpora and our new benchmarks. We argue that not artificially balancing the benchmarks is sound. Indeed, domain adaptation is not only about working with language from other domains, but also accounting for label distribution shifts. Tennis and Movie have many more answers to yes-no questions whose interpretation is *Middle* compared to Air (34.7% and 55.0% vs. 7.0%) and existing benchmarks (2.8% and 14.9%). We also observe that our benchmarks have fewer answers whose interpretation is *No*. Most answers in Air are interpreted as *Yes*; as discussed before most questions in Air come from a travel agent confirming travel arrangements rather than open-ended conversations. The substantial differences in the distribution of interpretations across existing and our benchmarks indicate that transfer learning across these domains might be challenging. As we shall see, however, we benefit from using distant supervision with all the dialogue corpora.

Inter-Annotator Agreements The three benchmarks were annotated in-house by two graduate students. Inter-annotator agreements (Cohen's κ) for Tennis, Movie, and Air are 0.68, 0.69, and 0.66 respectively. These coefficients indicate *substantial* agreement (Artstein and Poesio, 2008); above 0.8 would be (nearly) perfect. 87.0% of disagreements are between (a) *Yes* or *No* and (b) *Middle*, while only 13.0% are between *Yes* and *No*. These percentages suggest that most disagreements are minor. After annotating individually, the annotators discussed the disagreements in order to adjudicate them and create the final ground truth. We refer the reader to Appendix B for more details.

5.1 Model Training Strategies

We follow three strategies to build models to interpret answers to yes-no questions. The differences are which corpora we train with and the training procedure to combine the training corpora. All strategies start with the off-the-shelf RoBERTa transformer (Liu et al., 2019) released by Hugging Face (Wolf et al., 2020). This problem is harder than identifying yes-no questions, and we found it beneficial to use RoBERTa instead of BERT. We also experiment with BART (Lewis et al., 2020), however, RoBERTa outperforms BART on most benchmarks—the only exception is Air. Appendix C details the results with BART. All hyperparameters were tuned with the train and validation splits; we refer the reader to Appendix C for details.

The first strategy is to fine-tune a RoBERTa classifier with the existing benchmarks (Circa and SWDA-IA)—the only ground truth available for training purposes to interpret yes-no questions. The other two strategies also fine-tune a RoBERTa classifier, but combine training data from (a) existing benchmarks and (b) additional instances from the

same corpora we created our benchmarks with. These additional instances were obtained using distant supervision as detailed below. Crucially, obtaining them does not require human involvement after generic patterns applicable to any dialogue corpora are defined. The second and third strategies differ in the fine-tuning methodology. The former merges the training data and proceeds to fine-tune with the combination. The latter uses blended training to phase out the training data from existing corpora as detailed below.

Blended Training We adopt the method by Shnarch et al. (2018) to blend training data from existing corpora (Circa and SWDA-IA) and the additional annotations obtained with distant supervision. The blending process consists of two phases: *m* blending epochs using all the additional annotations and a fraction of the training instances from existing corpora, and then *n* epochs only using all the additional annotations. The intuition is that existing corpora provide a good base to interpret answers to yes-no questions, but that it is beneficial to use instances closer to the domain we evaluate with as training progresses. In each blending epoch, the fraction of instances from existing corpora are fed randomly to the network. The blending factor $\alpha \in [0, 1]$ determines the fraction of instances from existing corpora to consider. The first blending epoch trains with all of them, and the ratio to phase out in each epoch is determined by α. The blending hyperparameters (α, m, and n) are tuned like any other hyperparameter (see Appendix C).

Distant Supervision The goal of distant supervision is to explore whether considering additional instances automatically labeled in the new dialogue domains is beneficial. Given the high precision of the patterns to identify yes-no questions (Section 4), using the strict rules and matching *yes* and *no* keywords to their corresponding answers is worth exploring. The aim of these patterns ought to be as precise as possible. Disregarding many yes-no questions and answers (i.e., low recall) is acceptable as the large amount of unannotated dialogue corpora still allows us to automatically label many instances and use them to train models. We consider the same patterns from Section 4. The keywords to label an answer as *Yes* (*yes*, *yea*, *yup*, *yep*, *yeah*, *sure*) or *No* (*no*, or *nope*) are limited. However, we found that adding other keywords leads to unnecessary noise. For example, *sure* appears at first sight to be a good keyword for *Yes*, although it often is not (e.g., *Q: Do you like Mexican food? A: Sure, if I run out of everything else I will eat it.*).

Distant supervision in the three new dialogue domains yields 380k instances for training purposes (Tennis: 19,055, Movie: 6,250, and Air: 355,549). We balanced the datasets before model training.

5.2 Experimental Results

We present the results in Table 4. Air is unbalanced (Table 3, Yes: 90.0%), and all models obtain similar results than the majority baseline.

Let us first discuss the results training only with existing corpora (second block; Circa, SWDA-IA, or both). Synthetic yes-no questions and answers are much easier to interpret (Circa: 0.93, Air: 0.84, both F1-score) than those coming from genuine dialogues (F1: 0.37–0.68), although out-of-domain evaluation shows that training with existing corpora outperforms the majority baselines with Tennis (F1: 0.34 vs. 0.52) and Movie (F1: 0.36 vs. 0.37).

