What's Mine becomes Yours: Defining, Annotating and Detecting Context-Dependent Paraphrases in News Interview Dialogs

Anna Wegmann[1], Tijs van den Broek[2] and Dong Nguyen[1]

[1]Utrecht University, Utrecht, The Netherlands
[2]Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
{a.m.wegmann, d.p.nguyen}@uu.nl, t.a.vanden.broek@vu.nl

Abstract

Best practices for high conflict conversations like counseling or customer support almost always include recommendations to paraphrase the previous speaker. Although paraphrase classification has received widespread attention in NLP, paraphrases are usually considered independent from context, and common models and datasets are not applicable to dialog settings. In this work, we investigate paraphrases across turns in dialog (e.g., Speaker 1: "That book is mine." becomes Speaker 2: "That book is yours."). We provide an operationalization of context-dependent paraphrases, and develop a training for crowd-workers to classify paraphrases in dialog. We introduce ContextDeP, a dataset with utterance pairs from NPR and CNN news interviews annotated for context-dependent paraphrases. To enable analysis on label variation, the dataset contains 5,581 annotations on 600 utterance pairs. We present promising results with in-context learning and with token classification models for automatic paraphrase detection in dialog.

1 Introduction

Repeating or paraphrasing what the previous speaker said has time and time again been found to be important in human-to-human or human-to-computer dialogs: It encourages elaboration and introspection in counseling (Rogers, 1951; Miller and Rollnick, 2012; Hill, 1992; Shah et al., 2022), can help deescalate conflicts in crisis negotiations (Vecchi et al., 2005; Voss and Raz, 2016; Vecchi et al., 2019), can have a positive impact on relationships (Weger Jr et al., 2010; Roos, 2022), can increase the perceived response quality of dialog systems (Weizenbaum, 1966; Dieter et al., 2019) and generally provides tangible understanding-checks to ground what both speakers agree on (Clark, 1996; Jurafsky and Martin, 2019).

Fortunately, in NLP, paraphrases have received wide-spread attention: Researchers have created

Guest: And people always prefer, of course, to see the pope as the principal celebrant of the mass. So that's good. That'll be tonight. And it will be his 26th mass and it will be the 40th or, rather, the 30th time that this is offered in round the world transmission. And it will be my 20th time in doing it as a television commentator from Rome so.
Host: Yes, you've been doing this for a while now.

Figure 1: **Context-Dependent Paraphrase in a News Interview.** The interview host paraphrases part of the guest's utterance. It is only a paraphrase in the current context (e.g., *doing something 20 times* and *doing something for a while* are not generally synonymous). Our annotators provide word-level highlighting. The color's intensity shows the share of annotators that selected the word. Here, most annotators selected the same text spans, some included "from Rome" as part of what is paraphrased by the host. We underline the paraphrase identified by our fine-tuned DeBERTa token classifier.

numerous paraphrase datasets (Dolan and Brockett, 2005; Zhang et al., 2019; Dong et al., 2021; Kanerva et al., 2023), developed methods to automatically identify paraphrases (Zhang et al., 2019; Wei et al., 2022a; Zhou et al., 2022), and used paraphrase datasets to train semantic sentence representations (Reimers and Gurevych, 2019; Gao et al., 2021) and benchmark LLMs (Wang et al., 2018; bench authors, 2023). However, most previous work (1) has focused on context-independent paraphrases, i.e., texts that are semantically equivalent independent from the given context, and has not investigated the automatic detection of paraphrases across turns in dialog, (2) has classified paraphrases at the level of full texts even though paraphrases often only occur in portions of larger texts (see also Figure 1), (3) uses a small number of 1–3 annotations per paraphrase pair (Dolan and Brockett, 2005; Kanerva et al., 2023), (4) only annotate text pairs that are "likely" to include paraphrases using heuristics such as lexical similarity (Dolan and Brockett, 2005), although, especially for the dialog setting, we can not expect lexical similarity to be

Dataset	Agreement Acc.	α	Single Example with High Variation Shortened Example	Vote
BALANCED	0.71	0.32	**Guest:** [...] Maybe the money will help. **Host:** It can't hurt, let's put it that way.	9/20
RANDOM	0.72	0.23	**G:** So both parties agree that we need to stop horrific acts of violence against animals. But everyone is standing behind this. It is time to stop horrific acts of brutality on animals. **H:** Britain's Queen Elizabeth's senior dresser writes "If her majesty is due to attend an engagement in particularly cold weather from 2019 onwards fake fur will be used to make sure she stays warm." it's a very stark example of a monarch following public opinion in the U.K. which is moving away from fur and it very much embraces prevention of cruelty to the animals.	7/15
PARA	0.65	0.19	**G:** [...] it could be programmed in. But again, you'd have to set that up as part of your flight plan. **H:** So you'd have to say I'm going to drop to 5,000 feet, then go back up to 35,000 feet, and you would have had to have done that at the beginning.	8/15

Table 1: **Agreement Scores as an Indicator of Plausible Variation.** For each dataset, we display the "accuracy" with the majority vote (Acc.) which is the mean overlap of a rater's classification with the majority vote classification excluding the current rater and Krippendorff (1980)'s alpha (α) for the binary classifications by all raters over all pairs. The relatively low K's α scores can be explained by pairs where either label is plausible. We display such an example for each dataset with the share of annotators classifying it it as a paraphrase (Vote).

high for all or even most paraphrase pairs (e.g., the pair in Figure 1 only overlaps in two words) and (5) either use short annotation instructions (Dolan and Brockett, 2005) that rely on annotator intuitions or long and complex instructions (Kanerva et al., 2023) that limit the total number of annotators.

We address all five limitations with this work. First, we are, to the best of our knowledge, the first to focus on operationalizing, annotating and automatically detecting **context-dependent paraphrases across turns in dialog**. Dialog is a setting that is uniquely sensitive to context (Grice, 1957, 1975; Davis, 2002), e.g., "doing this for a while now" and "20th time [...] as a television commentator" in Figure 1 are not generally semantically equivalent. Second, instead of classifying whether two complete texts A and B are paraphrases of each other, we focus on classifying whether there exists a selection of a text B that paraphrases a selection of a text A, and **identifying the text spans that constitute the paraphrase pair** (e.g., Figure 1). Third, we collect a **larger number of annotations** of up to 21 per item in line with typical efforts to address plausible human label variation (Nie et al., 2020; Sap et al., 2022). Even though context-dependent paraphrase identification in dialog might at first seem straight forward with a clear ground truth, similar to other "objective" tasks in NLP (Uma et al., 2021), human annotators (plausibly) disagree on labels (Dolan and Brockett, 2005; Kanerva et al., 2023). For example, consider the first

text pair in Table 1. "[The money] can't hurt" can be interpreted in at least two different ways: as a statement with approximately the same meaning as "the money will help" or as an opposing statement meaning the money actually won't help but at least "It can't hurt" either. Fourth, instead of using heuristics to select text pairs for annotations, we choose a dialog setting where paraphrases are relatively likely to occur: transcripts of **NPR and CNN news interviews** (Zhu et al., 2021) since in (news) interviews paraphrasing or more generally active listening is encouraged (Clayman and Heritage, 2002; Hight and Smyth, 2002; Sedorkin et al., 2023). While the interview domain shows some unique characteristics limiting generalizability (e.g., hosts using paraphrases to simplify the guest's statements for the audience), the interview domain is is suitable to demonstrate our new task and includes a diverse set of topics and guests. Fifth, we develop an annotation procedure that goes beyond relying on intuitions and is scalable to a large number of annotators: an accessible **example-centric, hands-on, 15-minute training** before annotation.

In short, we operationalize context-dependent paraphrases in dialog with a definition and an iteratively developed hands-on training for annotators. Then, annotators classify paraphrases and identify the spans of text that constitute the paraphrase. We release ContextDeP (Context-Dependent Paraphrases in news interviews), a dataset with 5,581 annotations on 600 utterance

What?	Shortened Examples
Clear Contextual Equivalence $\subseteq$ **CP**	Guest: I know they are cruel. Host: You know they are cruel.
	G: We have been the punching bag of the president. H: The president has been using Chicago as a punching bag.
Approximate Contextual Equivalence $\subseteq$ **CP**	G: I'm like, "Fortnite", what is that? I don't even know what it is – H: So, you weren't even familiar?
	G: My wife is going through the same thing herself. H: She's also looking for work.

Table 2: **Contextual Paraphrases (CP).** We include text spans ($\subseteq$ CP) that range from clear to approximate equivalence for the given context. Few examples are very clear. Deciding between approximate equivalence and non-equivalence turns out to be a difficult task. In our dataset, annotator agreement scores can be used as a proxy for the ambiguity of an item.

What?	Shortened Example
Additional Conclusions or Facts $\not\subseteq$ **CP**	Guest: If you're not in our country, there are no constitutional protections for you. Host: So, you don't have a problem with Facebook giving the government access to the private accounts of people applying to enter the U.S.?
Isolated Equivalence $\not\subseteq$ **CP**	G: There are militant groups out there firing against <u>the military</u>. H: Why did <u>the army</u> decide today to move in and clear out the camp?

Table 3: **Non-Paraphrases in Dialog.** We do not include text pairs ($\not\subseteq$ CP) that are semantically related but where the second speaker does not actually rephrase a point the first speaker makes. Frequent cases are text spans that might only be considered approximately equivalent when taken out of context (<u>underlined</u>) and pairs that have too distant meanings, for example, when the interviewer continues with the same or a related topic but adds further-reaching conclusions or new facts.

pairs from NPR and CNN news interviews. We use in-context learning (ICL) with generative models like `Llama 2` or `GPT-4` and fine-tune a `DeBERTa` token classifier to detect paraphrases in dialog. We reach promising results of F1 scores from 0.73 to 0.81. Generative models perform better at classification, while the token classifier provides text spans without parsing errors. We hope to advance dialog based evaluations of LLMs and the reliable detection of paraphrases in dialog. Code[1], annotated data[2,3] and the trained model[4] are publicly available for research purposes.

2 Related Work

Paraphrases have most successfully been classified by encoder architectures with fine-tuned classification heads (Zhang et al., 2019; Wahle et al., 2023) and more recently using in-context learning with generative models like `GPT-3.5` and `Llama 2` (Wei et al., 2022a; Wang et al., 2022c; Wahle et al., 2023). To the best of our knowledge, only Wang et al. (2022a) go beyond classifying paraphrases at the complete sentence level. They use a `DeBERTa` token classifier to highlight text spans that are not part of a paraphrase, i.e., the reverse of our task.

Paraphrase taxonomies commonly go beyond binary classifications to make more fine-grained distinctions between paraphrase types, often including considerations w.r.t. the context of the text pairs. Bhagat and Hovy (2013) and Kovatchev et al. (2018) describe substitutions and other lexical operations that result in paraphrases in a given sentential context. Shwartz and Dagan (2016) show that context information can reverse semantic relations between phrases. Vila et al. (2014) discuss text pairs that are equivalent when one presupposes encyclopedic or situational knowledge (e.g., referents or intentions[5]), but exclude them as non-paraphrases. Further, to the best of our knowledge, most previous work annotate sentence pairs without considering the document context, with Kanerva et al. (2023) being the only exception, and no previous work looking at detecting paraphrases in dialog.

Dialog act taxonomies aim to classify the communicative function of an utterance in dialog and commonly include acts such as `Summarize/Reformulate` (Stolcke et al., 2000; Core and Allen, 1997). However, generally, communicative function can be orthogonal to meaning equivalence. For example, the paraphrase from Table 2 "So you weren't even familiar?" would probably be a `Declarative Yes-No-Question` dialog act (Stolcke et al., 2000), while the non-paraphrase "So you don't have a problem with … ?" in Table 3 would also be a `Declarative Yes-No-Question`. We see paraphrase detection in dialog as more ele-

---

[5] cases like 'Close the door please' and "There is air flow"

mentary and complementary to investigating communicative function of utterances.

