ClaimVer: Explainable Claim-Level Verification and Evidence Attribution of Text Through Knowledge Graphs

Anonymous ACL submission

Abstract

In the midst of widespread misinformation and disinformation through social media and the proliferation of AI-generated texts, it has become increasingly difficult for people to validate and trust information they encounter. Many fact-checking approaches and tools have been developed, but they often lack appropriate explainability or granularity to be useful in various contexts. A text validation method that is easy to use, accessible, and can perform fine-grained evidence attribution has become crucial. More importantly, building user trust in such a method requires presenting the rationale behind each prediction, as research shows this significantly influences people's belief in automated systems. Localizing and bringing users' attention to the specific problematic content is also paramount, instead of providing simple blanket labels. In this paper, we present *ClaimVer, a human-centric framework* tailored to meet users' informational and verification needs by generating rich annotations and thereby reducing cognitive load. Designed to deliver comprehensive evaluations of texts, it highlights each claim, verifies it against a trusted knowledge graph (KG), presents the evidence, and provides succinct, clear explanations for each claim prediction. Finally, our framework introduces an attribution score, enhancing applicability across a wide range of downstream tasks.

1 Introduction

Misinformation and disinformation are longstanding issues, but the proliferation of AI tools that can generate information on demand has amplified these issues. Tools for fact-checking are not keeping pace with sophisticated text generation techniques. Even when they are effective, they lack appropriate explainability and granularity to be useful to users. Studies have shown that explanations are crucial for users to build trust in AI systems

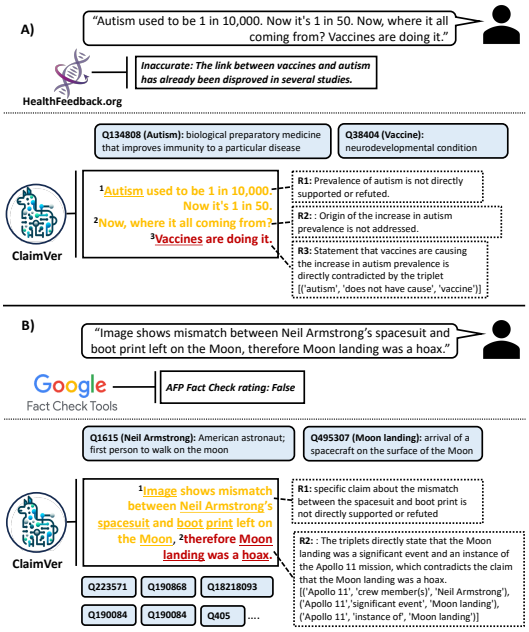

Figure 1: Demonstration of ClaimVer for claim verification and evidence attribution. (A) Text labeled as *Inaccurate* by HealthFeedback and ClaimVer's predictions, rationale, and evidence. (B) Text labeled as *False* by Google Fact Check Tools and ClaimVer's outputs. Predictions are color-coded (amber: extrapolatory, red: contradictory); R_i: rationale; related wiki entities are displayed in boxes.

(Rechkemmer and Yin, 2022; Weitz et al., 2019; Shin, 2021). Therefore, there is a need for a novel human-centric approach to text verification that offers usable and sufficiently granular explanations to inform and educate the user.

Most fact-checkers, including widely used ones in deployment, issue blanket predictions that can lead to user misunderstandings. For instance, in Figure 1 (A), we observe that HealthFeedback[1], a fact-checker for medical text, indicates that a misleading statement about the increase in Autism is inaccurate. However, there are multiple claims made in that text, which are not addressed by this tool. In fact, research does show that Autism cases have

[1] https://healthfeedback.org/

increased, but this is mostly attributed to increased testing (Russell et al., 2015). Our method accurately breaks down the text into multiple claims and shows that the specific claim that vaccines are causing autism is indeed incorrect, attributing it to a fact from the Wikidata (Vrandečić and Krötzsch, 2014). It also provides a clear rationale as to why the first two claims cannot be determined, as there's no conclusive evidence present in the KG. Such granular predictions, supported by justifications, significantly improve user confidence (Rechkemmer and Yin, 2022; Weitz et al., 2019; Shin, 2021).

Similarly, in Figure 1 (B), we notice that Google Fact Check Tools[2] simply provides a blanket label for an utterance denying the moon landing. In contrast, ClaimVer identifies the exact text span that can be conclusively proven incorrect and proceeds to provide specific information about the Apollo 11 mission and its crew members to refute the claim. All verified entities present in the text, along with their Wiki IDs and descriptions, are displayed for user reference.

Prior research (Rashkin et al., 2023; Yue et al., 2023; Thorne et al., 2019; Aly et al., 2021) typically validates text at the paragraph or sentence level without adequately enhancing user awareness by supplying key details such as rationale, match scores, or evidence. A KG-based approach allows for finer granularity, aiding in pinpointing specific inaccuracies like hallucinations in LLM-generated text or false claims in misleading text. Furthermore, if needed, broader-level metrics can be extracted from this detailed attribution.

The assumption of one-to-one mapping between input and reference texts, prevalent in previous methods (Rashkin et al., 2023; Yue et al., 2023; Thorne et al., 2019; Aly et al., 2021), does not hold if the given text consists of claims that can be mapped to more than one source. In contrast, utilizing a KG, which represents a consolidated body of knowledge, results in a more comprehensive evaluation. While most previous methods may not support scenarios with information spread across various references, querying a KG can yield triplets originally sourced from multiple documents. Additionally, procuring the specific spans of text required to evaluate claims from large text sources that may span several pages presents many challenges. In contrast, a KG captures only the most important relationships as nodes and links, providing a more efficient way to evaluate the claims.

[2]https://toolbox.google.com/factcheck/explorer

2 Related Work

Research on validating text has been ongoing for the past decade, while the concept of evidence attribution has gained increased attention in recent years, following the advent of generative models.

Our method integrates fact verification and evidence attribution. In this section, we discuss recent advancements in both domains.

2.1 Fact Verification

Fact verification is a task that is closely related to natural language inference (NLI) (Conneau et al., 2017; Schick and Schütze, 2020), in which given a premise, the task is to verify whether a hypothesis is an entailment, contradiction, or neutral. Similarly, in fact verification, the task is to check if a given text can be supported, refuted, or indeterminable, given a reference text. Recent studies in this domain show that LLMs can achieve high performance, and can be considerably reliable for verification tasks, even though they are prone to hallucations (Guan et al., 2023).

In Lee et al. (2020), the authors show that the inherent knowledge of LLMs could be used to perform fact verification. Other works (Yao et al., 2022; Jiang et al., 2023b) have shown that using external knowledge is helpful for many reasoning-intensive tasks, and report enhanced performance on HotPotQA (Yang et al., 2018) and FEVER (Thorne et al., 2018). A wide variety of studies have established that LLMs are suitable for fact verification. For example, (Dong and Smith, 2021) enhanced accuracy of table-based fact verification by incorporating column-level cell rank information into pre-training. In FactScore, authors (Min et al., 2023), introduce a new evaluation that breaks a long-form text generated by large language models (LMs) into individual atomic facts and calculates the proportion of these atomic facts that are substantiated by a credible knowledge base.