Second, training strategies combining existing training data and the additional instances obtained via distant supervision are beneficial. In particular, it is beneficial to consider *all* instances (Tennis, Movie, and Air) regardless of which domain we evaluate with. Finally, we observe that blending (fourth block; 0.49–0.86) yields better results than merging the additional annotations (third block; 0.42–0.85). Most importantly, for two benchmarks (Movie and Air), the improvements are statistically significant (McNemar's test (McNemar, 1947), $p < 0.05$) when compared to training with the existing data. Thus the proposed distant supervision is successful at adapting a model to interpret yes-no questions to new domains. This is true across all labels despite distant supervision only identifies additional training instances with *Yes* and *No* interpretations. Appendix C provides additional results (F1 score) per label and more metrics (Precision, Recall, and F1 score) that complement Table 4.

5.3 Error Analysis

We also conduct an error analysis to identify the most common error types made by the best-performing model (i.e., bottom row in Table 4). We analyze 100 errors from the three dialogue domains our best model makes the most errors with SWDA-IA (F1: 0.68), Tennis (F1: 0.59), and Movie (F1: 0.49). Note that we make few errors with Circa and Air (F1: 0.93 and 0.86).

	Existing Benchmarks		Our Benckmarks		
	Circa	SWDA-IA	Tennis	Movie	Air
Majority Baseline	0.43	0.32	0.34	0.36	0.84
RoBERTa, training with					
Circa	0.93	0.51	0.40	0.34	0.84
SwDA-IA	0.69	0.63	0.42	0.37	0.63
Circa + SwDA-IA	0.93	0.68	0.52	0.37	0.84
RoBERTa, training with					
in-domain instances and					
Circa	n/a	n/a	0.43	0.24	0.85
SwDA-IA	n/a	n/a	0.41	0.31	0.84
Circa + SwDA-IA	n/a	n/a	0.48	0.34	0.85
all additional instances and					
Circa	0.93	0.56	0.44	0.24	0.82
SwDA-IA	0.74	0.67	0.51	0.28	0.82
Circa + SwDA-IA	0.93	0.70	0.53	0.42	0.84
RoBERTa, blended training with					
in-domain instances and					
Circa	n/a	n/a	0.37	0.31	0.83
SwDA-IA	n/a	n/a	0.48	0.36	0.84
Circa + SwDA-IA	n/a	n/a	0.57	0.39	0.84
all additional instances and					
Circa	0.93	0.54	0.42	0.44	0.82
SwDA-IA	0.71	0.64	0.50	0.40	0.85
Circa + SwDA-IA	0.93	0.68	0.59	0.49*	0.86*

Table 4: Results (F1) interpreting answers to yes-no questions. *In-domain* refers to additional training instances from the *same* domain we evaluate with. *Air* is heavily unbalanced (Table 3) and limited to airline bookings; no model substantially outperforms the majority baseline. Training with the proposed distant supervision approach is (a) never detrimental if training data in the same domain is available (Existing Benchmarks) and (b) always beneficial otherwise (Our Benchmarks). The improvements on Movie and Air are statistically significant (McNemar's test (McNemar, 1947), $p < 0.05$; indicated with an asterisk).

We identify four frequent error types across the three dialogue domains (first block in Table 5). First, unresponsive answers should almost always be interpreted as *Middle* as they do not address the question, yet the model routinely (18% of errors) mispredicts *Yes* or *No*. Similar mispredictions occur for a specific kind of unresponsive answer: answering with a question (13%). Intricate, long answers account for 7% of errors. In the example, the answer has 335 tokens; it gives background and explanations but it never addresses the questions (Gold: *middle*). Note that in Tennis interviews, most of the conversation turns are rather long. Lastly, we found that 5% of errors in all three dialogue domains occur when the question has a negation—regardless of the answer.

We also identify six error types in at least two of the dialogue domains. In Tennis and SWDA-IA, polar distractors (i.e., *yes* or *no* indicators in an answer whose interpretation is *No* or *Yes*) account for 18% of errors. In the examples, the model is misguided by *can*, which indicates *Yes* despite the answer ought to be interpreted as *No*. Extremely short answers are somewhat common in Movie and SWDA-IA and account for 13% of errors. We define confrontational and uninformative answers as answers that are hostile towards the author of the questions and avoid providing an answer while not changing the topic of conversation (unlike unresponsive answers, which are discussed above). Uninformative answers are always to be interpreted as *middle*. Finally, we also identify that interpreting answers sometimes requires external knowledge such as world and commonsense knowledge (5% of errors in Movie and SWDA-IA). In the example, being from *New York* means someone is not from *L.A.*; however, being from *Hollywood* would mean the opposite. Finally, we identify conditions and contrasts—even in short answers—are present in 4% of errors in Tennis and SWDA-IA. In the example, the answer states that *he was able to do everything himself until the operation*, implying that he is not able anymore. Thus, the ground truth is *No*. The model is unable to see the contrast between the past and current situation.

