UNH at Check That! 2025 Task 2: Fine-tuning Vs Prompting

Joe Wilder[1], Nikhil Kadapala[1], Benji Xu[1], Mohammed Alsaadi[1], Mitchell Rogers[1], Palash Agrawal[1], Adam Hassick[1] and Laura Dietz[1]

[1]University of New Hampshire, Durham, NH, 03824, USA

Abstract

We participate in CheckThat! Task 2 and explore various methods of prompting and in-context learning, including few-shot prompting and fine-tuning with different LLM families, with the goal of extracting check-worthy claims from social media passages. Our best METEOR score is achieved by fine-tuning a FLAN-T5 model. However, we observe that higher-quality claims can sometimes be extracted using other methods, even when their METEOR scores are lower.

1. Introduction

The CheckThat! Lab aims to build computational infrastructure to support human fact checkers, consisting of a pipeline for verifying the truthfulness of social media text.

Task 2 [1] addresses a specific challenge in this pipeline:

Task Statement   Given the text of a social media post, extract the main claim in succinct and concise language suitable for a human fact checker to verify.

We foresee a number of unique challenges in this task and analyze them in our experiments.

Multiple equally relevant claims.   Often, multiple extracted claims are topically relevant. However, the task requires focusing on a single fact for manual verification. Analyzing the training and validation data, we found that many gold claims include multiple facts and could be rewritten more succinctly. For example: "Joe Biden lives in a large estate bought on a senator's salary."

Missing multi-modal content.   We found that many gold claims refer to information not available to participants, such as photos or videos included with the original post but removed during pre-processing. While we accept that claims cannot be based on unseen data, we instruct our models to make educated guesses about the topic.

Faithful extractions.   We are especially concerned about extracted claims where LLM hallucinations introduce content not present in the original post. An example of a hallucinated and overly verbose claim follows:

Social Media Post:

```
The salary of a U.S. Senator is $174,000 per year. This is Joe Biden's house...
seems legit :) The salary of a U.S. Senator is $174,000 per year. This is Joe
Biden's house...  seems legit :) The salary of a U.S. Senator is $174,000 per
year. This is Joe Biden's house...  seems legit :)
```

Hallucinated Extracted Claim

```
Joe Biden's house, purchased for an amount significantly exceeding the cumulative
value of his annual U.S. Senator's salary of $174,000, raises questions about
potential additional, undisclosed sources of income that may have contributed
to the down payment, mortgage payments, property taxes, insurance premiums, and
ongoing maintenance costs associated with the property.
```

Our emphasis is on exploring the design space across different LLMs and methods—fine-tuning and few-shot prompting—in search of the best trade-off between optimizing the METEOR score and producing claims that are genuinely useful for human fact-checkers.

We conduct a broad exploration of methods, prompts, and LLMs, casting a wide net. Our approaches fall into three overarching categories:

1. Fine-tuning approaches,

2. Prompting approaches, and

3. "Frustratingly easy" baselines.

We describe all explored approaches and submit those performing best on the validation set in terms of METEOR.

We only use resources provided by the Task 2 organizers and publicly available large language models from Hugging Face and the Together.AI API service.

2. Approaches: Fine-Tuning

In this section, we describe methods relying on fine-tuning across LLMs of different parameter scales.

Our key takeaway: Flan-T5 Large offered the best compromise between raw capability and practical fine-tuning feasibility under our hardware and time constraints.

2.1. Finetuned Flan-T5 Large

This approach fine-tuned the Flan-T5 Large model on the CLEF 2025 Task 2 training dataset to align its outputs more closely with the gold-standard claims. Fine-tuning was performed using the `huggingface` transformers library without advanced techniques such as LoRA or PEFT.

Due to resource limitations, billion-parameter models were out of scope. We opted for Flan-T5 Large (783M parameters), which can run locally and is more manageable to train due to its smaller size. A straightforward task-specific prompt was prepended to training examples:

```
Please read the following social media post and extract the claim made within it.
Normalize the claim by rephrasing it in a clear and concise manner.
Post: $text
Extracted Claim:
```

Training ran for 10 epochs on an NVIDIA 4060 GPU, taking nearly four days to complete. This approach's strength lies in its ability to internalize extraction patterns not easily expressible via prompting alone. It achieved an average validation-set METEOR score of 0.5569.

