Training LLMs to Recognize Hedges in Dialogues about Roadrunner Cartoons

Amie J. Paige[*Ψ], **Adil Soubki**[*💻🖩], **John Murzaku**[*💻🖩], **Owen Rambow**[💬🖩], **Susan E. Brennan**[Ψ]

[💻]Department of Computer Science, [💬]Department of Linguistics, [Ψ]Department of Psychology [🖩]Institute for Advanced Computational Science, Stony Brook University

[*]These authors contributed equally to this study.

amie.paige@stonybrook.edu, {asoubki,jmurzaku}@cs.stonybrook.edu

Abstract

Hedges allow speakers to mark utterances as provisional, whether to signal non-prototypicality or "fuzziness", to indicate a lack of commitment to an utterance, to attribute responsibility for a statement to someone else, to invite input from a partner, or to soften critical feedback in the service of face-management needs. Here we focus on hedges in an experimentally parameterized corpus of 63 Roadrunner cartoon narratives spontaneously produced from memory by 21 speakers for co-present addressees, transcribed to text (Galati and Brennan, 2010). We created a gold standard of hedges annotated by human coders (the *Roadrunner-Hedge corpus*) and compared three LLM-based approaches for hedge detection: fine-tuning BERT, and zero and few-shot prompting with GPT-4o and LLaMA-3. The best-performing approach was a fine-tuned BERT model, followed by few-shot GPT-4o. After an error analysis on the top performing approaches, we used an *LLM-in-the-Loop* approach to improve the gold standard coding, as well as to highlight cases in which hedges are ambiguous in linguistically interesting ways that will guide future research. This is the first step in our research program to train LLMs to interpret and generate collateral signals appropriately and meaningfully in conversation.

1 Introduction

The virtuosity of LLMs such as ChatGPT has led some to the impression that AI already converses (or will soon be able to converse) as people do. But as language users, LLMs and humans are quite different. The underlying foundations for learning by these distinct kinds of language users share little in common: Humans learn as infants to interact with others well before they learn their first words, and once word learning begins, they can pick up a new word in one or just a few exposures, whereas LLMs are pre-trained on humanly unfathomable quantities of text without ever learning to interact. Transformer-based chat programs can generate paragraphs-worth of text remarkably well without modeling the coordination between agents–but is this conversation?

Whether a sequence of prompts and responses exchanged in a dialogue between an LLM agent and a human counts as truly (rather than superficially) "conversational" depends on how conversation is conceptualized. Conversation is often presumed to be the passing back and forth of messages (a "message model"); but that does not explain phenomena common to spontaneous conversation such as incremental turns, clarifications, and repair. Here we conceptualize conversation as a collaborative process of grounding meanings (seeking and providing evidence) during which two or more partners signal, coordinate, and align their beliefs or cognitive states (Brennan, 2005; Clark and Wilkes-Gibbs, 1986). This leads to a broader research agenda that we hope will push generative AI to model phenomena such as a partner's knowledge or theory of mind, mutual beliefs or common ground, as well as when to take initiative in a dialogue.

The main contributions of this work include:
(i) After grounding the project in psycholinguistic theory (Section 2) and related work (Section 3), we present the Roadrunner-Hedge Corpus (Section 4), a corpus of spontaneous face-to-face narratives annotated for hedging.[1]
(ii) We describe a set of experiments on this corpus using zero-shot, few-shot, and fine-tuning methods on modern LLMs (Section 5).
(iii) We perform a detailed error analysis pinpointing where LLMs fail in detecting hedges (Section 6). With this analysis, we take an LLM-in-the-Loop approach to correcting gold annotations, reducing errors in our top performing systems.

We conclude with a discussion and implications of our results in Section 7, limitations and the future

[1] https://github.com/cogstates/hedging

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 204–215
September 18–20, 2024.

of our work in Section 8, and a final summary of our salient contributions in Section 9.

2 Theoretical Foundations from Psycholinguistics

In conversation, people communicate not only about the purpose or topic at hand, but they also communicate meta-information about what they're saying within the context of interaction, or *collateral signals* (Clark, 1996). Along with providing evidence for grounding in conversation, about whether a prior turn has been understood as intended (Clark and Brennan, 1991), collateral signals can also provide information about the speaker's relationship with the content of their message—how confident they are in what they are saying, whether it is difficult to recall or express, and whether they would welcome input from their partner. In this project, we focus on a particular kind of collateral signal used for coordination, *hedges*.

2.1 Why Speakers Hedge

There have been several proposals for why speakers hedge. Hedges have been claimed to characterize powerless "feminine" language (Lakoff, 1973) or to serve a politeness function by minimizing threat to a partner's "face" (Brown and Levinson, 1987); see also (Fraser, 2010). Hedges have also been thought to convey a certain "fuzziness" of category membership when a speaker means to describe a non-prototypical member of a category (e.g., a penguin belonging to the bird category; Lakoff, 1975). Prince et al. (1982) suggested that hedges play two functions: First, to make propositional content less exact (*approximators*, e.g. "sort of") and second, to change the relationship a speaker has to the content of their message (*shield hedges*). Shield hedges are further divided into *plausibility* shields that signal a lack of commitment to the content of a message ("I think his feet were blue," Prince et al., 1982, p. 5), and *attribution* shields that assign responsibility for a message to a source other than the speaker or writer themself ("According to her estimates..." Prince et al., 1982, p. 13).

