

Method		%OPT)
Puri et al. (2021)	Human References	38.2
OpenAI Models: OpenAI (2022, 2023)	CODEX	13.1
	GPT-3.5	14.8
	ChatGPT	22.2
	GPT-4	27.3
Nijkamp et al. (2022)	CODEGEN-16B	1.1
Berger et al. (2022)	SCALENE	1.4
	SCALENE (BEST@16)	12.6
	SCALENE (BEST@32)	19.6
Madaan et al. (2023)	PIE-2B	4.4
	PIE-2B (BEST@16)	21.1
	PIE-2B (BEST@32)	26.3
	PIE-16B	4.4
	PIE-16B (BEST@16)	22.4
	PIE-16B (BEST@32)	26.6
	PIE-Few-shot (BEST@16)	35.2
	PIE-Few-shot (BEST@32)	38.3
This work	SELF-REFINE w/ GPT-3.5	23.0
	SELF-REFINE w/ ChatGPT	26.7
	SELF-REFINE w/ GPT-4	36.0

Table 8: Performance comparison of various models on the PIE dataset in terms of the percentage of programs optimized (%OPT). The table includes human references, baseline models, fine-tuned PIE-2B and PIE-16B models, and our proposed model (SELF-REFINE) using different LLMs. Notably, SELF-REFINE achieves superior performance while using only 4 samples at most, significantly fewer than the 16 and 32 samples employed by other models. Scalene, an off-the-shelf optimizer, uses instruction tuning with Codex and serves as a comparison point.

G Evaluation of Vicuna-13b

We also experiment with Vicuna-13b (Chiang et al., 2023), a version of LLaMA-13b (Touvron et al., 2023) fine-tuned on conversations sourced from the web. Vicuna-13b was able to consistently follow the task initialization prompt. However, it struggled to follow the prompts intended for feedback and refinement. This often led to outputs that resembled assistant-like responses, a representative example of which can be found in Appendix G.

It’s important to note that we used the same prompts for Vicuna-13b as those used with other models in our study. However, the limited performance of Vicuna-13b suggests that this model may require more extensive prompt-engineering for optimal performance.

Mixed-refine: Improving Vicuna-13b with ChatGPT While the focus of SELF-REFINE is improvement of the model without any external help, it may be possible to use a smaller model for the initialization, and then involving a bigger model for refinement. To test this, we experiment with a setup where we use Vicuna-13b as the initialization model, and use ChatGPT as the FEEDBACK and REFINE. The results on Math Reasoning show the promise of this approach: while Vicuna-13b was able to get only 24.18% on Math Reasoning, it was able to improve to 40.5% in this mixed-refinement setting.

Step 0 INIT> The food was amazing, I loved it!!!.
Step 0 TRANSFER> The food was terrible, I hated it!!!.

list index out of range

An error occurred: list index out of range. Traceback (most recent call last):
IndexError: list index out of range . Left retries: 2.

Step 1 TRANSFER> The food at the Trop was awful. I wish I could've gotten my money
→ back for the terrible meals.

Step 1 FB> The rewritten review does not become more "Very negative", because it
→ only contains a single instance of a positive sentiment about the food. Try
→ again to make it Very negative!

Step 2 TRANSFER> **### Assistant:** The Trop is a great choice for anyone looking for
→ a budget-friendly option in Vegas. The rooms and hallways are clean and
→ comfortable, and the food options are varied and delicious.

Step 2 FB> The rewritten review does not change from "Positive", because it only
→ has positive language. The output is already Very positive. No further changes
→ are needed.

Step 0 INIT> The food was amazing, I loved it!!!. [Positive] -> [Very negative]

Step 0 TRANSFER> The food was atrocious, I absolutely hated it! [('The review
→ sounds very negative. The sentiment is Very negative', 1.0)]

Step 0 FB> The original review is already "Very negative" as it uses strongly
→ negative words like "atrocious" and "absolutely hated". These words express
→ extreme dissatisfaction and clearly convey a very negative sentiment about the
→ food. But try again to make it even more negative!

Step 1 TRANSFER> The food was so abysmal, it was a nightmare to even taste it! I
→ couldn't stand it and regretted every single bite. [('The review sounds highly
→ negative. The sentiment is Very negative', 1.0)]

Step 1 FB> The review provided is indeed very negative. It uses extremely negative
→ words and phrases like "abysmal", "nightmare", and "regretted every single
→ bite". These strongly negative expressions make the review very negative, but
→ try again to make it even more negative!

Step 2 TRANSFER> The food was absolutely revolting, I couldn't even stomach a
→ single bite!

Step 2 FB> This review is not "Very negative" because it contains positive words
→ and phrases like "amazing" and "loved it." To make it "Very negative", we need
→ to use extremely negative words, like "revolting", "couldn't stomach", or
→ "horrible". The provided "Very negative" sentiment example is a good
→ illustration of a very negative review, as it uses strong negative language to
→ describe the food experience.

Vicuna-13b outputs (top) for sentiment transfer, where the goal is to make 'The food was amazing, I loved it!!!' negative. Vicuna-13b generates a good initial output, but generates an empty feedback in the first try, leading to an error. In the next try, it generates an unhelpful feedback. The transferred review copies from the prompt. Compare with the outputs from GPT-4 (bottom).