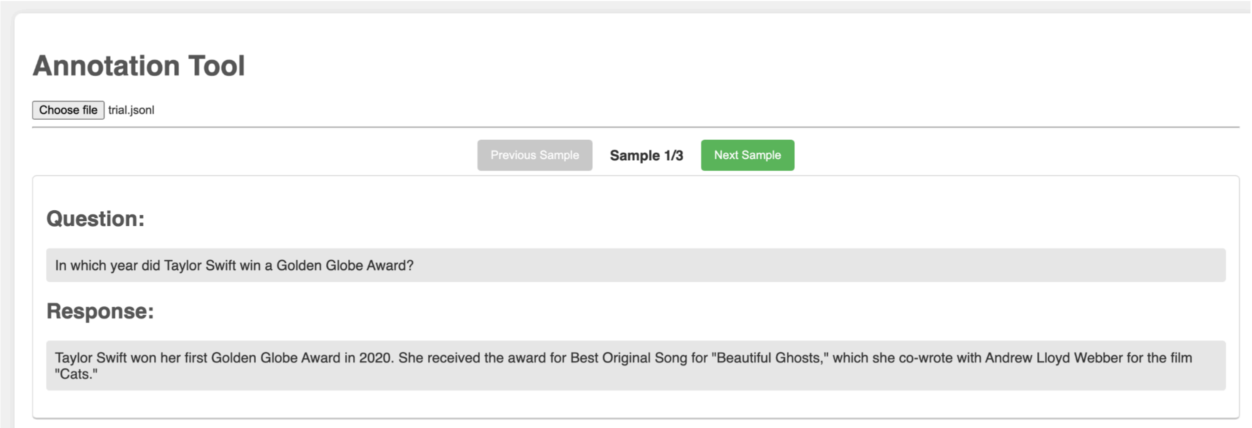
Annotation Protocol for FaStfact

Background

In this annotation task, you will review and revise a factuality evaluation pipeline of long-form generations from 4 sampled models on a few questions. To evaluate a model generation, the evaluation pipeline consists of three subprocesses: (1) claim extraction and preverification, (2) evidence collection, and (3) claim verification.

(1) **claim extraction and preverification**: A claim extractor decomposes a long-form generation into `atomic claims', by extracting verifiable factual statements from the model generation. If there is no extractable claim in the response, the extraction output would be "No verifiable claim.";
(2) **evidence collection**: retrieve or search for the evidence related to each claim;
(3) **claim verification**: verify each claim's factual correctness based on the evidence.

Lastly, each claim's classification label is aggregated together as a factuality score.



Question:

In which year did Taylor Swift win a Golden Globe Award?

Response:

Taylor Swift won her first Golden Globe Award in 2020. She received the award for Best Original Song for "Beautiful Ghosts," which she co-wrote with Andrew Lloyd Webber for the film "Cats."

Tasks

For each sample, you will review the outputs of (1) claim extraction and preverification and (3) claim verification subprocesses of the evaluation pipeline, and perform two tasks across two screens:

TASK 1: REVISE EXTRACTED CLAIMS

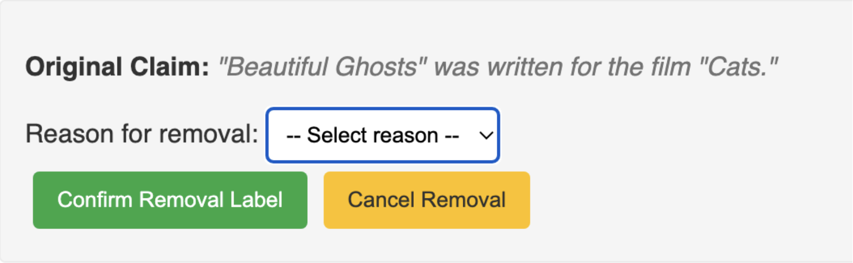
During this subprocess, the extractor extracts as many verifiable atomic factual claims mentioned in the model generation as possible. **Verifiable claims** describe a single event or state with all necessary modifiers (e.g., spatial, temporal, or relative clauses) that help denote entities or contextualize events in the real world, and can be verifiable given reliable external world knowledge (e.g., via Wikipedia).

- **Your input**: question + model response + a list of extracted claims.
- **Your task**: Based on the model generation and existing extracted claims, Task 1 is to annotate the quantity and quality of the claim extraction subprocess, by removing/adding/revising claims.

(1) **Remove/Add.** For a question-response pair, there should be a certain number of atomic facts that are included in the response, not more or less. Therefore, you should remove those that are {redundant claims, unextracted claims, not a (verifiable) claim}, and add missing claims that were contained in the model response.

