XATU: A Fine-grained Instruction-based Benchmark for Explainable Text Updates

Haopeng Zhang†, Hayate Iso‡, Sairam Gurajada‡, Nikita Bhutani‡

IFM Lab, UC Davis† Megagon Labs‡

https://github.com/megagonlabs/xatu

Outline

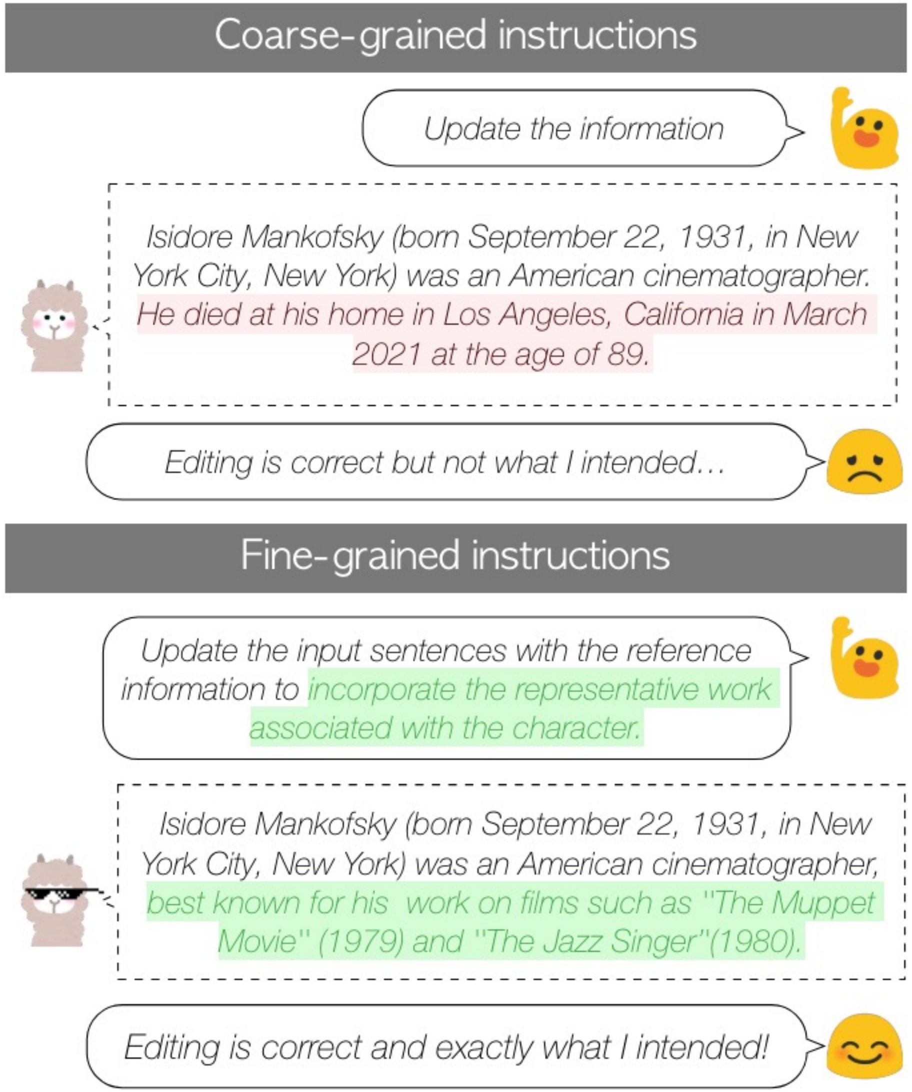
- Introduction
- XATU Benchmark
  - Data Source
  - Annotation Process
  - Benchmark Usage
- Experiment
  - Setup
  - Zero-shot Results
  - Fine-tuning Results
  - Discussion
- Conclusion

Introduction

- Text editing: modify text to better align with user intents

- Recently formulated as an iterative process

- Coarse-instruction vs. Fine-grained instruction



Coarse-grained instructions

Update the information

Isidore Mankofsky (born September 22, 1931, in New York City, New York) was an American cinematographer. He died at his home in Los Angeles, California in March 2021 at the age of 89.

Editing is correct but not what I intended...

Fine-grained instructions

Update the input sentences with the reference information to incorporate the representative work associated with the character.

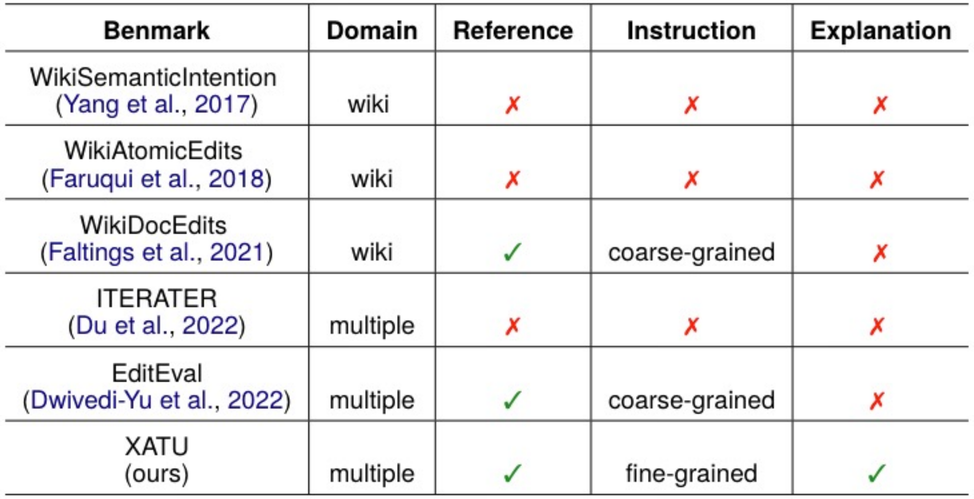
Isidore Mankofsky (born September 22, 1931, in New York City, New York) was an American cinematographer, best known for his work on films such as "The Muppet Movie" (1979) and "The Jazz Singer"(1980).