Error Type	%	TMS	Example	G, P
Unresponsive answer	18	✓✓✓	Q: Do you think it is a little early? A: I brought you something . . . From the library.	Middle, Yes
Answer has question	13	✓✓✓	Q: Really? Do you have the money with you? A: Do you have the material?	Middle, Yes
Intricate Answer	7	✓✓✓	Q: [. . .] do you think the English players have it easier? A: [335 tokens] There was a lot of talk about the lack [. . .]	Middle, Yes
Question has negation	5	✓✓✓	Q: Don't you like cats? A: Well, I like cats. This, this cat is a, uh, more like a dog.	Yes, No
Polar distractor in answer	18	✓ ✓	Q: You may be familiar with [. . .]. Have you ever, A: [. . .], you can tell me a little bit more about it [. . .]	No, Yes
Short question or answer	13	✓✓	Q: Were you A: Really is.	Middle, Yes
Confrontational answer	6	✓✓	Q: Did my father tell you not to talk about it? A: Come on. you brought it up.	Middle, Yes
Uninformative answer	5	✓ ✓	Q: Is it twenty percent? A: I, I have no idea, I just. My dad does it all for me.	Middle, No
External knowledge	5	✓✓	Q: You're from L.A., huh? A: New York.	No, Middle
Condition or contrast	4	✓ ✓	Q: Is he able to, uh, still do everything himself pretty well? A: Well, he was until this operation. He has arthritis.	No, Yes

Table 5: Most common error types in *T*ennis, *M*ovie, and *S*WDA-IA with our best model. Percentages are the average in the corpora where the error was observed, indicated with checkmarks. The last column indicates the *G*old and (wrong) *P*redictions.

	Tennis	Movie	Air
RoBERTa, training with additional instances and			
Circa	0.41	0.24	0.84
SWDA-IA	0.34	0.27	0.83
Circa + SWDA-IA	0.55	0.42	0.85
RoBERTa, blended with additional instances and			
Circa	0.48	0.42	0.84
SWDA-IA	0.41	0.29	0.85
Circa + SWDA-IA	0.60	0.48	0.86

Table 6: Results (F1) obtained with RoBERTa trained with a set of additional instances that is the same size as the in-domain instances. The results remain similar compared to the model trained with *all additional instances* (Table 4), demonstrating that the performance gains are mostly due to training instances from the new domains rather than just more training instances.

5.4 Ablation Study: More Instances or Cross-Domain Instances?

To further understand the source of the performance gains—whether they are due to more training instances or having cross-domain instances—we conduct an ablation study. This involves training a RoBERTa model with additional instances that are equivalent in size to the in-domain instances for each benchmark. Table 6 presents the results.

For the second training strategy (Section 5.1), the results are never detrimental compared to training with all additional instances (Table 4, Tennis: 0.53 vs. 0.55, Movie: unchanged 0.42, and Air: 0.84 vs. 0.85), demonstrating that the performance gains are mostly from the cross-domain instances. A similar trend is observed for the third training strategy (blending; Tennis: 0.59 vs. 0.60, Movie: 0.49 vs. 0.48, and Air: unchanged 0.86).

5.5 A Note on Large language Models

Recent works have shown that prompting with large language models achieves better results in many tasks (Mishra et al., 2022) compared to supervised approaches using substantially smaller models. This is not the case with the problem of interpreting indirect answers to yes-no questions.

We explore whether large language models outperform our RoBERTa-based classifier at interpreting indirect answers to yes-no questions. Specifically, we experiment with three LLM models: GPT-3.5 (Brown et al., 2020), Alpaca-7B (Taori et al., 2023), and Llama 2-7B (Touvron et al., 2023). We manually map the models' output to our interpretations of indirect answers (*Yes*, *No*, or *Middle*) for evaluation purposes.

We evaluate with GPT-3.5 using Microsoft Azure API. For Alpaca and Llama, we host them locally. However, we are only able to evaluate them in up to 4-shot prompting because of resource limitations. Table 7 shows the results. In general, 4-shot prompting yields improvements compared to 0-shot prompting. Surprisingly, GPT-3.5 obtains

	Tennis	Movie	Air
	0-shot		
GPT-3.5	0.39	0.29	0.23
Alpaca-7B	0.35	0.22	0.28
Llama 2-7B	0.32	0.19	0.32
	4-shot		
GPT-3.5	0.50	0.31	0.59
Alpaca-7B	0.43	0.40	0.77
Llama 2-7B	0.33	0.21	0.73
Our best model	0.59	0.49	0.86

Table 7: Results (F1) obtained with GPT-3.5, Alpaca-7B, and Llama 2-7B. We evaluate them with the test split (240 instances per benchmark) in 0-shot and 4-shot prompts. Despite their much larger model size, none of them outperforms our best model.

worse results than the other two models with Movie and Air despite it being a much larger model. We hypothesize the higher results with Tennis are due to GPT-3.5's better performance with longer texts. Tennis has on average much longer answers than the other two benchmarks. Most importantly, all three models fail to match our best results.

To further investigate the reason behind the poor performance, we conduct an error analysis with the results obtained with GPT-3.5 using 4-shot prompts. We calculate the error distributions by gold label and (wrong) predictions. In addition, we list a few examples. The results can be found in Appendix C, which also contains details about the prompts and experimental setup.

6 Conclusions

Indirect answers to yes-no questions in dialogue are common. In this paper, we have presented an approach to identify yes-no questions in dialogues (distant supervision and BERT), and more importantly, to interpret indirect answers to yes-no questions. Experimental results show for the first time that the identification problem is rather simple. The second problem, on the other hand, remains challenging—F1 scores are 0.49 and 0.59 with Movies and Tennis. These results lead to the conclusion that synthetic dialogues may not be representative of more open-ended conversations.

