WikiContradict: A Benchmark for Evaluating LLMs on Real-World Knowledge Conflicts from Wikipedia (Supplement)

Yufang Hou[1]**, Alessandra Pascale**[1]**, Javier Carnerero-Cano**[1]**, Tigran Tchrakian**[1]
Radu Marinescu[1], **Elizabeth Daly**[1], **Inkit Padhi**[2], **Prasanna Sattigeri**[2]
[1] IBM Research Europe - Ireland
[2] IBM Research, Thomas J. Watson Research Center, Yorktown Heights, USA
{yhou|apascale|tigran|radu.marinescu|elizabeth.daly}@ie.ibm.com
{javier.cano|inkpad}@ibm.com, psattig@us.ibm.com

A Annotation Guidelines

Here we include the details about the pre-defined contradiction taxonomy and the full annotation guideline. The annotation interface was developed using Label Studio[1].

---

[1] https://labelstud.io/

Wikipedia Contradiction Annotation Guideline

Table of Contents

1. Setup the Label Studio environment

1) install Label Studio locally
   Github: https://github.com/HumanSignal/label-studio
   Here's the command I used to install it in my laptop (make sure python>=3.8):

   ```
   conda create --name label-studio
   conda activate label-studio
   conda install psycopg2
   pip install label-studio
   ```

   About LS version: I'm using Label Studio 1.8.2. Normally all new versions later than 1.8.2 should work, I recommend using google Chrome to do annotation.

2) run "label-studio" to start the server at http://localhost:8080
3) create an account and log in
4) create a new project called "WikipediaContradict" or any other names
5)

Fig 1: Create a new project

6) import the annotation tasks into the platform: first, click "upload files" and choose "AnnotationFilesSplit_new/inconsistentArticleTags_all_X.json"; next, click "save" to finish importing the 170 annotation tasks. Please refer to the excel file (AnnotationAssignment.xlsx) to identify which file you need to import.

Fig 2: Import annotated data

7) setup the annotation UI:
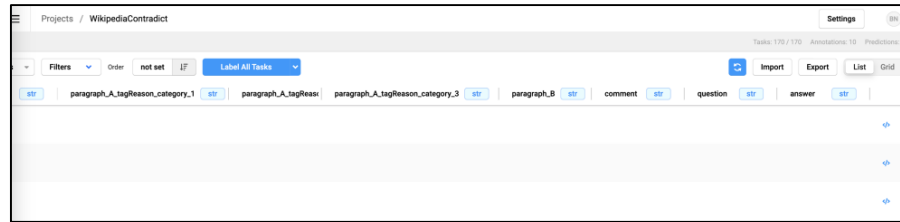   a) click "settings" on the upper right corner:

Fig 3: Setup the annotation UI

b) click "labelling interface", copy-paste the following UI code into the box (or the code from the "WikipediaContradiction_LabelSstudioUI"), then click "save". After this, you can start annotating by clicking "Label All Tasks" or any task in the panel.