Editing is correct and exactly what I intended!

Introduction

- Coarse instruction: may not accurately reflect the true editing capabilities of LLM systems

- Lack of explainability



Benmark	Domain	Reference	Instruction	Explanation
WikiSemanticIntention (Yang et al., 2017)	wiki	✗	✗	✗
WikiAtomicEdits (Faruqui et al., 2018)	wiki	✗	✗	✗
WikiDocEdits (Faltings et al., 2021)	wiki	✓	coarse-grained	✗
ITERATER (Du et al., 2022)	multiple	✗	✗	✗
EditEval (Dwivedi-Yu et al., 2022)	multiple	✓	coarse-grained	✗
XATU (ours)	multiple	✓	fine-grained	✓

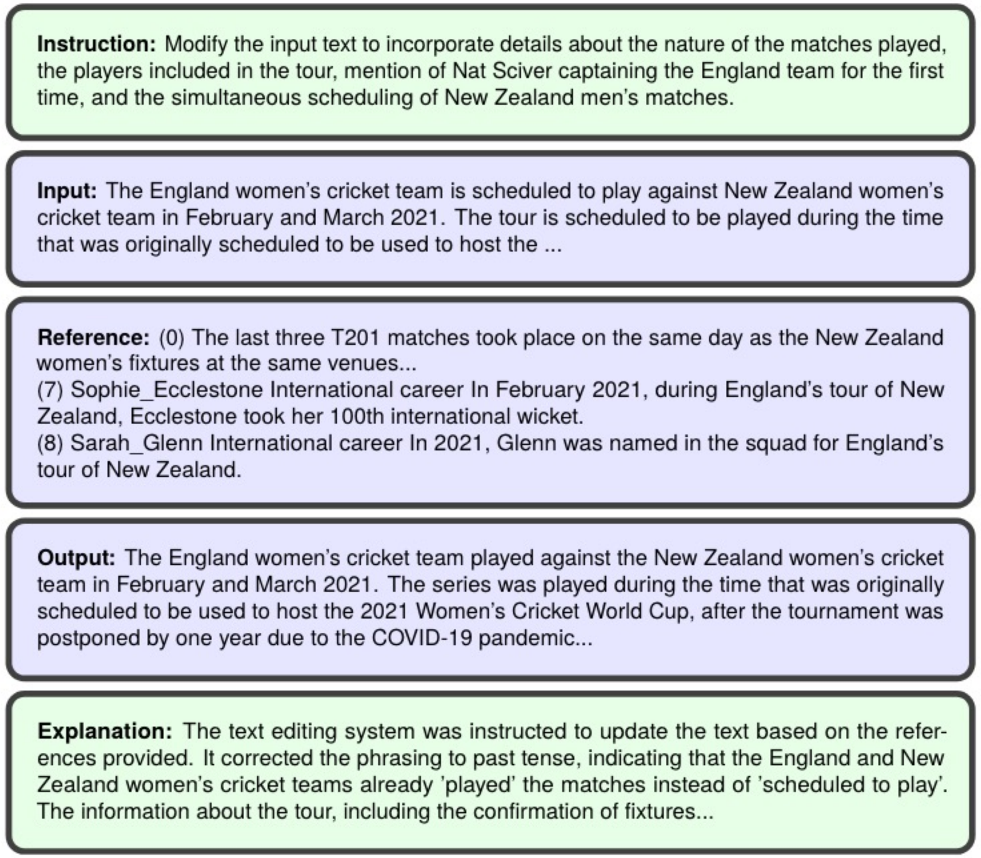
Table 1: A detailed comparison of XATU with existing text editing datasets and benchmarks.

XATU Benchmark

- XATU: explainable text updates

- first text editing dataset with **fine-grained** instructions and **explanations**.

- Instance format on the right



Instruction: Modify the input text to incorporate details about the nature of the matches played, the players included in the tour, mention of Nat Sciver captaining the England team for the first time, and the simultaneous scheduling of New Zealand men's matches.

Input: The England women's cricket team is scheduled to play against New Zealand women's cricket team in February and March 2021. The tour is scheduled to be played during the time that was originally scheduled to be used to host the ...

Reference: (0) The last three T201 matches took place on the same day as the New Zealand women's fixtures at the same venues...
(7) Sophie_Ecclestone International career In February 2021, during England's tour of New Zealand, Ecclestone took her 100th international wicket.
(8) Sarah_Glenn International career In 2021, Glenn was named in the squad for England's tour of New Zealand.

Output: The England women's cricket team played against the New Zealand women's cricket team in February and March 2021. The series was played during the time that was originally scheduled to be used to host the 2021 Women's Cricket World Cup, after the tournament was postponed by one year due to the COVID-19 pandemic...

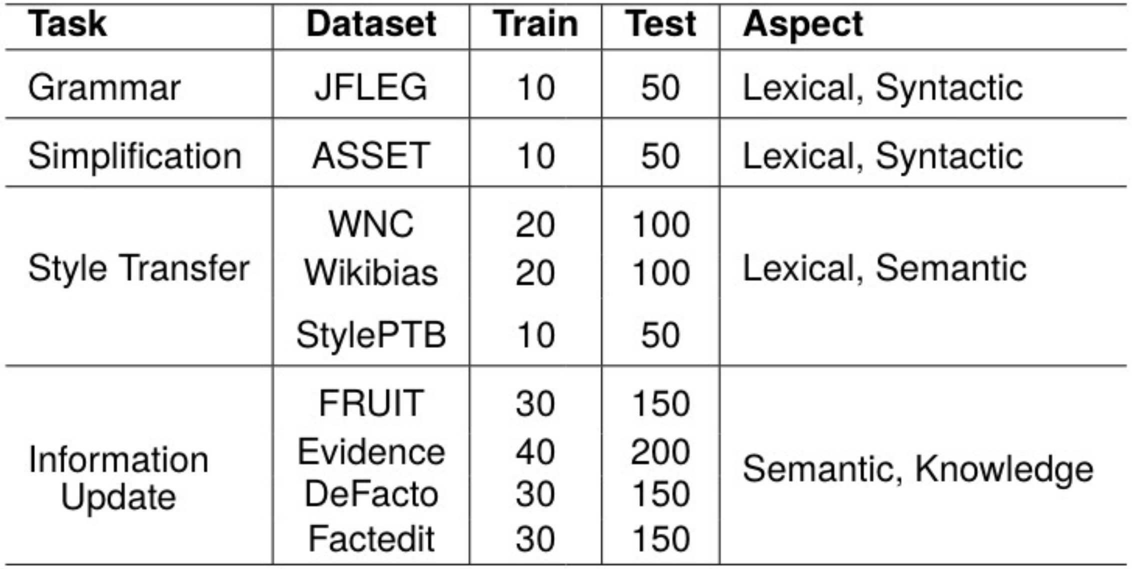
Explanation: The text editing system was instructed to update the text based on the references provided. It corrected the phrasing to past tense, indicating that the England and New Zealand women's cricket teams already 'played' the matches instead of 'scheduled to play'. The information about the tour, including the confirmation of fixtures...