(a) **Remove**: the two following types of claims should be removed. If removed a claim, you will be asked right away to label the reason for removal:

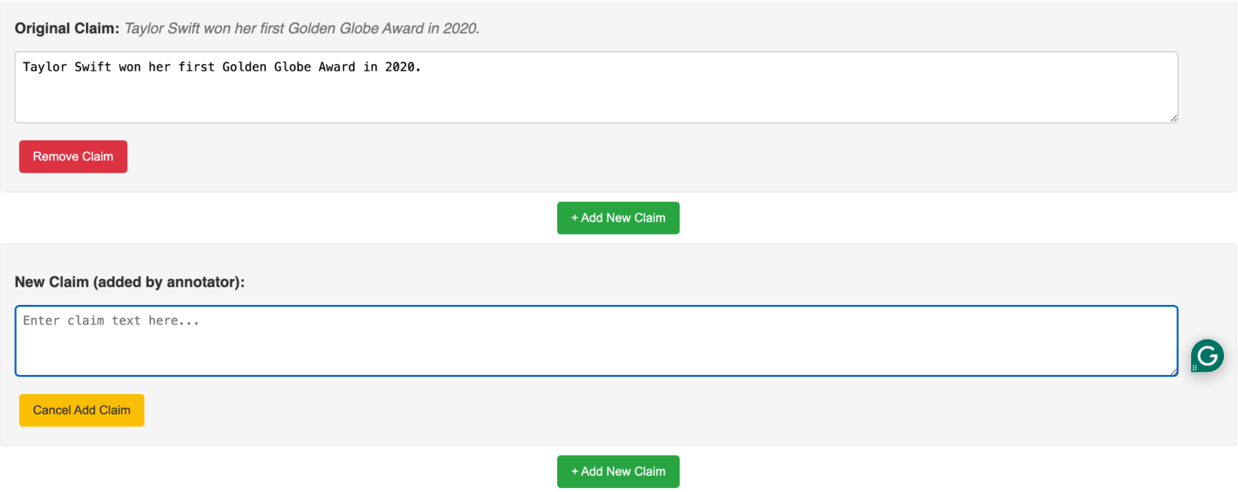
- Redundant: The meaning of the current claim overlaps exactly with part of another claim.
- Not a claim: The claim is about a story, personal experiences, hypotheticals (e.g., "would be" or subjunctive), subjective statements (e.g., opinions), future predictions, suggestions, advice, instructions, and other such content.



Original Claim: *"Beautiful Ghosts" was written for the film "Cats."*

Reason for removal: -- Select reason --

(b) **Add**: Add any missing verifiable claims in the exact sequence order corresponding to the response text.



Original Claim: *Taylor Swift won her first Golden Globe Award in 2020.*

Taylor Swift won her first Golden Globe Award in 2020.

New Claim (added by annotator):

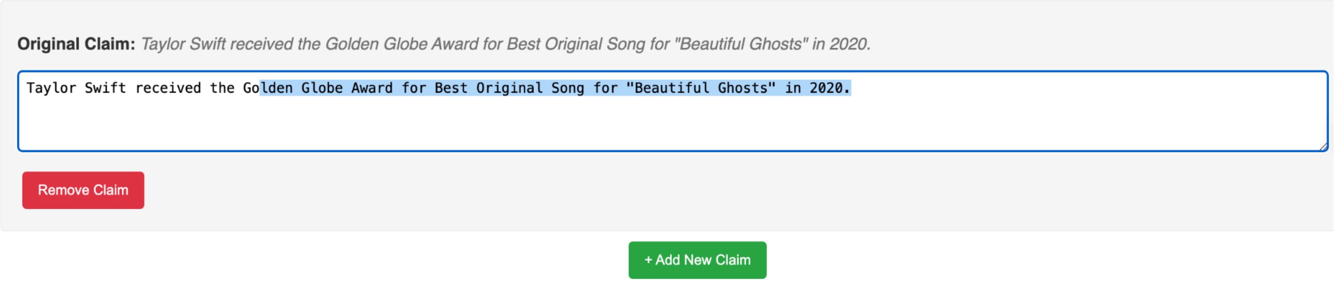
Enter claim text here...

(2) **Revise.** A verifiable claim should be context-independent and easily grounded with evidence search. Therefore, please make revisions to the extracted claim if a claim statement:

(a) Is ambiguous or contains equivocal information. Each extracted claim should also be describing either one single event (e.g., "Nvidia is founded in 1993 in Sunnyvale, California, U.S.") or single state (e.g., "UMass Amherst has existed for 161 years.") with necessary time and location information.