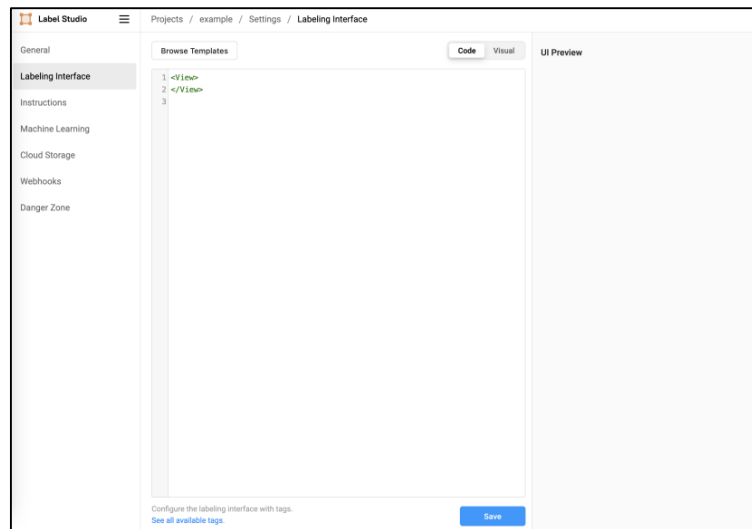Fig 4: Setup the annotation UI - Continue

```
<View>
<!-- <header name="articletitle" value="Wikipedia article: $title"/> -->
 <HyperText clickableLinks="true" name="articlelink" inline="true" target="_blank" value="">
   <h2><a target="_blank" href="$url">Wikipedia article: $title</a></h2>
 </HyperText>

 <View style="box-shadow: 2px 2px 5px #999;          padding: 20px; margin-top: 2em;          border-radius: 5px;">
 <Header value="Inconsistence or contradictory tag"/>
 <Text name="wikitag" value="$paragraph_A"/>
 </View>

 <View style="box-shadow: 2px 2px 5px #999;          padding: 20px; margin-top: 2em;          border-radius: 5px;">
 <Header value="* Step1: Is this tag valid?"/>
 <Choices name="wikitag_label_valid" toName="wikitag" choice="single" showInLine="true">
   <Choice value="Valid"/>
   <Choice value="Invalid"/>
 </Choices>
 <Header value="Additional comment" />
   <TextArea name="valid_comment" toName="wikitag"
       showSubmitButton="true" maxSubmissions="1" editable="true"
       required="false" />

</View>

<View visibleWhen="choice-selected" whenTagName="wikitag_label_valid" whenChoiceValue="Valid" style="box-shadow: 2px
2px 5px #999;          padding: 20px; margin-top: 2em;  border-radius: 5px; ">
<Header value="* Step 2: Copy the inconsistent sentences, paragraphs, or information (e.g., table or infobox rows) from the
wikipedia article(s) into the following boxes. Modify these copy-paste messages to ensure they form a coherent and valid stand-
alone text passage, if necessary."/>
<Header value="* Article A title" />
<TextArea name="paragraphA_article" toName="wikitag"
```

```
        showSubmitButton="true" maxSubmissions="1" editable="true"
        required="true" />

<Header value="* Article A relevant information (Passage 1)" />
  <TextArea name="paragraphA_information" toName="wikitag"
        showSubmitButton="true" maxSubmissions="1" editable="true"
        required="true" />
<Header value="[Modify] Article A relevant information_stand-alone (Passage 1)" />
  <TextArea name="paragraphA_information_standalone" toName="wikitag"
        showSubmitButton="true" maxSubmissions="1" editable="true"
        required="false" />


<Header value="* Article B title" />

  <TextArea name="paragraphB_article" toName="wikitag"
        showSubmitButton="true" maxSubmissions="1" editable="true"
        required="true" />

<Header value="* Article B relevant information (Passage 2)" />
  <TextArea name="paragraphB_information" toName="wikitag"
        showSubmitButton="true" maxSubmissions="1" editable="true"
        required="true" />
<Header value="[Modify] Article B relevant information_stand-alone (Passage 2)" />
  <TextArea name="paragraphB_information_standalone" toName="wikitag"
        showSubmitButton="true" maxSubmissions="1" editable="true"
        required="false" />

  <Header value="* Are Passage 1 and Passage 2 the same?"/>
  <Choices name="wikitag_label_samepassage" toName="wikitag" choice="single" showInLine="true">
    <Choice value="Same"/>
    <Choice value="Different" selected="true"/>
  </Choices>
</View>

<View visibleWhen="choice-selected" whenTagName="wikitag_label_valid" whenChoiceValue="Valid" style="box-shadow: 2px
2px 5px #999;          padding: 20px; margin-top: 2em;  border-radius: 5px; ">
<Header value="* Step 3: Contradiction reason" />
<Header size = "8">* Passage 1 (states that/contains):  </Header> <TextArea name="relevantInfo_comment_A" toName="wikitag"
        showSubmitButton="true" maxSubmissions="1" editable="true"
        required="true" />
<Header size = "8">* However, passage 2 (states that/contains):  </Header>   <TextArea name="relevantInfo_comment_B"
toName="wikitag"
        showSubmitButton="true" maxSubmissions="1" editable="true"
        required="true" />
<Header size = "8">If possible, copy the contradicted span from passage 1:  </Header> <TextArea
name="relevantInfo_comment_A_Span" toName="wikitag"
        showSubmitButton="true" maxSubmissions="1" editable="true"
        required="false" />

<Header size = "8">If possible, copy the contradicted span from passage 2:  </Header>   <TextArea
name="relevantInfo_comment_B_Span" toName="wikitag"
        showSubmitButton="true" maxSubmissions="1" editable="true"
        required="false" />

</View>

<style>
 .center-text {
   text-align: center;
 }
</style>


<View visibleWhen="choice-selected" whenTagName="wikitag_label_valid" whenChoiceValue="Valid" style="box-shadow: 2px
2px 5px #999;          padding: 20px; margin-top: 2em;  border-radius: 5px; ">
<Header value="* Step 4: Choose all options that can describe the above contradictory information"/>
  <Taxonomy name="taxonomy" toName="wikitag" required="true">
    <Choice value="Contradict type I">
      <Choice value="(PhraseLevel) Entity - Date/time" />
      <Choice value="(PhraseLevel) Entity - Location/GPE (Non-GPE locations, mountain ranges, bodies of water, and Countries,
cities, states)" />
      <Choice value="(PhraseLevel) Entity - Number" />
      <Choice value="(PhraseLevel) Entity - Organization (Companies, agencies, institutions, etc.)" />
      <Choice value="(PhraseLevel) Entity - Person" />
      <Choice value="(PhraseLevel) Entity - NORP (Nationalities or religious or political groups)" />
      <Choice value="(PhraseLevel) Entity - FAC (Buildings, airports, highways, bridges, etc.)" />
      <Choice value="(PhraseLevel) Entity - Work-of-Art (Titles of books, songs, etc.)" />
      <Choice value="(PhraseLevel) Entity - Product (Titles of books, songs, etc.)" />
      <Choice value="(PhraseLevel) Entity - Law (Named documents made into laws)" />
```

```
                    <Choice value="(PhraseLevel) Entity - Language (Any named language)" />
8                   <Choice value="(PhraseLevel) Entity - Event (Named hurricanes, battles, wars, sports events, etc.)" />
                    <Choice value="(PhraseLevel) Entity - Other" />
                    <Choice value="(PhraseLevel) NP-related (non-entity)" />
                    <Choice value="(PhraseLevel) Event/Relation (e.g., verb)" />
                    <Choice value="(DiscourseLevel) NP-related " />
                    <Choice value="(DiscourseLevel) Event/Relation-related " />
                  </Choice>
                  <Choice value="Contradict type II">
                    <Choice value="Text - Text" />
                    <Choice value="Text - Infobox/table" />
                    <Choice value="Infobox/table - Infobox/table" />
                    <Choice value="Other" />
                  </Choice>
                  <Choice value="Contradict type III">
                    <Choice value="Within the same article" />
                    <Choice value="Across different articles" />
                  </Choice>
                  <Choice value="Contradict type IV">
                    <Choice value="Explicit" />
                    <Choice value="Implicit (reasoning required)" />
                  </Choice>
                </Taxonomy>
              <Header value="Additional comment" />
                <TextArea name="contradict_comment" toName="wikitag"
                    showSubmitButton="true" maxSubmissions="1" editable="true"
                    required="false" />

            </View>

            <View visibleWhen="choice-selected" whenTagName="wikitag_label_valid" whenChoiceValue="Valid" style="box-shadow: 2px
            2px 5px #999;          padding: 20px; margin-top: 2em;  border-radius: 5px; ">
            <Header value="Step 5: Write at least one question that highlights the contradictions between Passage 1 and Passage 2, eliciting
            different responses based on each passage."/>
            <Header value="* Question 1" />
                <TextArea name="question1" toName="wikitag"
                    showSubmitButton="true" maxSubmissions="1" editable="true"
                    required="true"  />
            <Header value="* Answers for Question 1 (add the answers based on Passage 1 and Passage 2 seperately)" />
                <TextArea name="question1_answer1" toName="wikitag"
                    showSubmitButton="true" maxSubmissions="1" editable="true"
                    required="true" value = "add the answer based on Passage 1" />
                <TextArea name="question1_answer2" toName="wikitag"
                    showSubmitButton="true" maxSubmissions="1" editable="true"
                    required="true" value = "add the answer based on Passage 2"/>

            <Header value="Question 2" />
                <TextArea name="question2" toName="wikitag"
                    showSubmitButton="true" maxSubmissions="1" editable="true"
                    required="false" />
            <Header value="Answers for Question 2" />
                <TextArea name="question2_answer1" toName="wikitag"
                    showSubmitButton="true" maxSubmissions="1" editable="true"
                    required="false" value = "add the answer based on Passage 1" />
                <TextArea name="question2_answer2" toName="wikitag"
                    showSubmitButton="true" maxSubmissions="1" editable="true"
                    required="false" value = "add the answer based on Passage 2"/>


            <Header value="Additional comment" />
                <TextArea name="qa_comment" toName="wikitag"
                    showSubmitButton="true" maxSubmissions="1" editable="true"
                    required="false" />

            </View>

            <View visibleWhen="choice-selected" whenTagName="wikitag_label_valid" whenChoiceValue="Valid" style="box-shadow: 2px
            2px 5px #999;          padding: 20px; margin-top: 2em;  border-radius: 5px; ">
            <Header value="* Step 6: How confident do you feel about this annotation"  />
            <Rating name="confidence" toName="wikitag" defaultValue="5" required = "true"/>



            </View>
            </View>
```

Another way is to choose the assigned tasks (e.g., 31 - 40) and click "Label 10 Tasks", this will open the annotation window for these 10 tasks.

Fig 5: Choosing the annotation tasks

2. Task annotation

Please note that attributes marked with * are required.

Step1: Check whether the inconsistent tag is valid.
1) Open the Wikipedia article by clicking the corresponding link (Fig 6), identify the paragraph tagged with the inconsistent tag by searching "inconsistent" (Fig 7).

Fig 6: Annotation – step 1

In April 1956 the division was disbanded and transferred to the People's Liberation Army Navy and reorganized as 7th Institute of the Department of Defense, except its 3rd Battalion, 38th Garrison Regiment.[inconsistent]

This statement is inconsistent with other parts of the article. (December 2018)

Fig 7: Read the Wikipedia article

2) Read the Wikipedia article to check whether the tag makes sense. In the above example, the Wikipedia editor who added this tag didn't specify the reason as an attribute of the tag. Although this is the recommended template, they simply didn't follow the rules, so we need to investigate further. We know that this tag was added in December 2018, so we can check the editing history of this article to see under what condition this tag was added. In the revision history (see the figure below) we see that the editor put the reason in the edit comment "disbanded then 3 more years of events?" (last line in Fig 8). After checking this reason, if we agree that this inconsistent tag is valid, we go back to label studio and choose "valid" and put the reason in the additional comment box, as shown in Fig 1.

