Tailoring Rumor Debunking to You: Diversifying Chinese Rumor-Debunking Passages with an LLM-Driven Simulated Feedback-Enhanced Framework

Xinle Pang[1,2], **Danding Wang**[1,2*], **Qiang Sheng**[1,2], **Yifan Sun**[1,2], **Beizhe Hu**[1,2], **Juan Cao**[1,2]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
pangxinle23@mails.ucas.ac.cn,
{wangdanding, shengqiang18z, sunyifan23z, hubeizhe21s, caojuan}@ict.ac.cn

Abstract

Social media platforms have become primary sources for news consumption due to their real-time and interactive nature, yet they have also facilitated the widespread proliferation of misinformation, negatively impacting public health, social cohesion, and market stability. While professional fact-checking is essential for debunking rumors, the process is time-consuming, necessitating automation to effectively combat fake news. Existing approaches, such as extractive methods, often lack coherence and context, whereas abstractive methods leveraging large language models (LLMs) can generate more readable and informative debunking passages. However, readability alone is insufficient for effective misinformation correction; user acceptance is critical. Recent advancements in LLMs offer new opportunities for personalized debunking, as these models can generate context-sensitive responses and adapt to user profiles. Building on this, we propose the **MU**lti-round **R**efinement and **S**imulated f**E**edback-enhanced framework (**MURSE**), which generates Chinese user-specific debunking passages by iteratively refining outputs based on simulated user feedback. Specifically, MURSE-generated user-specific debunking passages were preferred twice as often as general debunking passages in most cases, highlighting its potential to improve misinformation correction and foster positive dissemination chains.

1 Introduction

Social media platforms are increasingly preferred over traditional media due to their real-time and interactive qualities, becoming primary sources for news consumption. However, this shift has facilitated the proliferation of fake news across platforms in various domains, especially after generative AI techniques have made significant progress (Liu et al., 2024; Hu et al., 2025). Such misinformation negatively impacts public health (Pierri et al., 2022), social cohesion (Shu et al., 2019), and market stability (Micevičienė et al., 2024), making its moderation a priority for maintaining a healthy information ecosystem.

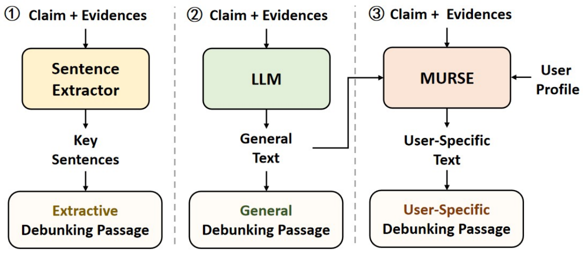


Figure 1: Paradigm comparison between existing ① extractive, ② abstractive approaches, and our proposed method ③ MURSE. Our proposed MURSE considers the user profile to personalize the generation of debunking passages, which can more effectively help vulnerable populations clarify misconceptions.

In cases where rumors have already gained widespread traction, simply labeling information as a rumor without providing specific explanations is insufficient. An effective explanation to debunk rumors is fact-checking reports written by professional fact-checkers. However, the fact-checking reports this time-consuming process necessitates automation to effectively curb fake news proliferation and its harmful effects. While some existing approaches have employed extractive methods to identify key sentences as debunking information (Yang et al., 2022; Atanasova et al., 2020; Russo et al., 2023b), these extracted sentences often lack coherence and comprehensive context. Given the impressive capabilities of LLMs, abstractive approaches now can leverage LLMs to generate readable and informative debunking passages.

However, debunking passages with adequate readability alone is insufficient for widespread dis-

[*]Corresponding author

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 586–597
March 25-27, 2026 ©2026 Association for Computational Linguistics

semination to achieve the purpose of misinformation correction. For misinformation correction to be truly effective, it must transcend basic readability standards and focus on user acceptance (Ma et al., 2023). Through the collection and analysis of user feedback, fact-checkers can precisely identify the cognitive tendencies and psychological needs of their target audience, thereby optimizing the presentation of corrective information (Basol et al., 2020). This user-centered approach to debunking not only enhances the persuasiveness of the content but also ensures that information effectively reaches those most susceptible to misinformation (Pennycook and Rand, 2019). When corrective content aligns with the audience's comprehension abilities, concerns, and value systems, individuals become more willing to proactively share such information, creating positive dissemination chains that ultimately help vulnerable populations clarify misconceptions and resist the adverse effects of false information (Guo et al., 2020; De keersmaecker and Roets, 2017; Sun et al., 2025). He et al. (2023) have adopted response generation methods for debunking misinformation, but they fail to take into account users personal profiles for personalized debunking.

In this context, the rapid development of large language models (LLMs) offers new perspectives and technical support for addressing this issue. LLMs are capable of generating responses through role-playing, as they possess a strong ability to comprehend human instructions and produce high-quality text. Furthermore, they can adapt their responses based on interactions with different individuals, enabling more personalized and context-sensitive debunking strategies. Russo et al. (2023a) has utilized LLM to generate corresponding debunking short texts for rumors with different emotions and styles specific to social media platforms, exploring strategies for debunking passages.

We propose a **MU**lti-round **R**efinement and **S**imulated f**E**edback-enhanced framework (**MURSE**) for generating debunking passages. **MURSE** is based on rumors and evidence, utilizing LLM to generate debunking passages and iteratively refining them through multi-round revisions based on simulated user feedback. Fig. 1 illustrates the distinctions between the extractive approach, the abstractive approach, and our proposed MURSE framework. Our MURSE framework is capable of generating user-specific debunking passages. We evaluated the generated user-specific debunking passages using three quantifiable criteria and conducted human evaluations on corresponding profiles. Through multiple rounds of iteration, the MURSE framework improves the performance of these metrics. Moreover, in human evaluations, the user-specific debunking passages generated by MURSE were preferred twice as often as general debunking passages in most cases.

