AFaCTA: Assisting the Annotation of Factual Claim Detection with Reliable LLM Annotators

Jingwei Ni[1], **Minjing Shi**[1], **Dominik Stammbach**[1], **Mrinmaya Sachan** [1], **Elliott Ash**[1], **Markus Leippold**[2, 3]

[1]ETH Zürich [2]University of Zürich [3]Swiss Finance Institute (SFI)

{jingni, msachan, ashe}@ethz.ch, shimin@student.ethz.ch,
markus.leippold@bf.uzh.ch

Abstract

With the rise of generative AI, automated fact-checking methods to combat misinformation are becoming more and more important. However, factual claim detection, the first step in a fact-checking pipeline, suffers from two key issues that limit its scalability and generalizability: (1) inconsistency in definitions of the task and what a claim is, and (2) the high cost of manual annotation. To address (1), we review the definitions in related work and propose a unifying definition of factual claims that focuses on verifiability. To address (2), we introduce **AFaCTA** (**A**utomatic **Fa**ctual **C**laim de**T**ection **A**nnotator), a novel framework that assists in the annotation of factual claims with the help of large language models (LLMs). AFaCTA calibrates its annotation confidence with consistency along three predefined reasoning paths. Extensive evaluation and experiments in the domain of political speech reveal that AFaCTA can efficiently assist experts in annotating factual claims and training high-quality classifiers, and can work with or without expert supervision. Our analyses also result in PoliClaim, a comprehensive claim detection dataset spanning diverse political topics.[1]

1 Introduction

The explosion of mis- and disinformation is a growing public concern, with misinformation being widely shared (Vosoughi et al., 2018). Manual fact-checking is an important counter-measure to misinformation (Lewandowsky et al., 2020). However, fact-checking is a time-consuming and expensive endeavor, and computational remedies are required (Vlachos and Riedel, 2014).

A first step to identify mis- and disinformation consists of factual claim detection, which filters out the claims with factual assertions that need checking (Arslan et al., 2020; Alam et al., 2021a; Stammbach et al., 2023b). Considering the sheer amount of daily online content and LLMs' generative capability, we argue that a valid factual claim detection system should be efficient and easily deployable to monitor misinformation consistently. Therefore, we need a way to produce high-quality resources to build transparent, accurate and fair models to automatically detect such claims. However, there are two major challenges in the data collection process.

Discrepancies in task and claim definitions. By now, arguably, several different claim definitions exist, which confuse practitioners. What is a *claim* is unclear, leading to various *claim detection tasks*, e.g., in automated fact-checking and argument mining. For example, Alam et al. (2021a) dismiss all opinions from factual claims, but Gupta et al. (2021) includes "opinions with social impact" as factual claims. Many studies (Arslan et al., 2020; Nakov et al., 2022) aim at detecting "check-worthy" claims while Konstantinovskiy et al. (2020) argues the definition of "check-worthiness" is highly subjective and political. Such variances reflect a lack of clarity in conceptualizing critical distinctions, such as the overlap between opinions and verifiable facts (refer to Table 1 row 1), and the separate nature of verifiability and check-worthiness in the context of factual claim detection (see Table 1 rows 2 and 3). To address these inconsistencies, we propose a definition of factual claims based on verifiability: factual claims present verifiable facts; a fact is verifiable only if it provides enough specificity to guide evidence retrieval and fact-checking. We focus on verifiability to maximize the definition's objectivity and clearly delineate facts from opinions.

Manual annotations are expensive. All existing datasets are manually annotated, which is time-consuming and expensive. Thus, most existing resources are inevitably restricted to certain topics for which it is feasible to annotate claims manually. Such examples include presidential debates (Has-

[1] https://github.com/EdisonNi-hku/AFaCTA.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1890–1912
August 11-16, 2024 ©2024 Association for Computational Linguistics

Type	Examples and Explanations
Facts entangled with Opinions	***Example 1:*** We are tackling other needed projects to increase capacity like six-laning I-10 in West Mobile from Theodore to Irvington. *Fact part: The sentence presents a clear and explicit fact about a project. Opinion part: the project's necessity is a subjective judgment.* ***Example 2:*** We are so thankful that we haven't suffered any loss of life, and it's always heartening to see and hear stories of Alaskans pitching in to help each other. *Fact part: no people die in the storm (according to contexts).* ***Example 3:*** I thank the legislature for standing with my administration and the people of Alaska by funding this effort. *Fact part: they fund the effort of resource development (according to contexts).*
Checkworthy but NOT verifiable	***Example 1:*** Democrats and the Media need to stop using the #Coronavirus to politicize things and scare people. It's irresponsible. This is not the time to try and gain political points or headlines from scaring people! *This tweet is labeled as check-worthy by CheckThat!-2021 (Nakov et al., 2021) since it is a polarized political opinion. However, the Democrats' and Media's intention is subjectively interpreted and cannot be verified by objective evidence.* ***Example 2:*** Trump's preference for well-done steaks topped with ketchup. *This is an unverifiable personal preference. However, it is politicized and used to criticize political figures, thus making it checkworthy.*
Verifiable but (maybe) NOT check-worthy	***Example 1:*** Italy's Prime Minister Giuseppe Conte has announced that the whole of the country is being put on lockdown in an attempt to contain the #coronavirus outbreak. *This tweet with verifiable fact is labeled as NOT checkworthy by CheckThat!-2021.* ***Example 2:*** Zee News: Petrol price reduced by Rs 2.69 CNN: Petrol price reduced by Rs 2.69 BBC: Petrol price reduced by Rs 2.69 NDTV: China is sending Corona Virus to the world via mails and WhatsApp. *This tweet cites news with verifiable facts. But it is labeled NOT checkworthy by CheckThat!-2021.*
Context of Claims	***Example:*** ...Those with schizophrenia spectrum and psychosis disorders, many self-medicating with drugs or alcohol addictions. That's precisely what our encampment resolution grants and our new CARE Court seek to address. Getting people off the streets, out of tents, and into housing and treatment is essential to making our streets safe for everyone, but public safety certainly isn't just about homelessness... *This claim defines the duty of CARE but is not self-contained. It is hard to determine its verifiability without the full semantic information in context.*

Table 1: Examples that are not well-defined according to definitions in related work, illustrating the definition of factual claim detection is hard and controversial. Example claims are highlighted in yellow. Explanations are written in *italics*.

san et al., 2015), COVID-19 tweets (Alam et al., 2021a), biomedical (Wührl and Klinger, 2021) and environmental claims (Stammbach et al., 2023a). This potentially limits models' ability to generalize to future topics. However, manually annotating datasets with new topics is too expensive. In light of this, we propose **AFaCTA**, a multi-step reasoning framework that leverages LLMs to assist in claim annotation, making annotation more scalable and generalizable while rigorously following our factual claim definition.

In fact-checking, it is essential to have high annotation accuracy. However, LLM annotators are far from perfect (Ziems et al., 2023; Pangakis et al., 2023). Thus, to ensure the reliability of LLM annotations, AFaCTA calibrates the correctness of the annotations based on the consistency of different paths. Our evaluation shows that AFaCTA outperforms experts by a large margin when all reasoning paths achieve perfect consistency but fails to achieve expert-level performance on inconsistent samples. Nevertheless, we argue that AFaCTA can be an efficient tool in assisting factual claim annotation: perfectly consistent samples can be labeled automatically by the tool, which roughly saves 50% of expert time (see GPT-4-AFaCTA's perfect consistency rate in Table 3). However, inconsistent ones may need expert supervision.

Using AFaCTA, we annotate **PoliClaim**, a high-quality claim detection dataset covering U.S. political speeches across 25 years, spanning various political topics. We split the 2022 speeches as the test set and the 1998 to 2021 speeches as the training set to imitate the real-world use case where a model learns from the past and predicts future claims. We evaluate hundreds of classifiers trained on various data combinations, finding that AFaCTA's annotated data with perfect consistency can be a strong substitute for data annotated by human experts. In summary, our contributions include:

1. We review the regular misconceptions and confounders in claim definition, proposing a claim definition for fact-checking focusing on verifiability.
2. We propose AFaCTA, an LLM-based framework that assists factual claim annotation and ensures its reliability by calibrating annotation quality with consistency along different reasoning paths.
3. We annotate PoliClaim, a high-quality factual claim detection dataset covering political speeches of 25 years and various topics.

2 Claim Definition for Fact-checking

In this section, we first provide an overview of the discrepancies in claim definitions in prior work.

Then, we propose our definition of a factual claim with respect to existing discrepancies.

2.1 Discrepancies in Prior Work

Claim conceptions: The term "claim detection" is used not only in fact-checking but also in other areas of research, for example, argument mining (Boland et al., 2022). However, this term refers to different concepts in different research areas. In fact-checking, claim detection aims at identifying objective information in statements, which can be ruled factually wrong or correct according to evidence (Thorne et al., 2018; Arslan et al., 2020; Gangi Reddy et al., 2022), and unverifiable subjective statements are usually not considered as factual claims. In contrast, in argument mining, claim detection aims at identifying the core argument or point of view referring to what is being argued about (Habernal and Gurevych, 2017). Therefore, both objective and subjective information can be identified as claims depending on their role in the discourse (Daxenberger et al., 2017; Chakrabarty et al., 2019). The intermixing of such concepts has led to dataset misuse issues in research: for instance, Gupta et al. (2021) annotate a claim detection dataset for fack-checking COVID-19 tweets. However, the dataset is jointly trained and evaluated with claim detection datasets for argument mining (Peldszus and Stede, 2015; Stab and Gurevych, 2017, inter alia), which potentially harms the soundness of the results.

Discrepancies in task definitions: Some prior work defines factual claim detection as identifying check-worthy claims (Arslan et al., 2020; Nakov et al., 2021, 2022; Stammbach et al., 2023b) while others aim at distinguishing factual claims and non-claims (Konstantinovskiy et al., 2020; Gupta et al., 2021). Alam et al. (2021a) and Arslan et al. (2020) have both check-worthiness and claim vs non-claim labels. However, Konstantinovskiy et al. (2020) posits that the definition of check-worthiness is subjective, depending on an annotator's knowledge or political stance about a topic. For example, the statement "human-induced climate change is an immediate and severe threat" might be deemed self-evident by climate scientists but as checkworthy by others who are skeptical of climate models or prioritize economic growth. Some might argue that claims like this, which are subject to disagreement regarding their importance, are check-worthy due to their controversial nature. However, it requires background knowledge outside the claim itself to determine the controversy. This could involve factors such as who made the claim and why it is controversial, making the task impossible to solve at the sentence level.

Check-worthiness labels also suffer from another serious problem of future prediction. Training a model detecting past check-worthy claims (e.g., about COVID-19) may fail to detect check-worthiness in future claims whose sociopolitical context and controversy are unknown.

Blurry boundaries between factual claims and non-claims: In related work, personal opinions are usually defined as non-factual claims (Arslan et al., 2020; Alam et al., 2021a). However, many opinions are explicitly based on verifiable facts, lying between the definition of factual claims and non-factual claims. For example: "Hydroxychloroquine cures COVID." is a verifiable factual claim. But "I believe Hydroxychloroquine cures COVID." becomes a personal opinion based on a verifiable fact. Alam et al. (2021a) excludes all opinions from factual claims, which is not a good practice. A false claim can be harmful in political speeches and social media, no matter if it is enclosed by "I believe" or not. Gupta et al. (2021) defines 'opinions with societal implications as factual claims", where societal implications is again an ambiguous definition.

The first row of Table 1 showcases the prevalent entanglement of subjective and objective information. To the best of our knowledge, no previous work in factual claim detection discusses the intersection of opinions and facts and how to delineate facts from opinions.

