First-AID: the first Annotation Interface for grounded Dialogues

Stefano Menini[1], **Daniel Russo**[1,2], **Alessio Palmero Aprosio**[2], **Marco Guerini**[1]
[menini,drusso,guerini]@fbk.eu a.palmeroaprosio@unitn.it

[1]Fondazione Bruno Kessler, Trento, Italy
[2]University of Trento, Trento, Italy

Abstract

The swift advancement of Large Language Models (LLMs) has led to their widespread use across various tasks and domains, demonstrating remarkable generalization capabilities. However, achieving optimal performance in specialized tasks often requires fine-tuning LLMs with task-specific resources. The creation of high-quality, human-annotated datasets for this purpose is challenging due to financial constraints and the limited availability of human experts. To address these limitations, we propose First-AID, a novel human-in-the-loop (HITL) data collection framework for the knowledge-driven generation of synthetic dialogues using LLM prompting. In particular, our framework implements different strategies of data collection that require different user intervention during dialogue generation to reduce post-editing efforts and enhance the quality of generated dialogues. We also evaluated First-AID on misinformation and hate countering dialogues collection, demonstrating (1) its potential for efficient and high-quality data generation and (2) its adaptability to different practical constraints thanks to the three data collection strategies.

Content warning: this paper contains unobfuscated examples some readers may find offensive

1 Introduction

The rapid progress in large language models (LLMs) has enabled their use across a multitude of tasks and domains, thanks to their remarkable generalization abilities. However, simple prompting does not suffice for optimal performance in specialized tasks. Consequently, researchers have concentrated on developing resources tailored to fine-tune the LLMs for specific tasks (Liu et al., 2022b). Nonetheless, some of these datasets remain inaccessible to the public due to legal restrictions, including issues of privacy and data ownership (Abowd and Vilhuber, 2008; Goyal and Mahmoud, 2024). Additionally, creating datasets curated solely by humans poses challenges, particularly in dialogical contexts. This approach is constrained by both financial considerations and the scarcity of human experts, as well as potentially restricted diversity, biases, and annotation artifacts in the content produced (Geva et al., 2019; Gururangan et al., 2018; Chmielewski and Kucker, 2020).

To overcome these limitations, researchers have recently leveraged LLMs to automatically generate synthetic datasets (Long et al., 2024). This approach not only reduces costs (Honovich et al., 2023) but also enables a more in-depth study of real-world domains by emulating privacy-constrained data, such as health or social media data (Kurakin et al., 2023). However, LLMs still tend to generate factual inaccuracies (Augenstein et al., 2024) and struggle with coherence and consistency, particularly for complex tasks (Dou et al., 2022). Furthermore, creating synthetic data that exhibits both diversity and complexity remains a challenging task (Liu et al., 2022a).

To address these limitations and improve LLMs' training, researchers have proposed hybrid data collection approaches that combine LLMs' generation capabilities with human experts' post-editing efforts. The *human-in-the-loop* generation strategy (HITL henceforth) has been proven to reduce costs and time, alleviate the workload of human post-editors, overcome the limitations of LLMs, and facilitate the generation of high-quality data (Tekiroğlu et al., 2022). The presence of a human in the loop during data collection is particularly crucial for knowledge-driven tasks, where accuracy and faithfulness to context must be ensured (Russo et al., 2023).

To accelerate synthetic data collection, several frameworks and tools have been proposed to either automatically generate datasets according to specific prompts and requirements (Daniel and Fran-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 563–571
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

cisco, 2023; Patel et al., 2024) or to facilitate post-editing and labeling of generated data (Tkachenko et al., 2020). However, most existing data generation tools have limited control over dialogue generation, typically requiring the model to produce an entire dialogue without human intervention between turns (Bonaldi et al., 2022). This can lead to cascading errors, resulting in increased post-editing efforts and potentially reduced data diversity.

