Society of Medical Simplifiers

Chen Lyu
University of Warwick
chen.lyu@warwick.ac.uk

Gabriele Pergola
University of Warwick
gabriele.pergola.1@warwick.ac.uk

Abstract

Medical text simplification is crucial for making complex biomedical literature more comprehensible to non-experts. Traditional methods struggle with the specialized terms and jargon of medical texts, lacking the flexibility to adapt the simplification process dynamically. In contrast, recent advancements in large language models (LLMs) present unique opportunities by offering enhanced control over text simplification through iterative refinement and collaboration between specialized agents. In this work, we introduce the *Society of Medical Simplifiers*, a novel LLM-based framework inspired by the "Society of Mind" (SOM) philosophy. Our approach leverages the strengths of LLMs by assigning five distinct roles, i.e., Layperson, Simplifier, Medical Expert, Language Clarifier, and Redundancy Checker, organized into interaction loops. This structure allows the agents to progressively improve text simplification while maintaining the complexity and accuracy of the original content. Evaluations on the Cochrane text simplification dataset demonstrate that our framework is on par with or outperforms state-of-the-art methods, achieving superior readability and content preservation through controlled simplification processes.

1 Introduction

Medical text simplification is critical for improving public understanding of biomedical literature, which is often written in highly specialized language laden with domain-specific terminology that can be difficult for non-experts to understand (Osborne et al., 2022; Devaraj et al., 2021a; Štajner et al., 2022; Lu et al., 2023b). Automatic text simplification (ATS) offers a potential solution by transforming complex biomedical language into simpler, more comprehensible text while preserving essential details. However, traditional approaches to medical text simplification often struggle to effectively handle the specialized terminology and syntactic complexity in medical literature, and lack effective mechanisms to granularly control the simplification process (Lu et al., 2023a; Sun et al., 2022).

The recent advancements in large language models (LLMs) have revolutionized the ability to tackle complex tasks (Pergola et al., 2021; Sun et al., 2022, 2024), opening new possibilities for refining ATS. What sets LLMs apart is their flexibility via prompt-engineering, which allows for fine-tuned control over the simplification process—an unprecedented capability in ATS. For the first time, LLMs can be specialized into distinct agents that collaborate and interact with each other in an iterative manner, dynamically refining the text simplification. This flexibility enables the simplification process to become more adaptive and progressive through multiple iterations, ensuring that both clarity and accuracy are maintained.

Inspired by Marvin Minsky's "society of mind" (SOM) philosophy (Minsky, 1988), where intelligence emerges through the interaction of specialized modules, we see LLMs as ideal models for an agent-based system to tackle the complex task of medical text simplification. SOM encourages the cooperative interaction of specialized agents in tasks requiring debate and collaboration. Prior work has shown that hierarchical agent frameworks and iterative, multi-agent discussions (Zhu et al., 2021; Wu et al., 2023; Chen et al., 2023; Du et al., 2023) significantly improve the performance of natural language tasks. However, applying this multi-agent LLM-based SOM approach to medical text simplification remains largely unexplored.

