

Compact Language Models with Iterative Text Refinement for Health Dialogue Summarization

Kellen Tan Cheng^{♣*}, Ganesh Ramesh^{◇*}, Nafiul Rashid[♣],
Geoffrey Jay Tso[♣], Jilong Kuang[♣]

[♣]Microsoft [♣]Samsung Research America [◇]Nutanix

[♣]kellencheng@microsoft.com, [◇]ganesh.ramesh@nutanix.com

[♣]{n.rashid, geoff.tso, jilong.kuang}@samsung.com

Abstract

Health wellness agents typically rely on large language models (LLMs) for response generation, where contextual information from a user’s conversation history can be used for response grounding and personalization. High-quality conversation summaries are one such method which can reduce the number of input tokens during response generation, decreasing overhead and inference latency. However, directly purposing LLMs for this task is infeasible due to the scale of the task, the compute overhead, and health data compliance regulations. Furthermore, ground truth for real-world datasets is scarce due to privacy concerns and the high cost of health expert annotators. These factors necessitate the development of small, potentially on-device, language models capable of health dialogue summarization, particularly in the absence of ground truth labels. In this paper, we first present a comprehensive empirical study that benchmarks a variety of state-of-the-art smaller language models to better understand their baseline capabilities. Second, we present an unsupervised method that uses the summaries from multiple models, refined with iterative feedback, to generate high-quality summaries of health dialogues. Experiments illustrate that our method is able to outperform baseline on both open-source and proprietary benchmarks. Notably, our method can be run viably on local compute without a GPU, using just a single Macbook with 16 GB of memory.

1 Introduction

Recent advancements in large language models (LLMs) have spurred the development of health assistants capable of automating clinical workflows, generating diagnoses and advice, outputting user-personalized responses, acting as voice assistants, summarizing health contexts, and more (Qiu et al., 2024; Abbasian et al., 2024; Yang et al., 2024a;

Liu and Wang, 2023; Yang et al., 2024b; Singhal et al., 2023; Kim et al., 2024; Cosentino et al., 2024; Mukherjee et al., 2024). For effective health wellness agents, contextual information from prior conversations can be utilized to help ground and personalize the agent’s response to the user. Since appending all prior conversations during response generation can be costly, high-quality conversation summaries can be used to succinctly represent the same information at a lower token cost, reducing memory overhead and response latency. As a result, these summaries are an important component in the development of effective health wellness assistants.

Despite the impressive capabilities of state-of-the-art LLMs, directly adopting them for health conversation summarization presents several challenges. First, these models incur substantial memory compute overhead, and introduce latency and monetary cost considerations with their usage, on top of increased carbon emissions. Second, health data compliance regulations may restrict, or even ban, the ability to send user health information to externally hosted LLMs to perform health dialogue summarization. Such factors raise a potential question: *can we perform health conversation summarization without LLMs, using models that can either be ran on-device or hosted locally?*

Towards this objective, we first benchmark a suite of smaller language models (LMs) to gauge and analyze their baseline capabilities for health dialogue summarization. Instead of blindly selecting a model to use, we evaluate a variety of state-of-the-art smaller LMs, ranging from 0.5B up to 4B parameters. We utilize multiple aspects of evaluation, namely both automatic metrics (ROUGE, BLEU, METEOR, BERTScore), when applicable, as well as winrate comparisons (Lin, 2004; Papineni et al., 2002; Banerjee and Lavie, 2005; Zhang* et al., 2020; Zheng et al., 2023).

We also present an approach to generate high-quality health dialogue summaries in an unsuper-

* Work done while at Samsung Research America.

vised manner. Our method removes the dependency on ground truth labels or LLMs, and can be directly applied to real-world data even in the absence of expert labels. More specifically, we extract several agent-generated summaries for a given conversation, before then refining each summary using text feedback in a silo. These refined summaries are then merged together to generate the final output. Our expectation is that through this approach, we can cultivate and refine diverse yet accurate conversation summaries. We tailor our method to various models by benchmarking on differing numbers of agents and rounds, in order to determine the most optimal setting for each model. We then use these settings to run our experiments.

For a comparison across multiple settings, we evaluate our results on both open-source benchmarks (Ben Abacha et al., 2023b; Wang et al., 2024a) as well as real-world health wellness conversations. Our contributions are as follows:

- (1) We perform evaluation benchmarking on a suite of smaller LMs for health conversation summarization, ranging from 0.5B to 4B parameters.
- (2) We present an unsupervised method for summary generation from health conversations. Our approach generates several summaries for each dialogue and iteratively refines these summaries using textual feedback. The final summary is then generated from all the refined summaries.
- (3) We provide a comprehensive analysis of our summaries, investigating its viability as a ground truth generation method for fine-tuning, as well as a comparison to conventional multiagent debate.

2 Related Work

2.1 Health Dialogue Summarization

Health and medical dialogue summarization is an important tool to improve the accuracy and personalization of an agent’s responses to a user/patient. This has been studied in the context of generating clinical notes (Yim and Yetisgen, 2021; Michalopoulos et al., 2022), medical topic-based methods (Liu et al., 2024; Nair et al., 2023), synthetic dialogue generation (Chintagunta et al., 2021; Sharma et al., 2023), or through the introduction of shared tasks (Ben Abacha et al., 2023a). We study health dialogue summarization through the lens of developing better health wellness assistants. Unlike prior work, we focus particularly on small LMs that can be ran locally, removing the need for LLMs at all. Additionally, our approach does not

require expert annotators or specialized medical domain knowledge.

2.2 Summarization From Feedback

Prior work has demonstrated the utility of incorporating feedback to fine-tune and align LMs (Ouyang et al., 2022). Text feedback has shown promise in enhancing model capabilities (Yuan et al., 2024), particularly for summarization (Chen et al., 2024; Fernandes et al., 2023; Zhang et al., 2023; Stiennon et al., 2020; Bai et al., 2022). These works typically utilize human annotators for verbose text feedback generation, whereas our proposed method is entirely unsupervised, removing the need for annotators entirely. Furthermore, automatic processes typically utilize LLMs directly and/or incorporate feedback generation during inference (Kirstein et al., 2025; Song et al., 2025), which introduces significant compute overhead. With smaller LMs, our approach can bypass this challenge through our iterative text feedback process, as well as aggregating multiple refined summaries to generate the final output.

2.3 Multiagent Frameworks

The adoption of multiagent frameworks in recent years has led to improvements in the text generation and reasoning capabilities of LLMs (Du et al., 2024). Multiagent frameworks incorporating feedback have been utilized towards improving LLM critique quality (Lan et al., 2024; Kar Gupta et al., 2025), enhancing agent collaboration (Wan et al., 2025), and table refinement (Yu et al., 2025). Concurrent work (Fang et al., 2025) has also demonstrated the viability of these frameworks for text summarization. Unlike true multiagent frameworks, our method iteratively improves agent generated summaries within a silo, removing the inter-agent communication step within each round. We motivate this approach because summarization is inherently subjective, and different viewpoints may summarize the same content differently (Li et al., 2021; Steen and Markert, 2024; Fabri et al., 2021). In this manner, we aim to model this setting by preserving and refining summaries that may originate from different agents, rather than having all agents converge onto the same summary. This process allows the final output to incorporate information from a variety of refined summaries. Crucially, our approach also differs by focusing on small LMs capable of being ran locally or even on-device, eliminating the need for LLMs entirely.

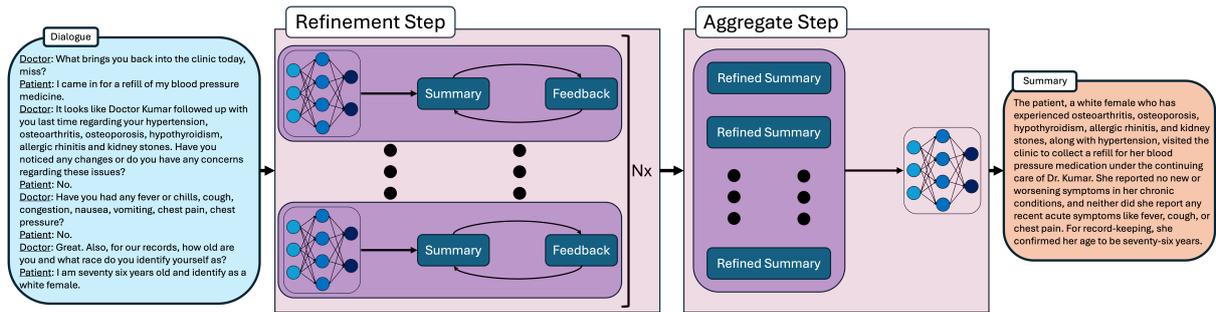


Figure 1: An overview of our method for a sample from the MTS-Dialog dataset. For a given conversation, each agent generates a summary, which is then iteratively refined through textual feedback for a variable number of rounds. After all rounds, we then aggregate all of these cultivated summaries to produce the final conversation summary. N denotes the number of agents.

3 Method

An overview of our approach to generate high-quality health dialogue summaries is depicted in Figure 1. For simplicity, our implementation uses the same base model for each agent. Please refer to Appendix A for the relevant prompts and Appendix C for the relevant generation parameters.

3.1 Summary Refinement with Text Feedback

For any conversation, we first generate initial summaries of the health dialogue using k agents. If the number of rounds is strictly greater than one, then the summaries are refined with text feedback. In this step, each agent provides text feedback on the summary that it generated. The summary feedback that we ask for looks to remove redundant information from the summary, while checking to see if there was anything in the original conversation that was not included in the initial summary. We remark that this feedback is unrestricted, since different parts of a summary may be important depending on future contexts, an aspect that makes health dialogue summarization difficult. For example, mentioning that you “went to a concert” may not typically be relevant in a health conversation summary, but may be relevant if a user is asking about how they might have contracted Covid or other infectious diseases.

Then, each agent will incorporate its feedback to update its output, resulting in a refined version of the original summary. This process is then repeated until we have completed all feedback rounds. Note that in the case of one round, no feedback is generated since there is no update step (which requires at least two rounds), and instead the summaries are directly fed into the aggregation step. Importantly, during the feedback and update steps, we note that

each agent does not see the output summaries of the other agents. One example of a feedback and update step can be found in Table 19.

3.2 Aggregating Refined Summaries

Whereas prior research approaches this step using specialized blending modules (Zhang et al., 2021; Jiang et al., 2023b) or through some combination of intermediate summaries (Laskar et al., 2023; Zhang et al., 2022), we treat all the summaries, which may have been iteratively refined, as good candidate examples to be included in the context.

More specifically, we append these refined summaries, output from the k individual agents, as candidate references the model can utilize during its final generation. We then ask the final agent to generate the final summary using all of these refined summaries as context. Essentially, these candidate references are included as in-context learning references. By incorporating siloed iterative refinement, our objective is to ensure that a diverse set of candidate references may be included in the aggregation step when generating the final summary.

4 Experimental Setup

4.1 Open-Source Benchmarks

We benchmark our method on two open-source datasets: MTS-Dialog (Ben Abacha et al., 2023b) and NoteChat (Wang et al., 2024a). We evaluate MTS-Dialog on its two test splits, with each one containing 200 samples. We evaluate on NoteChat’s ChatGPT split, which contains roughly 167k samples. Due to the sheer size of the NoteChat dataset, we evaluate our results on 100 randomly chosen samples, without overlap, across 3 different runs. Further details on conversation

| Model | Settings | |
|------------------|-------------|-------------|
| | Num. Agents | Num. Rounds |
| Qwen-2.5-0.5B-IT | 3 | 3 |
| Llama-3.2-1B-IT | 4 | 2 |
| Gemma-3-1B-IT | 2 | 1 |
| DS-Qwen-1.5B | 3 | 1 |
| Gemma-2-2B-IT | 4 | 1 |
| Llama-3.2-3B-IT | 3 | 3 |
| Phi-3-Mini-IT | 4 | 3 |

Table 1: The optimal number of agents and rounds that we selected for each model.

statistics for these datasets may be found in Table 22.

4.2 Internal Real-World Data

To provide a realistic overview of real-world viability, we also benchmark our experiments on two internal health wellness conversation datasets, which we title “InternalA” and “InternalB”. InternalA consists of 140 health dialogues drawn from 10 different users. InternalB consists of a larger set of 536 conversations, with 161 of those contained in its test split. For our purposes, we mainly use InternalA to determine the optimal hyperparameters (i.e. number of agents and number of rounds) for each model, leaving the InternalB dataset for the main evaluation. To determine the optimal settings for each model, we conduct a grid search by benchmarking all possible combinations of agents (2,3,4) and rounds (1,2,3) on our InternalA dataset. These optimal settings are displayed in Table 1. For generation details about the InternalA dataset, please refer to Appendix B. For an accurate breakdown and description of the full results for each combination, please refer to Appendix F.