To address these limitations, we propose a novel data collection framework called *First-AID* (First Annotation Interface for grounded Dialogues) for the automatic knowledge-driven generation of synthetic dialogues that incorporates a human-in-the-loop. Our framework implements different HITL strategies, leveraging user input during the generation of the dialogue to reduce post-editing efforts and improve the quality of generated dialogues. We designed an interactive interface enabling users to also employ external context and automatically associate pieces of this context with dialogue turns (necessary for RAG-based approaches), post-edit each turn before generating the next, and drive the generation process dynamically. This interface connects to a customizable API that allows for personalized dialogue generation to cover different topics and roles, and to configure the LLM and retrievers used in the interactions. Our interface can be specifically tailored for a wide range of use cases where high-quality dialogue generation is critical. We tested the interface for the creation of dialogues to counter hate speech and misinformation.

2 Related Work

As LLMs continue to advance in their ability to generate human-like text and generalize across a wide range of tasks, researchers are increasingly leveraging them for the automatic generation of synthetic data. This approach enables the reduction of data collection costs by minimizing or eliminating the need for human annotation efforts, promoting data diversity, and mitigating potential annotation artifacts (Lu et al., 2024).

Two primary approaches have emerged for synthetic dialogue generation: *LLM-only* methods, where one or more LLMs generate data entirely on their own (Chen et al., 2023; Penzo et al., 2025), and hybrid approaches that integrate human feedback or corrections into the dialogue generation process through a HITL strategy.

Most existing works leverage human intervention in a post-processing phase to correct possible errors and adjust the generated dialogue (Bonaldi et al., 2022; Occhipinti et al., 2024). Conversely, Lu et al. (2024) proposed the DialGen framework, which enables human feedback within the dialogue generation process itself. This allows a human reviewer to modify the dialogue at each turn and for the system to automatically generate the next turn accordingly.

While this framework can mitigate and correct LLM errors, it heavily relies on expert intervention to correct potential factual inaccuracies. The model requires generating based on a short story generated from ontology triplets (Kim et al., 2023), which may limit its application to domains requiring up-to-date knowledge that is not readily available in an ontology format or cannot be easily shaped into that form, such as misinformation detection, hate speech mitigation, or company-specific use cases.

To address this limitation, we built upon the DialGen framework (Lu et al., 2024) by proposing a novel knowledge-driven dialogue generation framework with HITL capabilities. This framework enables the generation of dialogues based on information provided in textual documents. At each turn, a human can revise and modify the generated text if needed, before proceeding to generate the next turn. To promote diversity during the generation phase while minimizing human effort, we require the model to generate three different versions of a specific turn. Our framework also allows humans to select the relevant text portions used for generating a turn, making it suitable for training more sophisticated knowledge-driven pipelines that integrate retrieval and reranking components in a RAG scenario (Lewis et al., 2020).

Furthermore, we recognized the need for a comprehensive tool that enables seamless dialogue generation and post-editing capabilities. To address this need, we created an intuitive interface that empowers end-users to automatically generate, edit, and customize knowledge-grounded dialogues in a flexible and user-friendly manner.

3 Task Description

First-AID is an annotation interface that aims to provide an environment for creating dialogical RAG-structured data in a human-AI collaboration setting (i.e., using HITL methodology). The tool focuses on the creation of scenario-specific dialogues starting from domain documents. The RAG-

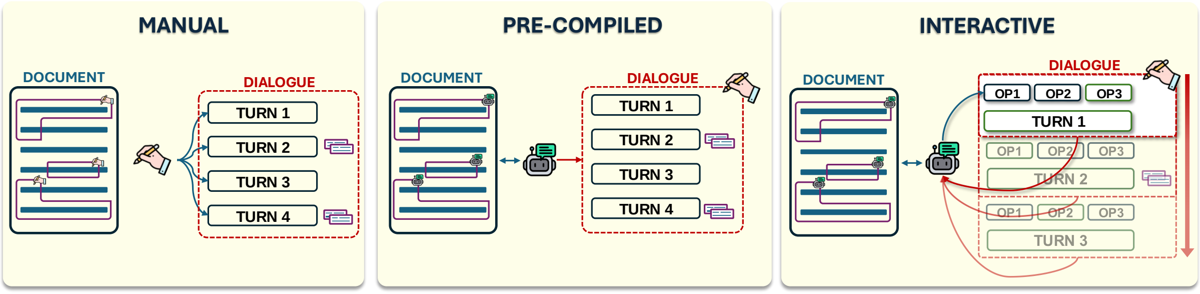


Figure 1: Graphical representation of the three data collection strategies supported by First-AID platform.

oriented connotation of the tool is given by allowing the annotator to link each turn in the dialogues to a specific document and, more in detail, to the portions of that document (a sentence, paragraphs, or custom spans) containing the information on which the turn is based. Each dialogue and individual turn can be associated with multiple ground texts, allowing the data created through the platform to be used to train not only a dialogical language model but also retriever components.

