AMBIFC: Fact-Checking Ambiguous Claims with Evidence

Max Glockner[1,5], **Ieva Staliūnaitė**[2], **James Thorne**[3], **Gisela Vallejo**[4], **Andreas Vlachos**[2] and **Iryna Gurevych**[1,5]
[1]UKP Lab, Department of Computer Science, Technical University of Darmstadt, Germany
[2]Department of Computer Science and Technology, University of Cambridge, UK
[3]KAIST AI, South Korea, [4]The University of Melbourne, Australia, [5]hessian.ai, Germany
{max.glockner,iryna.gurevych}@tu-darmstadt.de,
{irs38,av308}@cam.ac.uk, thorne@kaist.ac.kr,
gvallejo@student.unimelb.edu.au

Abstract

Automated fact-checking systems verify claims against evidence to predict their veracity. In real-world scenarios, the retrieved evidence may not unambiguously support or refute the claim and yield conflicting but valid interpretations. Existing fact-checking datasets assume that the models developed with them predict a single veracity label for each claim, thus discouraging the handling of such ambiguity. To address this issue we present AMBIFC,[1] a fact-checking dataset with 10k claims derived from real-world information needs. It contains fine-grained evidence annotations of 50k passages from 5k Wikipedia pages. We analyze the disagreements arising from ambiguity when comparing claims against evidence in AMBIFC, observing a strong correlation of annotator disagreement with linguistic phenomena such as underspecification and probabilistic reasoning. We develop models for predicting veracity handling this ambiguity via soft labels, and find that a pipeline that learns the label distribution for sentence-level evidence selection and veracity prediction yields the best performance. We compare models trained on different subsets of AMBIFC and show that models trained on the ambiguous instances perform better when faced with the identified linguistic phenomena.

1 Introduction

In Natural Language Processing, the task of automated fact-checking is given a claim of unknown veracity, to identify evidence from a corpus of documents, and predict whether the evidence supports or refutes the claim. It has received considerable attention in recent years (Guo et al., 2022) and gained renewed relevance due to the hallucination of unsupported or even false statements in natural language generation tasks, including information-seeking dialogues (Dziri et al., 2022; Ji et al., 2023).

Automated fact-checking is closely related to natural language inference (NLI) where the evidence is considered given (Thorne et al., 2018; Wadden et al., 2020; Schuster et al., 2021). Several studies (Pavlick and Kwiatkowski, 2019; Nie et al., 2020; Jiang and Marneffe, 2022) have shown that NLI suffers from inherent ambiguity leading to conflicting yet valid annotations. To address this, recent work has focused on utilizing these conflicting annotations, especially when aggregated labels are not considered to adequately represent the task (Plank, 2022; Leonardelli et al., 2023).

Many fact-checking datasets are purpose-made rather than naturally occurring, similar to those used in NLI; their claims are often created by manipulating sentences from the evidence documents (Thorne et al., 2018; Jiang et al., 2020; Aly et al., 2021). As a result, they are unlikely to represent real-world information needs, as they are written with knowledge of the evidence. On the other hand, in datasets with real-world claims evidence is often used without manual annotation, assuming that it is sufficient (Glockner et al., 2022). If evidence annotation is performed, datasets include artificially created incorrect claims, ensuring that the used evidence contradicts the claims (Wadden et al., 2020; Saakyan et al., 2021), or exhibits low annotator agreement (Hanselowski et al., 2019; Diggelmann et al., 2020) without

[1]https://github.com/CambridgeNLIP/verification-real-world-info-needs.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1–18, 2024. https://doi.org/10.1162/tacl_a_00629
Action Editor: Katrin Erk. Submission batch: 6/2023; Revision batch: 9/2023; Published 1/2024.

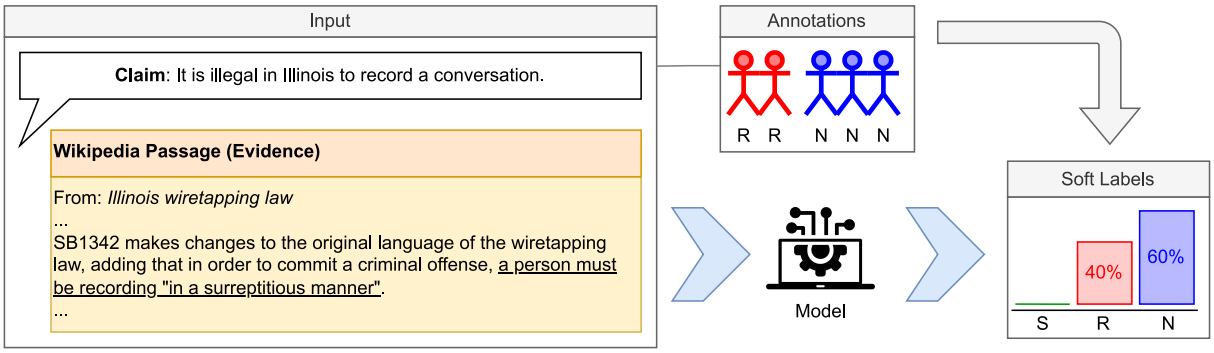

Figure 1: An example of an instance of claim and Wikipedia passage which is ambiguous due to underspecification. We consider all supporting (S), refuting (R), and neutral (N) annotations as valid perspectives. Given a claim and a Wikipedia passage, the model must predict soft labels derived from these annotations.

attempts to handle ambiguity. However, even human fact-checkers often disagree, particularly in ambiguous cases (Lim, 2018).

More concretely, the claim that *''it is illegal in Illinois to record a conversation''* in Figure 1 seems clear on its own, yet becomes ambiguous when compared to the evidence, as it is underspecified. The claim does not explicitly state whether the recording was done surreptitiously (i.e., secretively), allowing for various interpretations: (a) as refuting the claim since it is legal if not done surreptitiously, and (b) as neutral as it is impossible to determine whether it refutes or supports the claim without information about the recording intent. Surreptitious recording only pertains to a specific case and none of the annotators deemed it as prominent enough to provide overall support for the claim.

In this study we aim to investigate the presence of such ambiguities in fact-checking using realistic claims and evidence. To this end, we present AmbiFC, a large fact-checking dataset derived from real-world information needs, sourced from real-world yes/no questions of BoolQ (Clark et al., 2019). AmbiFC contains evidence annotations at the passage and sentence level from full Wikipedia pages, from a minimum of five annotations per instance. Unlike previous fact-checking datasets we consider each annotation as a valid perspective of the claim's veracity given a Wikipedia passage as evidence, and task models to predict the veracity via soft labels that consider all annotations. We provide explanations for the annotator disagreement via our annotations of linguistic phenomena, inspired by Jiang and Marneffe (2022), adding inference types idiosyncratic to fact-checking. Further, we experiment with three established methods to model annotator disagreement. Our work emphasizes the importance of ambiguity within automated fact-checking and takes a step towards incorporating ambiguity into fact-checking models.

2 Related Work

Disagreement among humans are often studied in computational argumentation. Habernal and Gurevych (2017) create a realistic dataset for mining arguments from online discussions, covering various topics. Perspectrum (Chen et al., 2019) gathers different perspectives supported by evidence and their stance on claims. However, computational argumentation focuses on controversial topics with diverse legitimate positions, while automated fact-checking focuses on claim factuality.