Crucially, we have shown that distant supervision to obtain additional examples with *direct Yes* and *No* answers is beneficial to interpret *indirect* answers. Indeed, combining the additional instances with blended training is never detrimental and yields substantial improvements with our new, out-of-domain benchmarks (Tennis, Movie, and Air). In other words, the proposed methodology can be used to adapt to new domains without requiring substantial human involvement, unlike annotating additional examples.

Our future plans include addressing the most common errors. In particular, we believe that exploring dialogue coherence and pretrained language models customized to the dialogue domain are two lines of research worth exploring.

Limitations

We adopt distant supervision to obtain additional data for training purposes. However, we only design rules to identify yes-no questions with direct answers, which means the extra (and noisy) training instances only have interpretations *yes* and *no*—there are no additional instances with *middle* interpretations. We notice that in some cases, the improvements in results are mostly brought by better results predicting indirect answers that are labeled *yes* and *no*, and only to a smaller degree by those labeled *middle*. Detecting hesitation or non-answers (or simply answers who indicate 50/50) could be critical in some domains and our distant supervision approach does not provide much benefit with *middle*.

We annotate our new benchmarks with 3 labels (*yes*, *no* and *middle*). Some previous works use finer-grained label sets to interpret answers to yes-no questions. For example, Sanagavarapu et al. (2022) use five labels including *Probably yes* and *Probably no*, and Louis et al. (2020) use nine labels including *Sometime yes*, *In the middle*, or *I am not sure how to interpret [the answer to the question]* (only 3.6% of the answers receive these three labels, however). Considering that (a) there is no universal agreement about the possible ways to interpret answers to yes-no questions, (b) other works also use three labels (Sulem et al., 2022), and (c) it is unclear which interpretations may be more useful in a real-world application, we argue that three labels are as sound as five or nine—or at least not worse.

We tune several hyperparameters (including the blending factor α and the amount of additional instances to train; third and fourth block in Table 4) with the train and development splits, and report results with the test set. The results are taken from the output of one run. We acknowledge that the average of multiple runs (e.g., 10) would be more

reliable, but they also require much more computational resources (literally, 10 more times).

Ethics Statement

Data biases. Our work only focuses on yes-no questions and answers in English dialogues. Researchers (Lafford, 2004) have shown that other languages (e.g., Spanish) tend to communicate more directly, thus the vagueness of indirect answers may not be as big of a problem. Our future plans include targeting the same problem in multiple languages.

Data sources and collection. We collect the Movie, Tennis, Friends and SWDA datasets from Convokit (Chang et al., 2020); and samples from MWOZ, under MIT license. Samples from MRDA are used under GPL-3.0 license. Samples from DailyDialog are used under CC BY-NC-SA 4.0. Samples from AirDialogue are used under Apache License 2.0.

Acknowledgements

We thank the reviewers for their insightful comments. We also thank the Chameleon platform (Keahey et al., 2020) for providing computational resources. The Accelerating Foundation Models Research program by Microsoft provided Azure credits to conduct this research.

This material is based upon work supported by the National Science Foundation under Grant No. 2310334. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Penelope Brown and Stephen C Levinson. 1978. Universals in language usage: Politeness phenomena. In *Questions and politeness: Strategies in social interaction*, pages 56–311. Cambridge University Press.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Yu-Hsin Chen and Jinho D. Choi. 2016. Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100, Los Angeles. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87, Portland, Oregon, USA. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Scott Grimm, and Christopher Potts. 2009. Not a simple yes or no: Uncertainty in indirect answers. In *Proceedings of the SIGDIAL 2009 Conference*, pages 136–143, London, UK. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2010. "was it good? it was provocative." learning the meaning of scalar adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 167–176, Uppsala, Sweden. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Julia Bell Hirschberg. 1985. *A theory of scalar implicature (natural languages, pragmatics, inference)*. Ph.D. thesis, University of Pennsylvania.

Beth Ann Hockey, Deborah Rossen-Knill, Beverly Spejewski, Matthew Stone, and Stephen Isard. 1997. Can you predict responses to yes/no questions? yes, no, and stuff. In *Fifth european conference on speech communication and technology*. Citeseer.

Kate Keahey, Jason Anderson, Zhuo Zhen, Pierre Riteau, Paul Ruth, Dan Stanzione, Mert Cevik, Jacob Colleran, Haryadi S Gunawi, Cody Hammock, et al. 2020. Lessons learned from the chameleon testbed. In *2020 USENIX annual technical conference (USENIX ATC 20)*, pages 219–233.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. (QA)^2: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8466–8487, Toronto, Canada. Association for Computational Linguistics.

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.

Barbara A Lafford. 2004. The effect of the context of learning on the use of communication strategies by learners of spanish as a second language. *Studies in second language acquisition*, 26(2):201–225.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Fu Liye, C Danescu, and Lillian Lee. 2016. Tie-breaker: Using language models to quantify gender bias in sports journalism. *Computation and Language*.