Data collection speed optimization: The goal of the First-AID platform is not only the creation of dialogical data, but also to provide an environment that allows testing and customising different data collection strategies to find the most appropriate for each specific data collection scenario. The tool adopts a HITL approach, proposing different strategies with different levels of automation and human control. This allows for tuning the human effort in the annotation process, identifying the best annotation setting to improve the data collection speed without sacrificing the quality of the output.

Multiple data collection strategies: The platform supports three different data collection strategies to create dialogues starting from one or more reference documents. In all three configurations, each turn, if needed, is grounded in the documents by pairing it with the relevant passages. A graphical representation is provided in Figure 1. The three main configurations that we implemented in the platform are:

1. **Manual**: the dialogue is manually written from scratch based on the provided document(s). The portions of the text on which the turn is based are manually linked.

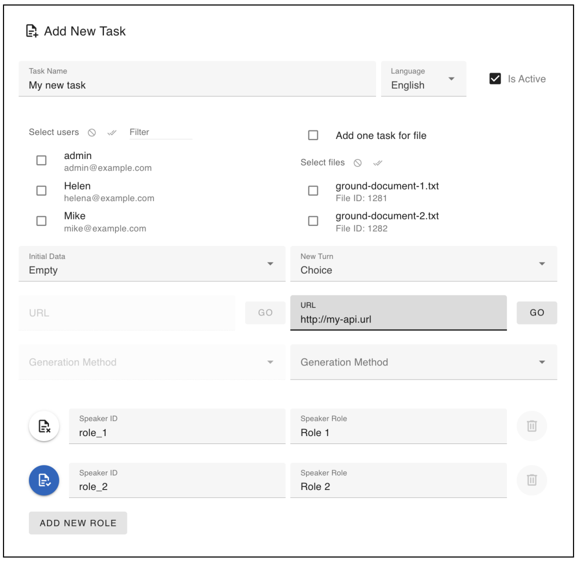
2. **Pre-compiled**: starting from the documents, we use an LLM to automatically generate a full dialogue based on the sources. The turns are automatically paired with the portions of text on which they are grounded. The annotator reviews the generated dialogue and makes any necessary edits to the turns (i.e., editing the text, removing unnecessary turns, or adding new ones) and to the ground (i.e., changing the boundaries of the text, adding new grounds, or removing the incorrect ones).

3. **Interactive**: the annotator is assisted by an LLM that suggests multiple options for each new turn in the dialogue. The annotator selects and edits one of the suggestions or, at will, can propose a new turn from scratch. Each turn is built upon the previous ones, and the human's choices guide the progression and direction of the dialogue.

Multiple annotation layers: In addition to creating dialogues from scratch, the platform allows users to post-edit them. Dialogues created using any of the three annotation modalities can serve as a starting point for other users' post-editing. This allows for multiple layers of annotations, for instance, to generate variations of the same dialogue or to have experts acting as curators to review other annotators' data.

Iterative model improvement: The *interactive* and *pre-compiled* task configurations can be used to iteratively refine a specific NLG model by *i)* generating dialogues with that model, *ii)* post-editing them, and *iii)* incorporating the edited data into subsequent model training iterations. Keeping a human in the loop through iterative feedback can help improve the model over time.

Customizable LLMs: When creating a pre-compiled or interactive dialogue, the system can be linked to custom APIs that generate the turns. This enables full customization of the models, the



Add New Task

Task Name: My new task

Language: English

☑ Is Active

Select users

□ admin admin@example.com

□ Helen helena@example.com

□ Mike mike@example.com

□ Add one task for file

Select files

□ ground-document-1.txt File ID: 1281

□ ground-document-2.txt File ID: 1282

Initial Data: Empty

New Turn: Choice

URL: http://my-api.url

Speaker ID: role_1 | Speaker Role: Role 1

Speaker ID: role_2 | Speaker Role: Role 2

Figure 2: The task creation screen

prompts, the actors involved in the dialogue, and their behavior or style.