4.3 Models

We collect results for a suite of small LMs (less than 4B parameters) that can feasibly be ran locally or on-device. We selected the following models: Qwen-2.5-0.5B-Instruct (Qwen et al., 2025), Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct (Grattafiori et al., 2024), Gemma-2-2B-Instruct (Team et al., 2024), Gemma-3-1B-Instruct (Team et al., 2025b), Deepseek-R1-Distill-Qwen-1.5B (DeepSeek-AI et al., 2025), and Phi-3-Mini-Instruct (Abdin et al., 2024a). In our tables, we use “IT” as a shorthand for “Instruct”, and denote the Deepseek-R1-Distill-Qwen-1.5B model as DS-Qwen-1.5B.

We utilize Gemini-2.0-Flash (Team et al., 2025a)

as our LLM comparison model, as it was the latest iteration of the Gemini series (at the time) and was readily available for our use. Preliminary results on a set of larger LMs may be found in Appendix H.

5 Results

Winrate evaluation primarily uses Claude-3.5-Sonnet (Anthropic, 2024) as our judge, given its strong performance on winrate benchmark datasets compared with other contemporary LLMs (Tan et al., 2025). Since prior work has noted the possibility of positional bias with these LLMs, we account for this factor by running each comparison twice, swapping the position of each output in the second run (Wang et al., 2024b; Pezeshkpour and Hruschka, 2024). We count a sample as a “win” for a model if and only if our judge prefers it in both comparisons. If the preferred summary differs with the swapped comparison, we count the two summaries as “comparable”, since it was selected at least once by the judge. For brevity, results computed using automated metrics are displayed in Appendix G.

5.1 Proprietary Data

For the InternalA dataset, we see that our refined summaries generally outperform the baseline approach for all models. From Table 2, the largest gain was found on Gemma-2-2B-Instruct, with a performance gain of 31.43% (i.e. the winrate of Gemini-2.0-Flash decreased by 31.43%). Furthermore, our approach improved the model’s own winrate from 17.14% to 51.43%, an increase of 34.29%. Even for Llama-3.2-3B-Instruct, which performs well in the baseline setting, we see that our approach still yields additional benefit, improving the model winrate by 24.29% from 63.57% to 87.86%. Across all models, we observe an average performance gain of 14.08%, with all improvements being statistically significant.

On the InternalB dataset, we observe the same general trend with the exception of Llama-3.2-1B-Instruct. As seen in Table 3, Deepseek-R1-Distill-Qwen-1.5B nearly doubles its own winrate, jumping from 3.11% to 6.21%, while its comparable winrate also improved from 5.59% to 8.07%, with an overall performance gain of 5.59%. Even for models where Gemini’s winrate remains similar to baseline summarization, we see that our refined summaries are able to shift a larger proportion of its winrate from “comparable” to preferred,

| Model | Baseline Winrates | | | Refined Winrates | | |
|------------------|-------------------|------------|----------------|------------------|----------------|----------------|
| | Model (↑) | Gemini (↓) | Comparable (↑) | Model (↑) | Gemini (↓) | Comparable (↑) |
| Qwen-2.5-0.5B-IT | 2.14% | 89.29% | 8.57% | 2.14% | 82.86%* | 15.00% |
| Llama-3.2-1B-IT | 24.29% | 57.14% | 18.57% | 25.71% | 45.71%* | 28.57% |
| Gemma-3-1B-IT | 17.14% | 58.57% | 24.29% | 27.86% | 50.71%* | 21.43% |
| DS-Qwen-1.5B | 2.86% | 82.14% | 15.00% | 12.86% | 67.14%* | 20.00% |
| Gemma-2-2B-IT | 17.14% | 58.57% | 24.29% | 51.43% | 27.14%* | 21.43% |
| Llama-3.2-3B-IT | 63.57% | 13.57% | 22.86% | 87.86% | 4.29%* | 7.86% |
| Phi-3-Mini-IT | 57.14% | 27.86% | 15.00% | 76.43% | 10.71%* | 12.86% |

Table 2: Winrate of model outputs compared to summaries generated by Gemini-2.0-Flash for the InternalA dataset. Percentages are rounded to 2 decimal places. The best performance, according to the winrate of Gemini, is **bolded**, and * denotes that the result is statistically significant with a minimum of 90% confidence (one-sided).

| Model | Baseline Winrates | | | Refined Winrates | | |
|------------------|-------------------|----------------|----------------|------------------|----------------|----------------|
| | Model (↑) | Gemini (↓) | Comparable (↑) | Model (↑) | Gemini (↓) | Comparable (↑) |
| Qwen-2.5-0.5B-IT | 0.00% | 98.76% | 1.24% | 0.00% | 98.76% | 1.24% |
| Llama-3.2-1B-IT | 8.07% | 77.64%* | 14.29% | 2.48% | 88.20% | 9.32% |
| Gemma-3-1B-IT | 5.71% | 81.37% | 13.66% | 12.42% | 80.75% | 6.83% |
| DS-Qwen-1.5B | 3.11% | 91.30% | 5.59% | 6.21% | 85.71%* | 8.07% |
| Gemma-2-2B-IT | 19.88% | 57.14% | 22.98% | 32.92% | 57.14% | 9.94% |
| Llama-3.2-3B-IT | 48.45% | 34.16% | 17.39% | 52.80% | 31.06% | 16.15% |
| Phi-3-Mini-IT | 36.02% | 50.31% | 13.66% | 29.81% | 50.31% | 19.88% |

Table 3: Winrate of model outputs compared to summaries generated by Gemini-2.0-Flash for the InternalB dataset. Percentages are rounded to 2 decimal places. The best performance, according to the winrate of Gemini, is **bolded**, and * denotes that the result is statistically significant with a minimum of 90% confidence (one-sided).

with Gemma-3-1B-Instruct, Gemma-2-2B-Instruct, and Llama-3.2-3B-Instruct winrates improving by 6.71%, 13.04%, and 4.35%, respectively. This is especially notable for Gemma-2-2B-Instruct, where Gemini’s winrate remains at 57.14% across both approaches, but our method is able to push the smaller model’s winrate up from 19.88% to 32.92%.

5.2 Open-Source Medical Datasets

On the MTS-Dialog dataset, our approach generally outperforms baseline summarization on the first test split, with the exception of the two Llama models (although only Llama-3.2-1B-Instruct is statistically significant). Otherwise, from Table 4, our refined summaries outperform baseline on all other models, with all but Qwen-2.5-0.5B-Instruct displaying statistically significant improvements. The largest gains are seen in the Gemma-2 and Gemma-3 models, as well as Phi-3-Mini-Instruct, with improvements of 29.00%, 13.00%, and 12.50%, respectively. The second test set is more divided: among the statistically significant improvements, Llama-3.2-1B-Instruct, Deepseek-R1-Distill-Qwen-1.5B, and Llama-3.2-3B-Instruct perform better in baseline summarization, while Gemma-3-1B-Instruct, Gemma-2-2B-Instruct, and Phi-3-Mini-Instruct perform better

with our method. These results mirror the trends we observed in the first test split, although the average improvement of the models that perform better in baseline is notably less than the average improvement of the models that perform better with our approach (6.50% versus 17.50%). Across both test splits, Gemma-2-2B-Instruct has the highest average improvement of 27.25% when using our approach. On the other hand, Llama-3.2-1B-Instruct is the worst performing model across both test splits, with its baseline summaries actually outperforming our refined summaries by 8.75%. This may in part be explained due to its prompt sensitivity, which results in the model being unable to accurately follow our proposed approach correctly, where such an issue has been noted in prior work to a degree (Arabzadeh and Clarke, 2025; Zhuo et al., 2024). Interestingly, the Deepseek-R1-Distill-Qwen-1.5B model actually exhibits different behavior on each test split, with refined summaries performing better on the first test split but baseline summaries performing better on the second test split, with both instances actually being statistically significant.

For the NoteChat dataset, we also see that LMs generally perform better with our refined summaries approach. From Table 5, we see that

| Model | Baseline Winrates | | | Refined Winrates | | |
|------------------|-------------------|----------------|----------------|------------------|----------------|----------------|
| | Model (↑) | Gemini (↓) | Comparable (↑) | Model (↑) | Gemini (↓) | Comparable (↑) |
| Qwen-2.5-0.5B-IT | 0.00% | 100.00% | 0.00% | 0.00% | 99.50% | 0.50% |
| Llama-3.2-1B-IT | 4.50% | 88.50%* | 7.00% | 0.00% | 97.50% | 2.50% |
| Gemma-3-1B-IT | 4.00% | 87.50% | 8.50% | 7.50% | 74.50%* | 18.00% |
| DS-Qwen-1.5B | 0.50% | 99.00% | 0.50% | 1.50% | 96.00%* | 2.50% |
| Gemma-2-2B-IT | 16.50% | 56.50% | 27.00% | 42.00% | 27.50%* | 30.50% |
| Llama-3.2-3B-IT | 43.00% | 30.00% | 32.00% | 38.50% | 36.00% | 25.50% |
| Phi-3-Mini-IT | 22.50% | 49.50% | 28.00% | 33.00% | 37.00%* | 30.00% |
| Qwen-2.5-0.5B-IT | 0.50% | 97.50% | 2.00% | 1.00% | 98.00% | 1.00% |
| Llama-3.2-1B-IT | 4.50% | 90.00%* | 5.50% | 0.00% | 98.50% | 1.50% |
| Gemma-3-1B-IT | 2.00% | 91.00% | 7.00% | 5.50% | 82.50%* | 12.00% |
| DS-Qwen-1.5B | 2.00% | 93.50%* | 4.50% | 1.00% | 97.50% | 1.50% |
| Gemma-2-2B-IT | 15.00% | 53.50% | 31.50% | 42.50% | 28.00%* | 29.50% |
| Llama-3.2-3B-IT | 42.50% | 29.50%* | 28.00% | 40.00% | 37.50% | 22.50% |
| Phi-3-Mini-IT | 23.00% | 57.00% | 20.00% | 39.50% | 38.50%* | 22.00% |

Table 4: Winrate of model outputs compared to summaries generated by Gemini-2.0-Flash for the MTS-Dialog dataset test splits. The top half displays results for the first test split, and the bottom half displays results for the second test split. Percentages are rounded to 2 decimal places. The best performance, according to the winrate of Gemini, is **bolded**, and * denotes that the result is statistically significant with a minimum of 90% confidence (one-sided).

Qwen-2.5-0.5B-Instruct, Gemma-3-1B-Instruct, Deepseek-R1-Distill-Qwen-1.5B, Llama-3.2-3B-Instruct, and Phi-3-Mini-Instruct perform better in all three of the NoteChat test splits, with an average improvement of 1.67%, 5.00%, 2.00%, 14.33%, and 15.33% across all splits, respectively. Additionally, Gemma-2-2B-Instruct performs better with refined summaries on two of the three test splits, with an average improvement of 3.67% across all splits. In general, it appears that the larger LMs typically exhibit the highest performance improvement with our refined summaries, in particular Llama-3.2-3B-Instruct and Phi-3-Mini-Instruct, as they are able to improve their model winrates by 13.00% and 8.00% on average across all test splits. The one notable exception was Llama-3.2-1B-Instruct (as was the case for the MTS-Dialog and InternalB datasets), where baseline summarization performed better in all three test splits, with two of those results statistically significant. The results on MTS-Dialog and NoteChat in particular for our method are especially impressive, given that these datasets typically contain more medical jargon and specialized terminology than typical health wellness conversations.

5.3 Viability Towards Fine-Tuning

We also conducted a case study analysis to investigate the real-world potential of our refined summaries to act as ground truth for supervised fine-tuning of other models. For use cases where inference latency is highly desirable, supervised fine-tuning could be one technique to reduce response

generation times, distilling the high quality of our refined summaries into just a single model call. We use Llama-3.2-3B-Instruct to generate refined summaries to act as ground truth for the InternalB training split, as it exhibited the best performance on the InternalB test split according to Table 3. On the 375 training samples, these refined summaries exhibited a winrate of 51.73%, compared to 27.47% for Gemini summaries, as well as 20.80% which were deemed comparable. The results from Table 6 are promising, with all but Llama-3.2-3B-Instruct outperforming *both* baseline and refined summaries. All models apart from Qwen-2.5-0.5B-Instruct outperform Gemini (i.e. Gemini’s winrate is under 50%), with the largest improvements being Llama-3.2-1B-Instruct at 31.68%, Gemma-2-2B-Instruct at 31.67%, and Phi-3-Mini-Instruct at 22.98%. For Llama-3.2-3B-Instruct, its performance degradation can in part be attributed to the fact that it was fine-tuned on its own outputs, which prior work has shown could result in performance loss (Shumailov et al., 2024; Gillman et al., 2024).

