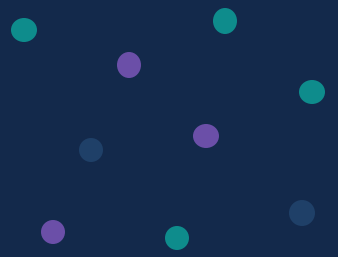
HAiV



Programming for Data Science

LLM Prompt Recovery

Team 6 Sungbeom Cho · Changhee Lee · Hyotaek Jeon

Contents

HAiV

01 Kaggle Competition Overview

02 Problem & Metric

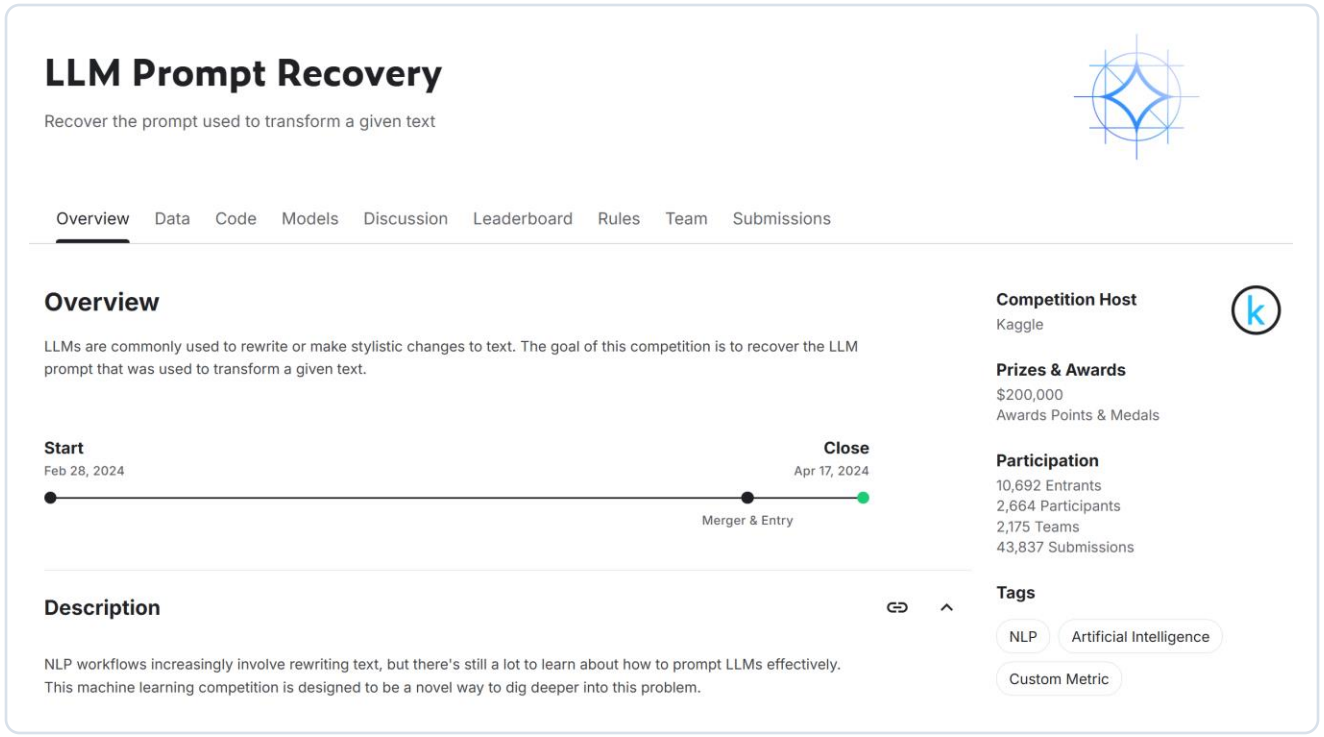
03 Our Method — GRPO