Fig 8: Checking the revision history of a Wikipedia article

Step2: find the inconsistent passages.

Copy the inconsistent sentences, paragraphs, or information (e.g., table or infobox rows) from the Wikipedia article(s) into the following boxes. Modify these copy-paste messages to ensure they form a coherent and valid stand-alone text passage, if necessary.

* Step 2: Copy the inconsistent sentences, paragraphs, or information (e.g., table or infobox rows) from the wikipedia article(s) into the following boxes. Modify these copy-paste messages to ensure they form a coherent and valid stand-alone text passage, if necessary.

* Article A title

10th Garrison Division (People's Republic of China)

* Article A relevant information (Passage 1)

In April 1956 the division was disbanded and transferred to the People's Liberation Army Navy and reorganized as 7th Institute of the Department of Defense, except its 3rd Battalion, 38th Garrison Regiment.

[Modify] Article A relevant information_stand-alone (Passage 1)

In April 1956 the 10th Garrison Division (People's Republic of China) was disbanded and transferred to the People's Liberation Army Navy and reorganized as 7th Institute of the Department of Defense, except its 3rd Battalion, 38th Garrison Regiment.

* Article B title

10th Garrison Division (People's Republic of China)

* Article B relevant information (Passage 2)

In March 1956 the division was reorganized as 3rd Garrison Division and was transferred to Guangzhou Military Region's control. Merely a month later, in April 1956 the division was further renamed as 3rd Machine-gun Artillery Division. The division then stationed in Zhongshan City, Guangdong. The machine-gun artillery division was also short-lived. In August 1956 the division was reduced and renamed as 10th Garrison Brigade. In May 1957 the brigade was inactivated and absorbed into Foshan Military Sub-district. In April 1959 the unit was re-activated from Foshan Military Sub-district as 10th Garrison Division. The division was then composed of: 37th Garrison Regiment; 38th Garrison Regiment; 39th Garrison Regiment; 1st Maritime Patrol Unit; 2nd Maritime Patrol Unit; 3rd Maritime Patrol Unit; 5th Maritime Patrol Unit. In December 1959, 39th Garrison Regiment and all four Maritime Patrol Units were transferred to Wanhu Fortress District's control.

[Modify] Article B relevant information_stand-alone (Passage 2)

In March 1956 the 10th Garrison Division (People's Republic of China) was reorganized as 3rd Garrison Division and was transferred to Guangzhou Military Region's control. Merely a month later, in April 1956 the division was further renamed as 3rd Machine-gun Artillery Division. The division then stationed in Zhongshan City, Guangdong. The machine-gun artillery division was also short-lived. In August 1956 the division was reduced and renamed as 10th Garrison Brigade. In May 1957 the brigade was inactivated and absorbed into Foshan Military Sub-district. In April 1959 the unit was re-activated from Foshan Military Sub-district as 10th Garrison Division. The division was then composed of: 37th Garrison Regiment; 38th Garrison Regiment; 39th Garrison Regiment; 1st Maritime Patrol Unit; 2nd Maritime Patrol Unit; 3rd Maritime Patrol Unit; 5th Maritime Patrol Unit. In December 1959, 39th Garrison Regiment and all four Maritime Patrol Units were transferred to Wanhu Fortress District's control.

* Are Passage 1 and Passage 2 the same?

☐ Same[3]   ☑ Different[4]

Fig 9: Annotation – step 2

For this annotation task, often "Article A title" and "Article B title" are the same. Note that "Article A relevant information (Passage 1)" and "Article B relevant information (Passage 2)" contain the original passage information from the Wikipedia, which means that you should copy-paste the original information into these boxes without modifying them. When copying the original passages into these boxes, please remove the citation marks and the inconsistent tags ({{inconsistent …}}). For "[Modify] Article A relevant information_stand-alone (Passage 1)" and "[Modify] Article B relevant information_stand-alone (Passage 2)", you are required to slightly modify the original passages to make them stand-alone (decontextualization). Normally, this requires you to resolve the coreference anaphors or the bridging anaphors in the first sentence. In Wikipedia, oftentimes the antecedents for these anaphors are the article titles themselves. If the original passage 1 or passage 2 are copy-pasted from tables/infoboxes, please write a stand-alone text to express the meaning of the copied part.

Example of resolving coreference anaphors:
In the example shown in Fig 9, we replace "the division" in the first sentence of both passages as "the 10th Garrison Division (People's Republic of China)"

Below is another example of resolving coreference anaphors:
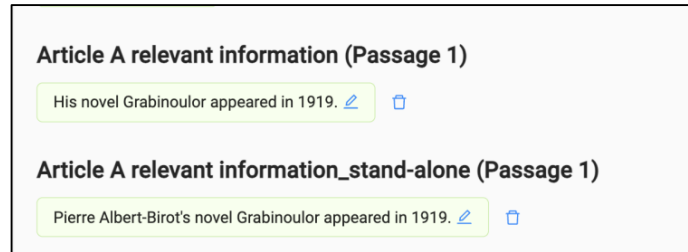
Fig 10: resolving coreference anaphors

<u>Example of resolving bridging anaphors (one of my favorite topics ☺)</u>:
In the following example as shown in Fig 11, we replace "The larvae" in the first sentence as "The larvae of Antherina"
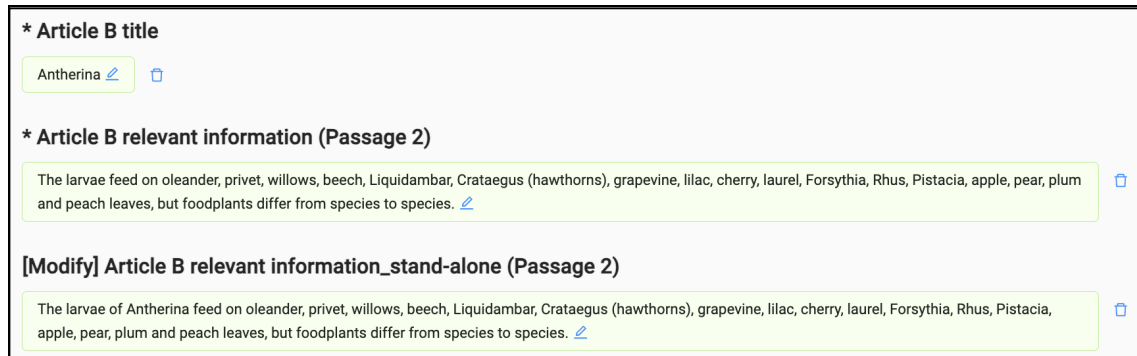
Fig 11: resolving bridging anaphors

By default, passsage1 and passage 2 are different. Sometimes it could be very difficult to clearly distinguish passage 1 from passage 2, choose "same" for the question "Are Passage1 and Passage the same?" Such cases often involve inference to figure out why there's a contradiction, as shown in the following figure:

13

Step3: Annotate the contradiction reason.