Several experimental studies have demonstrated how hedges can convey speakers' commitment to what they are saying. For example, in a question-answering task, people trying to recall the answers to trivia questions produced more disfluencies, longer latencies, more rising intonation, and more expressions of doubt when they reported having a low *feeling of knowing* about an answer. This metacognitive information was confirmed to be accurate when compared to the ground truth in the form of their answer to the same (multiple-choice) question later (Smith and Clark, 1993). Not only are hedges informative as collateral signals about what a speaker knows, but they are accurately interpreted as such by listeners (Brennan and Williams, 1995).

That hedges function as interactional signals in extended dialogue is evident from studies of referential communication. Typically in such studies, two partners who can't see each other converse in order to arrange and rearrange duplicate sets of objects in matching orders, with the objects needing to be distinguished from similar objects or consisting of Tangrams (abstract geometric shapes unassociated with any conventional or lexicalized labels). Hedges are common in initial referring expressions, where they tend to appear in wordy, disfluent, and often tentative descriptions, and then they drop out in repeated referring expressions once partners have reached a shared conceptualization for that object (marked by entrainment, or re-using the same shortened referring expression) (Brennan and Clark, 1996; Galati and Brennan, 2021), as in this sequence of repeated references to the same object over multiple rounds (adapted from Brennan and Clark, 1996, p. 1488):

Round 1: "a car, sort of silvery purple colored"
Round 2: "purplish car going to the left"
. . .
Round 5: "the purple car"

In another study that required triads of strangers to reach consensus while recalling the events from a movie clip that they had watched earlier, the speakers often hedged their contributions to the conversation, presumably to mark a lack of certainty about an utterance and an openness to being corrected by their partners (Brennan and Ohaeri, 1999). For example, from a triad that communicated by speaking face-to-face:

> *Yeah, they were sitting around the fireplace in the night... sort of like a bedtime story kind of thing*

People who did the same task by texting rather than speaking used fewer words, but still hedged:

> *We all agree it was a wreathy thingy on his neck???*

2.2 How Listeners React to Hedges

Hedges convey meaningful information that can affect listeners' subsequent behavior; a handful of psychological studies have measured the impacts of hedges on listeners. For example, children exposed to new words from a speaker who hedged learned fewer novel words compared to children exposed to a speaker who did not hedge (Sabbagh and Baldwin, 2001). Listeners rated utterances as more uncertain when they included shield hedges (e.g., "I think it was a mug"), and these ratings were related to speakers' ratings of their own uncertainty in identifying an image (Pogue and Tanenhaus, 2018). Moreover, addressees in a referential communication task expended more effort while grounding (they produced more low-confidence responses such as clarification questions) to demonstrate understanding when the speaker's description had contained a hedge (Dahan, 2023).

Hedges also influence which details are retold to another person; in one study, hedged details were less likely to be repeated to another addressee as compared to unhedged details (Liu and Fox Tree, 2012), although in the same study, hedged information presented in a story was more likely to be remembered by listeners; this was thought to stem from deeper engagement with hedged information when it was first presented (Liu and Fox Tree, 2012). And in tutoring dialogues, where face management can be particularly important, students were more successful at solving problems when their peer tutors used hedges (Madaio et al., 2017).

3 Related Computational Work

3.1 Hedging

Several research programs have examined hedges and the criteria for coding them, with computational goals that include automatic hedge detection. Hedging is domain-specific, meaning that their forms and frequencies vary across corpora; they are also context-specific, as they cannot be identified accurately simply by searching for strings (Prokofieva and Hirschberg, 2014). Hedges are distributed differently within different corpora (ibid).

Hedges are often ambiguous and difficult to code in the absence of dialogue context. In "I think it's a little odd," *I think* is often a hedge, but might not be when proffered in response to a question ("So what do *you* think?"). Hedges in spoken utterances may be disambiguated by stress and other intonational cues, as in "*I* think he'll win!" (not a hedge) vs. "I *think* he'll win?" (a hedge). Previous work found many cases of tokens that can serve as hedges as well as non-hedges, with systematic tests for coders to use in annotating them for gold standards (Prokofieva and Hirschberg, 2014; Ulinski and Hirschberg, 2019; Ulinski et al., 2018).

The coding of hedges is complicated by the fact that in spoken dialogue, they often co-occur with speech disfluencies. In some contexts, it may be difficult to distinguish these two kinds of signals (Prokofieva and Hirschberg, 2014), particularly since listeners can use disfluencies in much the same way they can use hedges to draw conclusions about the speaker's mental state (Arnold et al., 2003, 2007)

A strong motivation for computational work on hedging comes from work on computer-assisted learning by Cassell and colleagues, specifically tutoring dialogues (Abulimiti et al., 2023a,b; Raphalen et al., 2022). Most similar to our work is Raphalen et al. (2022), where the authors propose a model that combines rule-based classifiers and machine learning models with interpretable features such as unigram and bigram counts, part-of-speech tags, and LIWC categories to identify and classify hedge clauses. Our work differs in two major ways: first, our work operates on the token level rather than on the clause level. Token level classification makes possible a truly end-to-end approach (classifying all hedge and non-hedge tokens in utterances). Second, we include experiments with modern LLMs and offer a detailed error analysis into their mistakes; stemming from this error analysis, we use an LLM-in-the-Loop approach (Dai et al., 2023) to correcting gold standard hedge codings.