2.2. LoRA fine-tuning of Flan-T5 Base

The motivation behind this submission was to balance performance with efficiency. Full fine-tuning of T5-Base gave strong results but incurred high computational costs. Prompt tuning, while efficient, yielded limited gains—especially on larger models. LoRA provided a middle ground by updating only 0.4% of parameters, keeping the rest of the model frozen.

LoRA allowed us to adapt the model effectively with minimal overhead. We chose Flan-T5 Base due to its strong baseline performance, aiming to retain quality while reducing resource demands.

This model achieved a validation METEOR score of 0.3054 and was our third-best test-set run, with a test METEOR of 0.28.

2.3. Fine-tuned DeepSeek-R1-Distill-Llama-8b Approach

Here, we fine-tuned the `DeepSeek-R1-Distill-Llama-8b` causal language model on the training dataset to assess its ability to produce reference-style summaries. The training used the following system prompt:

```
Extract the verifiable claim as one sentence from the user input.
```

This approach achieved a validation METEOR score of 0.2541.

Fine-tuning leads to marginal improvements over a baseline model with a single-shot prompt on the validation set, but zero-shot prompting showed no gain. We conclude that even 8B-parameter LLMs exceed our capacity for effective fine-tuning under current constraints. We compare our fine-tuning methods to prompting-based and in-context learning approaches. These include variations on few-shot prompting, self-refinement, self-scoring, and post-processing.

2.4. Claimifying Social Media Posts with Self-Refinement

This method uses a combination of prompting strategies and an iterative Self-Refinement stage. In this stage, the same LLM that generates the initial claim provides feedback based on specific criteria evaluating the check-worthiness of the claim against the input text. This feedback, along with the input and initial claim, is fed back into the same LLM, which generates a refined version of the claim.

We tested both zero-shot and few-shot prompts, with and without a Chain-of-Thought (CoT) trigger phrase. The CoT phrase acts as a cue for the model to reason step by step before producing a final claim. We evaluated these configurations with and without one or more iterations of the self-refinement stage.

For the Task 2 submission, we used the step-by-step "Claimify" process [2]. After extracting the initial claim, we applied one iteration of SELF-REFINE [3].

We also evaluated a variant using a few-shot prompt followed by the CoT trigger phrase `Let's think step by step` [4].

Models explored: GPT-4.1-nano, Gemini-2.0-Flash, LLaMA-3.3-70B, Grok3
Highest Avg METEOR score: 0.332 (Grok3 + Few-shot-CoT)
Prompt variations tested:

1. Zero-shot:

   ```
   Identify the decontextualized, stand-alone, and verifiable central claim in
   the given post: ${post}
   ```

2. Zero-shot-CoT: Zero-shot + `Let's think step by step.`

3. Few-shot: Four examples from the training set followed by the Zero-shot instruction

4. Few-shot-CoT: Few-shot + `Let's think step by step.`

2.5. Keyword Few-Shot (KBFP) and Self-Refine

This method explores a smart selection of few-shot examples using keyword matching, combined with (or without) a Self-Refine step. All implementations used the LLaMA 3.3 70B model.

Keyword Few-Shot. The `Keyword Few-Shot` method selects relevant examples from the training set by matching keywords found in the target social media post. These examples are then used to construct a few-shot prompt.

For the example post:

```
The salary of a U.S. Senator is $174,000 per year.  This is Joe Biden's house...
seems legit :)
```

The method extracts a claim such as:

```
The main claim is that Joe Biden's house appears to be too expensive for him to
afford on a U.S. Senator's salary of $174,000 per year, implying that there may
be some other, potentially questionable, source of income.
```

Self-Refine. As an additional step, we apply one iteration of the Self-Refine procedure [3], using the following prompt:

```
Refine the following claim to make it more precise.
Here is the text: ${the claim}
Output only the refined claim and nothing else.
```

While the base Keyword Few-Shot method yields a higher METEOR score on average, we observe that the addition of Self-Refine often produces more concise and less redundant claims.

The refined version of the above claim is:

```
Joe Biden's house appears to be too expensive to be affordable solely on the basis
of his U.S. Senator's salary of $174,000 per year, suggesting that there may be an
additional, unreported, or unexplained source of wealth that contributed to its
purchase or maintenance.
```

We find that repeated applications of Self-Refine do not improve the results. On the contrary, multiple iterations tend to introduce verbosity and hallucinated facts not grounded in the original post.

