Toward the Automatic Detection of Word Meaning Negotiation Indicators in Conversation

Aina Garí Soler[1] **Matthieu Labeau**[2] **Chloé Clavel**[1]
[1]INRIA, Paris, [2]LTCI, Télécom-Paris, Institut Polytechnique de Paris, France
{aina.gari-soler,chloe.clavel}@inria.fr, matthieu.labeau@telecom-paris.fr

Abstract

Word Meaning Negotiations (WMN) are sequences in conversation where speakers collectively discuss and shape word meaning. These exchanges can provide insight into conversational dynamics and word-related misunderstandings, but they are hard to find in corpora. In order to facilitate data collection and speed up the WMN annotation process, we introduce the task of detecting WMN indicators – utterances where a speaker signals the need to clarify or challenge word meaning. We train a wide range of models and reveal the difficulty of the task. Our models have better precision than previous regular-expression based approaches and show some generalization abilities, but have moderate recall. However, this constitutes a promising first step toward an iterative process for obtaining more data.

1 Introduction

Miscommunication is an inherent part of linguistic interaction. Whether it stems from ambiguity, production errors, or differences in speaker knowledge, it provides a window into how language is processed and into the strategies used by speakers to resolve such situations (Clark, 1996; Bailey, 2004).

In this paper, we are interested in a specific type of miscommunication involving word meaning. These are cases where a speaker expresses the need to clarify the meaning of a word or phrase used previously in the conversation, engaging in a metalinguistic discussion of word meaning (see the example in Figure 1). These interactions, known as Word Meaning Negotiations (Myrendal, 2015, 2019) or WMN, offer a lens on word-related misunderstandings and misalignment and can complete and complement the study of repair mechanisms in dialog (Schegloff, 2007), from signaling to resolution strategies. They can also support the detection of unclear, ambiguous or controversial word usages (Garí Soler et al., 2025a); as well as contribute to the development of dialogue systems that can detect, avoid and resolve such lexico-semantic divergences as humans do.

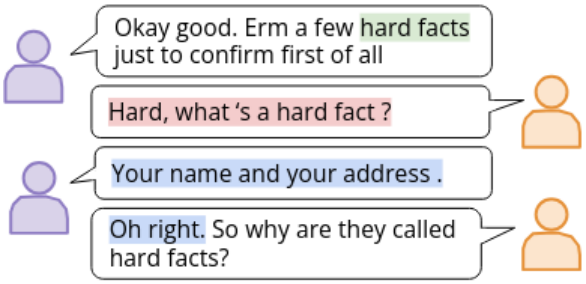

Figure 1: Example of a WMN sequence from the NeWMe corpus (Garí Soler et al., 2025b), consisting of a problematic word usage or *trigger* (green), an *indicator* (red) and a *negotiation* (blue).

Identifying and annotating WMN sequences in corpora, however, is challenging and costly in practice: existing annotated data is scarce, WMNs are hard to find, and regular expression-based methodologies used up to now (Garí Soler et al., 2025b) have low precision and require extensive manual filtering of irrelevant instances. Furthermore, the form of WMN interactions can differ substantially depending on the conversational modality (spoken or written), speaker relationship and familiarity as well as the communicative goal. Therefore, in order to study WMNs in their full complexity and diversity, it is crucial to collect a large number of WMN examples from a variety of conversational situations. This represents an additional challenge, as conversational corpora can differ substantially both in format and in register.

Our goal is to optimize the data collection process for WMN annotation, increasing the proportion of relevant retrieved examples (precision) without compromising their variety (recall). This can improve the efficiency of manual annotation, increasing the speed at which relevant instances are

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24580–24596
November 4-9, 2025

found and decreasing human effort.

To do so, we use the only existing dataset with WMN annotations to date to train and evaluate multiple models in a large number of settings on the task of identifying the part of a WMN that features the least variation: the *indicator* (in red in Figure 1). The indicator of a WMN is a (part of) an utterance that questions or challenges word meaning. Our evaluation and analysis highlight the difficulty of the task and the complementarity of our models with regular expression-based approaches. While our models have better precision and some incipient generalization abilities beyond previously considered patterns, we also identify and discuss several limitations.

Our main contributions are as follows: 1) we introduce the task of WMN indicator detection; 2) we benchmark supervised models and Large Language Models (LLMs) on this task for English; 3) we conduct a thorough analysis, comparing our best models to existing regular expression-based approaches and manually examining their predictions on new data; and 4) we identify key challenges and promising directions of improvement.

2 Background and Related Work

2.1 Word Meaning Negotiation

As depicted in Figure 1 and defined in Myrendal (2015), the typical WMN sequence has three parts: the *trigger* (a problematic word usage), the *indicator* and the *negotiation*, which consists of one or multiple turns where speakers address the issue, collaboratively shaping a shared understanding of the word's meaning. The trigger can be a word or a phrase, and the problem must involve semantic interpretation: referential unclarities or mishearing problemsdo not constitute a WMN. Myrendal (2015) identified two WMN types: those arising from non-understanding or misunderstanding (NONs) and those arising from disagreement about what a word should mean or how a word was used (DINs). See Table 6 in the Appendix for an example of a DIN.

WMN has been studied in different contexts, including non-native speaker conversations (Varonis and Gass, 1985) and online discussions (Myrendal, 2019; Noble et al., 2021). However, despite the extensive literature on the related phenomena of conversational repair (Drew; Bazzanella and Damiano, 1999; Mertens and De Ruiter, 2021; Dingemanse and Enfield, 2024; Ngo et al., 2024) and alignment (Brennan and Clark, 1996; Branigan et al., 2000), especially conceptual and lexico-semantic alignment (Schober, 2005; Garí Soler et al., 2023), there is not much work on this type of interaction. In this study, we rely on the only available dataset of WMN sequences, NeWMe (Garí Soler et al., 2025b), described in Section 3.

2.2 Utterance-level Classification

The NLP task most closely related to ours is Dialog Act Classification (DAC), which consists in assigning labels to utterances reflecting their communicative function (e.g., greeting, giving information, etc.) (Raheja and Tetreault, 2019). Our setting can be viewed as a simplified variant of DAC, as we treat indicator detection as a binary classification task focused on one specific dialogic function. Vielsted et al. (2022) train Transformer models for DAC on two social media datasets, including Reddit data, which is also one of the sources of NeWMe. They investigate the impact of lexical normalization and the inclusion of context on performance. While lexical normalization does not have a clear impact, including the text of the preceding utterance proved beneficial. Duran et al. (2023) run comparative experiments testing multiple types of sentence encodings and text pre-processing strategies for DAC. They find BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) to outperform dedicated supervised models. LLMs have also recently been found to obtain competitive results on DAC (Liu et al., 2025).