In this step, we use the template "passage 1 states that …, however, passage 2 states that …" to annotate the contradiction reason (see Fig 12). If possible, please copy-paste the exact contradicted spans (short phrases within a sentence) from both passages, such as "was disbanded" from passage 1 as shown in Fig 12. In this example, we leave the contradicted span from passage 2 empty since it involves a series of events across multiple sentences. Fig 13 shows an example in which we can easily identify the contradicted spans in both passages.

Fig 12: Annotate the contradiction reason

11

> *** Step 3: Contradiction reason**
>
> *** Passage 1 (states that/contains):**
>
> The northbound span of 14th Street bridges was renamed the Arland D. Williams Jr. Memorial Bridge in 1983. ✎   ⊓
>
> *** However, passage 2 (states that/contains):**
>
> The northbound span of 14th Street bridges was renamed the Arland D. Williams Jr. Memorial Bridge on March 13, 1985. ✎   ⊓
>
> **If possible, copy the contradicted span from passage 1:**
>
> 1983 ✎   ⊓
>
> **If possible, copy the contradicted span from passage 2:**
>
> March 13, 1985 ✎   ⊓

<p align="center">Fig 13: Annotate the contradiction reason – another example</p>

Step4: Annotate the contradiction types.

In this step, we assign the identified contradictory information to the appropriate types according to a pre-defined taxonomy. Back to our first example as shown in Fig 6/9/12, we assign it to the following contradiction types as shown in Fig 14

> **Step 4: Choose all options that can describe the above contradictory information**
>
> (DiscourseLevel) Event/Relation-related ✕   Text - Text ✕   Within the same article ✕   Implicit (reasoning required) ✕

<p align="center">Fig 14: Annotate the contradiction types</p>

Below we provide more details four the pre-defined four contradiction types.

    1) Contradiction type I: As shown in Fig 15, contradiction type I focuses on the fine-grained semantics of the contradiction.

> ∨ ⊟ Contradict type I   17
> ☐ (PhraseLevel) Entity - Date/time
> ☐ (PhraseLevel) Entity - Location/GPE (Non-GPE locations, mountain ranges, bodies of water, and Countries, cities, states)
> ☐ (PhraseLevel) Entity - Number
> ☐ (PhraseLevel) Entity - Organization (Companies, agencies, institutions, etc.)
> ☐ (PhraseLevel) Entity - Person
> ☐ (PhraseLevel) Entity - NORP (Nationalities or religious or political groups)
> ☐ (PhraseLevel) Entity - FAC (Buildings, airports, highways, bridges, etc.)
> ☐ (PhraseLevel) Entity - Work-of-Art (Titles of books, songs, etc.)
> ☐ (PhraseLevel) Entity - Product (Titles of books, songs, etc.)
> ☐ (PhraseLevel) Entity - Law (Named documents made into laws)
> ☐ (PhraseLevel) Entity - Language (Any named language)
> ☐ (PhraseLevel) Entity - Event (Named hurricanes, battles, wars, sports events, etc.)
> ☐ (PhraseLevel) Entity - Other
> ☐ (PhraseLevel) NP-related (non-entity)
> ☐ (PhraseLevel) Event/Relation (e.g., verb)
> ☐ (DiscourseLevel) NP-related

☑ (DiscourseLevel) Event/Relation-related   ...

Fig 15: Contradiction type I

<u>Contradiction Type 1 – Phrase level – Entity:</u> The contradictory information is around two named entities in passage 1 and passage 2. Normally phrase level contradiction can be easily fixed by changing one of the named entities (if we know which one is factually correct). We adapt OntoNotes named entity type definitions to describe the different types of contradicted named entities. For each named entity type, its and the explanation included in the parenthesis should provide a clear definition of the corresponding named entity type. The following example shown in Fig 16 was assigned to "(phraseLevel) – Entity-Date/time".

Step 3: Contradiction reason

Passage 1 states that:

The northbound span of 14th Street bridges was renamed the Arland D. Williams Jr. Memorial Bridge in 1983.

However, passage 2 states that:

The northbound span of 14th Street bridges was renamed the Arland D. Williams Jr. Memorial Bridge on March 13, 1985.

If possible, copy the contradicted span from passage 1:

1983

If possible, copy the contradicted span from passage 2:

March 13, 1985

Step 4: Choose all options that can describe the above contradictory information

(PhraseLevel) Entity - Date/time ✕   Text - Text ✕   Within the same article ✕   Explicit ✕

Fig 16: Contradiction type example

<u>Contradiction Type 1 – Phrase level – Non-entity NP:</u> The contradictory information is around two noun phrases that are not named entities in passage 1 and passage 2. In the following example as shown in Fig 17, the contradicted information are around two common nouns: monotypic (passage 1) and species to species (passage 2).

Step 3: Contradiction reason

Passage 1 states that:

Antherina is a monotypic. ## Additional Explanation: According to the world knowledge, a monotypic species is one that does not include subspecies.

However, passage 2 states that:

The foodplants for the larvae of Antherina differ from species to species.

If possible, copy the contradicted span from passage 1:

monotypic

If possible, copy the contradicted span from passage 2:

species to specie

Step 4: Choose all options that can describe the above contradictory information

(PhraseLevel) NP-related (non-entity) ✕   Text - Text ✕   Within the same article ✕   Implicit (reasoning required) ✕

Click to add... ∨

Fig 17: Contradiction type example

Contradiction Type 1 – Phrase level – Event/relation: The contradictory information is around two verb phrases that are describe two contradicted events or relations in passage 1 and passage 2. In the following example as shown in Fig 18, the contradicted information are around two verbs: "stimulates" (passage 1) and "inhibits" (passage 2).

Fig 18: Contradiction type example

Contradiction Type 1 – Discourse level – NP-related: The contradictory information is beyond the single sentence level from passage 1 and passage 2. The contradicted information from passage 1 can be anchored to a NP span, and the contradicted information from passage 2 is across multiple sentences. In the following example as shown in Fig 19, the contradicted information from passage 1 can be anchored to an NP: "five periods of time", and the contradicted information from passage 2 contains a few paragraphs that contains three sub-section headers indicating three time periods.

Fig 19: Contradiction type example

Contradiction Type 1 – Discourse level – Event/relation-related: The contradictory information is beyond the single sentence level from passage 1 and passage 2. The contradicted information from passage 1 can be anchored to a verb phrase, and the contradicted information from passage 2 is across multiple sentences. In the following example as shown in Fig 20, the contradicted information from passage 1 can be anchored to a VP: "was disbanded", and the contradicted information from passage 2 contains a few paragraphs that describes a series of events.

14

17

Fig 20: Contradiction type example

2) Contradiction type II: As shown in Fig 21, contradiction type II focuses on the modality the contradiction. It describes the modality of passage 1 and passage 2, whether the information is from a piece of text, or from a row an infobox or a table. Fig 22 shows an example of "contradict type II – infobox/table – Infobox/table".

Fig 21: Contradiction type II

Fig 22: Contradiction type example

3) Contradiction type III: As shown in Fig 23, contradiction type III focuses on the source the contradiction. It describes whether passage 1 and passage 2 are from the same article or not. For the inconsistent tags, most of contradictions are from the same article. In a few rare cases, the contradiction is from different articles. Fig 24 illustrates such an example: passage 1 is from the English version of the Wikipedia article "Pierre Albert-Birot" and passage 2 is from the French version of the article.