3 Context-Dependent Paraphrases in Dialog

In NLP, paraphrases typically are pairs of text that are approximately equivalent in meaning (Bhagat and Hovy, 2013), since full equivalence usually only applies for practically identical strings (Bhagat and Hovy, 2013; Dolan and Brockett, 2005) – with some scholars even claiming that different sentences can never be fully equivalent in meaning (Hirst, 2003; Clark, 1992; Bolinger, 1974). The field of NLP has mostly focused on paraphrases that are *context-independent*, i.e., approximately equivalent without considering a given context (Dolan and Brockett, 2005; Wang et al., 2018; Zhang et al., 2019). Some studies have operationalized paraphrases using more fine-grained taxonomies, where context is sometimes considered (Bhagat and Hovy, 2013; Vila et al., 2014; Kovatchev et al., 2018). However, only a few datasets include such paraphrases (Kovatchev et al., 2018; Kanerva et al., 2023) and to the best of our knowledge none that focus on context-dependent paraphrases or dialog data.

We define a **context-dependent paraphrase** as two text excerpts that are at least approximately equivalent in meaning in a given situation but not necessarily in all non-absurd situations.[6] For example, consider the first exchange in Table 2. In this situation, "I" uttered by the first speaker and "You" uttered by the second speaker are clearly signifying the same person. However, if uttered by the same speaker "I" and "you" probably do not signify the same person. The text pair in Table 2 is thus equivalent in at least one but not in all non-absurd situations. The text excerpts forming context-dependent paraphrases do not have to be complete utterances. In many cases they are portions of utterances, see highlights in Figure 1. Note that in dialog, the second speaker should rephrase part of the first speaker's point in the given situation (*context* condition) and not just talk about something semantically related (*equivalence* condition).

Context-dependent paraphrases range from clear (first example in Table 2) to approximate contextual equivalence (last example in Table 2). When the guest says "My wife is going through the same

thing", it seems reasonable to assume that the host is using contextual knowledge to infer that "the same thing" and "looking for a job" are equivalent for the given exchange. Even though in this last example the meaning of the two utterances could also be subject to different interpretations, we still consider such cases to be context-dependent paraphrases for two reasons: (1) similar to findings in context-independent paraphrase detection, limiting ourselves to very clear cases would mostly result in uninteresting, practically identical strings and (2) we ultimately want to identify paraphrases in human dialog, which is full of implicit contextual meaning (Grice, 1957, 1975; Davis, 2002).

We specifically exclude common cases of disagreements between annotators[7] that we consider not to be context-dependent paraphrases in dialog, see Table 3. First, we exclude text spans that might be considered approximately equivalent when they are looked at in isolation but do not represent a paraphrase of the guest's point in the given situation (e.g., "the military" and "the army" in Table 3). Second, we exclude text pairs that diverge too much from the original meaning when the second speaker adds conclusions, inferences or new facts. In an interview setting, journalists make use of different question types and communication strategies relating to their agenda (Clayman and Heritage, 2002) that can sometimes seem like paraphrases. For example in Table 3, the host's question "So, you ...?" could be read as a paraphrase with the goal of checking understanding with the guest. However, it is more likely to be a declarative conclusion that goes beyond what the guest said.

4 Dataset

Generally, people do not paraphrase each other in every conversation. We focus on the news interview setting, because paraphrasing, or more generally active listening, is a common practice for journalists (Clayman and Heritage, 2002; Hight and Smyth, 2002; Sedorkin et al., 2023). We therefore also only consider whether the journalist (the interview host) paraphrases the interview guest and not the other way around. We use Zhu et al. (2021)'s *MediaSum* corpus which consists of over 450K news interview transcripts and their summaries from 1999–2019 NPR and 2000–2020 CNN interviews.[8]

---

[6]definition combines elements from Kanerva et al. (2021) and Bhagat and Hovy (2013)

[7]derived from pilot studies, see also App. C.1 and specifically App. Table 15

[8]Released for research purpose, see https://github.com/zcgzcgzcg1/MediaSum?tab=readme-ov-file.

Dataset	size	# paraphrases	# anns/item
BALANCED	100	54	20.1
RANDOM	100	13	5.7
PARA	400	254	7.5
Total	600	321	9.3

Table 4: **Dataset Statistics.** For each dataset, we display the size, the number of paraphrases according to the majority vote and the average annotations per text pair.

Split	# (guest, host)-pairs	# annotations
Train	420	3896
Dev	88	842
Test	92	843
Total	600	5,581

Table 5: **Split of Dataset.** For each set, we show the number of text pairs and the total number of annotations.

4.1 Preprocessing

We only include two-person interviews, i.e., a conversation between an interview host and a guest. We remove interviews with fewer than four turns, utterances that only consist of two words or of more than 200 words, and the first and last turns of interviews (often welcoming addresses and goodbyes). Overall, this leaves 34,419 interviews with 148,522 (guest, host)-pairs. See App. B.1 for details.

4.2 Data Samples for Annotation

Even though paraphrases are relatively likely in the news interview setting, most randomly sampled text pairs still do not include paraphrases. To distribute annotation resources to text pairs that are likely to be paraphrase, previous work usually selects pairs based on heuristics like textual similarity features, e.g., word overlap, edit distance, or semantic similarity (Dolan and Brockett, 2005; Su and Yan, 2017; Dong et al., 2021). However, these approaches are systematically biased towards selecting more obvious, often lexically similar text pairs, possibly excluding many context-dependent paraphrases. For example, the guest and host utterance in Figure 1 have varying lengths, only overlap in three words and have a semantic similarity score of only 0.13[9]. Similar to Kanerva et al. (2023), we instead use a manual selection of promising text pairs for annotation: We (1) randomly sample a set of text pairs and (2) manually classify at each of them to (3) select three sets of text pairs that vary in their paraphrase distribution for the more resource-intensive crowd-sourced annotations: the RANDOM, BALANCED and PARA set.

Lead Author Annotation. We shuffle and uniformly sample 1,304 interviews. For each interview, we sample a maximum of 5 consecutive (guest, host)-pairs. To select promising paraphrase candidates, the lead author then manually classi-

fies all 4,450 text pairs as paraphrases vs. non-paraphrases (see App. B.2 for details).[10] In total, about 14.9% of the sampled text pairs are classified as paraphrases by the lead author. On a random set of 100 (guest, host)-pairs (RANDOM), we later compare the lead author's classifications with the crowd-sourced paraphrase classifications (see App. B.2). 89% of the lead author's classifications are the same as the crowd majority. Note that the lead author's classifications do not affect the quality of the annotations released with the dataset but only the text pairs that are selected for annotation. However, using lead author annotations instead of lexical level heuristics should increase paraphrase diversity in the released dataset beyond high lexical similarity pairs.

Paraphrase Candidate Selection. We sample three datasets for annotation that differ in their estimated paraphrase distributions (based on the lead author annotations): **BALANCED** is a set 100 text pairs sampled for equal representation of paraphrases and non-paraphrases. We annotate this dataset first with a high number of annotators per (guest, host)-pair, to decide on a crowd-worker allocation strategy that performs well for paraphrases as well as non-paraphrases. **RANDOM** is a uniform random sample of 100 text pairs. One main use of the dataset is to evaluate the quality of crowd-worker annotations on a random sample. **PARA** is a set of 400 text pairs with an estimated 84% of paraphrases designed to increase the variety of paraphrases in our dataset. Details on the sampling of the three datasets can be found in App. B.3.

5 Annotation

We first describe the annotation task (§5.1). Then, we discuss why the annotation task is difficult and a clear ground truth classification might not exist in many cases (§5.2). Therefore, we dynamically collect many judgments for text pairs with

---

[10]After experimenting with crowd-workers, having a first pass for selection done by one of our team seemed the best considering cost-performance trade-offs.

Dataset	Guest		Host	
	α	$\frac{A \cap B}{A \cup B}$	α	$\frac{A \cap B}{A \cup B}$
BALANCED	0.42	0.51	0.48	0.63
RANDOM	0.53	0.63	0.53	0.64
PARA	0.43	0.50	0.50	0.64

Table 6: **Agreement on highlights.** For pairs that at least two annotators classified a paraphrase, we display the average lexical overlap between the highlights (Jaccard Index displayed as $\frac{A \cap B}{A \cup B}$) and Krippendorff's unitizing α over all words for guest and host highlights, see Krippendorff (1995).

Shortened Examples
G: we don't really know what went into their algorithm to make it turn out that way.
H: We're talking about algorithms, but should we be talking about the humans who design the algorithms?
G: In Harrison County.
H: In Harrison County. Are you [...]

Table 7: **Low Quality Annotations.** We show human highlights that can be considered wrong or noisy. When absent, we underline the correct highlights.

high disagreements (§5.4). The annotation of utterance pairs takes place in two rounds with `Prolific` crowd-workers: (1) training crowd-workers (§5.3) and (2) annotating paraphrases with trained crowd-workers (§5.4 and §5.5).

5.1 Annotation Task

Given a (guest, host) utterance pair, annotators (1) classify whether the host is paraphrasing any part of the guest's utterance and, if so, (2) highlight the paraphrase in the guest and host utterance. This results in data points like the one in Figure 1. Note that our setup differs from prior work, which usually involves classifying whether an entire text B is a paraphrase of an entire text A (e.g., Dolan and Brockett, 2005). Instead, given texts A and B, our task is to determine whether there exists a selection of words from text B and text A, where the selection of text B is a paraphrase of the selection of text A. Our annotators are not only performing binary classification, but they also highlight the position of the paraphrase. To the best of our knowledge, we are the first to approach paraphrase detection in this way. Moreover, in contrast to previous work, the considered text pairs are usually longer than just one sentence and are contextualized dialog turns.

5.2 Plausible Label Variation

The task of annotating context-independent paraphrases is already difficult. Disagreements between human annotators are common (Dolan and Brockett, 2005; Krishna et al., 2020; Kanerva et al., 2023) — even with extensive manuals for annotators (Kanerva et al., 2023). In related semantic tasks like textual entailment,[11] disagreements have been linked to plausible label variations inherent to the task

(Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Jiang and de Marneffe, 2022).

Our task setup adds further challenges: First, instead of classifying full sentence pairs, annotators have to read relatively long texts and decide whether any portion of the text pair is a paraphrase. Second, while in previous work annotators usually had to decide if two texts are generally approximately equivalent, they now need to identify paraphrases in a highly contextual setting with often incomplete information.

As a result, similar to the task of textual entailment, we expect classifying context-dependent paraphrases in dialog to not always have a clear ground truth. We display examples of plausible label variation in Table 1. To handle label variation, common strategies are performing quality checks with annotators (Jiang and de Marneffe, 2022) and recruiting a larger number of annotators for a single item (Nie et al., 2020; Sap et al., 2022). We do both, see our approach in §5.3 and §5.4.

5.3 Annotator Training

When annotating paraphrases, the instructions for annotators are often short, do not explain challenges and rely on annotator intuitions (Dolan and Brockett, 2005; Lan et al., 2017).[12] In contrast, Kanerva et al. (2023) recently used an elaborate 17-page manual. However, they relied on only 6 expert annotators that might not be able to represent the full complexity of the task (§5.2). We aim for a trade-off between short intuition-based and long complex instructions that facilitates recruitment of a larger number of annotators: an accessible example-centric, hands-on 15-minute training of annotators that teaches our operationalization of context-dependent paraphrases (§3). We provide

---

[11]Paraphrase classification has been repeatedly equated to (bi-)directional entailment classification (Dolan and Brockett, 2005; Androutsopoulos and Malakasiotis, 2010)

[12]For example, instructions are to rate if two sentences "mean the same thing" (Dolan and Brockett, 2005) or are "semantically equivalent" (Lan et al., 2017).

Model	Classification				Highlighting		
	Extract ↓	F1 ↑	Prec ↑	Rec ↑	Extract ↓	Jacc Guest ↑	Jacc Host ↑
llama 2 7B	<u>1%</u>	0.66	0.49	**0.98**	59%	0.34	0.44
vicuna 7B	**1%**	0.29	0.67	0.19	32%	0.30	0.46
Mistral 7B Instruct v0.2	3%	0.62	0.66	0.58	66%	0.40	0.51
openchat 3.5	**0%**	0.66	0.76	0.58	64%	0.46	0.50
gemma 7B	<u>1%</u>	0.64	0.66	0.63	48%	0.24	0.51
Mixtral 8x7B Instruct v0.1	**0%**	<u>0.74</u>	0.73	0.74	65%	0.35	0.52
Llama 2 70B	**0%**	0.66	0.72	0.61	71%	0.29	0.56
GPT-4	**0%**	**0.81**	<u>0.78</u>	<u>0.84</u>	<u>17%</u>	**0.67**	**0.71**
DeBERTa v3 large AGGREGATED	-	0.73	0.67	0.81	-	<u>0.52</u>	0.66
DeBERTa v3 large ALL	-	0.66	**0.82**	0.56	-	0.45	0.64

Table 8: **Modeling Results.** We **boldface** the best and <u>underline</u> the second best performance. We display the extraction error of predictions from generative models and, for classification, the F1, precision and recall score as well as, for highlights, the Jaccard Index for the guest and host utterances. Higher values are better (↑) except for extraction errors (↓). `GPT-4` is the best classification model, while, overall, `DeBERTa` is the best highlight model as it does not lead to any extraction errors.

