Appendix

A. *Preparing annotation tasks for crowdsourcing*

The additional details for each context ensured that the pre-processed dataset included only citation contexts clearly linked to a specific cited paper, minimizing ambiguity for crowdsourcing workers. In cases where citation marks were not uniquely identifiable, we appended last name of the first author of the cited paper to each labeling item. Each entity to be labeled was composed of four parts:

- **Citation context**: The sentence(s) which contains the citation mark of the cited paper.
- **Related rep-study**: The reproducibility study corresponding to the citation context. We added this to identify the related rep-study to which the citation context belongs even after crowdsourced annotations process.
- **First author of the cited paper**: To resolve situations where multiple citation marks appear in the same context.
- **Context index**: An id to identify the specific citation context.

We organized the citation context entities in the data files according to the following format (see Fig. 3 for an example):

```
[Citation context] [Related rep-study] [First
author of the cited paper] [Context index]
```

The inclusion of the reproducibility study and context index made post-processing and analysis easier after crowdsourcing.

B. *Interface Design*

We designed a custom template for the labeling task GUI using *Mechanical Turk Crowd HTML Elements*[1], which abstracts HTML markup, CSS, and JavaScript functionality into HTML tags. This allowed for an efficient and visually consistent interface. As shown in Fig. 3, the GUI included a section containing detailed instructions for workers, along with examples for each label category. We provided definitions for key scientific terms, e.g., *citation*, *citation context*, and *cited paper*. The GUI iteratively displayed each citation context each with a label selection area on the right. Workers were required to choose a label from three options (*Positive*, *Negative*, *Neutral*) and click the *Submit* button to proceed to the next context. The instructions, examples, and definitions were consistently presented for every task to reinforce understanding and minimize errors. It took approximately 27 days to complete the labeling. We used *VS Code* as the Integrated Development Environment (IDE) to develop the template, and the *Mechanical Turk Developer Sandbox*[2] (a simulated environment for testing applications and tasks prior to publication) to test the GUI's functionality before deployment.

C. *Technical Validation*

The confusion matrix in Fig. 1 highlights the alignment and discrepancies between the crowdsourced annotations and our manually annotated labels for the randomly selected 244 citation contexts verification set.

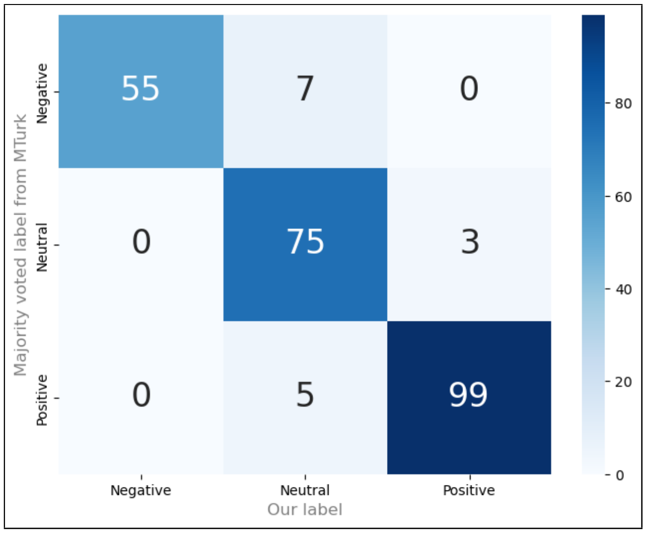


Fig. 1. Confusion matrix comparing MTurk annotator labels with ground truth labels in the verification set.

D. *LLM Fine-Tuning and RAG Configurations for CC30k*

To showcase the utility of the CC30k dataset, we experimented with multiple large language models using Low-Rank Adaptation (LoRA)–based parameter-efficient fine-tuning and retrieval-augmented generation (RAG). The corresponding hyper-parameters and configurations are listed in Tables I and II.