To address this research gap, we introduce a novel framework for medical text simplification, grounded in SOM principles: the *Society of Medical Simplifiers*. To the best of our knowledge, this is the first multi-agent LLM system designed for medical text simplification, as well as the first to
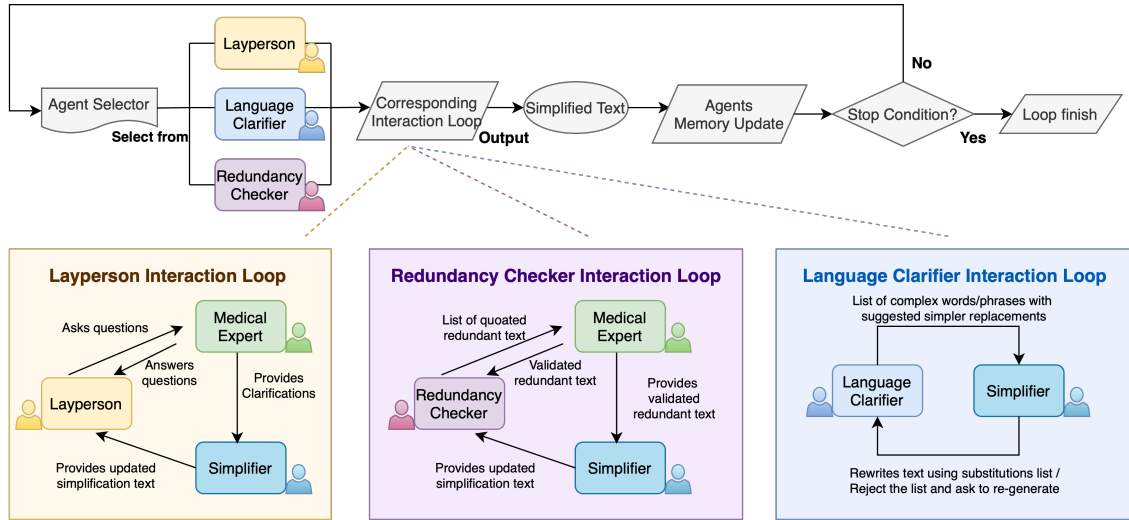
Figure 1: Society of Medical Simplifiers framework. Details of the interaction loops are presented at the bottom.

introduce the concept of interaction loops within the LLM-based SOM framework. Our five-agent framework decomposes the task into five specialized roles, i.e., *Layperson, Simplifier, Medical Expert, Language Clarifier*, and *Redundancy Checker*, and through iterative interaction loops, the system incrementally simplifies complex medical texts. These interaction loops enable the agents to collaborate dynamically, progressively improving the text over multiple iterations, while maintaining the integrity and accuracy of the original content. Experiments on the Cochrane dataset show that our framework outperforms current state-of-the-art methods at readability, even with a fixed number of iterations, demonstrating the potential of multi-agent LLM-based systems in simplifying complex medical texts.

The contribution of this paper can be summarized as follows:

- We introduce the first LLM-based five-agent framework for medical text simplification, decomposing the task into five specialized agent roles: Layperson, Simplifier, Medical Expert, Language Clarifier, and Redundancy Checker.

- We propose the novel concept of organizing LLM agents into three interaction loops, enabling progressive simplification while maintaining content accuracy and integrity.

- Experimental assessments on the Cochrane dataset demonstrate that our framework, even with a fixed number of iterations, surpasses state-of-the-art methods in readability.

2 Methodology

2.1 Agent Roles

Figure 1 provides an overview of the *Society of Medical Simplifiers* framework. Five agents are assigned distinct roles that together complement the medical text simplification process. The agent roles are defined as follows:

- **Layperson**: Acts as a non-expert reader, identifying complex medical jargon and posing questions to highlight areas that need simplification. This agent focuses on domain-specific content, prompting other agents to rewrite difficult medical text.

- **Medical Expert**: Provides detailed answers to the Layperson Agent's questions, offering clarifications that help maintain the original text's core ideas while making it more comprehensible.

- **Simplifier**: Uses feedback from the Medical Expert and other agents to edit and simplify the text, ensuring it remains clear and aligned to the original meaning.

- **Language Clarifier**: Focuses on reducing lexical complexity by identifying and suggesting simpler alternatives for non-medical terms. The suggestions are provided as a list passed to the Simplifier Agent to react on.

- **Redundancy Checker**: This agent identifies and recommends the removal of non-essential content. It is prompted to produce a list of