Annie Louis, Dan Roth, and Filip Radlinski. 2020. "I'd rather just go to bed": Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, Online. Association for Computational Linguistics.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Maryam Sadat Mirzaei, Kourosh Meshgi, and Satoshi Sekine. 2023. What is the real intention behind this question? dataset collection and intention classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13606–13622, Toronto, Canada. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to GPTk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612, Dublin, Ireland. Association for Computational Linguistics.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 113–120, Sydney, Australia. Association for Computational Linguistics.

Libo Qin, Tianbao Xie, Shijue Huang, Qiguang Chen, Xiao Xu, and Wanxiang Che. 2021. Don't be contradicted with anything! CI-ToD: Towards benchmarking consistency for task-oriented dialogue system. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2367, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Sudha Rao and Hal Daumé III. 2018. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2737–2746, Melbourne, Australia. Association for Computational Linguistics.

Abhilasha Ravichander, Matt Gardner, and Ana Marasovic. 2022. CONDAQA: A contrastive reading comprehension dataset for reasoning about negation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8729–8755, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Deborah Rossen-Knill, Beverly Spejewski, Beth Ann Hockey, Stephen Isard, and Matthew Stone. 1997. Yes/no questions and answers in the map task corpus.

Krishna Sanagavarapu, Jathin Singaraju, Anusha Kakileti, Anirudh Kaza, Aaron Mathews, Helen Li, Nathan Brito, and Eduardo Blanco. 2022. Disentangling indirect answers to yes-no questions in real conversations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4677–4695, Seattle, United States. Association for Computational Linguistics.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The ICSI meeting recorder dialog act (MRDA) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Anna-Brita Stenström. 1984. *Questions and responses in English conversation*. Krieger Pub Co.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Elior Sulem, Jamaal Hay, and Dan Roth. 2022. Yes, no or IDK: The challenge of unanswerable yes/no questions. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1075–1085, Seattle, United States. Association for Computational Linguistics.

Junya Takayama, Tomoyuki Kajiwara, and Yuki Arase. 2021. DIRECT: Direct and indirect responses in conversational text corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1980–1989, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wei Wei, Quoc Le, Andrew Dai, and Jia Li. 2018. AirDialogue: An environment for goal-oriented dialogue research. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3844–3854, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Xinyan Yu, Sewon Min, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. CREPE: Open-domain question answering with false presuppositions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10457–10480, Toronto, Canada. Association for Computational Linguistics.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. 2021. Retrospective reader for machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14506–14514.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.

A Additional Details to Identify Yes-No Questions

Dialogue Act Labels to Select Yes-No Questions The process to select yes-no questions uses dialogue act labels if available (Section 4).

The list of dialogue acts and their descriptions for SWDA is as follows:

- qh: Rhetorical question
- qy: Yes-no question
- qy^d: Declarative yes-no question
- ^g: Tag-Question
- qy^t: Yes-no question about task
- qy^r: Yes-no question repeat self
- qy^m: Yes-no question mimic other
- qy^h: Question in response to a question
- qy^c: Yes-no question about communication
- qy^2: Yes-no question collaborative completion
- qy(^q): Yes-no question quoted material
- qy^g: Yes-no question tag-question
- qy^g^t: Yes-no question tag-question about task
- qy^g^r: Yes-no question tag-question repeat self
- qy^g^c: Yes-no question tag-question about communication
- qy^d^t: Declarative yes-no question about task
- qy^d^r: Declarative yes-no question repeat self
- qy^d^m: Declarative yes-no question mimic other
- qy^d^h: Declarative yes-no question in response to a question
- qy^d^c: Declarative yes-no question about communication
- qy^d(^q): Declarative yes-no question quoted material
- qy^c^r: Yes-no question about-communication repeat self

The list of dialogue acts and their descriptions for MRDA is as follows (this corpus includes fewer dialogue act labels):

- qy: Yes-no question
- g: Tag-question

Details and Hyperparameters for BERT-based Classifier Referring to Section 4.2, we use an off-the-shelf BERT-base model (110M parameters) from Hugging Face (Wolf et al., 2020) to train a classifier that identifies yes-no questions. We run the experiments on a single NVIDIA Tesla V100 (32GB) GPU. It takes approximately 5 minutes to train 1 epoch. Table 8 shows the hyperparameters that yield the highest accuracy in identifying yes-no questions.

Hyperparameters	
Maximum Epochs	5
Batch Size	32
Optimizer	AdamW
Learning rate	5e-5

Table 8: Tuned hyperparameters for experiments to identify yes-no questions with BERT-base model.

B Additional Details about Benchmark Annotations

Annotation Instructions We conduct manual annotations to obtain ground truth interpretations (i.e., gold labels) for indirect answers to yes-no questions. We adopt three labels: *Yes*, *No*, and *Middle (Unknown)*, and we define them as follows for annotations:

- *Yes*: The answer shows (or implies) *yes*, *yes* under certain conditions or constraints (probably yes).
- *No*: The answer shows (or implies) *no*, *no* under certain conditions or constraints (probably no), conveys negative sentiment, or provides arguments for *no*.
- *Middle (Unknown)*: The answer is unresponsive (e.g., changes the topic) or uninformative (e.g., "I don't know" answer). It should imply or lean towards neither Yes nor No.