3.2 Belief

Hedging and the notion of belief (how committed the speaker is to the truth of an event) are closely related; hedges are often used by speakers to indicate a lack of belief or commitment towards what they say. Ulinski et al. (2018) improved belief classification using a hedge detector, yielding an improvement for the non-committed and reported belief labels.

Corpora Several corpora have been created that annotate the author's degree of belief (Diab et al., 2009; Prabhakaran et al., 2010; Lee et al., 2015; Stanovsky et al., 2017; Rudinger et al., 2018; Pouran Ben Veyseh et al., 2019; Jiang and de Marn-

Hedge Type	Example(s)
Like (not used as a simile, verb, or comparison)	"and then he like went over by..."
You know (not to communicate another's knowledge or as a discourse marker)	"and you know as he's falling down"
Just (not used to mean "only")	"he just jolts away"
Approximators/Rounders	"kind of", "about"
Proxies (for a detail the speaker cannot or chooses not to recall)	"thing," "whatever," "or something," "and everything"
Morpheme suffixes to content words	"circley," "springy"
Expressions of doubt attached to claims; self-speech	"I don't know," "maybe," "I guess," "what's it called?"
Tag questions and try markers	"he's standing there, right?"

Table 1: Coding scheme used to mark hedges in corpus.

effe, 2021). There are two corpora that further annotate nested beliefs of the sources mentioned in the text: FactBank (Saurí and Pustejovsky, 2009) and the Modal Dependency corpus (Yao et al., 2021).

Machine Learning Approaches Modern neural methods for belief detection include LSTMs with multi-task or single-task approaches (Rudinger et al., 2018), using BERT representations alongside a graph convolutional neural network (Pouran Ben Veyseh et al., 2019), or fine-tuning BERT with a span self-attention mechanism Jiang and de Marneffe (2021). Recent state-of-the-art work finds that fine-tuning RoBERTa (Murzaku et al., 2022) or fine-tuning Flan-T5 (Murzaku et al., 2023) yields the best performance on most corpora. For the label *Underspecified* (or, corresponding to no commitment and/or a hedge), these modern methods yield f-measures in the low to high 80s. We also have prior work exploring multi-modal approaches to belief detection (Murzaku et al., 2024).

4 The Roadrunner-Hedge Corpus

For training and testing, we obtained a corpus (Galati and Brennan, 2010) of spontaneous narratives produced from memory by 20 speakers who had watched a Roadrunner cartoon. Each speaker narrated the story face-to-face to an audience, a total of three times: first to a naïve addressee, a second time to the same addressee, and a third time to a new naïve addressee (with the latter two episodes counterbalanced for order). The original experiment was designed to detect differences in collateral signals (intelligibility vs. attenuation of speech and gestures) stemming from the speaker's vs. the addressee's knowledge states–that is, whether the story was new for the speaker (told for the first time) vs. old (retold), compared to the addressee's knowledge state (new vs. heard for the second time). Findings included that the attenuation of both referring expressions (Galati and Brennan, 2010) and gestures (Galati and Brennan, 2014) were driven by *both* speakers' and addressees' knowledge states–that is, shortened upon retelling the story to the same addressee, but lengthened upon retelling to a new addressee.

Gold Standard Coding. The original corpus transcribed the spontaneous narratives in detail, including speaking turns and disfluencies (for details, see Galati and Brennan, 2010), segmented into lines by installments that corresponded to narrative elementss in the cartoons. We annotated hedges on the original Roadrunner corpus to create the gold standard for hedge training and detection (the *Roadrunner-Hedge* corpus; see `https://github.com/cogstates/hedging` for the annotation codebook).

The Roadrunner-Hedge corpus is distributed as a csv file. It is structured as a total of 5,508 lines, over a quarter of which (N=1424) include one or more hedges. The first author annotated hedges in the corpus as in Table 1. Although disfluencies such as fillers (*uh, um*) and re-starts can function as hedges, we made a principled decision to not code them as such; hedges in our corpus are presumed to be shaped by the speaker's intention, whereas disfluencies are not necessarily under a speaker's control as a communicative signal, but may reflect difficulties in speaking (Grice, 1957; Clark, 1994). Overall word counts for hedges and non-hedges are 1,728 and 38,018 words respectively. Most hedges are one word, but a few cases contain many words. For each line in the csv file (corresponding to a narrative element), hedges are listed (separated by commas) in an adjacent cell. Each line has an average of 0.33 hedges.

Inter-Rater Reliability. To compute inter-rater reliability, a trained research assistant coded 7 randomly-selected transcripts with no overlapping speakers (10% of the corpus). We calculated Cohen's Kappa from each word marked as a hedge within each transcript. There was high agreement

between coders, with $\kappa = 0.985$.

Corpus Analysis. The Roadrunner-Hedge corpus, like the tutoring dialogues used by Abulimiti et al. (2023b); Raphalen et al. (2022), has fewer cases with hedges than without, but with more hedges per segment overall (25.85% of lines vs. 14.26% of turns respectively).

Over the three versions of the cartoon story produced by each speaker, hedges were most frequent in the first telling when the story was new to both speaker and addressee and least frequent when told to the same addressee a second time, consistent with the original findings from Galati and Brennan that collateral signals are affected by the knowledge states of both speaker and addressee.