Issues with the Gold Claim. It's worth noting that the gold-standard claims themselves can have shortcomings. For instance, the gold claim for the example post is:

```
Joe Biden lives in a large estate bought on a senator's salary.
```

This omits key details like the senator's actual salary ($174,000), which may be necessary for verification, and doesn't reflect the original post's implication that the estate seems unaffordable based on that salary alone.

2.6. Subclaim Extraction and Filtering with Refinement

We explore a multi-stage approach that begins by extracting several potential claims, so-called "subclaims", from each social media post. These are then scored and filtered before a final synthesis step generates the main predicted claim.

In the first stage, the LLM is prompted to extract multiple subclaims from the post. These subclaims represent possible interpretations or factual assertions implied by the content.

Next, we introduce a filtering stage. Rather than passing all subclaims to the claim synthesis step, we rank them using a self-assigned importance score (1 to 10) and retain only those scoring 7 or higher. This limits noise and reduces the cognitive load on the synthesis LLM.

Despite these refinements, the METEOR score did not improve significantly. To address this, we added a third step: post-synthesis revision. A final LLM call revisits the synthesized claim, comparing it with the original post. It performs a "quality check" focused on factual accuracy, emphasis, and eliminating

redundancy or verbosity. The prompt in this stage explicitly instructs the model to consolidate language while preserving core meaning.

This three-stage pipeline—subclaim extraction, importance-based filtering, and post-synthesis refinement—aims to balance comprehensiveness with clarity and precision.

2.7. Max Multi-Prompt

We observed that many social media posts are comments on images found online, while the claims in our dataset often describe those images. If we use a generic prompt, the extracted content won't align well with the dataset's gold claims. To address this, we designed a prompt that instructs the model to imagine searching for the referenced image online and then describe its likely content.

We also noted that many posts rely on metaphor or sarcasm, often targeting the government. For instance, when a user writes, "Biden's annual salary is only \$170K," the implication—delivered with irony—is that Biden must be supplementing his income through questionable means to afford a luxury home. Similarly, posts about epidemics often question vaccine policies sarcastically, implicitly accusing public health authorities of negligence or malice.

To account for these nuances, we created targeted prompts tailored to each type of rhetorical device. Empirical results were obtained using the LLaMA 3.3 70B model.

This approach demonstrates the potential benefits of intelligently triaging between multiple prompt templates. To simulate an upper bound on this strategy, we applied several different prompts to the same social media post and selected the resulting claim with the highest METEOR score.

3. Approach: Baseline

3.1. Regurgitation Baseline

To evaluate the significance of our METEOR scores, we designed a naive baseline that simply reuses the original social media post or a truncated version as a stand-in for actual claim extraction. Surprisingly, this sets a strong reference point for METEOR performance.

We explored the following variations:

- Full social media post

- Truncating after the first 100 characters, omitting partial words at the end

- Using only the nouns and verbs from the post

On the validation set, using the full post yields a METEOR score of 0.19. Truncating the post leads to improved results, with the best configuration achieving 0.24. This baseline scored 0.23 on the test set.

By contrast, using only the nouns and verbs significantly reduces METEOR performance.

4. Experimental Evaluation

4.1. Setup

We used only the datasets provided by the CheckThat! Lab Task 2 organizers, focusing exclusively on the English language subset.

Several methods relied on LLMs without additional training, using in-context learning only. In cases where few-shot examples were used, they were selected from the training set. Table 1 lists which LLM was used in each method that contributed to our empirical results.

For methods involving fine-tuning, all available training data was used.

All approaches were evaluated on the validation subset using the METEOR metric, as implemented in the NLTK toolkit. While we submitted methods with the highest METEOR score, we were interested in exploring methods with the following criteria.

Table 1

LLMs Models used in respective approach. LLMs were either used from huggingface or via the Together.AI API service.

Section	Method Name	Model
2.1	Finetuned FLAN-T5-Large	FLAN-T5-Large
2.2	LoRA fine-tuning of Flan-T5 Base	Flan-T5 Base
2.3	Fine-tuned DeepSeek-R1-Distill-Llama-8b	DeepSeek-R1-Distill-Llama-8b
2.4	Claimifying social media posts with self-refinement	Grok3
2.5	Self-Refine with KBFP	Llama 3.3 70B
2.6	Subclaim extraction and filtering with refinement	Llama 3.3 70B
2.7	Max Multi-Prompt / Single prompt score	Llama 3.3 70B

Table 3

Empirical evaluation of our different methods. We focus on METEOR on the validation set, and give standard error bars where available. We submitted the three best performing systems and the baseline to the official leaderboard. Corresponding results on the test set are provided for those methods.