Annotator Demographics Two annotators including a female and a male are recruited for the dataset annotation. Their ages range from 26 to 30 years old. Both of them are from Asia and with a graduate degree in Computer Science.

Annotation Agreements Figure 2 shows the percentages of disagreements between the annotators. Most disagreements are minor (between (a) Yes or No and (b) Middle). Recall that Cohen's κ inter-annotator agreements are between 0.66 to 0.69.

C Additional Details to Interpret Indirect Answers

Details and Hyperparameters Referring to Section 5.2, we use an off-the-shelf RoBERTa-base model (125M parameters) from Hugging Face (Wolf et al., 2020). We run the experiments on a single NVIDIA Tesla V100 (32GB) GPU. Depending on the sizes of the training datasets, it takes approximately 10 minutes to an hour to train one

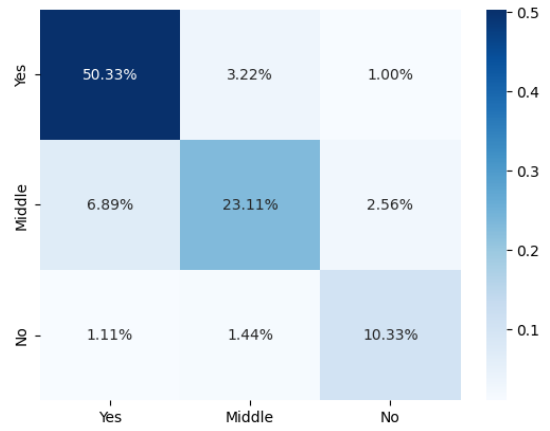

Figure 2: Heatmap of the inter-annotator agreements. The percentages are the average of three benchmarks (total: 900 instances). Most disagreements are between (1) *Yes* or *No* and (2) *Middle*.

epoch. Table 9 shows the hyperparameters that lead to the highest F1 score in interpreting indirect answers.

We also tune the number of additional training instances and blending factor α as other parameters. We choose the number of additional training instances from (2k, 5k, 10k) (or all instances if they are less than the number), and the α factor from (0.2, 0.5, 0.8). We report the tuned training size and the α factor that yields the highest F1 score in Table 10 and Table 11.

Additional Results and Metrics with RoBERTa To better interpret our experimental results, we report results (F1 score) per label in Table 12 and results with additional metrics (Precision, Recall, and F1 score) in Table 13. These results complement Table 4.

Additional Results with BART To minimize the variations by different models, we conduct experiments with BART-base using the same experimental setting as RoBERTa-base. Table 14 shows the results. Overall, BART underperforms RoBERTa on this task.

Experimental Details with LLMs Referring to Section 5.5, we test our benchmark with GPT-3.5 (*gpt-35-turbo*), Alpaca (7B parameters), and Llama 2 (7B parameters). Figure 3 shows the prompts. For GPT-3.5, we call the API from Microsoft Azure. We set the *temperature* to 0.1, *top_p* to 0.1, and *max_tokens* to 4 for optimal generation results. Both Alpaca and Llama are hosted locally using a single NVIDIA A100 (80GB) GPU.

```
Below is an instruction and a yes-no
question-answer pair input. Write a response
that appropriately completes the request.

### Instruction: I need you to help me
understand indirect answers to yes-no
questions. Indirect answers can be
interpreted with three meanings: Yes,
No, and Middle. Simply reply Yes, No or
Middle based on the question and answer.

### Input:

Question: "<Question from benchmarks>"

Answer: "<Answer from benchmarks>"

Does the answer mean Yes, No or Middle?

### Response:
```

Figure 3: Prompts used with GPT, Alpaca, and Llama.

Error Analysis on LLMs Results We conduct an error analysis for the results obtained with GPT-3.5 and the 4-shot prompt. Table 15 lists the error distributions and errro examples.

Hyperparameters	Training with extra data	Blended training
Maximum Epochs	30	20
Warmup steps	500	200
Batch Size	32	16
Optimizer	AdamW	AdamW
Learning rate	2e-5	2e-5
Weight decay	1e-2	1e-2
Gradient clipping	1.0	1.0

Table 9: Tuned hyperparameters for experiments to interpret indirect answers with RoBERTa-base model.

	Existing Benchmarks		Our Benckmarks		
	Circa	SWDA_IA	Tennis	Movie	Air
RoBERTa, training with					
in-domain instances and					
Circa	n/a	n/a	2k	2k	10k
SwDA-IA	n/a	n/a	2k	2k	5k
Circa+SwDA-IA	n/a	n/a	2k	2k	10k
all additional instances and					
Circa	5k	2k	5k	5k	2k
SwDA-IA	2k	2k	5k	2k	10k
Circa+SwDA-IA	10k	10k	10k	5k	10k

Table 10: Number of additional instances used in training. We report the number (in thousands) that yields the highest F1 score. This table complements the third block of Table 4.

	Existing Benchmarks		Our Benckmarks		
	Circa	SWDA_IA	Tennis	Movie	Air
RoBERTa, blended training with					
in-domain instances and					
Circa	n/a	n/a	0.5	0.2	0.5
SwDA-IA	n/a	n/a	0.5	0.8	0.8
Circa+SwDA-IA	n/a	n/a	0.2	0.2	0.5
all additional instances and					
Circa	0.5	0.5	0.2	0.2	0.8
SwDA-IA	0.2	0.2	0.2	0.8	0.8
Circa+SwDA-IA	0.2	0.5	0.5	0.8	0.8

Table 11: Tuned Blending factor (α). We report the α factor that yields the highest F1 score. This table complements the fourth block of Table 4.