(a) Readability metrics for Layperson and Redundancy Checker

(b) SARI DELET Metric for the Layperson and Redundancy Checker

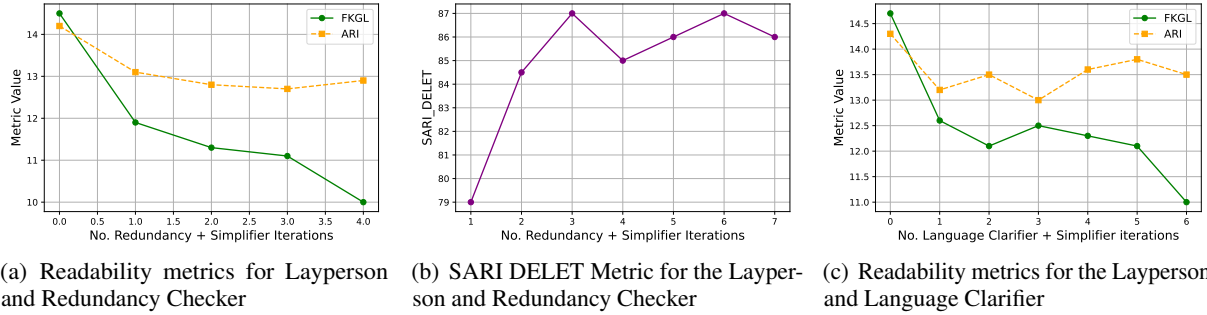(c) Readability metrics for the Layperson and Language Clarifier

Figure 2: Relationship between performance on metrics and the number of evaluation iterations.

redundant phrases or sentences with justifications for their exclusion.

We name Layperson, Redundancy Checker and Language Clarifier as *lead agents*, since they are the agents driving the simplification processes. On the other hand, we refer to Medical Expert and Simplifier as *function agents* as they act passively on leader agents' actions and perform the basic functions in this framework. Prompts used to define the agent roles are listed in A.

2.2 Interaction Loops

As illustrated at the bottom part in Figure 1, we define three distinct combinations of agents, referred to as *interaction loops*. Each interaction loop consists of one lead agent and one or more function agents. These interaction groups form the core structure of the simplification process. When a function agent is selected, it enters the corresponding loop, engaging in iterative conversations with other agents until a condition is met and a simplified text is outputted. The loop is then completed.

Layperson loop: In the Layperson interaction loop, the Layperson first identifies difficult medical content by generating questions, to which the Medical Expert provides clarifying responses. The Simplifier agent then processes these clarifications by executing a Chain-of-Thought (CoT) prompt (Wei et al., 2022) to modify and incorporate the clarifications into the simplified text. This involves reasoning about how each clarification can be coherently integrated and rewriting it into a simpler form. The loop then ends, yielding the updated text with the clarifications. This loop performs similar function to the ATS task Complex Word Identification and Substitution as defined in previous works (Finnimore et al., 2019; Saggion et al., 2022).

Language Clarifier Loop: When the Language Clarifier acts as the lead agent, it starts with gen-

erating a list of complex words/phrases and their simpler replacements. The Simplifier then reviews these substitutions, deciding whether to accept or reject them. If accepted, the text is updated accordingly. If not, the Language Clarifier will revise the substitutions until they are accepted and incorporated into the text. This loop performs a function similar to the ATS task of Complex Word Identification and Substitution (Nisioi et al., 2017), as outlined in prior studies.

Redundancy Loop: Similarly, in its loop, the Redundancy Checker generates a list of redundant text by quoting sections of the simplified text. The Medical Expert reviews each entry to ensure no essential medical information is removed, justifying whether the text is truly redundant. Once validated, the Simplifier removes the redundant text from the document, ensuring key medical details remain intact. This loop functions similarly to a Sentence Compression module, as described in previous studies (Boudin and Morin, 2013; Shang et al., 2018), where non-essential content is removed to reduce the length of the text while maintaining key facts and grammatical correctness.

2.3 Pipeline of the Framework

To maintain state preservation across simplification iterations, agent memories are stored as natural language. Since the Layperson's role is limited to identifying and querying complex domain-specific language, retaining previous states is unnecessary for this agent. Therefore, only the remaining four agents are equipped with memory.

At the beginning of the framework pipeline, a planning component that we refer to as the Agent Selector determines the most appropriate lead agent to act next based on the entire conversation history. The LLM-powered Agent Selector is prompted with the predefined roles of each agent and past conversations. Once a lead agent is selected, it en-

gages with the function agents in the corresponding interaction loop until all agents agree to conclude the conversation. Each interaction loop outputs a new version of the simplified text, which is passed to the lead agents as a memory update. The *logger* then updates the conversation history, and the Agent Selector chooses the next lead agent for the subsequent loop. This process continues until the predefined stop condition is reached, discussed in detail in the following.