Figure 2: The instance format of the data in XATU benchmark. Data in blue (Input, Output, Reference) are extracted from the corresponding data sources, and data in green (Fine-grained instruction and explanation) are obtained from joint automatic and human annotations.

Data Source

- text editing categorization:
  - Lexical
  - Syntactic
  - Semantic
  - Knowledge

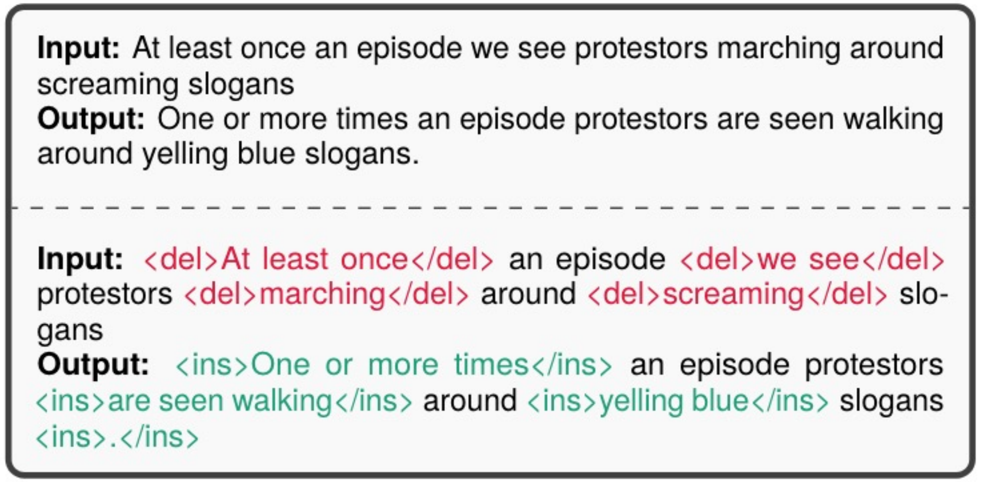
- XATU: 1000 annotated data from 9 data sources, 4 tasks



Task	Dataset	Train	Test	Aspect
Grammar	JFLEG	10	50	Lexical, Syntactic
Simplification	ASSET	10	50	Lexical, Syntactic
Style Transfer	WNC	20	100	Lexical, Semantic
	Wikibias	20	100	
	StylePTB	10	50	
Information Update	FRUIT	30	150	Semantic, Knowledge
	Evidence	40	200	
	DeFacto	30	150	
	Factedit	30	150	

Annotation Process

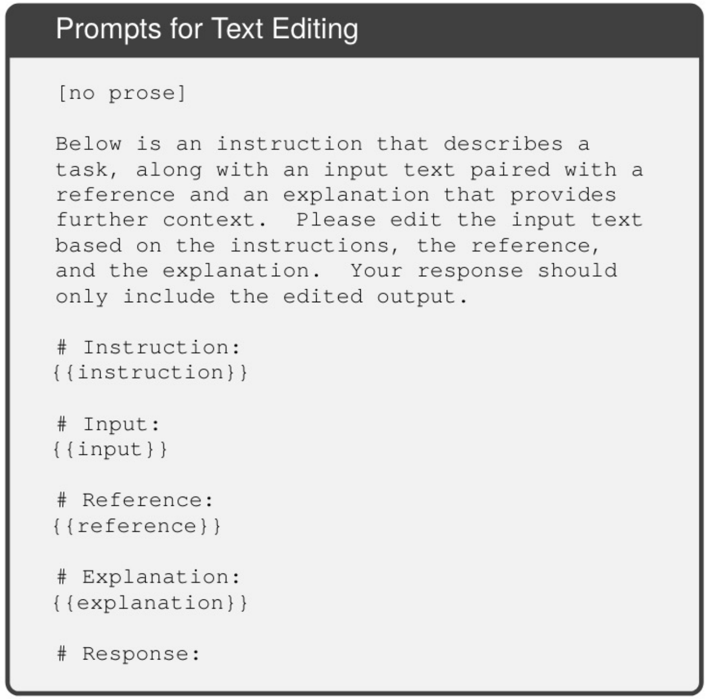
- LLM-in-the-loop annotation approach
  - Generate candidates of fine-grained instructions and explanations using LLMs
    - adding HTML tags
  - Candidate Validation by Human
    - human evaluation through the Appen platform
    - random test examples so select annotators
    - inter-annotator agreement of 80.27%



Input: At least once an episode we see protestors marching around screaming slogans
Output: One or more times an episode protestors are seen walking around yelling blue slogans.