As a quick remark, we did not report fine-tuning results for Gemma-3-1B-Instruct and Deepseek-R1-Distill-Qwen-1.5B. We had limited GPU access for fine-tuning, and at the time Gemma-3-1B-Instruct had not yet been released. For the Deepseek model, we were concerned that fine-tuning the model would disrupt its inherent reasoning ability, thus providing an unfair comparison of its results against other models. Fine-tuning

| Model | Baseline Winrates | | | Refined Winrates | | |
|------------------|-------------------|----------------|----------------|------------------|----------------|----------------|
| | Model (↑) | Gemini (↓) | Comparable (↑) | Model (↑) | Gemini (↓) | Comparable (↑) |
| Qwen-2.5-0.5B-IT | 0.00% | 100.00% | 0.00% | 0.00% | 100.00% | 0.00% |
| Llama-3.2-1B-IT | 3.00% | 92.00% | 5.00% | 3.00% | 95.00% | 2.00% |
| Gemma-3-1B-IT | 0.00% | 96.00% | 4.00% | 1.00% | 92.00% | 7.00% |
| DS-Qwen-1.5B | 1.00% | 99.00% | 0.00% | 1.00% | 94.00%* | 5.00% |
| Gemma-2-2B-IT | 8.00% | 68.00% | 24.00% | 16.00% | 70.00% | 14.00% |
| Llama-3.2-3B-IT | 23.00% | 62.00% | 15.00% | 28.00% | 54.00% | 18.00% |
| Phi-3-Mini-IT | 18.00% | 71.00% | 11.00% | 23.00% | 47.00%* | 30.00% |
| Qwen-2.5-0.5B-IT | 0.00% | 100.00% | 0.00% | 0.00% | 98.00% | 2.00% |
| Llama-3.2-1B-IT | 4.00% | 89.00%* | 7.00% | 1.00% | 95.00% | 4.00% |
| Gemma-3-1B-IT | 3.00% | 94.00% | 3.00% | 2.00% | 92.00% | 6.00% |
| DS-Qwen-1.5B | 0.00% | 97.00% | 3.00% | 1.00% | 97.00% | 2.00% |
| Gemma-2-2B-IT | 11.00% | 70.00% | 19.00% | 17.00% | 61.00% | 22.00% |
| Llama-3.2-3B-IT | 11.00% | 68.00% | 21.00% | 30.00% | 46.00%* | 24.00% |
| Phi-3-Mini-IT | 16.00% | 62.00% | 22.00% | 22.00% | 47.00%* | 31.00% |
| Qwen-2.5-0.5B-IT | 0.00% | 100.00% | 0.00% | 0.00% | 97.00%* | 3.00% |
| Llama-3.2-1B-IT | 5.00% | 87.00%* | 8.00% | 3.00% | 94.00% | 3.00% |
| Gemma-3-1B-IT | 1.00% | 99.00% | 0.00% | 6.00% | 90.00%* | 4.00% |
| DS-Qwen-1.5B | 1.00% | 99.00% | 0.00% | 1.00% | 98.00% | 1.00% |
| Gemma-2-2B-IT | 11.00% | 63.00% | 26.00% | 20.00% | 59.00% | 21.00% |
| Llama-3.2-3B-IT | 23.00% | 51.00% | 26.00% | 38.00% | 38.00%* | 24.00% |
| Phi-3-Mini-IT | 21.00% | 55.00% | 24.00% | 34.00% | 48.00% | 18.00% |

Table 5: Winrate of model outputs compared to summaries generated by Gemini-2.0-Flash for the NoteChat dataset on 100 random samples over 3 runs. The top half displays results for the first random seed, the middle half displays results for the second random seed, and the bottom half displays results for the final random seed. The random seeds were 42, 0, 12345. Percentages are rounded to 2 decimal places. The best performance, according to the winrate of Gemini, is **bolded**, and * denotes that the result is statistically significant (minimum of 90% confidence one-sided).

| Model | Fine-Tuned Winrates | | |
|-----------------|---------------------|--------|------------|
| | Model | Gemini | Comparable |
| Qwen-0.5B-IT | 3.73% | 91.93% | 4.35% |
| Llama-3.2-1B-IT | 27.95% | 45.96% | 26.09% |
| Gemma-2-2B-IT | 49.07% | 25.47% | 25.47% |
| Llama-3.2-3B-IT | 41.61% | 35.40% | 22.98% |
| Phi-3-Mini-IT | 57.76% | 27.33% | 14.91% |

Table 6: Winrate of fine-tuned model outputs compared to Gemini-2.0-Flash summaries for the InternalB dataset. Percentages are rounded to 2 decimal places.

hyperparameters can be found in Appendix D.

5.4 Comparison to Multiagent Debate

Additionally, we also compared our results on the InternalB, MTS-Dialog, and NoteChat datasets against summaries generated by conventional multiagent debate, a popular framework that is used to improve inference-time reasoning performance (Du et al., 2024). We modify multiagent debate by replacing the majority consensus step with a simple aggregation step, combining all agent-generated summaries to output a single final summary. For consistency, we use the same number of agents and rounds that we determined from Table 1. Please see Appendix A for the relevant prompts.

| Model | Multiagent Debate Winrates | | |
|------------------|----------------------------|---------|------------|
| | Model | Gemini | Comparable |
| Qwen-2.5-0.5B-IT | 0.00% | 100.00% | 0.00% |
| Llama-3.2-1B-IT | 4.97% | 83.85% | 11.18% |
| Gemma-3-1B-IT | 4.35% | 81.99% | 13.66% |
| DS-Qwen-1.5B | 0.62% | 92.55% | 6.83% |
| Gemma-2-2B-IT | 11.80% | 77.64% | 10.56% |
| Llama-3.2-3B-IT | 61.49% | 14.91% | 23.60% |
| Phi-3-Mini-IT | 4.35% | 90.06% | 5.59% |

Table 7: Winrate of model outputs generated by multiagent debate compared to summaries generated by Gemini-2.0-Flash for the InternalB dataset. Percentages are rounded to 2 decimal places.

For the InternalB dataset, as seen in Table 7, multiagent debate outperforms our refined summaries for both Llama-3.2-1B-Instruct and Llama-3.2-3B-Instruct, with gains of 4.35% and 16.15%, respectively. We see that Llama-3.2-3B-Instruct is also able to beat baseline summarization, although this is not the case for Llama-3.2-1B-Instruct. Overall however, multiagent debate performs notably worse for all other models, with an average performance loss of 13.91% compared to our refined summaries. Phi-3-Mini-Instruct is the most heavily hit, performing 39.75% worse than our refined sum-

| Model | Multiagent Debate Winrates | | |
|------------------|----------------------------|---------|------------|
| | Model | Gemini | Comparable |
| Qwen-2.5-0.5B-IT | 0.00% | 100.00% | 0.00% |
| Llama-3.2-1B-IT | 2.00% | 95.50% | 2.50% |
| Gemma-3-1B-IT | 0.50% | 94.50% | 5.00% |
| DS-Qwen-1.5B | 1.00% | 97.00% | 2.00% |
| Gemma-2-2B-IT | 32.00% | 41.00% | 27.00% |
| Llama-3.2-3B-IT | 31.00% | 41.00% | 28.00% |
| Phi-3-Mini-IT | 2.00% | 87.50% | 10.50% |
| Qwen-2.5-0.5B-IT | 0.00% | 100.00% | 0.00% |
| Llama-3.2-1B-IT | 1.00% | 95.50% | 3.50% |
| Gemma-3-1B-IT | 2.50% | 95.00% | 2.50% |
| DS-Qwen-1.5B | 1.00% | 93.50% | 5.50% |
| Gemma-2-2B-IT | 27.00% | 44.00% | 26.00% |
| Llama-3.2-3B-IT | 41.00% | 35.50% | 23.50% |
| Phi-3-Mini-IT | 1.00% | 91.50% | 7.50% |

Table 8: Winrate of model outputs generated by multiagent debate (using the same blended settings) compared to summaries generated by Gemini-2.0-Flash for the MTS-Dialog dataset. The top half displays results for the first test split, and the bottom half displays results for the second test split. Percentages are rounded to 2 decimal places.

maries, while Gemma-2-2B-Instruct is next, with a 20.50% degradation compared to our approach.

This trend also holds true for MTS-Dialog, where in Table 8 Llama-3.2-1B-Instruct is the only model where multiagent debate summaries outperform refined summaries across both test splits, with an average improvement of 2.50%, while Llama-3.2-3B-Instruct and Deepseek-R1-Distill-Qwen-1.5B outperform our refined summaries on the second test split (2.00% and 4.00% gains). Otherwise, all other models perform notably better than multiagent debate summaries, such as Phi-3-Mini-Instruct, Gemma-2-2B-Instruct, and Gemma-3-1B-Instruct, where refined summaries perform 51.75%, 14.75%, and 16.25% better on average.

NoteChat exhibits the same trends in comparison, as seen from Table 9, where refined summaries generally outperform multiagent debate summaries on all models apart from Llama-3.2-3B-Instruct. Once again, the discrepancy is especially pronounced for Phi-3-Mini-Instruct, where refined summaries perform 42.00% better than multiagent debate summaries, on average.

5.5 Alternative Judges

To guard against over reliance on a single LLM judge, we also conducted several case study evaluations using both Llama-3-70B-Instruct (Grattafiori et al., 2024) as well as 4 medical expert human annotators. On MTS-Dialog (second test split), our

| Model | Multiagent Debate Winrates | | |
|------------------|----------------------------|---------|------------|
| | Model | Gemini | Comparable |
| Qwen-2.5-0.5B-IT | 0.00% | 100.00% | 0.00% |
| Llama-3.2-1B-IT | 1.00% | 96.00% | 3.00% |
| Gemma-3-1B-IT | 1.00% | 98.00% | 1.00% |
| DS-Qwen-1.5B | 0.00% | 99.00% | 1.00% |
| Gemma-2-2B-IT | 11.00% | 77.00% | 12.00% |
| Llama-3.2-3B-IT | 35.00% | 34.00% | 31.00% |
| Phi-3-Mini-IT | 4.00% | 91.00% | 5.00% |
| Qwen-2.5-0.5B-IT | 0.00% | 100.00% | 0.00% |
| Llama-3.2-1B-IT | 2.00% | 94.00% | 4.00% |
| Gemma-3-1B-IT | 0.00% | 100.00% | 0.00% |
| DS-Qwen-1.5B | 1.00% | 99.00% | 0.00% |
| Gemma-2-2B-IT | 10.00% | 79.00% | 11.00% |
| Llama-3.2-3B-IT | 40.00% | 37.00% | 23.00% |
| Phi-3-Mini-IT | 3.00% | 90.00% | 7.00% |
| Qwen-2.5-0.5B-IT | 0.00% | 100.00% | 0.00% |
| Llama-3.2-1B-IT | 3.00% | 95.00% | 2.00% |
| Gemma-3-1B-IT | 0.00% | 100.00% | 0.00% |
| DS-Qwen-1.5B | 1.00% | 99.00% | 0.00% |
| Gemma-2-2B-IT | 11.00% | 72.00% | 17.00% |
| Llama-3.2-3B-IT | 44.00% | 33.00% | 23.00% |
| Phi-3-Mini-IT | 5.00% | 87.00% | 8.00% |

Table 9: Winrate of model outputs generated by multiagent debate (using the same blended settings) compared to summaries generated by Gemini-2.0-Flash for the NoteChat dataset on 100 random samples over 3 runs. The top half displays results for the first random seed, the middle half displays results for the second random seed, and the bottom half displays results for the final random seed. The random seeds were 42, 0, 12345. Percentages are rounded to 2 decimal places.

refined summaries are preferred 48% of the time, compared to just 22.50% for Gemini, with 29.50% being comparable, as shown in Table 10. On the NoteChat dataset (random seed 0), our summaries are preferred 44.00% of the time, compared to just 31.00% for Gemini, with 25.00% being comparable. This same trend also persists for the NoteChat random seed 12345 split, with our summaries being preferred 46.00% of the time, compared to just 26.00% for Gemini, with 28.00% being comparable. These results reinforce the effectiveness of our approach, exhibiting even higher winrates compared to those computed using Claude-3.5-Sonnet. On MTS-Dialog (second test split), our refined winrate increases from 33.00% to 48.00%, while on NoteChat (random seeds 0, 12345), our refined winrates increase from 22.00% to 44.00% and from 34.00% to 46.00%, respectively.

Additionally, we also conducted a small manual annotation study to verify the results from our LLM judges. Specifically, we asked 4 medical expert human annotators to evaluate the winrates of

| Model | Refined Winrates | | |
|---------------|------------------|--------|------------|
| | Model | Gemini | Comparable |
| Phi-3-Mini-IT | 48.00% | 22.50% | 29.50% |
| Phi-3-Mini-IT | 44.00% | 31.00% | 25.00% |
| Phi-3-Mini-IT | 46.00% | 26.00% | 28.00% |

Table 10: Winrate of Phi-3-Mini-Instruct model outputs compared to summaries generated by Gemini-2.0-Flash. Results are computed using Llama-3-70B-Instruct as our LLM judge. The top section displays results for the MTS-Dialog first test split, and the bottom two sections display results for the NoteChat dataset random seeds 0 and 12345. Percentages are rounded to 2 decimal places.

| Model | Refined Winrates | | |
|---------------|------------------|--------|------------|
| | Model | Gemini | Comparable |
| Phi-3-Mini-IT | 53.33% | 46.67% | 0.00% |
| Phi-3-Mini-IT | 30.00% | 23.33% | 46.67% |
| Phi-3-Mini-IT | 70.00% | 30.00% | 0.00% |
| Phi-3-Mini-IT | 40.00% | 33.33% | 26.67% |

Table 11: Winrate of Phi-3-Mini-Instruct model outputs compared to summaries generated by Gemini-2.0-Flash for the NoteChat dataset (random seed 12345) on a drawn split of 30 samples. Human annotation results for 4 different medical experts are shown. Percentages are rounded to 2 decimal places.

our refined summaries, generated from Phi-3-Mini-Instruct, for the NoteChat (random seed 12345) dataset. We selected 30 samples from this dataset, which were drawn such that there is an even amount that is preferred in each category. This means that Llama-3-70B-Instruct preferred our refined summaries in 10 of these samples, Gemini summaries in 10 of these samples, and considered both summaries comparable in 10 of these samples. From Table 11, we see that all 4 medical experts preferred our refined summaries over Gemini’s summaries, with an average winrate of 48.33% for our refined summaries, compared to just 33.00% for Gemini, with 18.33% being deemed comparable.