Context Unavailable: Related work focusing on sentence-level factual claim detection in political speech fails to discuss that sometimes sentences are not self-contained (Arslan et al., 2020; Barrón-Cedeño et al., 2023). However, resolving the co-references is essential for semantic understanding. The last row of Table 1 shows such an example.

2.2 Our Definition of Factual Claims

To avoid **claim misconceptions**, we always use "factual claim" or "claim detection for fact-checking" to specify our focus on fact-checking rather than argument mining. We define facts focusing on verifiability following Arslan et al. (2020) and Alam et al. (2021a):

Fact: *A fact is a statement or assertion that can*

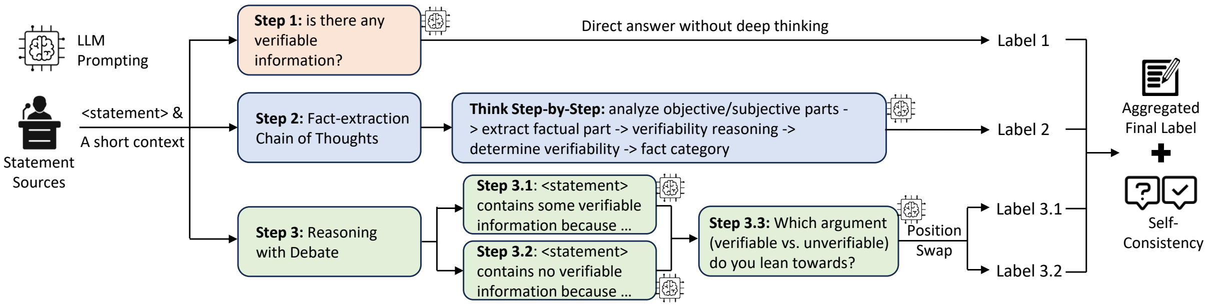
LLM Prompting
Statement Sources
<statement> &
A short context
Step 1: is there any verifiable information?
Direct answer without deep thinking
Label 1
Step 2: Fact-extraction Chain of Thoughts
Think Step-by-Step: analyze objective/subjective parts -> extract factual part -> verifiability reasoning -> determine verifiability -> fact category
Label 2
Step 3: Reasoning with Debate
Step 3.1: <statement> contains some verifiable information because ...
Step 3.2: <statement> contains no verifiable information because ...
Step 3.3: Which argument (verifiable vs. unverifiable) do you lean towards?
Position Swap
Label 3.1
Label 3.2
Aggregated Final Label
Self-Consistency

Figure 1: AFaCTA Pipeline. All steps that need LLM prompting are annotated with the brain icon. Besides the target statement, a short context (if available) is also provided to help the model understand the statement.

be objectively verified as true or false based on empirical evidence or reality.

To have **a clear and objective task definition**, we follow Konstantinovskiy et al. (2020) to focus on verifiability (factual vs. not factual claim) instead of check-worthiness (check-worthy vs. not check-worthy). Whether a sentence contains a verifiable fact or not depends only on its content (and sometimes on a little context surrounding it to clarify key statements), regardless of political or social contexts not captured by the text itself. This differs from many related works that annotate political opinions without verifiable facts as check-worthy and verifiable facts as not check-worthy. Examples of differences in checkworthiness and verifiability are showcased in rows two and three of Table 1. Controversial political opinions and interpretations are usually considered check-worthy due to their potential societal implications. However, they are often open to debate and can hardly be verified against certain evidence. Therefore, we argue that checkworthiness and verifiability are perpendicular dimensions of factual claim detection. In this work, we focus on verifiability for the scalability of data annotation and transferability to easy-to-deploy smaller models.

To address the **opinion-with-fact problem** that is overlooked by prior work, we define opinions and factual claims as:

Opinion: *An opinion is a judgment based on facts, an attempt to draw a reasonable conclusion from factual evidence. While the underlying facts can be verified, the derived opinion remains subjective and is not universally verifiable.*

Factual claim: *A factual claim is a statement that explicitly presents some verifiable facts. Statements with subjective components like opinions can also be factual claims if they explicitly present objectively verifiable facts.*

How to define verifiability? The verifiability of information is not trivial to define because many assertions can be interpreted either subjectively or objectively. For instance, "MIT is one of the best universities in the world" can be either expressing the speaker's subjective feeling about MIT, which is not verifiable, or it can be asserting a verifiable fact, which can be checked with evidence like university rankings and public survey results. For clarity, we define a statement as verifiable if **it provides enough specific information to guide fact-checkers in verification**. Therefore, the above MIT claim is verifiable. Generally, we observe that a statement is verifiable when it provides specific details for evidence search. For example, "MIT is a good university" is less verifiable than "MIT is one of the best universities according to the QS ranking".

3 AFaCTA

This section introduces AFaCTA for assisting factual claim annotation. AFaCTA consists of three prompting steps and an aggregation step (illustrated in Figure 1), inspired by Kahneman (2011) and our claim definitions. The prompts can be found in Appendix C.

Step 1: Direct Classification. We ask LLMs to answer whether a statement contains verifiable information without any chain of thought (CoT, Wang et al., 2023). This step corresponds to a human expert's fast decision-making at first sight of a statement without deep thinking.

Step 2: Fact-Extraction CoT. We instruct LLMs to conduct step-by-step reasoning over a statement: firstly, analyze the objective and subjective information covered; secondly, extract the factual part;

Dataset	|Sample|	|Claim|	Supervision	Split
PoliClaim$_{test}$	816	521	100%	Test
CheckThat!-2021-dev	140	114	100%	Test
PoliClaim$_{gold}$	1953	1154	53%	Train
PoliClaim$_{silver}$	4336	2959	0%	Train
PoliClaim$_{bronze}$	5320	2661	0%	Train

Table 2: |Sample| and |Claim| indicate the numbers of samples and positive samples. **Supervision** indicates the portion of the labels with human supervision. **Split** indicates if the dataset is used for training or test.

thirdly, reason why it is verifiable or unverifiable; and finally, determine whether the factual part is verifiable. This step aims at identifying verifiable facts entangled with subjective opinions (row 1 of Table 1). The prompt and an illustrative example of this step can be found in Appendix C.3.

Step 3: Reasoning with Debate. We note that the verifiability of many statements depends on their interpretation. Ambiguity between verifiable and unverifiable statements often arises from a lack of specificity, as shown in the examples in Appendix A.

Imitating a critical thinking process, we first prompt LLMs to argue that the statement contains some (or does not contain any) verifiable information. Then we pass the debating arguments to another LLM call to judge which aspect it leans towards. To address the position bias of LLM-as-a-judge (Zheng et al., 2023), we prompt the final judging step twice, each time with the positions of the verifiable and unverifiable arguments swapped. The prompts and an illustrative example of this step can be found in Appendix C.4.

Final Step: Results Aggregation. We aggregate the results of three steps through majority voting. Labels from steps 1 and 2 each contribute one vote, while two position-swapped labels from step 3 contribute 0.5 votes apiece (3 votes in total). Samples with more than 1.5 votes are classified as positive samples (factual claims), and others as negative samples. See Appendix D for a discussion on tie-breaking. Idealy, if all steps have perfect consistency (0 or 3 votes), the annotation accuracy should be high.

4 PoliClaim Dataset

We obtain a large political speech data from Picard and Stammbach (2022), which mainly consists of State of the State (SOTS) speeches (already cleaned and split into sentences). These speeches are governors' major public addresses of the year, thus including meaningful political topics. We randomly sample two speeches from each year, from 1998 to 2021, as training data and four speeches from 2022 as test data.[2] This design has two considerations: (1) We aim to replicate the real-world scenario where models are trained on previous claims (e.g., from 1998 to 2021) and used to predict future claims on potentially unseen topics (e.g., in 2022). (2) The test set will be used to evaluate the annotation performance of AFaCTA, and the 2022 speeches are likely unseen by June LLM checkpoints we use to better replicate the future-claim-detection scenario.

The PoliClaim test set (PoliClaim$_{test}$) was annotated by two human experts[3], who had no access to AFaCTA's output when annotating. The experts achieved a substantial Cohen's Kappa of 0.69 in independent annotation before the discussion. Then, they had meetings to resolve disagreements and develop gold labels. Disagreements were mainly caused by ambiguous verifiability, see Appendix A for disagreement resolving. Our annotation guideline, an instantiation of our factual claim definition, can be found in Appendix B.

To test AFaCTA's annotation performance on different domains, we re-annotate the development set of CheckThat!-2021 (Nakov et al., 2021), which originally contained check-worthiness labels of COVID-19 tweets, following the same annotation process (Cohen's Kappa 0.58). Due to budget limitations, our explorations and annotations mainly focused on the domain of political speech. We leave the extensive study on the social media domain (and other potential domains for factual claim detection) to future work.

After verifying the performance of AFaCTA using the test sets (see more in Section 5.1), we annotated the training set with the tool's assistance, imitating its expected use case of assisting annotation. The perfectly consistent samples were labeled directly with GPT-4 AFaCTA, while the inconsistent samples were left for human annotation. We randomly sampled 8 speeches and manually relabeled the inconsistent annotations from AFaCTA, leading to PoliClaim$_{gold}$ where all annotations are labeled with perfect consistency or human supervision. The perfectly consistent samples in the rest

[2] We do speech-level random sampling to keep the sentence distribution of full speeches.

[3] PhD students who are familiar with the domain of political speeches in the U.S. and COVID-related claims and have good knowledge of the literature on claim detection.

of the speeches fall into PoliClaim$_{silver}$ while the inconsistent samples fall into PoliClaim$_{bronze}$. The statistics of datasets can be found in Table 2.

5 Experiments

Since AFaCTA is an LLM-agnostic prompting framework, we test both GPT-3.5 (Ouyang et al., 2021) and GPT-4 (OpenAI, 2023) as the backbone LLM. We also test open-sourced LLMs which does not work well due to high position bias in Step 3 (see Appendix F). Detailed settings are in Appendix G to ensure reproducibility.

5.1 AFaCTA Annotation Performance

It is unlikely for LLMs to produce expert-level annotation on all samples S. Therefore, AFaCTA (with LLM $\mathcal{M}$) calibrates its performance with self-consistency, dividing S into two subsets: $S^{\mathcal{M}}_{con}$ with perfect consistency across all steps (0 or 3 votes) and $S^{\mathcal{M}}_{inc}$ with inconsistency among some steps (0.5 to 2.5 votes). We use two criteria to compare AFaCTA with human experts: (1) Accuracy: AFaCTA's accuracy vs. experts' average accuracy, both are computed against gold labels; (2) Agreement (Cohen's Kappa): AFaCTA's average agreement to experts vs. agreement between experts. Both metrics should be compared on S, $S^{\mathcal{M}}_{con}$, and $S^{\mathcal{M}}_{inc}$ to evaluate AFaCTA's reliability on entire, perfectly consistent, and inconsistent samples. See Appendix E for formulas and implementations of all metrics.

The results are presented in Table 3. On the full test set S, even GPT-4 AFaCTA underperforms the average performance of human experts on both accuracy and agreement. However, if we only consider the subset where AFaCTA has perfect consistency ($S^{\mathcal{M}}_{con}$), GPT-4 outperforms human experts by a large margin on accuracy (98.49% > 94.85%) and achieves better agreement with experts (0.833 > 0.743). On the contrary, LLMs achieve worse annotation performance than human experts on inconsistent subsets ($S^{\mathcal{M}}_{inc}$). Comparable inter-human agreement is achieved on both subsets, but the accuracy and agreement on $S^{\mathcal{M}}_{con}$ are higher, indicating that $S^{\mathcal{M}}_{con}$ is slightly less challenging than $S^{\mathcal{M}}_{inc}$.