2 Related Works

Explanation Generation for Fact Checking. According to Russo et al. (2023b), fact-checking tasks are often divided into two parts: one part involved *Veracity Prediction*, while the other is the more challenging task of explanation generation for the verdict (*Justification Production*). Since the inputs are often rumors and evidence, explanation generation for fact-checking is typically done by summarizing(Kotonya and Toni, 2020a; Eldifrawi et al., 2024), which is further divided into approaches of *extractive approach* like (Atanasova et al., 2020) and (Yang et al., 2022), and *abstractive approach* like (Kotonya and Toni, 2020b). The extractive approach often adopts joint learning of veracity prediction and summary sentence extraction to generate explanations. However, this approach tends to yield debunking passages that lack coherence, are not content-rich, and have poor readability. In contrast, the abstractive approach bases evidence to generate new sentences for explanation. With the development of LLMs, this approach has become more promising than simply extracting sentences (Yue et al., 2024; Russo et al., 2025). Therefore, we employ the abstractive approach based on iterative feedback to the LLM-based passage generation module.

Simulated Human Feedback on Social Media Platform. Human feedback is a crucial signal for related work, such as rumor detection and stance detection (Ma et al., 2018; Zhang et al., 2021), as it provides an additional perspective about how the crowd reacts to the described event and indicates the consequences that the message creator intends to make (Wang et al., 2025b,a). Jiang and Wilson (2018) analyze linguistic signals in user comments on social media posts associated with misinformation and fact-checking. Gatto et al. (2023) use the chain-of-thoughts embedding for stance detection on social media platforms. And Nan et al. (2025) distill the comment information to a content-only detector to facilitate early detection. With the emergence of LLMs, recent studies have explored using these models to simulate human feedback by generating role-specific comments (Qiu et al., 2025). In the domain related to fake news detection, Wan et al. (2024) use LLM to generate comments for social graphs and simulate the social media platform in the real world. Nan et al. (2024) generate comments from multiple subpopulations within diverse views and make veracity judgments. However, their ultimate goal is to improve the detection performance of misinformation. In this work, we exploit the simulation of human feedback based on LLM-driven role-playing to refine rumor-debunking passages.

3 Method

Existing studies have demonstrated that LLMs possess the capability to generate readable debunking passages (Yue et al., 2024; Kim et al., 2024). However, these studies have not adequately addressed the need to personalize debunking passages based on different user characteristics. Fig. 2 shows an overview of our framework. In our approach, after generating the initial debunking passage, we introduce a Simulated Feedback Module that provides refinement advice based on user responses to rumors. Subsequently, we employ a multi-round Passage Refinement Module to iteratively modify these passages, enhancing their effectiveness. The debunking passage undergoes continuous improvement until it satisfies our established criteria.

3.1 Debunking Passage Initialization

For each rumor, there exist several corresponding pieces of evidence provided by professional fact-checking organizations such as CHEF (Hu et al., 2022). But these pieces of evidence are always too long to read fast for most people, making them harder to gain widespread circulation. Unlike these official evidences, in this paper, we propose generating personalized debunking passages tailored to specific user characteristics, designed for broader dissemination across social media platforms to help vulnerable populations clarify misconceptions. The rumor, along with its related evidence, is input into the LLM to generate initial debunking passages for refinement. The prompt is shown in Appendix A.

3.2 MURSE Framework

Our framework consists of two main modules: the Simulated Feedback Module and the Passage Refinement Module. There are also three roles simulated by LLM: **Commenter**, **Advisor**, and **Editor**. The input is the initial debunking passage, and the output is a refined, user-specific debunking passage that meets our established evaluation criteria. For the following modules, the utilized prompts are provided in Appendix A.

3.2.1 Simulated Feedback Module

In this module, we utilize the **Commenter** to simulate target user responses to rumors on social media. The **Advisor** then analyzes the user feedback alongside the debunking passage to generate advice for further refinement.

The **Commenter** simulates real users on social media platforms who, when exposed to a circulating rumor, choose to believe it and post their own comments. For user profiles, we follow Nan et al. (2024), selecting gender, age, and education as key attributes. Specifically, these three attributes are categorized as follows:

- **Gender:** male; female.
- **Age:** adolescent; young adult; the middle-aged; the elderly.
- **Education:** college graduate; has not graduated from college; has a high school diploma or less.

Similarly, we combined these three attributes, resulting in 24 possible combinations of user profiles. The user profile, along with the rumor, is then used to prompt the **Commenter** to generate simulated comments.

The **Advisor** analyzes simulated user feedback regarding rumors and proposes further refinement advice for debunking passage (either initial or iteratively refined versions). The advice is then forwarded to the **Editor** for subsequent modifications.

3.2.2 Passage Refinement Module

In this module, we utilize the **Editor** to modify the debunking passage based on advice from the **Advisor**. We propose three criteria to evaluate whether the modified debunking passage meets the requirements. If qualified, it is output as the final user-specific debunking passage; if unqualified, the modified debunking passage is input back into the previous stage for iterative improvement.

The **Editor** is responsible for modifying the debunking passage according to the advice proposed by the **Advisor**. To ensure that throughout the iterative process, the debunking result maintains fidelity to the original facts, we also provide the rumor and supporting evidence to the **Editor**.