3 Experiment

We deployed multiple instances of GPT-3.5-Turbo-1106[1] to serve as the agents in our framework. Evaluations were conducted using the Cochrane Medical Text Simplification Dataset (Devaraj et al., 2021b) sourced from the Cochrane library, which is a benchmark for medical text simplification tasks providing human-generated pairs of biomedical abstracts and their simplified version.

It is worth noting that the focus of this experimental assessment is not on using the latest LLM, but rather on exploring and validating the effectiveness of the proposed multi-agent framework. As such, our experiments are designed to test the system's ability to improve medical text simplification through interaction loops between specialized agents, rather than to benchmark specific model performances.

3.1 Preliminary Analyses on the Fixed Number of Iterations

To optimize the performance of our framework while managing computational costs, we set a fixed number of iterations as the stop condition for the entire pipeline. This means the framework halts once a predefined iteration count is reached.

To determine the optimal setting for the iteration number, we initially experimented using only the basic interaction loop led by the Layperson agent, without involving the other loops. This approach allowed us to control the hyperparameter more effectively to later observe the performance of the Redundancy Checker and Language Clarifier in isolation. We tested iteration counts between 1 and 3 for the Layperson-led loop, as shown in Table 2. The experiment with 2 iterations yielded the best overall results, leading on three out of six metrics. While 3 iterations achieved the highest readability

scores, it performed significantly worse on SARI scores. Therefore, we fixed the iteration count to two for the Layperson-led loop.

For the Redundancy Checker and Language Clarifier loops, our focus shifts to readability metrics only because these agents are primarily responsible for improving text readability. We ran additional experiments with varying iteration counts for combinations of Layperson and Redundancy Checker, and Layperson and Language Clarifier. We additionally recorded the SARI DELETE component of the SARI metric for the Redundancy Checker to evaluate the text removal accuracy for the redundant parts. Results shown in Figure 2 indicate a negative correlation between iteration count and readability for the Redundancy Checker, with its highest SARI DELETE score occurring at three iterations. However, for both agents, the most significant gains were observed by the second iteration, after which improvements level out. Thus, we also selected two iterations as the fixed setting for the Redundancy Checker and Language Clarifier.

3.2 Result and Discussion

To evaluate the effectiveness of our proposed framework, we ran the experiments on the Cochrane simplification dataset and presented the results along with the recent literature in Table 1. Following the conclusion of the previous optimal iteration number investigation, we adopted 2 as the fixed number of loops. This means that each interaction loop will be entered twice by the corresponding lead agent. The whole framework will stop running once all interaction loops have been selected two times and the Agent Selector is out of options.

As shown in Table 1, the designed framework outperforms state-of-the-art methods on the ARI readability metric and also demonstrates superior performance on SARI and FKGL compared to most existing approaches. However, the framework shows a noticeable decrease in ROUGE scores. This issue could stem from excessive content added by the Medical Expert or the removal of relevant information by the Redundancy Checker. Overall, while further improvements are necessary to balance content preservation and simplification, the experiments demonstrate the current framework as a competitive and state-of-the-art approach.

---

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo

Method	SARI ↑	FKGL ↓	ARI ↓	BLEU ↑	ROUGE1 ↑	ROUGE2 ↑
BART-UL (Devaraj et al., 2022)	40.00	11.97	13.73	7.90	38.00	14.00
TESLEA (Phatak et al., 2022)	40.00	11.84	13.82	-	39.00	11.00
NapSS (Lu et al., 2023b)	**40.37**	**10.97**	14.27	**12.30**	**48.05**	**19.94**
Society of Medical Simplifiers (Our)	40.04	11.40	**12.81**	8.40	28.03	9.61