Section	Method name	METEOR on Validation Set		on Test Set
2.1	Finetuned FLAN-T5-Large	**0.5569**	±0.02	**0.37**
2.2	LoRA fine-tuning of FLan-T5 Base	0.3054		0.28
2.3	Fine-tuned DeepSeek-R1-Distill-Llama-8b (One-shot)	0.2541		
2.4	Claimifying social media posts with self-refinement	0.3310	±0.007	0.33
2.5	Keyword Based Few-shot Prompt (KBFP)	0.2943		
2.5	Self-Refine with KBFP	0.2392		
2.6	Subclaim extraction and filtering with refinement	0.2290	±0.005	
2.7	Max multi-prompt (simulated)	0.3277		
2.7	Single prompt score	0.2944		
3.1	Regurgitation Baseline (full post length)	0.1944	±0.008	
3.1	Regurgitation Baseline cut to 100	0.2429	±0.008	0.23

- Validation METEOR score

- Novelty and insightfulness of the approach

- Our subjective preference for the style of the extracted claims

5. Evaluation Results on Validation and Test Set

Table 3 presents METEOR evaluation results were obtained by our methods on the validation set provided for Task 2. We submitted the best methods and baseline to the challenge, for which test set results are reported as well.

Since Max multi-prompt is a simulated method to explore potential gains, we did not submit it to the challenge.

Discussion of Results. The best-performing method was fine-tuning Flan-T5-Large (Section 2.1), which achieved a METEOR score of 0.5569. This result highlights that larger models do not always guarantee better outcomes—Flan-T5-Large struck the best balance in our experiments. On the test set, this is still our best approach despite the performance drop.

Table 5
Our team ranked 9th on the overall leaderboard.

Rank	User	Test Set Results
1	tatiana.anikina	0.4569 (1)
2	DSGT-CheckThat	0.4521 (2)
3	Bharatdeep_Hazarika	0.4114 (3)
4	AKCIT-FN	0.4058 (4)
5	pratuat.amatya	0.4049 (5)
6	rohan_shankar	0.3920 (6)
7	manan-tifin	0.3881 (7)
8	MazenYasser74	0.3841 (8)
9	**UNH (Our Team)**	**0.3737**(9)
10	Ather-Hashmi	0.3565 (10)
11	teamopenfact	0.3370 (11)
12	Nikhil_Kadapala	0.3321 (12)
13	aryasuneesh	0.3153 (13)
14	Soumodeepsahaa	0.3098 (14)
15	uhh_dem4ai	0.2612 (15)
16	tomasbernal01	0.1660 (16)
17	VSE	0.0070 (17)

The runner-up is a prompting-based methods based on Claimify with Self-Refine, described in Section 2.4. It achieved a METEOR score of 0.331 on validation set, which manifested in a 0.33 on the official test set.

The third-best was T5-Base fine-tuned using LoRA. Although it achieved a lower score of 0.3054, its advantage lies in being a smaller and more efficient model. On the test set, it still obtains a reasonable 0.28 in METEOR.

The fourth method was Keyword-Based Few-Shot Prompting (KBFP), which used few-shot examples with LLaMA 3.3 70B. We did not submit this method to the challenge.

Despite Fine-tuning methods obtain the highest METEOR scores, we believe that somewhat better claims can be extracted with other approaches.

Subjectively, as we discuss below in Section 5.2, the most useful claims were generated by one iteration of Self-Refine combined with KBFP or Claimify, particularly for the Joe Biden house example. This approach correctly highlighted the assertion that the house seemed too expensive for his reported salary. In contrast, other methods either focused only on stating the salary amount ($174,000) or made vague claims about the size of the house. Several outputs hallucinated or speculated about visual content in the image, which was not part of the dataset.

We also observed that the gold claim for this example was not ideal: it omitted both the assertion about affordability and the actual salary figure—both of which are critical for verifying the claim.