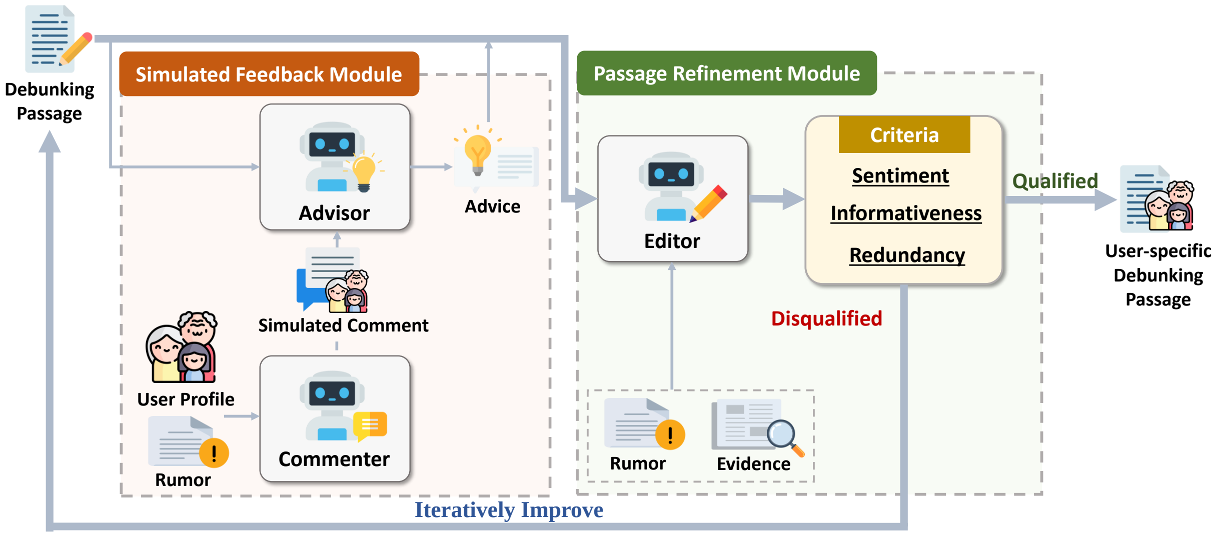


Figure 2: Illustration of the proposed **MURSE** framework. MURSE consists of a Simulated Feedback Module and a Passage Refinement Module. The Simulated Feedback Module utilizes a Commenter to simulate target user responses to rumors and an Advisor to analyze user feedback and provide advice. The Passage Refinement Module uses an Editor to modify the debunking passage based on this advice. The debunking passage undergoes continuous improvement until it satisfies our established criteria.

Given the modified debunking passage generated by the **Editor**, we introduced three criteria—**sentiment**, **informativeness** and **redundancy** to evaluate the quality of the debunking passage:

- **Sentiment:** Emotion plays an important role in the field of news dissemination (Kjerstin Thorson and Ekdale, 2010; Steffens et al., 2019; Zhang et al., 2021; Russo et al., 2023a;). To mitigate user defensiveness, we ensure that debunking passages maintain a positive tone.
- **Informativeness:** Chan et al. (2017) demonstrate that debunking passages containing more evidence-based information are more effective in reducing misconceptions and increasing user acceptance. Therefore, we measured the informativeness of debunking passages to ensure superior debunking outcomes.
- **Redundancy:** Lewandowsky et al. (2012) identify the familiarity backfire effect, where repeatedly mentioning misinformation during corrections actually reinforces false claims. Similarly, Lazer et al. (2018) demonstrate that restating erroneous information increases its familiarity, potentially causing debunking efforts to backfire. Thus, we require low redundancy between the debunking passages and the rumor to reduce the familiarity backfire effect.

When these three criteria reach the specified thresholds, we consider the modified debunking passage to have met the requirements and can be output as the user-specific debunking passage. If the requirements are not met, the modified debunking passage is resubmitted to the **Advisor** for a new iteration of revision advice. This process iterates continuously until the requirements are satisfied or the maximum number of iterations is reached.

Table 1: Domain distribution and statistics for the test dataset of CHEF.

Society	Culture	Health	Science	Politics	Total
117	12	143	31	30	333

Statistical Indicator	#
Avg #Words in Rumor	25
Avg #Words in Evidence	4,018
Avg #Words in Gold Evidence	124

4 Experiment

We experimentally answer the following questions:

- **EQ1**: Is the multi-round iterative improvement and user simulation module in MURSE effective?
- **EQ2**: How does the MURSE framework's response to the target user reflect in human evaluations?

4.1 Dataset

We conduct the experiment on the public dataset CHEF (Hu et al., 2022), which stands as the only Chinese real-world evidence-based fact-checking dataset annotated with human-labeled gold evidence. This dataset contains data points that include rumor, verdict, domain, evidence, and annotated gold evidence sentences from the evidence. The gold evidence sentences can be considered a form of extractive approach for debunking passage generation. The evidence and golden evidence sentences come from the annotation team of CHEF. The annotation team has 25 members, all annotators are native Chinese speakers. The data points in CHEF are categorized into three types: supported (SUP), refuted (REF), and not enough information (NEI). Since our framework focuses on rumors, we selected the rumors labeled with REF in its test set, totaling 333 rumors. The domain distribution and statistics are shown in Table 1.

4.2 Experimental Settings

4.2.1 Compared Baselines

We compared MURSE with the following three baselines:

- **Gold Evidence**: Use manually annotated gold evidence as the debunking passage.
- **General**: Input the claim and evidence into the LLM to generate a general-purpose summary, which is used as the debunking passage.
- **Single-Round**: Directly use the debunking passage obtained in the first round of MURSE without performing iterative refinement.

4.2.2 Implementation Details

In our MURSE framework, the LLM we use for prompting is GLM-4-Air (Team GLM, 2024), which is employed to generate general debunking passages and simulate the roles of commenter, adviser, and editor. We set the sampling temperature to 0.95 to increase diversity. For generating general debunking passages and simulating the adviser and editor, we set the max tokens to 200, while for simulating a commenter, we set it to 100. This is because we consider that user comments on social media are typically relatively short. Besides, we use automated methods to evaluate the metrics. The implementation details for each dimensionare as follows:

- **Sentiment** We use HanLP (He and Choi, 2021) to assess the sentiment polarity of the debunking passage. The sentiment polarity S is a value between [-1, 1], where the sign of the value represents positive or negative emotions, and the absolute value represents the intensity of the emotion.

