# DIS2DIS: Explaining Ambiguity in Fact-Checking

**Ieva Raminta Staliūnaitė** and **Andreas Vlachos**
Department of Computer Science and Technology, University of Cambridge, UK
{irs38,av308}@cam.ac.uk

## Abstract

Ambiguity is a linguistic tool for encoding information efficiently, yet it also causes misunderstandings and disagreements. It is particularly relevant to the domain of misinformation, as fact-checking ambiguous claims is difficult even for experts. In this paper, we argue that instead of predicting a veracity label for which there is genuine disagreement, it would be more beneficial to explain the ambiguity. Thus, this work introduces claim disambiguation, a text-editing task, to explain ambiguous claims in fact-checking. This involves specifying an interpretation that can then be unequivocally supported by the given evidence. We collect a dataset of 1501 such claim revisions and conduct experiments with sequence-to-sequence models. The performance is compared to a simple copy baseline and a Large Language Model baseline. The best results are achieved by employing Minimum Bayes Decoding, with a BertScore F1 of 92.22. According to human evaluation, the model successfully disambiguates the claims 72% of the time.

## 1 Introduction

Ambiguity is a property of language that allows utterances to have multiple possible meanings, serving communicative purposes such as efficiency (Piantadosi et al., 2012). However, it also causes some complications. Ambiguity is not always perceived by listeners or readers (Rodd, 2018), with interpretations depending on context and motivation (Voss et al., 2008), and implicit meanings are difficult to argue with (Henderson and McCready, 2018). Recent work has also indicated that ambiguity is difficult not only for humans, but NLP models too. Liu et al. (2023) observed that Large Language Models (LLMs) are not good at detecting ambiguity in language, including very large models fine-tuned on human feedback such as GPT-4 (OpenAI, 2023).

Cognitive science research has shown that underspecified statements can lend themselves to misinformation due to the human cognitive predisposition to powerful inferences with little evidence (Cimpian et al., 2010). Misinformation refers to claims that are verifiably non-factual, however many claims lie in between the true/false dichotomy, due to the inherent ambiguity in language (Uscinski and Butler, 2013; Adams et al., 2023). Even expert fact-checkers often disagree on the factuality of claims, mainly in cases with ambiguous or partially true claims (Lim, 2018). Fact-checking is a particularly interesting domain for studying ambiguity, since claims are often presented for fact-checking out of context. In addition, annotation disagreement in fact-checking has been shown to be largely caused by ambiguous language (Glockner et al., 2024). To illustrate, disagreement in the top example in Table 1 stems from the vagueness of the term *'power'*, which could mean *'political power'* or *'influence'*. Under the former interpretation the claim is refuted, however under the latter it is neutral with regard to the evidence. Recent work has also shown that labels alone are not sufficiently informative for end-users of automated fact-checking systems (Schlichtkrull et al., 2023a). Research on explainability in fact-checking provides explanations for the fact-checking labels (Kotonya and Toni, 2020; Stammbach and Ash, 2020; Krishna et al., 2022; Atanasova, 2024), however they do not focus on ambiguity.

In this work, we generate explanations for ambiguous claims, which have been largely understudied. We propose the disambiguation of a claim as an explanation for why its factuality may be debatable, in the paradigm of elaborative simplification (Srikanth and Li, 2021), positing that adding content can ease reasoning about the causal links in the text. In our context, the disambiguation makes the implicit interpretation that is supported by the evidence explicit. That is, a claim C is ambiguous,



*Findings of the Association for Computational Linguistics:*
*NAACL 2025*, pages 246–267

| | |
|---|---|
| S | **Original claim:** A Quiet Place has subtitles for the sign language.<br>**Evidence:**[...] Producers Andrew Form and Bradley Fuller said that they initially planned not to provide on-screen subtitles for sign-language dialogue while providing only "context clues," but they realized that subtitles were necessary for the scene in which the deaf daughter and her hearing father argue about the modified hearing aid. [...]<br>**Revised claim:** A Quiet Place has subtitles for the sign language. |
| R | **Original claim:** Gold is the highest an album can go.<br>**Evidence:** [...] In 1975, the additional requirement of 500,000 units sold was added for Gold albums. Reflecting growth in record sales, the Platinum award was added in 1976, for albums able to sell one million units, and singles selling two million units. The Multi-Platinum award was introduced in 1984, signifying multiple Platinum levels of albums and singles. Reflecting additional growth in music sales, the Diamond award was instituted in 1999 for albums or singles selling ten million units. [...]<br>**Revised claim:** Diamond is the highest an album can go. |
| A | **Original claim:** The king of Cambodia does have power.<br>**Evidence:** [...] Under the Constitution, the King has no political power, but as Norodom Sihanouk was revered in the country, his word often carried much influence in the government. [...]<br>**Revised claim:** The king of Cambodia has no political power, but has had influence |
| U | **Original claim:** No one died in the Tacoma Narrows Bridge collapse.<br>**Evidence:** [...] The weather system that caused the bridge collapse went on to cause the Armistice Day Blizzard that killed 145 people in the Midwest. [...] The Armistice Day storm and the strong winds that earlier had caused the Tacoma Narrows Bridge to oscillate, twist, and collapse into the waters below. [...]<br>**Revised claim:** It is not clear from the evidence whether anyone died in the Tacoma Narrows Bridge collapse. |

Table 1: Examples of S(UPPORTED), R(EFUTED), A(AMBIGUOUS) and U(NSUBSTANTIATED) claims in DIS2DIS.

because it would only be supported by the evidence if we take the rewrite C' as its interpretation. The disambiguation is not intended to represent the intention of the speaker.

The claim and evidence pair is the input, and the unambiguously supported revised claim is the expected output. Annotator disagreement is used as signal for item ambiguity. We collect the DIS2DIS (Disagreement to Disambiguation) dataset, with annotators labeling claims as SUPPORTED, REFUTED, AMBIGUOUS or UNSUBSTANTIATED by the evidence, and then revising the claims to be supported. Multiple rounds of revisions are needed to reach consensus on a claim being supported. Sequence-to-sequence (seq2seq) models are trained on the ensuing dataset. The best results are achieved with Minimum Bayes Risk (MBR) decoding (Freitag et al., 2022) for finding the disambiguations that represent the model consensus. Our best-performing model achieved 92.22 BertScore micro F1, and according to human evaluation, successfully disambiguates the claim 72% of the time [1].

## 2 Related Work

### 2.1 Linguistic Phenomenon: Ambiguity

Lexical ambiguity has been studied for decades (Bunescu and Pasca, 2006; Ide and Véronis, 1998; Mitkov, 2014; Ng and Cardie, 2002), and discourse information has been successfully integrated (Asher and Lascarides, 1995), disambiguation of entire discourses has not received as much attention. Some recent work has studied the linguistic phenomena that underpins ambiguity. Pragmatic inference has been proposed as a task either in its own right (Pandia et al., 2021; Nizamani et al., 2024), or as a by-product of other tasks such as natural language inference (Jeretic et al., 2020). Other work has focused on making implicit meanings explicit. Quan et al. (2019) perform ellipsis and coreference resolution in dialogue turns, essentially disambiguating utterances by making the omitted or referred expressions explicit. Choi et al. (2021) define the task of decontextualization, which consists of rewriting sentences to be interpretable out of context. Similarly, Wu et al. (2023) generate *Questions under Discussion (QUDs)* for sentences in dialogue to make explicit the underlying drivers of discourse, while Yu et al. (2023) edit loaded questions in order to remove implicit or explicit presuppositions, and Min et al. (2020) disambiguate questions in open-domain question answering. Some recent work has also explored the ability of LLMs to detect ambiguity, and improved their near-random performance by instruction-tuning, showing that this task can benefit from specialised data (Ruis et al., 2024). However, to the best of our knowledge, such discourse-expounding methods have not yet been applied in the context of fact-checking.

[1] https://github.com/ieva-raminta/Dis2Dis

### 2.2 Method: Text Editing

Text simplification and error correction relate to disambiguation as they use edits to clarify text. Both grammatical error correction and text simplification are often tackled with seq2seq or sequence-to-edit supervised training methods (Chandrasekar et al., 1996; Dahlmeier and Ng, 2012; Yuan and Briscoe, 2016; Al-Thanyyan and Azmi, 2021). Most simplification models do not generate elaborative simplifications, and those that do tend to hallucinate (Srikanth and Li, 2021). Factual error correction and claim-editing are also approached with seq2seq models (Cao et al., 2020), distant supervision (Thorne and Vlachos, 2021), and hypernetworks (Chen et al., 2023). The work in factual error correction has also replicated the limited binary factuality judgment framework, and is therefore limited to correcting REFUTED items to be SUPPORTED, without considering ambiguity.

### 2.3 Domain: Fact-Checking

Recent work has looked into the insufficiency of the SUPPORTED, REFUTED and NEUTRAL label scheme. Schlichtkrull et al. (2023b) add a category "conflicting evidence/cherry-picking" in order to characterise cases where the evidence provides reasons to both support and refute a claim. However, cherry picking is only one particular type of ambiguity, which bears an intentional connotation. Glockner et al. (2024) provide an analysis of the varied linguistic phenomena which cause disagreement over the traditional ternary labels, showing a statistically significant correlation between various pragmatic and discourse inference types, and annotator agreement over the labels. Consequently, they model the fact-checking task with soft labels, predicting a distribution rather than a single gold target, in order to account for the difference in interpretations of the ambiguous items. However, soft labels are not easily interpretable.