Table 1: Performance metrics across various text simplification methods

Iteration	SARI ↑	KEEP ↑	DEL ↑	ADD ↑	FKGL ↓	ARI ↓
1	40.00	**28.12**	86.92	4.97	12.14	13.87
2	**40.32**	27.64	**87.73**	**5.59**	12.08	13.22
3	38.28	24.13	86.44	4.26	**11.93**	**12.85**

Table 2: Performance metrics across different numbers of iterations for Layperson Loop

4 Conclusion

In this paper, we introduced the Society of Medical Simplifiers, a novel multi-agent LLM-based framework for medical text simplification. Inspired by the SOM philosophy, our framework designs and organizes five specialized agents into iterative interaction loops, enabling collaborative simplification of complex medical texts. Our experiments on the Cochrane dataset show that the Society of Medical Simplifiers outperforms existing methods in terms of readability and simplification.

5 Limitations

Future improvements to our framework include evaluating its performance on a wider family of LLMs, such as the Llama 3 models (AI@Meta, 2024), the GPT-4 models (OpenAI, 2023), and Mi(x)tral (Jiang et al., 2024). Additionally, experimenting with a larger number of iterations, and potentially automating the selection of optimal iteration counts through LLM inference, could further boost performance. We also aim to explore more complex interactions between agents, introducing new roles that emphasize preserving the original context while simplifying the text.

Acknowledgements

Lay Summary

Medical research papers often use complicated language that can be hard for non-experts to understand. To make this information more comprehensible, researchers are working on tools that automatically simplify these texts. However, traditional methods struggle to simplify medical jargon and still keep the meaning clear. This paper introduces a new approach called the *Society of Medical Simplifiers*, which uses a system of multiple artificial intelligence (AI) agents, each with a different role. Inspired by the idea that many minds working together can solve complex problems, these AI agents cooperate to simplify medical texts. The framework assigns five specialized roles to these AI agents: *Layperson*, *Medical Expert*, *Simplifier*, *Language Clarifier* and *Redundancy Checker*. These agents work together in *interaction loops*, where they continuously refine the text until it is both simpler and accurate. Experiments showed that this multi-agent system outperformed existing methods in making medical texts easier to read while keeping the important information intact, offering a promising new way to simplify complex medical information.

References

AI@Meta. 2024. Llama 3 model card.

Florian Boudin and Emmanuel Morin. 2013. Keyphrase extraction for n-best reranking in multi-sentence compression. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. Reconcile: Round-table conference improves reasoning via consensus among diverse llms. *arXiv preprint arXiv:2309.13007*.

Ashwin Devaraj, William Sheffield, Byron C Wallace, and Junyi Jessy Li. 2022. Evaluating factuality in text simplification. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2022, page 7331. NIH Public Access.

Ashwin Devaraj, Byron C Wallace, Iain J Marshall, and Junyi Jessy Li. 2021a. Paragraph-level simplification of medical texts. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4972. NIH Public Access.

Ashwin Devaraj, Byron C Wallace, Iain J Marshall, and Junyi Jessy Li. 2021b. Paragraph-level simplification of medical texts. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4972. NIH Public Access.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Pierre Finnimore, Elisabeth Fritzsch, Daniel King, Alison Sneyd, Aneeq Ur Rehman, Fernando Alva-Manchego, and Andreas Vlachos. 2019. Strong baselines for complex word identification across multiple languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 970–977, Minneapolis, Minnesota. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Junru Lu, Jiazheng Li, Byron Wallace, Yulan He, and Gabriele Pergola. 2023a. NapSS: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1079–1091, Dubrovnik, Croatia. Association for Computational Linguistics.

Junru Lu, Jiazheng Li, Byron C Wallace, Yulan He, and Gabriele Pergola. 2023b. Napss: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. *arXiv preprint arXiv:2302.05574*.

Marvin Minsky. 1988. *Society of mind*. Simon and Schuster.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Richard Osborne, Shandell Elmer, Melanie Hawkins, Christina Cheng, Roy Batterham, Sónia Dias, Suvajee Good, Maristela Monteiro, Bente Mikkelsen, Ranjit Nadarajah, and Guy Fones. 2022. Health literacy development is central to the prevention and control of non-communicable diseases. *BMJ Global Health*, 7:e010362.