Table 2: Average values of three criteria. The highest-performing results are indicated in **bold**, while the second-best results are underlined.

Method	Sentiment ↑	Informativeness ↑	Redundancy ↓
Gold Evidence	0.1197	**3.3970**	0.1370
General	0.2581	2.7612	0.0974
Single-Round	0.3989	2.8869	0.0802
MURSE	**0.5118**	2.9670	**0.0690**

- **Informativeness** We use LLM to calculate the perplexity of debunking passages. We initialize the MiniCPM-2B-128k model (Hu et al., 2024), and I is calculated by using Equation 1. I is greater than 0, with a higher I indicating more informativeness. Θ denotes the parameters of the LLM (Sachan et al., 2022).

$$I = \frac{1}{|d|}\sum_{t} \log p(d_t|d_{<t};\Theta). \qquad (1)$$

- **Redundancy** There are typically ROUGE-1, ROUGE-2, and ROUGE-L (Atanasova et al., 2020) in explanation generation. Given that the purpose of this metric is to prevent the content of the rumor from appearing coherently in the debunking passage, which would undermine the debunking effect, we use F1 score of ROUGE-L as the R. A lower ROUGE-L F1 score indicates lower redundancy. The calculation formula is shown by Equation 2:

$$R = \frac{2 \times \mathrm{LCS}(c, d)}{|c| + |d|}, \qquad (2)$$

  where LCS denotes Longest Common Subsequence.

When the debunking passage is refined in the second round, MURSE begins to calculate the differences in these three criteria between the debunking passage produced in the current round and the one from the previous round. If $|\Delta S| < 0.1$, $|\Delta N| < 0.2$ < and $|\Delta R| < 0.05$, the iteration stops. Beginning with the general debunking passage, MURSE modifies it up to 10 rounds.

4.3 Effectiveness of MURSE (EQ1)

To assess the effectiveness of the user-simulation module and the iterative improvements in the MURSE framework, we evaluate MURSE's performance on all 24 profiles and compare it against several baseline methods, as shown in Table 2. The average values of S we calculate involve first averaging the values for each rumor, and then averaging the values across different profiles.

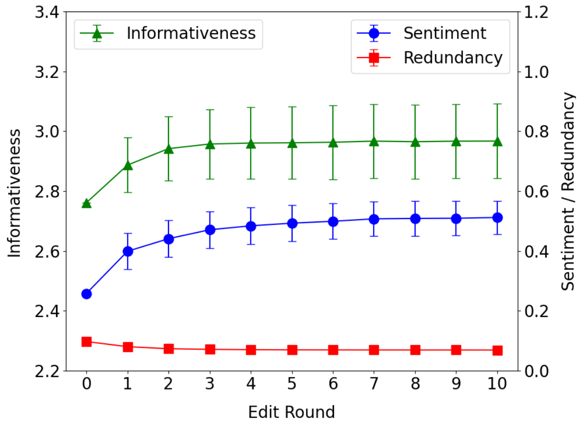


Figure 3: Three criteria changes with the iteration. Round 0 corresponds to the general debunking passage.

In comparison with the three abstractive-based methods, Gold Evidence exhibits notable shortcomings. On the one hand, it falls behind all abstractive-based methods in terms of Sentiment and Redundancy metrics, reflecting the advantages of abstractive-based approaches in generating emotionally balanced and less redundant content. On the other hand, Gold Evidence outperforms abstractive-based methods in Informativeness, likely due to the fact that sentences in Gold Evidence are directly extracted from human-written sources, which results in higher perplexity when evaluated by an open-source LLM.

Compared to other abstractive-based methods, MURSE demonstrates comprehensive superiority over the other two approaches, highlighting its effectiveness. Specifically, the performance of the Single-Round method surpasses that of the General method, indicating that the modification advice provided by the **User-Simulated Module** contributes significantly to enhancing the quality of debunking passages. Furthermore, MURSE outperforms the Single-Round method, underscoring the effectiveness of the **Iterative Improvement** module.

To further analyze the relationship between the **Iterative Improvement** module and the number of iterations, Fig. 3 illustrates the detailed changes in metrics after each round of modification. It shows that, starting from the 6th round onward, the metrics largely stabilize, indicating that the MURSE framework has reached its optimal capability for generating user-specific debunking passages.

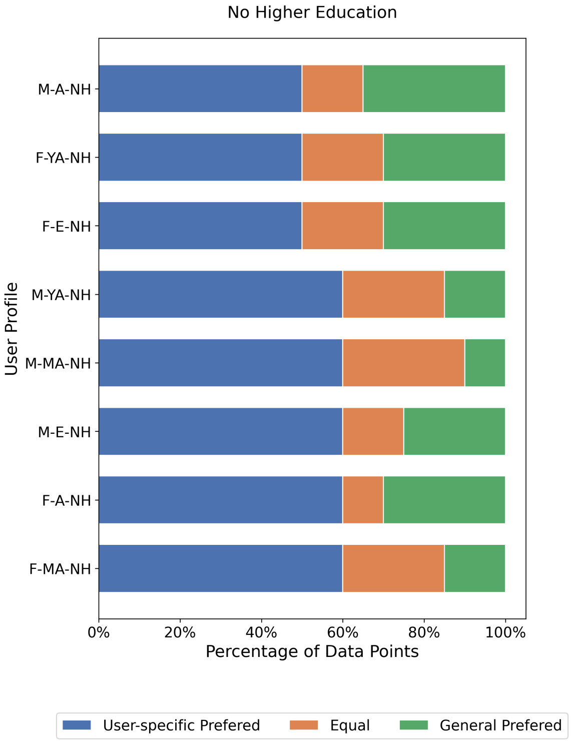


Figure 4: Results of human evaluation (No Higher Education). “Preferred” means that more annotators favored that debunking passage. “Equal” means that the number of participants who liked each of the two debunking passages was the same. (M-Male; F-Female; YA-Young Adult; E-The Elderly; MA-The Middle Aged; A-Adult; NH-has not graduated from college)

4.4 Evaluation of Debunking Passage for Target User Profile (EQ2)

To verify whether the user-specific debunking passages generated by the MURSE framework can attract the corresponding target profile users on social media platforms, we conducted a human evaluation experiment. For each profile, we recruited 20 eligible participants to provide annotations. In the questionnaire, the annotators were required to determine which debunking passage was more appealing. The prompt is shown in Appendix A. The display order of the options and questions is random to avoid the position bias.