(1) a short paraphrase definition, (2) examples of context-dependent paraphrases showing clear and approximate equivalence (c.f. Table 2), (3) examples of common difficulties with paraphrase classification in dialog (c.f. Table 3 and §3), and use (4) a hands-on approach where annotators have to already classify and highlight paraphrases after receiving instructions. Only once they make the right choice on what is (Table 2) and is not a paraphrase (Table 3) and highlight the correct spans they are shown the next set of instructions. Only annotators that undergo the full training and pass two comprehension and two attention checks are part of our released dataset. Overall, 49% of the annotators who finished the training passed it. See App. C for the instructions and further details.

5.4 Annotator Allocation

To the best of our knowledge, text pairs in paraphrase datasets receive a fixed number of 1, up to a maximum of 5 annotations (Kanerva et al., 2023; Zhang et al., 2019; Lan et al., 2017; Dolan and Brockett, 2005). However, this might not be enough to represent the inherent plausible variation to the task (§5.2). We have each pair in BAL-ANCED annotated by 20–21 trained annotators to simulate different annotator allocation strategies (App. C.4). Then, for RANDOM and PARA, we use a dynamic allocation strategy: Each pair receives at least 3 annotations. We dynamically collect more annotations, up to 15, on pairs with high disagreement (i.e., entropy > 0.8). Overall, this results in an average of 9 annotations per text pair across our released dataset.

5.5 Results

We discuss annotations results (tables 1, 4, 6) on our datasets BALANCED, RANDOM and PARA.

Classification agreement as an indicator of variation. Agreement for classification is relatively low (Table 1). We inspect a sample of 100 annotations on the RANDOM set and manually assess annotation quality. 90% of the annotations can be said to be at least plausible (see Table 7 for low quality and Table 1 for plausible variation examples), which is in line with the fact that we only use high quality annotators (§5.3). Further, we manually analyze the 42 annotations of ten randomly sampled annotators: Nine annotators consistently provide high quality annotations, while the other annotator chooses "not a paraphrase" a few times too often (see Appendix C.7 for details). As a result, we assume that most disagreements are due to the inherent plausible label variation of the task (§5.2).

Higher agreement on paraphrase position. Krippendorff's unitizing α on the highlights is higher than in other areas[13] (see Table 6). We also calculate the "Intersection-over-union" between the highlighted words (i.e., Jaccard Index), a common and interpretable evaluation measure for annotator highlights (Herrewijnen et al., 2024; Mendez Guzman et al., 2022; Mathew et al., 2021; Malik et al., 2021). It seems that while annotations vary on whether there is a paraphrase or not, they agree frequently on the position of the possible paraphrase. On average, at least 50% of the highlighted words

---

[13]E.g., 0.41 for hate speech (Carton et al., 2018) or 0.35 for sentiment analysis (Sullivan Jr. et al., 2022). Because of the different tasks these values are not exactly comparable.

	Preds		Shortened Examples
T	**G**	**D**	
✗	✗	✓	**G:** He was <u>the most famous guy</u> in the world of sports... **H:** <u>The most famous Italian...</u>
✓	✗	✓	**G:** A lot of them were the Bay Area influx that came up and <u>bought homes to flip</u>. You know what flipping is, right? **H:** Mm-hmm. Buying a house, improving it, selling it out of profit.

Table 9: **Model Errors.** We show examples of prediction errors made by DeBERTa (D) and GPT-4 (G). We display model predictions (D/G) for paraphrases (✓) and non-paraphrases (✗) and compare it to the crowd-majority (T). If one model predicted a paraphrase the corresponding text spans are <u>underlined</u>. For comparison, we also display the crowd majority highlights.

Shortened Example
G: ... then he goes on and references and **makes mention of Rudy Giuliani three times** in this conversation   **H:** And **Rudy Giuliani** was a private lawyer not a government official, so why is **he coming up** so **much in this** conversation between two world leaders?

Table 10: **Highlighting Differences.** We show examples of highlights made by DeBERTa, **GPT-4** and human highlights. Lower intensity means less human annotators selected the word. While GPT-4 struggles with providing highlights at all (c.f. extraction error in Table 8), DeBERTa highlights tend to be too sparse (just "Rudy Giuliani", "coming" and "conversation" in the host utterance). Here, we highlight words, when the softmax probability is > 0.44[17] instead of ≥ 0.5. On the complete test set, this also increases the mean Jaccard Index (by 0.06/0.01 for guest/host compared to Table 8).

are the same between annotations.[14] Agreement is higher on the host utterance, because on average the host utterance is shorter than the guest utterance ($33 < 85$ words).

Label variation is highest for paraphrases. Between the datasets, classification agreement is lowest for PARA. This is what we expected since it has the largest portion of "hard" non-repetition paraphrases (see App. B.3). Krippendorff's α is lower for the RANDOM than the BALANCED set, even though we expected the RANDOM set to include easier decisions for annotators (RANDOM includes more unrelated non-paraphrases, see App. B.3). As the other agreement heuristic is relatively high on RANDOM, the lower α values could be a result of Krippendorff's measure being sensitive to imbalanced label distributions (Riezler and Hagmann, 2022), see also Table 4 displaying the imbalanced distribution for RANDOM.

6 Modeling

In Table 5, we do a random 70, 15, 15 split of our 5,581 annotations, along the 600 unique pairs.

Token Classifier. Similar to Wang et al. (2022a), we fine-tune a large DeBERTa model[15] (He et al., 2020) on token classification to highlight the paraphrase positions (for hyperparameters, see App. D.2). We train two models: using all 3,896 training annotations ("ALL" in Table 8) and using the majority aggregated training annotations over

the 420 unique (guest, host) training pairs ("AGGREGATED" in Table 8). We consider a model to have predicted a paraphrase for a pair if at least one token is highlighted with softmax probability ≥ 0.5 in both texts. For each model, we average performances over three seeds.

In-Context Learning. We further prompt the following generative models (see URLs in App. D.1) to both classify and highlight the position of paraphrases: Llama 2 7B and 70B (Touvron et al., 2023), Vicuna 7B (Zheng et al., 2023), Mistral 7B Instruct v0.2 (Jiang et al., 2023), Openchat 3.5 (Wang et al., 2023), Gemma 7B (Team et al., 2024), Mixtral 8x7B Instruct v0.1 (Jiang et al., 2024) and GPT-4[16] (Achiam et al., 2023). We design the prompt to be as close as possible to the annotator training using a few-shot setup (Brown et al., 2020; Zhao et al., 2021) with all 8 examples shown during annotator training. We also provide explanations in the prompt (Wei et al., 2022b; Ye and Durrett, 2022) and use self-consistency by prompting the models 10 (GPT-4 and Llama 70B: 3) times (Wang et al., 2022b). For the prompt and further hyperparameter settings see App. D.1.

Results. For evaluation, we consider a pair to contain a paraphrase if it has been classified by a majority of crowd-workers and a word to be part of the paraphrase if it has been highlighted

---

[14]100% overlap in highlighting is uncommon. DeYoung et al. (2020) consider two highlights a match if Jaccard is greater than 50%.

[15]microsoft/deberta-v3-large

[16]API calls where performed using the "gpt-4" model id in March 2024.

[17]We tried a few different thresholds > 0.40 with 0.44 getting the biggest gain in the Jaccard Index on the test set.

by a majority of crowd-workers. We leave soft-evaluation approaches to future work (Uma et al., 2021), among others because of challenges in extracting label distributions for in-context learning in a straight-forward way (Hu and Levy, 2023; Lee et al., 2023). See Table 8 for test set performances. Performances for the token classifier are the mean over three seeds. Performances for the generative models is the majority vote for the 3–10 self-consistency calls. We display the F1 score for classification and, as before (§5.5), Intersection-Over-Union of the highlighted words for guest and host utterance highlights (Jaccard Indices), see, for example, DeYoung et al. (2020). For in-context learning, we also display how often we could not extract the highlights or classifications from model responses. Note that the test set contains 93 elements, so differences between models might appear bigger than they are.

Overall, GPT-4 and Mixtral 8x7B achieve the best results in paraphrase classification. In highlighting, our DeBERTa token classifiers and GPT-4 achieve the best overlap with human annotations. However, due to problems with extracting highlights from model responses (e.g., hallucinations, see App. D.3), our fine-tuned DeBERTa token classifiers are probably the best choice to extract the position of paraphrases. While the DeBERTa AGGREGATED model achieves higher F1 scores, the DeBERTa ALL model has the highest precision out of all models. We provide our best-performing DeBERTa AGGREGATED model (model with seed 202 and F1 score of 0.76) on the Hugging Face Hub[18] and use it in the following error analysis.

Error Analysis. We consider the best-performing classification and highlighting models for error analysis, i.e., GPT-4 and DeBERTa AGGREGATED. We manually analyze a sample of misclassifications, for examples see Table 9. Overall, the classification quality is better for GPT-4. The DeBERTa classifier finds more paraphrases (note that DeBERTa AGGREGATED for seed 202 has a recall of 0.86) but also predicts more false positives than GPT-4. For both models, the items with incorrect predictions also show higher human disagreement. The average entropy for human classifications is lower for the correct (0.45 for DeBERTa, 0.45 for GPT-4) than for the incorrect model predictions (0.59 for DeBERTa, 0.67 for GPT-4). DeBERTa

highlights shorter spans of text (on average 6.6/6.2, compared to 16.7/10.9 for GPT-4 for guest/host respectively), while GPT-4 usually highlights complete (sub-)sentences. GPT-4 highlights are largely of good quality, however they often can not be extracted (see App. D.3). The DeBERTa highlights can seem "chopped up" and missing key information (e.g., the original host highlights in Table 10 are just "Rudy Giuliani", "coming" and "conversation"). We recommend performing a classification of an utterance pairs as a paraphrase when there exist softmax probabilities ≥ 0.5 for both guest and host utterance, but then selecting the highlights also based on softmax probabilities lower than 0.5. Alternatively, the best DeBERTa ALL model[19] provides fewer but seemingly more consistent highlights (see Appendix D.3). One possible reason for this could be that DeBERTa ALL was trained on individual highlights provided by single annotators, rather than on aggregated highlights.

7   Conclusion

A majority of work on paraphrases in NLP has looked at the semantic equivalence of sentence pairs in context-independent settings. However, the human dialog setting is highly contextual and typical methods fall short. We provide an operationalization of context-dependent paraphrases and an up-scalable hands-on training for annotators. We demonstrate the annotation approach by providing 5,581 annotations on a set of 600 turn pairs from news interviews. Next to paraphrase classifications, we also provide annotations for paraphrase positions in utterances. In-context learning and token classification both show promising results on our dataset. With this work, we contribute to the automatic detection of paraphrases in dialog. We hope that this will benefit both NLP researchers in the creation of LLMs and social science researchers in analyzing paraphrasing in human-to-human or human-to-computer dialogues on a larger scale.

Limitations

Even though the number of our unique text pairs is relatively small, we release a high number of high quality annotations per text pair (5,581 annotations on 600 text pairs). Releasing more annotations on fewer "items" (here: text pairs), has increasingly been more common in NLP (Nie et al., 2020; Sap

---

[18]https://huggingface.co/AnnaWegmann/Highlight-Paraphrases-in-Dialog

[19]https://huggingface.co/AnnaWegmann/Highlight-Paraphrases-in-Dialog-ALL

et al., 2022). Further, big datasets become less necessary with better generative models: Using only eight paraphrases pairs in our prompt already led to promising results. We further use the full 3,896 annotations from the training set to train a token classifier showing competitive results with the open generative models. However, the token classifier and other potential fine-tuning approaches would probably profit from a bigger dataset.