In automated fact-checking, earlier works constructed complex claims from question answering datasets (Jiang et al., 2020; Tan et al., 2023; Park et al., 2022) or knowledge graphs (Kim et al., 2023). Our work is most comparable to FaVIQ (Park et al., 2022), which was also generated from real-world information needs questions. Unlike AmbiFC, it lacks evidence annotations and utilizes disambiguated question-answer pairs from AmbiQA (Min et al., 2020), hence excluding the natural ambiguity of claims based on real-world information needs, studied in this work.

Other works gathered claims from credible sources such as scientific publications or Wikipedia, using cited documents as evidence. This provides realistic claims which are only supported by evidence, and requires the generation

of artificial refuted claims (Sathe et al., 2020; Wadden et al., 2020; Saakyan et al., 2021), or only distinguishes between different levels of support (Kamoi et al., 2023). Another line of research collects claims from professional fact-checking organizations. These works often face disagreement among annotators but do not handle ambiguity (Hanselowski et al., 2019; Sarrouti et al., 2021), or do not provide annotated evidence (Augenstein et al., 2019; Khan et al., 2022). The recently published AVeriTeC dataset (Schlichtkrull et al., 2023) reconstructs the fact-checkers' reasoning via questions and answers from evidence documents. In AMBIFC we consider claims that are interesting according to the search queries used in constructing BoolQ (Clark et al., 2019), not claims deemed check-worthy by fact-checkers. Additionally, we provide passage- and sentence-level annotation, and address uncertainty and disagreement.

In the domain of NLI, Nie et al. (2020, ChaosNLI) presented a comprehensive annotation of NLI items, involving 100 annotators for each item. Jiang and Marneffe (2022) further investigate the causes of disagreement in ChaosNLI, categorizing them into pragmatic and lexical features, as well as general patterns of human behavior under annotation instructions. Our work extends the existing work in NLI to fact-checking, by examining the types of linguistic phenomena common in the two tasks. Plank (2022) and Uma et al. (2022) provide overviews of the current state of modeling and evaluation techniques for data with annotation variance. They highlight various methods, such as calibration, sequential fine-tuning, repeated labeling, learning from soft labels, and variants of multi-task learning.

3 Preliminaries

Each instance (c, P) comprises a claim c and a passage P from Wikipedia. A passage $P = [s_1, s_2, \ldots, s_n]$ is composed of n sentences s_i. Annotations are collected for the entire passage P and for each individual $s_i \in P$, indicating their stance towards c as *supporting*, *refuting*, or *neutral*. These *ternary* sentence-level annotations expressing stance towards the claim can be mapped to *binary* annotations by treating non-neutral annotations as ''evidence'' regardless of stance. We do not aggregate passage-level annotations into hard veracity labels. Instead, for each (c, P) we use soft labels, representing the veracity as a distribution of the passage-level annotations given a claim.

We specifically focus on the fact-checking subtasks of Evidence Selection (Ev.) and Veracity Prediction (Ver.) for each claim-passage instance (c P). We consider each sentence s_i as part of the evidence E for c if at least one non-neutral annotation for it exists. For the evidence selection subtask, the model must select all evidence sentences $s_i \in E$ in P. In the veracity prediction subtask, the model must predict the veracity of c given P using soft labels that represent the annotation distribution at the passage level (Figure 1). In addition to comparing the predicted and human label distributions, we assess the models using less stringent metrics (outlined in §6.2.2) to accommodate potential annotation noise.

4 The AMBIFC Dataset

To create the claims and annotate them with evidence, we followed a two-step process. First, crowd-workers transformed questions from BoolQ into assertive statements. Second, the crowd-workers labeled evidence sentences and passages from a Wikipedia page to indicate whether they support or refute the corresponding claim.

Claims BoolQ comprises knowledge-seeking user queries with yes-or-no answers, similar to fact-checking intentions. Dataset instances are generated by rephrasing these queries into claims. Two annotators on Mechanical Turk rephrase each BoolQ question as a claim, with instructions to retain as many tokens from the original question as possible. In case the claims by the two annotators were different, they were included in the dataset after manual review. The crowd-workers underwent a qualification round evaluated by the authors. A total of 512 unique annotators with a 95% acceptance rate completed the task; 20% of HITs were used for worker qualification and training, 80% form the final dataset.

Evidence Annotation For each claim, the full Wikipedia page from BoolQ containing the answer to the yes/no question was used as evidence. To prevent positional bias, where annotators concentrate on a page's beginning, and annotator

fatigue, pages were divided into multiple passage-level annotation tasks (capped at 20 contiguous sentences). Annotators assessed each sentence in a passage as *supporting*, *refuting*, or *neutral* towards the claim, and provided an overall judgment of the claim's veracity given the passage. Passages without evidence sentences were labeled neutral. In anticipation of potentially low inter-annotator agreement as observed in comparable annotation tasks (Hanselowski et al., 2019; Diggelmann et al., 2020), we introduce a second level of passage annotation to indicate uncertainty: If annotators chose ''neutral'' they could additionally flag passages as ''relevant'' to differentiate it from entirely unrelated passages. Non-neutral passage annotations could be flagged as ''uncertain'' by the annotators. We treat both of these additional labels (''relevant'' for ''neutral'' instances and ''uncertain'' for non-neutral ones) as indicators of unclear decision boundaries. Passages received two initial annotations, with an additional three for passages with at least one supporting or refuting initial annotation, resulting in five annotations per instance in these cases. Instances with identical claims (from identical paraphrasing of questions by different annotators) and passages were merged, resulting in instances with more than five annotations.

Quality Controls Annotators underwent a 3-stage approval process consisting of a qualification quiz, manual review of their first 100 HITs and continuous manual review. Errors were communicated to them to provide formative feedback. A batch of claims was sampled daily for continuous manual review during annotation. The authors reviewed and accepted 12,137 HITs (5.2% of all annotation tasks), while corrections were provided for additional 400 HITs, indicating a 3.2% error rate where annotators deviated from guidelines, *not* due to differences in opinion. The number of HITs reviewed for each annotator was proportional to the annotator's error rate and the number of annotations submitted. Annotation times were used to calibrate worker hourly pay at $22.

Agreement The inter-annotator agreement in terms of Krippendorff's α on the collected data is 0.488 on the passage veracity labels and 0.394 on the sentence level. The disagreement implies that single labels cannot capture all valid viewpoints, necessitating the use of soft labels for evaluation. Fully neutral samples have only two annotations (as per dataset construction), which is insufficient for reliable evaluation of soft labels, unless we can ensure that they are indeed 100% neutral. We estimate the probability of misclassifying an instance as fully neutral when only seeing two annotations, by randomly selecting two annotations from samples with 5+ annotations. The likelihood of wrongly assuming an instance as fully neutral when observing two neutral annotations is 0.9% for the entire dataset but it increases up to 20.9% when sampling from uncertain instances. Using this estimate, we omit fully neutral (but ''relevant'') instances from our experiments, while retaining them in our linguistic analysis in §5.2.