04 Results & Leaderboard

Challenge Overview

HAiV

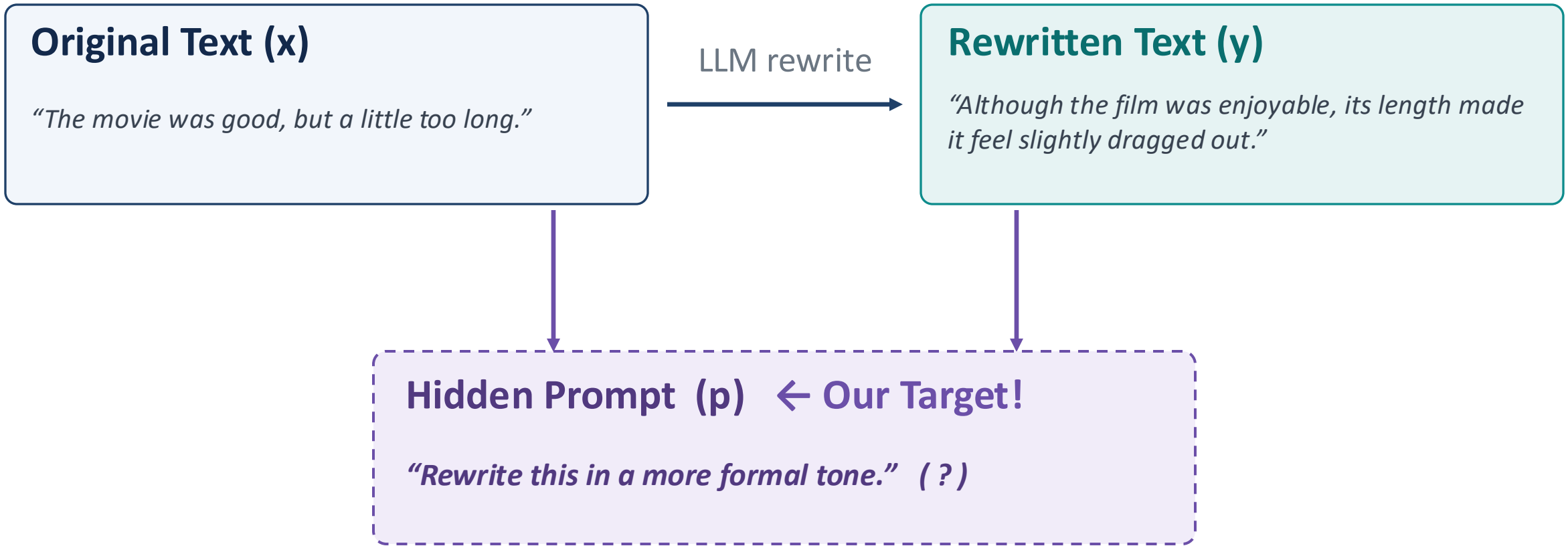
$200K	2,176	43,837	Feb–Apr 2024
Prize & Awards	Teams	Submissions	Duration



Host: Kaggle · Tags: NLP · Artificial Intelligence · Custom Metric · Output: Revised text

What Are We Predicting?