Even though our dataset of news interviews showed frequent, different and diverse occurrences of paraphrasing, it might not be representative of paraphrasing behavior in conversations across different contexts and social groups. In the future, we aim to expand our dataset with further out-of-domain items.

Our data creation process was not aimed at scalability. While our developed annotator training procedure can easily be scaled to a larger group of crowd-workers, we manually selected text pairs for annotation. Future work could scale this by skipping manual selection and accepting a more imbalanced dataset or using our trained classifiers as a heuristic to identify likely paraphrases.

Even though we carefully prepared the annotator training and took several steps to ensure high-quality annotations, there remain several choices that were out of our scope to experiment with, but might have improved quality even more. For example, experimenting with different visualizations of paraphrase highlighting, text fonts, giving annotators an option to add confidence scores for classifications and so on.

We only use one prompt that is as close as possible to the instructions the human annotators receive. We use the same prompt with the exact same formatting for all different generative LLMs. However, experimenting with different prompts might improve performance (Weng, 2023) and some models might benefit from certain formatting or phrasing. We leave in-depth testing of prompts to future work. Further, it might be possible to improve the performance of our DeBERTa model, through providing contextual information (like speaker names and interview summary). Currently, these are only provided to the generative models.

In this work we collect a high number of human annotations per item and highlight the plausible label variation in our dataset. However, we use hard instead of soft-evaluation approaches (Uma et al., 2021) for the computational models. We do this because, among others, extracting label distributions for in-context learning is challenging (Hu and Levy, 2023; Lee et al., 2023). We leave the development of a soft evaluation approach to future work but want to highlight the potential of our dataset here: The high number of annotations per item enables the modeling of classifications and text highlights as distributions, similar to Zhang and de Marneffe (2021). Further, our dataset provides anonymized unique ids for all annotators and enables modeling of different perspectives, e.g., with similar methods to Sachdeva et al. (2022) and Deng et al. (2023).

We do not differentiate between different communicative functions, intentions or strategies that affect the presence of paraphrases in a dialog. This is relevant as paraphrases might, for example, be a more conscious choice by interviewers (Clayman and Heritage, 2002) or a more unconscious occurrence similar to the linguistic alignment of the references for discussed objects (Xu and Reitter, 2015; Garrod and Anderson, 1987). With this work, we hope to provide an outline of the general class of context-dependent paraphrases in dialog that lays the groundwork for further, fine-grained distinctions.

Ethical Considerations

We hope that the ethical concerns of reusing a public dataset (Zhu et al., 2021) are minimal. Especially, since the CNN and NPR interviews are between public figures and were broadcast publicly, with consent, on national radio and TV.

Our dataset might not be representative of English paraphrasing behavior in dialogs across different social groups and contexts as it is taken from U.S. news interviews with public figures from two broadcasters. We caution against using our models without validation on out-of-domain data.

We performed several studies with U.S.-based crowd-workers as part of this work. We payed participants a median of $\approx 11.41\$/h$ which is above federal minimum wage. Crowd-workers consented to the release of their annotations. We do not release identifying ids of crowd-workers.

We confirm to have read and that we abide by the ACL Code of Ethics. Beside the mentioned ethical considerations, we do not foresee immediate risks of our work.

Acknowledgements

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *Computing Research Repository*, arXiv:2303.08774.

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Rahul Bhagat and Eduard Hovy. 2013. Squibs: What is a paraphrase? *Computational Linguistics*, 39(3):463–472.

Dwight Bolinger. 1974. Meaning and form. *Transactions of the New York Academy of Sciences*, 36(2 Series II):218–233.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2018. Extractive adversarial networks: High-recall explanations for identifying personal attacks in social media posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3497–3507, Brussels, Belgium. Association for Computational Linguistics.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

Eve V Clark. 1992. Conventionality and contrast: Pragmatic principles with lexical consequences. In *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*, pages 171–188. Lawrence Erlbaum Associates.

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

Steven Clayman and John Heritage. 2002. *The news interview: Journalists and public figures on the air*. Cambridge University Press.

Mark G Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Aaction in Humans and Machines*, volume 56, pages 28–35. Boston, MA.

Wayne A Davis. 2002. *Meaning, expression and thought*. Cambridge University Press.

Naihao Deng, Xinliang Zhang, Siyang Liu, Winston Wu, Lu Wang, and Rada Mihalcea. 2023. You are what you annotate: Towards better models through annotator representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12475–12498, Singapore. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Justin Dieter, Tian Wang, Arun Tejasvi Chaganty, Gabor Angeli, and Angel X. Chang. 2019. Mimic and rephrase: Reflective listening in open-ended dialogue. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 393–403, Hong Kong, China. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. ParaSCI: A large scientific paraphrase dataset for longer paraphrase generation. In *Proceedings of the*

16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 424–434, Online. Association for Computational Linguistics.

Sean P. Engelson and Ido Dagan. 1996. Minimizing manual annotation cost in supervised training from corpora. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 319–326, Santa Cruz, California, USA. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, 27(2):181–218.

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

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

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced bert with disentangled attention. *Computing Research Repository*, arXiv:2006.03654.

Elize Herrewijnen, Dong Nguyen, Floris Bex, and Kees van Deemter. 2024. Human-annotated rationales and explainable text classification: a survey. *Frontiers in Artificial Intelligence*, 7:1260952.

Joe Hight and Frank Smyth. 2002. *Tragedies & journalists: A guide for more effective coverage*. Dart Center for Journalism and Trauma.

Clara E Hill. 1992. An overview of four measures developed to test the Hill process model: Therapist intentions, therapist response modes, client reactions, and client behaviors. *Journal of Counseling & Development*, 70(6):728–739.

Graeme Hirst. 2003. Paraphrasing paraphrased. In *Keynote address for The Second International Workshop on Paraphrasing: Paraphrase acquisition and Applications*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel,

Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *Computing Research Repository*, arXiv:2310.06825.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *Computing Research Repository*, arXiv:2401.04088.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Dan Jurafsky and James H Martin. 2019. Speech and language processing (3rd ed. draft).

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Aurora Piirto, Jenna Saarni, Maija Sevón, et al. 2021. Annotation guidelines for the Turku paraphrase corpus. *Computing Research Repository*, arXiv:2108.07499.

Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Aurora Piirto, Jenna Saarni, Maija Sevón, and et al. 2023. Towards diverse and contextually anchored paraphrase modeling: A dataset and baselines for finnish. *Natural Language Engineering*, page 1–35.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Venelin Kovatchev, M. Antònia Martí, and Maria Salamó. 2018. ETPC - a paraphrase identification corpus annotated with extended paraphrase typology and negation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Klaus Krippendorff. 1980. *Content analysis: An introduction to its methodology*. Sage publications.

Klaus Krippendorff. 1995. On the reliability of unitizing continuous data. *Sociological Methodology*, pages 47–76.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762, Online. Association for Computational Linguistics.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*,

pages 1224–1234, Copenhagen, Denmark. Association for Computational Linguistics.

Noah Lee, Na Min An, and James Thorne. 2023. Can large language models capture dissenting human voices? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4585, Singapore. Association for Computational Linguistics.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian legal documents corpus for court judgment prediction and explanation. 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 4046–4062, Online. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Erick Mendez Guzman, Viktor Schlegel, and Riza Batista-Navarro. 2022. RaFoLa: A rationale-annotated corpus for detecting indicators of forced labour. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3610–3625, Marseille, France. European Language Resources Association.

William R Miller and Stephen Rollnick. 2012. *Motivational interviewing: Helping people change*. Guilford press.

Yixin Nie, Xiang Zhou, and Mohit Bansal. 2020. What can we learn from collective human opinions on natural language inference data? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9131–9143, Online. Association for Computational Linguistics.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

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*

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stefan Riezler and Michael Hagmann. 2022. *Validity, reliability, and significance: Empirical methods for NLP and data science*. Springer Nature.

Carl Ransom Rogers. 1951. *Client-centered therapy: Its current practice, implications, and theory*. Houghton Mifflin, Boston.

Carla Roos. 2022. *Everyday Diplomacy: dealing with controversy online and face-to-face*. Ph.D. thesis, University of Groningen.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: econometric and statistical modeling with python. *SciPy*, 7:1.

Gail Sedorkin, Amy Forbes, Ralph Begleiter, Travis Parry, and Lisa Svanetti. 2023. *Interviewing: A guide for journalists and writers*. Routledge.

Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. 2022. Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–24.

Vered Shwartz and Ido Dagan. 2016. Adding context to semantic data-driven paraphrasing. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 108–113, Berlin, Germany. Association for Computational Linguistics.

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

Yu Su and Xifeng Yan. 2017. Cross-domain semantic parsing via paraphrasing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1235–1246, Copenhagen, Denmark. Association for Computational Linguistics.

Jamar Sullivan Jr., Will Brackenbury, Andrew McNutt, Kevin Bryson, Kwam Byll, Yuxin Chen, Michael Littman, Chenhao Tan, and Blase Ur. 2022. Explaining why: How instructions and user interfaces impact annotator rationales when labeling text data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 521–531, Seattle, United States. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *Computing Research Repository*, arXiv:2403.08295.

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

Alexandra N Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Gregory M Vecchi, Vincent B Van Hasselt, and Stephen J Romano. 2005. Crisis (hostage) negotiation: current strategies and issues in high-risk conflict resolution. *Aggression and Violent Behavior*, 10(5):533–551.

Gregory M Vecchi, Gilbert KH Wong, Paul WC Wong, and Mary Ann Markey. 2019. Negotiating in the skies of hong kong: The efficacy of the behavioral influence stairway model (BISM) in suicidal crisis situations. *Aggression and violent behavior*, 48:230–239.

Marta Vila, M Antònia Martí, and Horacio Rodríguez. 2014. Is this a paraphrase? What kind? Paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, 4(01):205.

Chris Voss and Tahl Raz. 2016. *Never split the difference: Negotiating as if your life depended on it*. Random House.

Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2023. Paraphrase types for generation and detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12148–12164, Singapore. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. OpenChat: Advancing open-source language models with mixed-quality data. *Computing Research Repository*, arXiv:2309.11235.

Shuohang Wang, Ruochen Xu, Yang Liu, Chenguang Zhu, and Michael Zeng. 2022a. ParaTag: A dataset of paraphrase tagging for fine-grained labels, NLG evaluation, and data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7111–7122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *Computing Research Repository*, arXiv:2203.11171.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022c. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Harry Weger Jr, Gina R Castle, and Melissa C Emmett. 2010. Active listening in peer interviews: The influence of message paraphrasing on perceptions of listening skill. *International Journal of Listening*, 24(1):34–49.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In

Advances in Neural Information Processing Systems, volume 35, pages 24824–24837. Curran Associates, Inc.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Lilian Weng. 2023. Prompt engineering. *lilianweng.github.io*.

Ka Wong and Praveen Paritosh. 2022. k-Rater Reliability: The correct unit of reliability for aggregated human annotations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–384, Dublin, Ireland. Association for Computational Linguistics.

Yang Xu and David Reitter. 2015. An evaluation and comparison of linguistic alignment measures. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 58–67, Denver, Colorado. Association for Computational Linguistics.

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pages 30378–30392. Curran Associates, Inc.

Xinliang Frederick Zhang and Marie-Catherine de Marneffe. 2021. Identifying inherent disagreement in natural language inference. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4908–4915, Online. Association for Computational Linguistics.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Chao Zhou, Cheng Qiu, and Daniel E Acuna. 2022. Paraphrase identification with deep learning: A review of datasets and methods. *Computing Research Repository*, arXiv:1503.06733.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.

	Preprocessed		Sampled		Released	
	# i	# gh	# i	# gh	# i	# gh
all	34419	148522	1304	4450	480	600
NPR	11506	49065	423	1550	167	218
CNN	22913	99457	881	2900	313	382

Table 11: **Dataset Statistics.** Number of interviews (#i) and (guest, host)-pairs (# gh) respectively after preprocessing (§4.1), random sampling (§4.2) and the selection of paraphrase candidates for annotation (§4.2).

A Context-Dependent Paraphrases in Dialog

Should one include repetitions? Repetitions have been typically included in paraphrase taxonomies (Bhagat and Hovy, 2013; Zhou et al., 2022) even though, e.g., Kanerva et al. (2023) asked annotators to exclude such pairs as they considered them uninteresting paraphrases. However, distinguishing repetitions from paraphrases turns out to be especially hard in dialog: speakers tend to leave words out when they repeat and adapt the pronouns to match their perspective (e.g., I -> you). We therefore include repetitions in our definition of context-dependent paraphrases. In fact, those mainly make up the "Clear Contextual Equivalence" Paraphrases (see Table 2).