	AMBIFC^C		AMBIFC^U
	2-4 Ann.	*5+ Ann.*	*5+ Ann.*
Claims	6,241	4,613	9,380
Wiki Pages	3,418	2,732	4,789
Cl./Passage	18,214	6,475	26,680
Cl./Sentence	141,079	49,497	223,370
Pass. Ann.			
Has N	100%	38.7 %	93.1 %
Has S	0%	82.2 %	78.4 %
Has R	0%	29.0 %	42.0 %
Has S & R	0 %	11.3 %	21.6 %

Table 1: AMBIFC statistics including passages containing **S**upporting, **R**efuting, and/or **N**eutral annotations.

Subsets of AMBIFC We partitioned instances into subsets based on the additional labels ''relevant'' (for neutral passages) and ''uncertain'' (for non-neutral passages) provided by the annotators. Instances marked with either of these labels by any annotator form the ''uncertain'' subset (AMBIFC^U), while the remaining instances form the ''certain'' subset (AMBIFC^C). We split AMBIFC into train/dev/test splits with the proportions of 70/10/20 for both AMBIFC^C and AMBIFC^U based on instances (c, P). We ensure each Wikipedia page only exists in one split, and that the claims and Wikipedia pages occur in the same split regardless of their belonging to AMBIFC^C or AMBIFC^U. The entire AMBIFC includes 51,369 instances (c, P) with 10,722 unique claims and 5,210 unique Wikipedia pages (Table 1). Similar

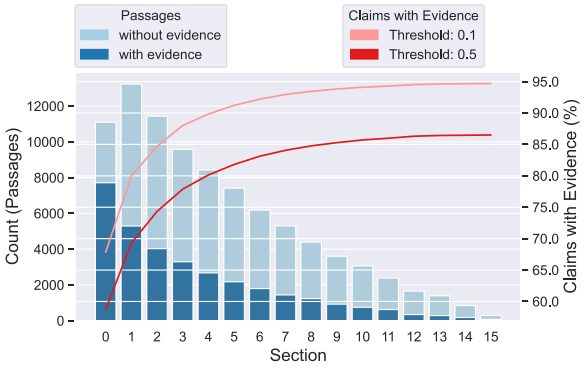

Figure 2: Evidence by section for claims and passages.

Label	Samples	*Krippendorff's* α AMBIFC^C	*Krippendorff's* α AMBIFC^U
Sentence			
binary	all	0.607	–
binary	5+ Ann.	0.563	0.314
ternary	all	0.595	–
ternary	5+ Ann.	0.560	0.302
Passage			
stance	all	**0.815**	–
stance	5+ Ann.	0.553	**0.206**

Table 2: Krippendorff's α over different subsets. Samples in the **bold** are used for AMBIFC.

to VitaminC (Schuster et al., 2021), each claim is annotated based on different evidence passages. Consequently, the same claim may have differing veracity labels depending on the passages. This helps diminish the influence of claim-only biases (Schuster et al., 2019), and allows the same claim to be present in both subsets with different evidence passages. For 7,054 claims (65.8%), (c, P) instances exist in both subsets, AMBIFC^U and AMBIFC^C. Passages with contradictory veracity labels are substantially more frequent in AMBIFC^U than in AMBIFC^C (21.6% vs 11.3%). Instances in AMBIFC^U have at least one non-neutral annotation, as per the dataset annotation process. However, 93.1% of them additionally contain at least one neutral annotation, indicating possibly insufficient evidence. Thus, models cannot achieve high performances when relying on spurious correlations within the claim only (Schuster et al., 2019; Hansen et al., 2021).

Positional Analysis Wikipedia pages have general information in the introduction and more specific details in later sections. In contrast to FEVER, which only uses introductions, our approach involves utilizing passages from entire Wikipedia pages. Figure 2 visualizes the detected evidence per Wikipedia section, revealing that a substantial number of passages from later sections contain evidence for or against the claim. The curves show cumulatively the number of claims with evidence found per section when considering passages with at least 10% or 50% of non-neutral annotations as evidence. While most claims have sufficient evidence in the early sections, there are still many claims that require later sections to be verified.

5 Disagreement Analysis

5.1 Quantitative Analysis

Agreement over Subsets Table 2 shows the agreement results for both subsets. We compared ternary and binary evidence labels at the sentence level. The inter-annotator agreement for the instances (''all'') in AMBIFC^C is 0.607 when calculated with binary labels and 0.595 with ternary labels. For the utilized instances in AMBIFC^U, the agreement is 0.314 with binary labels and 0.302 with ternary labels. The minor differences in agreement under both labeling schemes suggest that annotators with conflicting interpretations may emphasize different evidence sentences rather than assigning opposing labels to the same sentences. The agreement is consistently higher for the certain subset compared to the uncertain subset. The passage-level agreement for the certain subset, measured by Krippendorff's α, is 0.815. When computed over instances with 5+ annotations (removing neutral instances with perfect agreement as they did not receive annotations beyond the first two) the agreement drops to 0.553. We observe much poorer agreement (0.206) on AMBIFC^U. The difference in inter-annotator agreement between these two subsets, based on the annotators' own judgment, signals their awareness of alternative interpretations on these instances.

Agreement over Sections Continuing from the positional analysis (§4), we explore whether the position of evidence passages within sections affects annotator disagreement. Figure 3 visualizes the number of instances (solid) and average

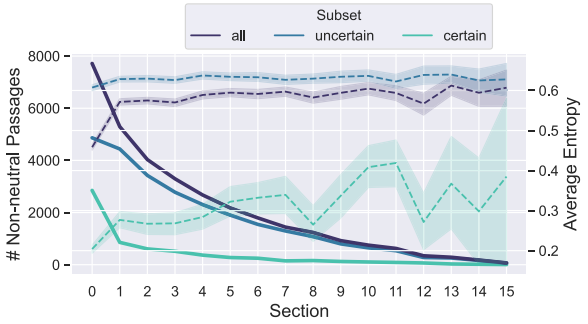

Figure 3: Passage-level annotation entropy (dashed) and count (solid) per section over samples with 5+ annotations.

Subset	Size	Agreement per	α
AMBIFC^C	1,000	Instance	.404
		Veracity	.601
AMBIFC^U	14,814	Instance	.250
		Veracity	.561
AMBIFC	15,814	Instance	**.264**
		Veracity	**.565**

Table 3: Sentence-level Krippendorff's α of evidence annotations between all annotators of the same **Instance**, or **Veracity** interpretation on the same instance.

passage-level annotation entropy (dashed), separated by subset. We only consider passages with 5+ annotations. The entropy is relatively stable within each subset, but substantially different between them. Instances from AMBIFC^C mostly contain evidence in the first section, with few samples in later sections considered certain by annotators. In contrast, instances from AMBIFC^U appear throughout most sections.

Agreement per Veracity Interpretation We aim to determine if different annotators focus on different, or on the same sentences of a passage when assigning contradictory veracity labels to a claim.[2] To examine this, we calculate the agreement among sentence annotations over binary evidence labels in two scenarios: (1) between all annotations of the same instance (c, P), and (2) between all annotations of the same instance when annotators assigned the same veracity label y_p to the claim (c, P, y_p). The results are reported in Table 3. To compare, we need at least two annotators per instance and veracity label. This yields annotations for 15,814 (c, P) instances (32.3% of all instances with 5+ annotations). Due to this selection, this subset represents a highly ambiguous subset of AMBIFC. As expected, the evidence inter-annotator agreement computed at the instance level is poor. When only comparing the evidence annotations among annotators who assigned the same veracity label, the agreement is substantially higher. This suggests that annotators deemed different sentences as important when assigning different veracity labels.