### 2.4 Aim: Explainability

In the field of explainability of fact-checking, different types of explanations have been proposed. Using saliency maps to indicate the most relevant parts of the input is the most straigtforward approach (Atanasova et al., 2022). Atanasova (2024) use the explanations provided by fact-checkers themselves as justification for their judgment. Similarly, Kotonya and Toni (2020) collect expert data and generate free-form explanations, including explanations for and against a given claim if the evidence is mixed between SUPPORTED and REFUTED. However, they do not separate ambiguous items from those that have conflicting evidence. Stammbach and Ash (2020) generate summaries of the evidence with regard to the given claim as explanations, and demonstrate their utility by predicting the veracity label from the summaries.

### 2.5 Data signal: Disagreement

Research on various NLP tasks has shown that disagreement over labels in classification tasks, as well as diversity of outputs in generation tasks, is informative of the difficulty of items (Uma et al., 2021), beneficial for training better models (Jiang and Marneffe, 2022), and valuable in evaluation (Pavlick and Kwiatkowski, 2019). However, disagreement has not been used as signal for disambiguation.

By and large, in the current paper we address the issues that have been raised by previous work, which have not been combined into one dataset as of yet, as summarised in Table 2.

| | S/R/A/U Distinction | Explanations | Ambiguity Explanations |
|---|---|---|---|
| (Thorne et al., 2018) | ✗ | ✗ | ✗ |
| (Stammbach and Ash, 2020) | ✗ | ✓ | ✗ |
| (Kotonya and Toni, 2020) | ✗ | ✓ | ✓ |
| (Thorne and Vlachos, 2021) | ✗ | ✗/✓ | ✗ |
| (Schlichtkrull et al., 2023b) | ✓ | ✗ | ✗ |
| (Glockner et al., 2024) | ✓ | ✗ | ✗ |

Table 2: Dataset Comparison

## 3 DIS2DIS: Disagreement to Disambiguation

This section presents the disambiguation task and dataset, from definition and data collection to quality evaluation.

### 3.1 Task Definition

The task of disambiguation is, given a claim and evidence, to generate a disambiguated claim that is fully supported by that evidence. The expected disambiguation is different depending on the relation between the original claim and the evidence. Claims can be SUPPORTED, REFUTED, AMBIGUOUS or UNSUBSTANTIATED by the evidence. If the claim is already SUPPORTED, then no changes are required, while REFUTED claims should be negated. The AMBIGUOUS class has items that could be either supported or refuted by the evidence depending on the interpretation, such as in the third example from the top in Table 1. If the claim is

ambiguous, the revision should lay out an interpretation of the original claim which is supported by the evidence, such as *"The king of Cambodia has no political power, but has had influence"* in this case. The UNSUBSTANTIATED class contains items where the evidence does not answer the Question Under Discussion (QUD) of the claim. For instance, the claim in the bottom row of Table 1 is UNSUBSTANTIATED, because while the evidence mentions the blizzard casualties, it does not specify whether anyone died in the bridge collapse. That is, the evidence does not answer the question *"Did anyone die at the Tacoma bridge collapse?"* Thus the disambiguation should state that "*It is not clear from the evidence whether the claim is true*".

### 3.2 Annotation Scheme

We collected a dataset for this task by using claims and evidence from the AmbiFC (Glockner et al., 2024) dataset, which reportedly had a high annotator disagreement due to ambiguity. In order to get as many ambiguous items as possible, we mostly select claims from AmbiFC with the highest entropy of labels, motivated by the relationship between label entropy and annotator certainty shown in (Glockner et al., 2024).

The annotations were collected using the Prolific platform.[2] The open-source annotation tool 'Potato' (Pei et al., 2022) was used to design the interface. The annotators were provided with explanations and examples of all the possible label classes and the expected respective disambiguations. The annotators are asked to select a label for the original claim, revise the claim, and highlight the parts of the input that they deem the most informative for the label they selected. The annotation guidelines are presented in Appendix A. In addition, a pre-tester question was used to ensure the annotators understood and followed the instructions. The annotators were asked to label the pre-tester item in the second row of Table 1 in order to take part in the annotation task.

The main task for the annotators was to revise the claim to be unambiguously supported by the evidence. Interestingly, many revisions for the pre-tester item paraphrased the following undesirable claims: *"Platinum is the highest an album can go"* (15%), *"Multi-Platinum is the highest an album can go"* (6%). This result shows that annotators were likely to stop reading once they reached the part of the evidence that was sufficient to reject the claim, namely the mention of the Platinum award, providing an insufficiently disambiguated revision. This provided an incentive to run multiple rounds of annotations of the same item by different annotators, as it indicated that single edits often do not suffice. A revised claim from the first annotator would be passed on to a second annotator as an original claim for a classification and disambiguation. This is repeated until an annotator labels the claim as SUPPORTED, which we take to mean that the claim has been fully disambiguated. If no consensus is reached after the third revision has been evaluated, we interpret this as an impossible disambiguation, therefore assigning it to the UNSUBSTANTIATED class. The iteration process is displayed in the flowchart in Figure 1. Some items are also annotated multiple times from scratch, in order to see the variation of disambiguations and acquire multiple references for a subset of the dataset.

The Sankey diagram in Figure 2 illustrates the paths through different labels that claims go through until a consensus is reached. A single edit is sufficient to disambiguate about half of the claims, however the remaining items require a few iterations until different annotators assign it the same label. The figure illustrates that the AMBIGUOUS class is particularly difficult to tease apart from UNSUBSTANTIATED, since multiple rounds of annotations are sometimes required to reach consensus on an ambiguous item.

[2] https://www.prolific.co/

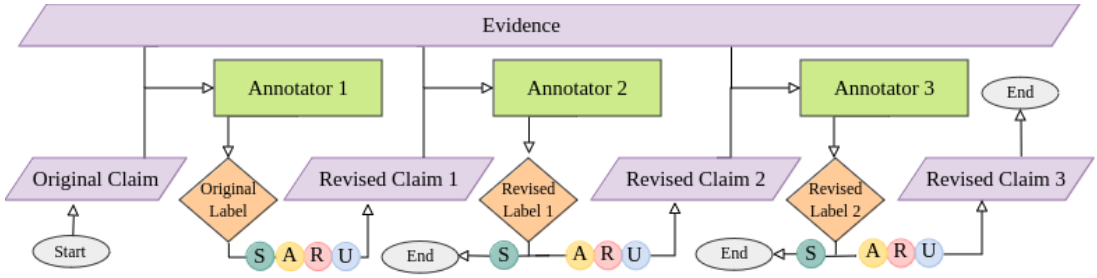


Figure 1: Flowchart illustrating multiple rounds of annotation.

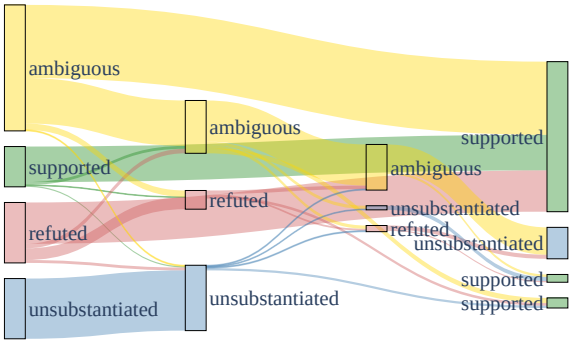


Figure 2: The labels assigned to claim revisions as they are iteratively edited.

### 3.3 DIS2DIS Dataset

To generate a dataset for the task of disambiguation, the original claim, any intermediate claims, the evidence and the final revised claim are put together to form an instance of the DIS2DIS dataset. If a single edit was sufficiently disambiguating, the original claim and the first edit form a *(source, target)* pair. Otherwise, in the case that the original claim is agreed on by more than one annotator as SUPPORTED, then the *(source, target)* pair is (original claim, original claim), while if the original claim is agreed on by more than one annotator as REFUTED, then it is (original claim, *"It is not true that "*+original claim). Alternatively, if multiple edits were required, the original as well as the intermediate claims are used as the *source* claim, while the final disambiguated claim is the *target*. Finally, if three edits still did not lead to agreement on the label, the original claim is treated as the *source*, while the *target* is formulated as *"It is not clear from the evidence whether "*+original claim.

The resulting dataset contains 1501 items (see Table 3 for dataset statistics, and Appendix F for the Dataset Datasheet). The split into training, development and test sets is performed by firstly retaining all the items with multiple references for the test set, and then applying stratified sampling to ensure that disambiguations that stem from the same AmbiFC (Glockner et al., 2024) claim or evidence do not get separated into different splits, ensuring no contamination of data from the training set in evaluation. The test set contains on average 1.48 references. The dataset has an 'AMBIguous' subset for experimenting only with items that are ambiguous, which contains 762 items that take at least one and no more than three edits to reach consensus on the veracity of the claim. This is the focus part of our study, however we include the other cases in the dataset to allow the model to learn different behaviors depending on the initial relationship of the claim and the evidence, which is not a given.

Table 4 illustrates the different types of ambiguities present in the dataset, and their respective disambiguations. The elaborations for disambiguating underspecified claims take the form of relative clauses and subordinate clauses (e.g. conditional or contrastive). On the other hand, hyponyms or shorter modifiers such as adjectives can be sufficient to disambiguate a vague claim.