5 Experiments

5.1 Experimental Setup

In this section, we present our hedge classification experiments on the Roadrunner-Hedge corpus, conducted by fine-tuning BERT and performing zero-shot and few-shot experiments with state-of-the-art LLMs. For all experiments, we performed five-fold cross validation using a fixed seed (42), splitting the corpus into a 80/20 train/test split. For our fine-tuning experiments, we did not perform any hyperparameter tuning, and therefore do not have a validation set.

We performed all zero-shot, few-shot, and fine-tuning experiments on the fold's respective test sets and report the average and standard deviation over all five folds test sets for F1, precision, and recall.

5.2 Zero Shot and Few Shot

For the zero-shot and few-shot experiments, we used GPT-4o (OpenAI, 2024) and LLaMA-3-8B-Instruct (AI@Meta, 2024), as these two LLMs have achieved state-of-the-art results in many zero-shot or few-shot benchmark tasks.

We conducted two classes of zero-shot and few-shot experiments: count/list generation and BIO tag generation. Both prompts began with an instruction detailing the specific task, and a random example. In our few-shot experiments, we provided three fixed hand-crafted examples. For our count/list generation, we prompted the models to list the integer number of hedges present in the utterance and then generated a list of the exact hedge words. For our BIO tag generation, we generated the tokens and their respective tags, where label *B* represents the beginning of a hedge token or span, *I* represents the inside of a hedge span, and *O* represents another token, all separated by *"/"*. For example, given the utterance *It is like warm*, we prompted the model to generate *It/O is/O like/B warm/O*.

We provide our exact prompts with their corresponding instructions in Appendix A. For our GPT-4o experiments, we used the default OpenAI API hyperparameters and a temperature of 1.0.

5.3 Fine-tuning

We performed all fine-tuning experiments using BERT (Devlin et al., 2019), specifically bert-base-uncased. We also performed experiments with the large variants of the model (bert-large), newer encoder-only models like RoBERTa (Liu et al., 2019) and DeBERTa-v3 (He et al., 2021), and encoder-decoder models like Flan-T5 (Chung et al., 2022), but got either worse or closely similar results.

Task Description All experiments followed a standard BIO token labelling approach to classify hedge tokens (B), tokens inside of hedge spans (I), and all other tokens (O). In other words, given an input utterance of n tokens, the respective BIO labels were output for each of the n tokens. Following the same example as described in our zero-shot and few-shot experiments in Section 5.2, we fine-tuned BERT to classify the tokens as *It/O is/O like/B warm/O*.

Hyperparameters We followed a standard fine-tuning approach, fine-tuning for a fixed 5 epochs. We set the batch size to 16 and learning rate to 2e-5. We performed five-fold cross validation and test on each folds respective test set. We did not perform any hyperparameter tuning.

5.4 Results

The performance of the models is shown in Table 2, which reports average precision (P), recall (R), and F1 over the five-folds. For our zero-shot, few-shot, and fine-tuning experiments, these metrics are calculated on each fold's test set and then averaged.

Despite its much smaller parameter count, BERT fine-tuned for BIO tagging outperforms even the best scoring prompting approaches by nearly 20 points in F-measure. This is consistent with a general trend in the literature of more parameter efficient fine-tuning approaches outperforming larger zero-shot and few-shot methods (Liu et al., 2022), though the gap here is larger than one might expect.

Model	Training	Prompt	Precision (P)	Recall (R)	F1 Score (F1)
BERT	Finetuned	-	0.883 ± 0.015	0.934 ± 0.012	0.908 ± 0.010
GPT-4o	Few-Shot	List	0.613 ± 0.027	0.848 ± 0.018	0.712 ± 0.021
LLaMA-3	Few-Shot	List	0.518 ± 0.035	0.799 ± 0.022	0.628 ± 0.031
GPT-4o	Few-Shot	BIO	0.514 ± 0.024	0.766 ± 0.036	0.616 ± 0.030
GPT-4o	Zero-Shot	List	0.430 ± 0.014	0.711 ± 0.004	0.536 ± 0.012
GPT-4o	Zero-Shot	BIO	0.436 ± 0.026	0.618 ± 0.033	0.510 ± 0.028
LLaMA-3	Few-Shot	BIO	0.298 ± 0.018	0.625 ± 0.016	0.404 ± 0.019
LLaMA-3	Zero-Shot	BIO	0.167 ± 0.014	0.428 ± 0.019	0.240 ± 0.017
LLaMA-3	Zero-Shot	List	0.274 ± 0.023	0.146 ± 0.010	0.190 ± 0.011

Table 2: Average performance metrics over the five folds with standard deviations for different models, training methods, and prompt types, ordered by F1 score.

In comparisons of the zero-shot and few-shot prompting methods, the few-shot models, unsurprisingly, performed better. The few-shot experiments averaged an F1 of 0.59, 22 points higher than the zero-shot models average of 0.37.

Of the two output formats prompted for, listing and BIO, the listing approach performed better. On average, models instructed to output a list had an F1 of 0.52 compared to 0.44 for those instructed to perform BIO tagging.

Among the two LLMs prompted, GPT-4o always performed best. Across all models and approaches, including fine-tuned BERT, precision tended to be lower than recall, with a mean of 0.46 for precision compared to 0.65 for recall. In other words, the models over-predicted the presence of hedges.