[2]Contradictory sentence- and passage-level annotations by the same annotator occur only in 0.8% (389 instances).

5.2 Linguistic Analysis

To assess the extent to which annotator disagreement in AMBIFC can be attributed to ambiguity, we conduct a statistical analysis that examines various forms of linguistic inference in the data and their relationship to annotator disagreement on veracity labels. We hypothesize that lexical, discourse, and pragmatic inference contributes to disagreements. The inference classes considered are *Implicature, Presupposition, Coreference, Vagueness, Probabilistic Enrichment*, and *Underspecification*, and examples of each type are shown in Table 4. *Implicature* concerns content that is suggested by means of conversational maxims and convention, but not explicitly stated (Grice, 1975). A *Presupposition* refers to accepted beliefs within a discourse (Karttunen, 1974). *Coreference* is used here as a shorthand for difficulty in resolving coreference of ambiguous denotations (Hobbs, 1979), *Vagueness* describes terms with fuzzy boundaries (Kenney and Smith, 1997), and *Probabilistic Enrichment* is a class for inferences about what is highly likely but not entailed. These classes closely follow the framework of Jiang and Marneffe (2022), with changes as follows.

Experimental research has explored the issue of *Underspecification* in generic statements in relation to human cognitive predispositions (Cimpian et al., 2010). They show that generic statements are inconsistently interpreted, suggesting a potential for discourse manipulation. We found many instances of generic underspecified claims in AMBIFC, such as the last example in Table 4. The claim is false *in Britain*, but ill-defined elsewhere, leading to disagreement on

Example	Inference interpretation	Annotations
IMPLICATURE		
Claim: *Red eared sliders can live in the ocean.* Evidence: *In the wild, red-eared sliders brumate over the winter at the bottoms of ponds or shallow lakes.*	Listing the types of bodies of water that red ear sliders brumate in implies that the ocean is not one of them.	[N, N, R, R, R, R]
PRESUPPOSITION		
Claim: *The Queen Anne's Revenge was a real ship.* Evidence:*On June 21, 2013, the National Geographic Society reported recovery of two cannons from Queen Anne's Revenge.*	The evidence presupposes that Queen Anne's Revenge is an existing ship by stating that parts of it were recovered.	[S, S, S, S, N]
COREFERENCE		
Claim: *Steve Carell will appear on the office season 9.* Evidence: *This is the second season not to star Steve Carell as lead character Michael Scott, although he returned for a cameo appearance in the series finale.*	Whether the claim is supported or refuted depends if 'series finale' and 'season 9' have the same referent.	[S, S, S, S, S, R, R]
VAGUENESS		
Claim: *Gibraltar coins can be used in the UK.* Evidence: *Gibraltar's coins are the same weight, size and metal as British coins, although the designs are different, and they are occasionally found in circulation across Britain.*	The veracity judgment depends on the meaning of the word 'can' being interpreted as 'be able to' or 'be legally allowed to'.	[S, N, N, N, N, R, R]
PROBABILISTIC ENRICHMENT		
Claim: *It is rare to have 6 wisdom teeth.* Evidence: *Most adults have four wisdom teeth, one in each of the four quadrants, but it is possible to have none, fewer, or more, in which case the extras are called supernumerary teeth.*	The fact that most adults have 4 wisdom teeth makes it likely that having 6 is rare.	[S, S, S, N, N]
UNDERSPECIFICATION		
Claim: *You cannot have a skunk as a pet.* Evidence: *It is currently legal to keep skunks as pets in Britain without a license.*	The claim is false under specific conditions of location, and underspecified otherwise.	[S, N, R, R, R, R, R]

Table 4: Examples of claim and relevant evidence which require different types of inference to resolve, and their corresponding veracity annotations at the passage level: **R**efuted, **N**eutral, and **S**upported.

the veracity label for the generic statement. This inference type is the reverse of ''*Accommodating Minimally Added Content*'' in hypotheses in Jiang and Marneffe (2022), as the claim (the counterpart to the hypothesis in NLI) in our case is less specific than the evidence. NLI data is usually collected by hypotheses being written for given premises (Williams et al., 2018), whereas the claims in realistic fact-checking data are generated independently from evidence, which leads to different inference types being encountered.

Annotation Scheme We employ stratified sampling to select 384 items, ensuring coverage of both rare and frequent veracity annotation combinations. Each claim is then evaluated with respect to its evidence sentences to determine whether the veracity judgment depends on a specific type of inference or is explicit. Initially, a subset of 20 items was double-annotated to assess the consistency of the guidelines, resulting in a Cohen's κ agreement of 0.67. Subsequently, an additional 364 items were annotated by one of the authors with graduate training in Linguistics.

Variables and Statistics Measured We perform ANOVA to examine the relationship between the independent variables (inference types) and the dependent variable (annotator agreement on veracity labels). Interactions are not included due to the non-overlapping nature of the independent variables in the linguistic inference annotation scheme. Confounders, such as the length of evidence and claim, presence of negation in the claim, and presence of quantifiers in the claim, are added to account for variance unrelated to the hypothesized independent variables. These confounders aim to capture aspects of annotator behavior, as increased cognitive load from negation or longer input length might negatively impact annotation quality, while quantifiers could make the claims clearer to the annotators.

Independent variable	coefficient	P-value
Implicature	−0.2989	0.000*
Presupposition	−0.3800	0.012*
Coreference	−0.2961	0.009*
Vagueness	−0.6469	0.000*
Underspecification	−0.6111	0.000*
Probabilistic reasoning	−0.5590	0.000*
Evidence length	−0.0356	0.966
Claim length	1.2642	0.600
Negation in claim	−0.2209	0.001*
Quantifier in claim	0.1255	0.119

Table 5: Results of ANOVA showing linguistic inference effects on Krippendorff's α in AMBIFC. The significant effects are marked with an asterisk.

Results The variance analysis showed an R^2 value of 0.367, indicating that a significant portion of the variation in annotator disagreement could be explained by annotators' sensitivity to non-explicitly communicated content in claims or evidence as captured by the independent variables. Table 5 presents the significant effects observed in the correlation between the presence of inference types and the level of disagreement. The coefficients in the table reveal that ambiguous content is significantly linked to agreement scores, with the presence of negation in the claim also having a significant effect, likely due to confusion regarding polarity. This corroborates the results of previous work, showing that ambiguity is inherent to linguistic data and therefore annotator disagreement on labels should be incorporated in NLP models.

6 Experiments

6.1 Evidence Selection (Ev.)

The system is tasked with identifying the sentences in a given Wikipedia passage $P = [s_1, \ldots, s_n]$ that serve as evidence $s_i \in E \subseteq P$ for a claim c. We use the F1-score over claim-sentence pairs (c, s_i) for evaluation. We compare four evidence selection methods using binary/ternary evidence annotations with hard or soft labels. Evaluation is performed on AMBIFC^C and AMBIFC datasets.