B Dataset

Topic of the Dataset. The topics of the CNN and NPR news interviews (Zhu et al., 2021) are mostly centered around U.S. politics (e.g., presidential or local elections, 9/11, foreign policy in the middle east), sports (e.g., baseball, football), domestic natural disasters or crimes and popular culture (e.g., interviews with book authors).

Utterance Pair IDs. We use unique IDs for utterance pairs. For example, for NPR-4-2, "NPR-4" is the ID used for interviews[20] as done in Zhu et al. (2021), "2" is the position of the start of the guest utterance in the utterance list as separated into turns by Zhu et al. (2021), in this case "Thank you.".

B.1 Preprocessing

We give details on the three preprocessing steps (see §4.1).

1. Filtering for 2-person interviews. We filter 49,420 NPR and 414,176 CNN interviews from Zhu et al. (2021) for 2-person interviews only.

This can be challenging: In the speaker list, authors sometimes have non-unique identifiers (e.g., 'STEVE PROFFITT', 'PROFFITT' or 'S. PROFFITT' refer to the same speaker). If one author identifier string is contained in the other we assume them to be the same speaker.[21] We generally assume the first speaker to be the host. We remove 538 NPR and 1,917 CNN interviews because the identifier of the second speaker includes the keywords "host" or "anchor" — thus contradicting our assumption. This leaves 14,000 NPR and 50,301 CNN 2-person interviews.

2. Removing first and last turns of an interview. The first turns in our 2-person interviews are usually (reactions to) welcoming addresses and acknowledgments by host and guest[22], while the last often contain goodbyes or acknowledgments[23]. We remove the first two and the last two (guest, host)-pairs. This step removes 2,409 NPR and 26,419 CNN interviews because they are fewer than 5-turns long. For the remaining interviews, this removes 34,773 NPR and 71,646 CNN (guest, host)-pairs.

3. Removing short and long utterances. We further remove short guest utterances of 1–2 words as they leave not much to paraphrase.[24] 3,540 NPR and 12,675 CNN pairs are removed like this. We also remove pairs where the host utterance consists of only 1–2 words.[25]. 2,940 NPR and 11,389 CNN pairs are removed like this. We also remove pairs

---

[20]In this case referring to https://www.npr.org/templates/story/story.php?storyId=16778438

[21]There might be other cases where different string identifiers in the dataset refer to the same speaker although they are not substrings of the other (e.g., 'S. PROFFITT' and 'STEVE PROFFITT'). For a randomly sampled selection of 44 interviews that were identified as more than 2 person interviews, 12 contained errors in the matching. 2/12 were the result of typos and 10/12 were the result of additions to the name like "(voice-over)" or "(on camera)".

[22]For example, "I'm Farai Chideya." "Welcome." "Thank you."

[23]For example the last 3 turns in the considered NPR-4 interview: "Well, Dr. Hader. Thanks for the information.", "Well, thank you for helping share that information [...]", "Well, thanks again. Dr. Shannon Hader [...]"

[24]We manually looked at a random sample of $0.3\% \approx 48$ such pairs. The 1-2 token guest utterances are mostly (40/48) assertions of reception by the guest (e.g., "Yes.", "Exactly. Exactly.", "That's right"). Some are signals of protest (4/48) (e.g., "Hey, man.", "Yes, but...", "Hold on."). None of them were reproduced by the host in the next turn.

[25]We manually looked at a random sample of $0.3\% \approx 37$ such pairs. The 1–2 tokens host utterances are mostly (28/37) assertions of reception by the host (e.g., "Yeah.", "Yes.", "Sure.", "Right.", "Right. Right.", "Ah, okay."). Some are requests for elaboration (5/37) (e.g., "How so?", "Like?", "Four?") or reactions (3/37) (e.g., "Wow!", "Oh, interesting."). Only one example "Four?" was reproducing content in the form of a repetition.
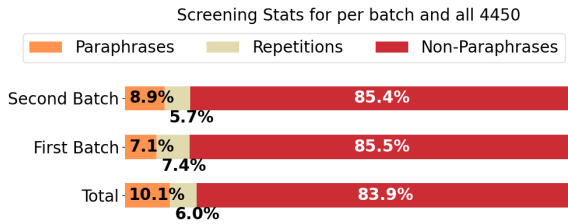
Figure 2: **Label distribution after first author annotations performed in two batches.** First author label classification was performed in two batches. The first batch consists of 750 text pairs, the second of 3,700.

where guest or host utterance consist of more than 200 words.[26] Overall, this leaves 148,522 (guest, host)-pairs in 34,419 interviews for potential annotation, see Table 11.

B.2 First Author Annotations

We provide more details on the first author annotations for selecting paraphrase candidates (§4.2).

Deciding on first author annotations. Since the share of paraphrases in randomly sampled (guest, host)-pairs was only at around 5-15% in initial pilots with lab members, similar to previous work, we opted to do a pre-selection of text pairs before proceeding with the more resource-intensive paraphrase annotation (c.f. §5.5 and App. C). However, commonly used automatic heuristics were not suitable for the highly contextual discourse setting (c.f. §4.2). Instead, we experimented with discarding obvious "non-paraphrases" through crowd-sourced annotations and compared it to manual annotations by the lead author, ultimately deciding on using lead author annotations. One of the reasons was that discarding obvious "non-paraphrases" was more resource intensive and difficult for crowd-workers than expected, making the resources needed for discarding non-paraphrases too close to annotating paraphrases themselves – which defeats the purpose of doing a pre-selection in the first place.

Changing lead author annotations from discarding obvious non-paraphrases to keeping interesting paraphrases. On an initial set of 750 random (guest, host)-pairs, we remained with the initial idea of discarding obvious non-paraphrase pairs. However, due to a resulting high share of uninteresting or improbable paraphrase pairs, we

---

Paraphrase	88
High Lexical Similarity	59
Repetition	45
Perspective-Shift	10
Directional	17
Difficult Decision	16
Non-Paraphrase	**519**
High Lexical Similarity	> 18
Partial	> 24
Unrelated	> 103
Topically Related	> 83
Conclusion	46
Ambiguous	**18**
Missing Context	**125**

Table 12: **Statistics Labels First Batch.** For 750 manually reviewed pairs, we also labeled several other categories. We found 88 paraphrases, 519 non-paraphrases, 18 ambiguous cases and 125 where the missing context impeded a definite decision. Note that we tried to not assign ambiguous if we were leaning to one category over another. Other categorizations include: "perspective-shift" (the perspective shifts between guest and host, e.g., "you" -> "I"), "directional" (guest or host utterance is entailed from or subsumed in the other), "partial" (a subsection could be understood as a paraphrase, but the overall larger section is clearly not a paraphrase), "related" (two utterances are closely related but no paraphrases), "conclusion" (host draws a conclusion or adds an interpretation that goes beyond a paraphrase). Some labels were only added in the last 200 annotations and therefore include the ">" indication.

Dataset	Overlap Lead and Crowd
BALANCED	0.72
RANDOM	0.89
PARA	0.72

Table 13: **Lead vs. Crowd Classifications.** We display the average overlap between the lead author's classifications and the majority vote of the crowd. The overlap is the highest on the RANDOM set. Probably because we keep all obvious non-paraphrases for classification and the annotators face less ambiguous (guest, host)-pairs to classify.

opted to classify paraphrases vs. non-paraphrases instead of possible paraphrases vs. obvious non-paraphrases. The lead author re-annotated the initial set of 750 paraphrase candidates and annotated 4450 additional (guest, host)-pairs for paraphrase vs. non-paraphrase. In the first batch, the lead author additionally labeled a variety of different paraphrase types/difficulties (e.g., high lexical similarity, missing context, unrelated), see also Table 12, in the second batch this was restricted to repe-

Type (guest, host)-pair	#	acc.
Paraphrase	46	0.80
High Lexical Similarity	24	0.92
Repetition	16	0.88
Context-Dependent		
Perspective-Shift	10	0.90
Directional	12	0.67
Other Difficult Cases	12	0.58
Non-Paraphrase	54	0.81
Unrelated utterances	13	1.00
More Difficult	41	0.76
Topically related	24	0.67
High Lexical Similarity	11	0.64
Partial	10	0.80
Conclusion	11	0.55

Table 14: **Selection of 100 Paraphrase Candidates for detailed Annotation.** The sample was selected based on assigned categories during paraphrase candidate annotation. Categories within Paraphrase and Non-Paraphrase can overlap. We display "accuracy" w.r.t. first author annotations.

tition paraphrase, paraphrase and non-paraphrase. The distribution of these three categories is displayed in Figure 2.

Relation to with Crowd Majority Annotations. We display the overlap between the lead author's paraphrase classifications and the released classifications of the crowd majority in Table B.2.

B.3 Paraphrase Candidate Selection

Based on the lead author classifications into paraphrase, non-paraphrase and repetition, we build three datasets for annotation (main paper §4.2). We display the first author classification distribution for the three datasets in Figure 3.

BALANCED. The BALANCED set is a sample of 100 (guest, host)-pairs that were randomly sampled based on the first batch of lead author annotations (§B.2). We had additional lead author labels available for this set, see Table 14 for the distribution of these on the BALANCED set. Constraints were 50 paraphrases and 50 non-paraphrases. In order to include more complex cases, we sampled more difficult than unrelated non paraphrase pairs and we limited the number of repetition paraphrases (51% of paraphrases are repetitions in the full batch, but only 33% of paraphrases in BALANCED are repetitions). Due to a sampling error, we ended up with a 46/56 split. Later, we calculate the majority vote of the 20–21 annotations per (guest, host)-pair on this set, and then evaluate it by
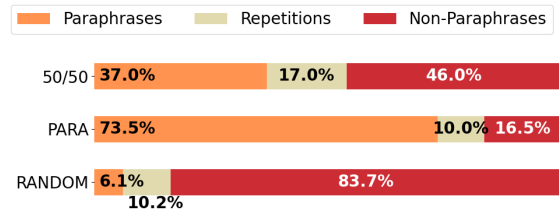

Figure 3: **Distribution of Labels by Lead Author.** We display the estimated number of (non-)paraphrases from the lead author annotations for the random subsample (RANDOM), the BALANCED sample and the wider paraphrase variety sample (PARA). Note, RANDOM consists of 100 elements, however only 98 are included in this statistic here (leading to numbers like 6.1). 2 pairs were not classified by the lead author because they were too ambiguous or were missing context information to reach a decision. We exclude such pairs in all other samples.

comparing it against the lead author classification, see "acc." column.

RANDOM. The random set is a sample of 100 (guest, host)-pairs that was uniformly sampled from the second batch of lead author annotations (§B.2).

PARA. After selecting the RANDOM set, the PARA set of 400 (guest, host)-pairs was sampled to reach a specified total 350 paraphrases and 150 non-paraphrases together with the RANDOM set.[27] The PARA set was selected to make the total number of non-repetition paraphrases together with RANDOM reach 300, while limiting the amount of repetition paraphrases to 50. Conversely, non-paraphrases were sampled to add up to 150. This led to 66 non-paraphrases and 334 paraphrases being sampled for the PARA set.

---

[27]RANDOM and PARA were undergoing annotation together in a second annotation round, after BALANCED had already been annotated. The aim was to reach a higher distribution of paraphrases in our released dataset. The 350/150 split was somewhat arbitrary. It could have easily been 400/100 or 300/200 as well.

C Annotations

C.1 Development of Annotator Training.

The eventual study design used in this work (see §5) is the product of iterative improvement with lab members, other volunteers and `Prolific` annotators. They iterative steps can roughly be separated into:

(1) The lead author repeatedly annotated the same set of (guest, host)-pairs with a time difference of one week. See an example of early self-disagreement in Table 15.

(2) With insights from (1) and our definition of context-dependent paraphrases, we created annotator instructions. We iteratively improved instructions while testing them with volunteers, lab members and `Prolific` crowd-workers. See examples of disagreements that led to changes in Table 15.