6 Error Analysis

While the fine-tuned BERT model performed fairly well, a certain number of cases did not align with the gold labels in the data. We performed error analysis to understand whether there were any systematic deviations from the corpus annotation.

We conducted an error analysis on the top two performing models, the fine-tuned BERT model and the GPT-4o Few-shot List (FSL) model (F1 = 0.91 and 0.71, respectively). Starting with the first fold, we selected the first hundred errors to categorize. These errors are broadly divided into instances where the models failed to detect a hedge (false negatives) and instances where models returned cases that were not annotated hedges (false positives). The remaining errors fell into two other categories: a gold error category, wherein errors in the (human) annotation were discovered, and an "other" category.

Of the hundred errors sampled from the BERT model, approximately the same number of errors were false negatives (26) as false positives (29). Of the hundred errors sampled from the GPT-4o FSL model, 66 were false positives and 25 were false negatives (reflecting the low precision and higher recall for this approach; see Table 3 and 4 for full error descriptions for BERT and GPT-4o FSL models).

Although the corpus annotation does not include the *type* of hedge (only the presence or absence of hedge tokens), our error analysis looked at hedge types in order to tease apart model behaviors. We observed systematic differences between models in their types of mismatches with the gold standard.

False Positives. First, the GPT-4o FSL model inaccurately classified disfluencies (e.g., "uh") as hedges in 37 of the 66 false positives reviewed, whereas BERT did not. Second, BERT showed quite a different pattern of mismatches than GPT-4o when classifying "like", returning false positives that always turned out to be comparatives (e.g., "it's like an open elevator"). These we considered to be true errors in their text form, although some may be ambiguities that could be resolved prosodically.

False Negatives. Tokens denoting approximator hedges (e.g. "that's *basically* it") were frequently misclassified as false negatives by BERT (9 of 26 false negatives reviewed), but never by the GPT-4o FSL model.

In addition, **Other** emerged as a category type for situations that could not clearly be described as false positives, false negatives, or gold errors. In the BERT model, these cases were typically segmentation errors (i.e., an inner token mislabeled as a beginning token).

Notably, the largest class of errors for the BERT

Gold Errors		**False Negative**		**False Positive**		**Other**	
Like	13	Approximator	9	*Like*	13	*I* should be *B*	4
Proxy	12	Proxy	8	*Just*	8	*O* should be *I*	3
Just	7	Self-talk	4	False proxy	4	*B* should be *I*	2
Approximator	1	*Like*	3	*You know*	2	Other	2
Other	1	*Just*	1	Misc. word	2		
		Morpheme	1				
Total	**34**		**26**		**29**		**11**

Table 3: Expanded error analysis on the BERT fine-tuned model, by hedge type.

Gold Errors		**False Negative**		**False Positive**		**Other**	
Approximator	4	*Just*	12	Disfluency tag	37	Other	1
Just	1	Proxy	8	Misc. word	15		
Like	1	*Like*	3	*Like*	7		
Proxy	1	Morpheme	1	Approximator	3		
Self-talk	1	Self-talk	1	Intensifiers	3		
				You know	1		
Total	**8**		**25**		**66**		**1**

Table 4: Expanded error analysis on the GPT-4o FSL model, by hedge type.

model was the **Gold Error** category (34 of 100). This was not the case for the GPT-4o model (only 9 gold errors). The BERT fine-tuned model revealed mistakes made by the human annotators for hedges denoted by "like", "just", and proxy hedges (e.g. "and stuff"). Upon closer inspection, some of these cases were ambiguous. For example, "he just hits the ground" could be taken to mean that the only action performed was hitting the ground (where "just" means only) or "just" might function to reduce the speakers' certainty (as in Madaio et al., 2017). Again, the text format of the storytelling corpus leaves some interpretations ambiguous that could be clarified with signals such as timing and prosodic stress.

The number of Gold Errors identified by the BERT model allowed us to modify the original gold annotation with missed cases and to re-evaluate the performance of our models more accurately – a sort of *LLM-in-the-Loop* approach (see Table 5).

7 Discussion

The results show that even enormous, recently released LLMs cannot reliably recognize hedges. There is no "emergent" ability in LLMs to understand full human linguistic behavior. On the other hand, when we explicitly train a small, rather old LLM (BERT) to perform our task by fine-tuning it, it performs quite well. What this shows is that detecting hedges is a capability that can be learned, but it cannot be learned in the manner that LLMs are taught, namely by simply ingesting large amounts of varied data. We interpret this to mean that if we want to make LLMs able to converse with humans as humans do, we need to understand what capabilities LLMs need and how to provide them with the ability to do so.

The prevalence of gold errors discovered by the BERT model raises two interesting points for discussion. First, some of these discrepancies identified by the BERT model were clearly errors made by the human coders; this was true in particular for proxies, which BERT coded for hedges more consistently than did human coders. This error analysis allowed us to iteratively improve the human coding before the final analysis, essentially deploying an LLM-in-the-Loop approach. Second, the discrepancies between BERT and gold coding on the tokens *just* and *like* highlight that these types of hedges have high potential for ambiguity–perhaps the very sort of ambiguity that could be resolved by prosody.