### 3.4 Agreement and Evaluation Metrics

For evaluating the quality of the collected dataset, as well as selecting the best automatic metrics for training and evaluating models on the data, we perform a blind human evaluation on the generated disambiguations. Two annotators with graduate training in language sciences review a set of 27 original claims and 108 of their revised versions, labeling each as SUPPORTED, REFUTED, AMBIGUOUS or UNSUBSTANTIATED by the evidence. The instructions to the evaluators provide the same information as the original annotators to keep the annotation scheme consistent, apart from the *'unsubstantiated'* label. Due to the fact that the task of the evaluators is to judge the change between the original and revised claim, when asked about the revised claim the annotators are required to choose the UNSUBSTANTIATED label if the revised claim does not address the same QUD as *either* the evidence *or* the original claim. This difference is necessary due to the fact that disambiguations which drift away from the point being made in the original claim are not truly disambiguations, even if they are factual. The annotation guidelines for the human evaluation are presented in Appendix C. For example, if the claim in the final row of Table 1 is revised to read

| | Train | Dev | Test | Original Label | | | | Mean Claim Length | | Mean # of Revisions |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | S | R | A | U | Original | Revised | |
| AMBI | 537 | 64 | 161 | 71 | 206 | 403 | 82 | 12.5 | 18.2 | 1.35 |
| ALL | 1128 | 136 | 237 | 219 | 317 | 403 | 562 | 12.6 | 18.6 | 1.96 |

Table 3: DIS2DIS dataset statistics. Original Label corresponds to Original Label in Figure 1.

| Source | Claim | Evidence | Revised claim |
|---|---|---|---|
| **Vagueness** | eric clapton did sing knocking on heaven's door | "Knockin' on Heaven's Door" is a song by Bob Dylan, written for the soundtrack of the 1973 film Pat Garrett and Billy the Kid. [...] The song became one of Dylan's most popular and most covered post-1960s compositions, spawning covers from Guns N' Roses, Eric Clapton, Randy Crawford and more. [...] | eric clapton did sing a cover of knocking on heaven's door |
| **Under-specification** | you can get held back in 7th grade | Norway, Denmark and Sweden do not allow grade retention during elementary school and junior high school (1-10th grade). In the United Kingdom, a similar streaming system to New Zealand's is used (see above). Germany, Italy, Austria, Netherlands, France, Finland and Switzerland use grade retention. Greece allows grade retention if [...] | you can get held back in 7th grade in certain countries |
| **Pre-supposition** | there is such a thing as a bladder transplant | On January 30, 1999, scientists announced that a lab-grown bladder had been successfully transplanted into dogs. These artificial bladders worked well for almost a year in the dogs. In 2000, a new procedure for creating artificial bladders for humans was developed. [...] In 2006, the first publication of experimental transplantation of bioengineered bladders appeared in The Lancet. [...] | There is such a thing as a bladder transplant for dogs, experiments for humans are being run |
| **Implicature** | there is such thing as over drinking water | Marathon runners are susceptible to water intoxication if they drink too much while running. This is caused when sodium levels drop below 135 mmol/L when athletes consume large amounts of fluid. This has been noted to be the result of the encouragement of excessive fluid replacement by various guidelines. [...] | there is such thing as over drinking water, particularly in marathon runners |
| **Probabilistic enrichment** | something has been to the bottom of the ocean | On 26 March 2012, Canadian film director James Cameron made a solo manned descent in the DSV Deepsea Challenger to the bottom of the Challenger Deep. [...] At 07:52, Deepsea Challenger arrived at the bottom. The descent lasted 2 hours and 36 minutes and the recorded depth was 10,908 metres (35,787 ft) when Deepsea Challenger touched down. [...] | something has been to the bottom of the Challenger Deep |
| **Coreference** | there is a red light district in canada | There is no official red-light district, although the definition of the boundaries has varied according to both the source and the time period. [...] Prostitution, gambling and drinking were more prevalent in this area because of its proximity to the city centre, which is often a major tourist attraction, and the high density of liquor shops (taverns, bars, night clubs, cabarets, etc). | It is not clear from the evidence whether there is a red light district in canada |

Table 4: Examples of different sources of ambiguity and their disambiguations.

"*145 people died in the the Armistice Day Blizzard*", it no longer answers the question of whether anyone died in the bridge collapse, and therefore is not a true disambiguation of the original claim.

A heuristic combines judgments on individual claims into an overall score for the quality of the edit, as shown in Figure 3. The agreement between the two evaluators on their individual labels assigned to original and revised claims, as well as the binary score between 0 (not disambiguated or poorly disambiguated) and 1 (disambiguated), is measured with Cohen's $\kappa$. We observe substantial agreement at $\kappa$ values of 0.66 and 0.69 respectively.

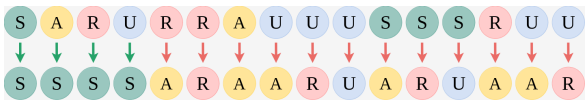


Figure 3: The overall score of 1 (green arrow) for disambiguated items, 0 (red arrow) for not disambiguated or poorly disambiguated, depending on the label of the original claim (top) and the revised claim (bottom) being S(UPPORTED), A(MBIGUOUS), R(EFUTED) or U(NSUBSTANTIATED).

The overall scores of the evaluators are then compared to automated metric scores in order to select the most appropriate metric for the task. The metrics commonly used in text generation tasks such as machine translation or text simplification are tested: ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), BertScore (Zhang et al., 2019), Comet (Rei et al., 2020) and SARI (Xu et al., 2016). Both neural and token-matching metrics are used, some of which support the inclusion of the source into the evaluation, which is valuable in a task such as disambiguation, where the original claim as well as the evidence text are relevant to evaluating the quality of the generated sequence. Table 5 presents the correlation scores for ALL items as well as the AMBIguous subset, using Pearson (Sedgwick, 2012) correlation coefficient. The correlation is strongest with BertScore values.

The neural metrics are better correlated with human judgments than the token-based metrics, especially for the AMBIguous subset. This is expected given that the ambiguous items have more nuanced edits which are less tied to exact token matches. It is interesting to note that BertScore Precision (p) stands out with a much higher correlation coefficient with human judgments for the AMBIguous subset compared to the rest of the items. This could

| | BLEU | ROUGE | | | | SARI | | | | BertScore | | | Comet |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | L | L-sum | mean | keep | add | delete | F1 (micro) | p | r | |
| Neural | ✗ | ✗ | | | | ✗ | | | | ✓ | | | ✓ |
| Source | ✗ | ✗ | | | | ✓ | | | | ✗ | | | ✓ |
| AMBI | | | | | | | | | | | | | |
| PCC | 0.41 | 0.52 | 0.48 | 0.52 | 0.52 | 0.41 | 0.44 | 0.33 | 0.31 | 0.54 | **0.58** | 0.44 | 0.55 |
| ALL | | | | | | | | | | | | | |
| PCC | 0.44 | 0.50 | 0.47 | 0.50 | 0.50 | 0.45 | 0.45 | 0.42 | 0.33 | **0.52** | **0.52** | 0.46 | 0.50 |

Table 5: Pearson's Correlation Coefficient (PCC) between automatic metrics and human judgment across ALL items and for the AMBIguous subset. 'Neural' marks scores which are based on neural models, and 'Source' marks whether the score takes the source input into account.

be explained by the fact that for items which are already unambiguous, adding false positive tokens (words which are not in the reference), might not affect the relationship between the claim and the evidence. That is, if the claim is already supported, adding irrelevant details to make the claim more specific does not affect the ambiguity or the factuality. On the other hand, for ambiguous items, adding the exact disambiguating details is key.

## 4 Experiments

The baselines and models trained to perform the disambiguation task as described in this section.

### 4.1 Baselines

#### 4.1.1 Copy Baseline

A common text editing baseline is copying the input as is. For items which are supported, the copied claims would be identical to the reference claims, while for other items they would be similar to the targets, as the disambiguations are comprised of a few token changes only. This could be expected to yield relatively high evaluation scores, especially on automatic metrics.

#### 4.1.2 LLM Baseline

We run a zero-shot and few-shot experiments with the Llama3, 8 billion parameter model (AI@Meta, 2024), in order to evaluate the out-of-the-box LLM performance on the task of disambiguation. For the few-shot scenario we provide the model with 4 or 8 examples, covering all 4 class types (SUPPORTED, REFUTED, AMBIGUOUS, UNSUBSTANTIATED). The model is given the instruction to "*please make the following claim less ambiguous with regard to the following evidence*." The examples are presented in random order, separated by newlines, and punctuated with *'Claim:'*, *'Evidence:'* and *'Revised Claim:'* tags. The reader may find the inputs in Appendix B. The model output is constrained to generate a single sentence by stopping generation at a newline token, replicating the way that the few-shot examples are fed to the model. The model is expected to perform well due its large size and large training set, however we hypothesize that the task is still hard enough for the model to incur some errors, given the scarcity of direct examples of disambiguation during training and the linguistic complexity of the ambiguity relations.

### 4.2 Model

To evaluate how well the collected data can serve for training seq2seq models, we finetuned a Flan-T5 base model with 250 million parameters (Chung et al., 2024). The model input is the same as for the LLM baseline. Additionally, decoding techniques are used to improve model performance by guiding it to select the specific tokens in the evidence which would help disambiguate a claim if added to the revised version. Length penalty, vocabulary forcing, and MBR (Freitag et al., 2022) are experimented with. The simplest of such methods is length penalty, which penalises the model for short generations. This is expected to improve results as disambiguations typically require an addition of a modifier, conditional clause or other specifying details, which make the reference length longer than the source. The method of vocabulary forcing restricts the decoder to a set of tokens. Using vocabulary forcing on this model leverages the fact that the modifying phrases needed for disambiguation can be generally found in the evidence. We therefore constrain the generation to include at least one of the tokens that appears in the evidence but does not appear in the source claim.