Fig 23: Contradiction type III

Fig 24: Contradiction type example

4) Contradiction type IV: As shown in Fig 25, contradiction type IV focuses on the reasoning aspect. It describes whether the contraction is explicit or implicit. Implicit contradiction requires us do to some reasoning to understand why passage 1 and passage 2 are contradicted. In the example shown in Fig 17, to understand that "monotypic (passage 1)" and "species to species (passage 2)" are contradicted, we need to carry out additional reasoning steps, i.e., first we need to know that according to commonsense knowledge, a monotypic species is one that does not include subspecies; then "species to species" entails that there are more than one specie. Therefore, we can conclude that passage 1 is contradicted with passage 2.

Fig 25: Contradiction type IV

Note: If the contradictory information exhibits other important attributes that are not covered by the existing taxonomy, using the additional comment box to describe it. Figure 26 shows such an example: passage 1 is from the English Wikipedia article, and passage 2 is from the French Wikipedia article, therefore in the additional comment for Step 4 we put "across different languages"

19

Fig 26: Additional contradiction Type

Step5: Annotate the question and answers.

In this step, we formulate at least one question that highlights the contradictions between Passage 1 and Passage 2, eliciting different responses based on each passage. Fig 27 shows a question and answers example about the contradicted information in Fig 16.

Fig 27: write a question and different answers

Step6: Annotate the confidence level for all annotations associated with the valid tag

3. Export annotations

After finishing the assigned annotation tasks, go back to the task pool panel, click "Export" and choose "JSON" as the format (Fig 28). Please rename the exported file using the following template: wikipediaContradict_10_<your name>.json

Fig 28: Export annotations

B   Details on the Prompt Templates 5.1 - 5.4 and Human Evaluation Results

Here we include the details of the Prompt Templates 5.1 - 5.4 and Human Evaluation Results (see Section 4 of the paper).

Prompt 5.1-5.4

Prompt 5.1

Provide a short answer for the following question based on the given context. Carefully investigate the given context and provide a concise response that reflects the comprehensive view of the context, even if the answer contains contradictory information reflecting the heterogeneous nature of the context.

Question: {Question}
Context: {Passage 2} + {Passage 1}

Prompt 5.2

Context: {Passage 1} + {Passage 2}

Does the above provided context contain conflicting information that could result in different answers to the question {Question}?" Provide a short answer followed by a concise explanation.

Prompt 5.3

Provide a short answer for the following question based on the given context. Carefully investigate the given context and provide a concise response that reflects the comprehensive view of the context, even if the answer contains contradictory information reflecting the heterogeneous nature of the context.

Question: {Question}
Context: {Passage 1'} + {Passage 2'}

Prompt 5.4

Context: {Passage 1'} + {Passage 2'}

Does the above provided context contain conflicting information that could result in different answers to the question {Question}?" Provide a short answer followed by a concise explanation.

Additional Human Evaluation Results

	Mistral-7b-instruct			Mixtral-8x7b-instruct			Llama-2-70b-chat			Llama-3-70b-instruct			GPT-4		
Prompt Template 5.1															
correct	22.9	30.0	11.8	14.6	16.7	5.9	17.0	27.6	0.0	54.2	70.0	23.5	12.5	16.7	5.9
partially correct	72.9	66.7	82.4	70.8	73.3	70.6	80.9	72.4	94.1	45.8	30.0	76.5	81.3	83.3	76.5
incorrect	4.2	3.3	5.9	14.6	10.0	23.5	2.1	0.0	5.9	0.0	0.0	0.0	6.3	0.0	17.6
Prompt Template 5.2															
correct	35.4	40.0	29.4	4.2	6.7	0.0	76.6	89.7	52.9	77.1	93.3	47.1	89.6	96.7	76.5
partially correct	14.6	20.0	0.0	6.3	6.7	5.9	17.0	10.3	29.4	6.3	3.3	11.8	2.1	3.3	0.0
incorrect	50.0	40.0	70.6	72.9	76.7	70.6	6.4	0.0	17.6	14.6	3.3	35.3	8.3	0.0	23.5
Prompt Template 5.3															
correct	87.5	93.3	82.4	85.4	83.3	88.2	87.2	96.6	76.5	89.6	96.7	82.4	87.5	96.7	70.6
partially correct	6.3	3.3	5.9	4.2	6.7	0.0	2.1	0.0	5.9	2.1	0.0	5.9	2.1	0.0	5.9
incorrect	6.3	3.3	11.8	10.4	10.0	11.8	10.6	3.4	17.6	8.3	3.3	11.8	10.4	3.3	23.5
Prompt Template 5.4															
correct	66.7	70.0	64.7	83.3	83.3	82.4	12.8	6.9	23.5	47.9	50.0	41.2	43.8	50.0	35.3
partially correct	4.2	6.7	0.0	2.1	3.3	0.0	23.4	24.1	17.6	4.2	3.3	5.9	14.6	10.0	23.5
incorrect	29.2	23.3	35.3	2.1	3.3	0.0	63.8	69.0	58.8	45.8	43.3	52.9	41.7	40.0	41.2

C   Further Details on the Judge LLM Prompt

Here we include further details on the judge LLM prompt (see Section 5 of the paper).

System prompt:

Please evaluate the response to a question under relaxed evaluation, where hallucinations, outdated information are allowed, as long as the primary answer is accurate. Each response is evaluated as correct, partially correct, or incorrect. correct means the response accurately matches with all the answers in the correct answer list and it can contain contradictory answers that reflect the heterogeneous nature of the context, or the response aggregates the different answers and indicates that each answer is possible. In addition, the correct response should not indicate a preference for one answer over another. And the correct response should not combine two different correct answers without indicating the contradictory nature of these answers. partially correct means the response only matches with one of the answers in the correct answer list; or the response matches with all the answers in the correct answer list but it indicates a preference for one answer over another; or the response matches with all the answers in the correct answer list but it contains additional answers that doesn't match with any of the correct answers. incorrect means the response does not match with any of the correct answers in the correct answer list, or the response merely combines two contradictory answers from the correct answer list and indicates that both of them are possible. Note that for each question, there are multiple correct answers based on different sources even though these correct answers contradict each other. Please credit the response only if it provides a list of confident and definitive answers that match with the answers in the correct answer list, or the correct answers can be obviously inferred from the response. The primary or final answers when standing alone must be accurate. Any additional information that is provided must not contradict the primary answers or reshape one's perception of them. For answers that involve names of entities (e.g., people), complete names or commonly recognized names are expected. Regarding numerical answers, approximate numbers are generally not accepted unless explicitly included in the ground-truth answers. We accept responses that contain hallucinated or outdated information that does not significantly impact the primary answers.