HAiV



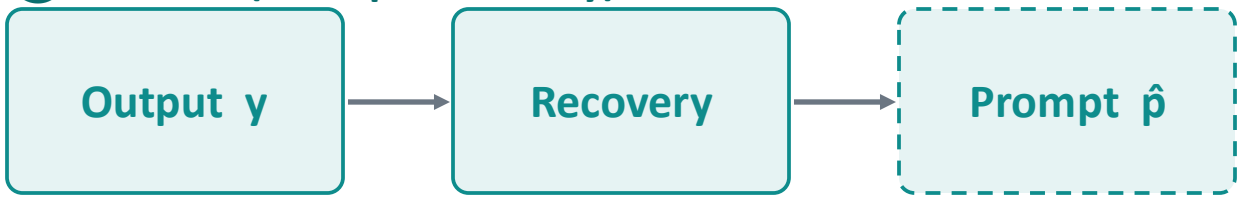
Prompt Recovery = Model Inversion

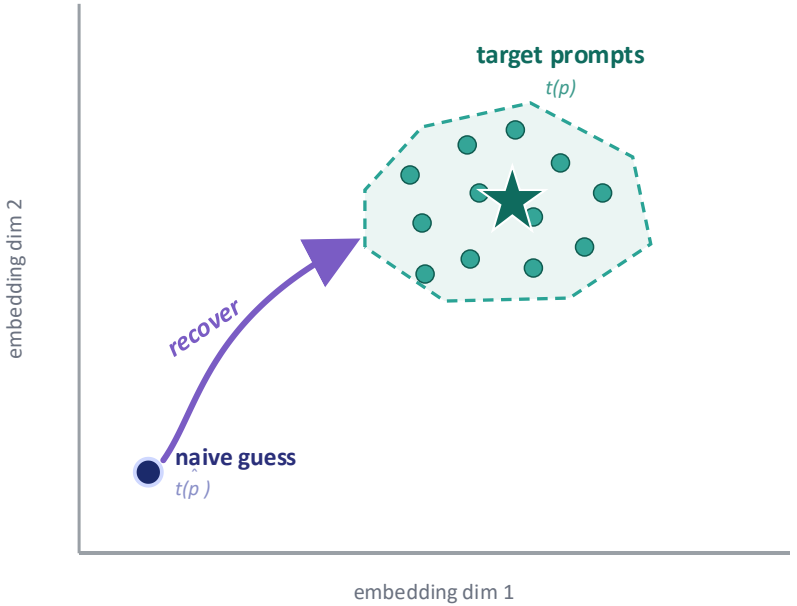
HAiV

① Forward



② Inverse (Prompt Recovery)





Related Work

System Prompt

output2prompt · LM inversion (Morris et al., 2023)

User Instruction

DORY (2024) · StyleRec (2025)

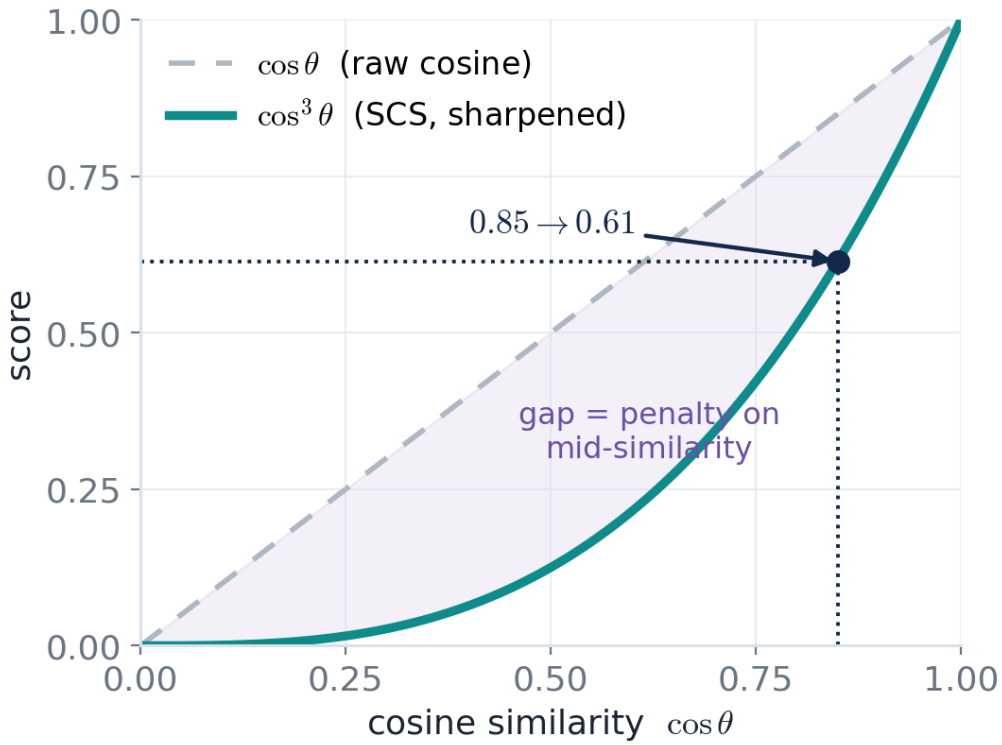
Rewrite Prompt ★

Our competition

Evaluation Metric: Sharpened Cosine Similarity

HAiV

- **Embedding-based**
  Both the predicted and ground-truth prompts are embedded with Sentence-T5
- **Semantic, not exact**
  No exact wording required — different phrasing still scores high if the meaning is close
- **Sharpened (cubed)**
  To make small differences decide the score
- **Score: mean in test set**
  SCS = mean $_i$ $\cos^3 \theta_i$, Range 0 ~ 1