Finally, the application of MBR to this task is inspired by the idea that disambiguation is tied to decreasing disagreement, which is reflected in the way the dataset was collected. Intrinsic uncertainty and ambiguity are related to the inadequacy of the mode sought by greedy and beam search decoding

(Stahlberg et al., 2022). The MBR method generates a number of hypothesis sequences as well as pseudo references, and uses a utility function to find the best hypothesis. In our case, the best hypothesis would be the one that would reach highest agreement amongst humans, therefore we try to find the hypothesis which has the highest BertScore value when compared to the pseudo references.

The model is trained on a single NVIDIA TU102 GPU with batch size 8, for a maximum of 30 epochs, using early stopping by monitoring the BertScore metric, which has the highest correlation with human judgment. A hyperparameter search is performed (please see Appendix E).

## 5 Results and Analysis

### 5.1 Results

Table 6 presents the best single run results of the models described in Section 4.2 after the hyperparameter search, and the baselines from Sections 4.1.1 and 4.1.2. The copy baseline predictions receive the highest scores on automatic metrics, however a careful inspection of the outputs of the models indicates that this result is not representative of the real ranking. The length of the generations also indicates a discrepancy between the appropriateness of the generation and its BertScore values. Length penalty, vocabulary forcing and the 0-shot LLama3-8B model all overshoot the target by generating lengthy claims which are not actually helpful disambiguations. The models perform relatively on par across the different classification labels, with the largest differences between approaches seen in the 'ambiguous' class.

We perform a human evaluation on a random subset of 50 test items with crowdworkers on the Prolific platform. The annotators are asked whether the revised claim is a good disambiguation of the original claim. The annotation guidelines are presented in Appendix D. As unreliability of the automatic metrics for evaluating disambiguations is corroborated by the results on the test set in Table 7 as well, which shows that the ranking order based on automatic metrics does not match the ranking order of human evaluation at all. Interestingly, the models trained on DIS2DIS perform better on the AMBIguous dataset than overall, exhibiting specialised knowledge, while the LLM baseline shows the reverse. The annotator agreement is 0.56, measured with Cohen's $\kappa$. All models perform statistically significantly better than the copy baseline and worse than humans, as shown in Table 8. The largest difference in performance between the models appears in the 'ambiguous' class, as shown in the breakdown in Table 9 .

### 5.2 Analysis

Table 10 presents different generations to the same original claim containing underspecification. The LLama3-8B model baseline fails to disambiguate the claim, leaving it as is. Human and MBR model

| | | Copy Baseline | Llama3-8B 0-shot | Llama3-8B 4-shot | Llama3-8B 8-shot | Flan-T5-250M Base | Flan-T5-250M Length Penalty | Flan-T5-250M Vocab Forcing | Flan-T5-250M MBR | Human | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bert | F1 (micro) | 94.17 | 91.39 | 93.29 | **94.42** | 93.98 | 92.72 | 92.76 | 93.36 | 100 | AMBI |
| Score | p | **95.50** | 91.16 | 94.52 | 95.28 | 94.92 | 92.42 | 93.00 | 93.88 | 100 | |
| len (tokens) | | 12.5 | 86.53 | 14.53 | 15.95 | 15.86 | 56.08 | 19.89 | 16.30 | 17.33 | |
| Bert | F1 (micro) | 94.38 | 90.91 | 94.39 | 94.25 | 94.13 | 94.18 | 93.59 | **94.81** | 100 | ALL |
| Score | p | **95.99** | 91.23 | 95.79 | 95.27 | 93.66 | 93.52 | 93.13 | 94.87 | 100 | |
| len (tokens) | | 12.6 | 86.01 | 15.82 | 21.46 | 21.61 | 24.39 | 45.13 | 18.74 | 18.43 | |

Table 6: Model performance and baseline scores on the development set, for the AMBIguous subset and ALL items.

| | Copy Baseline | Llama3-8B 8-shot | Flan-T5 Base | Flan-T5 MBR | Human |
|---|---|---|---|---|---|
| | | AMBI | | | |
| Bert F1 | **93.85** | 93.12 | 93.51 | 92.22 | 97.57 |
| Score p | **95.49** | 93.47 | 94.35 | 92.36 | 97.51 |
| Human | 0.10 | 0.60 | 0.62 | **0.74** | 0.85 |
| | | ALL | | | |
| Bert F1 | **94.10** | 93.30 | 93.07 | 93.05 | 98.05 |
| Score p | **95.65** | 94.01 | 93.19 | 93.24 | 98.03 |
| Human | 0.27 | 0.67 | 0.60 | **0.72** | 0.82 |

Table 7: Model performance and baseline scores with human evaluation on ALL items in the test set, and its AMBIguous subset.

| Model 1 | Model 2 | BertScore | | | Human Evaluation | | |
|---|---|---|---|---|---|---|---|
| | | t-statistic | p-value | <0.05 | t-statistic | p-value | <0.05 |
| Base | Human | -2.46 | 0.015 | ✓ | -17.46 | $5.03e-53$ | ✓ |
| Llama3-8B | Human | -3.52 | 0.000 | ✓ | -17.88 | $5.66e-55$ | ✓ |
| MBR | Human | -1.68 | 0.094 | ✗ | -18.52 | $5.60e-58$ | ✓ |
| Llama3-8B | Base | -1.03 | 0.306 | ✗ | 0.71 | 0.466 | ✗ |
| MBR | Base | 0.77 | 0.445 | ✗ | -0.04 | 0.965 | ✗ |
| MBR | Llama3-8B | 1.8 | 0.073 | ✗ | -0.79 | 0.43 | ✗ |
| Base | Copy Baseline | -3.74 | 0.000 | ✓ | 8.277 | $9.65e-15$ | ✓ |
| Llama3-8B | Copy Baseline | -2.92 | 0.004 | ✓ | 6.90 | $4.80e-11$ | ✓ |
| MBR | Copy Baseline | -3.51 | 0.000 | ✓ | 9.37 | $6.56e-18$ | ✓ |
| Human | Copy Baseline | 16.82 | $4.47e-50$ | ✓ | 11.96 | $4.78e-26$ | ✓ |

Table 8: The results of the T-test statistical significance test comparing different disambiguation methods.

| Metric | Class | Base | MBR | Llama3-8B | Copy Baseline | Human |
|---|---|---|---|---|---|---|
| BertScore | supported | 94.40 | 93.66 | 93.87 | 95.58 | 98.67 |
| | refuted | 93.00 | 93.03 | 93.31 | 93.79 | 98.14 |
| | ambiguous | 91.99 | 92.35 | 92.53 | 93.85 | 97.35 |
| | unsubstantiated | 94.36 | 94.05 | 94.42 | 93.97 | 98.91 |
| | all | 93.07 | 93.05 | 93.30 | 94.10 | 98.05 |
| Human Eval | supported | 0.90 | 0.90 | 0.85 | 1.00 | 0.90 |
| | refuted | 0.44 | 0.44 | 0.69 | 0.06 | 0.75 |
| | ambiguous | 0.57 | 0.67 | 0.36 | 0.10 | 0.79 |
| | unsubstantiated | 0.82 | 0.86 | 0.77 | 0.09 | 0.86 |
| | all | 0.60 | 0.72 | 0.67 | 0.27 | 0.82 |

Table 9: The breakdown of the results by class.

| | | |
|---|---|---|
| **Original claim:** You can name your kid anything in America.<br>**Evidence:** [...] Traditionally, the right to name one's child or oneself as one chooses has been upheld by court rulings and is rooted in the fourteenth Amendment and the First Amendment, but a few restrictions do exist. Several states limit the number of characters that can be used. A few states ban the use of obscenity. Restrictions vary by state, Kentucky for instance, has no naming laws whatsoever.[...] | | |
| **Revised claims:** | Llama3-8B 8-shot: | You can name your kid anything in America. |
| | Flan-T5 MBR: | You can name your kid anything in America, but restrictions exist. |
| | Human: | You can name your kid anything in Kentucky, while other states have some restrictions on length or obscenity. |

Table 10: Example target, baseline and model outputs.

generations both provide suitable disambiguations, where the nuance of restrictions to naming are mentioned in both, with the human generation providing a more detailed explanation. This represents the general tendency observed in qualitative analysis, with MBR model generations providing better disambiguations than other models and baselines, however not reaching the full potential of human revisions.

Based on a qualitative analysis, the most common errors for all models include a) not changing the claim at all when a revision is required, b) mixing up the types of edits needed for the 'unsubstantiated' and the 'ambiguous' classes, c) hallucinating details not present in the evidence, d) missing or superfluous negation. The Base model exhibits the highest number of a), b) and d) type errors, while the Llama3-8B baseline suffers the most from c).

## 6 Discussion and Conclusion

The results of this study provide evidence that ambiguity is difficult to detect and remove for humans as well as language models. We argue that the fact that humans find detecting ambiguity and disambiguation difficult, calls for work on disambiguation. Apart from the application to explainability in fact-checking, disambiguation could also be applied to assisting in writing less ambiguously, or providing less ambiguous summaries.

Future research may involve experimenting with multi-step disambiguation as well as exploring the utility of highlighted inputs for model training. Future directions could also include exploring the link between disagreement and ambiguity by directly using disagreement as feedback for disambiguations through reinforcement learning strategies.

## Limitations

Our approach is limited in handling certain types of ambiguity, namely the ones which are prominent in the fact-checking data we used: underspecification, vagueness, implicature, presupposition, probabilistic enrichment, coreference. This may not cover other types of ambiguity that could be more common in different domains. In addition, we only focused on English due to dataset availability. Our work was limited to claims with evidence from Wikipedia, however fact-checking and ambiguity are pervasive in various platforms of communication. This study is also limited to the fact-verification step of fact-checking, studying the impact of ambiguity when the evidence is given. This setup only handles a single piece of gold evidence for the disambiguation step, which should also be expanded in future work.