A data point consists of a rumor, a user-specific debunking passage, and a general debunking passage. In Fig. 4, we can observe that for each profile, more than half of the data points indicate that the user-specific debunking passages are preferred. Additionally, the percentage of “User-specific Preferred” cases is 60% or higher in most profiles. The percentage of participants who preferred the general debunking passage generally did not exceed 30%. It should be noted that the evaluation pre-

Table 3: A Case in Male-Young Adult-College Graduate

Rumor: Foods containing additives are all bad; only natural, additive-free, and preservative-free foods are high-quality.
User-specific Debunking Passage: The Truth About Food Additives: Safety and Misconceptions You Need to Know Before Eating! Buddy, your perspective on food additives might need an update. The Red Bull in your hand or the PowerBar after your workout—they all rely on food additives. But are they really as bad as the internet claims?...
General Debunking Passage: From Me to You! It seems you have concerns about food additives, which is completely normal. However, not all foods containing additives are bad. When used properly, food additives can enhance the color, aroma, and taste of food, as well as extend its shelf life...

sented here is based on participants without higher education, while the complete evaluation results are provided in the Fig. 7. In Table 3, we can observe that user-specific debunking passages are more appealing.

5 Conclusion

This paper presents MURSE, a framework for combating misinformation through personalized debunking passage generation. MURSE generates user-specific debunking passages that are both context-sensitive and highly effective. Experimental results show that MURSE-generated passages are preferred twice as often as general debunking content, underscoring the importance of personalization in misinformation correction.

Acknowledgment

The authors would like to thank the anonymous reviewers for their insightful comments. This work is supported by the National Natural Science Foundation of China (62203425, 62406310).

Limitations

In this paper, we selected gender, age, and education as attributes to model the key characteristics of a user. However, these three attributes may not fully represent users, indicating limitations in user modeling depth. In the future, we plan to explore more effective modeling approaches. Furthermore, our framework was implemented by using multiple commercial and open-source LLMs, and we did not conduct an exhaustive model selection process. We will consider more economic and effective LLM integration solutions in future exploration.

Ethical Consideration

In this paper, we propose to tailor rumor-debunking passages for targeted user groups to improve the reading willingness and experience, which could contribute to the ultimate rumor-debunking effects to some extent and provide a new automatic solution based on large language models for improving social good.

In the human evaluation, we recruit annotators from a public third-party platform. During the evaluation, no private information that reveals personal identities is obtained by our team, and the annotators know and understand their rights and responsibilities by agreeing to the platform's user policy. By clearly describing the annotation task and provide author-verified rumor-debunking passages, we do our best to avoid any misleading materials individually during the annotation process. We do not receive any complaints about the task contents.

Due to the fact that large language models are trained on large-scale general corpora, it is inevitable that the commenters played by LLMs in our simulated feedback module would entail some common impressions of specific user groups. This somewhat benefits the tailoring of rumor-debunking passages, but also brings a potential that a specific person in the target group does not favor the output passages because of their unique preferences. We advocate deeper research in this direction to better shape such preferences to generate higher-quality rumor-debunking passages.

References

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating fact checking explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7352–7364, Online. Association for Computational Linguistics.

Melisa Basol, Jon Roozenbeek, and Sander van der Linden. 2020. Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of Cognition*.

Man-pui Sally Chan, Christopher R Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science*, 28(11):1531–1546.

Jonas De keersmaecker and Arne Roets. 2017. 'fake news': Incorrect, but hard to correct. the role of cognitive ability on the impact of false information on social impressions. *Intelligence*, 65:107–110.

Islam Eldifrawi, Shengrui Wang, and Amine Trabelsi. 2024. Automated justification production for claim veracity in fact checking: A survey on architectures and approaches. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6679–6692. Association for Computational Linguistics.

Joseph Gatto, Omar Sharif, and Sarah Preum. 2023. Chain-of-thought embeddings for stance detection on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4154–4161, Singapore. Association for Computational Linguistics.

Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. 2020. The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys*, 53(4).

Bing He, Mustaque Ahamad, and Srijan Kumar. 2023. Reinforcement learning-based counter-misinformation response generation: A case study of covid-19 vaccine misinformation. In *Proceedings of the ACM Web Conference 2023*, pages 2698–2709. Association for Computing Machinery.

Han He and Jinho D. Choi. 2021. The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577. Association for Computational Linguistics.

Beizhe Hu, Qiang Sheng, Juan Cao, Yang Li, and Danding Wang. 2025. Llm-generated fake news induces truth decay in news ecosystem: A case study on neural news recommendation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 435–445. Association for Computing Machinery.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *Preprint*, arXiv:2404.06395.

Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. 2022. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States. Association for Computational Linguistics.

Shan Jiang and Christo Wilson. 2018. Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23.

Kyungha Kim, Sangyun Lee, Kung-Hsiang Huang, Hou Pong Chan, Manling Li, and Heng Ji. 2024. Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate. *Preprint*, arXiv:2402.07401.

Emily Vraga Kjerstin Thorson and Brian Ekdale. 2010. Credibility in context: How uncivil online commentary affects news credibility. *Mass Communication and Society*, 13(3):289–313.