In work concurrent to ours, Ngo et al. (2025) address a closely related but more general task: the detection of other-initiated self-repair, where one speaker signals some trouble and prompts the other to resolve it. They use a multimodal model that combines hand-crafted linguistic and prosodic features.

Intent classification also bears some resemblance with our task (Zhang and Zhang, 2019; Farfan-Escobedo et al., 2021). In this case, the goal is to classify the intent behind a user's utterance, but it is typically framed within a restricted domain (e.g., "book a flight"), while we aim for a domain-agnostic model.

3 Task and Data

We propose to address the task of WMN indicator detection, which we frame as a binary classification problem where a model needs to learn whether the

input text is signaling the need to discuss or clarify a word's meaning. Indicators are the most easily identifiable part of a WMN, as triggers and negotiation exhibit large variation. The dataset used in our experiments, the Indicators dataset, consists of utterances extracted from NeWMe.

3.1 The NeWMe Corpus

The NeWMe (**Ne**gotiating **W**ord **Me**aning) corpus (Garí Soler et al., 2025b) contains annotations on conversations from three different English corpora, encompassing both online and oral conversations: the Switchboard Dialog Act Corpus (SWDA) (Stolcke et al., 2000),[1] with dyadic phone conversations; the British National Corpus (BNC) (Consortium, 2007), consisting of a mix of conversations, interviews and lectures, among others; and the Winning Arguments Corpus (Reddit) (Tan et al., 2016), with Reddit data from the r/ChangeMyView subreddit.

Data collection To collect sequences that might contain a WMN, the authors used regular expressions (regexes) to identify utterances potentially containing a WMN indicator. These regexes covered a set of expressions that may be used in indicators, such as "What do you mean by ...?" or "This is not...". Additionally, they considered utterances repeating words said in the previous turn with an additional question mark (e.g., "Hard facts?"), which is referred to as the *repetition pattern*. We refer to these automatically selected utterances as *potential indicators*.

Annotations and phenomena NeWMe was annotated by two expert annotators based on a set of guidelines. Agreement was calculated on a portion of the dataset and is estimated to be between 86% and 96%. Every potential indicator comes annotated with the type of conversational phenomenon it signals (WMN or other) and, if applicable, text spans corresponding to the trigger, indicator and negotiation were annotated in the surrounding conversation. Table 6 in the Appendix contains examples of every type of phenomenon in NeWMe. 404 WMNs were identified, along with related phenomena: 203 NON-PURSUED sequences (failed attempts at initiating a WMN where no one pursued the negotiation); 42 cases of "Self-Initiated Meaning Negotiation" (SIMN), where the trigger and the indicator come from the same speaker;[2] and 12 instances "Without Trigger" (WT), where the problematic word does not come from the conversation itself, but from, for example, the surrounding situational context. While these are not complete WMNs, they contain an indicator that fits our task definition.

The annotation also includes two types of *distractors*, i.e., phenomena that look like WMN but do not meet its definitional requirements: 28 Reference/Named Entity (R/NE) cases where the problem stems from an unclear referent and not the meaning of a word,[3] and 173 "Other kinds of clarification requests" (OKOCR), where there is a request for semantic clarification but it is targeted at a broader level (e.g., a whole utterance), rather than at a word or phrase. All these phenomena are annotated with an indicator span. Uninteresting instances were labeled as NOTHING.[4]

3.2 The Indicators Dataset

Here we describe our criteria to distinguish between positive and negative instances and how we split the data for the Indicators dataset.

3.2.1 Positive and negative instances

We need examples of utterances that contain real indicators as well as negative instances where no indication of misunderstanding or disagreement about word meaning is present.

We consider 661 instances to be **positive**: all available WMNs as well as related phenomena (NON-PURSUED, SIMN and WT).

As **negative instances**, we use all distractors, i.e., R/NE and OKOCR. These are particularly tricky, as they resemble a WMN indicator, but the clarification request is not directed at the lexical meaning of a word. We also include instances annotated as NOTHING, which are generally easier to distinguish from positive instances, as they usually have nothing to do with word meaning. Since no indicator span was marked for NOTHING instances, we take the utterance that was marked as a potential indicator by the regular expression search in the corpus creation process. The dataset is heavily imbalanced: we use a total of 7,118 negative instances.

[1] All utterances in SWDA are simplified to remove symbols related to spoken phenomena.

[2] Simplified example: **A**: Have you heard of half life? **B**: Yeah. **A**: What does that mean?

[3] Simplified example: **A**: They don't have that here in Texas. **B**: You mean they don't have the smog alerts?

[4] For example, "Oh at the back?" or "Why eleven?".

		Train	Dev	Test
Positive	SWDA	36	2	5
	BNC	153	10	29
	Reddit	340	22	64
	Total	529	34	98
Negative	SWDA	443	30	783
	BNC	1,272	73	2,287
	Reddit	934	54	1,242
	Total	2,649	157	4,312
	Distractors	165	13	24
	NOTHINGs	2,484	144	4,288
Total		3,178	191	4,410

Table 1: Composition of the Indicators dataset.

3.2.2 Splits and dataset composition

We split the data into training, development and test sets while ensuring a similar distribution of corpora across subsets (see Table 1). Due to the class imbalance, we opt for an imbalanced split, with the training and development sets containing a higher proportion of positive instances.[5]

First, we determined the size and composition of the test set as well as the split of positive and non-NOTHING instances into each subset. We experiment with two sampling strategies for selecting NOTHING negative instances for the training and development sets: random selection (RD) and regex-aware selection (RX). We identified the three most productive regular expressions, i.e., the most common regular expressions found in the positive instances of the training set.[6] In RX, we favor the sampling of NOTHING instances containing these regexes. This results in 24% of "Nothing" instances in the RX training set containing one of these expressions, compared to 12% in the RD training set. Both variants, RX and RD, have the same class and corpus distribution, and vary only in their selection of NOTHING cases. The goal of RX selection is to force models to rely less on correlations between surface form and class; and instead focus more on utterance meaning.

[5]This choice was motivated by preliminary experiments showing that a smaller but balanced training set led to catastrophic performance. We then reallocated positive instances from the test set to the training and development sets.

[6]These correspond to (variants of) the expressions "you mean," "can you define," and "definition of." We restricted the selection to the first 3 because the 4th one was "what is," which is a very generic regular expression observed in many negative instances too.

4 Experimental Setup

In this section we describe how input is presented to the models (§4.1); how we train supervised models (§4.2) and prompt LLMs (§4.3) for the task; and we present a regex-based baseline (§4.4).