Models We experiment with four evidence selection approaches. First, following Thorne et al. (2018) and Wadden et al. (2020) we model (Ev.) as a binary classification task. Second, we train a model to predict the ternary label for each (c, s_i). For training, the majority of ''supporting'' and ''refuting'' annotations determines the ternary label, with the overall majority (''supporting'') as tiebreaker, and ''neutral'' assigned if only neutral annotations exist. Ternary predictions are mapped to binary evidence labels for evaluation. We refer to these evidence selection models as *binary* or *ternary*, respectively. To handle the different perspectives by the annotators, one intuitive approach is to mimic the annotation distribution using distillation (Hinton et al., 2015; Fornaciari et al., 2021). Annotation distillation is achieved by minimizing the soft cross-entropy loss between human and predicted distributions. Previous studies directly modeled human annotation probabilities for each class (Peterson et al., 2019), or applied softmax over annotation counts (Uma et al., 2020). We calculate human probabilities by dividing the frequency of annotations per class by the total number of annotations per instance, as this method proved most effective for AMBIFC in our initial experiments. We refer to models that distill these probabilities as *distill* models. A sentence is classified as evidence if the sum of predicted probabilities for ''supporting'' and ''refuting'' exceeds a threshold chosen by maximizing the evidence F1-score on the dev set, with values ranging from 0 to 0.3 in intervals of 0.01. Lastly, we experiment with a regression approach for evidence selection. We calculate the estimated probability p_i for a sentence s_i being part of the evidence set E based on the ratio of annotators who assigned a non-neutral label. We train a regression model (denoted as *regr*) to predict the probabilities p_i by minimizing the MSE loss.

6.2 Veracity Prediction (Ver.)

We experiment with soft labels on the entire AMBIFC and with aggregated labels only on AMBIFC^C. Previous studies in fact-checking have used two model architectures. The *Pipeline* approach predicts the claim's veracity solely based on selected evidence, as seen in approaches for FEVER. Following Wadden et al. (2020), we randomly sample one to two sentences during training only

when no evidence sentence exists. During inference, if no evidence is selected, the prediction defaults to neutral. The second architecture is the *Full-text* approach, where veracity is directly predicted based on the entire evidence document(s) as by Augenstein et al. (2019) or Park et al. (2022).

Fact-checking tasks typically assume single veracity labels (Thorne et al., 2018; Schuster et al., 2021; Park et al., 2022). However, aggregated labels cannot capture the ambiguity in AMBIFC. Therefore, our evaluation based on aggregated labels is only applied on AMBIFC^C, which exhibits higher annotator agreement for the veracity label. We experiment with soft labels using the entire dataset AMBIFC = AMBIFC^C ∪ AMBIFC^U.

6.2.1 Single Label Veracity Prediction

To aggregate the passage-level veracity annotations we employ the Dawid-Skene (Dawid and Skene, 1979) method using the implementation of Ustalov et al. (2021) on AMBIFC^C. Models are assessed based on their accuracy of in predicting the veracity for each (c, P). Similar to FEVER-score (Thorne et al., 2018), we require models to correctly predict the evidence and veracity label (Ev.+Ver.). We score models via the averaged instance-level product of the evidence F1-score with the accuracy of the veracity label. This results in scores of zero when either the veracity or evidence is incorrect, thereby penalizing the model if it doesn't perform well in both tasks.

Models We compare pipeline and full-text models for single-label veracity prediction (SINGLE). We also evaluate a self-correcting version of the pipeline (CSINGLE), which removes selected evidence if it predicts ''neutral'' as veracity. Baseline models utilize selected sentences from the ternary evidence selection approach: The MAX baseline selects the stance with the highest probability, while the MAJ baseline uses majority voting. Sentences are only considered if the predicted probability for a non-neutral label reaches a threshold $t = 0.95$. We determine the threshold t by optimizing the accuracy on the dev set over values ranging from 0 to 1 at intervals of 0.05.

6.2.2 Soft Labels Veracity Prediction

Incorporating diverse annotations in model evaluation is still an open challenge (Plank, 2022). We use four metrics adapted from recent literature (Baan et al., 2022; Jiang and Marneffe, 2022), to score models: The Human Entropy Calibration Error (*EntCE*) assesses the difference in indecisiveness between humans and model predictions by comparing their distribution entropies at the instance level. The Human Ranking Calibration Score (*RankCS*) evaluates the consistency of label rankings between predicted and human probabilities at the instance level. We modify RankCS introduced by Baan et al. (2022) to handle multiple valid rankings for veracity labels identically. The Human Distribution Calibration Score (*DistCS*) is derived from Baan et al. (2022) and quantifies the total variance distance (TVD) between the predicted distribution $\hat{y}$ and the human label distribution y. It is calculated as DistCS $= 1 - \text{TVD}(\hat{y}, y)$ at the instance level and is the strictest of our metrics.

Our annotations may not fully capture the true human distribution. Hence, we treat veracity prediction as a multi-label classification task. Following Jiang and Marneffe (2022), we require models to predict all veracity labels chosen by at least 20% of the annotators. We evaluate models using the sample-averaged F1-score (*F1*). For the joint evaluation (Ev.+Ver.), we calculate the point-wise product of the evidence F1-score with the sample-averaged F1-score (w-F1) and DistCS (w-DistCS).

Models We examine four models that incorporate different annotations to different extents. The first model, referred to as SINGLE (from §6.2), assumes a single veracity label for each (c, P) instance. Additionally, similar to §6.1 we train annotation distillation models (denoted as DISTILL) to learn the human annotation distribution. When no evidence is selected for the pipeline, the prediction defaults to 100% neutral. Third, we apply temperature scaling (Guo et al., 2017) as a method to recalibrate models by dividing the logits by a temperature parameter t before the softmax operation. This technique has demonstrated effectiveness in various NLP tasks (Desai and Durrett, 2020). We choose t based on the highest DistCS score on the dev set for the trained SINGLE models. This calibrated model is denoted as TEMP. SCALING. In the case of the pipeline model, if no evidence is selected, the predicted distribution defaults to 100% neutral. Finally, we explore a multi-label classification approach. Following

Jiang and Marneffe (2022), we estimate the probability of each class by applying the sigmoid function to the model's logits. Classes with a probability of $p \geq 0.5$ are considered as predicted. When necessary for computing metrics, we generate probability distributions by replacing the sigmoid function with softmax during inference. We use evidence selection models with ternary labels and annotation distillation as baselines. We combine the predicted probabilities of the labels ''supporting'' (S) and ''refuting'' (R) by summing them, resulting in $p^{S+R} = 1 - p^{N}$, where p^{N} represents the predicted probability for ''neutral''. We use the predictions based on the sentence with the highest p^{S+R} as the veracity prediction and refer to this baseline as MAX-EVID. We only consider sentences with $p^{S+R} \geq t$, where the threshold t is optimized for DistCS on the development set.

6.3 Implementation

We employ DeBERTaV3$_{\text{large}}$ (He et al., 2021) from the Transformers library (Wolf et al., 2020) for both (Ev. and Ver.) tasks, including Pipeline and full-text variants. DeBERTaV3$_{\text{large}}$ has achieved exceptional performance on the SuperGlue benchmark, including MNLI (Williams et al., 2018) and RTE (Dagan et al., 2006), related to fact-checking. We use fixed hyperparameters (`6e-6` learning rate, batch size of 8)[3] and train for 5 epochs, selecting the best models based on evidence F1-score (Ev. classification), MSE (Ev. regression), accuracy (Ver. single-label), micro F1-score (Ver. multi-label), and negative cross-entropy loss (distillation). DeBERTaV3$_{\text{large}}$ accommodates both short text snippets and longer sequences, enabling fair comparisons between all variants. In initial experiments, we observed that including the Wikipedia entity and section title enhances performance. We input all to the model via `[CLS] claim [SEP] evidence @ entity @ title [SEP]` and feed [CLS] embeddings to linear layer for predictions.