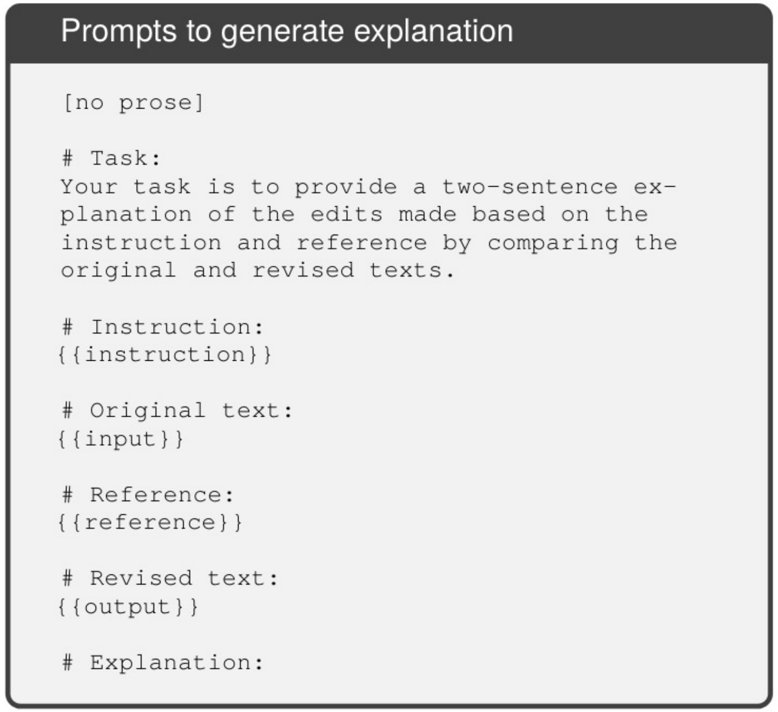
Input: <del>At least once</del> an episode <del>we see</del> protestors <del>marching</del> around <del>screaming</del> slogans
Output: <ins>One or more times</ins> an episode protestors <ins>are seen walking</ins> around <ins>yelling blue</ins> slogans <ins>.</ins>

Annotation Process



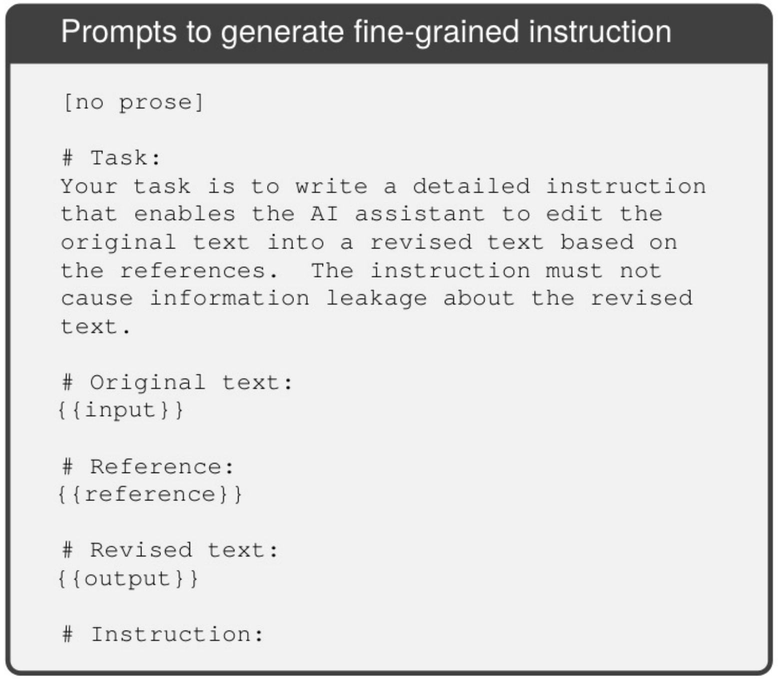
Prompts for Text Editing

```
[no prose]

Below is an instruction that describes a
task, along with an input text paired with a
reference and an explanation that provides
further context.  Please edit the input text
based on the instructions, the reference,
and the explanation.  Your response should
only include the edited output.

# Instruction:
{{instruction}}

# Input:
{{input}}

# Reference:
{{reference}}

# Explanation:
{{explanation}}

# Response:
```