8 Limitations and Future Work

This work represents the first step in our research program that aims to train LLMs to use collateral signals in support of human-LLM dialogue. Once hedges can be recognized by an LLM, it remains to be shown that they can be meaningfully interpreted and generated. Relevant work by Cassell and col-

Model	Original Gold F1	LLM-in-the-Loop Gold F1	Error Reduction (%)
BERT	0.908 ± 0.010	0.925 ± 0.019	18.5%
GPT-4o Few-Shot	0.712 ± 0.021	0.721 ± 0.020	3.1%
GPT-4o Zero-Shot	0.510 ± 0.028	0.551 ± 0.011	8.4%

Table 5: F1 scores with standard deviations on the original corpus, F1 scores with standard deviations obtained on the corpus corrected after LLM-in-the-Loop, and the change in average performance for our top performing models.

leagues has shown that it is possible to generate hedges in tutoring dialogues, but not always positioned where they are most probable or useful (Abulimiti et al., 2023a). In future work, we plan experiments using top-performing models such as BERT and GPT-4o in high- and low-probability situations that systematically vary the certainty associated with prompted-for information (where hedges can be most useful). It is already clear from our pilot trials using ChatGPT 3.5 that LLMs hedge somewhat superficially (hedging where humans wouldn't and failing to hedge where humans would).

Domains of Dialogue. Here we have used human-generated dialogue from a single domain, retelling stories from Roadrunner cartoons; the training data are text transcripts of speech. Because the initiative was unbalanced in this collaborative task, most of the speaking in each triad was done by the the partner who viewed and retold the cartoon stories in series to the two co-present addressees.

A more balanced domain in which partners continuously monitor each other's understanding to do a physical task–such as matching pictures of difficult-to-describe objects–could yield more hedges, distributed differently. We plan to conduct similar tests to replicate the current results on such referential communication corpora collected previously in our lab.

It is interesting that despite the fact that there is not a single instance of dialogue in Roadrunner cartoons (apart from Roadrunner's smug, trademark "meep meep" upon escaping from Coyote), speakers who retell the story in a dramatic and humorous way do a great deal of what looks like quoting Coyote's and Roadrunner's reactions:

> *so then he's saying he's like gone all sad and stuff you know?*
>
> *and he's like whatever she's gonna be dead right?*

Such uses of *like* in this corpus match the quotation-as-demonstrations forms described by Clark and Gerrig (1990); they count as hedges in that the speaker marks what follows as *not verbatim*.

Training with audio input. Our results for detecting hedges in this transcribed spoken corpus are surprisingly strong, especially given that the LLMs we used were pre-trained primarily on originally written text. But it is well-known that features such as pausing and intonation are related to speakers' levels of commitment to and confidence in their utterances. We plan to incorporate audio into future hedging studies and will explore multi-modal neural architectures fusing both speech and lexical features as we did in (Murzaku et al., 2024) for belief recognition.

Reliability. It is critical to keep in mind that human and LLMs are very different sorts of agents. Psychometric tests show that individual humans are likely to respond consistently when tested repeatedly, whereas an LLM is not (Shu et al., 2024). LLMs have no sense of "self" and are likely to respond differently when re-prompted with the same prompt. To the extent that a hedge signals that a speaker does not wish to be held entirely accountable for what they're saying, hedging on the part of an LLM may actually be desirable as a way to encourage users to not assume they can hold it accountable. On the other hand, it may be desirable for an LLM to be able to signal its *confidence* – the reliability or quality (or lack thereof) of information it's presenting – through the presence or absence of hedges. Finally, it remains to be seen whether LLMs can learn about interaction through exposure to collateral signals in meaningful contexts.

9 Conclusion

Our project is grounded in psycholinguistic theory and aims to capture theory-of-mind aspects of hedging among discourse participants. We present the Roadrunner-Hedge corpus, with hedges annotated from naturally occurring dialogues by speakers describing Roadrunner cartoons. We use the corpus to

train and perform experiments on detecting hedges using BERT, GPT-4o, and LLaMA-3. We find that fine-tuning BERT significantly outperforms state-of-the-art LLMs in few-shot and zero-shot settings. With our systems outputs, we perform an error analysis and use an LLM-in-the-Loop approach to correct gold standard annotations. Our LLM-in-the-loop approach provided further error reductions on all models.

Ethical Considerations

The Roadrunner-Hedge corpus was collected with Institutional Review Board approval from undergraduate students who gave informed consent prior to participating in the experiments.

Acknowledgments

This material is based upon work supported in part by the National Science Foundation (NSF) under No. 2125295 (NRT-HDR: Detecting and Addressing Bias in Data, Humans, and Institutions) as well as by funding from the Defense Advanced Research Projects Agency (DARPA) under the CCU program (No. HR001120C0037, PR No. HR0011154158, No. HR001122C0034). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF or DARPA.

We thank both the Institute for Advanced Computational Science and the Institute for AI-Driven Discovery and Innovation at Stony Brook for access to the computing resources needed for this work. These resources were made possible by NSF grant No. 1531492 (SeaWulf HPC cluster maintained by Research Computing and Cyberinfrastructure) and NSF grant No. 1919752 (Major Research Infrastructure program), respectively.

We would also like to thank our reviewers for their helpful comments, as well as Kayla Hunt for assistance with reliability coding.