(b) Is contextually dependent. Revise the contextually dependent parts (e.g., pronouns) into verifiable named entities, based on the context information in the model response and the question. Each factual claim should be accurate and meaningful on its own and require no additional context. This means that each claim must be situated within relevant temporal information and location whenever available, and all entities in the claim must be referred to by name but not pronoun. Use the name of entities (e.g., 'Edward Jenner') rather than definite noun phrases (e.g., 'the doctor') whenever possible. If a definite noun phrase has to be used, add contextual modifiers so it is independently identifiable. Keep each factual claim to one sentence with zero or at most one embedded clause.

Note that the meaning of the extracted claims should adhere to the original model response, regardless of claim correctness. Your revision should not alter the meaning of the original response.

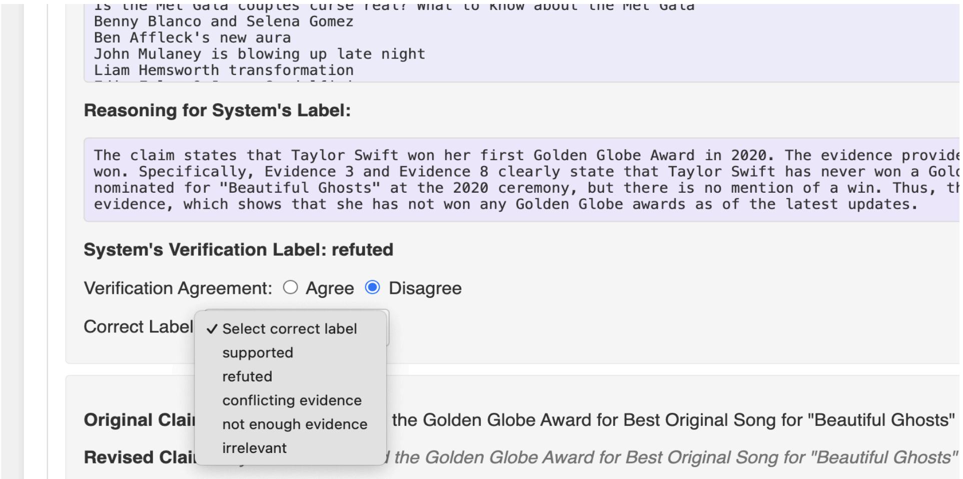


Original Claim: *Taylor Swift received the Golden Globe Award for Best Original Song for "Beautiful Ghosts" in 2020.*

```
Taylor Swift received the Golden Globe Award for Best Original Song for "Beautiful Ghosts" in 2020.
```

TASK 2. VERIFY THE REVISED CLAIMS

- **Your input**: Each (revised) claim + claim label + (OPTIONAL) searched evidence
- **Your task**: Decide whether you agree with the verification label for each revised claim; if not, give the correct verification label.
- **Your annotation labels:**
  - Agree
  - Disagree: if so, give the correct verification label based on your judgement.
    i. Irrelevant : A claim is completely irrelevant to the provided question.
    ii. Supported: A claim is supported if everything in the claim is supported and nothing is contradicted by the provided (or your own) search results. Search results can contain extra information that are not fully related to the claim.
    iii. Refuted: A claim is contradicted if a part of the claim is contradicted by a part of the provided (or your own) search results, and if no search result that supports the same part.
    iv. Conflicting evidence: A claim is controversial if a part of a claim cannot be verified by the provided (or your own) search results, because the claim is supported and contradicted by different mixed pieces of evidence in the search results.
    v. Not enough evidence: A claim is inconclusive if a part of a claim cannot be verified by the provided (or your own) search results, because there is not sufficient information in the search results related to the claim to make a verification.



```
Is the Met Gala couples curse real? What to know about the Met Gala
Benny Blanco and Selena Gomez
Ben Affleck's new aura
John Mulaney is blowing up late night
Liam Hemsworth transformation
```

Reasoning for System's Label:

```
The claim states that Taylor Swift won her first Golden Globe Award in 2020. The evidence provide
won. Specifically, Evidence 3 and Evidence 8 clearly state that Taylor Swift has never won a Golc
nominated for "Beautiful Ghosts" at the 2020 ceremony, but there is no mention of a win. Thus, th
evidence, which shows that she has not won any Golden Globe awards as of the latest updates.
```