Examples

Example 1
 question: How old is the world's oldest verified living person?
 correct_answers: 115 years old | 117 years old
 response: As of today the most up-to-date and relevant information regarding this query is as follows. The oldest verified living person is Maria Branyas Morera, who was born on March 4, 1907, making her 117 years old.
 comment: The primary answer in the response (117 years old) is partially correct as it only matched with with one of the correct answers (117 years old) and it doesn't mention another answer (115 years old). Thus, the response is partially correct.
 evaluation: partially correct

Example 2
 question: How old is the world's oldest verified living person?
 correct_answers: 115 years old | 117 years old
 response: According to the context, the oldest verified living person is Maria Branyas Morera, who is both 115 years old and 117 years old.
 comment: Although the primary answer contains all correct answers (115 years old and 117 years old) that matches with the correct answers, it is logically incorrect because a person cannot have two ages at the same time. Thus, the response is incorrect.
 evaluation: incorrect

Example 3
 question: How old is the world's oldest verified living person?,
 correct_answers: 115 years old | 117 years old

response: According to the context, one source claims that the oldest verified living person is Maria Branyas Morera, who is 117 years old. However, another source claims that she is 115 years old.

comment: The answer contains all correct answers: 115 years old and 117 years old, and it points out that these two answers are from different sources, which is logically possible. Thus, the response is correct.

evaluation: correct

Example 4

question: How old is the world's oldest verified living person?,

correct_answers: 115 years old | 117 years old

response: According to the context, the oldest verified living person is Maria Branyas Morera, who is either 117 years old or 115 years old.

comment: The answer contains all correct answers: 115 years old and 117 years old that aggregates the different answers from different sources, which is logically possible. Thus, the response is correct.

evaluation: correct

Example 5

question: How many books has Colleen Hoover published in 2020?

correct_answers: 26 books | 27

response: according to some sources, Colleen Hoover has published 26 books in 2020. However, other sources claim that it has published 27 books in 2020.

comment: The response accurately provides two answers (26 books and 27 books) based on different sources. It recognized the two sources that provide different answers. These two answers match with all correct answers (26 books, 27). Thus, the response is correct.

evaluation: correct

Example 6

question: How many books has Colleen Hoover published in 2020?

correct_answers: 26 books | 27

response: Colleen Hoover has published 26 books in 2020.

comment: The response only provides one answers (26 books) based on the given context, which matched with one of the answers in the correct answer list. Thus, the response is partially correct.

evaluation: partially correct

Example 7

question: How many member nations exist in the United Nations?

correct_answers: 193 members | 194 member states

response: As of my last update in April 2023, there were 191 member nations in the United Nations. Please note that this information may have changed since then, so it is advisable to check the latest information from the United Nations or other reliable sources.

comment: The primary answer in the response (191 member nations) does not match with any of the correct answers (193 member states or 194 members). Thus, the response is incorrect.

evaluation: incorrect

Example 8

question: How many member nations exist in the United Nations?

correct_answers: 193 members | 194 member states

response: Although some sources claim that there are 193 member nations in the United Nations, the UN is made up of 194 member states.

comment: Although the primary answer mentions two answers (193 members, 104 member states) that match with the correct answers, the response indicates a preference of one answer (194 member states) over the other (193 members). Thus, the response is partically correct.

evaluation: partically correct

Testing instance

question: {Question}

correct_answers: {Answer1} | {Answer2}

response: {LLM response}

comment:

D Additional Experimental Settings

To understand LLMs' behavior when faced with real-world inter-context conflicts (human evaluation), we tested *Mistral-7b-instruct*, *Mixtral-8x7b-instruct*, *Llama2-2-70b-chat*, *Llama3-70b-instruct*, and *GPT-4-turbo-2024-04-09*. For the judge LLMs (automatic evaluation), we used *Mistral-7b-instruct*, *Mixtral-8x7b-instruct*, *Llama2-2-70b-chat*, *Llama3-70b-instruct*, *GPT-4-turbo-2024-04-09*, and *GPT-4o-2024-05-13*. These models were selected due to their state-of-the-art performance in natural language processing tasks and their robustness across a wide range of applications. For *Mistral-7b-instruct*, *Mixtral-8x7b-instruct*, *Llama2-2-70b-chat*, and *Llama3-70b-instruct*, we set the maximum number of tokens to 250, the minimum number of tokens to 1, and the decoding method to greedy search. For *GPT-4-turbo-2024-04-09* and *GPT-4o-2024-05-13*, we used the OpenAI Chat Completions API with the following settings: temperature = 0, max_tokens = 250, seed = 5, messages = [{"role": "system", "content": ""}, {"role": "user", "content": prompt_1}].

E Data Format

WikiContradict is given in JSON format to store the corresponding information, so researchers can easily use our data. There are 253 instances in total.

An example of our JSON format is:

```
{
    "title": "",
    "url": "",
    "paragraph_A": "",
    "paragraph_A_clean": "",
    "tag": "",
    "tagDate": "",
    "tagReason": "",
    "annotationResult": {
        "wikitag_label_valid": "",
        "valid_comment": "",
        "paragraphA_article": "",
        "paragraphA_information": "",
        "paragraphA_information_standalone": "",
        "paragraphB_article": "",
        "paragraphB_information": "",
        "paragraphB_information_standalone": "",
        "wikitag_label_samepassage": "",
        "relevantInfo_comment_A": "",
        "relevantInfo_comment_B": "",
        "Contradict_type_I": "",
        "Contradict_type_II": "",
        "Contradict_type_III": "",
        "Contradict_type_IV": "",
        "taxonomy": [
            [
                ""
            ]
        ],
        "question1": "",
        "question1_answer1": "",
        "question1_answer2": "",
        "question2": "",
        "question2_answer1": "",
```

```
81              "question2_answer2": ""
82          }
83      }
```

The description of each key in the previous example is as follows:

- title: Title of article.

- url: URL of article.

- paragraph_A: Paragraph automatically retrieved (containing the tag).

- paragraph_A_clean: Paragraph automatically retrieved (removing the tag).

- tag: Type of tag of the article (Inconsistent/Self-contradictory/Contradict-other).

- tagDate: Date of the tag.

- tagReason: Reason for the tag.

- annotationResult: Results of the human data annotation.

    - wikitag_label_valid: Valid or invalid tag (Valid/Invalid).
    - valid_comment: Comment on the tag.
    - paragraphA_article: Title of article containing passage 1.
    - paragraphA_information: Relevant information of passage 1.
    - paragraphA_information_standalone: Decontextualized relevant information of passage 1.
    - paragraphB_article: Relevant information of passage 2.
    - paragraphB_information_standalone: Decontextualized relevant information of passage 2.
    - wikitag_label_samepassage: Boolean value stating whether passage 1 and passage 2 are the same (Same/Different).
    - relevantInfo_comment_A: Comment on the information of passage 1.
    - relevantInfo_comment_B: Comment on the information of passage 2.
    - Contradict type I: Contradiction type I focuses on the fine-grained semantics of the contradiction, e.g., date/time, location, language, etc.
    - Contradict type II: Contradiction type II focuses on the modality the contradiction. It describes the modality of passage 1 and passage 2, whether the information is from a piece of text, or from a row an infobox or a table.
    - Contradict type III: Contradiction type III focuses on the source the contradiction. It describes whether passage 1 and passage 2 are from the same article or not.
    - Contradict type IV: Contradiction type IV focuses on the reasoning aspect. It describes whether the contraction is explicit or implicit (Explicit/Implicit). Implicit contradiction requires some reasoning to understand why passage 1 and passage 2 are contradicted.
    - taxonomy: Array of key-values corresponding to contradict type I, contradict type II, contradict type III, and contradict type IV.
    - question1: Question 1 inferred from the contradiction.
    - question1_answer1: Gold answer to question 1 according to passage 1.
    - question1_answer2: Gold answer to question 1 according to passage 2.
    - question2: Question 2 inferred from the contradiction.
    - question2_answer1: Gold answer to question 2 according to passage 1.
    - question2_answer2: Gold answer to question 2 according to passage 2.

F Examples of `WikiContradict`

Here we show some examples of `WikiContradict` annotated through Label Studio, in Figure 1, 2, and 3. Note that data the Step 3 of the annotation (contradiction reason) is not included in the dataset.