Prompts to generate explanation

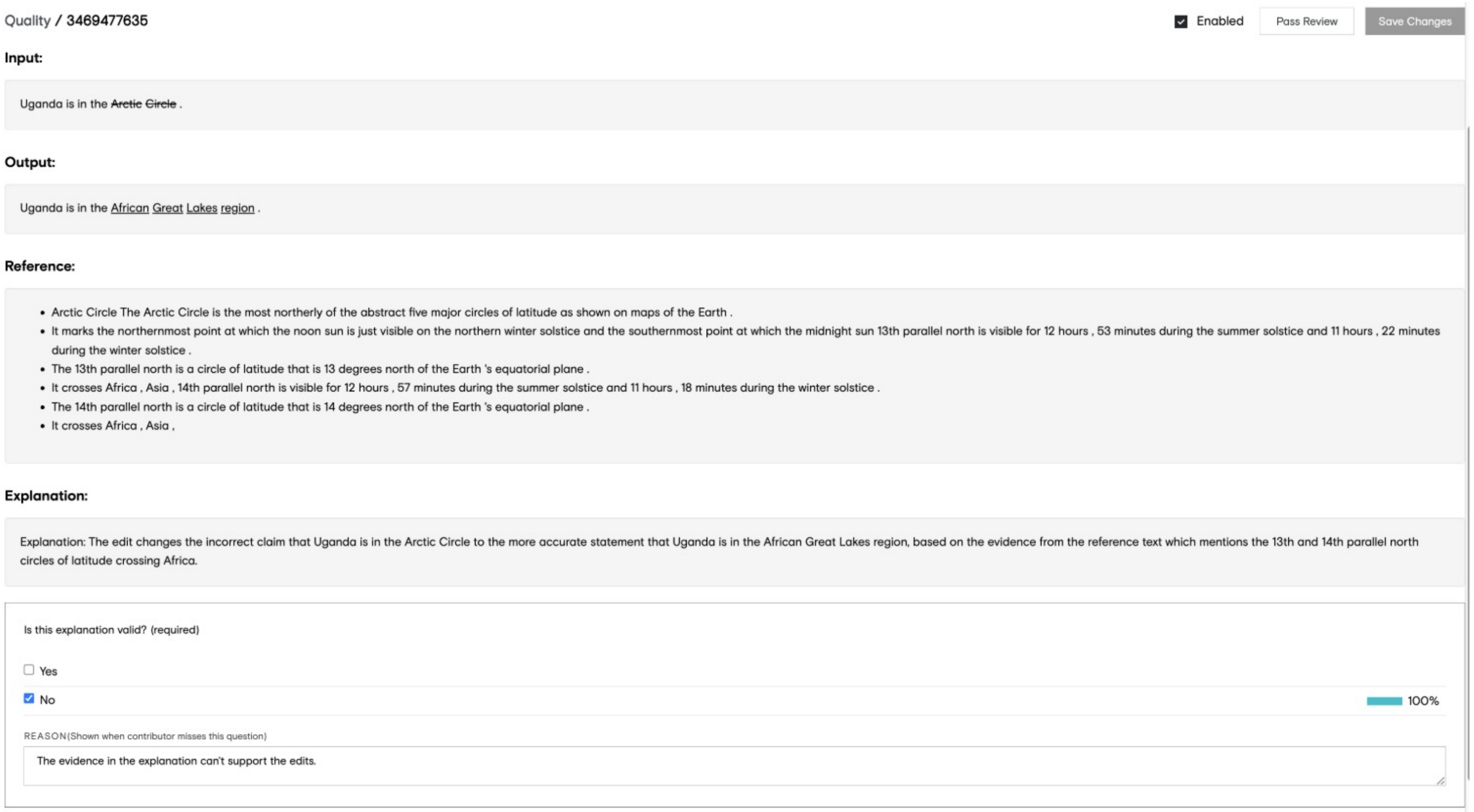
```
[no prose]

# Task:
Your task is to provide a two-sentence explanation of the edits made based on the
instruction and reference by comparing the
original and revised texts.

# Instruction:
{{instruction}}

# Original text:
{{input}}

# Reference:
{{reference}}

# Revised text:
{{output}}

# Explanation:
```



Prompts to generate fine-grained instruction

```
[no prose]

# Task:
Your task is to write a detailed instruction
that enables the AI assistant to edit the
original text into a revised text based on
the references.  The instruction must not
cause information leakage about the revised
text.

# Original text:
{{input}}

# Reference:
{{reference}}

# Revised text:
{{output}}

# Instruction:
```

Annotation Process



Input:

Uganda is in the ~~Arctic Circle~~ .

Output:

Uganda is in the African Great Lakes region .

Reference:

- Arctic Circle The Arctic Circle is the most northerly of the abstract five major circles of latitude as shown on maps of the Earth .
- It marks the northernmost point at which the noon sun is just visible on the northern winter solstice and the southernmost point at which the midnight sun 13th parallel north is visible for 12 hours , 53 minutes during the summer solstice and 11 hours , 22 minutes during the winter solstice .
- The 13th parallel north is a circle of latitude that is 13 degrees north of the Earth 's equatorial plane .
- It crosses Africa , Asia , 14th parallel north is visible for 12 hours , 57 minutes during the summer solstice and 11 hours , 18 minutes during the winter solstice .
- The 14th parallel north is a circle of latitude that is 14 degrees north of the Earth 's equatorial plane .
- It crosses Africa , Asia ,

Explanation:

Explanation: The edit changes the incorrect claim that Uganda is in the Arctic Circle to the more accurate statement that Uganda is in the African Great Lakes region, based on the evidence from the reference text which mentions the 13th and 14th parallel north circles of latitude crossing Africa.

Is this explanation valid? (required)

☐ Yes

☑ No 100%

REASON(Shown when contributor misses this question)

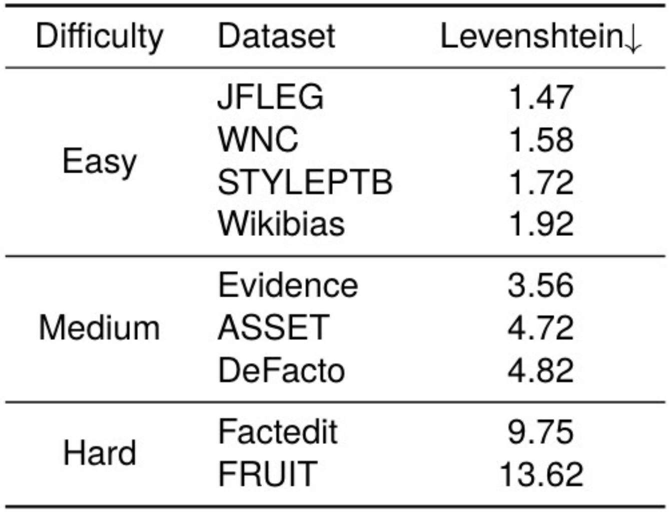
The evidence in the explanation can't support the edits.

Figure 6: The interface used for annotation.