4 System Description

First-AID is a multi-layered web-based system. An admin user is created automatically during setup. This user can access the interface to create projects and invite other users. Each project can be assigned to a group of users, with some designated as project managers. Within a project, tasks can be created and text documents can be assigned. Each task represents a single dialogue that may be automatically generated by the LLM. A user can then edit the dialogue, optionally using interactive suggestions to assist with the annotation. In doing so, the user can also select parts of the documents associated with the task, that represent the ground truth related to that turn (see Section 3). Once the user confirms the annotation, a project manager can assign the task to another user for further refinement. This process can be repeated as needed until the admin or a project manager closes the task.

During the task creation phase (see Figure 2), additional information is provided, such as the roles in the dialogue, the LLMs that have to be queried for the initial or for the interactive generation, and the documents that the annotator can use.

4.1 The annotation interface

The key innovation introduced by First-AID lies in its annotation interface (Figure 3), which allows users to write dialogue turns and link each one to specific source texts.

The interface is organized into three columns:

- The left column displays the source file(s). Annotators can highlight sections of the text and assign them to dialogue turns, indicating that the selected content was used to generate that part of the conversation.
- The middle column shows the dialogue itself, which the annotator can freely edit.
- The right column lists the text spans linked to each dialogue turn. Clicking on a span highlights the corresponding source text on the left, helping the annotator easily retrieve its context.

4.2 Development lifecycle

The software development cycle followed an iterative model, starting with the implementation of an initial version. Upon deployment, annotator feedback was gathered and evaluated for the subsequent development phases and feature enhancements.

This cyclical process ensured continuous improvement, as each iteration incorporated the user inputs to refine functionality, address issues, and align the product more closely with the evolving requirements.

4.3 Release

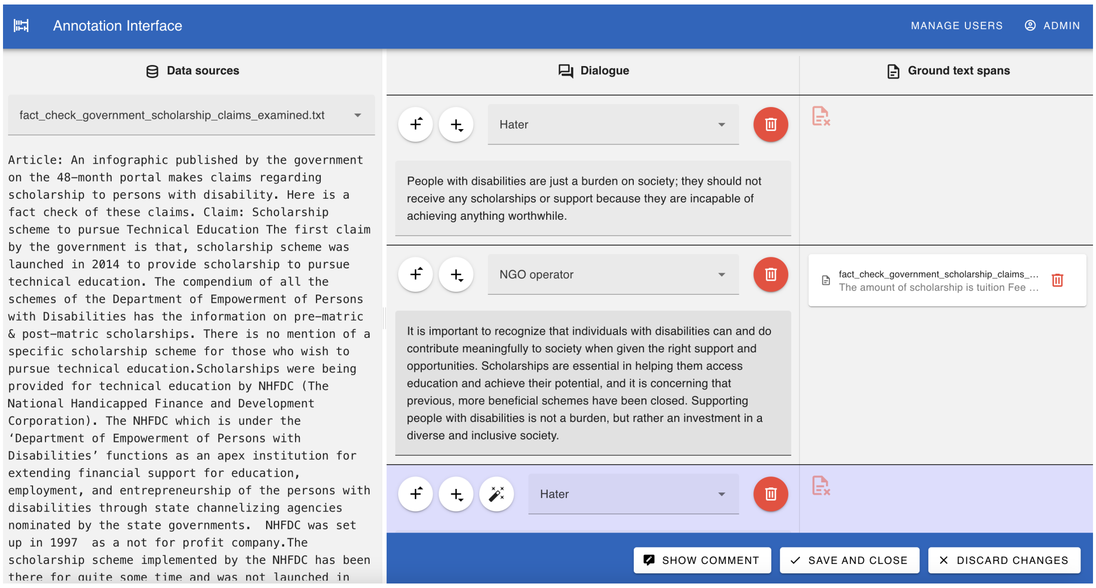
The software is implemented using VueJS (frontend) and Python/SQLite (backend). It is released on Github[1] as an open source package under the Apache license.