2.2 Evidence Attribution

The distinction between evidence attribution and fact verification lies in the emphasis on identifying a source that can be attributed to the information. This task is becoming increasingly important, as generative models produce useful and impressive outputs, but without a frame of reference to validate them. In (Rashkin et al., 2023), the authors present

a framework named *AIS* (Attributable to Identified Sources) that specifies annotation guidelines and underlines the importance of attributing text to an external, verifiable, and independent source. (Yue et al., 2023) demonstrate that LLMs can be utilized for automatic evaluation of attribution, operationalizing the guidelines presented in (Rashkin et al., 2023). However, both of these works are primarily designed for the question-answering (QA) task. In contrast, our method is not restricted to QA and is designed to work with text in general. Furthermore, while these previous studies focus on sentence or paragraph levels, our approach extends to a more detailed and granular level of analysis.

3 Methodology

In this section, we present the methodology for retrieving relevant triplets from the KG, fine-tuning LLM to process text at claim-level, verifying claims, tagging evidence for each prediction, and generating a rationale along with an attribution score that reflects the text's validity.

3.1 Preprocessing

Preprocessing involves multiple steps required to make the input text suitable for the subsequent operations. Since the nodes in a KG typically represent entities, performing Named Entity Recognition (NER) is necessary. In our work, we chose Wikidata (Vrandečić and Krötzsch, 2014) as the KG source; thus, we use an NER module suitable for Wiki entities (Gerber, 2023). However, the framework is sufficiently generic to support any kind of KG that models information in the form of triplets. As our analysis is performed at the claim level, coreference resolution (Lee et al., 2017) becomes a necessary step to form localized claims that are semantically self-contained. If input text exceeds the context length, which depends on design choices, compartmentalization would be required. As a final step in preprocessing, we perform KG entity linking. This step tags all entities in the text that are present in the KG as nodes.

3.2 Relevant Triplets Retrieval

Retrieving relevant triplets is a complex problem that has attracted attention from various research communities, and resulted in multiple approaches to address the challenge. While retrieving direct links between two given nodes in a KG is relatively straightforward, identifying complex paths that involve multiple hops is challenging. In our framework, we use Woolnet (Gutiérrez and Patricio, 2023), a multi-node Breadth-First Search (BFS) algorithm, to retrieve the most relevant triplets for a given claim present in the KG. This BFS algorithm initiates from multiple starting points and, at each step, searches for and processes all adjacent neighbors before advancing. It constructs a subgraph of visited nodes, tracking their origins, and distances from each BFS's start. The algorithm expands each search tree one node at a time until paths intersect or reach a predefined maximum length. Upon intersection, it assesses if the discovered path meets the length criteria. If so, it logs the route, utilizing backtracking to trace the path to its origins, while ensuring there are no repetitions or cycles, thus maintaining a connection to a starting node. In our experiments, we allow for a maximum of three hops between any two given nodes, and a maximum of four potential paths. Adopting less stringent conditions leads to less relevant triplets.

3.3 Objective Function

Previous works on evidence attribution tasks have established definitions for the categorization of input text with reference to a supporting source (Rashkin et al., 2023; Gao et al., 2023; Bohnet et al., 2022; Yue et al., 2023). Similar to the formulation in (Yue et al., 2023), we use three categories: *Attributable*, *Extrapolatory*, and *Contradictory*. However, there are two main differences that distinguish our approach from previous methods. First, we verify the input text against facts present in a KG, an aggregated information source constructed by integrating numerous data sources into a structure of triplets, instead of relying on a single reference. This approach eliminates the one-to-one dependency between the text and its information source. Second, we perform attribution with finer granularity, specifically at the claim level, involving a subtask of decomposing the input text into individual claims. We define our categories as follows:

- **Attributable**: Triplets fully support the claim.
- **Extrapolatory**: Triplets lack sufficient information to evaluate the claim.
- **Contradictory**: Triplets contradict the claim.

We formulate the objective function of our task

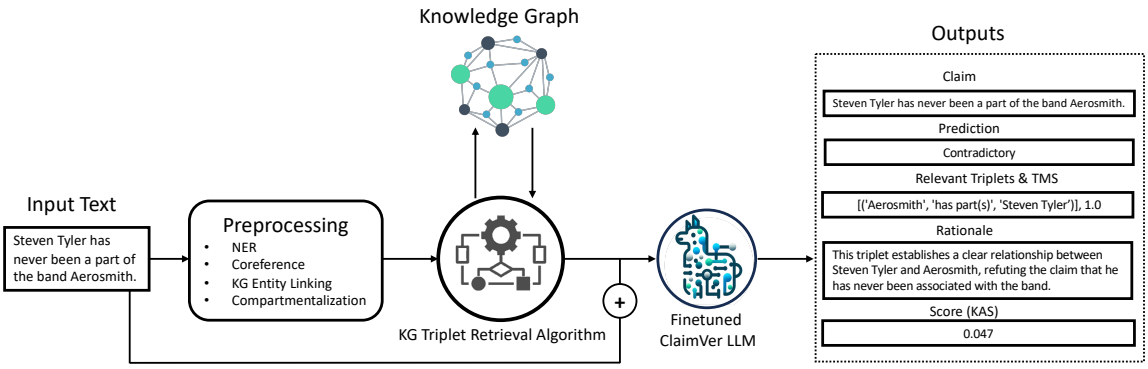

Figure 2: Flow of operations in the ClaimVer framework. Identified KG entity nodes during preprocessing inform the extraction of relevant triplets by the KG algorithm. Subsequently, these triplets and preprocessed text are then fed to a ClaimVer LLM, fine-tuned to operationalize the objective function. For each claim, the corresponding text span, prediction, relevant triplets, attribution scores, and rationale are generated.

as follows:

$$f(input\_text, ret\_triplets) = \{(claim\_span_i, claim\_pred_i, rel\_triplets_i, rationale_i)\}_{i=1}^{n} \quad (1)$$

where:

- $input\_text$: input text containing claim(s).
- $ret\_triplets$: *retrieved* triplets for the input text.
- $claim\_span_i$: i^{th} claim extracted as a substring from $input\_text$.
- $claim\_pred_i$: label predicted for $claim\_span_i$.
- $rel\_triplets_i$: *relevant* subset of $ret\_triplets$ for $claim\_span_i$.
- $rationale_i$: justification for $claim\_pred_i$.
- n: total number of claims in $input\_text$.

This objective function encompasses two main sub-tasks:

1. Decomposing input text into claims.
2. Generating prediction and corresponding rationale for each claim by identifying relevant supporting triplets.