5.1. Overall Leaderboard

Our best method, described in Section 2.1, placed us 9th on the leaderboard (Table 5). Notably, rank 12 was occupied by a test submission of the method from Section 2.4. Our naive baseline outperformed the final two teams in the rankings.

5.2. Examples of Claims Extracted on Validation Set

We manually reviewed extracted claims for the first few validation instances. We found considerable variation in claim styles: some were more actionable for fact-checking, while others focused more on the rhetorical tone or the poster's motivation.

Although these differences significantly impact utility for human fact-checkers, they are not well captured by the METEOR metric, which is limited to overlap with gold-standard claims.

Finetuned Flan-T5-Large. METEOR score: 0.5569, described in Section 2.1.

- Extracted Claim:

```
Joe Biden owns the house in this photo.
```

- Extracted Claim:

```
A Holocaust story told in the New York Times
```

- Extracted Claim:

```
This leopard cub's mother was killed by a trophy hunter
```

- Extracted Claim:

```
Video shows a crocodile spotted in a residential area in Hyderabad during the
ongoing heavy rains.
```

This method was able to closely mimic the patterns of the training dataset in some samples but still fails in other instances. It often does not produce proper claims, as it focuses solely on optimizing the METEOR score.

T5-base (LoRA, fine-tuned) METEOR Score: 0.3054, described in Section 2.2.

- Gold Claim: `Joe Biden lives in a large estate bought on a senator's salary`
  Extracted Claim: `Joe Biden's house is a fake`

- Gold Claim: `Photo shows Louis Armstrong as a child`
  Extracted Claim: `Jewish boy adopted by US Jewish family`

- Gold Claim: `Videos show current situation of Hyderabad amid heavy rain`
  Extracted Claim: `Video of a crocodile in Hyderabad`

This model extracts short, concise claims that are easy to verify. However, the claims are often overly literal and fail to capture the deeper meaning or intent behind the social media post.

Fine-tuned DeepSeek-R1-Distill-Lama-8b METEOR score: 0.2541 , described in Section 2.3.

- Extracted Claim:

```
The essential primary claim is that a U.S. Senator earns $174,000 per year,
and it is Joe Biden's house.
```

- Extracted Claim:

```
The Karnofsky family adopted a 7-year-old boy into their home, providing him
with food through his homework until he was 12, when they helped him buy his
first instrument, showcasing their support for his musical talent.
```

- Extracted Claim:

```
The essential primary claim is that trophy hunting is horrific.
```

- Extracted Claim:

```
The essential primary claim is that none of the listed items (Magarmacch,
Heavy Rain, Hyderabad, Crocodile, Alert) are the main focus.
```

- Extracted Claim:

```
The administration is now blaming the victims of today's deadly attacks in
Kabul for not leaving earlier.
```

The extracted claims are generally suitable for fact-checking and, in several cases, elaborate on the underlying message of the post.

We observed that the trained model often prefixes responses with phrases like "the essential primary claim is..." This model tends to extract short, concise claims that are easy to verify. However, these claims are frequently overly literal and fail to capture the deeper meaning or intent behind the social media post.

Claimifying social media posts with self-refinement   METEOR score: 0.332, Section 2.4.

- Gold Claim: ` Joe Biden lives in a large estate bought on a senator's salary `
  Extracted Claim: ` US Senator's annual salary is $174,000 `

- Gold Claim: ` Photo shows Louis Armstrong as a child `
  Extracted Claim:

```
A Lithuanian Jewish family employed a 7-year-old boy until he was 12 and gave
him money to buy his first instrument.
```

- Gold Claim: ` This leopard cub's mother was killed by a trophy hunter `
  Extracted Claim: ` Rescued animal's mother was killed by a hunter. `

- Gold Claim: ` Videos show current situation of Hyderabad amid heavy rain `
  Extracted Claim: ` Crocodile sighted in Hyderabad during heavy rain. `

- Gold Claim: ` Photo shows August 26, 2021 explosion near Kabul airport `
  Extracted Claim: ` Explosion outside Kabul airport kills 40 and injures 120. `

The generated claims are semantically closer to the gold-standard claims. In a separate analysis, we find that an average BERTScore (F1 mean) of 0.82 is achieved. Compared to the original input text, this indicates that the generated claims remain true to both the content and context of the topic, despite achieving a lower METEOR score.

Max multi-prompt.   METEOR score: (up to) 0.3277, described in Section 2.7.