Neema Kotonya and Francesca Toni. 2020a. Explainable automated fact-checking: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5430–5443. International Committee on Computational Linguistics.

Neema Kotonya and Francesca Toni. 2020b. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754, Online. Association for Computational Linguistics.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131.

Aiwei Liu, Qiang Sheng, and Xuming Hu. 2024. Preventing and detecting misinformation generated by large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3001–3004. Association for Computing Machinery.

Jing Ma, Wei Gao, and Kam Fai Wong. 2018. Detect rumor and stance jointly by neural multi-task learning. In *Companion Proceedings of the Web Conference 2018*, pages 585–593. Association for Computing Machinery.

Yingchen Ma, Bing He, Nathan Subrahmanian, and Srijan Kumar. 2023. Characterizing and predicting social correction on twitter. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 86–95. Association for Computing Machinery.

Diana Micevičienė, Kara Lina Guokė, Jan Rajchel, et al. 2024. Fake news in the socio-economic environment in the context of the war in ukraine. *Central European Journal of Security Studies*, 2(1):97–104.

Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, page 1732–1742. Association for Computing Machinery.

Qiong Nan, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Guang Yang, and Jintao Li. 2025. Exploiting user comments for early detection of fake news prior to users' commenting. *Frontiers of Computer Science*, 19(10):1910354.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Gordon Pennycook and David G. Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188:39–50. The Cognitive Science of Political Thought.

Francesco Pierri, Brea L Perry, Matthew R DeVerna, Kai-Cheng Yang, Alessandro Flammini, Filippo Menczer, and John Bryden. 2022. Online misinformation is linked to early covid-19 vaccination hesitancy and refusal. *Scientific reports*, 12(1):5966.

Zhongyi Qiu, Hanjia Lyu, Wei Xiong, and Jiebo Luo. 2025. Can llms simulate social media engagement? a study on action-guided response generation. *Preprint*, arXiv:2502.12073.

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

Daniel Russo, Stefano Menini, Jacopo Staiano, and Marco Guerini. 2025. Face the facts! evaluating RAG-based pipelines for professional fact-checking. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 846–865, Hanoi, Vietnam. Association for Computational Linguistics.

Daniel Russo, Serra Sinem Tekiroğlu, and Marco Guerini. 2023b. Benchmarking the generation of fact checking explanations. *Transactions of the Association for Computational Linguistics*, 11:1250–1264.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 312–320.

Maryke S Steffens, Adam G Dunn, Kerrie E Wiley, and Julie Leask. 2019. How organisations promoting vaccination respond to misinformation on social media: a qualitative investigation. *BMC public health*, 19:1–12.

Yifan Sun, Danding Wang, Qiang Sheng, Juan Cao, and Jintao Li. 2025. Enhancing the comprehensibility of text explanations via unsupervised concept discovery. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14695–14713, Vienna, Austria. Association for Computational Linguistics.

Team GLM. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. DELL: Generating reactions and explanations for LLM-based misinformation detection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2637–2667. Association for Computational Linguistics.

Zhengjia Wang, Qiang Sheng, Danding Wang, Beizhe Hu, and Juan Cao. 2025a. Bridging thoughts and words: Graph-based intent-semantic joint learning for fake news detection. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 3250–3260. Association for Computing Machinery.

Zhengjia Wang, Danding Wang, Qiang Sheng, Juan Cao, Siyuan Ma, and Haonan Cheng. 2025b. Exploring news intent and its application: A theory-driven approach. *Information Processing & Management*, 62(6):104229.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2608–2621. International Committee on Computational Linguistics.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5628–5643, Mexico City, Mexico. Association for Computational Linguistics.

Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, pages 3465–3476. Association for Computing Machinery.

A Prompt Templates

We list the five prompt templates used in the MURSE framework as follows:

Prompt 1: General Debunking Passage Generation Prompt

System Prompt: You are a writer of refutation summaries. Based on evidences, debunk the rumor. Return only the refutation summary.
Context Prompt: #Rumor: [*rumor*] #evidences: [*evidences*]

Prompt 2: Commenter Prompt

System Prompt: You are [*gender*]. Your education level is [*education level*]. Your age is [*age*]. You now believe this news. Please comment on it in the style of social media.
Context Prompt: #News: [*rumor*]

Prompt 3: Advisor Prompt

System Prompt: You are an advisor of refutation summaries. Based on the rumor and its comments, provide revision suggestions for the refutation summary to better align with the target audience. Return only the revision suggestions.
Context Prompt: #Rumor: [*rumor*] #Evidences: evidences: [*evidences*] #debunking passage: [*debunking passage*] #Target Audience Gender: [*gender*] Education Level: [*education level*] Age: [*age*]

Prompt 4: Editor Prompt

System Prompt: You are a refutation summary editor. Modify the refutation summary based on the provided revision suggestions. Return only the revised refutation summary.
Context Prompt: #Rumor: [*rumor*] #Evidences: evidences: [*evidences*] #debunking passage: [*debunking passage*] #Target Suggestions: [*feedback*]

Prompt 5: Questionnaire Prompt

While browsing the news, you come across the following headline: [*rumor*] According to feedback, this news contains misinformation. Which of the following replies would catch your attention at first glance?
[*user-specific debunking passage*]
[*general debunking passage*]

B Relationship between Profile Attributes and Criteria

To quantitatively assess the personalization capability of MURSE, we analyzed the correlation between configured user profile attributes and the corresponding automatic evaluation scores, as summarized in Table 4. The systematic discrepancies in the results across different demographic and behavioral profiles provide concrete evidence that our framework does not generate generic responses. Instead, it successfully produces tailored debunking passages that are adapted to the specific attributes of each user profile. This data-driven validation confirms that the personalization mechanisms within MURSE are functionally effective, enabling it to modulate various aspects of the generated text—such as tone, framing, or evidence selection—in response to different user contexts.

Table 4: The relationship between Profile Attributes and Criteria.