G Datasheets for `WikiContradict`

In this appendix, we provide the dataset documentation and intended uses following the framework *Datasheet for Datasets* [Gebru et al., 2021].

G.1 Dataset Documentation and Intended Uses

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description. The dataset was created to enable research on assessing LLM performance when dealing with retrieved passages containing real-world knowledge conflicts. The dataset was created intentionally with that task in mind, focusing on a benchmark consisting of high-quality, human-annotated instances.

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., 15 company, institution, organization)? The dataset was created by Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri. All authors are employed by IBM Research.

Who funded the creation of the dataset? There was no associated grant.

Any other comments? N/A.

G.2 Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description. The instances are extracted passages from Wikipedia articles. The data format and examples of `WikiContradict` can be found in Appendix E and F, respectively.

How many instances are there in total (of each type, if appropriate)? There are 253 instances in total.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable). The dataset contains all possible instances.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description. Each instance consists of a question, a pair of contradictory passages extracted from Wikipedia, and two distinct answers, each derived from on the passages. The pair is annotated by a human annotator who identify where the conflicted information is and what type of conflict is observed. The annotator then produces a set of questions related to the passages with different answers reflecting the conflicting source of knowledge.

Is there a label or target associated with each instance? If so, please provide a description N/A.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text. Each annotation instance contains at least one question and two possible answers, but some instances may contain more than one question (and the corresponding two possible answers for each question). Some instances may not contain a value for `paragraphA_clean`, `tagDate`, and `tagReason` (see Appendix E).

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit. N/A.

Are there recommended data splits (e.g., training, development/validation,testing)? If so, please provide a description of these splits, explaining the rationale behind them. N/A.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description. Since our dataset requires manual annotation, annotation noise is inevitably introduced.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? The dataset is entirely self-contained.

Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?If so, please provide a description. No.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why. No.

Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset. N/A.

Is it possible to identify individuals (i.e., one or more natural per- sons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how. N/A.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description. No.

Any other comments? None.

G.3 Collection process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how. The data was mostly observable as raw text. The raw data was retrieved from Wikipedia articles containing inconsistent, self-contradictory, and contradict-other tags. The first two tags denote contradictory statements within the same article, whereas the third tag highlights instances where the content of one article contradicts that of another article. In total, we collected around 1,200 articles that contain these tags through the Wikipedia

maintenance category "Wikipedia articles with content issues". Given a content inconsistency tag provided by Wikipedia editors, the annotators verified whether the tag is valid by checking the relevant article content, the editor's comment, as well as the information in the edit history and the article's talk page if necessary.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated? The authors modified the code of an existing Python package called wikitextparser, which allows users easily extract and/or manipulate templates, template parameters, parser functions, tables, external links, wikilinks, lists, etc. found in wikitexts. The authors parsed the relevant Wikipedia articles into clean text, and modified the code to keep the inconsistent, self-contradictory, and contradict-other tags.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? N/A.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? All the authors of this paper (Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, Tigran Tchrakian, Radu Marinescu, Elizabeth Daly, Inkit Padhi, and Prasanna Sattigeri) were involved in the data collection process.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the time-frame in which the data associated with the instances was created. The dataset was collected between February 2024 and June 2024 from Wikipedia.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation. N/A

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? N/A.

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented. N/A.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate). N/A.

Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation. N/A.

Any other comments? None.

G.4 Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the

questions in this section. The annotators were required to slightly modify the original passages to make them stand-alone (decontextualization). Normally, this requires to resolve the coreference anaphors or the bridging anaphors in the first sentence (see annotation guidelines). In Wikipedia, oftentimes the antecedents for these anaphors are the article titles themselves.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data. Yes. The dataset itself contains all the raw passages.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point. We have used Python language to implement data cleaning. We will share the scripts details in our codebase.

Any other comments? None.

G.5   Uses

Has the dataset been used for any tasks already? If so, please provide a description. The dataset has been used in the paper to assess LLMs performance when augmented with retrieved passages containing real-world knowledge conflicts.

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. We will provide links to the repository after acceptance.

What (other) tasks could the dataset be used for? The dataset could be used for improving the performance of LLMs when presented with conflicting sources of information, by augmenting the prompt or fine-tuning the model.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms? There is minimal risk for harm: the data was already public on Wikipedia.

Are there tasks for which the dataset should not be used? If so, please provide a description. N/A.

G.6   Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?If so, please provide a description. Yes, the dataset and its metadata will be publicly available on the repository after acceptance.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)? The dataset and DOI will be published after acceptance.

When will the dataset be distributed? The dataset will be released after acceptance.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions. WikiContradict is distributed

under an MIT[2] license. Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABIL-ITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAM-AGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHER-WISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions. No.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation. No.

G.7 Maintenance

Who is supporting/hosting/maintaining the dataset? Yufang Hou, Alessandra Pascale, Javier Carnerero-Cano, and Tigran Tchrakian are supporting/maintaining the dataset.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)? If you wish to extend or contribute to our dataset, please contact us via email: Yufang Hou (yhou@ie.ibm.com), Alessandra Pascale (apascale@ie.ibm.com), Javier Carnerero-Cano (javier.cano@ibm.com), Tigran Tchrakian (tigran@ie.ibm.com), Radu Marinescu (radu.marinescu@ie.ibm.com), Elizabeth Daly (elizabeth.daly@ie.ibm.com), Inkit Padhi (inkpad@ibm.com), and Prasanna Sattigeri (psattig@us.ibm.com).

Is there an erratum? If so, please provide a link or other access point. Any updates to the dataset will be shared via GitHub.

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g.,mailing list,GitHub)? If we find inconsistencies in the dataset or extend the dataset, we will release the new version on the website and GitHub.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? N/A.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users. All versions of WikiContradict will be continue to be supported and maintained on website. We will post the updates on the website and GitHub.

---

[2]https://www.mit.edu/~amini/LICENSE.md

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description. Yes. Please contact the authors of this paper for building upon this dataset.

G.8 Responsibility

The authors bear all responsibility in case of violation of rights, etc. We confirm that the dataset is licensed under MIT license.

H Explicit License

`WikiContradict` is distributed under an MIT[3] license. Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABIL-ITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS BE LIABLE FOR ANY CLAIM, DAM-AGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHER-WISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

I Ethics Statement

The authors bear all responsibility in the event of any violation of rights, the dataset will be released after acceptance under an MIT licence.

Biases Our data is downloaded from Wikipedia. As such, the data is biased towards the original content and sources. Given that human data annotation involves some degree of subjectivity we created a comprehensive 17-page annotation guidelines document to clarify important cases during the annotation process. The annotators were explicitly instructed not to take their personal feeling about the particular topic. Nevertheless, some degree of intrinsic subjectivity might have impacted the techniques picked up by the annotators during the annotation.

References

T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford. Datasheets for datasets. Communications of the ACM, 64(12):86–92, 2021.

---

[3]https://www.mit.edu/~amini/LICENSE.md

Wikipedia article: Mediterranean seas

Inconsistence or contradictory tag

*The Baltic Sea is a brackish inland sea, alleged to be the largest body of brackish water in the world (other possibilities include the Black Sea).{{Inconsistent|date=October 2019|reason=Baltic Sea is listed as a mediterranean sea in a section above. The Exceptions section seems to be meant to list seas that meet most criteria of mediterranean sea and yet are not considered so. Why is Baltic Sea listed in both exception and non-exception lists?}} It occupies a basin formed by glacial erosion.