- 1. ` The annual salary of a U.S. Senator is $174,000. `

- 2.

```
Owning such a home on that salary doesn't add up. This is a tongue-in-cheek
critique of perceived wealth versus official pay.
```

While the first prompt yields a claim that is objectively fact-checkable, we believe the second prompt better captures the motivation behind the social media post.

6. Overall Conclusions and Main Findings

We found that most LLMs produced claims related to the content of the social media posts. However, these often diverged from the gold standard claims, which negatively impacted their METEOR scores. When averaged across 1,170 validation examples, most methods converged on a METEOR score of approximately 0.27. The differences between methods only became apparent through manual inspection and checking for coverage of all key claim elements.

We also observed that many gold claims failed to capture all the critical assertions. In longer social media posts, multiple check-worthy claims were often present, making it difficult, even for human judges, to determine the primary claim without additional context about the user's information needs.

Several gold claims referenced images linked in the social media posts. Since these images were not included in the dataset, any claims based on them were speculative.

With sufficient fine-tuning, the smaller FLAN T5 model was able to approximate the method used to extract gold standard claims.

We found that multiple iterations of self-refinement often led LLMs to hallucinate or produce overly verbose responses. This pattern was consistent across a wide range of models, including LLaMA, GPT 4.1 nano, Gemini 2.0 Flash, and Grok3.

In many cases, the baseline outputs without self-refinement achieved higher METEOR scores and produced claims that more accurately reflected the content of the original post.

We also noted that instruction following was inconsistent. For example, directives such as omitting phrases like "The main claim is..." were often ignored, particularly by LLaMA models. Using structured outputs like JSON or pydantic formats improved adherence but frequently resulted in truncated outputs that were no longer valid JSON.

Acknowledgment

References

[1] M. Sundriyal, T. Chakraborty, P. Nakov, Overview of the CLEF-2025 CheckThat! lab task 2 on claim normalization, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 2025.

[2] D. Metropolitansky, J. Larson, Towards effective extraction and evaluation of factual claims, 2025. URL: https://arxiv.org/abs/2502.10855. arXiv:2502.10855.

[3] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, P. Clark, Self-refine: Iterative refinement with self-feedback, 2023. URL: https://arxiv.org/abs/2303.17651. arXiv:2303.17651.

[4] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, 2023. URL: https://arxiv.org/abs/2205.11916. arXiv:2205.11916.

A. Appendix

A.1. Prompts used in Claimifying social media posts with Self-refinement

Claim Extraction System Prompt

Identity / System Message

```
You are a helpful AI assistant and an expert in claim detection, extraction, and
normalization.
```

Instructions

```
 You are given a noisy social media post that contains only text but it might
  have been posted alongside a photo or video on the platform it was extracted
  from.

* Your task is to detect, extract, and respond with a normalized claim.

* A claim is a statement or assertion that can be objectively verified as true or
  false based on empirical evidence  or reality.

* Follow the below steps to analyze the input text and arrive at the final
  response.

* Step 1: Sentence Splitting and Context Creation

    * Start by splitting the post into individual sentences.
    * Now create or retrieve a context for each of those sentences by looking at
      the two preceding and two following sentences.

* Step 2: Selection

    * Now determine if each sentence contains any verifiable information based
      on the context created for the sentence in the previous step.
    * For each sentence, do the following:
        * If the sentence does not contain any verifiable information, discard
          that sentence.
        * If the sentence contains both verifiable and unverifiable information,
          rewrite the sentence retaining only verifiable information.
        * If the sentence doe not contain any unverifiable information, return
          the original sentence.

* Step 3: Disambiguation

    * The task here is to identify two types of ambiguity.
    * The first is referential ambiguity, which occurs when it is unclear what a
      word or phrase refers to. For example, in the sentence "They will update
      the policy next year", the terms "They," "the policy," and "next year" are
      ambiguous.
    * The second is structural ambiguity, which occurs when grammatical structure
      allows for multiple interpretations. For instance, the sentence "AI has
      advanced renewable energy and sustainable agriculture at Company A and
      Company B" can be interpreted as: (1) AI has advanced renewable energy and
      sustainable agriculture at both Company A and Company B, or (2) AI
      has advanced renewable energy at Company A, and it has advanced
      sustainable agriculture at Company B.
    * Now determine whether each instance of ambiguity can be resolved using the
```