3.4 Fine-tuning LLMs

The objective function shares similarities with the well-studied task of NLI (Conneau et al., 2017; Schick and Schütze, 2020). LLMs achieve state-of-the-art performance for NLI (Chowdhery et al., 2023), making them a suitable choice to operationalize the objective function. Additionally, (Yue et al., 2023) shows that LLMs can be used to automatically evaluate attribution to a given information source. However, these prior methods do not involve a complex sub-task, which is central to the proposed objective function, i.e., decomposing the input text into text spans that correspond to separate claims in the presence of multiple claims.

It is crucial to perform both claim decomposition and attribution for all claims in a single step, as processing each claim individually can lead to an exponential increase in LLM queries, leading to significantly higher computational costs and latency issues.

In order to perform attribution at the claim level, we need to fine-tune LLMs specifically for the proposed objective function (see §3.3) using a custom dataset. This is necessary because, as of this writing, even the state-of-the-art model, OpenAI's GPT-4 (Achiam et al., 2023), does not perform satisfactorily right out of the box. Our custom dataset, built using two sequential complex prompts with GPT-4, enables us to fine-tune significantly smaller models. This approach distills the performance of a large proprietary model using a multi-query prompt pipeline into small open-source models with a compact zero-shot prompt. We make the weights of the fine-tuned models publicly available[3].

We selected eight open-source LLMs with diverse sizes, ranging from 2B parameters to 10B parameters, to perform the fine-tuning: Gemma-2B-IT-Chat (Team et al., 2024), Phi-3-mini-4k-Chat (Javaheripi et al., 2023), Zephyr-7B-Beta-Chat (Tunstall et al., 2023), Mistral-7B-v0.3-Chat (Jiang et al., 2023a), Llama3-8B-Chat (Touvron et al., 2023), and Solar-10.7B-Chat (Kim et al., 2023). The models were fine-tuned using LoRA (Hu et al., 2021) with 4-bit quantization and adapters with rank 8 (Dettmers et al., 2024). The context length was set to 4096 tokens (for additional training de-

[3] anonymized link for peer-review

```
Analyze text against provided triplets, classifying claims as
"Attributable", "Contradictory", or "Extrapolatory".
Justify your classification using the following structure:
- "text_span": Text under evaluation.
- "prediction": Category of the text (Attributable /
Contradictory / Extrapolatory).
- "triplets": Relevant triplets (if any, else "NA").
- "rationale": Reason for classification.
For multiple claims, number each component (e.g., "text_span1",
"prediction1",..). Use "NA" for inapplicable keys.
Example:
"text_span1": "Specific claim",
"prediction1": "Attributable/Contradictory/Extrapolatory",
"triplets1": "Relevant triplets",
"rationale1": "Prediction justification",
...
Input for analysis:
-Text: {Input Text}
-Triplets: {Retrieved Triplets}
```

Figure 3: Instruction prompt for fine-tuned LLMs.

tails, refer §A.1) All models converged after 2 epochs, and high ROUGE-L (Lin, 2004) scores greater than 0.658 were achieved for each model. The instruction prompt used for fine-tuning is presented in Figure 3.

3.5 Computing Attribution Scores

For various downstream tasks, such as ranking and filtering, a continuous score that reflects the validity of a given piece of text with respect to a KG is desirable. We propose the KG Attribution Score (KAS), which accomplishes this task with a high level of granularity, and is detailed in this section.

3.5.1 Claim Scores

$$\text{cs}(y_i) = \begin{cases} 2 & \text{if } y_i = \text{Attributable} \\ 1 & \text{if } y_i = \text{Extrapolatory and } |\textit{rel\_triplets}_i| > 0 \\ 0 & \text{if } y_i = \text{Extrapolatory and } |\textit{rel\_triplets}_i| = 0 \\ 0 & \text{if } y_i = \text{No attribution} \\ -1 & \text{if } y_i = \text{Contradictory} \end{cases} \tag{2}$$

where, y_i is $claim\_pred_i$.

For each claim, we assign a score that reflects the level of its validity, ranging from -1 (*contradictory*) to 2 (*attributable*). If a claim is predicted to be *extrapolatory*, yet has one or more relevant triplets, we assign that claim a score of 1, as there is still relevant information available even though it may not be sufficient to completely support or refute the claim. However, if there are no triplets at all, along with an *extrapolatory* prediction, we assign 0 as it does not add any useful information. While decomposing claims, the model might occasionally omit words, typically stop-words, and we assign 0 in those cases as well.

3.5.2 Triplets Match Score (TMS)

This score reflects the extent of the match between the relevant triplets and the corresponding claim, and it can also serve as a proxy for the prediction confidence. Even though the prediction is made at the claim level, the triplets match score considers word-level matches in the computation. It can be computed as follows:

$$\begin{aligned} TMS(E(claim\_span_i), E(rel\_triplet_i)) = \\ \alpha \cdot SS(E(claim\_span_i), E(rel\_triplet_i)) \\ + \beta \cdot EPR(E(claim\_span_i), E(rel\_triplet_i)) \end{aligned} \tag{3}$$

where, $E(claim\_span_i)$ and $E(rel\_triplet_i)$ represent the sets of entities in $claim\_span_i$ and $rel\_triplet_i$, respectively. SS is the semantic similarity computed using the cosine similarity of text embeddings, and EPR represents the ratio of entities in $E(claim\_span_i)$ that are also present in $E(rel\_triplet_i)$. The parameters α and β can be adjusted as needed; in our experiments, we use 0.5 for both. In cases where examples of an entity retrieved from the KG are used to support the prediction, instead of the entity itself, we may not have a direct overlap, and thus semantic similarity would be helpful. EPR rewards the direct use of the entity, so a balance between both may be ideal in most cases.

3.5.3 KG Attribution Score (KAS)

For the final KG Attribution Score (KAS), a continuous score between 0 and 1 is desirable, as this facilitates various downstream applications such as ranking, fine-tuning, and filtering. This can be achieved using a Sigmoid function. However, the standard Sigmoid function treats positive and negative scores equally. In most cases, higher penalties should be assigned for erroneous text than rewards for valid text. This requirement can be met using a modified Sigmoid function that penalizes mistakes by a factor of γ:

$$\begin{aligned} \sigma_{\text{mod}}(x, \gamma) = \frac{1}{1 + e^{-\gamma \cdot x}}, \\ \text{where } \gamma = \begin{cases} \gamma = 3 & \text{if } x < 0, \\ \gamma = 1 & \text{if } x \geq 0, \end{cases} \end{aligned} \tag{4}$$

In our experiments, we set the value of γ to 3. Finally, the modified Sigmoid function, applied to the summation of triplet match scores and claim scores, is used to generate KAS:

$$\text{KAS} = \sigma_{mod}(\sum_{i=1}^{n}[TMS_i \cdot cs(y_i)], \gamma) \tag{5}$$

Split	Samples	Claims	Claim Labels		
			Att	Ext	Con
Train	3400	5342	998	3546	798
Test	1000	1677	316	1068	293

Table 1: Distribution of fine-tuning dataset. Att: Attributable, Ext: Extrapolatory, Con: Contradictory.