The study shows that automatic evaluation metrics are not reliable in evaluating the performance of different methods of disambiguation. As a result, a human evaluation is required, which is labour-intensive and time-consuming. In addition, the LLM baseline performance depends on the prompts used.

We recognise the potential risk that a disambiguation dataset, when misused, could be used to obscure rather than clarify claims, which could contribute to the spread of misinformation. We believe, however, that the benefits of learning about misinformation and ambiguity detection outweigh the drawbacks.

## Ethics Statement

The annotators in this study were selected on the basis of residing in the UK and being native English language speakers, and were paid an hourly rate above the minimum wage in the UK (£11.44), averaging at £13.28. The annotation protocol was approved by an ethics review board. The annotation instructions contained a disclaimer that the topics appearing in the claims in the study would contain content comparable to what one might encounter while browsing the internet, as the claims are sourced from common search engine queries (Clark et al., 2019). No personal information of the annotators was collected.

## Acknowledgments

The data collection in this work was supported through a gift from Google. Ieva Raminta Staliūnaitė is supported by Huawei. Andreas Vlachos is supported by the ERC grant AVeriTeC (GA 865958). The authors would like to thank Rui Cao, Rami Aly, Diana Galvan, Gabrielle Gaudeau, Yuan Gao, Michael Sejr Schlichtkrull, Georgi Karadzhov, Eric Chamoun, Nedjma Djouhra Ousidhoum, Yulong Chen, Mubashara Akhtar, Guy Emerson and Zebulon Goriely for their help with the design of the annotation scheme and insightful feedback.

## References

Zoë Adams, Magda Osman, Christos Bechlivanidis, and Björn Meder. 2023. (why) is misinformation a problem? *Perspectives on Psychological Science*, 18(6):1436–1463.

AI@Meta. 2024. Llama 3 model card.

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Nicholas Asher and Alex Lascarides. 1995. Lexical disambiguation in a discourse context. *Journal of Semantics*, 12(1):69–69.

Pepa Atanasova. 2024. Generating fact checking explanations. In *Accountable and Explainable Methods for Complex Reasoning over Text*, pages 83–103. Springer.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Diagnostics-guided explanation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 10, pages 10445–10453.

Razvan Bunescu and Marius Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *11th conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258.

Raman Chandrasekar, Christine Doran, and Srinivas Bangalore. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Jiangjie Chen, Rui Xu, Wenxuan Zeng, Changzhi Sun, Lei Li, and Yanghua Xiao. 2023. Converge to the truth: Factual error correction via iterative constrained editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 11, pages 12616–12625.

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Andrei Cimpian, Amanda C. Brandone, and Susan A. Gelman. 2010. Generic statements require little evidence for acceptance but have powerful implications. *Cognitive science*, 34 8:1452–1482.

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

Daniel Dahlmeier and Hwee Tou Ng. 2012. Better evaluation for grammatical error correction. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–572.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. Ambifc: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Robert Henderson and Elin McCready. 2018. How dogwhistles work. In *New Frontiers in Artificial Intelligence: JSAI-isAI Workshops, JURISIN, SKL, AI-Biz, LENLS, AAA, SCIDOCA, kNeXI, Tsukuba, Tokyo, November 13-15, 2017, Revised Selected Papers 9*, pages 231–240. Springer.

Nancy Ide and Jean Véronis. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational linguistics*, 24(1):1–40.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models imppressive? learning implicature and presupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705.

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.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. Proofver: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Chloe Lim. 2018. Checking how fact-checkers check. *Research & Politics*, 5(3):2053168018786848.

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

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Ruslan Mitkov. 2014. *Anaphora resolution*. Routledge.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 104–111.

Rashid Nizamani, Sebastian Schuster, and Vera Demberg. 2024. Siga: A naturalistic nli dataset of english scalar implicatures with gradable adjectives. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14784–14795.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 367–379.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the*

*40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. 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.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Steven T. Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. Gecor: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. 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 4547–4557.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Jennifer Rodd. 2018. Lexical ambiguity. *Oxford handbook of psycholinguistics*, pages 120–144.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2024. The goldilocks of pragmatic understanding: Fine-tuning strategy matters for implicature resolution by llms. *Advances in Neural Information Processing Systems*, 36.

Michael Schlichtkrull, Nedjma Ousidhoum, and Andreas Vlachos. 2023a. The intended uses of automated fact-checking artefacts: Why, how and who. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8618–8642.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023b. AVeritec: A dataset for real-world claim verification with evidence from the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Philip Sedgwick. 2012. Pearson's correlation coefficient. *Bmj*, 345.

Neha Srikanth and Junyi Jessy Li. 2021. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137.

Felix Stahlberg, Ilia Kulikov, and Shankar Kumar. 2022. Uncertainty determines the adequacy of the mode and the tractability of decoding in sequence-to-sequence models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8634–8645, Dublin, Ireland. Association for Computational Linguistics.

Dominik Stammbach and Elliott Ash. 2020. e-fever: Explanations and summaries for automated fact checking. *Proceedings of the 2020 Truth and Trust Online (TTO 2020)*, pages 32–43.

James Thorne and Andreas Vlachos. 2021. Evidence-based factual error correction. 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 3298–3309.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

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.

Joseph E Uscinski and Ryden W Butler. 2013. The epistemology of fact checking. *Critical Review*, 25(2):162–180.

Andreas Voss, Klaus Rothermund, and Jochen Brandtstädter. 2008. Interpreting ambiguous stimuli: Separating perceptual and judgmental biases. *Journal of Experimental Social Psychology*, 44(4):1048–1056.

Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. 2023. Qudeval: The evaluation of questions under discussion discourse parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5344–5363.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

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

Zheng Yuan and Ted Briscoe. 2016. Grammatical error correction using neural machine translation. In

*Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–386.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

# A Annotation Guidelines for Data Collection

## A.1 Instructions

**Procedures** In this study, you will be presented with pairs of claims and evidence, and your task will be to (1) label the the given claim as 'supported', 'refuted', 'ambiguous' or 'unsubstantiated' with regard to the provided evidence, (2) highlight parts of the text that support your choice, and (3) edit the claim to make the claim match the supported label better. You should make your decisions based on the information provided more than on your world knowledge. You can expect a larger portion of the provided items to be ambiguous, so please read carefully.

**Risks** The risks and discomfort associated with participation in this study are no greater than those ordinarily encountered in daily life, such as when surfing the internet.

**Benefits** There may be no personal benefit from your participation in the study but the knowledge gained may have academic or industrial value.

**Confidentiality** By participating in this research, you understand and agree that the researcher may be required to disclose your consent form, data, and other personally identifiable information as required by law, regulation, subpoena, or court order. Otherwise, your confidentiality will be maintained in the following manner: To protect your identity, the researchers will take the following steps: (1) Each participant will be assigned a number; (2) The researchers will record any data collected during the study by number, not by name; (3) Any original recordings or data files will be stored in a secured location accessed only by authorized researchers.

**Voluntary Participation** Your participation in this research is voluntary. You may discontinue participation at any time during the research activity.

**Navigating** You can use the right arrow to move forward, but you are not allowed to go backward. For highlighting text, multiple selections are allowed and encouraged, however they all have to correspond to the same veracity label that you have selected. If you wish to remove highlights that you have made, you can do so by clicking on them. The edits you are asked to do should be as minimal as possible, however they should not simply negate the original claim. The idea is to specify the claim more or change some details in the claim, to more closely match the 'supported' label.

## A.2 Annotation Scheme

Here are the explanations and examples of the 'supported', 'refuted', 'unsubstantiated' and 'ambiguous' labels, please read them carefully before moving forward.

**Supported** A claim is supported by its evidence if the evidence is sufficient to draw the conclusion that the claim is true. For instance, the evidence "The NATO summit will be hosted in Vilnius, Lithuania to discuss Ukraine" supports the claim "The NATO summit will be held in Eastern Europe ".

When highlighting the important parts of the input, you could emphasize the georgraphical references to the country and the larger region it belongs to.

In order to match the 'supported' label even better, the claim could be edited to read "The NATO summit will be held in Vilnius, which is in Eastern Europe" in order to remove any uncertainty about the georgraphical classification of the country. The edited claim is now even more supported by the evidence, because it clarifies the location of Vilnius for the readers who may not be aware of it.

**Refuted** In contrast, a claim is refuted by its evidence if the evidence is sufficient to draw the conclusion that the claim is false. For instance, the claim "Ukraine has a timeline for joining NATO " is refuted by the evidence stating that "Ukraine will not be offered timeline for NATO membership at the summit in July ".

When highlighting the relevant parts of the claim and evidence, you may want to consider what makes the claim and evidence contrast, such as the different time references and the negation.

The claim can be edited to match the supported label as such: "Ukraine's ~~has a~~ timeline for joining NATO has not been determined yet."

**Unsubstantiated** Alternatively, if the claim

is neither supported nor refuted by the evidence, the evidence may not provide enough information to draw a conclusion. For instance, the evidence "The NATO summit will be hosted in Vilnius, Lithuania to discuss Ukraine" is not enough to determine the veracity of the claim "Ukraine has a timeline for joining NATO". While the claim and the evidence discuss the same topic, the evidence here does not provide any answer as to whether the claim is true or false, therefore it should be marked as unsubstantiated.

In the case of an unsubstantiated claim, it would be good to highlight the parts of the input that refer to different aspects of the topic, such as the location of the summit in the evidence, and the NATO membership timeline in the claim.