Benchmark Usage

- Difficulty level: Levenshtein distance

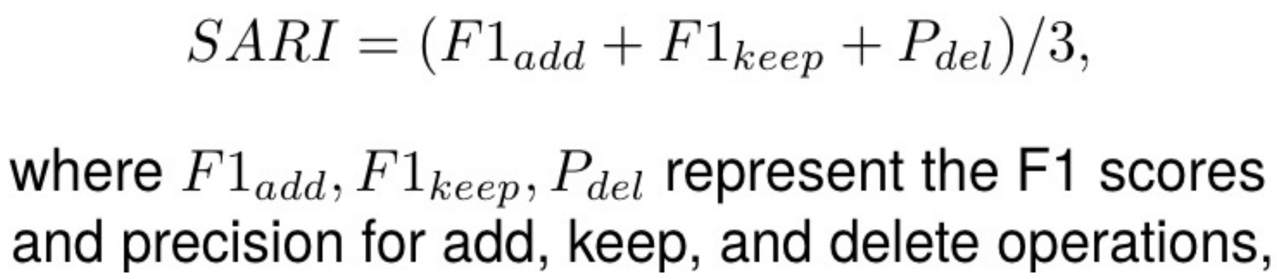
- Downstream Tasks:
  - Edit representation modeling
  - Automatic editing instruction generation
  - Editing span prediction
  - Explanation generation
  - Evidence retrieval



Difficulty	Dataset	Levenshtein↓
Easy	JFLEG	1.47
	WNC	1.58
	STYLEPTB	1.72
	Wikibias	1.92
Medium	Evidence	3.56
	ASSET	4.72
	DeFacto	4.82
Hard	Factedit	9.75
	FRUIT	13.62

Experimental Setup

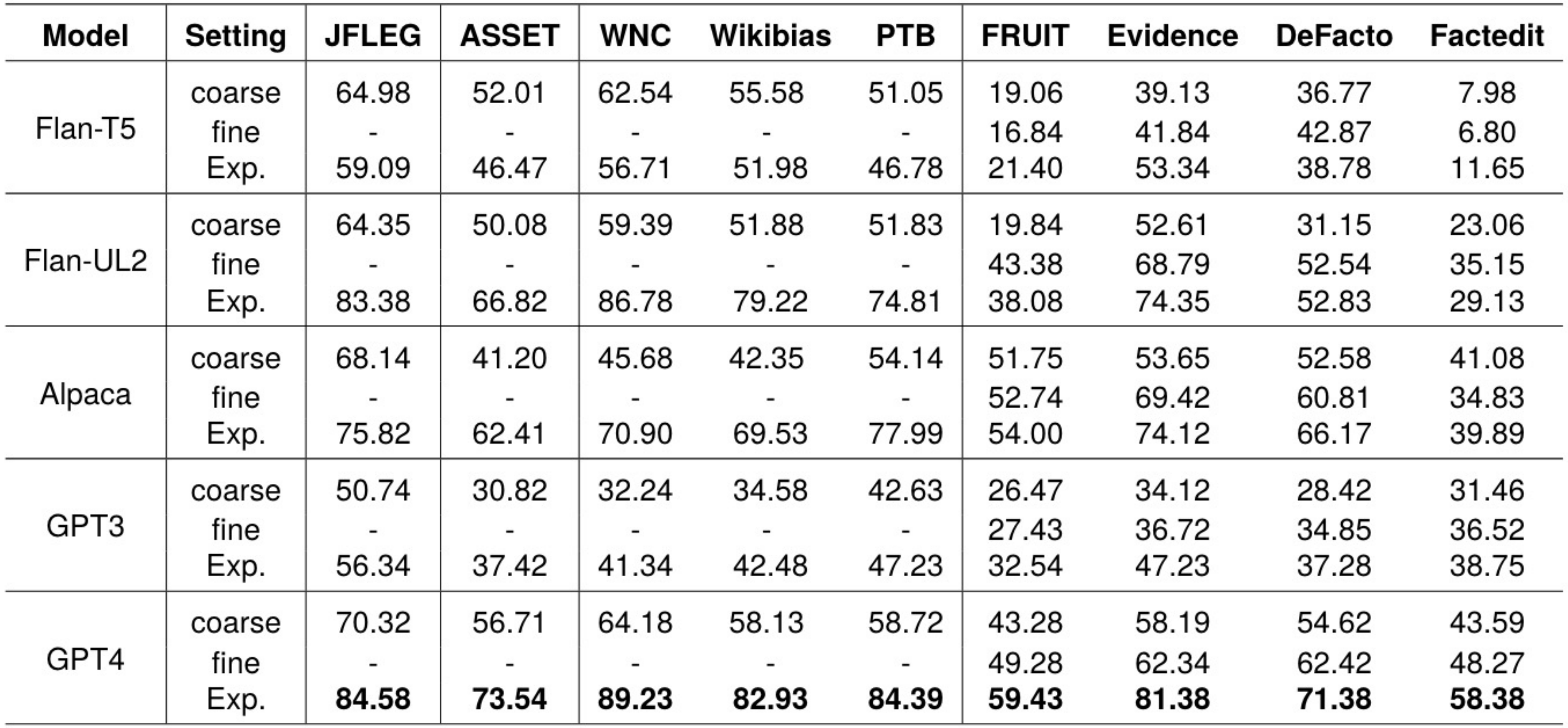
- Baseline LLMs: GPT-3, GPT-4, T5, FLAN-T5, UL2, FLAN-UL2, LLaMa, Alpaca
- Evaluation: SARI scores (n-gram based metric)



$$SARI = (F1_{add} + F1_{keep} + P_{del})/3,$$

where $F1_{add}, F1_{keep}, P_{del}$ represent the F1 scores and precision for add, keep, and delete operations,