	Circa				SWDA-IA			
	Yes	No	Mid	All	Yes	No	Mid	All
Majority Baseline	0.74	0.00	0.00	0.43	0.00	0.00	0.66	0.32
RoBERTa, training with								
Circa	0.95	0.92	0.50	0.93	0.72	0.25	0.12	0.51
SwDA-IA	0.75	0.64	0.15	0.69	0.77	0.52	0.29	0.63
Circa+SwDA-IA	0.95	0.93	0.38	0.93	0.79	0.57	0.44	0.68
RoBERTa, training with								
in-domain instances and								
Circa	——n/a——				——n/a——			
SwDA-IA	——n/a——				——n/a——			
Circa+SwDA-IA	——n/a——				——n/a——			
all additional instances and								
Circa	0.95	0.92	0.41	0.93	0.76	0.45	0.00	0.56
SwDA-IA	0.82	0.68	0.13	0.74	0.80	0.56	0.32	0.67
Circa+SwDA-IA	0.95	0.93	0.43	0.93	0.79	0.61	0.47	0.70
RoBERTa, blended training with								
in-domain instances and								
Circa	——n/a——				——n/a——			
SwDA-IA	——n/a——				——n/a——			
Circa+SwDA-IA	——n/a——				——n/a——			
all additional instances and								
Circa	0.95	0.93	0.51	0.93	0.74	0.37	0.05	0.54
SwDA-IA	0.80	0.61	0.04	0.71	0.77	0.51	0.35	0.64
Circa+SwDA-IA	0.95	0.92	0.48	0.93	0.79	0.61	0.41	0.68

	Tennis				Movie				Air			
	Yes	No	Mid	All	Yes	No	Mid	All	Yes	No	Mid	All
Majority Baseline	0.63	0.00	0.00	0.34	0.00	0.00	0.70	0.36	0.95	0.00	0.00	0.84
RoBERTa, training with												
Circa	0.62	0.38	0.13	0.40	0.50	0.50	0.21	0.34	0.91	0.26	0.00	0.84
SwDA-IA	0.61	0.33	0.23	0.42	0.49	0.29	0.34	0.37	0.70	0.07	0.00	0.63
Circa+SwDA-IA	0.72	0.55	0.25	0.52	0.53	0.43	0.27	0.37	0.93	0.17	0.00	0.84
RoBERTa, training with												
in-domain instances and												
Circa	0.70	0.42	0.07	0.43	0.55	0.46	0.02	0.24	0.92	0.30	0.00	0.85
SwDA-IA	0.68	0.45	0.05	0.41	0.57	0.41	0.15	0.31	0.92	0.27	0.00	0.84
Circa+SwDA-IA	0.67	0.49	0.22	0.48	0.60	0.43	0.18	0.34	0.93	0.21	0.00	0.85
all additional instances and												
Circa	0.68	0.37	0.24	0.44	0.50	0.43	0.05	0.24	0.90	0.12	0.00	0.82
SwDA-IA	0.70	0.46	0.27	0.51	0.59	0.42	0.07	0.28	0.90	0.17	0.00	0.82
Circa+SwDA-IA	0.67	0.50	0.36	0.53	0.57	0.47	0.33	0.42	0.92	0.22	0.00	0.84
RoBERTa, blended training with												
in-domain instances and												
Circa	0.65	0.38	0.00	0.37	0.59	0.42	0.13	0.31	0.90	0.24	0.00	0.83
SwDA-IA	0.69	0.45	0.23	0.48	0.58	0.48	0.22	0.36	0.92	0.23	0.00	0.84
Circa+SwDA-IA	0.70	0.55	0.42	0.57	0.57	0.49	0.26	0.39	0.92	0.15	0.00	0.84
all additional instances and												
Circa	0.63	0.36	0.19	0.42	0.53	0.48	0.38	0.44	0.90	0.18	0.00	0.82
SwDA-IA	0.69	0.47	0.27	0.50	0.62	0.41	0.29	0.40	0.93	0.30	0.00	0.85
Circa+SwDA-IA	0.69	0.55	0.49	0.59	0.59	0.42	0.48	0.49	0.94	0.24	0.00	0.86

Table 12: Detailed results (F1 score) obtained with RoBERTa per label. These results complement Table 4.