6 Discussion

Across all datasets, we observe that our approach of refining summaries via textual feedback outperforms both baseline summarization and classic multiagent debate frameworks. The one exception was Llama-3.2-1B-Instruct, where baseline summarization actually performed better for all but the InternalA dataset. As mentioned earlier, this may be due to the model’s prompt sensitivity. Indeed,

during our experiments, we found that Llama-3.2-1B-Instruct sometimes tended to continue the input conversation, generating a response as if it was replying directly to the user, rather than actually performing summarization on the input dialogue. While baseline summarization performed better for Llama-3.2-3B-Instruct on the MTS-Dialog dataset, our refined summaries are notably better on all other datasets.

Against multiagent debate, we see that our approach yields improvement on all models except for Llama-3.2-1B-Instruct, and in some instances Llama-3.2-3B-Instruct. However, for all other models, our refined summaries still perform notably better than multiagent debate summaries. Additionally, our fine-tuning results demonstrate the viability of training other LMs on our refined summaries, illustrating our method’s potential viability as a ground truth generator for unsupervised data.

Finally, we remark that our refined summaries are able to exhibit higher semantic diversity than baseline, without compromising in terms of text quality. While multiagent debate summaries can be even more diverse, this comes at the trade-off of reduced summary quality. We share detailed results and explain our findings further in Appendix E.

7 Conclusion

We conduct an empirical study analyzing the health conversation summarization capabilities of small LMs. We propose an approach to generate summaries by iteratively refining them using textual feedback, before merging these refined summaries to generate the final summary. Our results across both open-source and proprietary health wellness and medical conversation datasets demonstrate the viability of our method, outperforming both baseline summarization and conventional multiagent debate. Crucially, we focus primarily on small and compact LMs, and remark that our method can be run on as little as a single Macbook with 16 GB of memory, without the use of a GPU at all.

Limitations

One limitation of our work is that we evaluated health wellness conversation summarization quality for small LMs only, i.e. models below 4B parameters. For institutions and corporations with the necessary compute requirements, we acknowledge that results on a suite of much larger models would prove beneficial. For those who are inter-

ested however, we have included additional preliminary benchmarking results in Appendix H, for models ranging up to 9B parameters. Nevertheless, these results were excluded from the main body as our work primarily focused on models that can fit on-device, given the strict guidelines and compliance regulations on where personal medical/health data can be sent to, as well as its broad applicability. This aspect is underlined by the fact that all of our experiments, apart from the fine-tuned numbers displayed in Table 6, were conducted without the use of a GPU, making our method readily accessible.

Another limitation that may arise is the additional latency overhead, which is not incurred in baseline summarization approaches. For our purposes, this constraint was not as prominent as memory constraints, as our needs only require offline summarization, prioritizing quality over inference speed. Even in worst case scenarios (4 agents, 3 rounds, no GPUs), our method incurs roughly 61.87 seconds per sample. We deem this trade-off acceptable, particularly when working in the health wellness and medical domains, where accuracy is more highly valued than speed. Despite this, our results in Table 6 do illustrate one possible workaround: distilling our summary generation method into models via supervised fine-tuning. Then, one can utilize the fine-tuned models to perform summarization in one model call, thus eliminating the additional latency overhead introduced by our method. Realistically, we remark that this is a limitation that applies to all frameworks with multiple agents. Unlike existing approaches, however, we stress that our approach can feasibly be ran locally, and removing the dependence on LLMs entirely. We acknowledge that our approach incurs some latency, but for the purposes of our work, we were more concerned with summary quality, especially given the sensitive and decision-critical nature of the health and medical domains.

Finally, future work may also investigate the feasibility of heterogeneous models. We did not evaluate this approach due to the additional compute overhead that it requires, as our objective was instead on a method that can run even in the absence of GPU resources. Such an approach also raises additional concerns regarding model selection as well as model influence. Nevertheless, evaluating heterogeneous models may provide useful insights and comparisons with our approach, particularly in situations where GPU resources are abundant.

Ethical Considerations

We chose to conduct our open-source dataset evaluations on the MTS-Dialog and NoteChat datasets, which are open-source and publicly available. For the InternalA dataset, we note that these 140 conversations were generated synthetically, and were cleaned to ensure they do not contain personable user data or other identifiable information. As our internal data consists of health wellness conversations, the main filtering process simply removes names (data such as geographic locations were not present in our conversations). Given the nature of the task, health metrics were not removed, as that would defeat the purpose of a health wellness assistant. We note that all internal participants were made fully aware of the data collection procedure and its purposes for internal research only, and ensured consent was received before we used this data for either evaluation or fine-tuning. Additionally, our fine-tuned results on this dataset were performed privately in-house, in order to prevent data leakage and/or compromise. Due to the sensitive nature of the internal data, these datasets (presently) are unable to be released for public use.

We caution that for actual real-world use, all LM deployment systems, and particularly in the health and wellness domains, should undergo further rigorous testing before being exposed for public use, in order to guard against hallucinations in the summary. At our current stage, our results and models have not yet been exposed publicly and are currently only being used for research purposes. For those that wish to adopt our approach and method, we urge that they first rigorously test and verify on their own use cases/data.

As a quick remark, the Gemini-2.0-Flash model that we benchmarked our summaries against was hosted privately, rather than externally connected. We did this to guard against external data transfer, and model real-world settings where user data compliance regulations must be met.

References

- Mahyar Abbasian, Iman Azimi, Amir M. Rahmani, and Ramesh Jain. 2024. [Conversational health agents: A personalized llm-powered agent framework](#). *Preprint*, arXiv:2310.02374.
- Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck,

- Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024a. [Phi-3 technical report: A highly capable language model locally on your phone](#). *Preprint*, arXiv:2404.14219.
- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024b. [Phi-4 technical report](#). *Preprint*, arXiv:2412.08905.
- Anthropic. 2024. [Claude 3.5 sonnet model card addendum](#).
- Negar Arabzadeh and Charles L.A. Clarke. 2025. [A human-ai comparative analysis of prompt sensitivity in llm-based relevance judgment](#). *Preprint*, arXiv:2504.12408.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). *Preprint*, arXiv:2204.05862.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023a. [Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations](#). In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 503–513, Toronto, Canada. Association for Computational Linguistics.
- Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023b. [An empirical study of clinical note generation from doctor-patient encounters](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.
- Angelica Chen, Jérémy Scheurer, Jon Ander Campos, Tomasz Korbak, Jun Shern Chan, Samuel R. Bowman, Kyunghyun Cho, and Ethan Perez. 2024. [Learning from natural language feedback](#). *Transactions on Machine Learning Research*.
- Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. [Medically aware gpt-3 as a data generator for medical dialogue summarization](#). In *Proceedings of the 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, pages 354–372. PMLR.
- Justin Cosentino, Anastasiya Belyaeva, Xin Liu, Zhun Yang, Yun Liu, Shyam A. Tailor, Tim Althoff, John B Hernandez, Yossi Matias, Greg Corrado, Shwetak Patel, SHRAVYA SHETTY, Jiening Zhan, Daniel McDuff, and Cory Y McLean. 2024. [Towards a personal health large language model](#). In *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. [Improving factuality and reasoning in language models through multiagent debate](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763. PMLR.
- Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. [SummEval: Re-evaluating summarization evaluation](#). *Transactions of the Association for Computational Linguistics*, 9:391–409.
- Jiangnan Fang, Cheng-Tse Liu, Jieun Kim, Yash Bhedaru, Ethan Liu, Nikhil Singh, Nedim Lipka, Puneet Mathur, Nesreen K. Ahmed, Franck Dernoncourt, Ryan A. Rossi, and Hanieh Deilamsalehy. 2025. [Multi-llm text summarization](#). *Preprint*, arXiv:2412.15487.
- Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José G. C. de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, and André F. T. Martins. 2023. [Bridging the gap: A survey on integrating \(human\) feedback for natural language generation](#). *Transactions of the Association for Computational Linguistics*, 11:1643–1668.
- Nate Gillman, Michael Freeman, Daksh Aggarwal, Chia-Hong Hsu, Calvin Luo, Yonglong Tian, and Chen Sun. 2024. [Self-correcting self-consuming](#)

- loops for generative model training. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 15646–15677. PMLR.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. *The llama 3 herd of models*. *Preprint*, arXiv:2407.21783.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. *Mistral 7b*. *Preprint*, arXiv:2310.06825.
- Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023b. *LLM-blender: Ensembling large language models with pairwise ranking and generative fusion*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178, Toronto, Canada. Association for Computational Linguistics.
- Priyanka Kargupta, Ishika Agarwal, Tal August, and Jiawei Han. 2025. *Tree-of-debate: Multi-persona debate trees elicit critical thinking for scientific comparative analysis*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29378–29403, Vienna, Austria. Association for Computational Linguistics.
- Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. *Health-llm: Large language models for health prediction via wearable sensor data*. In *Proceedings of the fifth Conference on Health, Inference, and Learning*, volume 248 of *Proceedings of Machine Learning Research*, pages 522–539. PMLR.
- Frederic Thomas Kirstein, Terry Lima Ruas, and Bela Gipp. 2025. *What’s wrong? refining meeting summaries with LLM feedback*. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2100–2120, Abu Dhabi, UAE. Association for Computational Linguistics.
- Tian Lan, Wenwei Zhang, Chengqi Lyu, Shuaibin Li, Chen Xu, Heyan Huang, Dahua Lin, Xian-Ling Mao, and Kai Chen. 2024. *Training language models to critique with multi-agent feedback*. *Preprint*, arXiv:2410.15287.
- Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 2023. *Building real-world meeting summarization systems using large language models: A practical perspective*. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 343–352, Singapore. Association for Computational Linguistics.
- Lei Li, Wei Liu, Marina Litvak, Natalia Vanetik, Jiacheng Pei, Yinan Liu, and Siya Qi. 2021. *Subjective bias in abstractive summarization*. *Preprint*, arXiv:2106.10084.
- Chin-Yew Lin. 2004. *ROUGE: A package for automatic evaluation of summaries*. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Yanlin Liu and Jiayi Wang. 2023. *Ai-driven health advice: Evaluating the potential of large language models as health assistants*. *Journal of Computational Methods in Engineering Applications*, 3(1):1–7.
- Zhengyuan Liu, Siti Salleh, Pavitra Krishnaswamy, and Nancy Chen. 2024. *Context aggregation with topic-focused summarization for personalized medical dialogue generation*. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 310–321, Mexico City, Mexico. Association for Computational Linguistics.
- George Michalopoulos, Kyle Williams, Gagandeep Singh, and Thomas Lin. 2022. *MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations*. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Subhabrata Mukherjee, Paul Gamble, Markel Sanz Ausin, Neel Kant, Kriti Aggarwal, Neha Manjunath, Debajyoti Datta, Zhengliang Liu, Jiayuan Ding, Sophia Busacca, Cezanne Bianco, Swapnil Sharma, Rae Lasko, Michelle Voisard, Sanchay Harneja, Darya Filippova, Gerry Meixiong, Kevin Cha, Amir Youssefi, and 7 others. 2024. *Polaris: A safety-focused llm constellation architecture for healthcare*. *Preprint*, arXiv:2403.13313.
- Varun Nair, Elliot Schumacher, and Anitha Kannan. 2023. *Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models*. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 200–217, Toronto, Canada. Association for Computational Linguistics.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. *Training language models to follow instructions with human feedback*. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Pouya Pezeshkpour and Estevam Hruschka. 2024. [Large language models sensitivity to the order of options in multiple-choice questions](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.
- Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J. Topol. 2024. [Llm-based agentic systems in medicine and healthcare](#). *Nature Machine Intelligence*, 6(12):1418–1420.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Chantal Shaib, Joe Barrow, Jiuding Sun, Alexa F. Siu, Byron C. Wallace, and Ani Nenkova. 2025. [Standardizing the measurement of text diversity: A tool and a comparative analysis of scores](#). *Preprint*, arXiv:2403.00553.
- Ashwyn Sharma, David Feldman, and Aneesh Jain. 2023. [Team cadence at MEDIQA-chat 2023: Generating, augmenting and summarizing clinical dialogue with large language models](#). In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 228–235, Toronto, Canada. Association for Computational Linguistics.
- Iliia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2024. [The curse of recursion: Training on generated data makes models forget](#). *Preprint*, arXiv:2305.17493.
- Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaeckermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguerre y Arcas, and 12 others. 2023. [Towards expert-level medical question answering with large language models](#). *Preprint*, arXiv:2305.09617.
- Hwanjun Song, Taewon Yun, Yuho Lee, Jihwan Oh, Gihun Lee, Jason Cai, and Hang Su. 2025. [Learning to summarize from llm-generated feedback](#). *Preprint*, arXiv:2410.13116.
- Julius Steen and Katja Markert. 2024. [Bias in news summarization: Measures, pitfalls and corpora](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5962–5983, Bangkok, Thailand. Association for Computational Linguistics.
- Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. [Learning to summarize with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.
- Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chengguang Wang, Raluca Popa, and Ion Stoica. 2025. [Judgebench: A benchmark for evaluating LLM-based judges](#). In *The Thirteenth International Conference on Learning Representations*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025a. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025b. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- David Wan, Justin Chih-Yao Chen, Elias Stengel-Eskin, and Mohit Bansal. 2025. [Mamm-refine: A recipe for improving faithfulness in generation with multi-agent collaboration](#). *Preprint*, arXiv:2503.15272.
- Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024a. [NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15183–15201, Bangkok, Thailand. Association for Computational Linguistics.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024b. [Large language models are not fair evaluators](#). In *Proceedings*