4 Dataset

Open-domain Question Answering (QA) datasets, such as WikiQA (Yang et al., 2015), HotPotQA (Yang et al., 2018), PopQA (Mallen et al., 2022), and EntityQuestions (Sciavolino et al., 2021), as well as Fact Verification datasets like FEVER (Thorne et al., 2019), FEVEROUS (Aly et al., 2021), TabFacT (Chen et al., 2019), and SEM-TAB-FACTS (Wang et al., 2021a), provide texts along with corresponding reference contexts or attributable information sources. However, these datasets significantly differ from the type of data required to train and test our proposed objective function, primarily due to two major factors: (i) these datasets predominantly offer samples that are inherently *attributable*, and (ii) consist of atomic claims and/or one-to-one mappings between input and reference texts. To address the first limitation, prior work (Yue et al., 2023) in attribution evaluation introduced new samples by modifying correct answers to generate *contradictory* instances. Yet, this adjustment alone is not sufficient for our use case because our method requires attribution at the claim level, and necessitates the automatic decomposition input text to claims. Consequently, as this task represents a novel challenge, we developed a new dataset that enables effective training and testing of the objective function.

Considering the choice of our KG, which is Wikidata (Vrandečić and Krötzsch, 2014), we opted for WikiQA (Yang et al., 2015) as it is closely associated with the Wiki ecosystem. Given that our method is designed for text validation in general, and is not limited to question answering, we retain only answers and discard the questions. Subsequently, we processed the answers following the steps detailed in Section 3.1, selecting entries containing two or more Wiki entities. This approach resulted in the exclusion of most single-word answers and other responses that are dependent on their corresponding questions and may lack comprehensibility without them.

We utilize GPT-4 (Achiam et al., 2023) to generate the initial version of the ground truth. Although GPT-4 can adhere to the instructions (refer to Figure 3) to a reasonable degree and responds in the required format with all necessary keys, it still underperforms in the overall task. The most frequent issue observed is the erroneous assignment of prediction labels. To remedy this issue, we designed a detailed prompt tailored for the given task, incorporating techniques such as few-shot, chain-of-thought (Kojima et al., 2022), and other strategies (OpenAI, 2024; Nori et al., 2023) (full prompt in §A Figure 12). We also conducted manual checks to ensure only high-quality samples were retained, as research indicates that high alignment can be achieved with as few as 1,000 samples, provided they are of superior quality (Zhou et al., 2023).

The final dataset is comprised of two splits: the training split, based on the training split of WikiQA (Yang et al., 2015), and a test split, derived from both the test and validation splits. The training split contains 3,400 samples, and since some entries feature multiple claims, there are a total of 5,342 claims within this split. Similarly, the test split includes 1,000 samples and 1,677 claims. The label counts for the claims are tabulated in Table 1. The dataset is publicly shared to facilitate further research in this direction[4].

5 Experiments and Results

In this section, we present the evaluation of our claim-level attribution method. The performance metrics of the fine-tuned LLMs, which operationalize the objective function, are presented in Tables 3 and 4. In Table 3, we observe that all models converge and achieve sufficiently high ROUGE-L and ROUGE-1 scores, with *Mistral-7B-v0.3-Chat* achieving the highest of 0.694 and 0.719 respectively. We also observe that the smaller model, *Gemma-2B-IT-Chat* with just 2B parameters, is also sufficiently compatible for this task as it attained a decent ROUGE-L score of 0.667.

The first task of the proposed objective function (refer §3.3), decomposing text into multiple claims, is somewhat subjective, and there could be multiple valid approaches due to linguistic complexities. For instance, example 4 in Table 2 has been decomposed into three claims, but the first could arguably be further decomposed to verify whether Blagnac is in France, and whether it is a suburb of Toulouse. Controlling the precise manner of decomposition is challenging, and might necessitate an additional

[4] anonymized link for peer-review

	Input Text	Relevant Triplets	Prediction (TMS)	Rationale	KAS
1	[1]George O'Malley is a fictional character from the medical drama television series Grey's Anatomy, [2]which airs on the American Broadcasting Company (ABC) in the United States.	1: [("Grey's Anatomy", 'characters', "George O'Malley")] 2: [("Grey's Anatomy", 'original broadcaster', 'American Broadcasting Company'), ('American Broadcasting Company', 'country', 'United States of America')]	1: Attributable (0.852) 2: Attributable (0.637)	1: The triplet directly supports the claim that George O'Malley is a character in Grey's Anatomy. 2: The triplets confirm that Grey's Anatomy airs on ABC, which is based in the United States, directly supporting the claim about the show's broadcasting and location.	0.818
2	[1]Bane was portrayed as a tertiary villain by Robert Swenson in Batman & Robin , [2]directed by Joel Schumacher, [3]and Batman's Back	1: [('Batman & Robin', 'cast member', 'Robert Swenson')] 2: [('Batman & Robin', 'director', 'Joel Schumacher')] 3: NA	1: Attributable (0.788) 2: Attributable (0.882) 3: Extrapolatory (0.0)	1: The triplet directly supports the claim that Robert Swenson was involved in Batman & Robin, which is a requirement for the statement about his portrayal of Bane. 2: The triplet directly supports the claim about the director of Batman & Robin, which is relevant to the context of the film. 3: There are no triplets that directly support or refute the claim about Batman's back	0.752
3	[1]Crater Lake is the main feature of Crater Lake National Park [2]and famous for its deep blue color and water clarity.	1: [('Crater Lake', 'located in protected area', 'Crater Lake National Park')] 2: NA	1: Attributable (0.942) 2: Extrapolatory (0.0)	1: The triplet directly supports the claim that Crater Lake is a significant feature within Crater Lake National Park, as it is located within the protected area. 2: There are no triplets provided that directly support or refute the claim about the deep blue color and water clarity of Crater Lake.	0.719
4	[1]Based in Blagnac , France, a suburb of Toulouse, [2]and with significant activity across Europe, [3]airbus produces approximately half of the world's jet airliners .	1: [('Airbus Operations S.A.S.', 'country', 'France'), ('Airbus Corporate Jets', 'headquarters location', 'Toulouse'), ('Blagnac', 'country', 'France')] 2: NA 3: NA	1: Attributable (0.505) 2: Extrapolatory (0.0) 3: Extrapolatory (0.0)	1: The triplets confirm that Airbus Operations S.A.S. is in France, Airbus Corporate Jets is headquartered in Toulouse, and Blagnac is a suburb of Toulouse in France, supporting the statement about Airbus's location in France and its proximity to Toulouse. 2: The triplets do not provide information about Airbus's activity across Europe 3: The triplets do not provide any information about Airbus's production output or market share	0.583
5	[1]Pope Benedict XVI never appointed anyone significant within the Catholic Church, [2]nor did he ever teach the importance of understanding God's redemptive love.	1: [('Rutilio del Riego Jáñez', 'appointed by', 'Benedict XVI'), ('Rutilio del Riego Jáñez','religion or worldview', 'Catholic Church')] 2: [('God','said to be the same as', 'love')]	1: Contradictory (0.781) 2: Extrapolatory (0.065)	1: The triplets directly contradict the claim by showing that Pope Benedict XVI did indeed appoint someone (Rutilio del Riego Jáñez) who is associated with the Catholic Church, indicating that he did appoint significant individuals within the Church. 2: While the triplets indicate that God is equated with love, it does not directly address whether Pope Benedict XVI taught the importance of understanding God's redemptive love.	0.248
6	[1]Southwest Airlines has never operated any Boeing 737 models.	1: [('Boeing 737 MAX', 'operator', 'Southwest Airlines'), ('Boeing 737 #1491', 'operator', 'Southwest Airlines')]	1: Contradictory (0.933)	1: The triplets directly contradict the claim by indicating that Southwest Airlines has operated both the Boeing 737 MAX and Boeing 737 #1491, which are specific models of the Boeing 737. This refutes the statement that Southwest Airlines has never operated any Boeing 737 models.	0.057