Takeaway: With AFaCTA's self-consistency calibration, auto-annotation of perfectly consistent samples can be reliably adopted to reduce manual effort (also see Section 5.5). In the case of PoliClaim$_{test}$, only 51.22% needs further supervision, while 48.78% of manual effort is saved with GPT-4-AFaCTA.

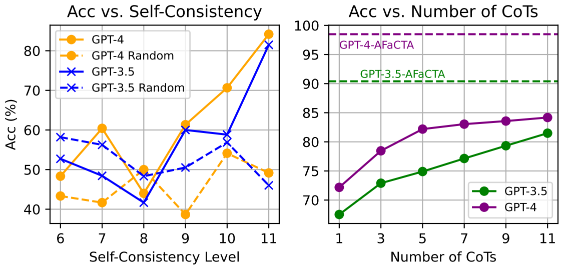

Figure 2: Left figure: accuracy vs. self-consistency levels achieved by 11 CoT calls. Self-consistency level x means there are x CoTs that agree on the label and $(11 - x)$ disagree. Solid and dashed lines denote the performance of LLMs and random guesses on subsets of different self-consistency correspondingly. Right figure: accuracy on the subset where all x CoTs achieve agreement vs. number of sampled CoTs x. Note that the subset of perfect consistency is getting narrower and narrower when sampling more CoTs.

5.2 Error Analysis

Annotation errors in the fact-checking domain may lead to downstream model inaccuracies. Therefore, we also analyze AFaCTA's errors within the perfectly consistent samples. We find that GPT-4 AFaCTA makes false positive errors due to oversensitivity to granular or implicit facts. It makes false negative errors due to context limitations. GPT-3.5 seems less capable of identifying implicit facts within opinions compared to GPT-4. It sometimes fails to identify facts that are specific enough for verification and asks for more "specific details". Roughly 97% of its errors are false negatives caused by misunderstanding verifiability and other hallucinations, indicating that its positive predictions are more reliable.

In Appendix N, we analyze all errors rather than provide isolated examples to avoid cherry-picking. We hope that this thorough analysis can benefit future research in manual/automatic annotation about factual claims.

5.3 Predefined Reasoning Paths Matter

Leveraging self-consistency to improve LLM reasoning is not new. Wang et al. (2023) show that LLMs can use self-sampled reasoning paths (i.e., CoTs) to improve predictions with self-consistency. In AFaCTA, we use pre-defined reasoning paths instead of LLM-sampled ones. To compare these approaches, we conduct self-consistency CoT with the prompt of Step 1: Direct Classification. Step 1

	S (100†/100‡)		$\mathbf{S}^{\mathcal{M}}_{\mathbf{con}}$ (43.38†/48.78‡)		$\mathbf{S}^{\mathcal{M}}_{\mathbf{inc}}$ (56.62†/51.22‡)	
	Agreement	**Accuracy**	**Agreement**	**Accuracy**	**Agreement**	**Accuracy**
GPT-3.5	0.510	76.47	0.754	90.40	0.331	65.80
GPT-4	0.615	86.27	**0.833**	**98.49**	0.418	74.64
Experts	**0.690**	**92.77**	0.746†/0.743‡	93.79†/94.85‡	**0.636**†/**0.629**‡	**91.99**†/**90.79**‡

Table 3: AFaCTA's performance on PoliClaim$_{test}$. "S", "$S^{\mathcal{M}}_{con}$", and "$S^{\mathcal{M}}_{inc}$" report scores on the full test set, perfectly consistent samples, and inconsistent samples correspondingly. The percentages (%) of "$S^{\mathcal{M}}_{con}$" and "$S^{\mathcal{M}}_{inc}$" samples are also reported in column titles. The **Experts** row reports inter-human agreement and average human annotation accuracy against gold labels. **GPT-3.5 (-4)** rows report AFaCTA's average agreement to both experts, and its accuracy score against gold labels. "†" and "‡" denote GPT-3.5 and GPT-4 reported $S^{\mathcal{M}}_{con}$ / $S^{\mathcal{M}}_{inc}$ correspondingly (i.e., $\mathcal{M}$ = GPT-3.5 / -4).

is chosen since it (1) directly addresses verifiability, which is the core of our factual claim definition; (2) contains no predefined CoT; and (3) is simple but achieves decent performance compared to Steps 2 and 3 (see Appendix H where we separately evaluate each step's performance).

We generate 11 CoTs (more details in Appendix I) for both GPT-3.5 and GPT-4 and then compute accuracy scores for different self-consistency levels. The results are illustrated in the left figure of Figure 2. We observe that self-consistency level, to some degree, calibrates accuracy: a higher self-consistency level generally indicates higher accuracy, and vice versa. However, self-consistency CoT underperforms AFaCTA on the perfectly consistent subset (84.18% < 98.49%) while the former samples 11 CoT reasoning paths, and the latter relies on only 3 predefined reasoning paths. One possible explanation is that the predefined paths encourage critical thinking and reasoning from different angles, making the achieved self-consistency more comprehensive. We also observe that AFaCTA and self-consistency CoT achieve perfect consistency on **48.78%** and **58.09%** of the data, respectively, indicating that the perfect-consistency in AFaCTA is only slightly harder to achieve than in self-consistency CoT.

Furthermore, we find that the accuracy on perfectly consistent samples grows with the number of CoT voters (see the right figure of Figure 2). This is intuitive as more consistent outputs indicate more confident predictions. However, the marginal benefit of adding more CoTs drops significantly: the accuracy of GPT-4 tends to converge to 85%. Since the accuracy of GPT-3.5 seems to grow linearly up to 11 CoTs, we further extend it to 19 CoTs and observe convergence to 84.1% (see Figure 5), which is still much lower than GPT-3.5 AFaCTA's 90.4%.

Takeaway: Auto-annotations with more self-consistency (especially the perfectly consistent ones) tend to be more accurate. However, the source of self-consistency needs to be diversified and well-defined to scale up annotation performance efficiently. In this case, we show that predefined reasoning paths with expertise outperform those automatically sampled by LLMs.

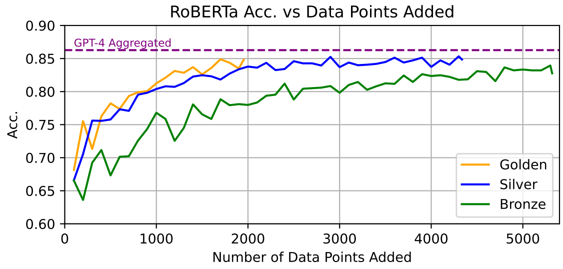

Figure 3: The performance of fine-tuned RoBERTa on PoliClaim$_{test}$ when gradually adding training data of different quality. "- -" denotes GPT-4's performance aggregating three AFaCTA reasoning steps.

5.4 Domain Agnostic AFaCTA

The reasoning logic of AFaCTA is not restricted to the political speech domain. To verify its performance on the social media domain, we conduct the analyses in Section 5.1 and Section 5.3 again on the CheckThat!-2021 (Nakov et al., 2021) development set. Experiment results are similar to those on PoliClaim$_{test}$ (see Appendix J). Therefore, AFaCTA may assist factual claim annotation in various domains.

5.5 AFaCTA Delivers Useful Annotations

To explore whether AFaCTA's annotation can replace or augment manual annotation in training classifiers, we train hundreds of classifiers with different combinations of PoliClaim$_{gold}$ (AFaCTA annotations + Human Supervision), PoliClaim$_{silver}$ (AFaCTA perfectly consistent annotations), and PoliClaim$_{bronze}$ (AFaCTA inconsistent annota-

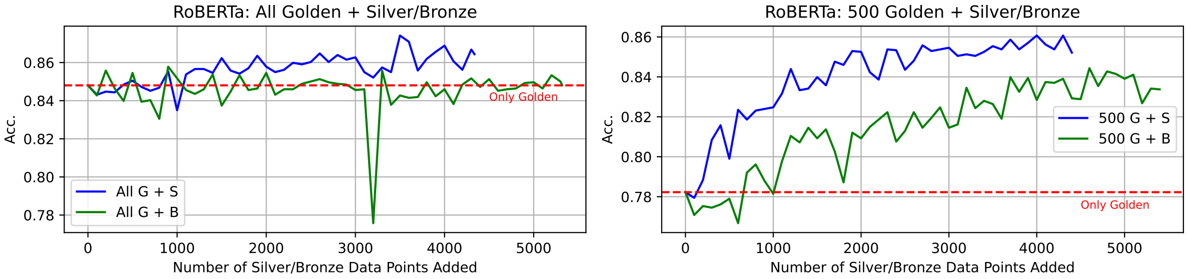

Figure 4: The performance of augmenting a limited number of PoliClaim$_{gold}$ data (left figure: all 1936 samples, right figure: 500 samples) with extra data from PoliClaim$_{silver}$ and PoliClaim$_{bronze}$. Experiments of augmenting 1000 and 1500 PoliClaim$_{gold}$ samples can be found in Appendix M. "- -" denotes the performance without augmentation. G, S, and B denote golden, silver, and bronze PoliClaim correspondingly.

tions). All results are averaged over random seeds of 42, 43, and 44, and are supported with statistical significance tests (see Appendix L). [4]

Using only gold, silver, or bronze data: We first gradually increase the number of training data points (by 100 per step) of the same quality. Results are shown in Figure 3. We observe the same phenomenon as previous work (Stammbach et al., 2023b) where the marginal accuracy gain drops while adding more data. The PoliClaim$_{gold}$ and PoliClaim$_{silver}$ curves roughly follow the same growing trend, approaching GPT-4's aggregated performance. This indicates that the perfectly consistent annotations (silver) from AFaCTA can strongly substitute for manually annotated data. The PoliClaim$_{gold}$ curve is slightly higher, showing that learning from human-supervised hard samples (inconsistent annotations of AFaCTA) is beneficial. The PoliClaim$_{bronze}$ curve is much lower, showing that the noisy, inconsistent annotations harm the classifier training.

Augmenting training with auto-annotated data: When the manual annotation budget is limited, can we augment the dataset with automatic annotation? In Figure 4, we gradually augment the PoliClaim$_{gold}$ data with automatically annotated ones (100 per step). It can be observed that: (1) The performance increases more with PoliClaim$_{silver}$ data augmentation, showing that the data quality is important in data augmentation. (2) Compared to augmenting the full PoliClaim$_{gold}$ dataset, augmentation results in more improvement when there are only 500 PoliClaim$_{gold}$ data. Therefore, high-quality automatic annotation is more helpful when the manual annotation budget is limited. (3) Combining gold and silver data leads to classifiers that outperform aggregated GPT-4 reasoning, demonstrating that extending training data with LLM annotation is a promising approach to achieving better performance. One of the best RoBERTa checkpoints trained on all PoliClaim$_{gold}$ and PoliClaim$_{silver}$ is available on HuggingFace[5].

6 Related Work

Claim Detection: The term "claim detection" has different definitions in various research fields (Boland et al., 2022). Even inside the field of fact-checking, its exact definition depends on the domain (Alam et al., 2021b; Stammbach et al., 2023b) or task objective (Arslan et al., 2020; Konstantinovskiy et al., 2020; Gangi Reddy et al., 2022) and is somewhat arbitrary. In this work, we propose a definition focusing on one important dimension of factual claims – verifiability, to minimize the conceptual uncertainty. Another important dimension of factual claims is check-worthiness (Arslan et al., 2020; Nakov et al., 2021, 2022; Barrón-Cedeño et al., 2023), whose definition is more arbitrary (Konstantinovskiy et al., 2020).