question and the context.
  * The standard for resolution is whether a group of readers would likely
    agree on the correct interpretation. For example, recall the sentence "AI
    has advanced renewable energy and sustainable agriculture at Company A and
    Company B." If the context specified that Company A builds solar panels and
    Company B reduces farms' water usage, readers would likely conclude that AI
    has advanced renewable energy at Company A and sustainable agriculture at
    Company B. Conversely, if the context only described both companies as
    "environmental pioneers," readers would have insufficient information to
    determine the correct interpretation.
  * If any ambiguity is unresolvable, discard the sentence even if it has
    unambiguous, verifiable components.
  * If all ambiguity is resolved, return a clarified version of the sentence.
  * If there is no ambiguity, retain the original sentence for the next step.

* Step 4: Decomposition

  * Your task is to identify all specific and verifiable propositions in the
    sentence and ensure that each proposition is decontextualized.
    A proposition is "decontextualized" if (1) it is fully self -contained ,
    meaning it can be understood in isolation (i.e., without the question , the
    context , and the other propositions), AND (2) its meaning in isolation
    matches its meaning when interpreted alongside the question , the context ,
    and the other propositions. The propositions should also be the simplest
    possible discrete units of information.

* If no verifiable claims are found, return an extractive summary of the central
  idea of the post in a single sentence.

* Use only the words found in the original input text when generating a response.

* The claim must be strictly extracted from the input without adding any inferred
  or assumed context.

* The claim should be a concise single sentence (up to a maximum of 25 words)
  that captures the main point of the post without any additional context or
  details. Prioritize the main claim if multiple claims are present.

* The claim should be a self-contained factual statement that can be verified. It
  should not contain any subjective opinions, speculations, or interpretations.

* Pay attention to negative sentiment, named entities, names of people, and
  linguistic features like assertions, hedges, implications, etc. If any one
  of these features are present in the post, it should be reflected in the claim.

* Do not include any additional information or explanations in your response.

* Minor clarifications (e.g., implied agent) are allowed if they are logically
  unavoidable and directly inferable from the input.

* If the input text contains any Named Entities, they must be included in your
  responses.

* Return your response in the style of a short caption or headline of a news
  bulletin.

* If the given input text is mostly likely to be referencing or directly talking

about or posted alongside a photo or video, return the response that starts
with either (1) Photo shows <your_response> or (2) Video shows <your_response>.

* Always return your response in English even if the original input is in a
  different language.