TABLE I
LoRA Fine-Tuning Hyperparameters for LLaMA 3–8B and Qwen1.5–7B

Setting	**LLaMA 3–8B**	**Qwen1.5–7B**
LoRA r	16	16
LoRA α	32	32
Dropout	0.05	0.05
Batch size	4	4
Grad. accumulation	4	4
Epochs	15	15
Learning rate	2×10^{-4}	2×10^{-4}
Max sequence length	1024	1024
Precision	`fp16`	`fp16`
Log steps	10	10
Save strategy	epoch	epoch
Save limit	2	2

TABLE II
GPT-4o LoRA Fine-Tuning and RAG Configuration

Setting	**Value**
Retrieval method	FAISS
Embedding model	SentenceTransformer
Top-K	5
Inference temperature	0
Max tokens	10

[1] https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMturkAPI/ApiReference_HTMLCustomElementsArticle.html

[2] https://docs.aws.amazon.com/AWSMechTurk/latest/AWSMechanicalTurkRequester/mturk-use-sandbox.html

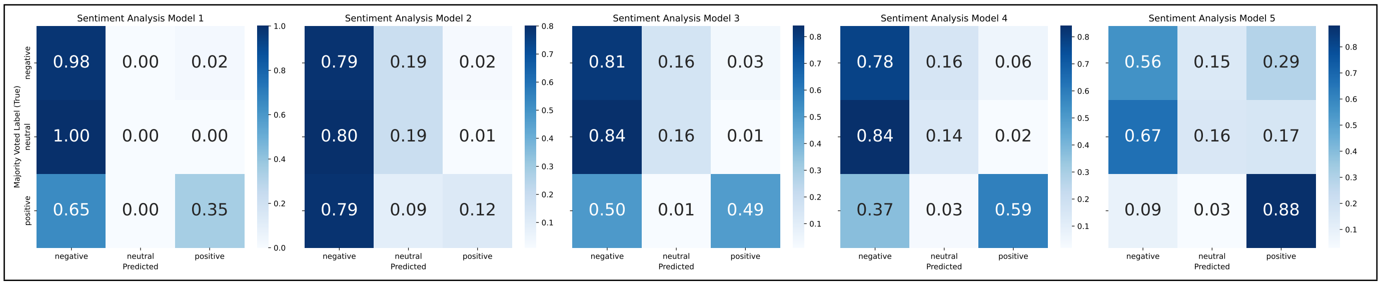
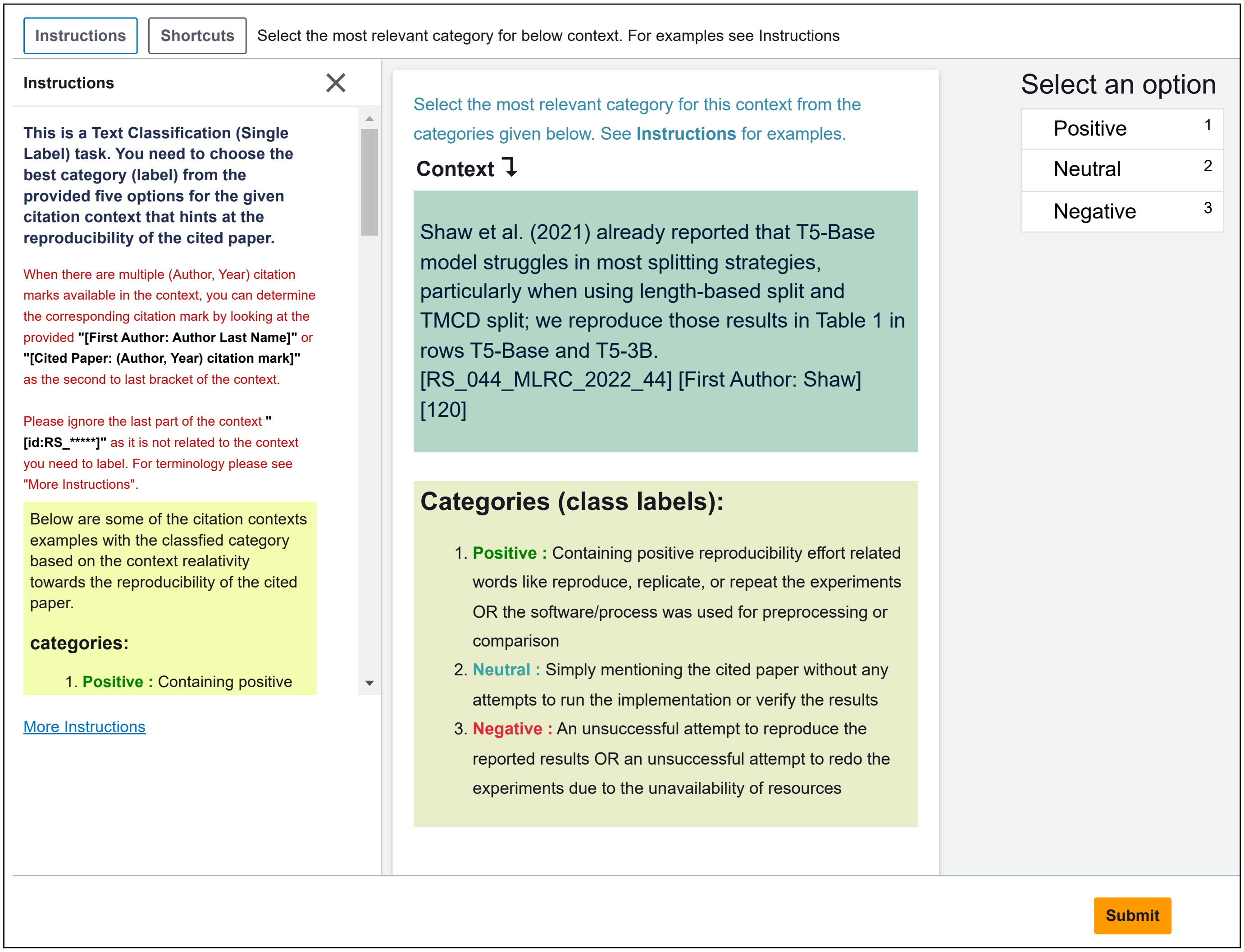