Automatic Annotation: Automatic data annotation using LLM is both promising (Pangakis et al., 2023) and necessary (Veselovsky et al., 2023). Early work observes that LLMs' annotation performance highly depends on tasks: LLMs outperform human annotators on some tasks (Gilardi et al., 2023; Zhu et al., 2023; Törnberg, 2023) but fails to achieve human-level performance on others (Ziems et al., 2023; Reiss, 2023). Therefore, we argue that

[4] This section presents RoBERTa (Liu et al., 2019) results. Appendix M presents similar DistilBERT (Sanh et al., 2019) results as side findings. Detailed fine-tuning settings are in Appendix K.

[5] https://huggingface.co/JingweiNi/roberta-base-afacta

a detailed task-specific study about LLM annotation reliability is essential.

Pangakis et al. (2023) recommend evaluating LLMs' annotation against a small subset that is not in the LLMs' training corpus and annotated by subject matter experts. We follow these suggestions in this work. Concurrent studies also explore self-consistency (Pangakis et al., 2023) and CoT (He et al., 2023) to improve the performance and reliability of LLM annotation. However, they do not compare predefined reasoning paths with automatically sampled CoTs.

7 Discussions

7.1 Check-Worthiness

The objective of factual claim detection is to prioritize claims that are both verifiable and checkworthy, maximizing the use of potentially limited fact-checking resources. However, in this project, we focus on verifiability without exploiting the other important aspect: checkworthiness. Konstantinovskiy et al. (2020) argues that the definition of check-worthiness is subjective. However, it is possible to define a claim's checkworthiness according to its context. For example, is the claimer an influential person or media? Is the topic controversial? There has already been work that takes some contextual information (e.g., claimer, topic, etc.) into account (Gangi Reddy et al., 2022). Future work may explore deterministic and efficient ways to define and annotate checkworthiness leveraging rich contextual information.

7.2 Only GPT-4 Is Reliable

We find that only GPT-4-AFaCTA outperforms human experts on perfectly consistent samples. GPT-3.5 achieves promising results but tends to produce false negative errors. Although GPT-4 is much cheaper than human supervision, it is close-sourced and is comparatively more expensive than other LLMs. Future work may study how to use open-sourced models to produce high-quality annotations. Specifically, future work may explore (1) training the model to better understand the annotation guideline; (2) leveraging internal certainties like output logits; and (3) extending the spectrum of self-consistency levels with cheaper inference.

8 Conclusion

We propose AFaCTA, which leverages LLMs to assist in the annotation of factual claim detection. It ensures reliability by calibrating annotation quality through consistency. AFaCTA's consistent annotation proves effective for training and data augmentation even without human supervision.

Limitations

AFaCTA Prompt. The design of AFaCTA prompts is inspired by the fast and slow thinking patterns (Kahneman, 2011) and prior knowledge of factual claim definition. However, we do not explore other techniques (e.g., few-shot prompting, in-context learning, and putting whole annotation guidelines in context etc.) to improve AFaCTA performance further, for two reasons: (1) the current AFaCTA's performance is good enough to show the potential of assisting claim detection annotation with LLMs; and (2) we annotated thousands of sentences with GPT-4-AFaCTA, which is very expensive. Extending the current prompts with more in-context information is not affordable for us.

Besides, AFaCTA step 2 and 3 cost (approximately) 6.5x and 8.5x more tokens than step 1. Although step 2 and 3 bring self-consistency calibration and performance gain through aggregation, the marginal benefit of API cost is far from perfect.

Social Media and Other Domains. In this work, we only conduct extensive experiments and analyses on the political speech domain, only exploring the social media domain with a small dataset (due to the definition discrepancy, we cannot evaluate our methods with prior datasets). We believe a comprehensive study on one domain can provide deeper insights, and the conclusions might be transferable to other domains. Therefore, we do not split our budget across various domains. Future work may consider extending the large-scale analyses to other domains that need fact-checking.

Limited Expert Annotators. We only evaluate AFaCTA's annotation performance against two experts, which may lead to potential bias. We fail to hire more expert annotators mainly because expert annotation is extremely expensive, and it is hard to find more experts with good knowledge about factual claim definitions. As compensation, we release all expert annotations and detailed error analyses where the potential bias can be analyzed. Besides, adding unsupervised LLM-annotated data continuously improves the accuracy on PoliClaim$_{test}$, demonstrating that our human labeling on PoliClaim$_{test}$ has very limited bias.

Ethics Statement

In this work, all human annotators are officially hired and have full knowledge of the context and utility of the collected data. We adhered strictly to ethical guidelines, respecting the dignity, rights, safety, and well-being of all participants.

There are no data privacy issues or bias against certain demographics with regard to the annotated data. Both original SOTS data (Picard and Stammbach, 2022) and CheckThat!-2021 (Nakov et al., 2021) datasets are widely used for NLP and other research. Our annotated datasets will also be publicly available for research purpose.

Acknowledgements

This paper has received funding from the Swiss National Science Foundation (SNSF) under the project 'How sustainable is sustainable finance? Impact evaluation and automated greenwashing detection' (Grant Agreement No. 100018_207800). It is also funded by grant from Hasler Stiftung for the Research Program Responsible AI with the project "Scientific Claim Verification."

References

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021a. Fighting the COVID-19 infodemic: Modeling the perspective of journalists, fact-checkers, social media platforms, policy makers, and the society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Firoj Alam, Shaden Shaar, Fahim Dalvi, Hassan Sajjad, Alex Nikolov, Hamdy Mubarak, Giovanni Da San Martino, Ahmed Abdelali, Nadir Durrani, Kareem Darwish, Abdulaziz Al-Homaid, Wajdi Zaghouani, Tommaso Caselli, Gijs Danoe, Friso Stolk, Britt Bruntink, and Preslav Nakov. 2021b. Fighting the COVID-19 Infodemic: Modeling the Perspective of Journalists, Fact-Checkers, Social Media Platforms, Policy Makers, and the Society. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 611–649, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fatma Arslan, Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2020. A Benchmark Dataset of Check-Worthy Factual Claims. *Proceedings of the International AAAI Conference on Web and Social Media*, 14:821–829.

Alberto Barrón-Cedeño, Firoj Alam, Andrea Galassi, Giovanni Da San Martino, Preslav Nakov, Tamer Elsayed, Dilshod Azizov, Tommaso Caselli, Gullal S. Cheema, Fatima Haouari, Maram Hasanain, Mucahid Kutlu, Chengkai Li, Federico Ruggeri, Julia Maria Struß, and Wajdi Zaghouani. 2023. Overview of the clef–2023 checkthat! lab on checkworthiness, subjectivity, political bias, factuality, and authority of news articles and their source. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18–21, 2023, Proceedings*, page 251–275, Berlin, Heidelberg. Springer-Verlag.

Katarina Boland, Pavlos Fafalios, Andon Tchechmedjiev, Stefan Dietze, and Konstantin Todorov. 2022. Beyond facts – a survey and conceptualisation of claims in online discourse analysis. *Semantic Web*, 13(5):793–827.

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019. IMHO fine-tuning improves claim detection. 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 558–563, Minneapolis, Minnesota. Association for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Revanth Gangi Reddy, Sai Chetan Chinthakindi, Zhenhailong Wang, Yi Fung, Kathryn Conger, Ahmed ELsayed, Martha Palmer, Preslav Nakov, Eduard Hovy, Kevin Small, and Heng Ji. 2022. NewsClaims: A New Benchmark for Claim Detection from News with Attribute Knowledge. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6002–6018, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. ArXiv:2303.15056 [cs].

Shreya Gupta, Parantak Singh, Megha Sundriyal, Md. Shad Akhtar, and Tanmoy Chakraborty. 2021. LESA: Linguistic Encapsulation and Semantic Amalgamation Based Generalised Claim Detection from Online Content. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3178–3188, Online. Association for Computational Linguistics.

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

Naeemul Hassan, Chengkai Li, and Mark Tremayne. 2015. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, page 1835–1838, New York, NY, USA. Association for Computing Machinery.

Xingwei He, Zhenghao Lin, Yeyun Gong, A.-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2023. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. ArXiv:2303.16854 [cs].

Daniel Kahneman. 2011. Thinking, fast and slow.

Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2020. Towards Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. ArXiv:1809.08193 [cs].

Stephan Lewandowsky, John Cook, Ullrich Ecker, Dolores Albarracin, Michelle Amazeen, Panayiota Kendeou, Doug Lombardi, Eryn Newman, Gordon Pennycook, Ethan Porter, David G. Rand, David N. Rapp, Jason Reifler, Jon Roozenbeek, Philipp Schmid, Colleen M. Seifert, Gale M. Sinatra, Briony Swire-Thompson, Sander van der Linden, Emily K. Vraga, Thomas J. Wood, and Maria S. Zaragoza. 2020. Debunking handbook 2020. `https://sks.to/db2020`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Hamdy Mubarak, Alex Nikolov, and Yavuz Selim Kartal. 2022. Overview of the clef-2022 checkthat! lab task 1 on identifying relevant claims in tweets. In *Conference and Labs of the Evaluation Forum*.

Preslav Nakov, Giovanni Da San Martino, Tamer Elsayed, Alberto Barrón-Cedeño, Rubén Míguez, Shaden Shaar, Firoj Alam, Fatima Haouari, Maram Hasanain, Watheq Mansour, Bayan Hamdan, Zien Sheikh Ali, Nikolay Babulkov, Alex Nikolov, Gautam Kishore Shahi, Julia Maria Struß, Thomas Mandl, Mucahid Kutlu, and Yavuz Selim Kartal. 2021. Overview of the clef–2021 checkthat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news.

OpenAI. 2023. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Bo Ouyang, Wenbing Huang, Runfa Chen, Zhixing Tan, Yang Liu, Maosong Sun, and Jihong Zhu. 2021. Knowledge representation learning with contrastive completion coding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3061–3073, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated Annotation with Generative AI Requires Validation. ArXiv:2306.00176 [cs].

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948, Lisbon, Portugal. Association for Computational Linguistics.

Léo Picard and Dominik Stammbach. 2022. Political metaphors in u.s. governor speeches. *SSRN Electronic Journal*.

Michael V. Reiss. 2023. Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark. ArXiv:2304.11085 [cs].

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Christian Stab and Iryna Gurevych. 2017. Parsing Argumentation Structures in Persuasive Essays. *Computational Linguistics*, 43(3):619–659.

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023a. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

Dominik Stammbach, Nicolas Webersinke, Julia Anna Bingler, Mathias Kraus, and Markus Leippold. 2023b. Environmental Claim Detection. ArXiv:2209.00507 [cs] version: 4.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct Distillation of LM Alignment. ArXiv:2310.16944 [cs].

Petter Törnberg. 2023. ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning. ArXiv:2304.06588 [cs].

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. ArXiv:2306.07899 [cs].

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA. Association for Computational Linguistics.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. ArXiv:2203.11171 [cs].

Amelie Wührl and Roman Klinger. 2021. Claim detection in biomedical twitter posts.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. ArXiv:2306.05685 [cs].

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks. ArXiv:2304.10145 [cs].

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023. Can Large Language Models Transform Computational Social Science? ArXiv:2305.03514 [cs] version: 1.

A Ambiguities in Verifiability

In political speeches and social media, not all statements are necessarily grounded with enough specific information and are undoubtedly verifiable. Many statements are a mixture of specificity and vagueness, which makes verifiability hard to define. The specificity required for verification may vary based on the topic. But generally, the more specific information a fact contains, the more verifiable it is. For example, a vague statement like "Birmingham is small" tends to be a not verifiable opinion since it lacks specificity (e.g., the standard of “being small”). In contrast, "Birmingham is small in terms of population compared to London" offers a clearer path for verification by comparing the population sizes of both cities. Such ambiguity in verifiability results in different expert annotations. To resolve disagreement and obtain gold labels, we have the experts debate “whether a statement provides enough specific information to guide fact-checkers in verification” to achieve agreement.

In the following list, we showcase some examples with vague verifiability. We rely on our experts’ critical thinking and common sense to determine their verifiability.