7 Results

Evidence Selection The results in Table 6 show that predicting ternary labels provides no advantage over binary evidence labels. This holds true for both training on the entire AMBIFC and the AMBIFC^{C} subset. However, integrating annotators' uncertainty in evidence selection consistently improves the overall scores. Training solely on AMBIFC^{C} leads to lower F1-score on the entire AMBIFC. A possible reason is the different distribution of evidence sentences: In AMBIFC^{C}, evidence sentences constitute only 8.9% of all sentences. These sentences contain on average 52.2% non-neutral annotations. In AMBIFC, evidence is found in 19.9% of all sentences. These sentences contain on average 38.8% non-neutral annotations. The *distill* approach trained on AMBIFC^{C} performs well in detecting evidence in AMBIFC^{C} (recall = 68.8%, precision = 65.2%), but struggles on AMBIFC as it fails to detect many evidence sentences on instances from AMBIFC^{U} (recall = 42.8%, precision = 76.5%). Training on all of AMBIFC improves the recall of selected evidence, reaching a recall of 80.4% / 64.3% and precision of 51.9% / 68.3% on AMBIFC^{C} and AMBIFC^{U}.

Training		*Evidence F1*	
Data	**Model**	**AMBIFC**	**AMBIFC^{C}**
AMBIFC	binary	64.1 ±0.2	64.4 ±1.2
	ternary	63.5 ±0.7	64.4 ±1.3
	regr	64.5 ±0.4	63.1 ±0.8
	distill	**65.3** ±0.3	63.0 ±1.5
AMBIFC^{C}	binary	56.4 ±1.2	66.2 ±0.6
	ternary	54.0 ±1.9	65.6 ±0.5
	regr	58.2 ±2.3	**66.9** ±0.4
	distill	57.9 ±2.0	**66.8** ±0.6

Table 6: Evidence F1-score averaged with standard deviation over five runs.

Single Veracity Labels We evaluate single label classification models for veracity prediction, selecting the best evidence selection methods from Table 7. The MAJ and SINGLE models achieve high accuracy when provided with oracle evidence. When using automatically selected evidence, SINGLE outperforms the baselines on (Ver.) but performs worse on the joint score (Ev.+Ver.): One possible explanation is that our baselines cannot predict ''neutral'' when evidence

[3]As proposed for MNLI: `https://huggingface.co/microsoft/deberta-v3-large`.

Model			Ver.	Ev.+Ver.
Train	**Ev.**	**Ver.**	**Acc.**	**w-Acc.**
–	*oracle*	MAJ.	*98.5*	*98.5*
–	*oracle*	SINGLE	97.1	97.1
AMBIFC^C	ternary	MAJ.	91.4	85.2
	ternary	MAX.	91.5	85.2
	regr	SINGLE	94.0	83.3
	regr	CSINGLE	94.0	**88.2**
	–	SINGLE	**94.1**	–
AMBIFC	ternary	MAJ.	89.0	82.8
	ternary	MAX.	89.1	82.9
	binary	SINGLE	94.1	77.0
	binary	CSINGLE	94.1	**88.0**
	–	SINGLE	**94.4**	–

Table 7: Averaged veracity prediction results on AMBIFC^C over aggregated single labels over five runs.

sentences are selected. This is beneficial on AMBIFC^C where 96.6% of all instances with evidence sentences have non-neutral veracity labels. The trained SINGLE model, however, can incorrectly predict ''neutral'' even when evidence is correctly identified. For comparison, assuming single labels on AMBIFC^U, 37.2% of instances have a neutral veracity along with supporting or refuting evidence sentences. Training on AMBIFC improves performance on aggregated labels in AMBIFC^C, especially for the full-text model that avoids errors from evidence selection.

High scores on aggregated labels may not comprehensively represent all valid perspectives (Prabhakaran et al., 2021; Fleisig et al., 2023). In the test set, 6.6% of annotations in AMBIFC^C are ignored by the aggregated labels (Figure 4; left). The single-label prediction of the full-text model trained on AMBIFC^C aligns with 87.3% of the veracity annotations. In comparison, aggregated veracity labels in AMBIFC^U would capture only 66.9% of all annotations (Figure 4; right). The AMBIFC-trained full-text model only agrees with 57.1% of them when predicting single labels (with a computed accuracy of 68.8%). Both highlight the importance of annotation-based evaluations throughout AMBIFC.

Soft Veracity Labels We report the results on AMBIFC in Table 8. While SINGLE models are

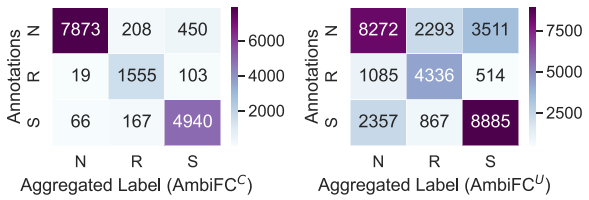

Figure 4: Annotations that are (not) considered by aggregated labels on the respective test sets.

not optimized for metrics over soft labels, they serve as informative baselines. Applying temperature scaling significantly boosts performance on most metrics, particularly EntCE. MULTI and DISTILL outperform other models on various metrics, with each excelling in metrics aligned with their respective optimization objectives. The pipeline approach is comparable to the full-text approach in terms of DistCS, while also providing a rationale for predictions and room for improvement through better evidence selection methods (as indicated by *oracle* evidence). The sentence-level baselines of annotation distillation perform well, but cannot compete with models trained for veracity prediction.

The performance of the top-performing pipelines (based on DistCS) is examined on different subsets in Table 9. Additionally training on ambiguous instances from AMBIFC^U improves performance across all subsets, except for AMBIFC^C. This discrepancy may be attributed to the abundance of fully neutral instances within AMBIFC^C—which do not exist in AMBIFC^U. Performance on instances with 5+ annotations benefits from the inclusion of ambiguous instances. The notable performance gap between AMBIFC and the ambiguous claims in AMBIFC^U underscores the challenge posed by these ambiguous cases.

8 Analysis

Errors by Linguistic Category Model performance varies depending on which lexical, pragmatic, and discourse inference types are present in the items. We compare the predictions of the best performing model (Annotation Distillation, last row in Table 8) trained on AMBIFC^C and AMBIFC, and separate the results per linguistic category (Figure 5). The results corroborate the analysis in §5.2, as the smallest difference between the models trained on AMBIFC^C and AMBIFC is seen with items without linguistic cues for ambiguity.

