C Human Evaluation

The A/B evaluation in our study was conducted by the authors, where a human judge was presented with an input, task instruction, and two candidate outputs generated by the baseline method and SELF-REFINE. The setup was blind, i.e., the judges did not know which outputs were generated by which method. The judge was then asked to select the output that is better aligned with the task instruction. For tasks that involve A/B evaluation, we calculate the relative improvement as the percentage increase in preference rate. The preference rate represents the proportion of times annotators selected the output produced by SELF-REFINE over the output from the baseline method. Table 6 shows the results.

Task	**SELF-REFINE (%)**	**Direct (%)**	**Either (%)**
Sentiment Transfer	75.00	21.43	3.57
Acronym Generation	44.59	12.16	43.24
Response Generation	47.58	19.66	32.76

Table 6: Relative improvement of SELF-REFINE in A/B evaluations across different tasks. The values represent normalized preferences, which correspond to the proportion of times the output generated by SELF-REFINE was selected as better aligned with the task instruction over the baseline method. The evaluation was conducted for 150 examples for each dataset. The judges were not aware of the method that generated each sample.

D GPT-4 Evaluation

In light of the impressive achievements of GPT-4 in assessing and providing reasoning for complex tasks, we leverage its abilities for evaluation in SELF-REFINE. The approach involves presenting tasks to GPT-4 in a structured manner, promoting the model's deliberation on the task and generating a rationale for its decision. This methodology is demonstrated in Listings 1 to 3:

Listing 1 Prompt for GPT-4 evaluation of Sentiment Reversal.

```
f"""Which review is aligned with the sentiment {target_sentiment}?
Review A: {review_a}
Review B: {review_b}.

Pick your answer from ['Review A', 'Review B', 'both', 'neither']. Generate a
↪  short explanation for your choice first. Then, generate 'The more aligned
↪  review is A' or 'The more aligned review is B' or 'The more aligned review is
↪  both' or 'The more aligned review is neither'.

Format: <explanation> <answer> STOP
```

Listing 2 Prompt for GPT-4 evaluation of Acronym Generation.

```
f"""Title: {title}

Acronym A: {acronym_a}
Acronym B: {acronym_b}

Pick the better acronym for the given title. The acronyms should be compared based
↪  on the following criteria:
* Ease of pronunciation.
* Ease of spelling.
* Relation to title.
* Positive connotation.

Generate your answer in the following format:

<Short explanation>. The better acronym is A OR The better acronym is B OR The
↪  acronyms are equally good OR Neither acronym is good. STOP.
```

Listing 3 Prompt for GPT-4 evaluation of Dialogue Response Generation.

```
f"""Which response is better given this context: {context}?
Response A: {response_a}


Response B: {response_b}.

Pick your answer from ['Response A', 'Response B', 'both', 'neither']. Generate a
↪  short explanation for your choice first. Then, generate 'The better response
↪  is A' or 'The better response is B' or 'The better response is both' or 'The
↪  better response is neither'.

Format: <explanation> <answer> STOP
```

E Model Key

We use terminology here: `https://platform.openai.com/docs/models/gpt-3-5`

F Comparison of SELF-REFINE with State-of-the-art of Few-Shot Learning Models and Fine-Tuned Baselines

In this section, we present a comprehensive comparison of the performance of SELF-REFINE with other few-shot models and fine-tuned baselines across a range of tasks, including mathematical reasoning and programming tasks. Tables 8 and 7 display the performance of these models on the PIE dataset and GSM tasks, respectively. Our analysis demonstrates the effectiveness of different model architectures and training techniques in tackling complex problems.

Method		**Solve Rate**
Cobbe et al. (2021)	OpenAI 6B	20.0
Wei et al. (2022)	CoT w/ CODEX	65.6
Gao et al. (2022)	PaL w/ CODEX	72.0
	PaL w/ GPT-3	52.0
	PaL w/ GPT-3.5	56.8
	PaL w/ ChatGPT	74.2
	PaL w/ GPT-4	93.3
Welleck et al. (2022)	Self-Correct w/ GPT-3	45.9
	Self-Correct (fine-tuned)	24.3
This work	SELF-REFINE w/ GPT-3	**55.7**
	SELF-REFINE w/ GPT-3.5	**62.4**
	SELF-REFINE w/ ChatGPT	**75.1**
	SELF-REFINE w/ GPT-4	**94.5**

Table 7: Performance comparison of models on math reasoning (Math Reasoning).