E1. *“I promised that our roads would be the envy of the nation.”* Analysis: “envy of the nation” seems to be an unverifiable subjective expression. However, this is a part of the speaker’s pledge about improving infrastructure and can be verified by comparing the roads with those in other states.

E2. *“Evil acts against innocent people in the places where we once ran errands or recreated have also made us feel less safe.”* Analysis: the speaker claims the existance of evil acts which seems verifiable. However, no specific details are mentioned and different people may interpret or define “evil act” differently. Therefore, it is hard to verify.

E3. *“In my budget proposals, we will fully fund our rainy-day accounts.”* Analysis: the "rainy-day account." seems to be an unspecific metaphor which is hard to verify. However, we know from the context that the speaker claims to fund emergency cases (i.e., rainy days). Therefore, it tends to be verifiable.

E4. *"Ensuring society provides a hand up when people need help."* Analysis: it seems that the speaker is pledging a helpful society. However, nothing specific is mentioned, making this claim hard to verify.

E5. *"Folks, no doubt, the last couple of years have been especially trying for our medical professionals."* Analysis: at the first glance, the medical professionals' personal feeling seems subjective and not verifiable. However, as COVID is a public event, this can be verified by checking data related to the workload, stress levels, and overal conditions of medical professionals.

E6. *"Authoritarian and illiberal impulses aren't just rising overseas, they've been echoing here at home for some time."* Analysis: it claims the arising of authoritarian and illiberal impulses. However, no specific events or details are mentioned thus different people may interpret those things differently, making it hard to verify.

E7. *"We are finally going to fix the darn roads."* Analysis: "darn roads" is a subjective expression. However, the speaker's pledge of improving (at least some) roads is verifiable.

E8. *"I'll call this nonsense what it is, and that is an un-American, outrageous breach of our federal law."* Analysis: the speaker interprets the COVID vaccination plan as "an un-American, outrageous breach of federal law", which seems verifiable by checking laws. However, this is a controversial issue where different people may have different interpretations of the laws. And importantly, no specific legal provisions are mentioned. Therefore, it leans towards unverifiable opinion.

We make all our experts' annotations publicly available. Challenging samples can be found by locating disagreements. Though we tried our best to make the annotation accurate, errors may still occur due to their challenging nature. We encourage future work to improve our definitions to resolve the existing vagueness.

B Annotation Guideline

The task is to select verifiable statements from political speeches for fact-checking. Given a statement from a political speech and its context, answer two questions following the guidelines. Your annotation will be used to evaluate an LLM-based annotation assistant for factual claim definition.

B.1 Guidelines

Context: Make sure to consider a small context of the target statement (the previous and next sentence) when annotating. Some statements require context to understand the meaning. For example:

E1. "*... Just consider what we did last year for the middle class in California, sending 12 billion dollars back – the largest state tax rebate in American history. But we didn't stop there. We raised the minimum wage. We increased paid sick leave. Provided more paid family leave. Expanded child care to help working parents ...*" Without the context, the underlined sentence seems an incomplete sentence. With the context, we know the speaker is claiming a bunch of verifiable achievements of their administration.

E2. "*... When I first stood before this chamber three years ago, I declared war on criminals and asked for the Legislature to repeal and replace the catch-and-release policies in SB 91. With the help of many of you, we got it done. Policies do matter. We've seen our overall crime rate decline by 10 percent in 2019 and another 18.5 percent in 2020! ...*" The underlined part claims that the policies against crimes have been "done", which is verifiable. It needs context to understand it.

Opinion with Facts: Opinions can also be based on factual information. For example:

E1. "*I am proud to report that on top of the local improvements, the state has administered projects in almost all 67 counties already, and like I said, we've only just begun.*" The speaker's "proud of" is a subjective opinion. However, the content of pride (administered projects) is factual information.

E2. "*I first want to thank my wife of 34 years, First Lady Rose Dunleavy.*" The speaker expresses their thankfulness to their wife. However, there is factual information about the first lady's name and the length of their marriage.

What is verifiable? The verifiability of the factual information depends on how specific it is. If there is enough specific information to guide a general fact-checker in checking it, the factual information is verifiable. Otherwise, it is not verifiable. For example:

E1. "*Birmingham is small.*" is not verifiable because it lacks any specific information for determining veracity. It leans more toward subjective opinion.

E2. "*Birmingham is small, compared to London*" is more verifiable than E1. A fact-checker can retrieve the city size, population size ... etc., of London and Birmingham to compare them. However, what to compare to prove Birmingham's "small" is not specific enough.

E3. "*Birmingham is small in population size, compared to London*" is more verifiable than E1 and E2. A fact-checker now knows it is exactly the population size to be compared.

When does an opinion explicitly present a fact? Many opinions are more or less based on some factual information. However, some facts are explicitly presented by the speakers, while others are not. Explicit presentation means the fact is directly entailed by the opinion without extrapolation:

E1. "*The pizza is delicious.*" This opinion seems to be based on the fact that "pizza is a kind of food". However, this fact is not explicitly presented.

E2. "*I first want to thank my wife of 34 years, First Lady Rose Dunleavy.*" The name of the speaker's wife and their year of marriage are explicitly presented.

Along with these guidelines, definitions in Section 2 are also presented to the annotators.

B.2 Annotation Questions

Q1. Does the target statement explicitly present any verifiable factual information?

- A - Yes, the statement contains factual information with enough specific details that a fact-checker knows how to verify it. E.g., Birmingham is small in population compared to London.
- B - Maybe, the statement seems to contain some factual information. However, there are certain ambiguities (e.g., lack of specificity) making it hard to determine the verifiability. E.g., Birmingham is small compared to London. (lack of details about what standard Birmingham is small)
- C - No, the statement contains no verifiable factual information. Even if there is some, it is clearly unverifiable. E.g., Birmingham is small.

If your answer to Q1 is B - Maybe, then please answer Q2 below:

Q2. Do you think this statement needs fact-checking of any degree? In other words, does it lean more to checkable facts or subjective opinions?

- A - Yes, it leans more to facts that need checking.
- B - No, it leans more toward subjective opinion and does not need a fact-check.

Samples labeled with A and B/A are positive samples, while those with C and B/B are negative samples.

C AFaCTA Prompts

Following are the prompts of AFaCTA. In all prompts, we always include the previous and next sentence of the target statement if the context is available. "{sentence}", and "{context}" are variables to be substituted with the target sentence and its contexts correspondingly. When annotating Twitter data, we simply change "political speech" to "Twitter" and remove the specifications about contexts (see exact prompts in our code base).

C.1 System Prompt

```
You are an AI assistant who helps fact-checkers
     to identify fact-like information in
     statements.
```

C.2 Step 1: Direct Classification

```
Given the <context> of the following <sentence>
      from a political speech, does it contain any
      objective information?

<context>: "...{context}..."
<sentence>: "{sentence}"

Answer with Yes or No only.
```

C.3 Step 2: Fact-Extraction CoT

In this prompt, we use the categorical definition for facts in Konstantinovskiy et al. (2020), removing the final category of "other statements you think are claims" to reduce uncertainty.

```
Statements in political speech are usually based
      on facts to draw reasonable conclusions.

Categories of fact:
C1. Mentioning somebody (including the speaker)
     did or is doing something specific and
     objective.
C2. Quoting quantities, statistics, and data.
C3. Claiming a correlation or causation.
```

```
C4. Assertion of existing laws or rules of
     operation.
C5. Pledging a specific future plan or making
     specific predictions about future.

Please first analyze the objective and
     subjective information that the following <
     statement> (from a political speech) covers.
Then extract the fact that the <statement> is
     based on.
Then carefully reason about if the extracted
     fact is objectively verifiable.
Finally answer if the fact falls into the above
     categories (C1 to C5) or not (C0).

Context for <statement> to help you understand
     it better: "{context}"
<statement>: "{sentence}"

Format your answer in JSON with the following
     keys in order:
{{
    "ANALYSIS": "What are the objective and
         subjective information that <statement>
         covers?",
    "FACT_PART": "The extracted fact.",
    "VERIFIABLE_REASON": "Detailed reason about
         the extracted fact's verifiability. Note
          that a fact lacks important details or
         can be interpreted differently is not
         objectively verifiable. Future plans/
         pledge (C5) that are specific and clear
         can be verifiable. Citing others' words
         is verifiable and falls into C1. ",
    "VERIFIABILITY": "A boolean value indicates
         the verifiability.",
    "CATEGORY": "C1 to C5, or C0."
}}
```

Following is an example showing GPT-4's analyses of a statement, given its context:

```
Target statement: ``We are so thankful that we
     haven't suffered any loss of life, and it's
     always heartening to see and hear stories of
      Alaskans pitching in to help each other.''

GPT-4's reply: \{
    "ANALYSIS": "The objective information in
         the statement is that there has been no
         loss of life due to the storms. The
         subjective information is the speaker's
         expression of gratitude and the
         heartening stories of Alaskans helping
         each other.",
    "FACT_PART": "There hasn't been any loss of
         life due to the storms.",
    "VERIFIABLE_REASON": "The fact that there
         hasn't been any loss of life due to the
         storms can be verified by checking
         official records and reports from
         relevant authorities such as emergency
         services, hospitals, and local
         government.",
    "VERIFIABILITY": true,
    "CATEGORY": "C1"
\}
```

C.4 Step 3: Reasoning with Debate

Argument for objectivity:

```
Concisely argue that the following <sentence>
     from a political speech does contain some
     objective information.

Context of <sentence> in the speech: "...{
     context}..."
<sentence>: "{sentence}"
```

Following is an example argument from GPT-4 that there is some objective information in the target statement:

```
Target statement: ``We are so thankful that we
     haven't suffered any loss of life, and it's
     always heartening to see and hear stories of
      Alaskans pitching in to help each other.''

GPT-4's reply: The sentence "The storms have
     required state disaster declarations
     covering almost half of our boroughs and the
      communities they include" does contain
     objective information. It provides specific
     details about the impact of the storms,
     stating that they have led to state disaster
      declarations in nearly half of the boroughs
      in the state. This is a factual claim that
     can be verified by checking the official
     records of disaster declarations.
```

Argument for subjectivity:

```
Concisely argue that the following <sentence>
     from a political speech does not contain any
      objective information.

Context of <sentence> in the speech: "...{
     context}..."
<sentence>: "{sentence}"
```

Following is an example argument from GPT-4 that there is NO objective information in the target statement:

```
Target statement: ``We are so thankful that we
     haven't suffered any loss of life, and it's
     always heartening to see and hear stories of
      Alaskans pitching in to help each other.''

GPT-4's reply: The sentence "The storms have
     required state disaster declarations
     covering almost half of our boroughs and the
      communities they include" does not contain
     objective information because it lacks
     specific details. It does not provide the
     exact number or names of the boroughs
     affected, the specific nature of the
     disaster declarations, or the precise extent
      of the damage. Without these details, the
     statement remains vague and subjective.
```

Final judgment:

```
Two AI assistants are debating about whether the
      following <sentence> (from a political
     speech) contains any objectively verifiable
     information.

Context of <sentence> in the speech: "...{
     context}..."
<sentence>: "{sentence}"

Assistant A's View: "{assistant_a}"

Assistant B's View: "{assistant_b}"

Based on the above, does <sentence> contain any
     objectively verifiable information? Which
     perspective do you align with more closely?
Please reply with "Lean towards A", or "Lean
     towards B" only.
```

Following is an example of GPT-4's judge given arguments of different perspectives:

```
Target statement: ``We are so thankful that we
     haven't suffered any loss of life, and it's
     always heartening to see and hear stories of
      Alaskans pitching in to help each other.''

GPT-4's reply: Lean towards A.

Here the objective argument is assigned to
     assistant A.
```

D AFaCTA Tie-Breaking

All samples with 1.5 votes are labeled as non-factual claims. This tie-break is randomly chosen for two reasons: (1) tie-break does not influence AFaCTA's performance as it does not affect the perfectly consistent subset (i.e., the subset annotated with best confidence and may not need further human supervision). (2) Samples with 1.5 votes are the most confusing ones reported by AFaCTA, which require human supervision the most. They should not be dogmatically classified in practice. We only do it here to compute AFaCTA's annotation performance on the test sets.