Profile Attribute	**Sentiment**	**Informativeness**	**Redundancy**
Gender			
Male	0.4676	2.9463	0.0705
Female	0.5560	2.9876	0.0675
Age			
Adolescent	0.5099	3.0466	0.0679
Young Adult	0.5195	3.0547	0.0686
The Middle-Aged	0.5009	2.8902	0.0693
The Elderly	0.5170	2.8764	0.0702
Education			
High School or Less	0.5416	3.0232	0.0692
Has Not Graduated From College	0.5127	3.0205	0.0685
A College Graduate	0.4813	2.8573	0.0694
Avg.	0.5118	2.9670	0.0690

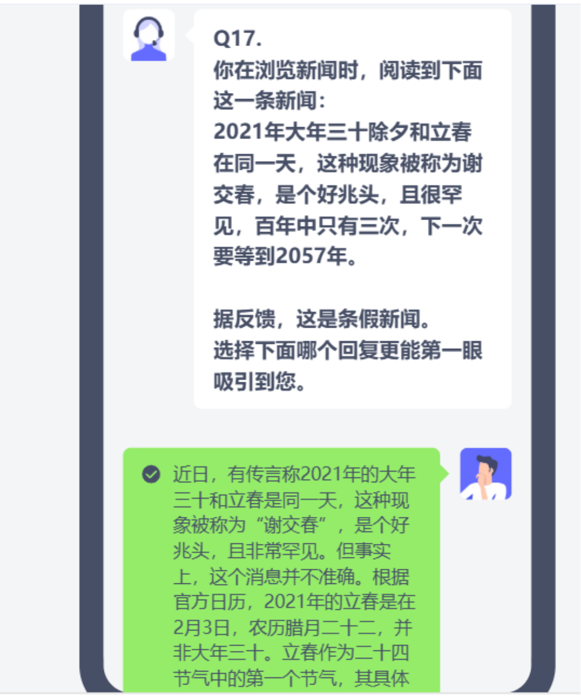
C Questionnaire Platform

Regarding evaluation metrics, our findings demonstrate that three criteria exhibit strong consistency with human cognitive judgments. This correlation was validated through our user study conducted via the Fengling platform[1]. The compensation is ¥4.15 per response at least. The threshold settings were also determined empirically. Our questionnaire is accessible on both mobile and desktop platforms. The interface and demo scenarios are illustrated in Fig. 5 and Fig. 6, respectively.

D More Information About Dataset and Baselines

For industrial deployment considerations, we exclusively evaluated our approach on the CHEF dataset for these key reasons: Our system primarily serves Chinese-language applications. Existing baselines are not directly comparable for this specific task, and the dataset provides gold-standard evidence sentences that establish an upper bound for extractive approaches (concatenated gold evidence serves as the extractive ceiling). In our experiments, the

[1] Fengling crowdsourcing platform: https://www.powercx.com/product



Q17.
你在浏览新闻时，阅读到下面这一条新闻：
2021年大年三十除夕和立春在同一天，这种现象被称为谢交春，是个好兆头，且很罕见，百年中只有三次，下一次要等到2057年。

据反馈，这是条假新闻。
选择下面哪个回复更能第一眼吸引到您。

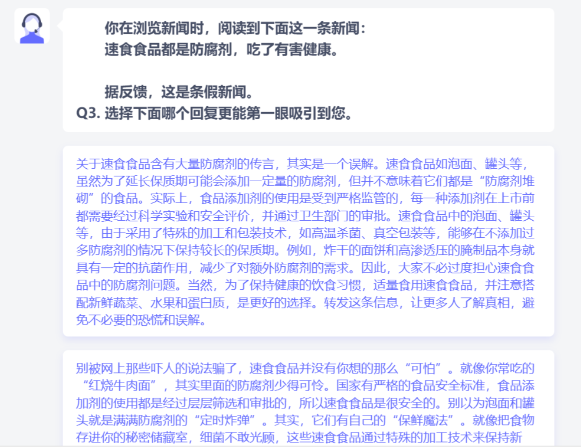
近日，有传言称2021年的大年三十和立春是同一天，这种现象被称为“谢交春”，是个好兆头，且非常罕见。但事实上，这个消息并不准确。根据官方日历，2021年的立春是在2月3日，农历腊月二十二，并非大年三十。立春作为二十四节气中的第一个节气，其具体

Figure 5: Questionnaire on the mobile client (in Chinese)

baseline strategies employed for comparison are all variants derived from this particular dataset.

E Analysis of Demographic Preference Distributions

From Fig. 7, we observe a consistent pattern: User-specific Preferred responses dominate across nearly all groups, indicating that personalized content is generally more favored than generic alternatives. In contrast, the proportion of General Preferred responses remains relatively small, though it varies noticeably across profiles, suggesting that certain user groups are more tolerant of non-personalized outputs. The Equal category shows the largest fluctuation among profiles, reflecting differences in how strongly various demographic groups distinguish between personalized and general content. Notably, the overall shapes of the male and female subgroups are highly similar, implying that gender itself is not the primary determinant of preference patterns; rather, variations are more pronounced along dimensions such as age and education level. These results collectively highlight that personalization exerts a robust and consistent influence on user preference, while demographic attributes modulate the degree to which users discriminate between personalized and general responses.