References

Alafate Abulimiti, Chloé Clavel, and Justine Cassell. 2023a. When to generate hedges in peer-tutoring interactions. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 572–583, Prague, Czechia. Association for Computational Linguistics.

Alafate Abulimiti, Chloé Clavel, and Justine Cassell. 2023b. How about kind of generating hedges using end-to-end neural models? *Preprint*, arXiv:2306.14696.

AI@Meta. 2024. Llama 3 model card.

Jennifer Arnold, Maria Fagnano, and Michael Tanenhaus. 2003. Disuencies signal theee, um, new information. *Journal of Psycholinguistic Research*, 32:25–36.

Jennifer Arnold, Carla Hudson Kam, and Michael Tanenhaus. 2007. If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of experimental psychology. Learning, memory, and cognition*, 33:914–30.

Susan E. Brennan. 2005. How conversation is shaped by visual and spoken evidence. In John Trueswell and Michael Tanenhaus, editors, *Approaches to studying world-situated language use: Bridging the language-as-product and language-as-action traditions*, pages 95–129. MIT Press.

Susan E. Brennan and Herbert H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Susan E. Brennan and J. O. Ohaeri. 1999. Why do electronic conversations seem less polite? the costs and benefits of hedging. In *ACM SIGSOFT Software Engineering Notes*.

Susan E Brennan and Maurice Williams. 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, 34(3):383–398.

P. Brown and S. C. Levinson. 1987. *Politeness: Some universals in language usage*. Cambridge University Press.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. H. chi, jeff dean, jacob devlin, adam roberts, denny zhou, quoc v. le, and jason wei. 2022. scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Herbert H. Clark. 1994. Managing problems in speaking. *Speech Communication*, 15:243–250.

Herbert H. Clark. 1996. *Using Language*. "Using" Linguistic Books. Cambridge University Press.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In L. B. Resnick, J. Levine, and S. D. Teasley, editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association.

Herbert H. Clark and Richard J. Gerrig. 1990. Quotations as demonstrations. *Language*, 66:764–805.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Delphine Dahan. 2023. Collaboration under uncertainty in unscripted conversations: The role of hedges. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 49:320–335.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging large language model for thematic analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9993–10001, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 68–73, Suntec, Singapore. Association for Computational Linguistics.

Bruce Fraser. 2010. Pragmatic competence: The case of hedging. *New Approaches to Hedging*, 9:15–34.

Alexia Galati and Susan E. Brennan. 2010. Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language*, 62:35–51.

Alexia Galati and Susan E. Brennan. 2014. Speakers adapt gestures to addressees' knowledge: implications for models of co-speech gesture. *Language, Cognition and Neuroscience*, 29:435 – 451.

Alexia Galati and Susan E. Brennan. 2021. What is retained about common ground? Distinct effects of linguistic and visual co-presence. *Cognition*, 215.

H Paul Grice. 1957. Meaning. *The philosophical review*, 66(3):377–388.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2021. He thinks he knows better than the doctors: BERT for event factuality fails on pragmatics. *Transactions of the Association for Computational Linguistics*, 9:1081–1097.

George Lakoff. 1975. Hedges: A study in meaning citeria and the logic of fuzzy concepts. *Journal of Philosophical Logic*, pages 458–508.

Robin Lakoff. 1973. Language and woman's place. *Language in Society*, 2(1):45–79.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648, Lisbon, Portugal. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Kris Liu and Jean Fox Tree. 2012. Hedges enhance memory but inhibit retelling. *Psychonomic bulletin & review*, 19:892–8.

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*.

Michael A. Madaio, Justine Cassell, and Amy E. Ogan. 2017. "i think you just got mixed up": confident peer tutors hedge to support partners' face needs. *International Journal of Computer-Supported Collaborative Learning*, 12:401–421.

John Murzaku, Tyler Osborne, Amittai Aviram, and Owen Rambow. 2023. Towards generative event factuality prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 701–715, Toronto, Canada. Association for Computational Linguistics.

John Murzaku, Adil Soubki, and Owen Rambow. 2024. Multimodal belief prediction. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2024*. International Speech Communication Association.

John Murzaku, Peter Zeng, Magdalena Markowska, and Owen Rambow. 2022. Re-examining FactBank: Predicting the author's presentation of factuality. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 786–796, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

OpenAI. 2024. Gpt-4o. `https://openai.com/index/hello-gpt-4o/`.

Amanda Pogue and Michael K. Tanenhaus. 2018. Learning from uncertainty: exploring and manipulating the role of uncertainty on expression production and interpretation. In *Proceedings of the 40th Annual Conference of the Cognitive Science Society*.

Amir Pouran Ben Veyseh, Thien Huu Nguyen, and Dejing Dou. 2019. Graph based neural networks for event factuality prediction using syntactic and semantic structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4393–4399, Florence, Italy. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Owen Rambow, and Mona Diab. 2010. Automatic committed belief tagging. In *Coling 2010: Posters*, pages 1014–1022, Beijing, China. Coling 2010 Organizing Committee.

E. F. Prince, J. Frader, and C. Bosk. 1982. On hedging in physician-physician discourse. In Robert Di Prieto, editor, *Linguistics and the Professions*, pages 83–97. Albex.

Anna Prokofieva and Julia Hirschberg. 2014. Hedging and speaker commitment. In *5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data*.