Table 2: Examples of claim-level attribution by the proposed method. The first column shows the numbered claims in the input text. Second column lists relevant triplets for each claim. Predictions and *Triplets Match Score (TMS)* are in the third column, while the rationale behind each prediction is in the fourth column. The *Knowledge Graph Attribution Score (KAS)* is shown in the last column. Model: *Solar-10.7B-Chat*.

step before the prediction step, involving separate processing for each claim. However, this option could prove to be impractical, as the number of LLM queries could increase exponentially.

To accurately compute classification performance, we impose a strict strategy: the text span of the claim, the identified relevant triplets, and the prediction label must all exactly match the ground truth to be considered accurate. In Table 4, the second column indicates number of claims with text spans exactly matching the ground truth responses. Columns 3 to 6 present the accuracy, precision, recall, and F1 scores for these matching claims. The most performant model is *Solar-10.7B-Chat*, with 1031 exact matches out of 1677 claims in the test set. Additionally, the classification scores in all metrics are above 89%, which clearly demonstrates that the model can reliably differentiate between the classes *attributable*, *extrapolatory*, and *contradictory*.

Table 2 showcases the claim-level attribution performed by our method. Each claim in the input text is numbered and color-coded to reflect its prediction: green for attributable, amber for extrapolatory, and red for contradictory. The examples are sorted in descending order by their KAS scores, which reflect the validity of the text. As expected, we observe more green at the top of the table and more amber and eventually red as we move down. Since the Wiki ecosystem is open-domain, we observe that the examples cover a wide range of topics, demonstrating that the method is adaptable to diverse inputs.

In the first example in 2, the input text is decomposed into two claims, both of which are attributable. The first claim is supported by a single triplet in the KG, while the second claim can be supported by combining two triplets. The second example presents more challenges for evaluation due to its complex sentence structure, but ClaimVer accurately identifies that the third claim regarding Batman's Back is neither supported nor refuted by the triplets, as indicated in the rationale. In the third example, we note that the first claim is predicted to be attributable with a high triplet match score of 0.942 since there is a triplet that clearly supports the location description of Crater Lake. However, as there is no information regarding the

Model	Size	ROUGE-L	ROUGE-1
Gemma-2B-IT-Chat	2B	0.667	0.692
Phi-3-mini-4k-Chat	4B	0.658	0.685
Zephyr-7B-Beta-Chat	7B	0.686	0.712
Vicuna-7B-v1.5-Chat	7B	0.676	0.700
Mistral-7B-v0.3-Chat	7B	**0.694**	**0.719**
Gemma-7B-IT-Chat	7B	0.678	0.703
Llama3-8B-Chat	8B	0.679	0.705
Solar-10.7B-Chat	10B	0.689	0.714

Table 3: ROUGE scores on the test set ($n = 1,000$).

Model	#MC	Acc	Prec	Rec	F1
Gemma-2B-IT-Chat	895	77.09	77.20	77.09	74.24
Phi-3-mini-4k-Chat	882	72.22	78.10	72.22	72.86
Zephyr-7B-Beta-Chat	978	85.89	87.41	85.89	86.16
Vicuna-7B-v1.5-Chat	898	79.62	78.83	79.62	78.84
Mistral-7B-v0.3-Chat	1002	86.63	87.03	86.63	86.73
Gemma-7B-IT-Chat	940	82.87	84.09	82.87	83.17
Llama3-8B-Chat	959	80.92	85.48	80.92	81.36
Solar-10.7B-Chat	**1031**	**89.23**	**89.52**	**89.23**	**89.30**

Table 4: Scores on matching claims in the test set ($n = 1677$). #MC: number of matching claims.

water characteristics, the second claim is categorized as extrapolatory. In the fourth example, the first claim alone requires three triplets combined as supporting evidence, illustrating the method's ability to handle complex multi-hop paths within the KG. The second and third claims are predicted to be extrapolatory, since there are no triplets concerning Airbus's market share, or its activities in Europe, as highlighted in the model's rationale. It is noteworthy that the context provided in the third claim is crucial for the first claim to be comprehensible, demonstrating why individual claim evaluation may be suboptimal. Interestingly, in the fifth example, the method identifies a specific instance from the KG to refute a general claim, citing the appointment of Rutilio del Riego Jáñez. Similarly, in the sixth example, the method provides specific instances, quoting two distinct Boeing 737 models to demonstrate contradiction with a high triplet match score.

6 Discussion

The susceptibility of LLMs to generating factually incorrect statements is an alarming concern as LLM-powered services become increasingly popular for seeking advice and information. The democratization of generative models has also had adverse effects, such as increasing misinformation (Monteith et al., 2024). To arm end-users with the tools necessary to combat being misinformed, it is crucial to develop text-validation methods that are human-centric, and prioritize user engagement, understanding, and informativeness. We design our method with these principles in mind: we make predictions at the claim level, and identify text spans within the given text, that can be color-coded and presented to the user. The proposed method also generates easily comprehensible explanations along with the prediction and evidence, thus reducing the cognitive burden on the end-user, and making the process user-friendly.

The usability and evaluation of these systems should align with human needs and capabilities. Chatbots, such as ChatGPT (Achiam et al., 2023), serve a wide array of tasks; therefore, the text validation method should be adaptable to various domains. While KGs like Wikidata (Vrandečić and Krötzsch, 2014) are considered open-domain, the implementation of more specialized KGs, along with corresponding routing algorithms may be necessary to support a broader range of topics. For instance, a common-sense KG (Hwang et al., 2020) would be more useful in validating non-factoid answers that involve logic.

Furthermore, the maintenance efficiency of our approach aligns well with the need for sustainable, long-term AI solutions. In a world where information is constantly evolving, the ability to update and maintain AI systems with minimal effort is not just a convenience, but a necessity. This directly ties into the ethical implications of AI, where outdated or incorrect information can lead to harmful decisions. By leveraging existing, well-maintained KGs, we can ensure that AI systems remain accurate and relevant over time.