5 Application Scenario

The tool can be applied to data collection for several application scenarios, summarized as follows:

1. **Training Retrieval-Augmented Generation (RAG) modules:** The interface allows linking each turn of the dialogue to specific passages of the source documents. This functionality is fundamental to train RAG models, which retrieve information from external sources to generate more accurate and contextually relevant responses.
2. **Training long context modules:** The ability to handle and link dialogues to source documents of varying lengths makes the tool useful to train models that can handle larger and

[1] https://github.com/LanD-FBK/first-AID



fact_check_government_scholarship_claims_examined.txt

Article: An infographic published by the government on the 48–month portal makes claims regarding scholarship to persons with disability. Here is a fact check of these claims. Claim: Scholarship scheme to pursue Technical Education The first claim by the government is that, scholarship scheme was launched in 2014 to provide scholarship to pursue technical education. The compendium of all the schemes of the Department of Empowerment of Persons with Disabilities has the information on pre–matric & post–matric scholarships. There is no mention of a specific scholarship scheme for those who wish to pursue technical education.Scholarships were being provided for technical education by NHFDC (The National Handicapped Finance and Development Corporation). The NHFDC which is under the 'Department of Empowerment of Persons with Disabilities' functions as an apex institution for extending financial support for education, employment, and entrepreneurship of the persons with disabilities through state channelizing agencies nominated by the state governments. NHFDC was set up in 1997 as a not for profit company.The scholarship scheme implemented by the NHFDC has been there for quite some time and was not launched in

Hater

People with disabilities are just a burden on society; they should not receive any scholarships or support because they are incapable of achieving anything worthwhile.

NGO operator

It is important to recognize that individuals with disabilities can and do contribute meaningfully to society when given the right support and opportunities. Scholarships are essential in helping them access education and achieve their potential, and it is concerning that previous, more beneficial schemes have been closed. Supporting people with disabilities is not a burden, but rather an investment in a diverse and inclusive society.

fact_check_government_scholarship_claims_...
The amount of scholarship is tuition Fee ...

Hater

Figure 3: The annotation screen

more complex dialogue contexts even without the use of RAG modules.

3. **Training on proprietary/specific use-cases:** The tool can be adapted to a wide range of use cases where high-quality dialogue generation is crucial and the use of API commercial tool is not an option.

4. **Improving existing LLMs via direct interaction** (and indirectly evaluating the quality of a system): Our platform can be linked directly to the LLM to be deployed. Thus, not only the generated and post-edited dialogues can be used to improve the performance of existing LLMs, but they also represent direct correction of their output.

5. **Intrinsic evaluation of the quality of a dialogue generation system**: By analyzing the amount and type of post-editing required, it is possible to understand the quality of the LLM being developed, as First-AID saves both the messages proposed by the LLMs and the edited ones.

6 Evaluation

First-AID showcases its versatility through both the range of implementable data collection strategies and the variety of dialogical domains it is capable of addressing. For instance, its application extends to critical areas such as healthcare (Zhou et al., 2021), education (Tack et al., 2023), public administration (Nirala et al., 2022), and increasingly important society-driven domains like misinformation and hate speech countering (Bonaldi et al., 2022).

To evaluate the First-AID platform, we organized four evaluation sessions with 50 experts on a specific task of misinformation and hate countering dialogues rooted in fact-checking articles. Participants came from four different European countries and were either fact-checkers from recognized organizations or NGO members devoted to hate speech countering. The evaluation included both *qualitative* (via interviews) and *quantitative* (via analysis of users' activity logs) aspects. Below we report a summary of the sessions structure and main findings.

Sessions structure. Each evaluation session lasted around two and a half hours. They started with a brief introduction of the evaluation and its aims, followed by a description of the tasks to be performed by the participants. After introducing the specific guidelines for the task, we presented the platform, together with the three interface modalities for data collection. Half an hour was dedicated to explaining each modality and to allowing participants to exercise with it. In particular, each session of data collection was introduced by a 10-minute tutorial on the specific modality usage, followed by a 20-minute hands-on annotation activity with the same modality. As a final step, we closed each evaluation session with a half-hour feedback discussion to gather issues, impressions, and suggestions from the participants on the tasks they performed and the platform as a whole.

	Avg. Time per Dialogue (sec)	Avg. Turns per Dialogue	Avg. Words per Dialogue	Words per minute	Turns per minute	Turns with Ground (%)	Avg. Number of Grounds per Turn
Manual	1006.06	4.74	132.79	7.92	0.28	70.90	1.56
Pre-compiled	825.37	6.25	162.28	11.80	0.45	83.15	2.18
Interactive	479.28	3.98	141.26	17.68	0.50	85.88	1.10

Table 1: Dialogues statistics over the data collected through the three different collection strategies.