E Details of Evaluation Metrics

This section presents the formulas of metrics used in Section 5. For conciseness, only formulas on perfectly consistent samples $S^{\mathcal{M}}_{con}$ are showcased. Similar formulas are applied for inconsistent samples $S^{\mathcal{M}}_{inc}$ and all samples S.

Average accuracy of human expert on perfectly consistent samples $S^{\mathcal{M}}_{con}$ is calculated as:

$$Acc^{H}_{con} = \frac{1}{2} \sum_{h \in \{h1, h2\}} acc\_score(G_{con}, P^{h}_{con}) \tag{1}$$

where G_{con} and P^{h}_{con} denote the gold labels and human-annotated labels of samples where AFaCTA achieves perfect self-consistency; and $h1$ and $h2$ denotes two human experts.

Accuracy of AFaCTA against gold label on $S^{\mathcal{M}}_{con}$ is calculated as:

$$Acc^{\mathcal{M}}_{con} = acc\_score(G_{con}, P^{\mathcal{M}}_{con}) \tag{2}$$

where $P^{\mathcal{M}}_{con}$ denotes AFaCTA's prediction on perfectly consistent samples.

Agreement (Cohen's Kappa) between human annotators on $S^{\mathcal{M}}_{con}$ is calculated as:

$$Kappa^{H}_{con} = cohen\_kappa(P^{h1}_{con}, P^{h2}_{con}) \tag{3}$$

Average Cohen's Kappa between AFaCTA and two human annotators on $S^{\mathcal{M}}_{con}$ is calculated as:

$$Acc^{M}_{con} = \frac{1}{2} \sum_{h \in \{h1, h2\}} cohen\_kappa(P^{h}_{con}, P^{M}_{con}) \tag{4}$$

We use Sci-Kit Learn's accuracy and Cohen's Kappa implementations to calculate all metrics.

F AFaCTA with Open-sourced LLMs

We tried AFaCTA framework on two popular open-sourced LLMs: Llama-2-chat-13b (Touvron et al., 2023) and zephyr-7b-beta (Tunstall et al., 2023). Results are presented in Table 4. For both models, we use the official checkpoints on huggingface and conduct greedy decoding when inference. We observe that both models suffer from heavy position bias in AFaCTA step 3: when putting arguments for verifiable and unverifiable to different positions, llama-2-chat-13b and zephyr-7b-beta predict inconsistently in 99% and 97% cases correspondingly. Therefore, there are seldom annotations with perfect consistency, and the consistency-based annotation strategy of AFaCTA does not help.

We also observe that zephyr-7b-beta achieves better performance than GPT-3.5 on CheckThat!2021-dev, showing the potential of using open-sourced LLMs as annotators. In future work, we will explore fine-tuning open-sourced LLMs to mitigate the position bias problem and improve annotation quality.

G Hyperparameter Settings

For OpenAI models, we always use gpt-3.5-turbo-0613 and gpt-4-0613. We use a temperature of 0, and top-p of 1 for all experiments except the self-consistency CoT (Wang et al., 2023) experiments where we use a temperature of 0.7. We make all LLM generations publicly available. We always use a random seed of 42 if not specified. For open-sourced LLM inference, we use greedy sampling, a top p of 1, and a maximum generation length of 3072.

H Performance of Each AFaCTA Step

We compute the annotation performance of each AFaCTA reasoning step. For Step 3, we average the scores of labels 3.1 and 3.2 (see Figure 1). The results are presented in Table 5. It can be observed that Step 1, though simple, achieves promising performance. It outperforms other steps by a wide margin with GPT-4.

I Self-Consistency CoT

We use the following prompt to generate Self-consistency CoT. It keeps most of the prompt template of AFaCTA Step 1 to make them comparable. We use a temperature of 0.7 to sample different CoTs.

```
Given the <context> of the following <sentence>
     from a political speech, does it contain any
      objective information?

<context>: "...{context}..."
<sentence>: "{sentence}"

Format your reply as follows:
```

	PoliClaim$_{test}$			CheckThat!2021-dev		
	Agreement	**Accuracy**	**Consistency**	**Agreement**	**Accuracy**	**Consistency**
zephyr-7b-β	0.205	66.18	0.49	0.539	77.86	5.00
llama-2-13b-chat	0.306	56.74	0.00	0.260	50.71	1.43
GPT-3.5	0.510	76.74	43.38	0.359	69.29	44.29
GPT-4	0.615	86.27	48.78	0.437	86.43	57.85

Table 4: The performance of AFaCTA with close- and open-source models. We report the average Cohen's Kappa with human experts for agreement, and the accuracy scores are in percentage. We also report the portion of perfectly consistent annotations reported by each model in percentage, which can be found in the consistency column.

	Step 1		**Step 2**		**Step 3**	
	Agreement	**Accuracy**	**Agreement**	**Accuracy**	**Agreement**	**Accuracy**
GPT-3.5	0.458	73.16	0.452	78.06	0.546	66.42
GPT-4	0.633	85.54	0.437	79.90	0.630	73.28

Table 5: The performance of each AFaCTA steps. Similar to Table 3, we report the average Cohen's Kappa with human experts for agreement, and the accuracy scores are in percentage.

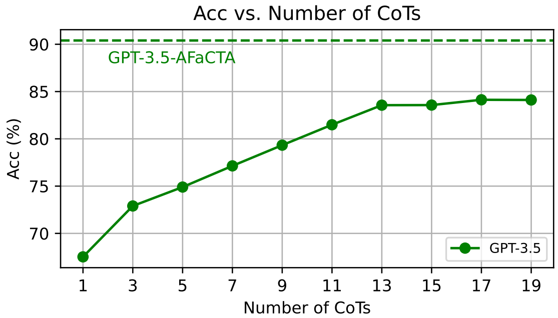

Figure 5: We notice that in Figure 2, GPT-3.5's accuracy on the perfectly consistent set does not seem to converge with 11 voters. So we extend the number of CoTs to 19, observing that the accuracy converges to 84.1%.

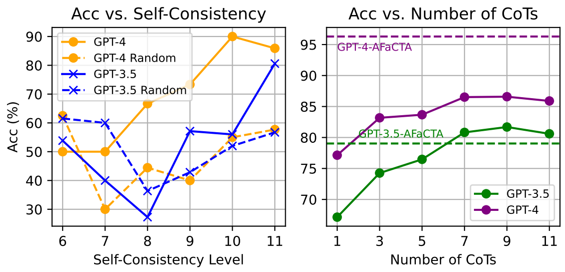

Figure 6: Self-consistency CoT experiments on CheckThat!-2021-dev. Same metrics are reported as Figure 2.

```
[Chain of thought]: your step-by-step reasoning
      about the question
[Answer]: a single word yes or no
```

J Experiments on Social Media Domain

We compare AFaCTA's annotation performance with human experts on the re-annotated CheckThat!-2021 development set. We have chosen this small set of social media data due to the limitation of the annotation budget.

Similar observations as PoliClaim$_{test}$ can be drawn. GPT-4 AFaCTA outperforms experts on perfectly consistent samples and underperforms on inconsistent samples. GPT-3.5 also achieves a moderate agreement with human experts on perfectly consistent samples. Error analysis shows that GPT-3.5's error concentrates on false negatives, similar to its behavior in the political speech domain (see Table 12).

We also conduct the self-consistency CoT experiments on CheckThat!-2021-dev to verify the importance of a diversified source of self-consistency. The results are shown in Figure 6. It can be observed that the level of self-consistency calibrates accuracy, and the 3 predefined reasoning paths outperform automatically generated ones. One discrepancy is that self-consistency CoT slightly outperforms GPT-3.5 AFaCTA when sampling more than 7 reasoning paths. We attribute this to GPT-3.5's heavier hallucinations on Twitter domain (see Table 12 where it fails to identify apparent factual information). Therefore, complicated reasoning paths like AFaCTA Step 3 might be challenging in many cases.

Importantly, due to the annotation budget, our experimental dataset on the social media domain is limited. We leave the extensive analysis of this domain to future work.

K Fine-tuning Settings

For all RoBERTa and DistilBERT fine-tuning experiments, we keep all settings the same except for the training data. All models are fine-tuned for 5 epochs with a batch size of 64. We do not

	$\mathbf{S}$ (100†/100‡)		$\mathbf{S}^{\mathcal{M}}_{\mathbf{con}}$ (44.29†/57.85‡)		$\mathbf{S}^{\mathcal{M}}_{\mathbf{inc}}$ (55.71†/42.15‡)	
	Agreement	**Accuracy**	**Agreement**	**Accuracy**	**Agreement**	**Accuracy**
GPT-3.5	0.359	69.29	**0.584**	79.03	0.205	61.54
GPT-4	0.437	86.43	0.566	**96.30**	0.280	72.89
Experts	**0.579**	**92.86**	0.514†/0.540‡	91.13†/95.68‡	**0.638**†/**0.536**‡	**94.23**†/**88.98**‡

Table 6: AFaCTA's performance on our re-annotated CheckThat!-2021-dev. Similar rows, columns, and scores are reported as Table 3.

conduct checkpoint selection. For other hyperparameters, we keep the default setting of huggingface TrainingArgument: a learning rate of 5e-5, a max_grad_norm of 1, no warm-up and weight decay, etc. We use the huggingface checkpoints of "roberta-base" and "distilbert-base-uncased". All experiments are conducted on a node with 4 32G V100 GPUs. It takes roughly 0.1 GPU hour to train a classifier. In this work, we always use Sci-kit Learn for score computing.

L Statistical Significance Test

We conduct a statistical significance test to show that different training set combinations of PoliClaim$_{gold}$, PoliClaim$_{silver}$, and PoliClaim$_{bronze}$ lead to statistically significant differences in fine-tuning claim detectors. We first conduct a Student-t test for each training combination based on the results of three random seeds and then aggregate p-values using Fisher's method. For example, to compare "only PoliClaim$_{gold}$" vs. only "PoliClaim$_{silver}$", we use the following formula:

$$p_{x00} = \text{Student-t}(\{Acc^{r}_{x00g}\}, \{Acc^{r}_{x00s}\}) \quad (5)$$

$$p_{agg} = \text{Fisher}(p_{100}, p_{200}, ..., p_{2000}) \quad (6)$$

where r denotes random seeds 42, 43, and 44; p_{x00} denotes the p-value of the x00 step; and p_{agg} denotes the aggregated p-value. The aggregated p-values of all comparisons are shown in Table 7. It can be seen that all observations in Section 5.5 and Appendix M are statistically significant. Scipy's implementations for Student-t test and Fisher's Method are used.

We do not conduct statistical tests on experiments of Section 5.1 as obtaining independent samples of human / GPT-4 annotation can be very costly, and OpenAI API does not support random seeds at the moment of experimenting.

M Further Fine-tuning Experiments

This section provides more supplementary results of the experiments in Section 5.5.

Comparison			**RoBERTa**	**DistilBERT**
Only S	<	Only G	5.54e-3*	8.89e-5**
Only B	<	Only S	2.39e-36**	5.79e-51**
Only B	<	Only G	1.88e-20**	6.03e-29**
500 G + B	<	500 G + S	1.50e-28**	1.82e-30**
1000 G + B	<	1000 G + S	8.13e-13**	3.30e-8**
1500 G + B	<	1500 G + S	2.19e-16**	1.69e-15**
All G + B	<	All G + S	3.68e-9**	1.36e-13**

Table 7: Statistical significance of performance difference with different train sets. G, S, and B denotes PoliClaim$_{gold}$, PoliClaim$_{silver}$, and PoliClaim$_{bronze}$ correspondingly. By * and **, we denote a p-value smaller than 0.01 and 0.001, respectively.