System's Verification Label: refuted

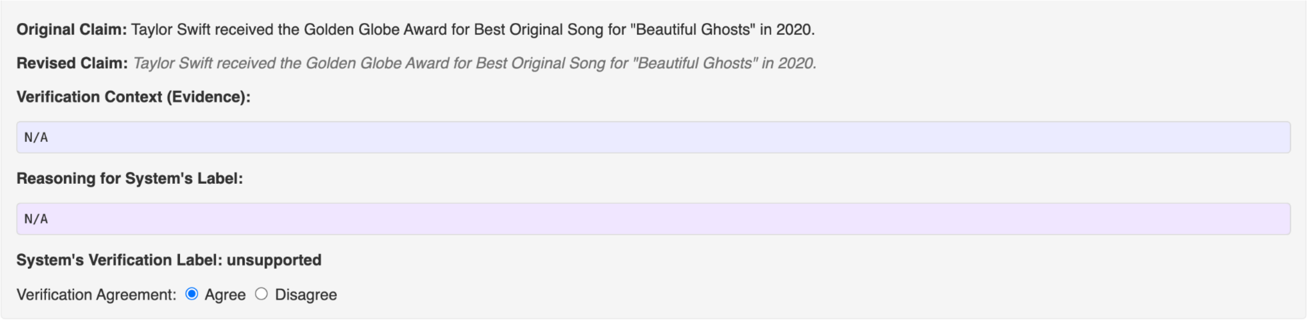
Verification Agreement: ○ Agree ● Disagree

Correct Label

Original Clai the Golden Globe Award for Best Original Song for "Beautiful Ghosts"

Revised Clai *d the Golden Globe Award for Best Original Song for "Beautiful Ghosts"*

- **Decision framework per claim**:
    a. Verifying a claim without searched evidence provided / Verifying your newly added claim:
        - You will decide whether to agree with the verification label without being provided any evidence. Where there is no explicit search evidence, the displayed verification labels would be only be [Irrelevant (i), Supported (ii), Unsupported (iii, iv, v)], with the Unsupported label encompassing the last three situations (iii, iv, v).
        - Manually search webs/wikipedia for as much evidence as needed, to decide the correct label of the claim. If you disagree, please choose from the 5 annotation labels above. You can click the 🔍 button beside the (revised) claim to automatically search about the claim on Google.



Original Claim: Taylor Swift received the Golden Globe Award for Best Original Song for "Beautiful Ghosts" in 2020.

Revised Claim: *Taylor Swift received the Golden Globe Award for Best Original Song for "Beautiful Ghosts" in 2020.*

Verification Context (Evidence):

```
N/A
```

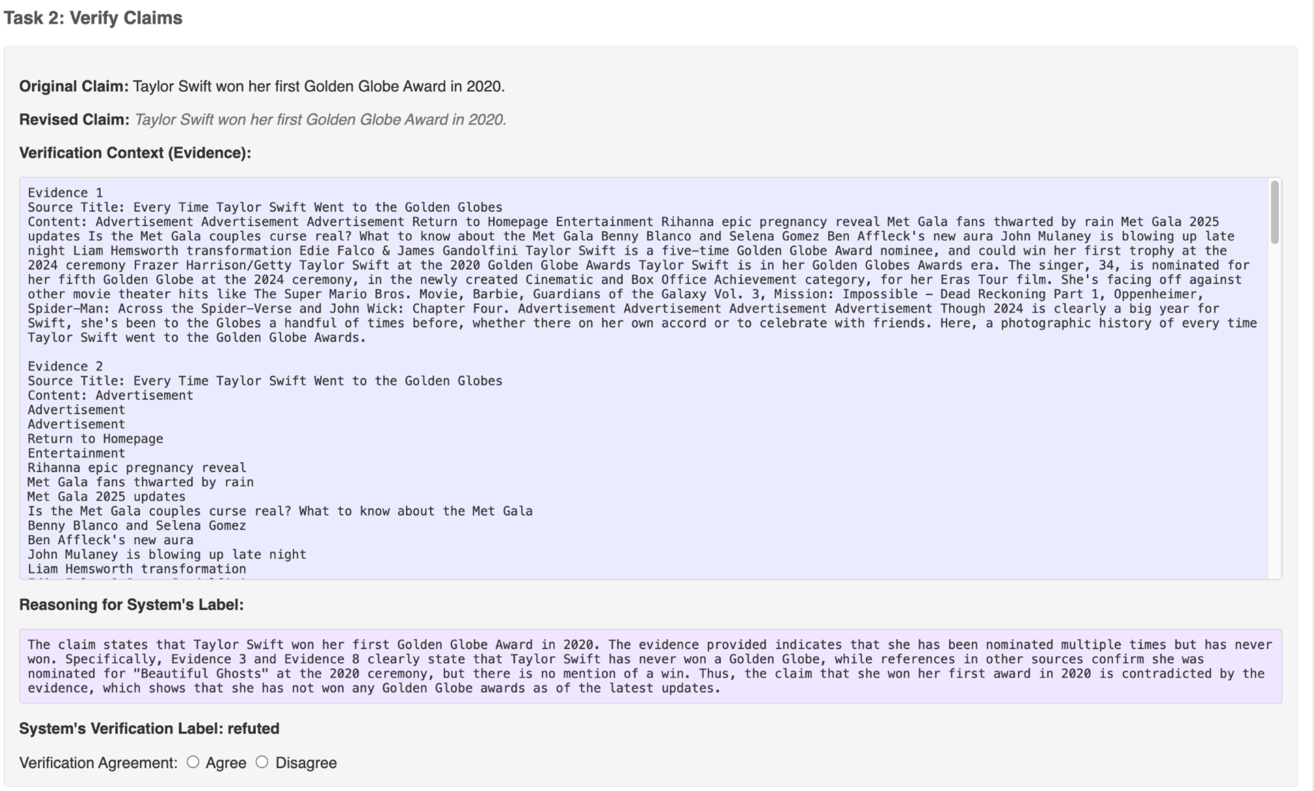
Reasoning for System's Label:

```
N/A
```