Fig. 2. Confusion matrices of five sentiment classification (3-class) models compared to MTurk annotator ground truth labels



Select the most relevant category for below context. For examples see Instructions

Instructions

This is a Text Classification (Single Label) task. You need to choose the best category (label) from the provided five options for the given citation context that hints at the reproducibility of the cited paper.

When there are multiple (Author, Year) citation marks available in the context, you can determine the corresponding citation mark by looking at the provided **"[First Author: Author Last Name]"** or **"[Cited Paper: (Author, Year) citation mark]"** as the second to last bracket of the context.

Please ignore the last part of the context **" [id:RS_*****]"** as it is not related to the context you need to label. For terminology please see "More Instructions".

Below are some of the citation contexts examples with the classfied category based on the context realativity towards the reproducibility of the cited paper.

categories:

1. **Positive :** Containing positive

Select the most relevant category for this context from the categories given below. See **Instructions** for examples.

Context ↴

Shaw et al. (2021) already reported that T5-Base model struggles in most splitting strategies, particularly when using length-based split and TMCD split; we reproduce those results in Table 1 in rows T5-Base and T5-3B. [RS_044_MLRC_2022_44] [First Author: Shaw] [120]

Categories (class labels):

1. **Positive :** Containing positive reproducibility effort related words like reproduce, replicate, or repeat the experiments OR the software/process was used for preprocessing or comparison
2. **Neutral :** Simply mentioning the cited paper without any attempts to run the implementation or verify the results
3. **Negative :** An unsuccessful attempt to reproduce the reported results OR an unsuccessful attempt to redo the experiments due to the unavailability of resources

Select an option

Positive	1
Neutral	2
Negative	3

Fig. 3. Crowdsourcing task interface using Crowd HTML elements.