4.1 Input Configurations

We experiment with different amounts of context given to the models: the 3 previous utterances (`3previous`) as well as one previous and one posterior utterance (`1pre-1post`).[7] We concatenate each turn, preceding them with "Speaker X: ", where X is a number starting with 1 and assigned based on speaker ID. Although Vielsted et al. (2022) found that the inclusion of context helped for DAC on social media, utterances coming from the BNC and Reddit are sometimes long and convoluted. Therefore, we also experiment with inputting only the target utterance (`utterance`) or only a sentence selected from the utterance (`sentence`). We base the choice of a sentence on the annotated indicator span or, for NOTHINGs, the potential indicator found with regex search. We use `stanza` (Qi et al., 2020) for segmentation into sentences. If the indicator span contains more than one sentence, we select the longest one.

4.2 Supervised Models

Following the findings of Duran et al. (2023) (Section 2), we use RoBERTa (Liu et al., 2019) base and large. Due to the conversational nature of our data, we also include DialogLED base and large (Zhong et al., 2022). DialogLED is an encoder-decoder model for dialogue understanding, pretrained on interview data and movie subtitles. We use its encoder part. We use the `transformers` library (Wolf et al., 2020).

Training setup We experiment with a regular **fine-tuning** of all weights of the model as well as two more training strategies:

- **Contrastive learning** is known to be helpful in 1-class or anomaly detection problems (Winkens et al., 2020). We first train the models in a contrastive setting to learn a better representation of instances of each class using a triplet loss.[8] Then, we fine-tune the resulting

[7]We also tried including only the previous utterance, but we have omitted these results from the paper for brevity as their behaviour was similar to the other two settings.

[8]We use the `sentence-transformers` library (Reimers and Gurevych, 2019).

models on the classification task.

- **Domain adaptation**. We follow Lemmens and Daelemans (2023), who run continued pretraining to improve models' performance on social media data. However, to prevent overfitting, we do it on utterance pairs from conversations present in the three source corpora but not in NeWMe. We select a total of 7,380 utterance pairs, balancing the representation of every corpus. We train RoBERTa and DialogLED with a Masked Language Modeling objective on this data and then fine-tune the models for classification.

More details on implementation and hyperparameters are found in Appendix B.

Training on separate corpora The heterogeneous nature of our dataset adds complexity to an already challenging scarce data scenario. To explore corpus-specific learning, we train separate models on BNC and Reddit data. We do not do this for SWDA due to the limited data available. Given to the large amount of experiments and the size of the models, these experiments are only run in a fine-tuning setting. For Reddit data, we also experiment with BERTweet (Nguyen et al., 2020), as it was found to give good results for DAC on Reddit data (Vielsted et al., 2022). Given its 128 token limit, we only try it with the `sentence` input configuration.

4.3 Large Language Models

We test three open-weights generative LLMs: OLMo-2-7B-Instruct (Groeneveld et al., 2024), Llama-3.2-3B-Instruct and Llama-3.3-70B-Instruct; the latter with 4-bit quantization. We try 18 different few-shot prompts and one zero-shot prompt and choose the best one for each model on the development set. Prompts vary with regard to the input format of each instance (Section 4.1) and whether or not they include an explanation of the examples' labels, as well as a hint about the low frequency of the phenomenon. More details and examples of prompts can be found in Appendix D.

4.4 Regex Baseline

We compare all models to a baseline that takes into account the success rate of every regex in the training set, based on the original set of 30 regexes used in NeWMe. Due to the class imbalance in the data, we adopt a sampling approach for its training: We randomly sample k instances of each class and identify the majority class of each regular expression in the sample. This process is repeated for t iterations. A regex is considered as indicative of a positive instance if, after this process, it is more often associated with the positive class than the negative class. To assign a class to an instance, we consider all regexes that matched it. If there was more than one, a simple majority vote is made between positive vs negative regexes. In case of a tie, a default class d is assigned. The best values of $k = \{100, 200, 300\}$, $t = \{3, 5, 10\}$ and d are determined on the development set.

5 Results and Analysis

In this section we present the results obtained by all models on the Indicators dataset (Section 5.1) and compare our method with the original collection method based on regexes (Section 5.2).

5.1 General Results

We evaluate models based on their F1 scores. In Table 2, we report the F1 scores obtained by all models that were trained on the full training sets. LLM and baseline results are presented in Table 3. For reference, performances obtained on the development set are provided in Appendix C.

Global results The first notable observation from Table 2 is the overall weak performance of all models, with a highest F1 score of only 0.49, achieved by RoBERTa-large. Most configurations yield results just over 0.30, and in a few cases, models fail to learn the task entirely (F1=0.00), predicting only the negative class.[9] These results highlight the difficulty of the task and suggest that more data is needed in order to learn it properly. LLM performance is particularly poor, with a highest F1 score of 0.12 obtained by OLMo, barely matching the regex baseline. An examination of the LLMs' predictions did not reveal any clear error patterns. OLMo's main weakness stems from poor precision (0.07), but recall is not very high either (0.34). This suggests that OLMo strongly over-predicts the positive class, despite mentioning the rarity of the phenomenon in the prompt. While larger models or adjusting the prompt might improve performance,

[9] All results reported in the paper were obtained with the same seed, but we explored two additional seeds for models with an F1 score of 0. Only two out of the 14 settings yielded at least one non-zero F1, suggesting that this is not simply a problem of an unfortunate initialization.

		RX			RD		
Model	**Input type**	**Fine-tuning**	**Contrastive**	**Dom.-adapted**	**Fine-tuning**	**Contrastive**	**Dom.-adapted**
RoBERTa-b	3previous	.28	.28	.28	.23	.29	.28
	1pre-1post	.31	.26	.24	.23	.00	.29
	utterance	.31	.33	.31	.30	.33	.31
	sentence	.35	.37	**.38**	**.41**	.36	.40
RoBERTa-l	3previous	.30	.00	.33	.30	.30	.30
	1pre-1post	.29	.00	.33	.30	.00	.29
	utterance	.30	.33	.32	.33	.00	**.37**
	sentence	.39	**.49**	.41	.36	.00	.35
DialogLED-b	3previous	.29	.28	.28	.33	.27	.28
	1pre-1post	.34	.24	.25	.27	.25	.27
	utterance	.31	.29	.29	.33	.27	.28
	sentence	**.48**	.37	.32	.33	**.39**	.32
DialogLED-l	3previous	.36	.31	.00	.30	.35	.00
	1pre-1post	.34	.28	.00	.34	.23	.00
	utterance	.33	.28	.00	.37	.32	.00
	sentence	**.39**	.37	.00	.33	**.39**	.00

Table 2: F1 scores of supervised models on the full test set. The best results obtained by each model architecture and on each training set are boldfaced.

Model	RX	RD
Regex baseline	.11	.12
OLMo-2-7B	.12	
Llama-3.2-3B	.06	
Llama-3.3-70B	.10	

Table 3: F1 scores by the regex baseline and the LLMs.

such low scores confirm the difficulty of the task for LLMs in a few-shot setting.