System's Verification Label: unsupported

Verification Agreement: ● Agree ○ Disagree

    b. Verifying a claim with searched evidence provided:
        - You will judge the model's verification label and its reasoning based on the searched evidence. The displayed verification labels would be from the same list of annotation labels above (i, ii, iii, iv, v).
        - Consider ONLY the evidence presented to decide the correct label of the current claim. Ignore any external knowledge that contradicts the evidence. If you disagree, please choose the correct label from the 5 annotation labels above.



Task 2: Verify Claims

Original Claim: Taylor Swift won her first Golden Globe Award in 2020.

Revised Claim: *Taylor Swift won her first Golden Globe Award in 2020.*

Verification Context (Evidence):

```
Evidence 1
Source Title: Every Time Taylor Swift Went to the Golden Globes
Content: Advertisement Advertisement Advertisement Return to Homepage Entertainment Rihanna epic pregnancy reveal Met Gala fans thwarted by rain Met Gala 2025
updates Is the Met Gala couples curse real? What to know about the Met Gala Benny Blanco and Selena Gomez Ben Affleck's new aura John Mulaney is blowing up late
night Liam Hemsworth transformation Edie Falco & James Gandolfini Taylor Swift is a five-time Golden Globe Award nominee, and could win her first trophy at the
2024 ceremony Frazer Harrison/Getty Taylor Swift at the 2020 Golden Globe Awards Taylor Swift is in her Golden Globes Awards era. The singer, 34, is nominated for
her fifth Golden Globe at the 2024 ceremony, in the newly created Cinematic and Box Office Achievement category, for her Eras Tour film. She's facing off against
other movie theater hits like The Super Mario Bros. Movie, Barbie, Guardians of the Galaxy Vol. 3, Mission: Impossible – Dead Reckoning Part 1, Oppenheimer,
Spider-Man: Across the Spider-Verse and John Wick: Chapter Four. Advertisement Advertisement Advertisement Advertisement Though 2024 is clearly a big year for
Swift, she's been to the Globes a handful of times before, whether there on her own accord or to celebrate with friends. Here, a photographic history of every time
Taylor Swift went to the Golden Globe Awards.

Evidence 2
Source Title: Every Time Taylor Swift Went to the Golden Globes
Content: Advertisement
Advertisement
Advertisement
Return to Homepage
Entertainment
Rihanna epic pregnancy reveal
Met Gala fans thwarted by rain
Met Gala 2025 updates
Is the Met Gala couples curse real? What to know about the Met Gala
Benny Blanco and Selena Gomez
Ben Affleck's new aura
John Mulaney is blowing up late night
Liam Hemsworth transformation
```

Reasoning for System's Label:

```
The claim states that Taylor Swift won her first Golden Globe Award in 2020. The evidence provided indicates that she has been nominated multiple times but has never
won. Specifically, Evidence 3 and Evidence 8 clearly state that Taylor Swift has never won a Golden Globe, while references in other sources confirm she was
nominated for "Beautiful Ghosts" at the 2020 ceremony, but there is no mention of a win. Thus, the claim that she won her first award in 2020 is contradicted by the
evidence, which shows that she has not won any Golden Globe awards as of the latest updates.
```

System's Verification Label: refuted

Verification Agreement: ○ Agree ○ Disagree