The claim could be rewritten to be supported by stating that "The status of Ukraine's ~~has a~~ timeline for joining NATO is not clear from the evidence"

**Ambiguous**

In contrast, there is an ambiguous relationship between the claim that "Ukraine's application to join NATO is being supported" and the evidence that says "France resolves to support Ukraine's NATO membership bid". The claim is partially true, as the application is supported by some countries, but it is not known whether it is supported by everyone. The generic statement in the claim is too broad.

In order to show the source of ambiguity, you should highlight the parts of the input that make the claim vaguer than the evidence, such as specifically naming France in the evidence in this case.

In order to match the claim to the 'supported' label, the claim could be rewritten as "Ukraine's application to join NATO is being supported by France.", as this removes the ambiguity from the original claim by specifying the country.

### A.3 Examples

The examples shown to the annotators are shown in Figures 4 and 5.

## B Instruction and Examples for LLM Baseline

### B.1 0-shot Instruction and Input Format

"*Please make the following claim less ambiguous with regard to the following evidence. Claim:* [CLAIM], *Evidence:* [EVIDENCE], *Revised Claim:"*

### B.2 4-shot Instruction, Examples and Input Format

"Please make the following claim less ambiguous with regard to the following evidence, as in the examples below.

Claim: bridges of madison county is a true story.

Evidence: The Bridges of Madison County (also published as Love in Black and White) is a 1992 best-selling romance novella by American writer Robert James Waller that tells the story of a married Italian-American woman (WW2 War bride) living on a Madison County, Iowa, farm in the 1960s. While her husband and children are away at the State Fair, she engages in an affair with a National Geographic photographer from Bellingham, Washington, who is visiting Madison County to create a photographic essay on the covered bridges in the area. The novel is presented as a novelization of a true story, but it is in fact entirely fictional. The novel is one of the bestselling books of the 20th century, with 60 million copies sold world-wide. It has also been adapted into a feature film in 1995 and a musical in 2013.

Revised claim: bridges of madison county is a fictional story

Claim: you can keep a gray wolf as a pet.

Evidence: Some wildlife centers housing captive wolves prohibit handlers from entering wolf enclosures if they happen to have a cold or other vulnerability which the wolves can detect. Captive wolves are generally shy and avoid eye contact with humans other than their owner, as well as not listening to any commands made by any other humans. They usually vacate rooms or hide when a new person enters the establishment. Even seemingly friendly wolves need to be treated with caution, as captive wolves tend to view and treat people as other wolves, and will thus bite or dominate people in the same situation in which they would other wolves. Ordinary pet food is inadequate, as an adult wolf needs 12.5 kg (25 lbs) of meat daily along with bones, skin and fur to meet its nutritional requirements. Wolves may defend their food against people, and react violently to people trying to remove it. The exercise needs of a wolf exceed the average dog's demand. Because of this, captive wolves typically do not cope well in urban areas. Due to their talent at observational learning, adult

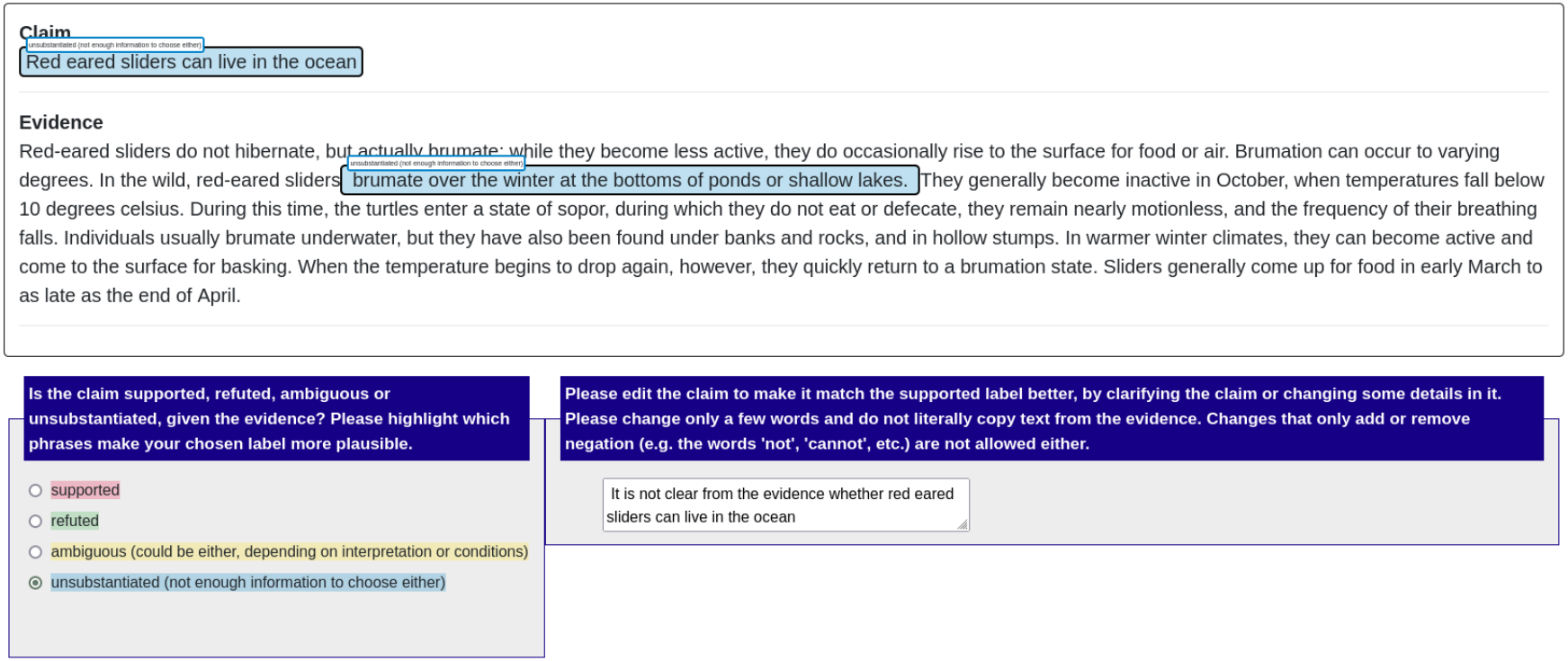


Figure 4: An example annotation with an ‘Unsubstantiated’ label.

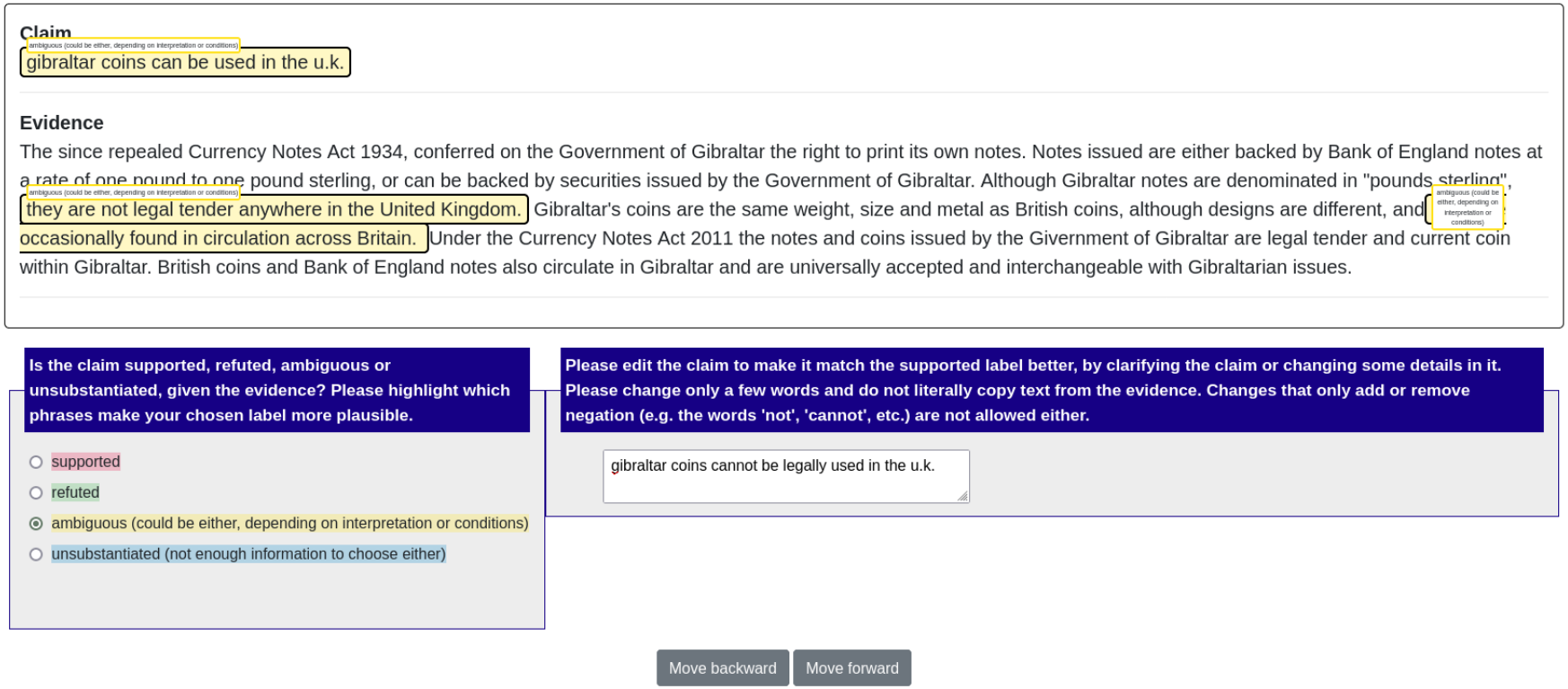


Figure 5: An example annotation with an 'Ambiguous' label.

captive wolves can quickly work out how to escape confinement, and require constant reinforcement by caretakers or owners, which makes raising wolves difficult for people who raise their pets in an even, rather than subordinate, environment.