Qualitative: Platform Feedback. Overall, the interface was deemed user-friendly, intuitive, and generally simple to interact with (*"The system operates with great fluidity, offering a smooth and seamless experience. The interface is responsive, making navigation easy and efficient, which enhances overall user satisfaction."*, *"The platform is intuitive. It provides a seamless user experience with easy navigation and simple interfaces that allow users to quickly engage with its features."*). However, some participants agreed on the need to improve the standards for the automatically generated text and for the types of articles included in the tasks (*"One of the main drawbacks of the app is that the counter-narrative it generates often follows a repetitive pattern, offering limited variation and struggling to address more nuanced forms of hate speech."*). This is not strictly related to the interface quality itself, but points to the need to properly craft the task within the platform.

Quantitative: Modalities Feedback. While we could have expected a clear-cut preference of some modalities over the others, the results of the interviews indicated a multifaceted evaluation that was taking into account three main variables/criteria: *locus of control* of the annotator (e.g. how much control they want to have on the unfolding of the conversation), quality of the *LLM output* that is connected to the various modalities, *interlocutors' rendering* (i.e., the quality of the output is good in term of grammaticality but the LLM is not able to proper render one of the interlocutors stances, such as hater's). Depending on how much a variable is relevant, the choice went to one of the modalities. In Table 2 we report the main values that drove the preference of the annotators.

From the interviews emerged that the manual strategy, while commended for not introducing *"biases beforehand"* and being *"a very good option for educational functionalities"*, was also less preferred for being *"more labor-intensive"* and requiring *"more time to complete the task"*. In contrast, the pre-compiled strategy was appreciated for its efficiency, with users finding it *"a rapid way to give an accurate response with fact-checked arguments"*. However, this speed came at the cost of repetition, with feedback indicating that *"some of the answers were pretty repetitive and the dialogue got stuck"*. The interactive strategy emerged as a promising middle ground, with users appreciating the *"flexibility to create answers"* and the *"accurate"* and *"useful"* AI-generated responses. One user particularly highlighted its advantage over the pre-compiled option, noting that it *"facilitates the job and makes it more efficient but still allows controls from a human"*. While largely positive, minor issues were noted, such as the system occasionally generating responses from the wrong persona.

Modality	Dimension	Feedback
Manual	LLM output Interlocutors' rendering Locus Control	↓↓ ↓ ↑↑
Pre-compiled	LLM output Interlocutors' rendering Locus Control	↑ ↑↑ ↓
Interactive	LLM output Interlocutors' rendering Locus Control	↑ ↓ ↑

Table 2: Expert preference for the various modalities according to locus of control, LLM output quality, interlocutor's rendering. ↑ indicates a positive correlation with the dimension, a ↓ a negative one.

Quantitative: Efficiency. The Interactive modality demonstrates the highest time efficiency, with an average annotation time per dialogue of 479.28 seconds and 17.68 words per minute. This is significantly faster than both the Manual (1006.06 seconds) and Pre-compiled (825.37 seconds) modalities. The reduced annotation time in the Interactive modality suggests that the ability to provide alternative turns to choose from streamlines the annotation process. The Manual modality, unsurprisingly, shows the lowest annotation speed (7.92 words/minute, 0.28 turns/minute), reflecting the inherent time constraints of manual annotation.

Quantitative: Dialogue Length. The Pre-compiled modality exhibits the longest dialogues, both in terms of average turns (6.25) and average words (162.28) per dialogue. This suggests that leaving it up to the LLM the possibility to create the whole material allows obtaining more articulated dialogues. In contrast, the Interactive modality has the shortest dialogues (3.98 turns, 141.26 words), indicating a more concise dialogue style. Still, the turns are longer than the manual modality.

Quantitative: Grounding Patterns. The average percentage of turns with grounds is quite consistent across all modalities, ranging from 70.90 (Manual) to 85.88 (Interactive). Turning to the average number of provided grounds (for the turns that have a ground), it can be noted that it is around 2.18 for the Pre-compiled, 1.56 for the manual, while it is notably lower for the Interactive modality having the lowest percentage (1.10, explained by the generation modality that is based on only one ground). This suggests that the ability to provide even a suboptimal list of grounds to choose from is helping annotators to provide more evidence per turn. This observation hints that further investigation into strategies for encouraging more explicit grounding is needed.