- of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- Dingkang Yang, Jinjie Wei, Dongling Xiao, Shunli Wang, Tong Wu, Gang Li, Mingcheng Li, Shuaibing Wang, Jiawei Chen, Yue Jiang, Qingyao Xu, Ke Li, Peng Zhai, and Lihua Zhang. 2024a. [Pediatricsgpt: Large language models as chinese medical assistants for pediatric applications](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 138632–138662. Curran Associates, Inc.
- Ziqi Yang, Xuhai Xu, Bingsheng Yao, Ethan Rogers, Shao Zhang, Stephen Intille, Nawar Shara, Guodong Gordon Gao, and Dakuo Wang. 2024b. [Talk2care: An llm-based voice assistant for communication between healthcare providers and older adults](#). *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(2).
- Wen-wai Yim and Meliha Yetisgen. 2021. [Towards automating medical scribing : Clinic visit Dialogue2Note sentence alignment and snippet summarization](#). In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20, Online. Association for Computational Linguistics.
- Peiyong Yu, Guoxin Chen, and Jingjing Wang. 2025. [Table-critic: A multi-agent framework for collaborative criticism and refinement in table reasoning](#). Preprint, arXiv:2502.11799.
- Weizhe Yuan, Kyunghyun Cho, and Jason Weston. 2024. [System-level natural language feedback](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2773–2789, St. Julian’s, Malta. Association for Computational Linguistics.
- Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2023. [SummIt: Iterative text summarization via ChatGPT](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10644–10657, Singapore. Association for Computational Linguistics.
- Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R. Gormley. 2021. [Leveraging pretrained models for automatic summarization of doctor-patient conversations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.
- Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. 2022. [Summⁿ: A multi-stage summarization framework for long input dialogues and documents](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1592–1604, Dublin, Ireland. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-bench and chatbot arena](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. [ProSA: Assessing and understanding the prompt sensitivity of LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976, Miami, Florida, USA. Association for Computational Linguistics.

A Model Prompts

We provide the full prompts which we used to refine health dialogue summaries through iterative feedback. The prompt that is used to generate the original summary from each agent is depicted in Table 12. The prompt that is used to elicit text feedback from each agent is depicted in Table 13. From here, each agent can then update their original summary, incorporating this text feedback through the prompt depicted in Table 14. At the final step, we then merge these refined summaries to generate the final conversation summary, according to the prompt displayed in Table 15.

We also modify the prompts present in multi-agent debate towards our health dialogue summarization task. The prompt used to generate the original summary is the same as our refined approach, depicted in Table 12. For the multiagent communication, we modified the original prompts¹ to that which is shown in Table 16. The final summary is then generated from multiagent debate using the prompt displayed in Table 17. Finally, our winrate computation uses the prompt illustrated in Table 18, which is then fed into our judge, Claude-3.5-Sonnet.

| Prompt |
|---|
| Summarize the following conversation in one clear and concise paragraph, capturing key ideas without missing critical points. |
| Conversation: |
| { <i>C</i> } |
| Please limit your response to a single paragraph. Your summary should be written in third-person. |
| Summary: |

Table 12: The prompt that we used to generate summaries from a model. Note that *C* denotes the input conversation to summarize.

B InternalA Generation Procedure

While some of the more proprietary details are unable to be disclosed, we briefly share some high-level details about the generation process of the InternalA dataset.

¹https://github.com/composable-models/llm_multiagent_debate/tree/main

| Prompt |
|--|
| Please provide feedback on the summarization quality for the following conversation. Is there any information in the summary that is redundant, or any information that is present in the conversation but missing from the summary? |
| Conversation: |
| { <i>C</i> } |
| Example Summary: |
| { <i>S</i> } |
| Please output only the feedback itself, and limit your response to one paragraph. |
| Feedback: |

Table 13: The prompt that we used to generate textual feedback to help refine summaries. Note that *C* denotes the given conversation to summarize, while *S* denotes the example generated summary.

The synthetic InternalA dataset is used in product development for an AI-based conversational agent for health wellness coaching. A large LLM is used to mimic conversations between various user profiles and the conversational agent, which is also a large LLM (unfortunately for proprietary reasons, the specific model(s) used for generation are not disclosed). These user profiles are also generated synthetically from a large LLM, and cover a wide range of general health and wellness aspects, such as physical activity, stress, and sleep. To create these user profiles, real health data, collected from users via wearable devices (e.g. smartwatches), are used as a reference, ensuring that the generated synthetic data accurately represents real-world data. However, we note that data from actual users are filtered out of the final dataset to ensure user privacy.

Once user profiles are generated, multi-turn conversations are generated between the synthetic users and the conversational agent, mimicking real-world conversations. Conversational topics revolve around user-specific health, inquiries, preferences, and health goals, where the agent usually provides wellness recommendations based on the user profiles. The agent is also allowed to ask follow-up questions to a user’s query, facilitating more personalized user recommendations. We use guardrails on both user queries and agent responses, keeping

Prompt

Summarize the following conversation in one clear and concise paragraph, capturing key ideas without missing critical points.

Conversation:

{*C*}

Below is one example summary of the conversation as well as its corresponding feedback. Please incorporate this feedback and update the summary.

Example Summary:

{*S*}

Feedback:

{*F*}

Please limit your response to a single paragraph. Your summary should be written in third-person.

Summary:

Table 14: The prompt that we used to refine a given summary with the associated textual feedback. Note that *C* denotes the given conversation to summarize, *S* denotes the example generated summary and *F* denotes the associated text feedback.

the conversation strictly focused on health wellness while preventing the agent from outputting medical suggestions (e.g. take this drug) or irrelevant responses.

C Generation Parameters

Summaries of our health conversations are generated utilizing the following parameters: maximum new tokens of 512, temperature of 0.7, and a random seed of 42, when applicable. Note that for the Deepseek models, we increased the maximum new tokens to 1024 in order to account for the increased token counts from the models’ reasoning chain outputs. For the winrate computation, we alter the temperature setting to 0.1 to maintain additional response consistency.

D Fine-Tuning Parameters

We conduct instruction fine-tuning for all models using 4-bit quantization with LoRA adapters, i.e.

Prompt

Summarize the following conversation in one clear and concise paragraph, capturing key ideas without missing critical points.

Conversation:

{*C*}

For reference, below are a few example summaries of the conversation. Please use these to generate your final summary.

Example Summary: {*S*₁}

Example Summary: {*S*₂}

Example Summary: {*S*₃}

Please limit your response to a single paragraph. Your summary should be written in third-person.

Final Summary:

Table 15: The prompt that we used to merge all refined summaries into the final output summary. The example shown here assumes a total of 3 agents, with their refined summary outputs denoted as *S*₁, *S*₂, *S*₃.

QLoRA (Dettmers et al., 2023). Our LoRA module uses the following settings: rank of 16, alpha of 8, dropout of 0.01, and targets all linear layers. Our procedure trains the model on the text completions only, rather than training on the entire input.

We use standardized hyperparameters for fine-tuning, with the following settings: learning rate of 2e-5, weight decay of 0.01, max sequence length of 2048, batch size of 4, random seed of 42, and a warmup ratio of 0.10, fine-tuned for a total of 10 epochs. Note that during inference, we simply load in the fine-tuned LoRA adapters into the module – we do not merge the adapter into the original model’s weights.

E Measuring Summary Diversity

We use 3 metrics to measure the semantic diversity of generated summaries: homogenization score (with Rouge-L), compression ratio, and the n-gram diversity score. These metrics are introduced and described in prior work, and are publicly available in Python’s diversity library package (Shaib et al., 2025).

From Table 23, we note that our refined summaries and multiagent debate summaries are typically more diverse than baseline summaries, and

| Prompt |
|--|
| Here are some summaries of the conversation given by other agents: |
| ````{S ₁ }``` |
| ````{S ₂ }``` |
| ````{S ₃ }``` |
| Using the summaries from other agents as additional advice, can you give an updated summary? Examine your solution and that other agents step by step. Please limit your response to a single paragraph. Your summary should be written in third-person. |

Table 16: The modified multiagent debate prompt which we used to append other agents’ summaries to the message context. The example shown here assumes a total of 3 agents, with their refined summary outputs denoted as S_1, S_2, S_3 .

from our prior results discussed in Section 5, our refined summaries are able to be more diverse without sacrificing for text quality. This is important because diversity itself is not necessarily equal to quality – a set of highly diverse summaries can be of either very high quality or of very low quality. Essentially, some amount of diversity is desired, but too much may be due to having a set of very low quality conversation summaries.

On the InternalB dataset, we observe that our refined summaries are the most diverse for both Llama models, where from Section 5 we recall that Llama-3.2-3B-Instruct outperforms baseline but Llama-3.2-1B-Instruct does not, where it performs worse than both baseline and multiagent debate approaches. This reflects on the point that diversity is desirable *up to a certain point*, after which we experience performance degradation. This point is further highlighted by the fact that multiagent debate summaries are the most diverse for Gemma-2-2B-Instruct, despite them having the worst performance when compared to both baseline and our refined summary approach. This is best exemplified by the diversity scores for Phi-3-Mini-Instruct, where our refined summaries perform better than both baseline summarization as well as multiagent debate. In this scenario, our refined summaries are more diverse than baseline summaries, but less diverse than multiagent debate summaries, which happen to perform the worst among the three.

These trends are also generally observed on the MTS-Dialog dataset. In the case of Llama-3.2-1B-Instruct, it exhibits the highest diversity across

| Prompt |
|---|
| Summarize the following conversation in one clear and concise paragraph, capturing key ideas without missing critical points. |
| Conversation: |
| {C} |
| For reference, here are some example summaries from other agents: |
| Example Summary: {S ₁ } |
| Example Summary: {S ₂ } |
| Example Summary: {S ₃ } |

Table 17: The prompt that we used to merge summaries to modify multiagent debate. Note that C denotes the input conversation, and S_1, S_2, S_3 denotes the example summaries generated by the other agents. The example shown above is for the case of 3 agents.

both test splits, despite performing worse than both baseline summarization and multiagent debate. Interestingly, there is an exception with Gemma-2-2B-Instruct, where our refined summaries are the most diverse while also significantly outperforming both baseline summarization as well as multiagent debate, as recalled from Table 4 and Table 8.

On the NoteChat dataset, two notable exceptions to this trend are Qwen-2.5-0.5B-Instruct and Gemma-2-2B-Instruct, where our refined summaries are typically the most diverse on all test splits while also achieving the highest quality (as recalled from Table 5 and Table 9). Interestingly, models like Llama-3.2-3B-Instruct typically show that our refined summaries are the most diverse on the NoteChat test splits, even though they do not necessarily have the worse quality (that would be baseline summarization). Nevertheless, we see that these same general trends hold true, where some amount of diversity in our text is desirable, but too much can begin to sacrifice summarization quality.

Overall, we note that for all datasets and across all models, our refined approach to generating summaries through iterative text feedback is able to improve the diversity of our health dialogue summaries while still maintaining high text quality. Approaches such as multiagent debate are able to further increase the summarization diversity in some scenarios, but at the cost of degrading summarization quality.

| Prompt |
|---|
| <p>Which of the following summaries does a better job of summarizing the most important points in the given conversation? A good summary contains informative details in a precise and concise manner without omitting any information. A good summary should be written in a single fluid paragraph, without containing bullet points or other extra formatting options.</p> <p>Conversation:</p> <p>{<i>C</i>}</p> <p>Summary A:</p> <p>{<i>S</i>₁}</p> <p>Summary B:</p> <p>{<i>S</i>₂}</p> <p>FIRST provide a one-sentence comparison of the two summaries, explaining which you prefer and why. SECOND, on a new line, state only “A” or “B” to indicate your choice. Your response should use the format:</p> <p>Comparison: <one-sentence comparison and explanation></p> <p>Preferred: <“A” or “B”></p> |

Table 18: The prompt that we used to compute the winrate between two example summaries. Note that each comparison prompts the judge twice, where the second comparison swaps the position of the two summaries S_1, S_2 to account for positional bias.