(3) Based on insights from (2), we introduced an intermediate annotator training that explains paraphrase annotation in a "hands-on" way: Annotators have to correctly annotate a teaching example to get to the next page instead of just reading an instruction. As soon as the correct selection is made, an explanation is show (e.g., Figures 6 and 10). After some testing rounds, we also require annotators to pass 2 attention (see Figure 12) as well as 2 comprehension checks (see Figures 5 and 11).

(4) We test the developed training on a selection of 20 (guest, host)-pairs out of which 10 were classified as clearly containing a paraphrase, and 10 as containing no paraphrase by the lead author, half of all examples we considered to be more difficult to classify (e.g., paraphrase with a low lexical overlap, non-paraphrase with a high lexical overlap). Two lab members reached pairwise Cohen of 0.51 after receiving training. Two newly recruited `Prolific` annotators reached average pairwise Cohen of 0.42 after going through training. Due to the inherent difficulty of the task and the good annotation quality when manually inspecting the 20 examples for each annotator, we carry on with this training setup.

C.2 Annotator Training.

We train participants to recognize paraphrases (see Figure 4–13 for the instructions they received). We presented (guest, host)-pairs with their MediaSum summaries, the date of the interview and the in-

terviewer names for context.[28] Participants were only admitted to the paraphrase annotation if they passed two attention checks (see Figure 12) and two comprehension checks (see Figure 5 and 11).

Comprehension Checks. Similar to examples in Table 2, they are presented with a clear paraphrase pair (App. Figure 5) and a less obvious context-dependent paraphrase pair (App. Figure 11) that they have to classify as a paraphrase. Additionally, they are only allowed to highlight the text spans that are a part of the paraphrase.

Training Stats. Of the initial 347 `Prolific` annotators who started the training, 95 aborted the study without giving a reason[29] and 126 were excluded from further studies because they failed at least one comprehension (29%) or attention check (24%) during training. Since annotators can perform annotations after training over a span of several days, we further exclude single annotation sessions, where the annotator fails any of two attention checks.

---

[28]The additional information of summary, date and speaker names increased reported understanding of context and eased difficulty of the task in pilot studies among lab members.

[29]Usually quickly, we assume that they did not want to take part in a multi-part study or did not like the task itself.

Who?	Example	see Instructions
Self-Disagreement	Guest: [..] So there was a <u>consensus organization last year that people from genetics and ethics law got together and said, in theory, it should be acceptable to try this in human beings.</u> The question will be, how much safety and evidence do we have to have from animal models before we say it's acceptable. Host: When it comes to this issue, let's face it, <u>while there are the concerns here in the United States</u>, it's happening in other countries.	**(C)** distinguish paraphrases from inferences, conclusions or "just" highly related utterances
Lab Members	Guest: Hey, it's going to be a long and a long week, and we're <u>going to</u> use every single minute of it to make sure that Americans know that Al Gore and Joe Lieberman are fighting for working families, right here in Los Angeles and across America. Host: And are you guys ready <u>to go</u>?	**(P)** short subselections of tokens might be "paraphrases" that do not adequately represent the content of the guest's utterance
	Guest: [...] There are militant groups out there firing against <u>the military</u>. And we just - we really don't know who is whom. Host: Why did <u>the army</u> decide today to move in and clear out the camp?	
	Guest: Police have indicated that they have been getting cooperation from the people involved, of course, they are looking at all of <u>her personal relationships</u> to see if there were any problems there. [...] Host: Well what have <u>family members</u> told you? I know you've talked to various <u>members of her family</u>. I understand she never missed her shifts at the restaurant where she worked. [...]	
	Guest: Yes, it is, all $640,000. Host: <u>That's a lot of dough.</u>	**(CD)** emphasize situational aspect to annotators, **(H)** ask for token-level accuracy of highlights
Prolific Annotators	Guest: [...] He was an employee that worked downtown Cleveland and saw it fall out of the armored car carrier, <u>and pick it up, and took it, and placed it in his car</u>. Host: And <u>he's been holding it ever since?</u>	similar to **(C)**
	Guest: [...] <u>Would I ever thought that this would be happening, no, it is, it's crazy? Just enjoy the moment.</u> Host: [...] , Magic Johnson was saying that when he first started taking meetings with investors or with business people, <u>they didn't take him seriously, but he thought maybe they just wanted his autograph.</u> [...]	**(AT)** use annotator screening to throw out annotators more likely to produce nonsensical pairs
	Guest: [...] they say, you, you must sue "Fortnite", <u>and I'm like, "Fortnite", what is that? I don't even know what it is</u> – Host: <u>So you weren't even familiar?</u>	**(AT)** throw out annotators that do not select obvious pairs

Table 15: **Examples of Disagreements in Paraphrase Annotation Pilots.** All of the presented examples were <u>highlighted</u> by at least one annotator and selected as not showing any paraphrases at all by at least one other annotator. We show examples from three different conditions: Self-disagreement for the lead author, disagreements between volunteers/lab members and disagreements between Prolific annotators. These disagreements informed later training instructions: For (C), see Figure 6; for (P), see Figure 9; for (CD), see Figure 10; for (H), see Figure 8; for (AT), we chose the separate training setup with attention and comprehension checks, see Figures 5, 11 and 12. Early on, we chose to include repetitions in our paraphrase definition since it turned out to be conceptually difficult to separate the two – especially in a context-dependent setting (e.g., is "You don't know." a repetition of "I do not know it." or not?), see Figure 4.

In every question, you will see a news interview summary, a statement said by the interview guest, followed by a reply from the interview host taken from a real news interview. We ask you to highlight the guest and host text for paraphrases.

For example:

Summary: Fresh Prince Star Alfonso Ribeiro Sues Over Dance Moves; Rapper 2 Milly Alleges His Dance Moves were Copied.

Guest: I guess it was season 5 when they premiered it in the game. A bunch of DMs, a bunch of Twitter requests, e-mails, everything was like, you, your game is in the dance, you need to sue, "Fortnite" stole it. Even like big artists, major artists like Joe Buttons and stuff, they have their own like show, daily struggle, they say, you, you must sue "Fortnite", and I'm like, "Fortnite", what is that? I don't even know what it is --

Host: So you weren't even familiar?

A **Paraphrase** is a **rewording or repetition** of content in the guest's statement. It rephrases what the guest said. You highlight the paraphrase in orange color.

We also ask you to highlight in red the section of the guest statement that the host paraphrases.

Figure 4: **Annotator Training (1).** Definition Paraphrase

The following is an example of a task we ask you to perform. Please select whether there is a paraphrase present according to the definition we presented to you on the previous page. If there is, we ask you to highlight the position of the paraphrase in the guest and host utterance. This is a __comprehension check__. You have 2 tries to get it right.

Guest (REP. RAUL LABRADOR (R), IDAHO)

Highlight Referred-to text here:

That's what we have been asking the president. We would like the senators to actually come and negotiate with us. So I think that would be a terrific idea.

Host (BLITZER)

Highlight Paraphrases here:

You say you want to negotiate, but what about the debt ceiling? Are you ready to see that go up without any strings attached, as the president demands?

In the reply, the host ...

- ◉ is paraphrasing something specific the guest says.
- ○ is not paraphrasing anything specific the guest said.

Figure 5: **Annotator Training (2).** Comprehension Check Paraphrase. Variations of the the shown highlighting are accepted.

Guest:

If you're not in our country, there are no constitutional protections for you.

Host:

So, you don't have a problem with Facebook giving the government access to the private accounts of people applying to enter the U.S.?

In the reply, the host ...

- ○ is paraphrasing something specific the guest says.
- ◉ is not paraphrasing anything specific the guest said.

You are correct, there is no paraphrase. Instead, **the host adds a conclusion to what the guest said** ("so you don't have a problem with ..."), **without rewording or repeating the content of the guest's utterance.** In a real news interview, the host most often says something related to the guest's statement that is not a paraphrase.

Figure 6: **Annotator Training (3).** Related but not a Paraphrase

---

Guest

Highlight Referred-to text here:

She was very pleasant. Talked about family life. They chatted about errands they need to run and things like that.

Host

Highlight Paraphrases here:

Well, she talked a lot about her family and her kids. And you get a personal sense for how they're living day by day.

In the reply, the host ...

- ◉ is paraphrasing something specific the guest says.
- ○ is not paraphrasing anything specific the guest said.

Great, your highlights capture what is paraphrased. **The host paraphrases two different sentences uttered by the guest.** You might have noticed that there is some ambiguity in what words exactly one can highlight as part of the paraphrase. More on that on the next page.

Figure 7: **Annotator Training (4).** Multiple Sentences.

There is some ambiguity in what words exactly one can highlight as being the target of the paraphrase in the guest statement or as being part of the paraphrase in the host reply. For example, looking back at the example from the beginning, you might have chosen to highlight:

Guest: Even like big artists, major artists like Joe Buttons and stuff, they have their own like show, daily struggle, they say, you, you must sue "Fortnite", and I'm like, "Fortnite", what is that? I don't even know what it is --
Host: So you weren't even familiar?

or

Guest: Even like big artists, major artists like Joe Buttons and stuff, they have their own like show, daily struggle, they say, you, you must sue "Fortnite", and I'm like, "Fortnite", what is that? I don't even know what it is --
Host: So you weren't even familiar?

or

Guest: Even like big artists, major artists like Joe Buttons and stuff, they have their own like show, daily struggle, they say, you, you must sue "Fortnite", and I'm like, "Fortnite", what is that? I don't even know what it is --
Host: So you weren't even familiar?

All these ways of highlighting are perfectly reasonable and acceptable. Don't worry about these ambiguities too much. We ask you to select what **you** think is a good representation of (1) what the host tried to paraphrase and (2) the actual paraphrase. We do ask you to be accurate on word level (e.g., a word like "So" might not be a relevant part of the paraphrase here but "you" and "I" should probably be present as they refer to the object of discussion)

Figure 8: **Annotator Training (5).** Highlighting

Guest (TONY HANSON, KYW NEWSRADIO):

Police have indicated that they have been getting cooperation from the people involved, of course, they are looking at all of her personal relationships to see if there were any problems there.

Host (PHILLIPS, host):

I know you've talked to various members of her family. What did they tell you?

In the reply, the host ...

○ is paraphrasing something specific the guest says.
◉ is not paraphrasing anything specific the guest said.

Correct. The host is not paraphrasing the guest. What makes it difficult is that "personal relationships" and "members of her family" could be seen as semantically related. However, here, **the host is not repeating the actual content of the guest's statement** (the guest talks about the police while the host talks about the guest, who is a reporter)**. The host rather continues on with a related topic.**

Figure 9: **Annotator Training (6).** Partial vs actual paraphrase

Sometimes it is helpful to have a bit more context information to understand whether something is a paraphrase. Next to the names of the speakers in brackets, we also provide a summary of the interview. Let's look at an example where that is helpful.

Given an interview on 2005-05-26 with the summary:

Lionel Tate is back in jail for allegedly holding up a pizza deliveryman. The 18-year-old Florida youth was the youngest person to be sent to prison for life in U.S. history. At age 12, he was accused of murdering his 6-year-old neighbor and friend Tiffany Eunick when he claimed he was demonstrating wrestling moves. Ed Gordon talks with Sgt. DeLacy Davis from New Jersey, a mentor and one of several supporters that put together a "re-entry into society plan" for Tate.

Guest and Host say the following:

Guest (DE LACY DAVIS):

Highlight Referred-to *text here:*

I think that, God willing, and then certainly if we were given another shot at this apple, I think the entire group would be amenable to shipping him here to me, which is what we felt would be a better environment to give him a new start. People, places and things needed to be changed, and consistently changed, and the plan adjusted based upon how he was faring.

Host (ED GORDON, host)

Highlight Paraphrases *here:*

So that would be, actually, coming to New Jersey and being under the auspices, frankly, of De Lacy Davis.

In the reply, the host ...

◉ is paraphrasing something specific the guest says.
○ is not paraphrasing anything specific the guest said.

Great. **The two highlighted sections describe the same content. Even though the perspective changes** ("here" -> "New Jersey", "me" -> "De Lacy Davis"). Here, the **name of the guest and the summary** contains the necessary information.

Figure 10: **Annotator Training (7).** Using context information

Given an interview on 2000-8-3 with the summary:

Kissinger: Bush is Fully Qualified for Foreign Policy Decisions

Guest and Host say the following:

Guest (HENRY KISSINGER, FMR. SECRETARY OF STATE)

Highlight `Referred-to` *text here:*

No, I haven't talked to him but I've talked to his secretary and we've passed messages back and forth between the family and me. And they tell me he's improved a lot. And `I'm to see him` tomorrow morning.