Effect of input configuration We observe a clear preference for the sentence input type across the board. This contrasts with previous findings that context is beneficial for DAC (Vielsted et al., 2022). In our case, instead, providing conversational context appears to confuse the models rather than help them. Except for DialogLED-large, which is pretrained on longer dialog sequences, inputting only the target utterance tends to be the second-best strategy. While a sentence often suffices to determine whether an utterance is an indicator, the absence of context can sometimes introduce ambiguity.[10] Our current models are not capable of effectively leveraging the broader context required to solve such cases. Their generally low performance suggests that there are more fundamental issues to address, but this kind of ambiguity may eventually constitute a performance ceiling for models relying solely on information from the target utterance.

[10]Consider the utterance "What does this mean?". It would likely be a positive instance if following "I have osteoarthritis", or a negative one if following "The statue has its eyes covered."

Effect of other parameters We do not find a consistent pattern with respect to the type of training setup: no approach is systematically better than the others. The best results are obtained with the contrastive setting, but fine-tuning is the only strategy that did not lead to learning failure. Regarding the two training sets, there is no clear preference between RX and RD, but the two best-performing models (F1=0.49 and 0.48) were trained on RX and are far ahead from the the third-best result (0.41), which was trained on RD. Interestingly, even the regex baseline shows only a minor difference between the two, suggesting that they may not have been distinct enough to impact model behavior.

Results on individual corpora We also examine results obtained on BNC and Reddit data individually. First, we compare the scores obtained on the two corpora by the best configuration of every model when trained on the full dataset. Full results can be found in Table 9 in the Appendix. We observe large differences between the two corpora, with Reddit consistently yielding higher scores and a performance gap ranging from 16 to 28 F1 points (e.g., RoBERTa large obtains F1=0.38 on the BNC and 0.54 on Reddit). This can partly be expected, given the higher number of positive instances coming from Reddit (cf. Table 1). It could also be related to the oral nature of the BNC, which contains ungrammatical or interrupted utterances, compared to the more standard, written form of the debates coming from Reddit. OLMo displays the same pattern, with a better performance on Reddit (F1 =

Test corpus	Training corpus	Model	Input type	Training strategy	Regex	F1
BNC	BNC	RoBERTa-base	sentence	fine-tuning	RX	.37
		RoBERTa-large	sentence		RX	.27
		DialogLED-base	utterance		RX	.29
		DialogLED-large	utterance		RD	.28
	all	RoBERTa-large	sentence	contrastive	RX	**.38**
Reddit	Reddit	RoBERTa-base	sentence	fine-tuning	RX	.58
		RoBERTa-large	sentence		RX	.52
		DialogLED-base	sentence		RD	.51
		DialogLED-large	sentence		RX	**.62**
		BERTweet	sentence		RD	.55
	all	DialogLED-base	sentence	contrastive	RD	.56

Table 4: Results of models trained on individual corpora, compared to the best per-corpus performance among models trained on all data.

0.25) compared to the BNC (F1 = 0.09).

In Table 4, we present results of models trained exclusively on data from one corpus. For comparison, we also include the best result on each corpus obtained by models trained on all corpora. Interestingly, the best performance on the BNC comes from a model trained on all data, while Reddit benefits from having a dedicated model. In the remainder of the paper, all results relating to an individual corpus are obtained with the best model for that corpus.

Results by type of phenomenon We break down performance of the best overall model by phenomenon type, focusing on phenomena with more than 20 instances in the test set. Among negative items, as expected, accuracy is much lower on distractors (58.3%) than on clear-cut NOTHINGs (99%). Among complete WMNs, disagreement (DIN) indicators are detected more reliably (79%) than those of NONs (41%). This discrepancy may stem from the fact that DINs are almost exclusively found in Reddit data (23 out of 24 in our test set), where models perform better. This is confirmed by a further breakdown of NONs: accuracy is lower on BNC NONs (33%, 21 instances) than on Reddit ones (64%, 14 instances). NON-PURSUED indicators, all from Reddit in our test set, are also detected with relative ease (63%). Full results are found in Table 10 in the Appendix.

5.2 Comparison to Original Data Collection

To put results into perspective, we compare the **precision** of our model to the original regex-based data collection method used for NeWMe. As shown in Table 1, the original methodology exhibits very low precision: 0.03, 0.05 and 0.16 on SWDA, the BNC and Reddit, meaning that only an average of 8% of retrieved instances contained actual indicators.[11] In contrast, our models have a precision of 0.50 (BNC), 0.55 (Reddit), and 0.44 overall. These figures are not directly comparable, since regex precision is measured on the entire dataset, whereas model precision is calculated on the test set.[12] They can still, however, offer meaningful insight. We run a chi-square test comparing the proportions of true and false positives between the two strategies, which shows a significant difference ($\alpha = 0.0001$). This confirms that our model retrieves a higher proportion of relevant instances than the regex-based strategy.

Regex	Regex (full dataset)	Model (test set)
What is	6.4%	9.1% (11)
this is not x	4.2%	20.0% (15)*
this isn't x	5.2%	66.7% (9)**
is that a/the x	4.3%	45.5% (11)**
repetition pattern	3.0%	0.0% (3)

Table 5: Precision of challenging regexes in the Indicators dataset and of our models on test instances containing them. Numbers in parentheses correspond to the number of predicted-positives by the model.* and ** indicate significance with $\alpha = 0.05$ and $\alpha = 0.001$.

In the creation of NeWMe, Garí Soler et al. (2025b) identified a few patterns which, despite capturing an important number of relevant instances, had a low overall precision, retrieving also many irrelevant ones that needed to be manually discarded. We therefore focus on these challeng-

[11] Our definition of precision differs from that of Garí Soler et al. (2025b), who considered distractors to be positive matches.

[12] It would be unfair to calculate the regexes' precision on the test set, as we artificially determined its class distribution.

ing patterns, where there is greater room for improvement. Table 5 provides comparative precision values. Chi-square tests reveal a significant precision gain for three of these five patterns. We run the same analysis with four regexes that were identified as being highly productive. Here too, we observe a significant improvement in precision for three of them. Full results, as well as details on the calculation of chi-square tests, are found in Table 11 and Section F in the Appendix.

To assess how many relevant instances our model is able to capture, it is also important to consider **recall**. Unfortunately, no direct comparison with the regex-based method is possible, as the true prevalence of indicators in the original corpora is unknown. However, our best model achieves a recall of 0.55 (0.31 on BNC, and 0.70 on Reddit), meaning that, on average, the model misses about half of the relevant instances captured by the regexes.