F Winrates Per Combination

Recall that we displayed the number of agents and number of rounds that we selected for each model in Table 1. We display the full breakdown of the winrate, as computed on a random split of 20 samples drawn from our InternalA dataset, in Table 21, where our selected settings are bolded. Settings were chosen first by those that minimized the number of Gemini-2.0-Flash summaries which are preferred and then by those that maximized the number of model-generated summaries which are preferred. In instances where two different settings had equally identical optimal results, we biased towards the setting with the larger number of feedback rounds, in order to ensure that more feedback would be available in the scenario that difficult-to-summarize conversations were encountered. This tiebreaker setting was only used to determine the optimal setting for Llama-3.2-3B-Instruct, where both 3 agents 1 round and 3 agents 3 rounds displayed the same results.

G Results Compared To Dataset Labels

To ensure that our analyses arise from multiple axes of evaluation, we also show results obtained on the InternalA and MTS-Dialog datasets using automated metrics, as these datasets have built-in ground truth labels readily available. Note that we do not report results on the NoteChat test splits, as the associated ground truth labels are physician SOAP notes rather than conversation summaries.

From Table 24, we see that Phi-3-Mini-Instruct and Llama-3.2-3B-Instruct compare favorably against Gemini-2.0-Flash on the InternalA dataset, whereas other models that appear to do the same (e.g. Llama-3.2-1B-Instruct, Gemma-2-2B-Instruct) exhibit much lower performance as evaluated through winrate comparisons. These multiple axes of evaluation are important to visualize the real differences in quality between the model-generated summaries. For example, Qwen-2.5-0.5B-Instruct has results on automatic metrics which are somewhat comparable with Gemini-2.0-Flash (although slightly worse), but exhibits a winrate of just 2.14% versus Gemini’s 82.86%, as recalled from Table 2. For the MTS-Dialog splits, we observe that Gemma-2-2B-Instruct and Phi-3-Mini-Instruct are somewhat comparable to Gemini, with Rouge-1 scores of 25.30 and 25.53 versus 27.60 (first test split), respectively. However, when we evaluate these two models in baseline summarization through winrate comparisons, we see that they only exhibit winrates of 16.50% and 22.50% (first test split).

Using automated metrics to evaluate our refined summaries, we observe in Table 25 that Llama-3.2-3B-Instruct shows improvement versus baseline summarization, with a Rouge-1 and Rouge-2 of 44.08 and 17.21, compared to 43.15 and 16.51, an improvement that is also reflected in our winrate results. This is also reflected for models like Gemma-2-2B-Instruct and Deepseek-R1-Distill-Qwen-1.5B. On the other hand, Phi-3-Mini-Instruct shows a significant degradation in performance when we use automated metrics, with its Rouge-1 falling from 44.07 to 36.77 and its METEOR falling from 35.87 down to 30.48. However, these results are in contrast to our winrate comparisons, which we recall from Table 2 showed a 17.15% performance improvement in samples that are considered preferred/comparable. On the MTS-Dialog test splits, we observe a general decrease when we evaluate using automatic metrics, which gener-

Original Conversation

Doctor: Good afternoon, sir. My chart here says that you are a fifty one year old white male, is that correct? Patient: Good afternoon, doctor. Yes, all of that is correct. Doctor: Great, thank you. Please, walk me through the history of your symptoms, when did everything begin. Patient: Well, I was diagnosed with P T L D towards the end of two thousand seven. Doctor: What symptoms were you experiencing? Patient: I had swollen lymph nodes, stomach pain, I was losing a lot of weight, and I wasn't really eating. Doctor: How have you been treated for this? Patient: Well, I didn't go to the doctor right away, but eventually I went to the hospital, and they did a biopsy of my groin lymph nodes, and they found large scale lymphoma. Doctor: Who's care were you under in the hospital? Patient: Um, it was the renal transplant services. Oh yeah, they talked about my case at the hematopathology conference. Doctor: That's interesting. After you were diagnosed, what treatments were you given? Patient: Um, I started chemo with, um, I have it written down, cyclophosphamide, daunorubicin, vincristine, and prednisone. Doctor: Did you have any complications with chemo? Patient: Um, yeah, I had sepsis during my first cycle, they said that I had it even with growth factor support. I was also disoriented. Doctor: How many cycles of chemo have you completed? Patient: I've done five to date. Doctor: Did you have your P E T Scan after your first three cycles? Patient: No, I wasn't feeling well at all, so I didn't go. Doctor: Has your chemotherapy been interrupted for any other reason? Patient: Yeah, I had an infection related to the skin cancer, and a chest infection, sir. Doctor: Did you have surgery for the skin cancer? Patient: Yes, I did. Doctor: Okay, and you're here for your sixth round of chemo, correct? Patient: Yes, that's correct. Doctor: How are you feeling today? Patient: Overall, I feel pretty well. Doctor: Are you having any cold or flu like symptoms today? Patient: No. Doctor: Any shortness of breath, or chest pain? Patient: No, doctor. Doctor: Finally, your E C O G score is one according to the nurse, is that correct?

Candidate Summary

A fifty-one-year-old white male patient with a history of large-scale lymphoma, diagnosed at the end of 2007 and initially treated by renal transplant services, underwent five chemotherapy cycles (including drugs like cyclophosphamide, daunorubicin, vincristine, and prednisone) after experiencing complications such as sepsis. The patient's treatment was occasionally interrupted due to skin cancer surgery and subsequent infections but resumed for his sixth round of chemotherapy. Currently on a low ECOG score of one, the patient reported overall well-being with no cold or flu symptoms, shortness of breath, or chest pain during his recent consultation with the doctor.

Text Feedback

The summary is concise and accurately captures the essential information from the conversation regarding the patient's diagnosis, treatment history, complications, and current health status without redundancy or omission of crucial details. However, it fails to mention that the PET scan was not performed after the first three cycles due to the patient feeling unwell, which is a significant detail in understanding the full context of his treatment timeline.

Next Round Updated Summary

A fifty-one-year-old white male with a large-scale lymphoma diagnosis at the end of two thousand seven, treated initially by renal transplant services, has undergone five chemotherapy cycles including cyclophosphamide, daunorubicin, vincristine, and prednisone. Despite experiencing sepsis during his first cycle, surgical intervention for skin cancer, and infections leading to temporary treatment breaks, he completed the sixth round of chemo with a low ECOG score of one. The patient reported no current symptoms like cold or flu, shortness of breath, or chest pain but had not undergone his scheduled PET scan after the initial three cycles due to ill health.

Table 19: One example of a conversation, summary, its corresponding text feedback, as well as the updated summary, taken from the Phi-3-Mini-Instruct model. As we see, the text feedback highlights specific actionable improvements to the summary, e.g. "However, it fails to mention that the PET scan was not performed after the first three cycles due to the patient feeling unwell, which is a significant detail in understanding the full extent of his treatment timeline". These suggestions are then incorporated into the updated summary.

| Model | Baseline Winrates | | |
|--------------------|-------------------|--------|------------|
| | Model | Gemini | Comparable |
| Phi-3.5-Mini-IT | 57.14% | 18.57% | 24.29% |
| Phi-4-Mini-IT | 17.14% | 60.71% | 22.14% |
| Gemma-3-4B-IT | 62.14% | 12.86% | 25.00% |
| Mistral-7B-IT-v0.3 | 47.14% | 30.71% | 22.14% |
| DS-Qwen-7B | 53.57% | 32.86% | 13.57% |
| Llama-3.1-8B-IT | 10.71% | 75.00% | 14.29% |
| DS-Llama-8B | 42.14% | 38.57% | 19.29% |
| Gemma-2-9B-IT | 42.86% | 32.86% | 24.29% |

Table 20: Winrate of model outputs generated by baseline approaches compared to summaries generated by Gemini-2.0-Flash for the InternalA dataset. Percentages are rounded to 2 decimal places.

ally clashes with our winrate comparisons. However, given the relatively low scores, evaluated via automatic metrics, for even the Gemini-2.0-Flash model, we hypothesize that this is simply due to the different format of the ground truth labels, which are generally single or double sentences, rather than fully-fledged summaries.

For multiagent debate, the results illustrated in Table 26 are even lower than both baseline summaries as well as our refined summaries, further reinforcing the efficacy of our refined summary approach. The lone exception is the Deepseek-R1-Distill-Qwen-1.5B model, which generally performs better on all metrics, as well as Llama-3.2-1B-Instruct, which exhibits a higher BLEU score (3.10 versus 2.52 on the first test split) but comparable ROUGE, METEOR, and BERTScore compared to our refined summaries. These trends hold true for both of the MTS-Dialog test splits.

H Preliminary Results on Larger LMs

We also ran preliminary experiments on the InternalA dataset for a wider variety of slightly larger LMs for this task, as well as results illustrating the winrates per combination for these LMs. While investigating larger LMs are outside the scope of our work, as we focus on smaller/compact LMs, we recognize that results on larger LMs may provide usefulness to others. Namely, these models included Phi-3.5-Mini-Instruct, Phi-4-Mini-Instruct (Abdin et al., 2024b), Gemma-3-4B-Instruct, Mistral-7B-Instruct-v0.3 (Jiang et al., 2023a), Deepseek-R1-Distill-Qwen-7B, Llama-3.1-8B-Instruct, Deepseek-R1-Distill-Llama-8B, and Gemma-2-9B-Instruct.

We first benchmark the winrate comparison results of baseline summarization approaches using

these LMs. In Table 20, we observe that all models apart from Llama-3.1-8B-Instruct and Phi-4-Mini-Instruct are able to outperform Gemini-2.0-Flash on the InternalA dataset (i.e. Gemini-2.0-Flash exhibits a winrate less than 50%), with Phi-3.5-Mini-Instruct, Gemma-3-4B-Instruct, and Deepseek-R1-Distill-Qwen-7B able to even exceed 50% winrate for their generated summaries.

When benchmarked using automatic metrics, as shown in Table 28, we observe that these models are actually relatively comparable to the results we observed in Table 24. In fact, only Gemma-3-4B-Instruct is able to generally outperform Gemini-2.0-Flash on the InternalA dataset, while for our original suite of models this was the case for both Llama-3.2-3B-Instruct and Phi-3-Mini-Instruct. However, in this instance, we see that Gemma-3-4B-Instruct also happens to exhibit the best results in the winrate comparison as well.

We also gathered results for each model on various combinations of number of agents and number of rounds, displayed in Table 27. We note that the larger LMs typically perform better with more agents and more feedback rounds, with the Deepseek-R1-Distill-Llama-8B model even able to achieve a perfect 100% winrate on the 20 sample random split from the InternalA dataset. Additionally, both Llama-3.1-8B-Instruct and Gemma-2-9B-Instruct perform well at higher settings, with their optimal combinations being A4R3 and A4R2, respectively. Furthermore, only Phi-4-Mini-Instruct has its optimal combination at a low number of feedback rounds, which makes sense given its poor baseline performance. In these scenarios, model-elicited feedback may be of low quality due to the poor baseline performance of the model, resulting in instances where additional feedback actually ends up degrading model performance as observed through winrate computations.

One thing to note is that we used the results from Table 20 to inform our decision to choose Phi-3-Mini-Instruct in our original set of models instead of Phi-4-Mini-Instruct, as Phi-3-Mini-Instruct outperforms Phi-4-Mini-Instruct on the InternalA dataset. Additionally, our winrate results on different combinations from Table 27 illustrate that the best setting for Phi-3-Mini-Instruct (A3R3) outperforms the best setting we could achieve on Phi-3.5-Mini-Instruct (A2R2). Furthermore, we were unable to gather results on Phi-3.5-Mini-Instruct for the A4R2 and A4R3 settings, due to an open issue with the Phi-3.5 series models in Ol-

| Model | Agent & Round Settings | | | | | | | | |
|------------------|------------------------|----------|----------|----------------|----------|-----------------|-----------------|-----------------|-----------------|
| | A2R1 | A2R2 | A2R3 | A3R1 | A3R2 | A3R3 | A4R1 | A4R2 | A4R3 |
| Qwen-2.5-0.5B-IT | (1,19,0) | (0,18,2) | (0,17,3) | (0,14,6) | (0,17,3) | (4,14,2) | (1,15,4) | (1,19,0) | (1,17,2) |
| Llama-3.2-1B-IT | (9,7,4) | (8,9,3) | (4,12,4) | (11,7,2) | (7,6,7) | (11,7,2) | (9,7,4) | (10,6,4) | (7,6,7) |
| Gemma-3-1B-IT | (9,5,6) | (4,13,3) | (6,10,4) | (4,8,8) | (5,11,4) | (3,9,8) | (5,8,7) | (8,6,6) | (5,6,9) |
| DS-Qwen-1.5B | (1,13,6) | (2,13,5) | (2,8,10) | (4,8,8) | (3,10,7) | (4,12,4) | (2,12,6) | (2,13,5) | (3,12,5) |
| Gemma-2-2B-IT | (13,4,3) | (12,5,3) | (15,3,2) | (12,6,2) | (13,4,3) | (15,3,2) | (15,2,3) | (11,4,5) | (14,4,2) |
| Llama-3.2-3B-IT | (17,2,1) | (17,0,3) | (17,1,2) | (19,0,1) | (17,0,3) | (19,0,1) | (17,0,3) | (17,0,3) | (18,2,0) |
| Phi-3-Mini-IT | (16,1,3) | (15,2,3) | (17,1,2) | (15,2,3) | (17,1,2) | (18,2,0) | (14,0,6) | (14,2,4) | (17,0,3) |

Table 21: Winrate results for each combination of agent and round settings, computed for a random split of 20 samples drawn from the InternalA dataset. Each result is in the format (A, B, C) , where A, B, C denote the number of model outputs that are preferred, the number of Gemini-2.0-Flash outputs that are preferred, and the number of samples that are comparable, respectively. Note that A2R2 for example denotes the setting using 2 agents and 2 rounds. The settings that we ended up selecting for each model are **bolded**.

| Avg. Num. of Turns | Shortest Num. of Turns | Longest Num. of Turns | Avg. Words Per Conversation |
|--------------------|------------------------|-----------------------|-----------------------------|
| 8.69 | 0 | 44 | 105.13 |
| 9.89 | 0 | 103 | 111.94 |
| 14.24 | 0 | 153 | 449.16 |
| 13.77 | 0 | 46 | 421.37 |
| 13.37 | 0 | 44 | 440.72 |

Table 22: Conversation statistics for the MTS-Dialog and NoteChat datasets. The top two sections correspond to the MTS-Dialog test split one and test split two. The bottom three sections correspond to the NoteChat dataset for random seeds of 42, 0, and 12345.

lama². These issues randomly caused degenerated text output by the Phi-3.5-Mini-Instruct model.