	Circa			SWDA-IA		
	P	R	F1	P	R	F1
Majority Baseline	0.34	0.59	0.43	0.24	0.49	0.32
RoBERTa, training with						
Circa	0.93	0.93	0.93	0.52	0.58	0.51
SwDA-IA	0.69	0.70	0.69	0.65	0.65	0.63
Circa+SwDA-IA	0.93	0.93	0.93	0.68	0.69	0.68
RoBERTa, training with						
in-domain instances and						
Circa	————n/a———-			————n/a————		
SwDA-IA	————n/a———-			————n/a————		
Circa+SwDA-IA	————n/a———-			————n/a————		
all additional instances and						
Circa	0.93	0.93	0.93	0.67	0.63	0.56
SwDA-IA	0.74	0.75	0.74	0.66	0.68	0.67
Circa+SwDA-IA	0.93	0.93	0.93	0.70	0.71	0.70
RoBERTa, blended training with						
in-domain instances and						
Circa	————n/a———-			————n/a————		
SwDA-IA	————n/a———-			————n/a————		
Circa+SwDA-IA	————n/a———-			————n/a————		
all additional instances and						
Circa	0.92	0.93	0.93	0.57	0.62	0.54
SwDA-IA	0.72	0.72	0.71	0.65	0.66	0.64
Circa+SwDA-IA	0.93	0.93	0.93	0.69	0.71	0.68

	Tennis			Movie			Air		
	P	R	F1	P	R	F1	P	R	F1
Majority Baseline	0.25	0.47	0.34	0.29	0.55	0.36	0.80	0.90	0.84
RoBERTa, training with									
Circa	0.53	0.47	0.40	0.60	0.41	0.34	0.84	0.85	0.84
SwDA-IA	0.61	0.33	0.42	0.53	0.40	0.37	0.83	0.53	0.63
Circa+SwDA-IA	0.60	0.58	0.52	0.57	0.40	0.37	0.83	0.86	0.84
RoBERTa, training with									
in-domain instances and									
Circa	0.69	0.51	0.43	0.72	0.34	0.24	0.84	0.86	0.85
SwDA-IA	0.70	0.52	0.41	0.73	0.38	0.31	0.84	0.85	0.84
Circa+SwDA-IA	0.60	0.53	0.48	0.60	0.41	0.34	0.84	0.88	0.85
all additional instances and									
Circa	0.75	0.49	0.44	0.70	0.35	0.24	0.83	0.82	0.82
SwDA-IA	0.65	0.57	0.51	0.58	0.36	0.28	0.84	0.81	0.82
Circa+SwDA-IA	0.67	0.57	0.53	0.59	0.46	0.42	0.22	0.00	0.84
RoBERTa, blended training with									
in-domain instances and									
Circa	0.30	0.48	0.37	0.59	0.39	0.31	0.83	0.82	0.83
SwDA-IA	0.60	0.53	0.48	0.64	0.43	0.36	0.83	0.85	0.84
Circa+SwDA-IA	0.70	0.55	0.57	0.66	0.45	0.39	0.83	0.85	0.84
all additional instances and									
Circa	0.57	0.44	0.42	0.62	0.47	0.44	0.83	0.82	0.82
SwDA-IA	0.66	0.54	0.50	0.68	0.45	0.40	0.84	0.87	0.85
Circa+SwDA-IA	0.60	0.60	0.59	0.62	0.50	0.49	0.84	0.88	0.86

Table 13: Results obtained with RoBERTa in Precision, Recall and F1 score. These results complement Table 4.

	Existing Benchmarks		Our Benckmarks		
	Circa	SWDA-IA	Tennis	Movie	Air
Majority Baseline	0.43	0.32	0.34	0.36	0.84
BART, training with					
Circa	0.92	0.55	0.41	0.27	0.84
SwDA-IA	0.72	0.55	0.43	0.35	0.83
Circa+SwDA-IA	0.92	0.66	0.46	0.36	0.84
BART, training with					
in-domain instances and					
Circa	n/a	n/a	0.45	0.24	0.78
SwDA-IA	n/a	n/a	0.40	0.24	0.78
Circa+SwDA-IA	n/a	n/a	0.46	0.30	0.85
all additional instances and					
Circa	0.92	0.54	0.35	0.32	0.83
SwDA-IA	0.59	0.61	0.43	0.39	0.86
Circa+SwDA-IA	0.92	0.63	0.45	0.41	0.88
BART, blended training with					
in-domain instances and					
Circa	n/a	n/a	0.41	0.28	0.86
SwDA-IA	n/a	n/a	0.38	0.25	0.82
Circa+SwDA-IA	n/a	n/a	0.49	0.37	0.86
all additional instances and					
Circa	0.92	0.52	0.41	0.31	0.86
SwDA-IA	0.58	0.65	0.36	0.32	0.87
Circa+SwDA-IA	0.92	0.66	0.52	0.42	0.88

Table 14: Results (F1 score) for interpreting indirect answers to yes-no questions with BART. BART underperforms RoBERTa on this task.

Gold	Prediction	%			Example
		Tennis	Movie	Air	
Y, N, M	Fail to predict due to content filtering	1	25	3	Q: Have you got to tell her your life story? A: I'll say what I **** please.
Yes	No	21	2	3	Q: Did you follow the Lance Armstrong stuff? A: A little bit.
Yes	Middle	38	35	92	Q: Are there any specifications? A: My departure time is evening.
No	Yes	0	1	0	Q: Sure, do you prefer any class? A: I am ok with any class.
No	Middle	4	19	1	Q: Are you working? A: Working? What do you mean, working? I'm walking.
Middle	Yes	1	2	0	Q: He went out and bought himself men's cologne the other day. Did I tell you that? A: Larry bought himself cologne?
Middle	No	35	16	1	Q: Any reason why you felt you were down in the first three sets in terms of quality? A: What's the question?

Table 15: Error distributions with GPT-3.5 and the 4-shot prompt. The error percentages are categorized by Gold label and (wrong) Predictions.