Revised claim: it is difficult to raise a wolf as a pet.

Claim: it is illegal to flash your headlights to warn off the police in the uk.

Evidence: Though not all of its rules represent law, the Highway Code states "Only flash your headlights to let other road users know that you are there. Do not flash your headlights in an attempt to intimidate other road users". Drivers warning others about speed traps have been fined in the past for "misuse of headlights". Headlight flashing in the United Kingdom is often used as a signal that the driver flashing you is offering to let you go first. Such use is however strongly discouraged because it can lead to accidents where the driver flashing has not seen the approach of another road user. Using it to indicate that you are coming through and the other driver must wait, could lead to an accident. Drivers should also be aware of the so-called "Flash-for-Cash" scam, in which criminals flash their lights to let other drivers out of a junction, then crash into them on purpose in order to make fraudulent insurance claims for damage and whiplash injury.

Revised claim: In the UK, you should only flash your headlights to let other drivers know you are there.

Claim: you do need intent to commit a crime
Evidence: In criminal law, intent is a subjective state of mind that must accompany the acts of certain crimes to constitute a violation. A more formal, generally synonymous legal term is scienter: intent or knowledge of wrongdoing.
Revised claim: you do need intent to commit some crimes

Claim: [CLAIM], Evidence: [EVIDENCE], Revised Claim:"

### B.3 8-shot Instruction, Examples and Input Format

"Please make the following claim less ambiguous with regard to the following evidence, as in the examples below.

Claim: bridges of madison county is a true story.

Evidence: The Bridges of Madison County (also published as Love in Black and White) is a 1992 best-selling romance novella by American writer Robert James Waller that tells the story of a married Italian-American woman (WW2 War bride) living on a Madison County, Iowa, farm in the 1960s. While her husband and children are away at the State Fair, she engages in an affair with a National Geographic photographer from Bellingham, Washington, who is visiting Madison County to create a photographic essay on the covered bridges in the area. The novel is presented as a novelization of a true story, but it is in fact entirely fictional. The novel is one of the bestselling books of the 20th century, with 60 million copies sold world-wide. It has also been adapted into a feature film in 1995 and a musical in 2013.

Revised claim: bridges of madison county is a fictional story

Claim: you can keep a gray wolf as a pet.

Evidence: Some wildlife centers housing captive wolves prohibit handlers from entering wolf enclosures if they happen to have a cold or other vulnerability which the wolves can detect. Captive wolves are generally shy and avoid eye contact with humans other than their owner, as well as not listening to any commands made by any other humans. They usually vacate rooms or hide when a new person enters the establishment. Even seemingly friendly wolves need to be treated with caution, as captive wolves tend to view and treat people as other wolves, and will thus bite or dominate people in the same situation in which they would other wolves. Ordinary pet food is inadequate, as an adult wolf needs 12.5 kg (25 lbs) of meat daily along with bones, skin and fur to meet its nutritional requirements. Wolves may defend their food against people, and react violently to people trying to remove it. The exercise needs of a wolf exceed the average dog's demand. Because of this, captive wolves typically do not cope well in urban areas. Due to their talent at observational learning, adult captive wolves can quickly work out how to escape confinement, and require constant reinforcement by caretakers or owners, which makes raising wolves difficult for people who raise their pets in an even, rather than subordinate, environment.

Revised claim: it is difficult to raise a wolf as a pet.

Claim: it is illegal to flash your headlights to warn off the police in the uk.

Evidence: Though not all of its rules represent law, the Highway Code states "Only flash your headlights to let other road users know that you are there. Do not flash your headlights in an attempt to intimidate other road users". Drivers warning others about speed traps have been fined in the past for "misuse of headlights". Headlight flashing in the United Kingdom is often used as a signal that the driver flashing you is offering to let you go first. Such use is however strongly discouraged because it can lead to accidents where the driver flashing has not seen the approach of another road user. Using it to indicate that you are coming through and the other driver must wait, could lead to an accident. Drivers should also be aware of the so-called "Flash-for-Cash" scam, in which criminals flash their lights to let other drivers out of a junction, then crash into them on purpose in order to make fraudulent insurance claims for damage and whiplash injury.

Revised claim: In the UK, you should only flash your headlights to let other drivers know you are there.

Claim: you do need intent to commit a crime

Evidence: In criminal law, intent is a subjective state of mind that must accompany the acts of certain crimes to constitute a violation. A more formal, generally synonymous legal term is scienter: intent or knowledge of wrongdoing.

Revised claim: you do need intent to commit some crimes

Claim: running with scissors is based on a true story.

Evidence: In 2005, the family of Dr. Rodolph H. Turcotte (1919ˇ20132000), of Massachusetts filed suit against Burroughs and his publisher, alleging defamation of character and invasion of privacy. They stated that they were the basis for the Finch family portrayed in the book but that Burroughs had fabricated or exaggerated various descriptions of their activities. It's still a memoir, it's marketed as a memoir, they've agreed one hundred percent that it is a memoir. The case was later settled with Sony Pictures Entertainment in October 2006, prior to the release of the film adaptation. Burroughs and his publisher, St. Martin's Press, settled with the Turcotte family in August 2007. The Turcottes were reportedly seeking damages of $2 million for invasion of privacy, defamation, and emotional distress; the Turcottes alleged Running with Scissors was largely fictional and written in a sensational manner. Burroughs defended his work as "entirely accurate", but agreed to call the work a "book" (instead of a "memoir") in the author's note, to alter the acknowledgments page in future editions to recognize the Turcotte family's conflicting memories of described events, and express regret for "any unintentional harm" to the Turcotte family. Burroughs felt vindicated by the settlement. "I'm not at all sorry that I wrote [the book]. And you know, the suit settled - it settled in my favor. I didn't change a word of the memoir, not one word of it. It's still a memoir, it's marketed as a memoir, they've agreed one hundred percent that it is a memoir". Future printings of Running with Scissors will contain modified language in the Author's Note and Acknowledgments pages. Where the Acknowledgments page had read: "Additionally, I would like to thank each and every member of a certain Massachusetts family for taking me into their home and accepting me as one of their own," the following was substituted: "Additionally, I would like to thank the real-life members of the family portrayed in this book for taking me into their home and accepting me as one of their own. I recognize that their memories of the events described in this book are different than my own. They are each fine, decent, and hard-working people. The book was not intended to hurt the family. Both my publisher and I regret any unintentional harm resulting from the publishing and marketing of Running with Scissors"

Revised claim: running with scissors is somewhat based on the recollections of part of the author's life

Claim: you can drink at any age in wisconsin.

Evidence: The drinking age in Wisconsin is 21. Those under the legal drinking age may be served, possess, or consume alcohol if they are with a parent, legal guardian, or spouse who is of legal drinking age. Those age 18 to 20 may also possess (but not consume) alcohol as part of their employment. In the early 70s the sale of alcohol was reduced to the age of 18. The 1983 Wisconsin Act 74, effective July 1, 1984, created a drinking age of 19. Meeting in special session at the call of the governor, the legislature enacted 1985 Wisconsin

Act 337, which raised the drinking age to 21 and brought the state into compliance with the NMDA (National Minimum Drinking Age) on September 1, 1986. The NMDA law was amended to permit an exception for those persons who were between ages 18 and 21 on the effective date of the law. Wisconsin 19- and 20-year-olds were grandfathered in by this exception after enactment of Act 337. In effect, the state did not have a uniform age of 21 until September 1, 1988.

Revised claim: you can drink at any age in wisconsin with someone who is of legal drinking age.

Claim: it is normal for your second toe to be longer.

Evidence: Morton's toe is the condition of having a first metatarsal which is short in relation to the second metatarsal (see diagram). It is a type of brachymetatarsia. The distal metatarsal bones vary in relative length compared to the proximal. For most feet, a smooth curve can be traced through the joints at the bases of the toes (the metatarsal-phalangeal, or MTP, joints). But in Morton's foot, the line has to bend more sharply to go through the base of the big toe, as shown in the diagram. This is because the first metatarsal, behind the big toe, is short compared to the second metatarsal, next to it. The longer second metatarsal puts the MTP joint at the base of the second toe further forward. If the big toe and the second toe are the same length (as measured from the MTP joint to the tip, including only the toe bones or phalanges), then the second toe will protrude farther than the big toe, as shown in the photo.

Revised claim: your second toe can be longer than your big toe.

Claim: baby sign language is the same as regular sign language.

Evidence: Baby sign involves enhanced gestures and altered signs that infants are taught in conjunction with spoken words with the intention of creating richer parent-child communication. The main reason that parents use baby sign is with hope that it will reduce the frustration involved in trying to interpret their pre-verbal child's needs. It can be considered a useful method of communication in the early developmental stages, since speech production follows children's ability to express themselves through bodily movement. Baby sign is distinct from sign language. Baby sign is used by hearing parents with hearing children to improve communication. Sign languages, including ASL, BSL, ISL and others, are natural languages, typically used in the Deaf community. Sign languages maintain their own grammar, and sentence structure. Because sign languages are as complex to learn as any spoken language, simplified signs are often used with infants in baby sign. Teaching baby signs allows for greater flexibility in the form of sign and does not require the parent to learn the grammar of a sign language. Baby signs are usually gestures or signs taken from the sign language community and modified to make them easier for an infant to form.

Revised claim: baby sign language is distinct from regular sign language.