I Software Implementation

Our method is implemented using Python, with Ollama quantized inference used for response generation. This approach ensures that our approach can be ran viably without the use of any GPU. Fine-tuning was implemented in Python, using the TRL library for supervised fine-tuning, training on completions only. For these experiments, we used a single GPU node with 40 GB of memory.

Although our approach utilizes multiple agents, we remark that we only require enough memory to store one model instance. Rather than instantiating a new model instance to represent each agent, we load one instance of the model and control for its generated responses with different random seeds. This design follows from the implementation of multiagent debate³

²<https://github.com/ollama/ollama/issues/6449>

³https://github.com/composable-models/llm_multiagent_debate/tree/main

| Model | Baseline Results | | | Refined Results | | | MAD Results | | |
|------------------|---------------------|---------------------|--------------------|---------------------|---------------------|--------------------|---------------------|---------------------|--------------------|
| | CR (\downarrow) | HS (\downarrow) | NGD (\uparrow) | CR (\downarrow) | HS (\downarrow) | NGD (\uparrow) | CR (\downarrow) | HS (\downarrow) | NGD (\uparrow) |
| Qwen-2.5-0.5B-IT | 3.89 | 0.12 | 2.23 | 3.68 | 0.11 | 2.37 | 3.67 | 0.10 | 2.43 |
| Llama-3.2-1B-IT | 4.08 | 0.17 | 2.05 | 3.67 | 0.14 | 2.41 | 4.18 | 0.18 | 2.12 |
| Gemma-3-1B-IT | 4.61 | 0.14 | 2.17 | 4.08 | 0.19 | 2.30 | 3.92 | 0.15 | 2.35 |
| DS-Qwen-1.5B | 3.31 | 0.11 | 2.71 | 3.46 | 0.12 | 2.62 | 4.90 | 0.13 | 1.76 |
| Gemma-2-2B-IT | 4.62 | 0.21 | 1.85 | 4.02 | 0.19 | 2.32 | 3.64 | 0.04 | 2.56 |
| Llama-3.2-3B-IT | 4.21 | 0.18 | 2.03 | 3.85 | 0.16 | 2.31 | 4.15 | 0.19 | 2.09 |
| Phi-3-Mini-IT | 3.93 | 0.18 | 2.19 | 3.38 | 0.14 | 2.70 | 3.13 | 0.13 | 2.79 |
| Qwen-2.5-0.5B-IT | 3.45 | 0.16 | 2.53 | 3.28 | 0.14 | 2.63 | 3.37 | 0.12 | 2.61 |
| Llama-3.2-1B-IT | 3.53 | 0.19 | 2.40 | 3.31 | 0.16 | 2.70 | 3.73 | 0.21 | 2.37 |
| Gemma-3-1B-IT | 3.38 | 0.10 | 2.66 | 3.39 | 0.17 | 2.66 | 3.29 | 0.14 | 2.70 |
| DS-Qwen-1.5B | 2.98 | 0.14 | 2.87 | 3.11 | 0.14 | 2.81 | 3.30 | 0.19 | 2.63 |
| Gemma-2-2B-IT | 3.32 | 0.17 | 2.55 | 3.28 | 0.16 | 2.69 | 3.25 | 0.16 | 2.69 |
| Llama-3.2-3B-IT | 3.40 | 0.18 | 2.52 | 3.32 | 0.16 | 2.63 | 3.47 | 0.18 | 2.50 |
| Phi-3-Mini-IT | 3.20 | 0.16 | 2.71 | 3.10 | 0.14 | 2.87 | 2.95 | 0.13 | 2.96 |
| Qwen-2.5-0.5B-IT | 3.41 | 0.15 | 2.51 | 3.35 | 0.13 | 2.59 | 3.35 | 0.13 | 2.61 |
| Llama-3.2-1B-IT | 3.45 | 0.18 | 2.44 | 3.23 | 0.15 | 2.76 | 3.61 | 0.20 | 2.44 |
| Gemma-3-1B-IT | 3.43 | 0.10 | 2.68 | 3.34 | 0.17 | 2.68 | 3.18 | 0.07 | 2.78 |
| DS-Qwen-1.5B | 3.00 | 0.15 | 2.87 | 2.94 | 0.14 | 2.92 | 3.29 | 0.19 | 2.61 |
| Gemma-2-2B-IT | 3.31 | 0.17 | 2.55 | 3.21 | 0.15 | 2.74 | 3.18 | 0.16 | 2.72 |
| Llama-3.2-3B-IT | 3.34 | 0.18 | 2.53 | 3.25 | 0.15 | 2.68 | 3.40 | 0.17 | 2.54 |
| Phi-3-Mini-IT | 3.10 | 0.16 | 2.72 | 3.07 | 0.14 | 2.89 | 2.92 | 0.13 | 2.96 |
| Qwen-2.5-0.5B-IT | 3.93 | 0.17 | 2.55 | 3.15 | 0.12 | 2.72 | 3.26 | 0.13 | 2.68 |
| Llama-3.2-1B-IT | 3.08 | 0.20 | 2.72 | 2.99 | 0.15 | 2.96 | 3.12 | 0.19 | 2.78 |
| Gemma-3-1B-IT | 3.14 | 0.17 | 2.77 | 3.20 | 0.19 | 2.77 | 2.97 | 0.16 | 2.89 |
| DS-Qwen-1.5B | 2.77 | 0.14 | 3.08 | 2.82 | 0.13 | 3.04 | 2.97 | 0.18 | 2.91 |
| Gemma-2-2B-IT | 3.09 | 0.19 | 2.68 | 2.98 | 0.16 | 2.92 | 2.98 | 0.16 | 2.90 |
| Llama-3.2-3B-IT | 3.09 | 0.19 | 2.92 | 3.02 | 0.15 | 2.89 | 3.11 | 0.18 | 2.75 |
| Phi-3-Mini-IT | 2.94 | 0.17 | 2.86 | 2.79 | 0.12 | 3.11 | 2.75 | 0.11 | 3.13 |
| Qwen-2.5-0.5B-IT | 3.46 | 0.16 | 2.46 | 3.28 | 0.13 | 2.63 | 3.27 | 0.12 | 2.67 |
| Llama-3.2-1B-IT | 3.06 | 0.20 | 2.73 | 2.97 | 0.15 | 2.98 | 3.14 | 0.19 | 2.78 |
| Gemma-3-1B-IT | 3.07 | 0.18 | 2.84 | 3.22 | 0.19 | 2.81 | 2.95 | 0.15 | 2.94 |
| DS-Qwen-1.5B | 2.77 | 0.14 | 3.08 | 2.81 | 0.13 | 3.06 | 2.98 | 0.18 | 2.91 |
| Gemma-2-2B-IT | 3.11 | 0.20 | 2.68 | 2.98 | 0.16 | 2.94 | 2.96 | 0.16 | 2.93 |
| Llama-3.2-3B-IT | 3.09 | 0.19 | 2.74 | 3.05 | 0.15 | 2.88 | 3.16 | 0.18 | 2.73 |
| Phi-3-Mini-IT | 2.97 | 0.17 | 2.89 | 2.79 | 0.12 | 3.11 | 2.76 | 0.12 | 3.12 |
| Qwen-2.5-0.5B-IT | 3.41 | 0.16 | 2.46 | 3.08 | 0.13 | 2.82 | 3.18 | 0.13 | 2.70 |
| Llama-3.2-1B-IT | 3.07 | 0.20 | 2.72 | 2.96 | 0.15 | 2.96 | 3.14 | 0.19 | 2.76 |
| Gemma-3-1B-IT | 3.13 | 0.18 | 2.75 | 3.22 | 0.20 | 2.76 | 2.50 | 0.00 | 3.11 |
| DS-Qwen-1.5B | 2.77 | 0.14 | 3.09 | 2.82 | 0.13 | 3.04 | 2.97 | 0.18 | 2.91 |
| Gemma-2-2B-IT | 3.06 | 0.19 | 2.69 | 3.00 | 0.16 | 2.91 | 2.97 | 0.17 | 2.90 |
| Llama-3.2-3B-IT | 3.12 | 0.19 | 2.69 | 3.05 | 0.16 | 2.87 | 3.15 | 0.19 | 2.71 |
| Phi-3-Mini-IT | 2.97 | 0.17 | 2.85 | 2.79 | 0.12 | 3.10 | 2.77 | 0.13 | 3.10 |

Table 23: Diversity metrics computed for the baseline approach, our approach, as well as multiagent debate (MAD). The top section displays results computed on the InternalB dataset, the next two sections display results computed on the MTS-Dialog dataset (test split one and two), and the last three sections display results computed on the NoteChat dataset (for random seeds of 42, 0, 12345). CR, HS, and NGD denote compression ratio, homogenization score, and n-gram diversity, respectively. Arrows indicate the direction of *more diversity*. Note that results are rounded to 2 decimal places. The most diverse scores, within each metric, model, and dataset are **bolded**.

| Model | ROUGE | | | | BLEU | METEOR | BERTScore | | |
|------------------|-------|-------|-------|--------|-------|--------|-----------|--------|-------|
| | R-1 | R-2 | R-L | R-LSum | | | Precision | Recall | F1 |
| Qwen-2.5-0.5B-IT | 37.09 | 12.48 | 21.86 | 31.68 | 7.72 | 29.27 | 87.52 | 84.60 | 86.02 |
| Llama-3.2-1B-IT | 41.51 | 15.03 | 23.47 | 35.10 | 9.37 | 35.37 | 87.34 | 85.34 | 86.32 |
| Gemma-3-1B-IT | 38.37 | 12.33 | 22.91 | 32.40 | 7.37 | 27.74 | 87.60 | 84.63 | 86.08 |
| DS-Qwen-1.5B | 34.70 | 11.21 | 21.36 | 29.28 | 5.61 | 23.94 | 88.16 | 83.86 | 85.95 |
| Gemma-2-2B-IT | 39.77 | 14.15 | 21.82 | 35.66 | 6.67 | 36.49 | 85.85 | 87.18 | 86.50 |
| Llama-3.2-3B-IT | 43.15 | 16.51 | 24.55 | 36.36 | 10.30 | 37.99 | 87.38 | 85.59 | 86.47 |
| Phi-3-Mini-IT | 44.07 | 16.02 | 25.49 | 36.87 | 10.53 | 35.87 | 87.79 | 85.29 | 86.52 |
| Gemini-2.0-Flash | 42.47 | 16.31 | 25.67 | 36.78 | 9.39 | 34.29 | 86.98 | 85.52 | 86.23 |
| Qwen-2.5-0.5B-IT | 20.15 | 5.82 | 14.90 | 14.91 | 3.94 | 24.42 | 86.29 | 87.62 | 86.91 |
| Llama-3.2-1B-IT | 21.63 | 7.24 | 15.10 | 15.11 | 4.86 | 25.91 | 85.60 | 88.37 | 86.92 |
| Gemma-3-1B-IT | 19.08 | 5.87 | 13.51 | 13.53 | 3.24 | 23.49 | 80.69 | 88.28 | 84.13 |
| DS-Qwen-1.5B | 20.69 | 6.32 | 15.18 | 15.30 | 3.59 | 24.88 | 85.83 | 87.88 | 86.80 |
| Gemma-2-2B-IT | 25.30 | 9.01 | 19.21 | 19.15 | 6.91 | 30.87 | 86.90 | 89.44 | 88.12 |
| Llama-3.2-3B-IT | 22.85 | 7.95 | 16.69 | 16.70 | 5.39 | 29.29 | 85.96 | 88.95 | 87.39 |
| Phi-3-Mini-IT | 25.53 | 8.88 | 19.11 | 19.08 | 6.21 | 30.18 | 86.94 | 89.11 | 87.97 |
| Gemini-2.0-Flash | 27.60 | 10.31 | 20.62 | 20.58 | 7.59 | 31.90 | 87.25 | 89.31 | 88.23 |
| Qwen-2.5-0.5B-IT | 19.13 | 5.03 | 13.77 | 13.77 | 3.40 | 24.57 | 85.84 | 87.57 | 86.66 |
| Llama-3.2-1B-IT | 21.04 | 6.51 | 14.71 | 14.71 | 4.54 | 27.55 | 85.46 | 88.35 | 86.84 |
| Gemma-3-1B-IT | 17.94 | 5.12 | 12.78 | 12.80 | 3.58 | 22.96 | 80.22 | 88.12 | 83.78 |
| DS-Qwen-1.5B | 20.26 | 5.24 | 14.74 | 14.76 | 3.88 | 24.41 | 85.75 | 87.75 | 86.70 |
| Gemma-2-2B-IT | 24.01 | 7.98 | 18.03 | 18.05 | 7.10 | 30.53 | 86.65 | 89.39 | 87.97 |
| Llama-3.2-3B-IT | 22.35 | 6.94 | 16.04 | 16.03 | 5.40 | 30.01 | 85.82 | 88.93 | 87.31 |
| Phi-3-Mini-IT | 25.30 | 7.90 | 19.16 | 19.18 | 6.79 | 30.77 | 86.94 | 89.26 | 88.05 |
| Gemini-2.0-Flash | 27.02 | 9.10 | 20.12 | 20.16 | 7.61 | 32.52 | 87.01 | 89.30 | 88.10 |