7 Conclusion

In this paper, we present ClaimVer, a framework for text verification and evidence attribution at the claim level by leveraging information present in KGs. In contrast to other methods, ClaimVer eliminates the one-to-one mapping between input and reference text, allowing for layered interpretation and handling of distributed information. In addition to these primary functions, ClaimVer incorporates human-centric design principles by offering clear, concise explanations for each claim prediction—an important characteristic for building user trust and enhancing usability. Furthermore, we introduce an attribution score, which enhances its applicability across a wide range of downstream tasks. Finally, we share ClaimVer fine-tuned LLMs to facilitate further exploration of this research direction.

8 Limitations

Limitations of LLMs for Fact Verification. Like most ML models, LLMs are prone to erroneous predictions, which is particularly concerning in sensitive applications such as handling misinformation. Despite this, they remain the most performant techniques for fact verification and related tasks like NLI (Yue et al., 2023; Wang et al., 2021b). Therefore, while it is reasonable to use the best option available, fact verification systems relying on LLMs should be utilized with caution and necessary validations.

Limitations of Knowledge Graph. While there are several advantages associated with using KGs, we also acknowledge the presence of known issues, such as knowledge coverage and the efforts required to keep these sources up-to-date. For our solution, we assume that the KG is up-to-date and possesses adequate coverage. However, this may not always be the case, and thus the most suitable technique should be adopted after considering the specific requirements of a particular use case. Another point to consider is that the proposed method does not provide traditional citations to articles, although it may be possible to retrieve that information from the KG, if information source mapping has been properly maintained.

Variations in Claim Decomposition. Decomposing text into multiple claims is a complex linguistic task that often results in multiple valid decompositions. Although this may not impact usability if the prediction, rationale, and text spans are comprehensible and supported by facts from the KG, it poses a challenge for evaluating model performance. One potential approach is to operate at the token level instead of the span level, but this would significantly complicate the problem space. Additionally, token-level verification and attribution would require substantially higher compute resources, necessitating further studies to assess their value and impact on system usability and reliability.

LLM Reasoning Errors. Previous works have demonstrated that using LLM reasoning for tasks like fact verification, evidence attribution, and NLI can yield impressive results, surpassing alternative approaches (Yue et al., 2023; Wang et al., 2021b). However, LLM reasoning can still be flawed. In our work, we impose validations to minimize LLM mistakes by performing membership checks for supporting triplets and string matching for text spans. However, validating reasoning remains an open problem with ongoing research efforts.

Fine-tuning Dataset Limitations. To build the fine-tuning dataset, we utilized GPT-4 with two detailed sequential prompts designed in accordance with OpenAI's recommendations (OpenAI, 2024) and previous works (Nori et al., 2023). Despite employing techniques like few-shot prompting with state-of-the-art LLMs, we still observe mistakes, indicating the complexity of this problem. To address this, we conducted manual checks to minimize errors and share the dataset with the research community for further improvement.

9 Ethical Concerns

Our work presents a scalable and interpretable framework for fact-checking textual claims. To promote the exploration of this important problem space, we fine-tune and share open-source LLMs that are well-aligned to the framework's objective function. While the models we provide perform better than the publicly available base models for our specific task, they still share similar weaknesses. To address this as best as we can, we have incorporated and described ways to mitigate these issues to the extent possible. We believe the benefits of this work outweigh any potential risks.

References

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

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

Bernd Bohnet, Vinh Q Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037*.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul

Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Rui Dong and David A Smith. 2021. Structural encoding and pre-training matter: Adapting bert for table-based fact verification. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2366–2375.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. 2023. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

Emanuel Gerber. 2023. spacy module for linking text to wikidata items. `https://github.com/egerber/spaCy-entity-linker`. Accessed: 2024-02-26.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2023. Language models hallucinate, but may excel at fact verification. *arXiv preprint arXiv:2310.14564*.

Torres Gutiérrez and Cristóbal Patricio. 2023. Sistema visual para explorar subgrafos temáticos en wikidata.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI Conference on Artificial Intelligence*.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. 2023. Phi-2: The surprising power of small language models. *Microsoft Research Blog*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023a. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023b. Active retrieval augmented generation. *arXiv preprint arXiv:2305.06983*.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, et al. 2023. Solar 10.7 b: Scaling large language models with simple yet effective depth up-scaling. *arXiv preprint arXiv:2312.15166*.

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

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Nayeon Lee, Belinda Z Li, Sinong Wang, Wen-tau Yih, Hao Ma, and Madian Khabsa. 2020. Language models as fact checkers? *arXiv preprint arXiv:2006.04102*.

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. When not to trust language models: Investigating effectiveness and limitations of parametric and non-parametric memories. *arXiv preprint arXiv:2212.10511*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

Scott Monteith, Tasha Glenn, John R Geddes, Peter C Whybrow, Eric Achtyes, and Michael Bauer. 2024. Artificial intelligence and increasing misinformation. *The British Journal of Psychiatry*, 224(2):33–35.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. 2023. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*.

OpenAI. 2024. Best practices for prompt engineering with the openai api.

https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api. Accessed:2024-01-11.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, pages 1–64.

Amy Rechkemmer and Ming Yin. 2022. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *Proceedings of the 2022 chi conference on human factors in computing systems*, pages 1–14.

Ginny Russell, Stephan Collishaw, Jean Golding, Susan E Kelly, and Tamsin Ford. 2015. Changes in diagnosis rates and behavioural traits of autism spectrum disorder over time. *BJPsych open*, 1(2):110–115.

Timo Schick and Hinrich Schütze. 2020. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146:102551.

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

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2019. The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

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

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Nancy XR Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021a. Semeval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (sem-tab-facts). *arXiv preprint arXiv:2105.13995*.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021b. Entailment as few-shot learner. *arXiv preprint arXiv:2104.14690*.

Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. " do you trust me?" increasing user-trust by integrating virtual agents in explainable ai interaction design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 7–9.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Xiang Yue, Boshi Wang, Kai Zhang, Ziru Chen, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. *arXiv preprint arXiv:2305.06311*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

A Appendix

A.1 Training Details

In this section, we present the training parameters used for fine-tuning each model, along with their corresponding loss plots. All models converged after two epochs, achieving ROUGE-L (Lin, 2004) scores greater than 0.658, with the highest model reaching 0.719.

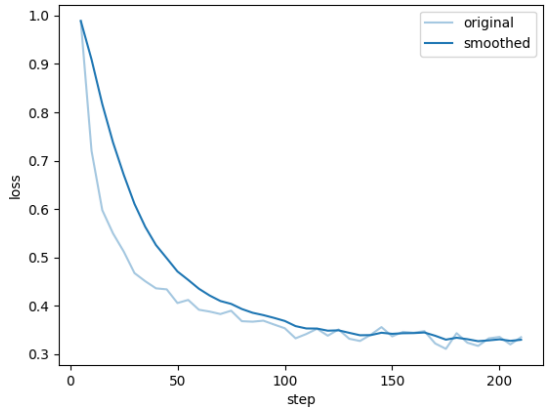

Figure 4: Fine-tuning loss plots for Llama3-8B-Chat.