Model		*Ver.*				*Ev. + Ver.*	
Ev.	**Ver.**	**EntCE↓**	**RankCS↑**	**DistCS↑**	**F1↑**	**w-DistCS↑**	**w-F1↑**
avg. distribution		.568	.529	.597	.747	.215	.267
ternary	MaxEvid.	.305 ±0.03	.644 ±0.01	.701 ±0.03	.730 ±0.03	.546 ±0.02	.506 ±0.03
distill	MaxEvid.	.223 ±0.01	.712 ±0.01	.793 ±0.00	.850 ±0.00	.574 ±0.01	.593 ±0.00
oracle	Single	.289 ±0.03	.779 ±0.01	.787 ±0.01	.800 ±0.01	.787 ±0.01	.800 ±0.01
oracle	Temp. Scaling	.175 ±0.00	.779 ±0.01	.840 ±0.00	.842 ±0.01	.840 ±0.00	.842 ±0.00
oracle	Multi	.244 ±0.01	.792 ±0.00	.810 ±0.00	**.915** ±0.00	.810 ±0.00	**.915** ±0.00
oracle	Distill	**.146** ±0.00	**.801** ±0.00	**.867** ±0.00	.891 ±0.00	**.867** ±0.00	.891 ±0.00
distill	Single	.306 ±0.02	.744 ±0.01	.760 ±0.01	.777 ±0.01	.552 ±0.01	.543 ±0.00
distill	Temp. Scaling	.244 ±0.01	.744 ±0.01	.795 ±0.00	.812 ±0.01	.584 ±0.00	.567 ±0.01
distill	Multi	.270 ±0.01	.755 ±0.01	.782 ±0.01	.881 ±0.01	.566 ±0.01	**.615** ±0.01
distill	Distill	**.214** ±0.00	.764 ±0.00	**.826** ±0.00	.862 ±0.00	**.603** ±0.01	.601 ±0.00
–	Single	.302 ±0.02	.755 ±0.00	.765 ±0.01	.782 ±0.01	–	–
–	Temp. Scaling	.264 ±0.00	.755 ±0.00	.783 ±0.01	.801 ±0.01	–	–
–	Multi	.249 ±0.01	.764 ±0.00	.795 ±0.00	**.884** ±0.00	–	–
–	Distill	.228 ±0.00	**.773** ±0.01	**.826** ±0.00	.867 ±0.00	–	–

Table 8: Results on AmbiFC averaged over five runs. All models are trained on AmbiFC.

	Trained	
Evaluated	**AmbiFC^C**	**AmbiFC**
AmbiFC^C	**.928**	.905
AmbiFC^C (5+)	.804	**.824**
AmbiFC^U	.642	**.751**
AmbiFC	.781	**.826**

Table 9: *DistCS*↑ evaluated across different subsets. AmbiFC^C *(5+)* refers to all instances of AmbiFC^C with at least five annotations.

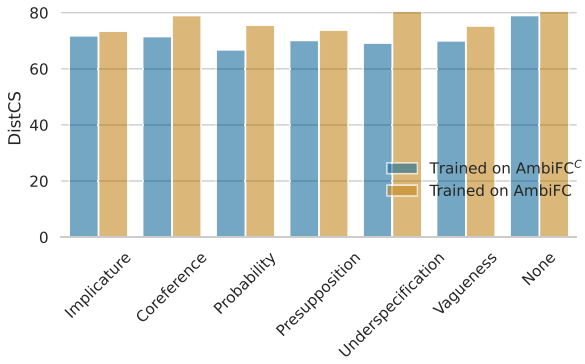

Figure 5: Performance of the Annotation Distillation model on different linguistic categories, separated by the training data used: AmbiFC^C and AmbiFC.

Furthermore, the largest difference appears in the subsets of the development set which contain Underspecification, Vagueness, Probabilistic Enrichment, and Coreference, and the first three of these categories have the strongest correlation with annotator disagreement, as seen in Table 5. This suggests that the model performs better on the more ambiguous items when it has seen such items in training. Furthermore, Underspecification, Vagueness, and Coreference have a lower agreement in the AmbiFC^C subset as compared to the overall agreement in the AmbiFC. This suggests that the annotators are often not aware of the presence of alternative interpretations in these classes, which could also be the reason for these items being more difficult for the model to learn.

Correct Probabilities by Veracity Labels We analyze how accurately the Distill pipelines trained on AmbiFC predict veracity label probabilities in Figure 6. Predictions are considered correct if the difference between human and predicted probabilities falls within the tolerance t on the x-axis. With a tolerance of $t = 0.15$, the pipeline accurately predicts the probability for 70% of instances across all labels in AmbiFC^C. However, the performance is consistently lower in AmbiFC^U, highlighting the greater challenge posed by this subset. The model performs best in predicting the probability for ‘‘refuting’’ labels on both subsets. This is likely because it assigns

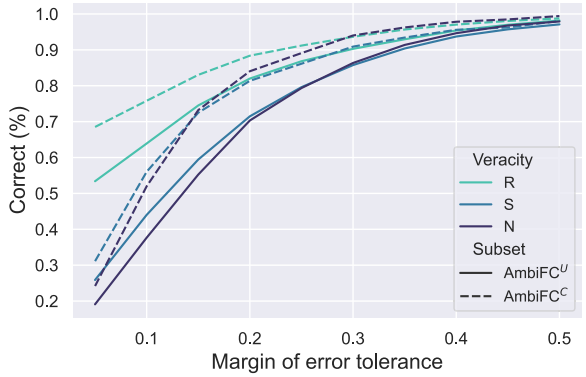

Figure 6: Correct Veracity Estimation by allowing Errors within the margin of the threshold.

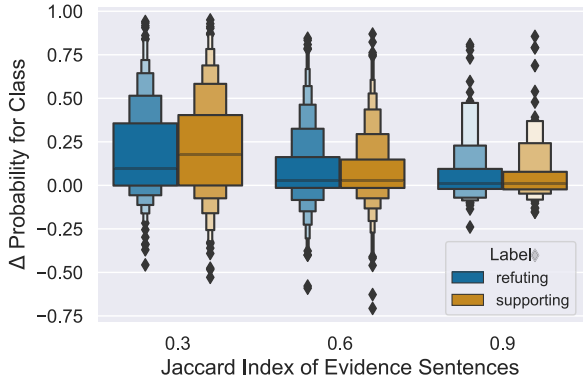

Figure 7: Probability differences of the correctly predicted class when only providing evidence for one of both veracity labels (*S* or *R*).

a lower probability to this less common label. When no refuting annotations exist, the average error is 0.04. However, when refuting annotations are present, the error increases to 0.19.

Contradictory Evidence Interpretations Following our observations in §5.1 we analyze whether models learn the subtle differences between different evidence sentences for different veracity interpretations. We analyze the predictions of a DISTILL pipeline model ($\mathcal{M}$) by inputting evidence sentences annotated with supporting (E^S) *or* refuting (E^R) veracity labels separately. A model that captures the subtle differences would assign high probabilities to the refuting veracity label R given E^R, and low probabilities to R given E^S. We input the claim c and evidence E into $\mathcal{M}$ to predict the probability p^R for the veracity label R via $p^R = \mathcal{M}(c, E)$. We measure the different effect of E^R and E^S for both veracity labels R and S as $\Delta p^R = \mathcal{M}(c, E^R) - \mathcal{M}(c, E^S)$ and $\Delta p^S = \mathcal{M}(c, E^S) - \mathcal{M}(c, E^R)$. In Figure 7, we examine all 1,352 test instances from AMBIFC with both supporting and refuting veracity annotations. To address cases where similar sentences are selected for E^R and E^S, we group samples based on their similarity using the Jaccard Index. Presenting only E^R or E^S generally increases the probability of the correct class. On average, the Δp score is at 11.9%, and decreases with more overlap between sentences in E^R and E^S.