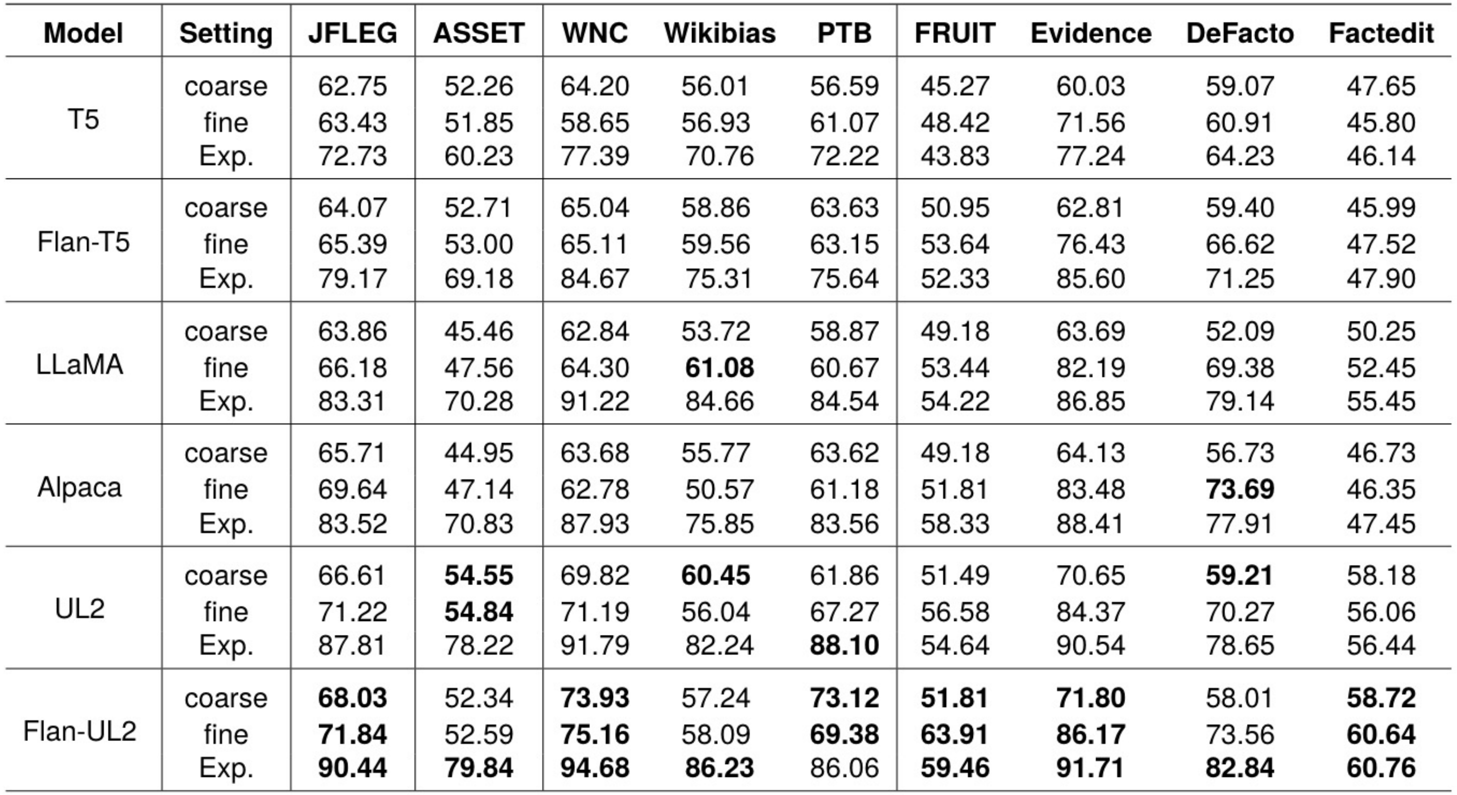
Zero-shot Results



Model	Setting	JFLEG	ASSET	WNC	Wikibias	PTB	FRUIT	Evidence	DeFacto	Factedit
Flan-T5	coarse	64.98	52.01	62.54	55.58	51.05	19.06	39.13	36.77	7.98
	fine	-	-	-	-	-	16.84	41.84	42.87	6.80
	Exp.	59.09	46.47	56.71	51.98	46.78	21.40	53.34	38.78	11.65
Flan-UL2	coarse	64.35	50.08	59.39	51.88	51.83	19.84	52.61	31.15	23.06
	fine	-	-	-	-	-	43.38	68.79	52.54	35.15
	Exp.	83.38	66.82	86.78	79.22	74.81	38.08	74.35	52.83	29.13
Alpaca	coarse	68.14	41.20	45.68	42.35	54.14	51.75	53.65	52.58	41.08
	fine	-	-	-	-	-	52.74	69.42	60.81	34.83
	Exp.	75.82	62.41	70.90	69.53	77.99	54.00	74.12	66.17	39.89
GPT3	coarse	50.74	30.82	32.24	34.58	42.63	26.47	34.12	28.42	31.46
	fine	-	-	-	-	-	27.43	36.72	34.85	36.52
	Exp.	56.34	37.42	41.34	42.48	47.23	32.54	47.23	37.28	38.75
GPT4	coarse	70.32	56.71	64.18	58.13	58.72	43.28	58.19	54.62	43.59
	fine	-	-	-	-	-	49.28	62.34	62.42	48.27
	Exp.	**84.58**	**73.54**	**89.23**	**82.93**	**84.39**	**59.43**	**81.38**	**71.38**	**58.38**

- GPT-4 demonstrates exceptional zero-shot editing performance
- Almost all models exhibit improvements when guided by the fine-grained instructions
- The underlying architecture (encoder-decoder vs. decoder-only) of language models significantly impacts the performance of different types of text editing tasks

Fine-tuning Results



Model	Setting	JFLEG	ASSET	WNC	Wikibias	PTB	FRUIT	Evidence	DeFacto	Factedit
T5	coarse	62.75	52.26	64.20	56.01	56.59	45.27	60.03	59.07	47.65
	fine	63.43	51.85	58.65	56.93	61.07	48.42	71.56	60.91	45.80
	Exp.	72.73	60.23	77.39	70.76	72.22	43.83	77.24	64.23	46.14
Flan-T5	coarse	64.07	52.71	65.04	58.86	63.63	50.95	62.81	59.40	45.99
	fine	65.39	53.00	65.11	59.56	63.15	53.64	76.43	66.62	47.52
	Exp.	79.17	69.18	84.67	75.31	75.64	52.33	85.60	71.25	47.90
LLaMA	coarse	63.86	45.46	62.84	53.72	58.87	49.18	63.69	52.09	50.25
	fine	66.18	47.56	64.30	**61.08**	60.67	53.44	82.19	69.38	52.45
	Exp.	83.31	70.28	91.22	84.66	84.54	54.22	86.85	79.14	55.45
Alpaca	coarse	65.71	44.95	63.68	55.77	63.62	49.18	64.13	56.73	46.73
	fine	69.64	47.14	62.78	50.57	61.18	51.81	83.48	**73.69**	46.35
	Exp.	83.52	70.83	87.93	75.85	83.56	58.33	88.41	77.91	47.45
UL2	coarse	66.61	**54.55**	69.82	**60.45**	61.86	51.49	70.65	**59.21**	58.18
	fine	71.22	**54.84**	71.19	56.04	67.27	56.58	84.37	70.27	56.06
	Exp.	87.81	78.22	91.79	82.24	**88.10**	54.64	90.54	78.65	56.44
Flan-UL2	coarse	**68.03**	52.34	**73.93**	57.24	**73.12**	**51.81**	**71.80**	58.01	**58.72**
	fine	**71.84**	52.59	**75.16**	58.09	**69.38**	**63.91**	**86.17**	73.56	**60.64**
	Exp.	**90.44**	**79.84**	**94.68**	**86.23**	86.06	**59.46**	**91.71**	**82.84**	**60.76**