Table 24: Results computed via automatic metrics for baseline summarization model outputs. Note that the top half displays results from the InternalA dataset, the middle half displays results for the MTS-Dialog first test split, and the bottom half displays results for the MTS-Dialog second test split. Results are rounded to 2 decimal places.

| Model | ROUGE | | | | BLEU | METEOR | BERTScore | | |
|------------------|-------|-------|-------|--------|-------|--------|-----------|--------|-------|
| | R-1 | R-2 | R-L | R-LSum | | | Precision | Recall | F1 |
| Qwen-2.5-0.5B-IT | 34.89 | 10.93 | 19.06 | 28.81 | 5.61 | 31.31 | 86.12 | 84.51 | 85.29 |
| Llama-3.2-1B-IT | 41.89 | 14.72 | 22.99 | 34.48 | 8.97 | 35.51 | 87.23 | 85.24 | 86.21 |
| Gemma-3-1B-IT | 38.05 | 12.15 | 22.44 | 32.06 | 7.20 | 27.94 | 87.42 | 84.31 | 85.83 |
| DS-Qwen-1.5B | 36.67 | 11.65 | 21.08 | 30.06 | 6.86 | 27.74 | 87.58 | 84.37 | 85.94 |
| Gemma-2-2B-IT | 41.03 | 14.55 | 24.10 | 34.57 | 9.56 | 33.58 | 87.66 | 84.92 | 86.26 |
| Llama-3.2-3B-IT | 44.08 | 17.21 | 24.90 | 36.52 | 10.34 | 39.72 | 87.37 | 85.96 | 86.66 |
| Phi-3-Mini-IT | 36.77 | 9.58 | 19.92 | 29.62 | 5.09 | 30.48 | 86.28 | 84.91 | 85.59 |
| Qwen-2.5-0.5B-IT | 17.53 | 4.80 | 11.23 | 11.51 | 2.49 | 23.19 | 83.88 | 87.63 | 85.67 |
| Llama-3.2-1B-IT | 17.09 | 4.58 | 11.13 | 11.09 | 2.52 | 21.74 | 84.05 | 87.74 | 85.81 |
| Gemma-3-1B-IT | 19.94 | 5.85 | 14.20 | 14.16 | 3.66 | 23.19 | 85.30 | 88.30 | 86.73 |
| DS-Qwen-1.5B | 20.18 | 5.74 | 14.71 | 14.68 | 3.01 | 24.63 | 85.72 | 88.00 | 86.80 |
| Gemma-2-2B-IT | 21.57 | 6.77 | 15.23 | 15.18 | 4.47 | 27.67 | 85.76 | 88.65 | 87.14 |
| Llama-3.2-3B-IT | 18.45 | 5.71 | 12.27 | 12.25 | 3.45 | 23.81 | 84.32 | 88.23 | 86.19 |
| Phi-3-Mini-IT | 19.02 | 5.16 | 12.74 | 12.72 | 2.70 | 24.14 | 84.48 | 88.44 | 86.38 |
| Qwen-2.5-0.5B-IT | 16.47 | 3.82 | 10.44 | 10.67 | 2.05 | 22.77 | 83.72 | 87.45 | 85.50 |
| Llama-3.2-1B-IT | 16.08 | 3.91 | 10.28 | 10.26 | 2.53 | 21.44 | 83.93 | 87.63 | 85.70 |
| Gemma-3-1B-IT | 19.61 | 5.39 | 14.10 | 14.13 | 4.15 | 24.46 | 85.21 | 88.31 | 86.69 |
| DS-Qwen-1.5B | 20.58 | 5.22 | 14.45 | 14.45 | 3.83 | 26.43 | 85.66 | 88.14 | 86.84 |
| Gemma-2-2B-IT | 21.06 | 6.68 | 14.75 | 14.79 | 5.48 | 28.04 | 85.63 | 88.71 | 87.10 |
| Llama-3.2-3B-IT | 17.63 | 4.95 | 11.49 | 11.51 | 3.28 | 24.59 | 84.18 | 88.23 | 86.12 |
| Phi-3-Mini-IT | 18.52 | 4.69 | 12.18 | 12.21 | 2.86 | 24.74 | 84.31 | 88.53 | 86.33 |

Table 25: Results computed via automatic metrics for summaries generated by blending via iterative feedback. Note that the top half displays results from the InternalA dataset, the middle half displays results for the MTS-Dialog first test split, and the bottom half displays results for the MTS-Dialog second test split. Results are rounded to 2 decimal places.

| Model | ROUGE | | | | BLEU | METEOR | BERTScore | | |
|------------------|-------|------|-------|--------|------|--------|-----------|--------|-------|
| | R-1 | R-2 | R-L | R-LSum | | | Precision | Recall | F1 |
| Qwen-2.5-0.5B-IT | 13.82 | 2.78 | 9.62 | 9.72 | 1.52 | 19.23 | 83.78 | 86.65 | 85.15 |
| Llama-3.2-1B-IT | 16.99 | 5.45 | 11.47 | 11.45 | 3.10 | 21.76 | 84.16 | 87.61 | 85.81 |
| Gemma-3-1B-IT | 15.47 | 4.33 | 10.92 | 11.10 | 2.51 | 19.42 | 78.12 | 87.25 | 80.53 |
| DS-Qwen-1.5B | 21.43 | 6.64 | 16.31 | 16.30 | 4.58 | 26.39 | 86.47 | 87.96 | 87.16 |
| Gemma-2-2B-IT | 21.33 | 6.87 | 15.28 | 15.28 | 4.24 | 27.01 | 85.42 | 88.97 | 87.13 |
| Llama-3.2-3B-IT | 17.73 | 5.65 | 12.06 | 12.07 | 3.41 | 23.56 | 84.00 | 88.34 | 86.08 |
| Phi-3-Mini-IT | 14.34 | 3.32 | 9.47 | 9.52 | 1.46 | 19.04 | 82.90 | 87.35 | 85.03 |
| Qwen-2.5-0.5B-IT | 13.13 | 3.16 | 9.33 | 9.42 | 1.56 | 18.38 | 83.68 | 86.52 | 85.03 |
| Llama-3.2-1B-IT | 16.60 | 4.78 | 11.30 | 11.28 | 3.19 | 22.32 | 83.94 | 87.77 | 85.77 |
| Gemma-3-1B-IT | 10.84 | 2.83 | 7.37 | 7.60 | 2.62 | 14.14 | 54.40 | 84.82 | 56.02 |
| DS-Qwen-1.5B | 20.05 | 5.74 | 15.03 | 15.06 | 4.49 | 25.99 | 86.14 | 87.77 | 86.90 |
| Gemma-2-2B-IT | 20.42 | 5.99 | 14.34 | 14.38 | 4.30 | 27.19 | 85.34 | 88.87 | 87.04 |
| Llama-3.2-3B-IT | 17.23 | 5.36 | 11.84 | 11.83 | 3.51 | 24.16 | 84.01 | 88.44 | 86.13 |
| Phi-3-Mini-IT | 14.00 | 3.01 | 9.16 | 9.20 | 1.51 | 19.04 | 82.72 | 87.27 | 84.90 |

Table 26: Results computed via automatic metrics for summaries generated by multiagent debate. Note that the top half displays results from the MTS-Dialog first test split, and the bottom half displays results for the MTS-Dialog second test split. Results are rounded to 2 decimal places.

| Model | Agent & Round Settings | | | | | | | | |
|--------------------|------------------------|-----------------|-----------------|----------------|----------|-----------------|-----------------|-----------------|-----------------|
| | A2R1 | A2R2 | A2R3 | A3R1 | A3R2 | A3R3 | A4R1 | A4R2 | A4R3 |
| Phi-3.5-Mini-IT | (12,3,5) | (14,1,5) | (16,3,1) | (14,4,2) | (11,5,4) | (14,2,4) | (12,5,3) | - | - |
| Phi-4-Mini-IT | (4,12,4) | (4,12,4) | (3,15,2) | (6,9,5) | (4,11,5) | (2,12,6) | (7,10,3) | (3,12,5) | (3,13,4) |
| Gemma-3-4B-IT | (18,0,2) | (15,2,3) | (18,1,1) | (17,0,3) | (17,0,3) | (19,0,1) | (17,1,2) | (19,0,1) | (18,1,1) |
| Mistral-7B-IT-v0.3 | (17,2,1) | (17,2,1) | (15,3,2) | (17,0,3) | (16,2,2) | (18,2,0) | (17,0,3) | (18,1,1) | (17,2,1) |
| DS-Qwen-7B | (16,2,2) | (16,3,1) | (18,2,0) | (15,2,3) | (13,3,4) | (15,2,3) | (16,2,2) | (16,2,2) | (16,2,2) |
| Llama-3.1-8B-IT | (19,1,0) | (19,1,0) | (19,1,0) | (18,1,1) | (19,1,0) | (18,1,1) | (18,1,1) | (18,1,1) | (19,0,1) |
| DS-Llama-8B | (17,1,2) | (18,2,0) | (17,2,1) | (17,2,1) | (18,0,2) | (16,2,2) | (14,2,4) | (17,1,2) | (20,0,0) |
| Gemma-2-9B-IT | (12,5,3) | (13,1,6) | (12,2,6) | (15,3,2) | (14,2,4) | (15,0,5) | (15,2,3) | (16,0,4) | (17,1,2) |

Table 27: Winrate results for each combination of agent and round settings, computed for a random split of 20 samples drawn from the InternalA dataset. Each result is in the format (A, B, C) , where A, B, C denote the number of model outputs that are preferred, the number of Gemini-2.0-Flash outputs that are preferred, and the number of samples that are comparable, respectively. Note that A2R2 for example denotes the setting using 2 agents and 2 rounds. The best performing settings for each model are **bolded**.

| Model | ROUGE | | | | BLEU | METEOR | BERTScore | | |
|--------------------|-------|-------|-------|--------|-------|--------|-----------|--------|-------|
| | R-1 | R-2 | R-L | R-LSum | | | Precision | Recall | F1 |
| Phi-3.5-Mini-IT | 39.80 | 12.54 | 21.97 | 32.04 | 7.03 | 34.58 | 86.24 | 85.21 | 85.72 |
| Phi-4-Mini-IT | 35.54 | 10.47 | 19.16 | 29.47 | 5.34 | 31.96 | 85.87 | 85.24 | 85.54 |
| Gemma-3-4B-IT | 42.91 | 16.11 | 26.10 | 36.41 | 10.79 | 33.07 | 87.56 | 85.38 | 86.45 |
| Mistral-7B-IT-v0.3 | 41.44 | 15.91 | 23.48 | 35.14 | 9.84 | 36.15 | 87.55 | 85.24 | 86.37 |
| DS-Qwen-7B | 40.75 | 13.95 | 23.37 | 35.97 | 9.06 | 33.08 | 87.42 | 85.09 | 86.23 |
| Llama-3.1-8B-IT | 40.15 | 14.90 | 22.80 | 36.12 | 8.82 | 35.93 | 86.88 | 86.19 | 86.43 |
| DS-Llama-8B | 40.07 | 14.74 | 23.51 | 33.67 | 9.05 | 33.14 | 87.55 | 84.99 | 86.24 |
| Gemma-2-9B-IT | 41.16 | 15.07 | 25.56 | 35.27 | 10.11 | 31.97 | 88.13 | 85.37 | 86.72 |

Table 28: Results computed via automatic metrics for summaries generated by baseline approaches, computed for the InternalA dataset. Note that results are rounded to 2 decimal places.