These results show that while our model offers a significant gain in precision, holding promise to optimize and speed up data collection, it does so at the cost of missing many interesting cases. This suggests that a hybrid strategy, combining the strengths of the model and the regexes, may be a more effective approach.

6 Model Behavior on New Data

We now explore our models' ability to generalize beyond the data retrieved by regexes. Specifically, we ask whether they can correctly identify and uncover indicators that are expressed in ways that were not previously considered by the previous regex-based methodology.[13] We use the best-performing model for each corpus.[14]

6.1 Data for Error Analysis

To assess the models' generalization ability, we focus on fully unseen data: conversations that were considered in the NeWMe corpus creation but are not present in it because they contained no regex matches. This ensures a clean separation from the Indicators dataset. For SWDA, this consists of 445 conversations with a total of 41,213 utterances. For the BNC, we have 123 conversations and 21,114 sentences; and for Reddit, 43 threads and 5,655 sentences.

We use our models to make predictions on this new data and we select 200 predictions per corpus for manual evaluation. Since the number of predicted positives is quite low on the SWDA and BNC data (32 instances each),[15] we include them all in the sample. For Reddit, we randomly pick 100 out of 819. While a random selection of negative instances would provide a fair assessment of model performance, it would also likely include mainly clear-cut NOTHINGs, offering little insight into the model's limitations. To allow for a richer error analysis, and a more challenging evaluation, we pick the lowest-confidence negative predictions to make up half of the sampled negative cases, and the rest are randomly selected. Confidence is quantified as the logits' negative entropy. One of the authors annotated the selected instances. They had access to conversational context, but could not see the model's predictions. Note that this does not constitute a complete WMN annotation like the one provided in NeWMe, with spans and phenomenon types, and concerns only the presence of an indicator. As such, it is faster and can be seen as a filtering preceding full-fledged annotation. However, the distinction between distractors and NOTHINGs is lost, as they are both labelled as negative instances.

6.2 Results and Analysis

The evaluation of model prediction against manual annotations reveals a small number of true positives: 3 in the SWDA sample, 1 in BNC, and 4 in Reddit, yielding precisions of 0.03, 0.00 and 0.03; and recalls of 0.33, 0.00 and 0.75, respectively. Despite these negative results on this new and challenging sample, our qualitative analysis yields interesting insights:

Model intuitions A closer look at predicted-positives and low-confidence negatives shows the models exhibit sensible intuitions. We find utterances expressing lack of knowledge (e.g., "I'm not familiar with that"), making word-centered re-

[13] Note, however, that not all indicators in NeWMe contain an expression matched by a regex. This can happen, for instance, when an utterance matching a regex was judged by the annotator as being part of a WMN's negotiation and not its indicator. See Appendix G for an example. While these cases are a minority (7% of positive instances in the training data), they do introduce some degree of variation. Thus, the data is not strictly limited to the initial set of 30 expressions and contains a few other phrasings found during annotation.

[14] For SWDA, this is the RoBERTa-large model trained only on RX BNC data with `1pre-1post` context.

[15] This could be expected, as the models were trained on a small and expression-specific dataset, and defaulting to negative predictions is a safer strategy when encountering an unfamiliar pattern. WMN also seem to be much more frequent in Reddit.

quests for more details ("Oh, fajitas, how do you make fajitas?"), discussing word choices and meaning; and they also include distractors ("I don't know who Mitch Schneider is") and cases that require context ("I think you might be confusing cause and effect."). We provide more such examples in Appendix H. This shows that the models have started acquiring relevant knowledge and patterns, as they are confused by tricky and semantically interesting utterances.

Uncovering new expressions Some true positives reflect previously unconsidered expressions that may be productive patterns to be used in future studies: e.g., "I've never heard of that one" or "To you, hope means inaction." At the same time, we also find more idiosyncratic cases, also among the challenging examples presented above and in Appendix H, showing the potential of models to go beyond patterns and focus on the sentence's meaning.

Short sentences We observe that short sentences are disproportionately predicted as positives or low-confidence negatives. For example, their average sentence length in the Reddit sample is 7.9±5.2 and 9.1±3.9 words, respectively; while for random negatives it is 23±9.4. This could be due to the fact that the sentences used in the Indicators dataset tend to be long (30.2±32.9; 33.4±18.9 for Reddit); and short sentences provide less context, being therefore more ambiguous and error-prone. Several of them are clearly irrelevant (e.g., "Thanks!", "Uh-huh"), but others resemble instances corresponding to the repetition pattern, which is very context-dependent. This outlines a clear direction of improvement, perhaps through the separate treatment of short utterances or certain patterns, or providing the models more explicit guidance on these sentences through active learning.

7 Conclusion and Future Work

We have introduced the task of Word Meaning Negotiation Indicator Detection and presented a first exploration using supervised models and few-shot prompting with LLMs. Our results show that this is a challenging problem which current open-source LLMs do not manage to solve, but for which supervised models show promise. Our best models achieve a higher precision than the previous regex-based approach, but their limited recall and our closer look at specific regexes suggest that a hybrid method combining the strengths of both approaches could offer a better compromise between efficiency and coverage. Our manual evaluation on challenging samples highlights that while the models have learned useful patterns and have some generalization abilities (i.e., they are able to correctly detect a few new indicators that were missed by the previous regex methodology), more data is needed; with special attention on the treatment of short sentences. This makes the task well-suited for an active learning approach (Cohn et al., 1994), where the most informative instances would be used to help the model better learn the task with as little additional annotation as possible.

More work and data is also needed to better incorporate conversational context to the models, and to explore their domain adaptation capability across corpora and dialogue settings, which can be challenging given the observed difference in results between corpora. In the longer term, future studies could explore jointly addressing indicator and trigger detection in a multi-task setting, as well as collecting more data and annotations from other modalities that also express communication issues (e.g., speech and gestures) to train multi-modal models.

This study lays a foundation toward the computational modeling and semi-automatic annotation of this interesting linguistic phenomenon. Improving its detection can support both discourse and linguistic analysis as well as conversational system development. We share our code to promote further research.[16]

Limitations

The main limitation of our work is the size of the dataset used, which constrains model performance as well as generalization. This is why we decided to start with a simplified, binary setting. Fine-grained classification of WMN types is a desirable long-term goal, but our findings give little promise for good results in a more complex task.

Indicator detection is also a task with a certain degree of subjectivity, which also contributes to the difficulty in achieving higher performance and highlights the need for more data. However, WMNs are expensive to annotate, and our goal is precisely to develop more efficient strategies of obtaining data that can, in turn, be used to itera-

[16] https://github.com/ainagari/WMNindicator_detection

tively improve model performance. In the paper, we make a first step in this direction.

Another limitation is that our study focuses exclusively on the English language. This is driven by data availability, but WMN and WMN indicators are not specific to English and may exhibit differences both cross-linguistically and cross-culturally; which should be explored in future work.