你在浏览新闻时，阅读到下面这一条新闻：
速食食品都是防腐剂，吃了有害健康。

据反馈，这是条假新闻。
Q3. 选择下面哪个回复更能第一眼吸引到您。

关于速食食品含有大量防腐剂的传言，其实是一个误解。速食食品如泡面、罐头等，虽然为了延长保质期可能会添加一定量的防腐剂，但并不意味着它们都是“防腐剂堆砌”的食品。实际上，食品添加剂的使用是受到严格监管的，每一种添加剂在上市前都需要经过科学实验和安全评价，并通过卫生部门的审批。速食食品中的泡面、罐头等，由于采用了特殊的加工和包装技术，如高温杀菌、真空包装等，能够在不添加过多防腐剂的情况下保持较长的保质期。例如，炸干的面饼和高渗透压的腌制品本身就具有一定的抗菌作用，减少了对额外防腐剂的需求。因此，大家不必过度担心速食食品中的防腐剂问题。当然，为了保持健康的饮食习惯，适量食用速食食品，并注意搭配新鲜蔬菜、水果和蛋白质，是更好的选择。转发这条信息，让更多人了解真相，避免不必要的恐慌和误解。

别被网上那些吓人的说法骗了，速食食品并没有你想的那么“可怕”。就像你常吃的“红烧牛肉面”，其实里面的防腐剂少得可怜。国家有严格的食品安全标准，食品添加剂的使用都是经过层层筛选和审批的，所以速食食品是很安全的。别以为泡面和罐头就是满满防腐剂的“定时炸弹”。其实，它们有自己的“保鲜魔法”。就像把食物存进你的秘密储藏室，细菌不敢光顾，这些速食食品通过特殊的加工技术来保持新

Figure 6: Questionnaire on the desktop client (in Chinese)

F Faithfulness Verification

To assess the factual consistency of the user-specific debunking passages produced by the editor, we evaluated their faithfulness with respect to the provided evidence. We employed GPT-4o (OpenAI, 2024) to rate each debunking passage on a 1–5 scale, where higher scores indicate stronger alignment with the ground-truth evidence. Table 5 reports the distribution of scores across different user groups. Overall, all user profiles achieve high average faithfulness scores of more than 4.00, suggesting that the editor-edited debunking passages maintain strong factual grounding regardless of user profile.

G Latency and Cost

Latency In our experiments, a single iteration of MURSE completes in approximately 12 seconds. A full run of 10 iterations has a total latency of only 2 minutes. Furthermore, the framework exhibits high scalability, as the core LLM inference step (using the GLM model) can efficiently handle concurrency levels ranging from 50 to 500.

Cost We utilize the GLM-4-Air model, which is priced at 5 RMB per million tokens. Each iteration of MURSE consumes approximately 4,000 tokens. Therefore, a complete 10-iteration run consumes about 40,000 tokens. The cost in RMB is calculated as: (40,000 tokens / 1,000,000 tokens) × 5 RMB = 0.2 RMB ($0.028 USD). This shows that our framework can run using cost-competitive LLMs and has potential to scale up when the requests increase.

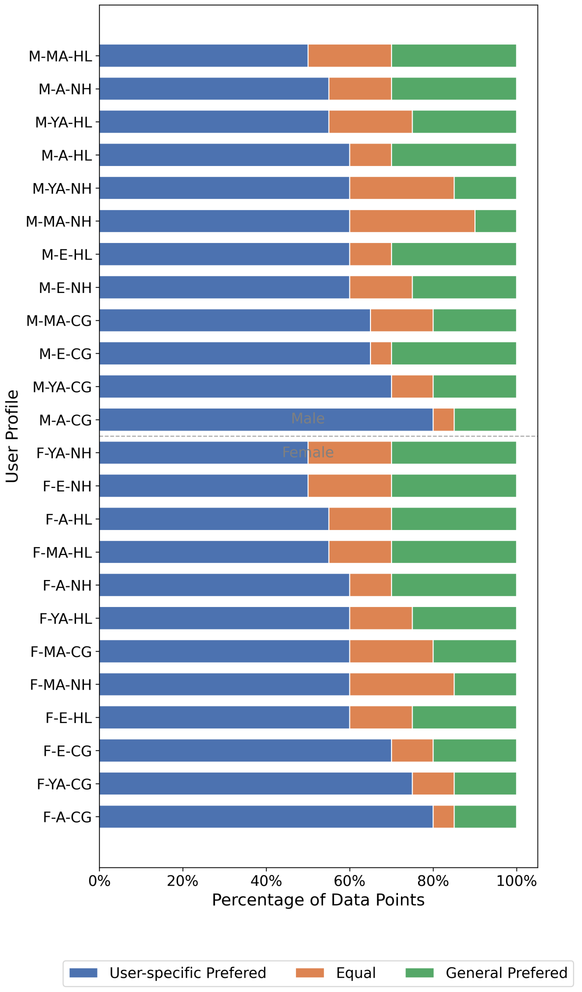


Figure 7: Results of all human evaluations

Table 5: Faithfulness score distribution (1–5) across all demographic groups.

User Profile	**5**	**4**	**3**	**2**	**1**	**Avg**
M-A-CG	87	147	23	8	5	4.12
M-A-HL	76	145	38	6	5	4.05
M-A-NH	78	150	36	6	3	4.06
M-YA-CG	88	146	24	7	5	4.13
M-YA-HL	88	144	27	3	7	4.13
M-YA-NH	84	148	33	5	3	4.07
M-MA-CG	86	145	26	7	6	4.10
M-MA-HL	78	146	34	5	6	4.07
M-MA-NH	77	149	38	6	3	4.05
M-E-CG	85	147	27	6	5	4.10
M-E-HL	77	147	38	5	3	4.07
M-E-NH	78	148	37	5	2	4.07
F-A-CG	86	149	20	9	5	4.13
F-A-HL	75	154	23	8	4	4.11
F-A-NH	76	155	29	7	3	4.08
F-YA-CG	89	148	19	9	5	4.14
F-YA-HL	76	159	24	7	3	4.11
F-YA-NH	77	158	30	7	2	4.09
F-MA-CG	86	147	26	6	5	4.12
F-MA-HL	74	155	28	6	4	4.10
F-MA-NH	75	154	33	7	4	4.06
F-E-CG	85	149	27	6	5	4.11
F-E-HL	75	156	26	6	4	4.11
F-E-NH	75	152	31	7	3	4.07