Parameter	Value
Base Model	meta-llama/Meta-Llama-3-8B-Instruct
ROUGE-L	TBD
ROUGE-1	TBD
Fine-Tuning Type	LoRA
LoRA Alpha	16
LoRA Rank	8
Cutoff Length	4096
Gradient Accumulation Steps	8
Learning Rate	5.0e-05
LR Scheduler Type	Cosine
Number of Training Epochs	2.0
Optimizer	AdamW
Quantization Bit	4

Table 5: Fine-tuning Parameters for Llama3-8B-Chat

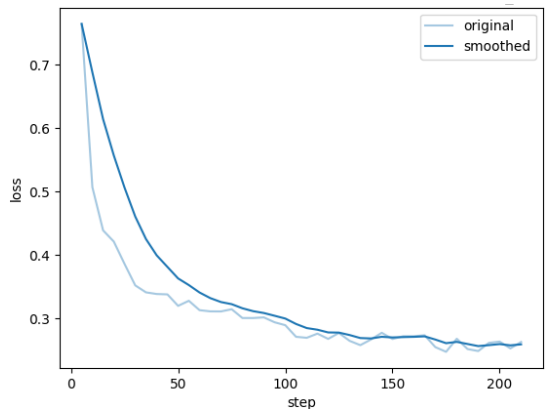

Figure 5: Fine-tuning loss plots for Mistral-7B-v0.3-Chat.

Parameter	Value
Base Model	mistralai/Mistral-7B-Instruct-v0.3
ROUGE-L	0.694
ROUGE-1	0.719
Fine-Tuning Type	LoRA
LoRA Alpha	16
LoRA Rank	8
Cutoff Length	4096
Gradient Accumulation Steps	8
Learning Rate	5.0e-05
LR Scheduler Type	Cosine
Number of Training Epochs	2.0
Optimizer	AdamW
Quantization Bit	4

Table 6: Fine-tuning Parameters for Mistral-7B-v0.3-Chat

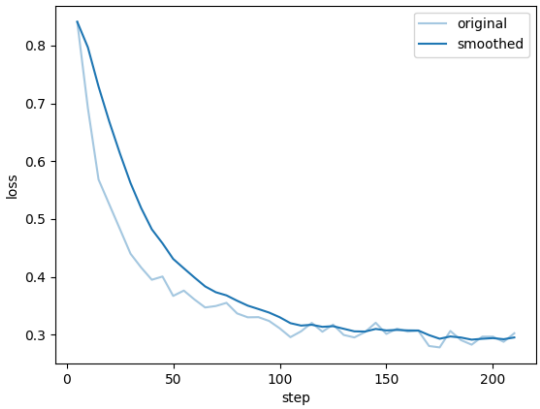

Figure 6: Fine-tuning loss plots for Phi-3-mini-4k-Chat.

Parameter	Value
Base Model	microsoft/Phi-3-mini-4k-instruct
ROUGE-L	0.658
ROUGE-1	0.685
Fine-Tuning Type	LoRA
LoRA Alpha	16
LoRA Rank	8
Cutoff Length	4096
Gradient Accumulation Steps	8
Learning Rate	5.0e-05
LR Scheduler Type	Cosine
Number of Training Epochs	2.0
Optimizer	AdamW
Quantization Bit	4

Table 7: Fine-tuning Parameters for Phi-3-mini-4k-Chat

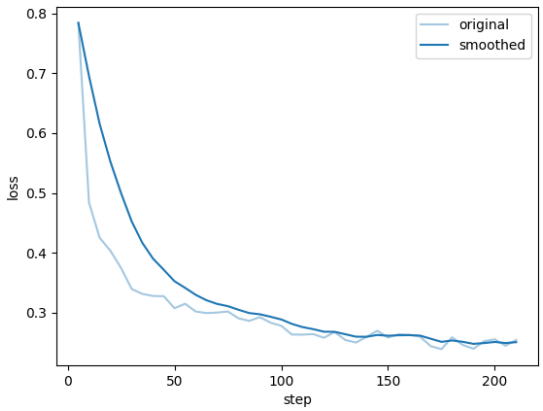

Figure 7: Fine-tuning loss plots for SOLAR-10.7B-Chat.

Parameter	Value
Base Model	upstage/SOLAR-10.7B-Instruct-v1.0
ROUGE-L	0.689
ROUGE-1	0.714
Fine-Tuning Type	LoRA
LoRA Alpha	16
LoRA Rank	8
Cutoff Length	4096
Gradient Accumulation Steps	8
Learning Rate	5.0e-05
LR Scheduler Type	Cosine
Number of Training Epochs	2.0
Optimizer	AdamW
Quantization Bit	4

Table 8: Fine-tuning Parameters for SOLAR-10.7B-Chat

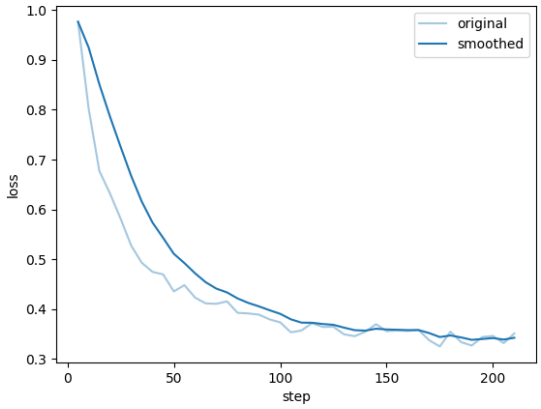

Figure 8: Fine-tuning loss plots for Vicuna-7B-v1.5-Chat.

Parameter	Value
Base Model	lmsys/vicuna-7b-v1.5
ROUGE-L	0.676
ROUGE-1	0.700
Fine-Tuning Type	LoRA
LoRA Alpha	16
LoRA Rank	8
Cutoff Length	4096
Gradient Accumulation Steps	8
Learning Rate	5.0e-05
LR Scheduler Type	Cosine
Number of Training Epochs	2.0
Optimizer	AdamW
Quantization Bit	4

Table 9: Fine-tuning Parameters for Vicuna-7B-v1.5-Chat

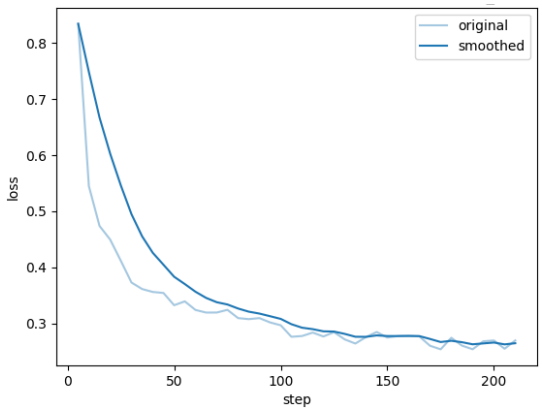

Figure 9: Fine-tuning loss plots for Zephyr-7B-Beta-Chat.