Finally, as discussed in the paper, our best-performing models were surprisingly those operating at the sentence level. This makes them unable to properly handle ambiguous, context-dependent indicators, the proportion of which is unknown. We hope that as we obtain more data, models will learn to properly use context to solve the task.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This research was partially funded by the Agence Nationale de la Recherche SINNet project (ANR-23-CE23-0033-01). Additional support was provided by the ANR under the France 2030 program PRAIRIE (ANR-23-IACL-0008).

References

Benjamin Bailey. 2004. Misunderstanding. *A companion to linguistic anthropology*, pages 395–413.

Carla Bazzanella and Rossana Damiano. 1999. The interactional handling of misunderstanding in everyday conversations. *Journal of Pragmatics*, 31(6):817–836.

Holly P Branigan, Martin J Pickering, and Alexandra A Cleland. 2000. Syntactic co-ordination in dialogue. *Cognition*, 75(2):B13–B25.

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.

Herbert H Clark. 1996. *Using language*. Cambridge University Press.

David Cohn, Les Atlas, and Richard Ladner. 1994. Improving generalization with active learning. *Machine learning*, 15:201–221.

BNC Consortium. 2007. The British National Corpus, XML Edition. *Oxford Text Archive*.

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.

Mark Dingemanse and Nick J Enfield. 2024. Interactive repair and the foundations of language. *Trends in Cognitive Sciences*, 28(1):30–42.

Paul Drew. 'Open'class repair initiators in response to sequential sources of troubles in conversation. *Journal of pragmatics*, 28(1).

Nathan Duran, Steve Battle, and Jim Smith. 2023. Sentence encoding for Dialogue Act classification. *Natural Language Engineering*, 29(3):794–823.

Jeanfranco D Farfan-Escobedo, Kelly Lopes, and Julio C Dos Reis. 2021. Active learning approach for intent classification in portuguese language conversations. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 227–232. IEEE.

Aina Garí Soler, Matthieu Labeau, and Chloé Clavel. 2023. Measuring Lexico-Semantic Alignment in Debates with Contextualized Word Representations. In *Proceedings of the First Workshop on Social Influence in Conversations (SICon 2023)*, pages 50–63, Toronto, Canada. Association for Computational Linguistics.

Aina Garí Soler, Matthieu Labeau, and Chloé Clavel. 2025a. Potentially Problematic Word Usages and How to Detect Them: A Survey. *Accepted at the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*.

Aina Garí Soler, Jenny Myrendal, Chloé Clavel, and Staffan Larsson. 2025b. The NeWMe Corpus: A gold standard corpus for the study of Word Meaning Negotiation. *PREPRINT (Version 1) available at Research Square*.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the Science of Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Jens Lemmens and Walter Daelemans. 2023. Combining Active Learning and Task Adaptation with

BERT for Cost-Effective Annotation of Social Media Datasets. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 237–250, Toronto, Canada. Association for Computational Linguistics.

Xuanqing Liu, Luyang Kong, Wei Niu, Afshin Khashei, Belinda Zeng, Steve Johnson, Jon Jay, Davor Golac, and Matt Pope. 2025. Learning LLM Preference over Intra-Dialogue Pairs: A Framework for Utterance-level Understandings. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 86–98, Albuquerque, New Mexico. Association for Computational Linguistics.

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*.

Julia Beret Mertens and Jan P De Ruiter. 2021. Cognitive and social delays in the initiation of conversational repair. *Dialogue & Discourse*, 12(1):21–44.

Jenny Myrendal. 2015. *Word meaning negotiation in online discussion forum communication*. Ph.D. thesis, University of Gothenburg.

Jenny Myrendal. 2019. Negotiating meanings online: Disagreements about word meaning in discussion forum communication. *Discourse studies*, 21(3):317–339.

Ang Ngo, Nicolas Rollet, Catherine Pelachaud, and Chloé Clavel. 2025. "Mm, Wat?" Detecting Other-initiated Repair Requests in Dialogue. *Accepted at the 2025 Conference on Empirical Methods in Natural Language Processing*.

Anh Ngo, Dirk Heylen, Nicolas Rollet, Catherine Pelachaud, and Chloé Clavel. 2024. Exploration of Human Repair Initiation in Task-oriented Dialogue: A Linguistic Feature-based Approach. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 603–609.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Bill Noble, Kate Viloria, Staffan Larsson, and Asad Sayeed. 2021. What do you mean by negotiation? annotating social media discussions about word meaning. In *Proceedings of the Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2021)*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Vipul Raheja and Joel Tetreault. 2019. Dialogue Act Classification with Context-Aware Self-Attention. 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 3727–3733, Minneapolis, Minnesota. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Emanuel A Schegloff. 2007. *Sequence organization in interaction: A primer in conversation analysis I*, volume 1. Cambridge University Press.

Michael F Schober. 2005. Conceptual Alignment in Conversation. *Other minds: How humans bridge the divide between self and others*, pages 239–252.

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):339–373.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

Evangeline Marlos Varonis and Susan Gass. 1985. Non-native/non-native conversations: A model for negotiation of meaning. *Applied linguistics*, 6(1):71–90.

Marcus Vielsted, Nikolaj Wallenius, and Rob van der Goot. 2022. Increasing Robustness for Cross-domain Dialogue Act Classification on Social Media Data. In *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 180–193, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. 2020. Contrastive Training for Improved Out-of-Distribution Detection. *arXiv preprint arXiv:2007.05566*.

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.

Leihan Zhang and Le Zhang. 2019. An ensemble deep active learning method for intent classification. In *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence*, pages 107–111.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. DialogLM: Pre-trained Model for Long Dialogue Understanding and Summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

A Examples of Phenomena in NeWMe

For convenience, we present an example of every kind of phenomenon in the NeWMe corpus in Table 6. Examples are extracted from the original NeWMe paper (Garí Soler et al., 2025b).

B Implementation Details

This section gives more details on the training and hyperparameters used for our supervised models described in Section 4.2. The contrastive learning setting is run with 5 epochs with a learning rate of 2×10^{-5}. Domain adaptation is run for 2 epochs with a learning rate of 1×10^{-5}. Then, all fine-tuning operations run for 5 epochs with a learning rate of 1×10^{-5}. The training batch size was 8 throughout all experiments. The best epoch is selected based on results on the development set. The development set reflects the training data: for models trained on data from one corpus, the development set is restricted to that corpus. All results reported in the paper were obtained with the same seed, to keep the experimental budget manageable.

C Results on the Development Set

Tables 7 and 8 provide results obtained on the development set. F1 scores are higher than on the test set, which is probably due to the more balanced class distribution of the dev set.