7 Conclusions

This paper introduces First-AID, a novel annotation interface designed to facilitate the knowledge-driven generation of synthetic dialogues with a HITL approach. The interface provides three distinct data collection strategies: manual writing, post-editing of pre-compiled dialogues, and interactive dialogue creation with LLM assistance. Evaluation results indicate that the interactive modality offers the highest time efficiency, while the pre-compiled modality generates the longest dialogues. User feedback highlights the platform's user-friendliness and intuitiveness, with suggestions for improvements in LLM-generated text quality and source article relevance. Future work will focus on refining the system based on user feedback and exploring its use in other domains.

Acknowledgments

This work was supported by Provincia Autonoma di Trento under grant agreement No. C49G22001020001 (AIxPA) and by the European Union's CERV fund under grant agreement No. 101143249 (HATEDEMICS).

References

John M. Abowd and Lars Vilhuber. 2008. How protective are synthetic data? In *Privacy in Statistical Databases*, pages 239–246, Berlin, Heidelberg. Springer Berlin Heidelberg.

Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, Eduard Hovy, Heng Ji, Filippo Menczer, Ruben Miguez, Preslav Nakov, Dietram Scheufele, Shivam Sharma, and Giovanni Zagni. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

Michael Chmielewski and Sarah C. Kucker. 2020. An mturk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science*, 11(4):464–473.

Vila-Suero Daniel and Aranda Francisco. 2023. Argilla - Open-source framework for data-centric NLP.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China. Association for Computational Linguistics.

Mandeep Goyal and Qusay H. Mahmoud. 2024. A systematic review of synthetic data generation techniques using generative ai. *Electronics*, 13(17).

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. Unnatural instructions: Tuning language models with (almost) no human labor. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. 2023. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022a. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022b. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965. Curran Associates, Inc.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Bo-Ru Lu, Nikita Haduong, Chia-Hsuan Lee, Zeqiu Wu, Hao Cheng, Paul Koester, Jean Utke, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2024. Does collaborative human–LM dialogue generation help information extraction from human–human dialogues? In *First Conference on Language Modeling*.

Krishna Kumar Nirala, Nikhil Kumar Singh, and Vinay Shivshanker Purani. 2022. A survey on providing customer and public administration based services using ai: chatbot. *Multimedia Tools and Applications*, 81:22215–22246.

Daniela Occhipinti, Michele Marchi, Irene Mondella, Huiyuan Lai, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2024. Fine-tuning with HED-IT: The impact of human post-editing for dialogical language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11892–11907, Bangkok, Thailand. Association for Computational Linguistics.

Ajay Patel, Colin Raffel, and Chris Callison-Burch. 2024. DataDreamer: A tool for synthetic data generation and reproducible LLM workflows. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3781–3799, Bangkok, Thailand. Association for Computational Linguistics.

Nicolò Penzo, Marco Guerini, Bruno Lepri, Goran Glavaš, and Sara Tonelli. 2025. Don't stop the multiparty! on generating synthetic multi-party conversations with constraints. *Preprint*, arXiv:2502.13592.

Daniel Russo, Shane Kaszefski-Yaschuk, Jacopo Staiano, and Marco Guerini. 2023. Countering misinformation via emotional response generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11476–11492, Singapore. Association for Computational Linguistics.

Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The BEA 2023 shared task on generating AI teacher responses in educational dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 785–795, Toronto, Canada. Association for Computational Linguistics.

Serra Sinem Tekiroğlu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114, Dublin, Ireland. Association for Computational Linguistics.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Meng Zhou, Zechen Li, Bowen Tan, Guangtao Zeng, Wenmian Yang, Xuehai He, Zeqian Ju, Subrato Chakravorty, Shu Chen, Xingyi Yang, Yichen Zhang, Qingyang Wu, Zhou Yu, Kun Xu, Eric Xing, and Pengtao Xie. 2021. On the generation of medical dialogs for COVID-19. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 886–896, Online. Association for Computational Linguistics.