Yann Raphalen, Chloé Clavel, and Justine Cassell. 2022. "You might think about slightly revising the title": Identifying hedges in peer-tutoring interactions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2174, Dublin, Ireland. Association for Computational Linguistics.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 731–744, New Orleans, Louisiana. Association for Computational Linguistics.

Mark Sabbagh and Dare Baldwin. 2001. Learning words from knowledgeable versus ignorant speakers: Links between preschoolers' theory of mind and semantic development. *Child Development*, 72:1054 – 1070.

Roser Saurí and James Pustejovsky. 2009. Factbank: a corpus annotated with event factuality. *Language resources and evaluation*, 43:227–268.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. *Preprint*, arXiv:2311.09718.

Vicki L. Smith and Herbert H. Clark. 1993. On the course of answering questions. *Journal of Memory and Language*, 32(1):25–38.

Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. 2017. Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 352–357, Vancouver, Canada. Association for Computational Linguistics.

Morgan Ulinski, Seth Benjamin, and Julia Hirschberg. 2018. Using hedge detection to improve committed belief tagging. In *Proceedings of the Workshop on Computational Semantics beyond Events and Roles*, pages 1–5, New Orleans, Louisiana. Association for Computational Linguistics.

Morgan Ulinski and Julia Hirschberg. 2019. Crowdsourced hedge term disambiguation. In *LAW@ACL*.

Jiarui Yao, Haoling Qiu, Jin Zhao, Bonan Min, and Nianwen Xue. 2021. Factuality assessment as modal dependency parsing. 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 1540–1550, Online. Association for Computational Linguistics.

A Prompting Details

The exact prompt templates used for the BIO and listing experiments are shown below.

```
Given an utterance, perform BIO tagging to↩
     classify hedges in the sentence. ``↩
    BIO" tagging is a method used in ↩
    named entity recognition where each ↩
    token (word) in the sentence is ↩
    tagged as follows:

B (Beginning): The token is the beginning ↩
    of a hedge.
I (Inside): The token is inside, but not ↩
    the first token of a hedge.
O (Outside): The token is not part of a ↩
    hedge.
Please assign one of these tags to each ↩
    token in the given utterance, ↩
    representing whether each word is ↩
    part of a hedge phrase or not. Format↩
     your response by listing each token ↩
    followed by its corresponding BIO tag↩
    .

Example:

If the utterance is ``I think maybe you ↩
    could try an approach like that" then↩
     ``I think" and ``maybe" are ↩
    identified as hedges so your output ↩
    should look like this:

Utterance:
I think maybe you could try an approach ↩
    like that

Tags:
```

```
I/B think/I maybe/B you/O could/O try/O an↩
    /O approach/O like/O that/O

Now given the following input, please ↩
    classify the hedges in the sentence.

Utterance:
{utterance}
```

```
Given a conversation, answer a question. ↩
    Be as precise and succinct as ↩
    possible. If asked for a number, ↩
    provide a numeric value.

Format the output as follows:
Number of Hedges: Integer number of ↩
    linguistic hedges (e.g. 0)
List of Hedges: List of hedges found (e.g.↩
     [``first hedge", ``second hedge", ↩
    etc...])

Conversation:
{utterance} <stop sign emoji>

Question:
At the line that ends with <stop sign ↩
    emoji>, how many linguistic hedges ↩
    are there? List all the linguistic ↩
    hedges using quotations. Do not add ↩
    any additional information.
```

B Glossary

Due to the interdisciplinary nature of this work, we provide below brief definitions for terms which may be unfamiliar. The numbers refer to the pages in this paper in which the term first appears.

BERT BERT (Devlin et al., 2018) stands for Bidirectional Encoder Representations from Transformers. BERT is a transformer-based model which produces contextual representations of text by conditioning on both the left and right surrounding words. 4

BIO BIO, short for Beginning, Inside, Outside, is a format for labeling chunks of tokens. Tokens are assigned B if they begin a sequence which should be labeled (e.g., a named entity), I if they belong to a previously begun sequence, and O otherwise. 5

Cohen's Kappa Measure of agreement between two raters that an item falls within a subjective category; higher values denote higher agreement. 4

epoch A single pass through the training data. 5

F1 The harmonic mean of precision and recall.

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

It is also called F-measure or F-score. Loosely speaking, the metric is a balance of how often the model is correct when it predicts a particular class (precision), and how often the model predicts that class when it would be correct to do so (recall). 5

LLM Large Language Models are large (typically by parameter count) models which take in text and produce a distribution over their vocabulary which can be used to predict the next token. 1

LSTM Long Short-Term Memory networks (Hochreiter and Schmidhuber, 1997) are a type of recurrent neural network designed to capture long-range dependencies. 4

narrative element Observable events in the Roadrunner cartoon that and were likely to be mentioned in narrations (see Galati and Brennan, 2010). Segmentation by narrative elements allowed for comparisons across speakers for elements realized in each narration. 4

precision The number of correct predictions (true positives) for a class divided by the number of times the model predicted that class (true positives + false positives). 5, 12

recall The number of correct predictions (true positives) for a class divided by the number of samples which belong to that class (true positives + false negatives). 5, 12

temperature A hyperparameter that modifies the next token distribution of language models. Larger temperature values increase the likelihood of lower probability tokens. 5

token The smallest unit of text, often words or subwords, which are used as the input for various NLP models. 3