Host (JOHN YDSTIE, host)

Highlight `Paraphrases` *here:*

I'm sure that will cheer `him` up, to `have a visit from you,` and the doctors did say, just a short while ago, he's expected make a full recovery.

In the reply, the host ...

- ⦿ is paraphrasing something specific the guest says.
- ◯ is not paraphrasing anything specific the guest said.

Figure 11: **Annotator Training (8).** Example of an accepted answer for the comprehension check at the end. Only annotators who highlighted similar spans are admitted to annotate unseen instances. Some of the admitted annotators additionally selected the pair "he's improved a lot" and "he's expected to make a full recovery".

The following is an attention check. Please select "is not paraphrasing anything specific the guest said."

Guest:

And so that's the main question I've been asking people here, is, was the price worth it?

Host:

You're telling me people on the ground don't see it that way.

In the reply, the host ...

- ◯ is paraphrasing something specific the guest says.
- ◯ is not paraphrasing anything specific the guest said.

The following is an attention check. Please select "is paraphrasing something specific the guest says". Then highlight exactly: (1) "I'm like, "Fortnite", what is that? I don't even know what it is" in the guest utterance and (2) "you weren't even familiar?" in the host reply.

Guest:

I guess it was season 5 when they premiered it in the game. A bunch of DMs, a bunch of Twitter requests, e-mails, everything was like, you, your game is in the dance, you need to sue, "Fortnite" stole it. Even like big artists, major artists like Joe Buttons and stuff, they have their own like show, daily struggle, they say, you, you must sue "Fortnite", and I'm like, "Fortnite", what is that? I don't even know what it is --

Host:

So you weren't even familiar?

In the reply, the host ...

- ◯ is paraphrasing something specific the guest says.
- ◯ is not paraphrasing anything specific the guest said.

Figure 12: **Annotator Training (10).** Two attention checks shown at different times during training.

	Short Description	Example
In the Guest Statement		
Highlight	Referred-To is the target of the paraphrase	see below
In the Host Reply		

Highlight	Paraphrases reword or repeat what the guest said
	Subtleties
	   - several ways of highlighting are acceptable
		Guest: they say, you, you must sue "Fortnite", and I'm like, "Fortnite", what is that? I don't even know what it is --
		   or: **Guest:** *they say, you, you must sue "Fortnite", and I'm like, "Fortnite", what is that? I don't even know what it is --*
		Host: So you weren't even familiar?

		Guest: That's what we have been asking the senators to actually come and negotiate with us. **Host:** I know you want to negotiate ...
		Guest: She was very pleasant. Talked about family life. They chatted about errands they need to run and things like that. **Host:** Well, she talked a lot about her family and her kids. And you get a personal sense for how they're living day by day.
	- can include perspective shifts (here -> New Jersey, I -> you)	**Guest (De Lacy Davis):** I think the entire group would be amenable to shipping him here to us, which is what we felt would be a better environment to give him a new start. **Host:** So that would be, actually, coming to New Jersey and being under the auspices, frankly, of De Lacy Davis.
Do Not Highlight	additional content like consequences, inferences or facts	**Guest:** if you're not in our country, there are no constitutional protections for you **Host:** So, you don't have a problem with Facebook giving the government access to the private accounts of people applying to enter the U.S.? **Guest:** ... And so that's the main question I've been asking people here, is, was the price worth it? **Host:** You're telling me people on the ground don't see it that way.
	vague references or (short) terms that are only related in content but are not actually rephrasing the guest's point	**Guest:** ... they are looking at all of her personal relationships to see if there were any problems there. **Host:** Well what have family members told you? I know you've talked to various members of her family.

Figure 13: **Annotator Training (9).** Overview Table shown to annotators

C.3 Annotation After Training.

Next, the trained annotators were asked to highlight paraphrases. See Figure 14 for an example of the annotation interface. Annotators had access to a summary of their training at all times, see Figure 13. We again included two attention checks. Answers failing either attention check are removed from the dataset.

C.4 Annotator Allocation Strategy

To the best of our knowledge, what constitutes a "good" number of annotators per item has not been investigated for paraphrase classification.

Summary. Based on the 20–21 annotations per item for the BALANCED set, we simulate fixed and dynamic strategies to recruit up to 20 annotations per item. We evaluate the different strategies w.r.t. closeness to the annotations of all 20–21 annotators. When considering resource cost and performance trade-offs, dynamic recruitment strategies performed better than allocating a fixed number of annotators for each item.

Details. We consider three different strategies for allocating annotators to an item: (1) using a fixed number for all items, (2) for each item, dynamically allocate annotators until n of them agree and (3) similar to Engelson and Dagan (1996), for each item, dynamically allocate annotators until the entropy is below a given threshold t or a maximum number of annotators has been allocated. We simulate each of these strategies using the annotations on BALANCED. We evaluate the strategies on (a) cost, i.e., the average number of annotators per item and (b) performance via (i) the overlap between the full 20 annotator majority vote (i.e., we assume this is the best possible result) and the predicted majority vote for the considered strategy and (ii) k-rater-reliability (Wong and Paritosh, 2022) — a measure to compare the agreement between aggregated votes. Note, for the dynamic setup we change the original calculation of kRR (Wong and Paritosh, 2022) by dynamically recruiting more or less annotators per item and thus aggregating the votes of a varying instead of a fixed number of annotators.

Results. See Figure 15 for the results. We selected a practical resource limit of an average 8 annotators per items and the requirement of at least 90% accuracy with the majority vote and 0.7 kRR (dotted lines). We decide on strategy (3) dynamically recruiting annotators (minimally 3, maxi-

Given an interview on 2011-12-13 with the summary:

Senators Exchange Gifts for Secret Santa

Guest and Host say the following:

Guest (JOE JOHNS, CNN SENIOR CORRESPONDENT)

Highlight Referred-to *text here:*

This is not good.

Host (BALDWIN)

Highlight Paraphrases *here:*

This is what you don't want happening with your menorah, folks.

In the reply, the host ...

◉ is paraphrasing something specific the guest said.
○ is not paraphrasing anything specific the guest said.

Figure 14: **Interface for highlighting categories.** Annotators are asked to highlight the categories on word level.

mally 15) until entropy is below 0.8. Also with other min/max parameters this was a good trade-off between accuracy, kRR and average # of annotators. The average number of annotators needed per item is then about 6.8. In this way, most items receive annotations from 3 annotators, while difficult ones receive up to 15.

C.5 Annotator Payment.

Via Prolific's internal screening system, we recruited native speakers located in the US. Payment for a survey was only withheld if annotators failed two attention checks within the same survey or when a comprehension check at the very beginning of the study was failed[30] in line with Prolific guidelines.[31] Across all `Prolific` studies performed for this work (including pilots), we payed participants a median of $8.98 £/h \approx 11.41\$/h$[32] which is above federal minimum wage in the US.[33]

---

[30]Technically, in line with Prolific guidelines, we do not withhold payment but ask annotators to "return" their study in this case. Practically this is the same, as all annotators did return such a study when asked.

[31]Prolific Attention and Comprehension Check Policy

[32]on March 20th 2024

[33]Federal minimum wage in the US is $7.25/h \approx 5.71 £/h$ according to https://www.dol.gov/agencies/whd/minimum-wage on March 20th 2024
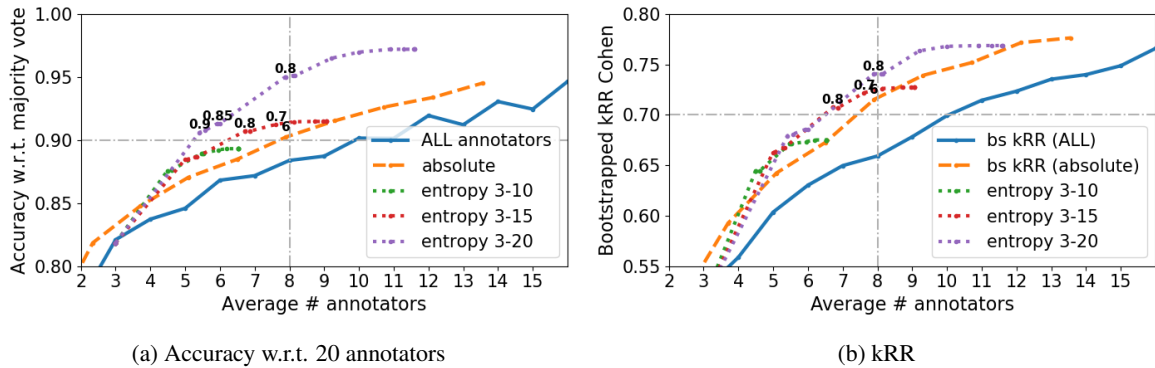
(a) Accuracy w.r.t. 20 annotators

(b) kRR

Figure 15: **Annotator Recruitment Strategies.** To decide the number of annotators for a specific item, we test three different strategies: (1) using a fixed number of annotators across all items (ALL), (2) increasing the number of annotators until at least n annotators agree for each item (absolute) and (3) increasing the number of annotators from 3 until the entropy is smaller than a given threshold (entropy) or a maximum of 10, 15 or 20 annotators is reached. We display the accuracy of the methods compared to using all 20 annotations in (15a) and the reliability measure kRR depending on the average number of annotators used (Wong and Paritosh, 2022) in (15b). We set a maximum average cost of 8 annotators per item and require a minimum accuracy of 90% as well as a minimum kRR of 0.70. When a strategy fulfills these requirements (i.e., falls in the upper left quadrants for (a) and (b)), we display the entropy thresholds for (3) and absolute number of annotators for (2).
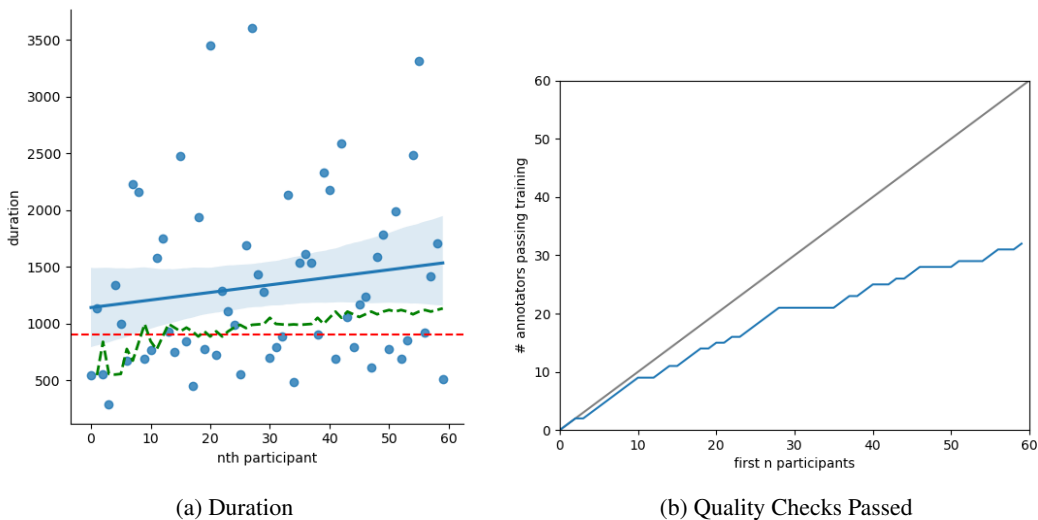

(a) Duration

(b) Quality Checks Passed

Figure 16: **On BALANCED, later training sessions take longer and pass fewer quality checks.** In 16a, we display the seconds the nth annotator needs to go through the training session. The annotators are ordered according to the dates they completed training. Annotations were distributed across 6 different days in June 2023. The green line represents the median duration time of the first n participants. The red line displays the initially estimated completion time of 900 seconds according to pilot studies. The blue line is a linear regression estimate of the duration and it's 95% confidence interval. On average, participants participating on a later date need more time to finish. In 16b, we display the summed number of the first n participants that passed the quality checks during training. The grey line represents the angle bisector, i.e., if every participant would pass all quality checks. Later participants are less likely to pass the quality checks.

C.6 Varying Annotator Behavior over Time.

