Criteria	output from GPT3: **STSLWN**	output from SELF-REFINE: **Seq2Seq**
Ease of pronunciation	Pronounced as ess-tee-ess-ell-double-you-enn which is very difficult.	Pronounced as seq-two-seq which is easy.
Ease of spelling	Very difficult to spell.	Easy to spell.
Relation to title	No relation to the title.	Mentions sequence which is somewhat related to the title.
Positive connotation	Meaningless acronym.	Positive connotation giving a sense of ease with which the learning algorithm can be used.
Well-known	Not a well-known acronym.	Close to the word sequence which is a well-known word.
Total score	5/25	20/25

Table 18: Comparison of acronyms for input = “Sequence to Sequence Learning with Neural Networks”

R Constrained Generation

In this work, we introduce a more challenging variant of the CommonGen task, dubbed “CommonGen-Hard,” designed to push the boundaries of state-of-the-art language models. CommonGen-Hard requires models to generate coherent and grammatically correct sentences incorporating 20-30 concepts, as opposed to the original task which presents a set of 3-5 related concepts. This significant increase in the number of concepts tests the model’s ability to perform advanced commonsense reasoning, contextual understanding, and creative problem-solving, as it must generate meaningful sentences that encompass a broader range of ideas. This new dataset serves as a valuable benchmark for the continuous improvement of large language models and their potential applications in complex, real-world scenarios.

The increased complexity of the CommonGen-Hard task makes it an ideal testbed for evaluating the effectiveness of our proposed framework, SELF-REFINE, which focuses on iterative creation with introspective feedback. Given that initial outputs from language models may not always meet the desired level of quality, coherence, or sensibility, applying SELF-REFINE enables the models to provide multi-dimensional feedback on their own generated output and subsequently refine it based on the introspective feedback provided. Through iterative creation and self-reflection, the SELF-REFINE framework empowers language models to progressively enhance the quality of their output, closely mimicking the human creative process and demonstrating its ability to improve generated text on complex and demanding natural language generation tasks like CommonGen-Hard (Figure 15).

S Prompts

We include all the prompts used in the experiments in Figures 16-35:

- **Acronym Generation:** Figures 16-18
- **Code Optimization:** Figures 19-21
- **Code Readability Improvement:** Figures 22-23
- **Constrained Generation:** Figures 24-26
- **Dialogue Response Generation:** Figures 27-29
- **Math Reasoning:** Figures 30-32
- **Sentiment Reversal:** Figures 33-35

Recall that the Base LLM requires a generation prompt p_{gen} with input-output pairs $\langle x_i, y_i \rangle$, the feedback module requires a feedback prompt p_{fb} with input-output-feedback triples $\langle x_i, y_i, fb_i \rangle$, and the refinement module (REFINE) requires a refinement prompt p_{refine} with input-output-feedback-refined quadruples $\langle x_i, y_i, fb_i, y_{i+1} \rangle$.

- **Sentiment Reversal** We create positive and negative variants of a single review from the training set and manually write a description for converting the negative variant to positive

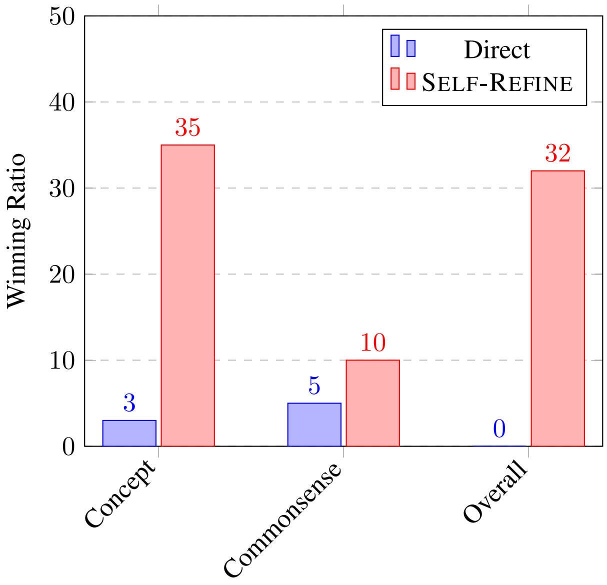

Figure 15: A comparison of SELF-REFINE and direct generation with GPT-3.5 on CommonGen-Hard.

and vice versa. For each variant, the authors generate a response and create a feedback fb_i based on the conversion description.

- **Dialogue Response Generation** We sample six examples as $\langle x_i, y_i \rangle$ for the few-shot prompt for the Base LLM. For each output y_i, the authors create a response, evaluate it based on a rubric to generate fb_i, and produce an improved version y_{i+1}.
- **Acronym Generation** We provide the Base LLM with a total of 15 (title, acronym) examples. Then, for one title (x_i) we generate an acronym (y_i) using ChatGPT. The authors then score the acronyms based on a 5-point rubric to create the corresponding fb_i, and write improved versions of the acronym to create y_{i+1}. 3 such examples are used for REFINE and FEEDBACK.
- **Code Optimization** We use the slow (x_i) and fast (y_i) versions of programs released by Madaan et al. (2023) for Base LLM. We use their provided explanations (Madaan et al., 2023) for FEEDBACK and REFINE.
- **Math Reasoning** The prompts for the Base LLM are sourced from PaL (Gao et al., 2022) as $\langle x_i, y_i \rangle$. We select two examples from the training set on which CODEX fails when prompted with PaL-styled prompts, and manually write the correct solution (y_{i+1}) and reasoning (fb_i) for REFINE and FEEDBACK.
- **Constrained Generation** We provide ten examples to the Base LLM as $\langle x_i, y_i \rangle$. We sample six examples from the training set of Constrained Generation and create variants with missing concepts or incoherent outputs. The missing concepts and the reason for incoherence form fb.
- **TODO:** Add relevant information for the remaining task.

```
Title: A Survey of Active Network Research
Acronym: SONAR

Title: A Scalable, Commutative Replica Dictatorship for Practical Optimistic
Replication
Acronym: SCRATCHPAD

Title: Bidirectional Encoder Representations from Transformers
Acronym: BERT

Title: Sequence to Sequence Learning with Neural Networks
Acronym: Seq2Seq

Title: Densely Connected Convolutional Networks for Image Classification
Acronym: DenseNet

Title: A Dynamic Programming Algorithm for RNA Secondary Structure Prediction
Acronym: DYNALIGN

Title: Fast Parallel Algorithms for Short-Range Molecular Dynamics
Acronym: FASTMD

Title: Real-Time Collaborative Editing Systems
Acronym: COCOON

Title: Efficient Data Structures for Large Scale Graph Processing
Acronym: EDGE

Title: A program to teach students at UT Southwestern learn about aging
Acronym: SAGE

Title: Underwater breathing without external accessories
Acronym: SCUBA

Title: An educational training module for professionals
Acronym: LEAP

Title: Teaching a leadership program
Acronym: LEAD
```

Figure 16: Initial generation prompt for Acronym Generation