* Step1: Is this tag valid?

☑ Valid[1]   ☐ Invalid[2]

Additional comment

Baltic Sea is listed as a mediterranean sea in a section above. The Exceptions section seems to be meant to list seas that meet most criteria of mediterranean sea and yet are not considered so. Why is Baltic Sea listed in both exception and non-exception lists?

* Step 2: Copy the inconsistent sentences, paragraphs, or information (e.g., table or infobox rows) from the wikipedia article(s) into the following boxes. Modify these copy-paste messages to ensure they form a coherent and valid stand-alone text passage, if necessary.

* Article A title

Mediterranean seas

* Article A relevant information (Passage 1)

List of mediterranean seas The mediterranean seas of the Atlantic Ocean • The Baltic Sea Types of mediterranean seas Dilution basin • A dilution basin has a lower salinity due to freshwater gains such as rainfall and rivers, and its water exchange consists of outflow of the fresher mediterranean water in the upper layer and inflow of the saltier oceanic water in the lower layer of the channel. Renewal of deep water may not be sufficient to supply oxygen to the bottom. o The Baltic Sea

[Modify] Article A relevant information_stand-alone (Passage 1)

The Baltic Sea is a mediterranean sea of the Atlantic Ocean. An example of a mediterranean sea which is a dilution basin is the Baltic Sea. A dilution basin has a lower salinity due to freshwater gains such as rainfall and rivers, and its water exchange consists of outflow of the fresher mediterranean water in the upper layer and inflow of the saltier oceanic water in the lower layer of the channel. Renewal of deep water may not be sufficient to supply oxygen to the bottom.

* Article B title

Mediterranean seas

* Article B relevant information (Passage 2)

Exceptions The Baltic Sea is a brackish inland sea, alleged to be the largest body of brackish water in the world (other possibilities include the Black Sea). It occupies a basin formed by glacial erosion.

[Modify] Article B relevant information_stand-alone (Passage 2)

The Baltic Sea is not a mediterranean sea because is a brackish inland sea, alleged to be the largest body of brackish water in the world (other possibilities include the Black Sea). It occupies a basin formed by glacial erosion.

* Are Passage 1 and Passage 2 the same?

☐ Same[3]   ☑ Different[4]

* Step 4: Choose all options that can describe the above contradictory information

Text - Text ×   Within the same article ×   Explicit ×   (DiscourseLevel) NP-related × ∨

Step 5: Write at least one question that highlights the contradictions between Passage 1 and Passage 2, eliciting different responses based on each passage.

* Question 1

Is the Baltic Sea considered a type of mediterranean sea?

* Answers for Question 1 (add the answers based on Passage 1 and Passage 2 seperately)

Yes

No

Figure 1: Example of annotation of Mediterranean seas.

Wikipedia article: Mitta Mitta River

Inconsistence or contradictory tag

The river flows through a magnificent valley that contains four small towns: Mitta Mitta, Eskdale, Dartmouth. {{Inconsistent|reason=text says 'four' but lists only 3}} Mitta Mitta is a small hamlet at the confluence of the River and Snowy Creek.

* Step1: Is this tag valid?

☑ Valid[1]   ☐ Invalid[2]

Additional comment

text says 'four' but lists only 3 ✎   🗑

* Step 2: Copy the inconsistent sentences, paragraphs, or information (e.g., table or infobox rows) from the wikipedia article(s) into the following boxes. Modify these copy-paste messages to ensure they form a coherent and valid stand-alone text passage, if necessary.

* Article A title

Mitta Mitta River ✎   🗑

* Article A relevant information (Passage 1)

The river flows through a magnificent valley that contains four small towns: Mitta Mitta, Eskdale, Dartmouth. ✎   🗑

[Modify] Article A relevant information_stand-alone (Passage 1)

Mitta Mitta River flows through a magnificent valley that contains four small towns: Mitta Mitta, Eskdale, Dartmouth. ✎   🗑

* Article B title

Mitta Mitta River ✎   🗑

* Article B relevant information (Passage 2)

The river flows through a magnificent valley that contains four small towns: Mitta Mitta, Eskdale, Dartmouth. ✎   🗑

[Modify] Article B relevant information_stand-alone (Passage 2)

Mitta Mitta River flows through a magnificent valley that contains four small towns: Mitta Mitta, Eskdale, Dartmouth. ✎   🗑

* Are Passage 1 and Passage 2 the same?

☑ Same[3]   ☐ Different[4]

* Step 4: Choose all options that can describe the above contradictory information

(PhraseLevel) Entity - Number ✕   Text - Text ✕   Within the same article ✕   Explicit ✕   ⌄

Step 5: Write at least one question that highlights the contradictions between Passage 1 and Passage 2, eliciting different responses based on each passage.

* Question 1

How many small towns does the valley in which the Mitta Mitta river flows contain? ✎   🗑

* Answers for Question 1 (add the answers based on Passage 1 and Passage 2 seperately)

four ✎   🗑

three ✎   🗑

Question 2

Does the valley in which the Mitta Mitta river flows contain four small towns? ✎   🗑

Answers for Question 2

Yes ✎   🗑

No ✎   🗑

Figure 2: Example of annotation of Mitta Mitta River.

Wikipedia article: Wolfoo

Inconsistence or contradictory tag

* Scout Skunk, a 34-year-old skunk who is a badger{{Inconsistent|date=June 2023|reason=When is a skunk a badger?}}

---

* Step1: Is this tag valid?

☑ Valid[1]　☐ Invalid[2]

Additional comment

When is a skunk a badger? ✎　🗑

---

* Step 2: Copy the inconsistent sentences, paragraphs, or information (e.g., table or infobox rows) from the wikipedia article(s) into the following boxes. Modify these copy-paste messages to ensure they form a coherent and valid stand-alone text passage, if necessary.

* Article A title

Wolfoo ✎　🗑

* Article A relevant information (Passage 1)

Scout Skunk, a 34-year-old skunk who is a badger ✎　🗑

[Modify] Article A relevant information_stand-alone (Passage 1)

Scout Skunk, a 34-year-old skunk who is a badger ✎　🗑

---

* Article B title

Wolfoo ✎　🗑

* Article B relevant information (Passage 2)

Scout Skunk, a 34-year-old skunk who is a badger ✎　🗑

[Modify] Article B relevant information_stand-alone (Passage 2)

Scout Skunk, a 34-year-old skunk who is a badger ✎　🗑

* Are Passage 1 and Passage 2 the same?

☑ Same[3]　☐ Different[4]

---

* Step 4: Choose all options that can describe the above contradictory information

(PhraseLevel) NP-related (non-entity) ✕　Text - Text ✕　Within the same article ✕
Implicit (reasoning required) ✕

Step 5: Write at least one question that highlights the contradictions between Passage 1 and Passage 2, eliciting different responses based on each passage.

* Question 1

What kind of animal is Scout Skunk? ✎　🗑

* Answers for Question 1 (add the answers based on Passage 1 and Passage 2 seperately)

skunk ✎　🗑

badger ✎　🗑

Question 2

Is Scout Skunk a skunk? ✎　🗑

Answers for Question 2

Yes ✎　🗑

No ✎　🗑

Figure 3: Example of annotation of Wolfoo.