For the BALANCED set, we performed separate training and annotation rounds. See Figure 16 for the completion times and share of passed quality checks of `Prolific` annotators in the training session. Participants that were recruited later performed worse: they pass less quality checks and need more time. This effect was noticeable but it is not quite clear to us why this happens. We recruit all participants at once for later studies and not iteratively as for the BALANCED set, to avoid effects that have to do with study age. The effect on the quality of the released annotations should be minimal as we discard annotators that do not

pass our quality checks. It does have an effect on the pay per hour for our participants, which we had initially estimated to be much higher.

C.7 Intra-Annotator Annotations Quality

We manually randomly sample ten annotators (with anonymized PROLIFIC ids 60, 6, 86, 84, 47, 31, 68, 88, 41, 92) and analyze 42 of their annotatations. Nine annotators consistently provide plausible annotations, while the other annotator chooses "not a paraphrase" a few times too often. We also noticed some other annotator-specific tendencies, for example, one annotator might tend to highlight fewer words, more words or prefer exact lexical matches.

C.8 Anonymization

We replace all `Prolific` annotator IDs with non-identifiable IDs. We only make the non-identifiable IDs public.

D   Modeling

D.1   In-Context Learning

Models.   We provide the Huggingface URLs to our used models.   Vicuna 7B: `https://huggingface.co/lmsys/vicuna-7b-v1.5`, Mistral 7B Instruct v0.2: `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2`, Openchat: `https://huggingface.co/openchat/openchat-3.5-0106`, Gemma 7B: `https://huggingface.co/google/gemma-7b-it`, Mixtral 8x7B Instruct v0.1: `https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1`, Llama 7B: `https://huggingface.co/meta-llama/Llama-2-7b-hf`   and   Llama 70B: `https://huggingface.co/meta-llama/Llama-2-70b-hf`.

Prompt.   We use a few-shot prompt that is close to the original annotator training and instructions, see Figure 17.   We use chain-of-thought like explanations, i.e., always starting with "Let's think step by step." and ending with "Therefore, the answer is", (Kojima et al., 2022) and a few-shot setup showing all 8 examples showed to annotators during training (Figures 4–12). For `GPT-4`, we use a temperature of 1, self-consistency through prompting the model 3 times (Wang et al., 2022b) and the default top_p nucleus sampling value of 1, a maximum of new tokens to 512.   For all the huggingface models, we use a temperature of 1, self-consistency through prompting the model 10 times (only 3 times for `Lllama 70B` due to resource limits) and a top_k sampling of the top 10 tokens, a maximum of new tokens of 400 for all other models. Note, there are many more prompts and choices we could have tried that are out-of-the scope of this work. Further steps could have included separating the classification and highlighting task, experimenting with further phrasings and so on. We leave this to future work.

D.2   Token Classification

We use settings very close to Wang et al. (2022a) and test different learning rates and number of epochs with 3 different seeds each.   We use the "save best model" option to save the model after the epoch which yielded the best result on the dev set. For the results, see Figure 16. We use a learning rate of 3e-3 and 12 epochs for further modeling.

Learning Rate	Epoch	F1
1e-3	8	0.61 ± 0.04
<u>3e-3</u>	8	0.64 ± 0.06
5e-3	8	0.52 ± 0.15
3e-3	4	0.65 ± 0.07
3e-3	<u>12</u>	**0.65** ± 0.00
3e-3	16	0.60 ± 0.10

Table 16: **Hyperparameter tuning on the DEV set.** We train a token classifier for learning rates 1e-3, 3e-3, 5e-3 and epochs 4, 8, 12 and 16 for 3 seeds. We keep learning rate fixed at 3e-3 when varying the number of epochs and epoch fixed at 8 when varyig the learning rates. Best options of learning rate and epoch are underlined. Best F1 score is **boldfaced**.

```
A Paraphrase is a rewording or repetition of content in the guest's statement. It rephrases what the guest said.
Given an interview on – with the summary: Fresh Prince Star Alfonso Ribeiro Sues Over Dance Moves; Rapper 2 Milly
    Alleges His Dance Moves were Copied.
Guest and Host say the following:
Guest (TERRENCE FERGUSON, RAPPER): I guess it was season 5 when they premiered it in the game. A bunch of DMs, a bunch
    of Twitter requests, e-mails, everything was like, you, your game is in the dance, you need to sue, "Fortnite"
    stole it. Even like big artists, major artists like Joe Buttons and stuff, they have their own like show, daily
    struggle, they say, you, you must sue "Fortnite", and I'm like, "Fortnite", what is that? I don't even know what it
    is --
Host (QUEST): So you weren't even familiar?
In the reply, does the host paraphrase something specific the guest says?

Explanation: Let's think step by step.
Terrence Ferguson says at the end of his turn that he didn't know Fortnite.
Quest, the host of the interview, repeats that the guest doesn't know Fortnite.
So they both say that the guest didn't know Fortnite. Therefore, the answer is yes, the host is paraphrasing the guest.
Verbatim Quote Guest: "I'm like, "Fortnite", what is that? I don't even know what it is"
Verbatim Quote Host: "you weren't even familiar?"
Classification: Yes.


Given an interview on 2013-10-1 with the summary: ...
Guest and Host say the following:
Guest (REP. RAUL LABRADOR (R), IDAHO): ...
Host (BLITZER): ...
In the reply, does the host paraphrase something specific the guest says?

Explanation: Let's think step by step. EXPLANATION Therefore, the answer is yes, host is paraphrasing the guest.
Verbatim Quote Guest: "We would like the senators to actually come and negotiate with us."
Verbatim Quote Host: "you want to negotiate"
Classification: Yes.


ITEM

Explanation: ...
Verbatim Quote Guest: None.
Verbatim Quote Host: None.
Classification: No.

ITEM

Explanation: ...
Verbatim Quote Guest: "She" "Talked about family life." "errands they need to run and things like that."
Verbatim Quote Host: "she talked" "about her family and her kids." "how they're living day by day."
Classification: Yes.

ITEM

Explanation: ...
Verbatim Quote Guest: None.
Verbatim Quote Host: None.
Classification: No.


ITEM

Explanation: ...
Verbatim Quote Guest: None.
Verbatim Quote Host: None.
Classification: No.


ITEM

Explanation: ...
Verbatim Quote Guest: "shipping him here to me"
Verbatim Quote Host: "coming to New Jersey and being under the auspices" "of De Lacy Davis."
Classification: Yes.


ITEM

Explanation: ...
Verbatim Quote Guest: "I'm to see him."
Verbatim Quote Host: "him" "have a visit from you"
Classification: Yes.


Given an interview on DATE with the summary: SUMMARY
Guest and Host say the following:
Guest (NAME): UTTERANCE
Host (NAME): UTTERANCE
Explanation: Let's think step by step.
```

Figure 17: **Prompt Template close to Annotator Instructions** The used prompt template is based closely on our annotator training and instructions. Phrasings were adapted to match the prompt-setting but kept the same where possible. See the full prompt in our Github Repository.

D.3 Highlighting Analysis

We compare the highlights provided by DeBERTa AGGREGATED[34] and DeBERTa ALL[35] on 10 text pairs from the test set that were classified as paraphrases by both models. We provide examples in Table 17. DeBERTa ALL highlights are shorter, often more on point and arguably more consistent than DeBERTa AGGREGATED highlights. We also manually analyzed 10 text pairs from the test set that GPT-4 classified as paraphrases. We provide examples of GPT-4 highlights in Table 18. Generally, they seem of good quality, but have the tendency to span complete sub-sentences, even if not all is relevant.

Hallucinations. One of the biggest problems for in-context learning are the extractions of the highlighting from the model responses which has errors in up to 71% of the cases in Table 8. Most of these errors can be split into two categories: (1) inconsistent highlighting, where the model classifies a paraphrase but does not highlight text spans in both, the guest and host utterance and (2) hallucinations, where the model highlights spans that do not exist in that form in the guest or host utterance. Hallucination is more prevalent than inconsistent highlighting for GPT-4, where in most cases it leaves out words (e.g., "coming back to a normal winter" vs. "coming back daryn to a normal winter"), in some other cases it adds or replaces words (e.g.,"he's a counterpuncher" vs. "he's counterpuncher"), uses morphological variation (e.g., "you've" vs. "you have") or quotes from the wrong source (e.g., from the host when considering the guest utterance). Most of these extraction errors seem to be resolvable by humans when looking at them manually, so it might be possible to address them in future work with a more advanced matching algorithm or by querying GPT-4 until one gets a parsable response. When looking at the classifications by GPT-4 they often seem plausible, even when counted as incorrect with the F1 score.

D.4 Computing Infrastructure

The fine-tuning of 18 DeBERTa token classifier, and the inference of 7 generative models took about approximately 260 GPU hours with one A100 card with 80GB RAM on a Linux computing cluster.

We use scikit-learn 1.2.2 (Pedregosa et al., 2011), statsmodels 0.14.1 (Seabold and Perktold, 2010) and krippendorff 0.6.1 (Castro, 2017) for evaluation.

E Use of AI Assistants

We used ChatGPT and GitHub Copilot for coding, to look up commands and sporadically to generate functions. Generated functions are marked in our code. Generated functions were tested w.r.t. expected behavior. We did not use AI assistants for writing.

---

[34]i.e., seed 202 with F1 score of 0.76, precision of 0.72 and recall of 0.84, see https://huggingface.co/AnnaWegmann/Highlight-Paraphrases-in-Dialog

[35]i.e., seed 201 with F1 score of 0.72, precision of 0.84 and recall of 0.63, see https://huggingface.co/AnnaWegmann/Highlight-Paraphrases-in-Dialog-ALL

AGG	ALL	C	Shortened Examples
✓	✓	✗	**G:** There are people that are in that age range where we know they're high risk, why are they going to the supermarket to **buy** their **own groceries?** Get the community, the neighborhood to go and help them. **H:** if you're going to help somebody by **helping them maybe get their groceries,** how long does the coronavirus live on surfaces?
✓	✓	✓	**G:** And people always prefer, of course, to see the pope as the principal celebrant of the mass. So that's good. That'll be tonight. And it will be his 26th mass and it will be the 40th or, rather, the 30th time that this is offered in round the world transmission. And it will be my 20th time in doing it as a television commentator from Rome so. **H:** Yes, **you've been doing this for a while** now.
✓	✓	✓	**G:** Well, what happened was we finally waved down a Coast Guard helicopter. And what they were looking for were people with disabilities and medical conditions, which none of us really had. They didn't lift any of us into the helicopter or anything. What they told us was to basically **walk** out of our house, up **the** street, trying to **fight against the current that was** going the **opposite** way of where we needed to go. **H:** So **you walked through that current to get to the higher ground** or get to a drier spot?
✓	✗	✗	**G:** They've now spent $6 million on this Benghazi investigation. They keep coming up with more and more interviews. **H:** On Benghazi, Trey Gowdy now says your committee has interviewed 75 witnesses.

Table 17: **DeBERTa ALL vs DeBERTa AGGREGATED highlights.** Paraphrase highlights predicted by the best **DeBERTa ALL** (i.e., seed 201 with F1 score of 0.72) and the best DeBERTa AGG model (i.e., seed 202 F1 score of 0.76, same as in the main paper). Even though DeBERTa AGG gets better F1 scores on classification, the DeBERTa ALL highlights are arguably more on point. For comparison, we also display the human highlights if they exist. Note, highlights can exist even if the crowd majority vote did not predict a paraphrase.

GPT-4	C	Shortened Examples
✓	✗	**G:** We also want to see what connections exist between pardons and potential gifts to the Clinton Library. **H:** Congressman, short of, though, having a thank-you note attached to a check that went to the Clinton Library, what is it exactly that is going to prove that there was a quid pro quo, that these pardons were actually bought?
✓	✗	**G:** They've now spent $6 million on this Benghazi investigation. They keep coming up with more and more interviews. **H:** On Benghazi, Trey Gowdy now says your committee has interviewed 75 witnesses.
✓	✓	**G:** [Trump] is appointing very young judges. **H:** [...] if you're 50-plus, you're probably too old for the Trump Administration to be seriously considered for a district court judgeship.

Table 18: **GPT-4 highlights.** Paraphrase highlights predicted by GPT-4. For comparison, we also display the human highlights if they exist. Note, highlights can exist even if the crowd majority vote did not predict a paraphrase.