Claim: [CLAIM], Evidence: [EVIDENCE], Revised Claim:"

## C Annotation Guidelines for the Initial Human Evaluation

The following are examples of claims which might be ambiguous with regard to the given evidence (original claims). Attempts have been made to disambiguate the claims by editing them to be more supported by the evidence (edited claims). Please read the original claims, the evidence, and the edited claims, and assess whether the original claim is a) supported by the evidence, b) refuted by the evidence, c) ambiguous with regard to the evidence, or d) the evidence does not address the question that is implied in the original claim. Then, assess the revised claim with regard to the evidence and determine whether it is a) supported by the evidence, b) refuted by the evidence, c) ambiguous with regard to the evidence, or d) the revised claim does not address the same question as the original claim or the evidence.

For example, the claim "you can keep a gray wolf as a pet" is ambiguous with regard to the following evidence: "[...] Captive wolves are generally shy and avoid eye contact with humans other than their owner, as well as not listening to any commands made by any other humans. [...] Ordinary pet food is inadequate, as an adult wolf needs 12.5 kg (25 lbs) of meat daily along with bones, skin and fur to meet its nutritional requirements. [...] The exercise needs of a wolf exceed the average dog's demand. Because of this, captive wolves typically do not cope well in urban areas. Due to

their talent at observational learning, adult captive wolves can quickly work out how to escape confinement, and require constant reinforcement by caretakers or owners, which makes raising wolves difficult for people who raise their pets in an even, rather than subordinate, environment." Depending if the claim is taken to mean that it is possible, legal or practical to keep a wolf as a pet, the reader might reach different conclusions about whether the evidence supports it. The evidence provides conflicting reasons for both a supported and a refuted label. Therefore, the annotator should choose option c) ambiguous with regard to the evidence.

Alternatively, if the original claim read “it is legal to keep a gray wolf as a pet", the annotator would have to choose d) the evidence does not address the question that is implied in the original claim. The original claim addresses the question of the legality of keeping a wolf as a pet, which the evidence does not cover. Differently from option c) above, this case does not provide conflicting evidence, but rather does not provide enough evidence to choose either way.

When it comes to determining the ambiguity of the original claim with regard to the evidence, feel free to rely on common knowledge to determine whether the claim and the evidence are talking about the exact same entities. For instance, while the claim above is about gray wolves, and the evidence talks more generally about wolves, the annotator may make the assumption that a gray wolf is a wolf, based on their general knowledge. Similar assumptions can be made about names, such as the name Lopez referring to Jennifer Lopez if the evidence mentions the song “Jenny from the block", or any other information that the annotator deems sufficient to determine the referrent. If the annotator does not feel confident about such coreferences, please treat such items as ambiguous with regard to the evidence.

A revised claim “keeping a gray wolf as a pet is very difficult" is supported by the evidence, as the evidence states that raising wolves requires constant reinforcement by the caretakers, which makes it difficult to keep them as pets. The claim addresses the same question as the original claim, as it still implicitly answers the question “can one keep a wolf as a pet?", just like the original claim. Therefore, the annotator should choose option a) supported by the evidence.

If, alternatively, the edited claim read “captive wolves are shy", it would not be addressing the same question as the original claim anymore, as it is not about whether one can keep a wolf as a pet, but about wolf personalities. Even though it would still be supported by the evidence and unambiguous, in this case the annotator should choose d) the revised claim does not address the same question as the original claim (or the evidence). Similarly, if the revised claim might be true but is not related to the evidence anymore, such as “wolves are predators", option d) is again the right choice.

In a different scenario, if the edited claim stated that “wolves are easy to care for as pets", this would still address the question of the difficulty of raising wolves as pets, but the annotator would have to mark it as b) refuted by the evidence. Similarly, if the revised claim states that “it is not clear from the evidence whether wolves are difficult to care for", the annotator should also choose b) refuted by the evidence, as the evidence does in fact specify the difficulty of care for wolves.

Finally, the edited claim might also not be properly disambiguated, for example, if it says that “you should not keep a wolf as a pet". This claim is even more ambiguous than the original one, as the evidence does not provide directives on whether one should attempt keeping such a pet. In this case, it should be noted as c) ambiguous with regard to the evidence.

If the original claim stated that “gray wolves eat meat", the annotator should choose a) supported by the evidence. If it was revised to the claim “gray wolves eat meat, including bones, skin and fur", this edited claim would still be addressing the question of what gray wolves eat, and it would still be a) supported by the evidence. If instead the claim was edited to read “gray wolves eat food", this would still address the question of wolf diets, but it would be c) ambiguous with regard to the evidence, as some of the elements of wolf diets may not be considered food by some readers.

## D Annotation Guidelines for the Final Human Evaluation

The following are examples of claims which might be ambiguous with regard to the given evidence (original claims). Attempts have been made to disambiguate the claims by editing them to be more supported by the evidence (edited claims). Please read the original claims, the evidence, and the edited claims, and assign a score of 0 or 1, where 1 means that the revised claim is now fully

supported by the evidence, whereas 0 means that it is still either ambiguous, unsubstantiated, irrelevant or refuted by the evidence.

For example, you are given the following Evidence: "[...] Captive wolves are generally shy and avoid eye contact with humans other than their owner, as well as not listening to any commands made by any other humans. [...] Ordinary pet food is inadequate, as an adult wolf needs 12.5 kg (25 lbs) of meat daily along with bones, skin and fur to meet its nutritional requirements. [...] The exercise needs of a wolf exceed the average dog's demand. Because of this, captive wolves typically do not cope well in urban areas. Due to their talent at observational learning, adult captive wolves can quickly work out how to escape confinement, and require constant reinforcement by caretakers or owners, which makes raising wolves difficult for people who raise their pets in an even, rather than subordinate, environment."

If the original claim is already unambiguously supported, then the only correct revision would be to keep it as is (minor changes in phrasing would be no problem):
1.
Original claim: "Captive wolves are shy"
Revised claim: "Captive wolves are generally shy"
Score: 1

If the original claim is refuted by the evidence, then the disambiguation should simply negate it:
2.
Original claim: "Captive wolves are not shy"
Revised claim: "Captive wolves are generally shy"
Score: 1

On the other hand, if the original claim is ambiguous, then the revision only gets a score of 1 if it is better supported by the evidence:
3.
Original claim: "You can keep a gray wolf as a pet"
Revised claim: "It may be possible to keep a gray wolf as a pet, but they are very difficult to manage"
Score: 1

Anything that makes the claim unsupported by the evidence, or change the main point of the original claim, would get a score 0:
4.
Original claim: "You can keep a gray wolf as a pet"
Revised claim: "Gray wolves avoid eye contact"
Score: 0 (irrelevant to the original claim)
5.
Original claim: "You can keep a gray wolf as a pet"
Revised claim: "Ordinary pet food is adequate for wolves"
Score: 0 (explicitly refuted by the evidence)

In other cases the evidence might not provide enough information to disambiguate the claim, in which case that should be stated:
6.
Original claim: "It is legal to keep a gray wolf as a pet"
Revised claim: "It is not clear from the evidence whether it is legal to keep a gray wolf as a pet"
Score: 1

However, if it is possible to disambiguate the claim, like in the ambiguous example 2. above, then it is not sufficient to say that it is not clear from the evidence, as there could be a better disambiguation:
7.
Original claim: "You can keep a gray wolf as a pet"
Revised claim: "It is not clear from the evidence whether it is legal to keep a gray wolf as a pet"
Score: 0 (the disambiguation should be provided as in example 3.)

## E Hyperparameters

| Parameter | Values | Model |
|---|---|---|
| lr | [$5x10^{-4}$,$5x10^{\mathbf{-5}}$,$5x10^{-6}$] | all |
| beam size | [**1**,5,10] | base |
| penalty | [**1**,2,3] | lp |
| # pseudo ref | [32,64,**128**] | MBR |
| # hypotheses | [32,64,**128**] | MBR |
| top p | [0.85,**0.9**,0.95] | MBR |
| top k | [40,**50**,60] | MBR |
| epsilon | [0.01,**0.02**,0.03] | MBR |

Table 11: Model hyperparameter values searched, for base, length penalty (lp) and Minimum Bayes Risk (MBR) models (best in bold).

## F Datasheet for Dataset

### F.1 Why was this dataset created?

The DIS2DIS dataset was created for a task of disambiguation. Disambiguation is intended as an alternative method to existing explainability approaches in fact-checking. The dataset was collected for training and testing models for this task. The intended use of the data is to study the phenomenon of ambiguity in the domain of fact-checking.

### F.2 Who funded the creation of the dataset?

[Anonymised]

### F.3 What preprocessing/cleaning was done?

Removal of instances was performed by a manual inspection of random samples from the dataset to ensure high quality annotations.

### F.4 If it relates to people, were they told what the dataset would be used for and did they consent?

No personal data was collected. The annotators consented to the following confidentiality terms: *"By participating in this research, you understand and agree that the researcher may be required to disclose your consent form, data, and other personally identifiable information as required by law, regulation, subpoena, or court order. Otherwise, your confidentiality will be maintained in the following manner: To protect your identity, the researchers will take the following steps: (1) Each participant will be assigned a number; (2) The researchers will record any data collected during the study by number, not by name; (3) Any original recordings or data files will be stored in a secured location accessed only by authorized researchers."*

### F.5 If so, how? Were they provided with any mechanism to revoke their consent in the future or for certain uses?

The annotators were provided the opportunity to revoke their consent at any point during the study. No option to revoke consent in the future was offered, due to the fact that the data collected was completely anonymized and it would not be possible to trace back the responses of a particular annotator.

### F.6 Will the dataset be updated? How often, by whom?

The dataset may be updated in the future, to include other domains or languages. This would be done by the authors of the paper.