- Few-shot fine-tuning is effective, even with a limited number of examples for text editing tasks (200 in this case)
- The instruction-tuned versions of LLM consistently outperform their base models. (Flan-T5 and Flan-UL2)

Discussion

- Fine-grained instruction models consistently outperform their coarse-grained counterparts.
- Alpaca demonstrates superior robustness

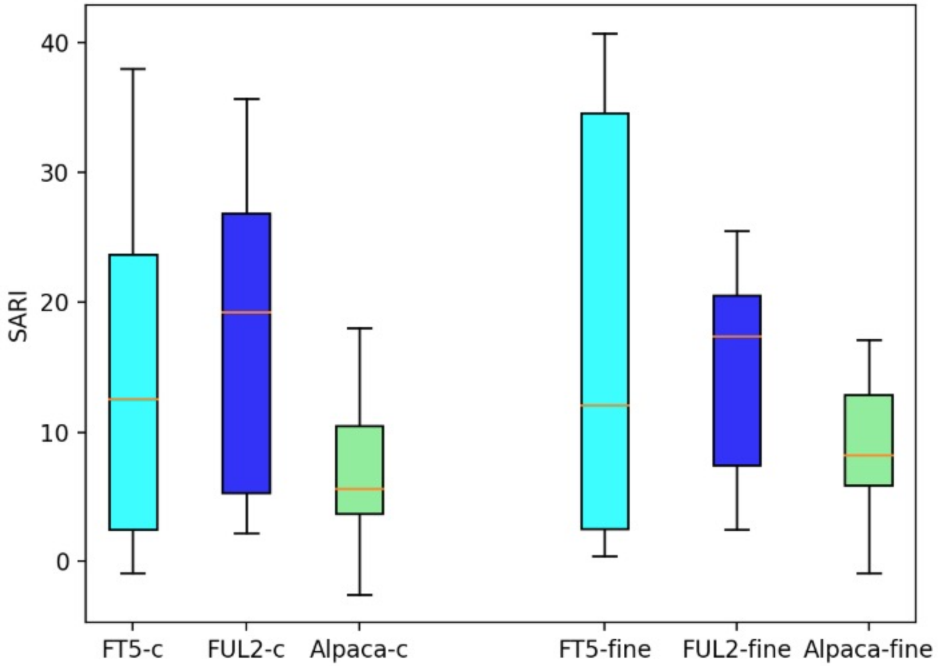


Figure 4: Fine-tuning with fine-grained instructions (-fine) vs. coarse instructions (-c).

Discussion

- Effectiveness of instruction tuning in improving the performance of language models across different settings

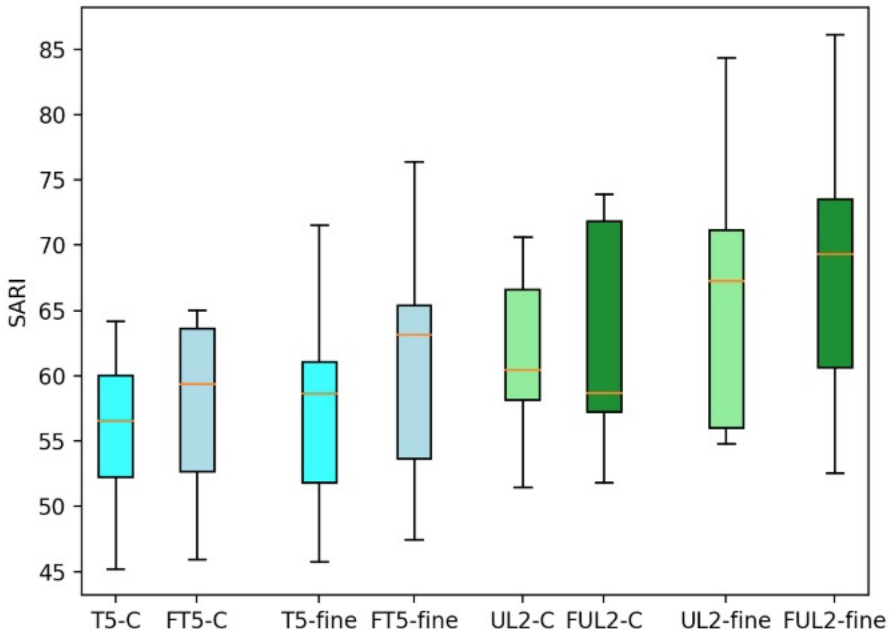


Figure 5: Boxplot comparing instruction-tuned LLMs (Flan-xx) vs. pre-trained counterparts with fine-grained (-fine) and coarse instructions (-c).

Conclusion

- This paper introduces XATU, the first benchmark for explainable text updates with fine-grained instructions.

- XATU is a diverse benchmark covering a wide range of topics and text types and leverages high-quality data sources from various existing sources.

- We compare existing open and closed instruction-tuned language models under both the zero-shot and fine-tuning settings and reveal their capabilities to edit text and follow instructions