9 Conclusions

We present AMBIFC, a fact-checking dataset with annotations for evidence-based fact-checking, addressing the inherent ambiguity in real-world scenarios. We find that annotator disagreement signals ambiguity rather than noise and provide explanations for this phenomenon through an analysis of linguistic phenomena. We establish baselines for fact-checking ambiguous claims, leaving room for improvement, particularly in the area of evidence selection. By publishing AMBIFC along with its annotations, we aim to contribute to research integrating annotations into trained models.

Limitations Claims in AMBIFC are based on real-world information needs. They are not collected from real-world sources and differ from claims seen as check-worthy by human fact-checkers. AMBIFC lacks evidence retrieval beyond the passage level. It contains different veracity labels for the same claim given different passages, without overall verdict. Models trained on AMBIFC are constrained to this domain and only address partial aspects of complete fact-checking applications, as defined by Guo et al. (2022).

Acknowledgments

This work was supported through a gift from Google as well as the donation of cloud compute credits. The authors wish to thank Dipanjan Das for his advice and support. The authors would like to thank the anonymous reviewers and the Action Editor for their valuable feedback and discussions. The authors would like to thank Jan Buchmann, Sukannya Purkayastha, and Jing Yang for their valuable feedback on an early version of this publication. Conditional scoring with F1 arose from

an idea during a conversation with Jonty Page. Max Glockner is supported by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. Ieva Staliūnaitė is supported by Huawei. James Thorne is supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program [KAIST]). Gisela Vallejo is supported by the graduate research scholarship from the Faculty of Engineering and Information Technology, University of Melbourne. Andreas Vlachos is supported by the ERC grant AVeriTeC (GA 865958) and the EU H2020 grant MONITIO (GA 965576).

References

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The Fact Extraction and Verification Over Unstructured and Structured information (FEVEROUS) Shared Task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.fever-1.1

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. 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 4685–4697, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1475

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.124

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle: Discovering diverse perspectives about claims. 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 542–557, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1053

Andrei Cimpian, Amanda C. Brandone, and Susan A. Gelman. 2010. Generic statements require little evidence for acceptance but have powerful implications. *Cognitive Science*, 348:1452–1482. https://doi.org/10.1111/j.1551–6709.2010.01126.x, PubMed: 21116475

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. https://doi.org/10.18653/v1/N19-1300

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11–13, 2005, Revised Selected Papers*, pages 177–190. Springer. https://doi.org/10.1007/11736790_9

Alexander Philip Dawid and Allan M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series*

C (Applied Statistics), 28(1):20–28. https://doi.org/10.2307/2346806

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.21

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. In *Tackling Climate Change with Machine Learning workshop at NeurIPS 2020*, Online.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022. FaithDial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490. https://doi.org/10.1162/tacl_a_00529

Eve Fleisig, Rediet Abebe, and Dan Klein. 2023. When the majority is wrong: Leveraging annotator disagreement for subjective tasks. *arXiv preprint arXiv:2305.06626v3*.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.204

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.397

H. P. Grice. 1975. Logic and conversation. *Foundations of Cognitive Psychology*, page 719. https://doi.org/10.1163/9789004368811_003

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206. https://doi.org/10.1162/tacl_a_00454

Ivan Habernal and Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179. https://doi.org/10.1162/COLI_a_00276

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 493–503, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/K19-1046

Casper Hansen, Christian Hansen, and Lucas Chaves Lima. 2021. Automatic fake news detection: Are models learning to reason? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 80–86, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-short.12

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543v3*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jerry R. Hobbs. 1979. Coherence and coreference. *Cognitive Science*, 3(1):67–90. https://doi.org/10.1207/s15516709cog0301_4

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38. https://doi.org/10.1145/3571730

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. https://doi.org/10.1162/tacl_a_00523

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.309

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. *arXiv preprint arXiv:2303.01432v1*.

Lauri Karttunen. 1974. Presupposition and linguistic context. *Theoretical Linguistics*, 1(1-3):181–194. https://doi.org/10.1515/thli.1974.1.1-3.181

Rosanna Kenney and Peter Smith. 1997. *Vagueness: A Reader*. The MIT Press. https://doi.org/10.7551/mitpress/7064.001.0001

Kashif Khan, Ruizhe Wang, and Pascal Poupart. 2022. WatClaimCheck: A new dataset for claim entailment and inference. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1293–1304, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.92

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. FactKG: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.895

Elisa Leonardelli, Alexandra Uma, Gavin Abercrombie, Dina Almanea, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, and Massimo Poesio. 2023. SemEval-2023 Task 11: Learning With Disagreements (LeWiDi). *arXiv preprint arXiv:2304.14803v1*. https://doi.org/10.18653/v1/2023.semeval-1.314

Chloe Lim. 2018. Checking how fact-checkers check. *Research & Politics*, 5(3):2053168018786848. https://doi.org/10.1177/2053168018786848

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. https://doi.org/10.18653/v1/2020.emnlp-main.466

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. https://doi.org/10.18653/v1/2020.emnlp-main.734

Jungsoo Park, Sewon Min, Jaewoo Kang, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. FaVIQ: FAct verification from information-seeking questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5154–5166, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.354

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694. https://doi.org/10.1162/tacl_a_00293

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9617–9626. https://doi.org/10.1109/ICCV.2019.00971

Barbara Plank. 2022. The ''problem'' of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.731

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 133–138, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.law-1.14

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. 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 2116–2129, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.165

Mourad Sarrouti, Asma Ben Abacha, Yassine Mrabet, and Dina Demner-Fushman. 2021. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3499–3512, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.297

Aalok Sathe, Salar Ather, Tuan Manh Le, Nathan Perry, and Joonsuk Park. 2020. Automated fact-checking of claims from Wikipedia. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6874–6882, Marseille, France. European Language Resources Association.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A dataset for real-world claim verification with evidence from the Web. *arXiv preprint arXiv:2305.13117v2*.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! Robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.52

Tal Schuster, Darsh Shah, Yun Jie Serene Yeo, Daniel Roberto Filizzola Ortiz, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. 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 3419–3425, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1341

Neset Tan, Trung Nguyen, Josh Bensemann, Alex Peng, Qiming Bao, Yang Chen, Mark Gahegan, and Michael Witbrock. 2023. Multi2Claim: Generating scientific claims from multi-choice questions for scientific fact-checking. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2652–2664, Dubrovnik, Croatia. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.eacl-main.194

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. https://doi.org/10.18653/v1/N18-1074

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2020. A case for soft loss functions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 173–177. https://doi.org/10.1609/hcomp.v8i1.7478

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470. https://doi.org/10.1613/jair.1.12752

Dmitry Ustalov, Nikita Pavlichenko, Vladimir Losev, Iulian Giliazev, and Evgeny Tulin. 2021. A general-purpose crowdsourcing computational quality control toolkit for Python. In *The Ninth AAAI Conference on Human Computation and Crowdsourcing: Works-in-Progress and Demonstration Track*, HCOMP 2021.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.609

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6