D LLM Prompts

This section provides additional information on the prompts used for the LLMs (presented in Section 4.3). An example of a prompt can be found in Figure 2. All prompts contain an introduction to the task and instruct the model to reply with "Yes" or "No." We tested a total of 18 few-shot prompts and one zero-shot prompt. As shown in the Figure, prompts differ with respect to their input configuration (Section 4.1), the mention of the higher frequency of the negative class, the set of examples shown to the model, and whether an explanation for the answer to each example is provided. We first ran all configurations with a set of examples A, and used the winning configuration (determined on the development set) to build one prompt using another set B and the zero-shot prompt. Both A and B consisted of three examples: a NON, a DIN and a distractor (see Figure 3). The best prompts for all models were few-shot, relied on `sentence` input and mentioned the low frequency of the phenomenon. For OLMo-2-7B-Instruct and LLama-3.3-70B-Instruct the examples were accompanied by an explanation, while in Llama-3.2-3B-Instruct they were not.

E Complete Results Tables

Tables 9, 10 and 11 contain the complete results mentioned in the text of Section 5. Note that in Table 10 we report accuracy instead of F1 scores because the type of phenomenon determines its class, rendering F1 scores on distractors and NOTHINGs uninformative.

F Chi-square Tests

This section clarifies the calculation of the chi-square tests of independence run as part of Section 5.2. For each individual regex, we build a 2×2 contingency table with, on the one side, the true positives and false positives captured by the regex in the full Indicators dataset; and the true positives and false positives of our model on test instances involving that regular expression. The test evaluates whether the observed proportions of correct and incorrect predictions differ significantly between the two approaches.

G NeWMe Indicators not Matched by Regular Expressions

As explained in Section 6, a 7% of indicators in NeWMe did not actually match a regular expression. This is because annotators could choose to annotate indicator spans in the vicinity of a regular expression match if they believed there was an indicator a few turns before or after it. For example,

Phenomenon	Corpus	Sequence
NON	SWDA	**S1**: But the problem is, that I am very liberal politically and so I hardly ever have anybody that wins that I vote for **S2**: Oh, liberal, by, what do you mean by liberal, um **S1**: Liberal politically, I'm, you know, like pretty left wing Democrat, so **S2**: Well see, I don't know anything about politics. **S1**: Oh you don't ? **S2**: Uh, what's the main difference between Republicans and Democrats? (...) **S1**: (...) Democrats usually are more supportive of public assistance programs (...) the big Republican thing is that they don't, they vote for less government, they want less government involvement in society (...)
DIN	Reddit	**S1**: True waffles are crisp on the outside and fluffy on the inside, and at least double the height of a pancake. **S2**: Let us look at the Merriam-Webster waffle definition. 1waf·fle wä-fl, wo - noun :a crisp cake of batter baked in a waffle iron (...) Well which definition are we supposed to go by according to your view? The dictionary definition mentions nothing about waffles being fluffy on the inside or being at least double the height of a pancake, yet your definition mentions bother of these things. So, which definition of a waffle should we be going by? **S1**: True, the pancake height was my own personal addition for dramatic effect, but that doesn't take away from the fact that Waffle House waffles are not crisp, and therefore do not fit the definition. (...) **S2**: So should I go by the dictionary definition of a waffle or your personal definition? If crispness is the sole determining factor then it seems like Waffle House may simply undercook their waffles, not that they aren't waffles at all.
WMN: Other	BNC	**S1**: Mhm . Do you so you see yourself the fact that you 're living here on your own, in the in the flats , makes you feel more vulnerable . [UNCLEAR] feel more vulnerable [UNCLEAR] living in here on your own ? **S2**: You mean frightened ? **S1**: Yeah I think yeah.
NON-PURSUED	Reddit	**S1**: (...) I am telling you race is not a scientific description and its use as a popular culture description causes hate and ignorance. (...) **S2**:(...) Do you know what you mean by "unscientific", or is it just a non-declarative placeholder speech act you use when you have attitudes of aversion? (...) **S1**: Wow you are putting alot of effort into avoiding my question. I was careful to put it into the most basic terms and yet there was still a communication failure, no problem, lets break out the crayola's. Do you hold the position that race is a scientific descriptor? Your answer should be framed as a "yes", "no" or "unknown". Seriously, dont over think this one, just a one word answer. You know what, let me pull it out of this paragraph and restate, again, no tricks, just a simple one word answer is all that required. Also, if you are worried about tricks go ahead and ask any questions you need, look up any references for clarification, just dont let it interfere with your response. Do you hold the position that race is an accepted scientific descriptor? Your answer should be framed as a "yes", "no" or "unknown".
SIMN	BNC	**S1**: Not a good idea. Erm have you heard of half life ? **S2**: Yeah . **S1**: What does that mean ? **S2**: that, that's how long it takes for half of the radioactive isotopes to disappear. **S1**: Great, yeah, that 's a wonderful definition .
WT	BNC	**S1**: (...) what do you think about this new menu for the canteen at Digby's Ballbearings ? **S2**: Crunchy nut salad, t , what 's tortellini ? **S1**: Pasta, stuffed with spinach and cheese (..)
R/NE	SWDA	**S1**: they'd have their smog alerts where you'd have to stay indoors for so many hours with an air conditioner. And, of course, they don't have that here in Texas. So, there's... **S2**: You mean they don't have the, uh, the smog alerts?
OKOCR	Reddit	**S1**: For this to change my view I'd need strong evidence that presumably lower interest rates on lons from banks have lead to a substantial increase in quality of life as they have increased. **S2**: I'm not sure what you mean here. Lower interest rates on a loan mean that your monthly payments are lower, and it's easier to pay off / you have more money available for other things. That's a quality of life increase.

Table 6: Examples of the different phenomena present in NeWMe, taken from Garí Soler et al. (2025b). Highlighted passages correspond to indicators. NON: Non-understanding WMN; DIN: Disagreement WMN; SIMN: Self-Initiated Meaning Negotiation; WT: Without trigger; R/NE: Reference/Named Entity; OKOCR: Other kinds of clarification requests.

LLM Prompt

Introduction	You are tasked with determining whether {an utterance | a sentence} signals the need to discuss or clarify a word's meaning. This may take the form of a clarification request or a challenge to the appropriateness or meaning of a word or short phrase.
Context (optional)	To help you decide, you are provided with {at most three previous utterances as context, followed by the target utterance | the previous utterance as context, followed by the target utterance and the subsequent turn, if present}.
Instruction	Respond with "Yes" if the target utterance signals the need to discuss or clarify a word's meaning, or "No" if it does not.
Examples	Below are some examples with the expected answer {and an additional explanation}. [EXAMPLES]
Frequency (optional)	In most cases, the response should be "No". [INSTANCE FOR PREDICTION]
Prompting for answer	Your response (Yes or No):

Figure 2: Template of the prompts provided to the LLMs. Passages inside curly brackets reflect options depending on the type of prompt used.