Claim Extraction User Prompt

```
# Here are some examples of how to identify a decontextualized, stand-alone,
  and verifiable central claim in a post.
<user_query id="example-1">
**Identify the decontextualized, stand-alone, and verifiable central claim in
the given post:** Lieutenant Retired General Asif Mumtaz appointed as
Chairman Pakistan Medical Commission PMC Lieutenant Retired General Asif
Mumtaz appointed as Chairman Pakistan Medical Commission PMC Lieutenant
Retired General Asif Mumtaz appointed as Chairman Pakistan Medical Commission
PMC None.
Let's think step by step.
</user_query>

<assistant_response id="example-1">
**Normalized claim:**  Pakistani government appoints former army general to
head medical regulatory body.
</assistant_response>
-------------------------------------------

<user_query id="example-2">
**Identify the decontextualized, stand-alone, and verifiable central claim in
the given post:** A priceless clip of 1970 of Bruce Lee playing Table Tennis
with his Nan-chak !! His focus on speed A priceless clip of 1970 of Bruce Lee
playing Table Tennis with his Nan-chak !! His focus on speed A priceless clip
of 1970 of Bruce Lee playing Table Tennis with his Nan-chak !! His focus on
speed None
Let's think step by step.
</user_query>

<assistant_response id="example-2">
**Normalized claim:**  Late actor and martial artist Bruce Lee playing table
tennis with a set of nunchucks.
</assistant_response>
-------------------------------------------

<user_query id="example-3">
**Identify the decontextualized, stand-alone, and verifiable central claim in
the given post:** Hydrate YOURSELF W After Waking Up Water 30 min Before a Meal
DRINK Before Taking a Shower 2192 2192 Before Going to Bed at the correct time
T A YE Helps activate internal organs Helps digestion Helps lower blood
pressure. Helps to avoid heart attack Health+ by Punjab Kesari.
Let's think step by step.
</user_query>

<assistant_response id="example-3">
**Normalized claim:** Drinking water at specific times can have different
health benefits
</assistant_response>
-------------------------------------------
```

```
<user_query id="example-4">
**Identify the decontextualized, stand-alone, and verifiable central claim in
the given post:** Eating vaginal fluids makes you immune to cancer, and other
diseases. Do it for health. Scientists at St. Austin University in North
Carolina, they investigated the benefits of vaginal or cervical mucus
consumption and the results were amazing. These fluids contain high levels of
active proteins up to 10 minutes after leaving the female body. The vaginal
fluid is rich in protein, sodium, vitamins like C1, C4, C4, vc and others.
This study confirms what was exposed by Dr. John d. Moore in his 2009 study of
the "equivalent exchange" theory, which indicates that women and men benefit in
the same way. The benefits of "swallowing" vaginal fluids are:
1. **Eliminates buttons and buttons**
2. **Stimulates the electrical charges of the cells**
3. **Prevents prostate cancer.**
4. **Improved digestion.**
5. **Very effective against constipation.**
6. **It makes teeth and bones stronger.**
7. **Helps the functioning of the kidneys Share men! Everything is for your health!
    Share it on all social networks.**
Let's think step by step.
</user_query>

<assistant_response id="example-4">
**Normalized claim:** St. Austin University North Carolina says eating vaginal
fluid makes you immune to cancer
</assistant_response>
-------------------------------------------

<user_query id="example-5">
**Identify the decontextualized, stand-alone, and verifiable central claim in
the given post:** Corona virus before it reaches the lungs it remains in the
throat for four days drinking water a lot and gargling with warm water &
salt or vinegar eliminates the virus $\ldots$
Let's think step by step.
</user_query>

<assistant_response id="example-5">
**Normalized claim:** Gargling water can protect against coronavirus
</assistant_response>
-------------------------------------------

<|User|>: Identify the decontextualized, stand-alone, and verifiable central
claim in the given post: {input} <|End_user|>
Let's think step by step.
```

Feedback Generation Prompt

Identify / System Message

```
You are a professional fact-checker and an expert in claim normalization.
```

Instructions

```
Your task is to provide detailed, constructive feedback on the generated
response based on the criteria provided to ensure that the normalized claims
are not only consistent with the original post, but are also self-contained and
```

verifiable.
We want to iteratively improve the above generated response. To help with this, please score the response on the following criteria using a 0-10 scale, and provide a brief justification for each score:

1. **Verifiability:** To what extent does the response contain claims that can be independently verified using reliable sources? (0 = not verifiable, 10 = fully verifiable)
2. **Likelihood of Being False:** How likely is it that the response contains false or misleading information? (0 = very unlikely, 10 = very likely)
3. **Public Interest:** How likely is the response to be of general public interest or relevance? (0 = not interesting, 10 = highly interesting)
4. **Potential Harm:** How likely is the response to be harmful, offensive, or cause negative consequences? (0 = not harmful, 10 = extremely harmful)
5. **Check-Worthiness:** How important is it to fact-check this response? (0 = not worth fact-checking, 10 = highly worth fact-checking)

For each criterion, provide:
- A score (0-10)
- Provide a short, precise justification in 1 sentence.

Optionally, suggest specific improvements to the response based on your evaluation.

Response/Normalized Claim: ${Extracted Claim}

Refined Claim Generation Prompt

Identity / System Message

You are a professional fact-checker and expert in claim normalization.

Instructions

* Your task is to refine the genrated response in light of the feedback provided.
* Using the feedback provided, return a refined version of the generated response, ensuring that the normalized claim is consistent with the original post, self-contained, and verifiable.
* Your response must only be based on the feedback provided.
* Do not speculate, provide subjective opinions, or add any additional information or explanations.
* Only include the refined, normalized claim in your response.
* If no meaningful refinement is necessary, re-output the original normalized claim as-is.
* If the response is not not decontextualized, stand-alone, and verifiable, improve the response by adding more context from the original post if needed.

<|user_query|>{original user query}<|end_of_user_query|>
<|assistant_response|>{Initial Claim}<|end_of_assistant_response|>
<|feedback|>{feedback}<|end_of_feedback|>
<|instruction|>Based on the feedback provided, please refine the above generated response/normalized claim.<|end_of_instruction|>