Parameter	Value
Base Model	HuggingFaceH4/zephyr-7b-beta
ROUGE-L	0.686
ROUGE-1	0.712
Fine-Tuning Type	LoRA
LoRA Alpha	16
LoRA Rank	8
Cutoff Length	4096
Gradient Accumulation Steps	8
Learning Rate	5.0e-05
LR Scheduler Type	Cosine
Number of Training Epochs	2.0
Optimizer	AdamW
Quantization Bit	4

Table 10: Fine-tuning Parameters for Zephyr-7B-Beta-Chat

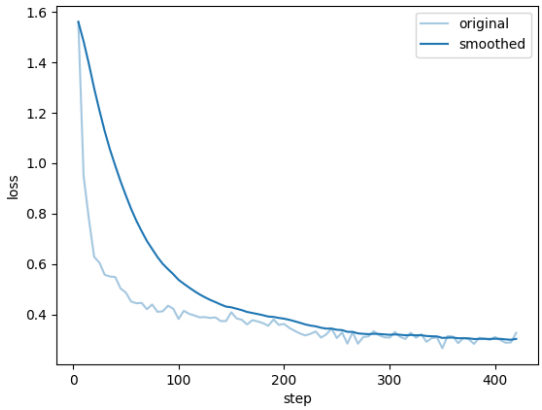

Figure 10: Fine-tuning loss plots for Gemma-7B-IT-Chat.

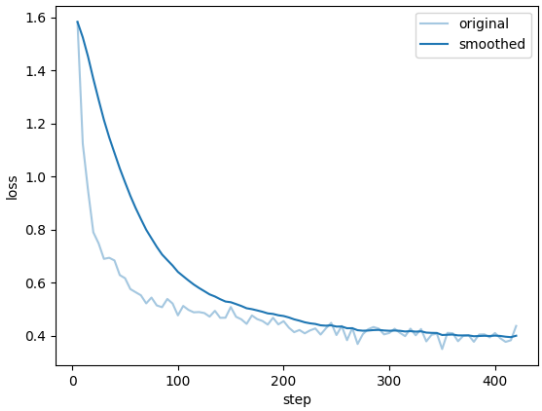

Figure 11: Fine-tuning loss plots for Gemma-2B-IT-Chat.

Parameter	Value
Base Model	google/gemma-7b-it
ROUGE-L	0.678
ROUGE-1	0.703
Fine-Tuning Type	LoRA
LoRA Alpha	16
LoRA Rank	8
Cutoff Length	4096
Gradient Accumulation Steps	8
Learning Rate	5.0e-05
LR Scheduler Type	Cosine
Number of Training Epochs	2.0
Optimizer	AdamW
Quantization Bit	4

Table 11: Fine-tuning Parameters for Gemma-7B-IT-Chat

Parameter	Value
Base Model	google/gemma-2b-it
ROUGE-L	0.667
ROUGE-1	0.692
Fine-Tuning Type	LoRA
LoRA Alpha	16
LoRA Rank	8
Cutoff Length	4096
Gradient Accumulation Steps	8
Learning Rate	5.0e-05
LR Scheduler Type	Cosine
Number of Training Epochs	2.0
Optimizer	AdamW
Quantization Bit	4

Table 12: Fine-tuning Parameters for Gemma-2B-IT-Chat

```
**Text Span Attribution Verification**

**Objective:** Predict whether the text span is "Attributable", "Contradictory", or "Extrapolatory" based on the information provided in the triplets.

**Instructions:**

1. **Read the Full Text:**
- Understand the context and content of the full text string.

2. **Examine the Text Span:**
- Determine the claims made within the text span.

3. **Analyze the Triplets:**
- Evaluate if the triplets support, refute, or neither support nor refute the claims in the text span.

4. **Make Your Prediction:**
- Classify the text span as "Attributable", "Contradictory", or "Extrapolatory" based on your analysis of the triplets.

5. **Provide Rationale:**
- Clearly explain your reasoning for the classification.

**Classification Criteria:**

- **"Attributable"**: The text span is sufficiently supported by the triplet(s). All claims in the text span are directly present in the triplet information.
- **"Contradictory"**: The text span is conclusively refuted by the triplet(s). All claims in the text span are directly contradicted by the triplet information.
- **"Extrapolatory"**: The triplet(s) can neither support nor refute the text span. The information provided is either irrelevant, indirect, or related but not sufficient to support or refute the text span.

**Example:**

**Full Text:** "Albert Einstein is widely recognized as the father of modern physics. He was awarded the Nobel Prize in Physics for his services to Theoretical Physics."

**Text Span:** "He was awarded the Nobel Prize in Physics."

**Triplets:** [("Albert Einstein", "award received", "Nobel Prize in Physics")]

**Sample Evaluation:**
- **Prediction:** "Attributable"
- **Rationale:** "The triplet directly supports the claim that Albert Einstein received the Nobel Prize in Physics."

**Example:**

**Full Text:** "Isaac Newton discovered the element radium."

**Text Span:** "Isaac Newton discovered radium."

**Triplets:** [("Marie Curie", "discovered", "radium")]

**Sample Evaluation:**
- **Prediction:** "Contradictory"
- **Rationale:** "The triplet states that Marie Curie discovered radium, contradicting the claim that Isaac Newton discovered it."

**Example:**

**Full Text:** "The Eiffel Tower is a wrought-iron lattice tower that was opened in 1889."

**Text Span:** "The Eiffel Tower is a wrought-iron lattice tower that was opened in 1889."

**Triplets:** [("Eiffel Tower", "located in", "Paris")]

**Sample Evaluation:**
- **Prediction:** "Extrapolatory"
- **Rationale:** "The triplet states that the Eiffel Tower is located in Paris, which is related but not sufficient to confirm or refute that it was opened in 1889."

**Verification Checklist:**

- [ ] The prediction accurately reflects the relationship between the text span and the triplets.
- [ ] The rationale clearly explains the classification based on the triplets.
- [ ] The explanation is free from irrelevant information.

**Response Format:**
Provide your evaluation in the following JSON format:
- "prediction": "Attributable", "Contradictory", or "Extrapolatory"
- "rationale": "Your comments here"

**Inputs to Evaluate**

**Full text:** "{full_text}"
**Text span:** "{text_span}"
**Triplets:** {triplets}
```

Figure 12: Prompt for the text span attribution verification task, guiding the model to classify text spans as "Attributable," "Contradictory," or "Extrapolatory" based on the provided triplets. The prompt design incorporates concepts such as few-shot learning, chain-of-thought reasoning (Kojima et al., 2022), and tailored prompt engineering (OpenAI, 2024; Nori et al., 2023)