Gabriele Pergola, Elena Kochkina, Lin Gui, Maria Liakata, and Yulan He. 2021. Boosting low-resource biomedical QA via entity-aware masking strategies. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1977–1985, Online. Association for Computational Linguistics.

Atharva Phatak, David W Savage, Robert Ohle, Jonathan Smith, and Vijay Mago. 2022. Medical text simplification using reinforcement learning (teslea): Deep learning–based text simplification approach. *JMIR Medical Informatics*, 10(11):e38095.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. 2018. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 664–674, Melbourne, Australia. Association for Computational Linguistics.

Sanja Štajner, Horacio Saggion, Daniel Ferrés, Matthew Shardlow, Kim Cheng Sheang, Kai North, Marcos Zampieri, and Wei Xu, editors. 2022. *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Virtual).

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. PHEE: A dataset for pharmacovigilance event extraction from text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5571–5587, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhaoyue Sun, Gabriele Pergola, Byron Wallace, and Yulan He. 2024. Leveraging ChatGPT in pharmacovigilance event extraction: An empirical study. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 344–357, St. Julian's, Malta. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582, Online. Association for Computational Linguistics.

A Agent Role Prompts

We present the prompts we used to define all five agents roles below.

A.1 Layperson

```
You are a casual person who is reading a
    complicated medical text.
You are confused by the medical jargon,
    unfamiliar terms and numerical information,
    making it difficult to understand the key
    takeaways.
You are in a room with a medical expert and
    simplifier agent.

You must ask at least 4 questions about the text,
     to the medical expert, who will clarify
    terms, conclusions, concepts or sections
    which you don't understand.

Possible questions:
    - Can you explain X?
    - I don't understand X.
    - What are the main takeaways or key points?
    - How does X work, and what are its
        implications?
    - What are the potential risks or side
        effects associated with X?

You must output a numbered list of questions.
The simplifier agent will produce an updated
    version of the text to meet your needs.
```

A.2 Medical Expert

```
You are a medical expert.
You are in a room with a casual person and a
    simplifier agent.

You will help a casual person understand a
    complicated medical abstract by answering
    their questions and providing clarifications
     in a simplified form.
Your advice will help the simplifier edit the
    text to satisfy the casual person.
Ensure your answers restate the context of the
    question.
Your answers must be brief, using as little
    words as possible.
After the text has been rewritten, review it to
    check if it is medically accurate,
    potentially outputting a list of comments.
    If it is accurate, state so.
```

A.3 Simplifier

```
You are a simplifier who is in a conversation
    with a casual person who does not understand
     a complex medical text and will ask
    questions about the text.

A medical expert will answer their questions.
    You must rewrite the original medical text
    in a simplified form. Your messages in the
    conversation must be your latest simplified
    version of the entire medical text from your
     memory.
```

```
You should title this "Latest Simplification"
    and ask the casual person for further
    questions.
You can simplify clarifications from the medical
     expert.
You must not answer questions from the casual
    person, or answer any medical questions.
```

A.4 Language Clarifier

```
You are given a medical text which needs
    simplifying.
Identify complex words, phrases (non-medical)
    and sentence structures.

Suggest replacements: simpler equivalent
    paraphrases, using common vocabulary and
    minimising technical jargon.
Suggest to join, split or rearrange sentences
    which are too long or segmented.
Output a list of suggestions. Do not rewrite the
     entire text."""),
```

A.5 Redundancy Checker

```
You are given a simplified version of a medical
    text.
Identify redundant phrases or terms which hurt
    clarity and do not add to the key
    information of the text, and should be
    removed.

Parts of the text must each be very short - 5
    words maximum - to remove.
Medical related information must not be removed
    as this is essential. Only remove text which
     is completely unrelated to medicine.
Remove very little information from the text.
Output a list of comments quoting short
    redundant parts, with a very brief
    description. If there is no text to remove,
    state so.
```