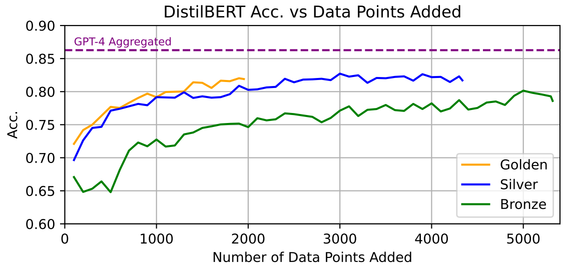

Figure 7: The performance of fine-tuned DistilBERT on PoliClaim$_{test}$ when gradually adding training data of different quality. Same scores are reported as Figure 3.

M.1 Only Golen, Silver, or Bronze

We gradually increase the size of golden, silver, and bronze training data to fine-tune DistilBERT. The results are shown in Figure 7. The same observations can be drawn from Figure 3: perfectly consistent (silver) data achieve a similar growing trend as manually supervised (golden) data, while accuracy grows slower when adding (bronze) inconsistent data.

M.2 Augmenting Gold Data with Silver/Bronze Data

We conduct the data augmentation experiments in Section 5.5 on both RoBERTa (Figure 8) and DistilBERT (Figure 9) with a different number of PoliClaim$_{gold}$ data (500, 1000, 1500, and 1936). Similar conclusions as Section 5.5 can be drawn: perfectly consistent (silver) data are better at aug-

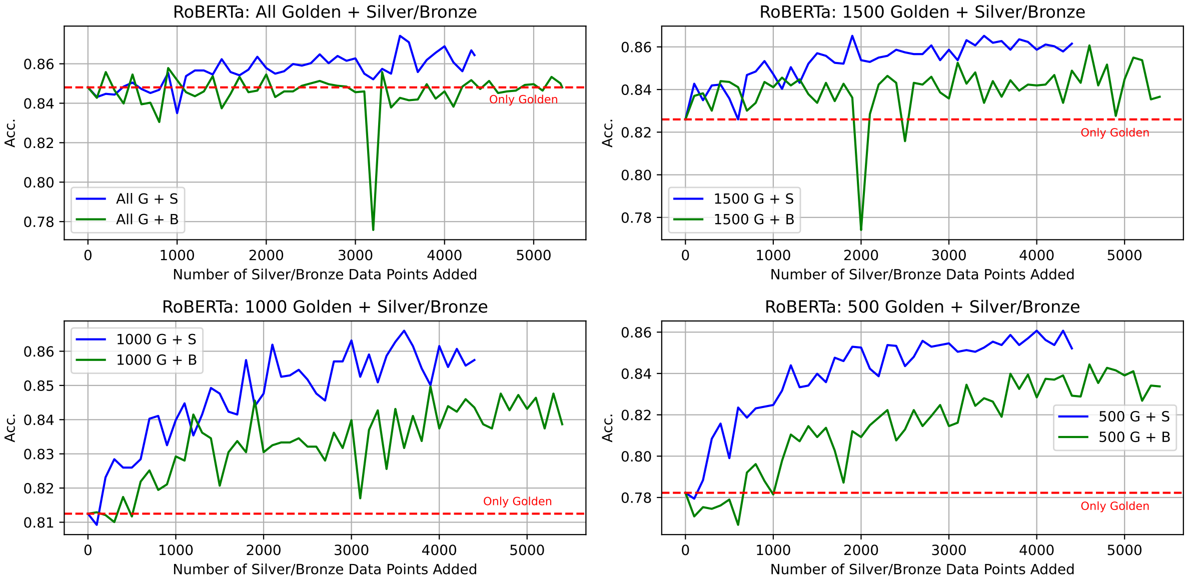

Figure 8: The RoBERTa performance of augmenting a limited number of PoliClaim$_{gold}$ data. An augmented version of Figure 4 with 1000 and 1500 Gold data experiments added.

mentation than inconsistent (bronze) data. Figure 10 also shows a clear trend. When the manual annotation budget is more restricted, more augmentation data are needed to achieve a comparable performance.

In all experiments, the marginal benefit of adding data decreases quicker on DistilBERT than on RoBERTa, as expected. However, we suspect adding more high-quality annotated and diversified data might boost weaker models to outperform stronger models, though the marginal accuracy gain is low. We leave this exploration to future work.

N Error Analyses

We conduct a thorough analysis on GPT-4 and GPT-3.5 AFaCTA. Errors on PoliClaim$_{test}$ can be found in Table 8, Table 9, and Table 10. Errors on CheckThat!-2021-dev can be found in Table 11 and Table 12.

In both domains, we observe that GPT-4 is good at disentangling factual information from speeches or tweets. But it also leads to false positive errors due to over-sensitivity towards factual information. It also makes negative errors due to the lack of full context of the statements. In general, GPT-4 only makes mistakes on confusing samples that lie between factual and non-factual claims.

GPT-3.5's errors concentrate on false negatives. It regularly hallucinates about personal experience and quotations which are explicitly defined in the prompts. It is very conservative in identifying anything as verifiable fact arguing there not enough "specific details" to determine verifiability. However, many facts are already specific enough for verification (see row 2 of Table 9). Sometimes, it also fails to identify facts entangled with opinions (see row 1 of Table 10 and row 1 of Table 12).

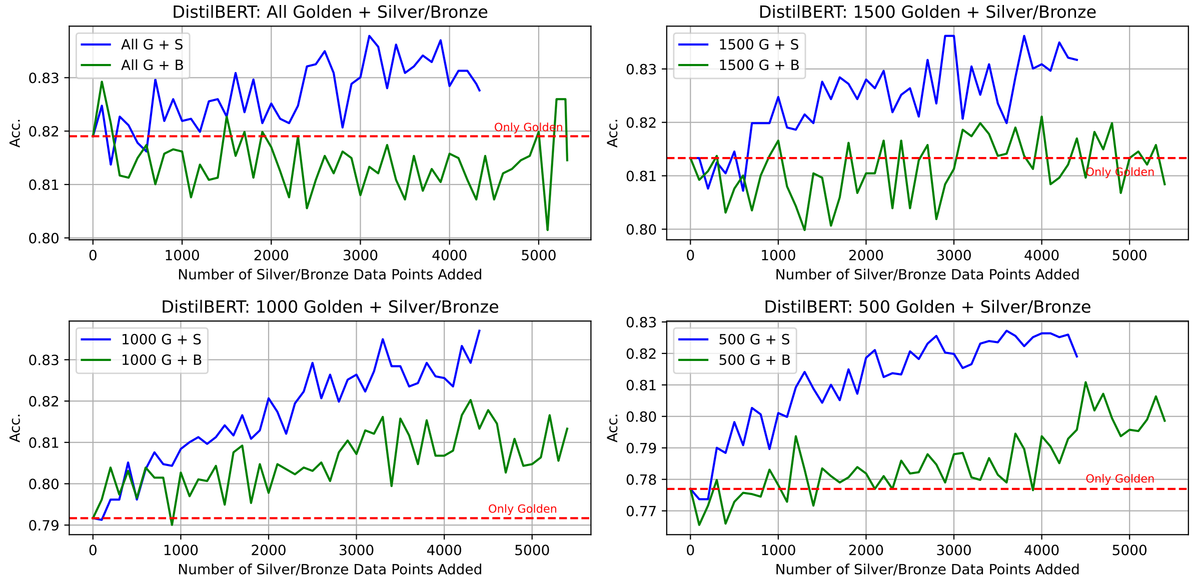

Figure 9: The DistilBERT performance of augmenting a limited number of PoliClaim$_{gold}$ data. The same scores are reported as Figure 8.

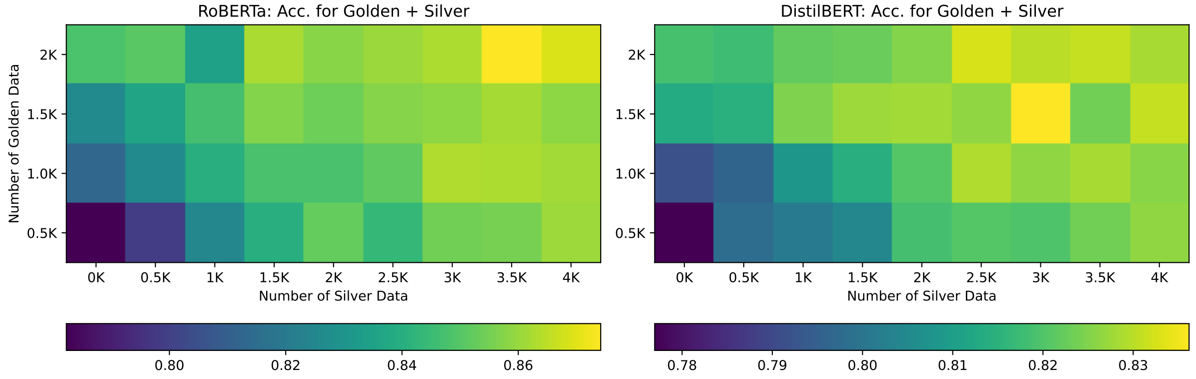

Figure 10: The performance of combining different amount of PoliClaim$_{gold}$ and PoliClaim$_{test}$.

Error Type	GPT-4 Errors and Explanations
False Positive: over-sensitive to granular, unspecific, or not-explicitly-presented facts	***Error 1:*** I just want to thank you, thank you members of the Legislature for all you did these past two years to keep us safe. *Error reason: recognizing ’members of the legislature did some thing’ as fact, which is too vague.* ***Error 2:*** It’s true from the Flatirons to Fishers Peak to Pikes Peak to Longs Peak and beyond. *Error reason: location names are recognized as facts.* ***Error 3:*** Sheriff Pelle, firefighters, and emergency responders, please stand so we can thank you for the lifesaving work that you do every day. *Error reason: people’s appearance at the event is recognized as a fact, which is not explicitly presented.* ***Error 4:*** I’m glad to be back at the capitol addressing the Legislature in person, and I thank you for the invitation to speak to you tonight. *Error reason: identify the speaker’s back and addressing legislature as facts.*
False Negative: not enough context	***Error 5:*** It’s the result of great investment decisions, policies, vision, and direction. *Error reason: “it” here refers to the return of pension fund, which is in a far context. But AFaCTA only considers a one-sentence context.* ***Error 6:*** Alaskans won’t accept that we can’t get anything done because it’s an election year. *Error reason: it claims a fact that this year is an election year, but the model comprehends this as a hypothetical condition, due to its lack of context that 2022 is the election year for Alaska.*

Table 8: All errors made by GPT-4 AFaCTA on PoliClaim$_{test}$. Statements are highlighted in yellow. The reasons for making errors are written in *italics*.

