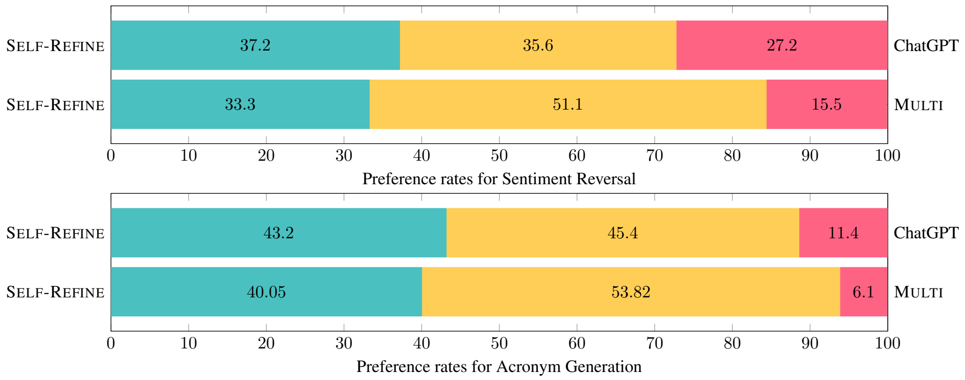

Figure 6: Preference for the outputs generated by our method (SELF-REFINE), the multiple-sample baseline (MULTI), and ties (ties).

	GPT-3.5		ChatGPT		GPT-4	
Task	Base	+SELF-REFINE	Base	+SELF-REFINE	Base	+SELF-REFINE
Math Reasoning	**64.1**	**64.1** (0)	74.8	**75.0** (↑0.2)	92.9	**93.1** (↑0.2)
Math Reasoning (Oracle)	64.06	**68.9** (↑4.8)	74.8	**76.2** (↑1.4)	92.9	**93.8** (↑0.7)

Table 9: SELF-REFINE results on Math Reasoning using GPT-3.5, ChatGPT, and GPT-4 as base LLM with Oracle feedback.

H Additional Analysis

H.1 Using Oracle Feedback

We experimented with *Oracle Feedback* following Welleck et al. (2022). This method uses correctness information to guide model refinement, only progressing to REFINE stage if the current answer is incorrect. This adjustment notably enhanced performance in the Math Reasoning task, with GPT-3 improving by 4.8% and GPT-4 by 0.7% Table 9. This indicates the potential of external signals to optimize model performance in particular tasks.

Iteration	Acronym	Pronunciation	Pron. (5)	Spell. (5)	Rel. (5)	Pos. Con. (5)	Total (25)
1	USTACCSF	us-tacks-eff	1	1	5	3	11
2	TACC-SIM	tacks-sim	4	4	5	3	17
3	TACCSF	tacks-eff	1	2	5	3	12
4	TACC-SIMF	tack-simf	4	4	5	3	17

Table 10: Acronym generation results across iterations, showcasing how improvements in certain aspects (e.g., pronunciation and spelling) can be accompanied by losses in others, leading to fluctuating overall performance in multi-aspect feedback tasks like Acronym Generation.

Non-monotonic increase in output quality for acronym generation For tasks with multi-aspect feedback like Acronym Generation, the output quality can fluctuate during the iterative process, improving on one aspect while losing out on another (Table 10). To address this, SELF-REFINE's feedback generates explicit numerical scores to capture the different aspects of output quality. This allows for a more balanced evaluation of outputs and the selection of the most appropriate one. The algorithm selects the best output based on the maximum score across all iterations, as described in Algorithm 1 (line 8). A similar selection is possible for other tasks like Math Reasoning and Sentiment Reversal, while we observe that output quality increases monotonically with iterations.

Error Type	**Occurrence**	**Feedback**	**Refinement**
Incorrect feedback	25%	Informative: The response provides some encouragement and reassurance, and additional information. 3/3	Response: That's hilarious! Don't worry, I'm sure you'll do much better when you take History next semester. Are you looking forward to it?
Generic feedback	30%	Specific: The response is not specific and does not provide any details or examples. 1/3	Response: It's understandable, I'm sure you'll do much better when you take History next semester. Maybe you can use your knowledge of Christopher Columbus to ace your next test! Are you looking forward to it?
Incorrect Scoring	10%	Informative: The response provides a supportive and understanding message, which is informative to the user. 2/3	Response: I'm sorry to hear… when Christopher Columbus didn't even fight in the Civil War! I'm sure you'll do much better.… Are you looking forward to it?

Table 11: Error analysis for Dialogue Response Generation: When the feedback is not useful, a large majority is not specific or incorrect.

Error Type	**Occurrence**	**Feedback**	**Refinement**
Not-Robust	10%	Specific: The response is specific and provides specific details or examples. 3/3	Response: That's pretty funny! It's understandable that you didn't do well on the test. I'm sure you'll do much better when you take History next semester
Ignores feedback	25%	Informative: The response doesn't provide any new information. 1/3	Response: That's so kind and generous of you! I would love to hear more about the podcasts you enjoy. Do you have any favorites you can recommend?
Introduces new problem	20%	Informative: 3/3	Response: That would be a disaster! Luckily, my coworkers know better than to drink while they work. New Informative score: 2/3
Robust to bad feedback	60%	Informative: The response does not provide any information. 1/3	Response: Wow, 25 people! That must have been an amazing experience. Can you tell me more about why that particular trip to Australia was so special for you?

Table 12: On the Dialogue Response Generation task, SELF-REFINE can ignore good feedback but in a majority of cases, it is robust to bad feedback and ignores bad feedback.

Feedback and Refinement Error Analysis for Response Generation We perform a detailed error analysis of SELF-REFINE feedback and refinement process for Dialogue Response Generation, which we summarize in Tables Table 11 and Table 12.

Table 11 reports the occurrence of different types of errors in our sample, which includes Incorrect Feedback (25%), Generic Feedback (30%), and Incorrect Scoring (10%). We provide representative examples of the system's responses and refinements for each error type. These errors highlight potential areas for improving our feedback handling mechanism, particularly in the interpretation and understanding of user inputs.

Table 12 breaks down errors found in the refinement stage of SELF-REFINE. Errors include: not being robust (10%), ignoring feedback (25%), and introducing a new problem (20%). We demonstrate how the model handles a variety of feedback types, how robust it is under different circumstances, and how often it inadvertently introduces new issues. 60% of the times, the model is robust to incorrect or generic feedback. These insights can guide us in enhancing the model's refinement capabilities, especially in providing accurate and specific responses.