Few-shot Examples

Set A

Target sentence: [END-CITE]No, what they are expressing is a sense of being unworthy and the recipient of a gift, a gift of talent, a gift of opportunity, a gift of trust from voters.
Expected response: Yes
Explanation: The speaker expresses a disagreement on the meaning of the word "humble".

Target sentence: But noise in what sense ? [UNCLEAR] what kind of noise are they talking about ?
Expected response: Yes
Explanation: The speaker asks for clarification about the word "noise".

Target sentence: I'm not trying to lampoon you, but that's what I'm getting from your posts and I figure you'd like to clarify that statement.
Expected response: No
Explanation: The speaker asks for a clarification, but it involves a broader semantic content rather than a specific word's meaning.

Set B

Target sentence: This isn't a functional or desirable definition of compassion.
Expected response: Yes
Explanation: The speaker expresses a disagreement on the meaning of the word "compassion".

Target sentence: Skidding control , you mean the antilock brake system ?
Expected response: Yes
Explanation: The speaker asks for clarification about the word "skidding control".

Target sentence: What do you mean despite them ?
Expected response: No
Explanation: The speaker asks for a clarification, but it involves a broader semantic content rather than a specific word's meaning.

Figure 3: Examples provided in the prompts to the LLMs (sentence version).

		RX			RD		
Model	**Input type**	**Fine-tuning**	**Contrastive**	**Dom.-adapted**	**Fine-tuning**	**Contrastive**	**Dom.-adapted**
RoBERTa-b	3previous	.68	.71	.67	.68	.72	.68
	1pre-1post	.73	.50	.24	.73	.00	.75
	utterance	.73	**.78**	.69	.76	.73	**.78**
	sentence	.69	.69	.67	.78	.75	.75
RoBERTa-l	3previous	.74	.00	**.78**	.74	.81	.82
	1pre-1post	.72	.00	.74	.66	.00	.78
	utterance	.73	**.78**	.74	.75	.00	.79
	sentence	.70	.70	.71	.77	.00	**.82**
DialogLED-b	3previous	.63	.65	.64	.64	.70	.72
	1pre-1post	.57	.54	.56	.66	.71	.65
	utterance	.68	**.72**	.65	**.77**	.72	.72
	sentence	.69	.68	.64	.69	.74	.73
DialogLED-l	3previous	**.78**	.72	.00	.82	.81	.00
	1pre-1post	.72	.70	.00	**.84**	.76	.00
	utterance	.75	.72	.00	.82	.32	.00
	sentence	.76	.71	.00	.79	.82	.00

Table 7: F1 scores of supervised models on the full *development* set. The best results obtained by each model architecture and on each training set are boldfaced. We underline the results corresponding to the best test set performance (which are in boldface in Table 2 of the main paper).

Model	RX	RD
Regex baseline	.41	.48
OLMo-2-7B	.50	
Llama-3.2-3B	.47	
Llama-3.3-70B	.50	

Table 8: F1 scores by the regex baseline and the LLMs.

	BNC	**Reddit**
RoBERTa-base	.28	.55
RoBERTa-large	.38	.54
DialogLED-base	.37	.55
DialogLED-large	.26	.54

Table 9: Illustrating the disparity of performance across corpora. Results are F1-scores obtained with the best model by each architecture, trained on the full dataset.

Phenomenon	**# in test**	**Accuracy**
Distractors	24	58.3%
NOTHING	4,288	98.6%
WMN: NON	39	41.0%
WMN: DIN	24	79.2%
NON-PURSUED	27	63.0%

Table 10: Accuracy by conversational phenomenon. We also report the frequency of each phenomenon on the test set.

Regex	**Regex (full dataset)**	**Model (test set)**
what is	6.4%	9.1% (11)
this is not x	4.2%	20.0% (15)*
this isn't x	5.2%	66.7% (9)**
is that a/the x	4.3%	45.5% (11)**
repetition pattern	3.0%	0.0% (3)
what do you mean by	62.4%	87.5% (16)
you mean	25.5%	69.2% (31)**
can you define	20.2%	64.7% (16)**
definition of	32.2%	66.7% (20)*

Table 11: Precision of different regexes in the Indicators dataset and of our models on test instances containing them. Numbers in parentheses correspond to the number of predicted-positives by the model.* and ** indicate significance with $\alpha = 0.05$ and $\alpha = 0.001$.

the following utterance is marked as an indicator in NeWMe but was not matched by any regular expression:

"For me introversion is not about shyness, it's about how after I have a social interaction, I need to be alone for awhile to recover. People tire me out, and I would rather be daydreaming, and like to catch up on my thoughts. But by your own logic, everyone is also introverted, because we could learn to have a negative self image."

Instead, the regular expression for "definition of" matched another utterance in a parallel subthread, specifically the sentence:

"This definition of introversion/extroversion is much different than shyness, which I think is what the OP is describing."

The latter utterance actually contains another indicator, expressing a disagreement about the same trigger ("introvert") but in a different subthread of

the same Reddit conversation.

H Examples of Low-confidence Negatives and Incorrect Positives

Table 12 contains a selection of examples of cases from new data (Section 6) that were predicted as positive or as low-confidence negatives by the model, together with our explanation as to why they constitute interesting examples.

Our characterization	Instance
Expressing lack of knowledge	"I'm not familiar with that."
Distractor (NE)	"I don't know who Mitch Schneider is."
Word-centered questions requesting more information, but not about meaning	"So when you say you went through, what what did you go through before you embarked on on surrogacy?" "Oh, fajitas, how do you make fajitas?"
Talking about ambiguity	"the surrogacy act was ambiguous"
Discussing word choices and/or meaning	"I shouldn't probably say control. I mean regulation. Control is something that I wouldn't want the Federal government to have." "I was trying to say hemorrhoids, no, I was trying to say hem- , hormones."
Tricky and context-dependent cases	"*That's* racism" "I think you might be confusing cause and effect." "But I don't know about these, uh, these, uh, uh, these pincers, these, now, what are they called. Pit bulls, pit bulls"
Expressing what a word or phrase means to them	"I guess invasion of privacy, uh, to me, for example would be unauthorized use of credit cards." "I guess, uh, when I think nursing home I do think of people that are not able to do, take care of themselves physically."
Providing more information about a word	"Oh , knishes, no you don't, you don't see a lot of, that's basically Eastern European Jewish food."
Suggesting someone's way of thinking of a concept is limited	"You seem to think only of wrong timing on emotional terms."
Seemingly challenging definitions	"Their criterion and definition for authority are vastly different from ours."

Table 12: Examples of interesting negative cases from the unlabeled data predicted to be positives or low-confidence negatives by our model.