Error Type	GPT-3.5 Errors and Explanations
False Positive: over-sensitive to unspecific facts	***Error 1:*** The fresh mountain air that so many people associate with Colorado isn't a given. *Error reason: identifying "the fresh air is not a given" as a fact, which is unspecific and leans towards unverifiable opinion.*
False Negative: not enough specific detail or context and thus not verifiable	***Error 2:*** When our federal government overreached, we found a way to fight back. *Error reason: the model argues it lacks details of "overreach".*
	Error 3: While our work is far from over, we have made significant progress thanks to the Rebuild Alabama Act. *Error reason: the model argues it lacks important details of "significant progress".*
	Error 4: Folks, no doubt, the last couple of years have been especially trying for our medical professionals. *Error reason: the model argues the "especially trying" lacks detail.*
	Error 5: I am proud that my Administration, with the support of the Legislature, is doing more to make significant improvements in mental health care than any since Governor Lurleen Wallace in the 1960s. *Error reason: the model argues the "significant improvements" lack detail.*
	Error 6: At times, her schoolwork and distance from her home state made her wonder if she should give up her Miss Alaska title. *Error reason: the model argues the "significant improvements" lack detail.*
	Error 7: It's the result of great investment decisions, policies, vision, and direction. *Error reason: not enough context about "it".*
	Error 8: Together with these partners, we'll build a stronger, more durable health care system in Alaska that can respond to most any situation. *Error reason: the model argues it lacks details about the plan.*
	Error 9: At the same time, our ability to increase production is under attack from Washington, DC, and federal courts that side with extremist environmental groups. *Error reason: the model argues it lacks details about the "attack".*
	Error 10: No state has been targeted more by the current administration than our Great State of Alaska. *Error reason: the model argues it lacks details about specific actions.*
	Error 11: No state has been targeted more by the current administration than our Great State of Alaska. *Error reason: the model argues it lacks details about specific actions.*
	Error 12: At every turn and since day one of the Biden Administration, this hostility has been perfectly clear. *Error reason: the model argues it lacks details about the "hostility".*
	Error 13: Because no president should have to beg for more oil from the Middle East or Russia's Arctic when we can produce it right here better and safer than anywhere else on the planet! This is common sense! *Error reason: the model argues it lacks details or evidence about "the US produces better oil".*
	Error 13: Many of them have been with us for so long that they've almost been normalized in Alaska, as almost unsolvable. *Error reason: lacking context and details about "long issues".*
	Error 14: I will always stand between Alaskans and a federal government that violates our God-given rights and exceeds its constitutional authority. *Error reason: the model argues it lacks details about specific actions of the federal government and the speaker's future action.*
	Error 15: I envision an Alaska where our cost of energy is no longer the second-highest in the nation, but one of the lowest. That's my vision. I hope it is yours as well. *Error reason: the model argues it lacks details about the definition of "second-highest" and "lowest".*
	Error 16: I've seen it in the men and women on the frontlines of this pandemic who have helped us achieve one of the shortest shutdowns and one of the lowest death rates in the country. *Error reason: the model argues it lacks details about "shortest" and "lowest".*
	Error 17: And because we want to lead by example, we are saving Coloradans money by making your State Government more efficient and effective. *Error reason: the model argues it lacks details about "efficient" and "effective".*
	Error 18: Just as an earthquake is followed by aftershocks, we know that the overarching crisis of the pandemic has led to many other crises, perhaps lesser seen, but no less important to address. *Error reason: the model argues it lacks details about "the crises".*
	Error 19: We owe it to the people of Colorado to improve safety and make Colorado truly one of the ten safest states in the nation over the next five years. *Error reason: the model argues it lacks details about the speaker's plan.*
	Error 20: No other place offers opportunity to so many from such diverse backgrounds. *Error reason: the model argues it lacks specific details.*
	Error 21: It's that, as our businesses grow, we don't leave our workers behind. *Error reason: the model argues it lacks specific details about business growth.*
	Error 22: By creating choices - real choices - for parents, and unprecedented support for their kids. *Error reason: the model argues it lacks specific details about the choices and supports.*

Table 9: The only false positive error and the major type of false negative errors made by GPT-3.5 AFaCTA on PoliClaim$_{test}$.

Error Type	GPT-3.5 Errors and Explanations
False Negative: understand facts as opinions or fail to identify facts entangled with opinions	***Error 23:*** They plan their lives around hunting season, or fishing season; construction season, or tourism season.But not election season. *Error reason: the model misunderstands it as the speaker's opinion. But people's lifestyles and priorities can be verified with related surveys or studies.* ***Error 24:*** A future where a dynamic, multi-modal transportation system meets the needs of our growing population. *Error reason: the model fails to identify "our growing population" as a fact.* ***Error 25:*** When I was elected Governor, I knew that I would be remembered not for who I was, where I came from, or even what I said at events like this, but for what I did to make a meaningful, measurable, positive impact on the lives of Coloradans. *Error reason: the model fails to identify that "the speaker is elected as the governor" is a fact.* ***Error 26:*** But over time, we've learned we can't solve big problems like climate change situationally, with short-term thinking. *Error reason: the model fails to identify the causality claim about short-term thinking and big problems.* ***Error 27:*** But at a time, when we've been heating and burning up, one thing we cannot do is repeat the mistakes of the past by embracing polluters. *Error reason: the model fails to recognize the fact of embracing polluters in the past.*
False Negative: hallucinate about personal experience and citation	***Error 28:*** At times, her schoolwork and distance from her home state made her wonder if she should give up her Miss Alaska title. *Error reason: the model argues the personal experience is subjective.* ***Error 29:*** "A lot of people," she said, "don't recognize that their low points are what are going to propel them to their future. *Error reason: subjective personal experience.* ***Error 30:*** I agree with former Governor Jay Hammond that the government should never take more from the Permanent Fund than is distributed to the people of Alaska. *Error reason: fail to detect the citation.* ***Error 31:*** She is in a healthy marriage and is reconnecting with her children. *Error reason: consider personal experience as unverifiable.* ***Error 32:*** "Dad," Catherine said, "Alaska has so much to offer." *Error reason: fail to detect the citation.* ***Error 33:*** Still, she found the strength to take down the shooter, ending his violent killing spree and saving many precious lives. *Error reason: consider personal experience as unverifiable.*
False Negative: hallucinate about rhetoric	***Error 34:*** They're wondering how we've come to a place where the PFD is nothing more than what's left over after government takes the lion's share. *Error reason: fail to understand the metaphor.*

Table 10: Other types of false negative errors made by GPT-3.5 AFaCTA on PoliClaim$_{test}$ other than not-enough-detail/context.

Error Type	GPT-4 Errors and Explanations
False Positive: over-sensitive to granular, unspecific, or not-explicitly-presented facts	***Error 1:*** Requesting to work from home because of the #coronavirus is what's called a "reasonable accommodation." You have disabled people to thank for that. Remember this moment in history the next time you think Accessibility laws are too "burdensome" to be abided. *Error reason: the model recognizes the concept of "reasonable accommondation" and the existence of "accessibility laws" as facts, which are not explicitly presented by the post.*
False Negative: misunderstand verifiable fact as subjective interpretation	***Error 2:*** "Last week Trump told aides he's afraid journalists will try to purposefully contract #coronavirus to give it to him on Air Force One." https://t.co/sS1MZR6D7w *Error reason: GPT-4 understands it as the tweet author's subjective interpretation of Trump's words. However, we think that it can be verified by checking whether Trump said the words or not.* ***Error 3:*** Due to #coronavirus, media advises the economy must tank, the people must panic, Trump must be blamed, Biden must be secreted away from the public, and Bernie must cease rallies. I wonder why people do not trust the media's motives on this? *Error reason: GPT-4 understands it as the tweet author's subjective interpretation of the media's advice. However, we think it can be verified by checking if there are media suggesting such information.*

Table 11: All errors made by GPT-4 AFaCTA on CheckThat!-2021-dev.

Error Type	GPT-3.5 Errors and Explanations
False Negative: fail to identify facts entangled with opinions	***Error 1:*** Who would you prefer to lead our nation's response to the growing #coronavirus threat? *Error reason: fail to identify "the growing coronavirus threat".* ***Error 2:*** It was a really really really really really really really really really really really really really really really really really really really bad idea to elect Donald Trump President of the United States. #TrumpVirus #TrumpCrash #TrumpRecession #COVID19 #coronavirus *Error reason: fail to identify "elected Donald Trump Predisent of the US".* ***Error 3:*** If people who are infected by corona virus in SA were black, their names, homes street: pictures will be all over Social Media. White privileges goes a long way. Wait for a case of a black person, they will mention even his location they won't say WC, they'll say Gugulethu ext 5 *Error reason: fail to identify the correlation between the infected persons' race and their suffers.* ***Error 4:*** @realDonaldTrump On a morning when Americans are terrified, the markets are gonna historically crash and we need LEADERSHIP...all you've done is hate-tweet BULLSHIT about Sanders, Warren, Biden, Democrats, Schumer. the media and now Obama. Your incompetence is staggering.... #Trump #coronavirus *Error reason: although have subjective interpretations, the facts that "markets are gonna historically crash" and Trump commented something about others is verifiable.* ***Error 5:*** Dear BBC, I want to fight for you. I know you're more than news (which has been questionable) you're also great drama, documentaries, kids tv etc But don't make me question that by inviting Farage on to talk about Corona FFS!! Show you'll fight for your integrity yourselves! *Error reason: fail to identify the verifiable fact that BBC invited Farage is verifiable.* ***Error 6:*** Thread 1: One wonders about the racial politics of this corona outbreak. What would have happened had it been blacks who came into the country with the virus? Would they hav been allowed to "self quarantine"? If the virus was from the continent; wouldn't travels be banned by now? *Error reason: fail to identify "this corona outbreak" and the practice of "self-quarantine".* ***Error 7:*** The total Iranian #COVID19 case-count is in the hundreds of thousands, perhaps millions, according to my estimates (detailed at the link). This raises an important question: if there are two million cases, where are all the bodies? https://t.co/nHYbQlXlVC *Error reason: GPT-3.5 understands it as the tweet author's subjective interpretation about the numbers. However, it can be verified by checking details in the link and reliable data source.* ***Error 8:*** Due to #coronavirus, media advises the economy must tank, the people must panic, Trump must be blamed, Biden must be secreted away from the public, and Bernie must cease rallies. I wonder why people don't trust the media's motives on this? *Error reason: GPT-3.5 understands it as the tweet author's subjective interpretation about the media's advice. However, we think it can be verified by checking if there are medias suggesting such information.*
False Negative: fail to comprehend claims about attached links	***Error 9:*** Public Safety Announcement Fighting #CoronaVirus. We have to do this together. Wishing good health to all of you! Love, Vijay. https://t.co/fbafmmtq8S *Error reason: this tweet claims that the link contains a public safetly announcement fighting COVID, which is verifiable.* ***Error 10:*** This thread needs to fly. It shows how the legacy media is USING covid-19 as a political weapon and even how the SAME reporters are contradicting themselves. This. Is. SICK. https://t.co/Werq544xii *Error reason: this tweet claims the content of the link, which is verifiable.*
False Negative: fail to identify personal experience or citation	***Error 11:*** Beware the spread of coronavirus and Fox News pandemic propaganda. Trish Regan's melodramatic rant decrying Dems and MSM for allegedly exploiting #COVID19 as "another attempt to impeach the President." Yank this dangerous shrew off the air. #Trumpdemic https://t.co/6B60RLMIS0 *Error reason: fail to identify the personal experience of Trish Regan.* ***Error 12:*** I keep bumping into this problem. I want to be able to stand up and unequivocally defend the BBC. But it has repeatedly shut out radical voices and crucial issues while providing a massive platform for the alt-right to spout ill-informed nonsense. It is hard to love. https://t.co/A1pQMsqDxV *Error reason: The interpretation about BBC's behavior is subjective. But the tweet author's previous stance might be verifiable by checking their previous statements.* ***Error 13:*** @keywilliamss One African man actually got Corona and was cured in like a week. Health care officials are "baffled as to why Africa is virtually unscathed." Which is... kinda racist that they expected it to be but lemme hush URL: https://t.co/yvP0DUXDiX *Error reason: fail to identify the African man's experience and the quotation of health care officials' speech.*

Table 12: All errors made by GPT-3.5 AFaCTA on CheckThat!-2021-dev.