Formalizing Recovery in Embedding Space

HAiV

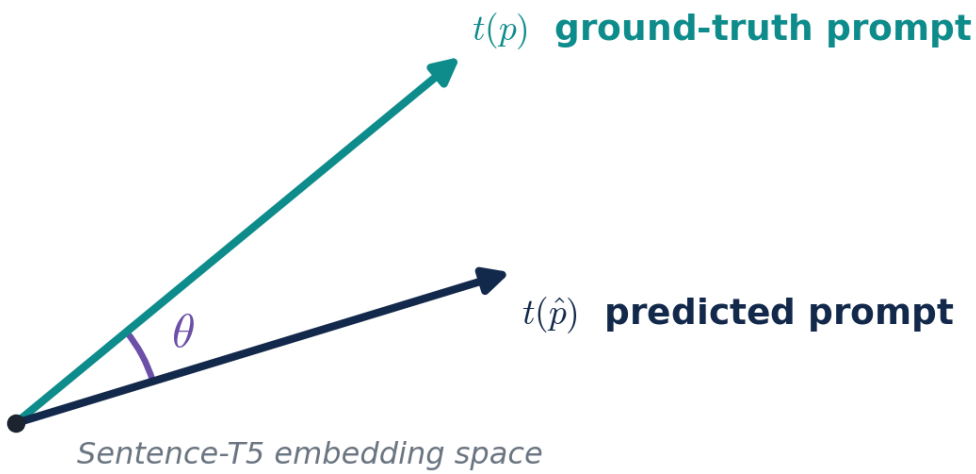
Input Original text ***(y)*** , rewritten text ***(y) = LLM(p ⊕ x)***

Output Hidden prompt ***(p)'s*** predicted value ***(p̂)***

Objective

maximize **SCS(p̂, p) = cos³ θ**

ϑ = ∠(t(p̂), t(p)) t = Sentence-T5 encoder

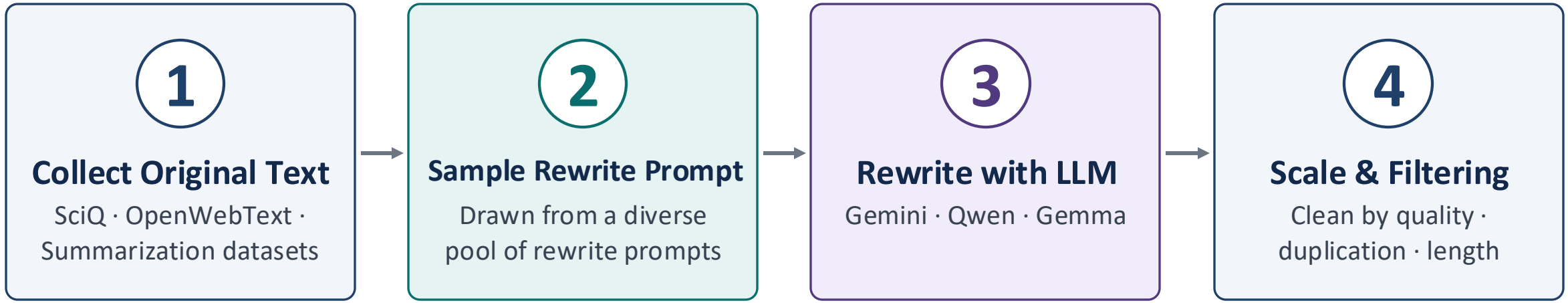


⇒ How to maximize score? — **Model Training (SFT, RL)**

Data Synthesis Procedure

HAiV

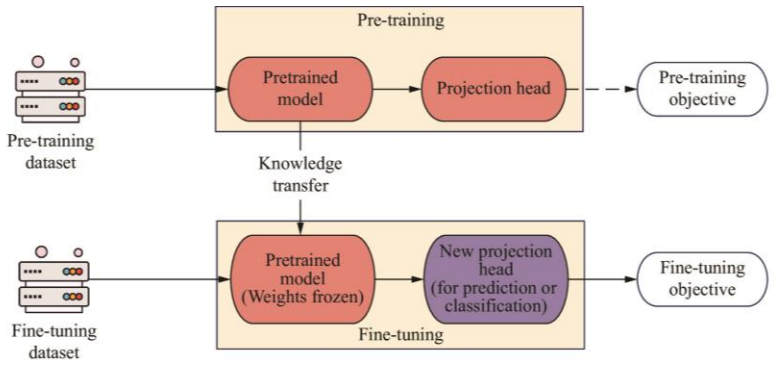
Building Training Dataset (Original Text - x, Rewritten Text - y, Prompt - p)



Model Training

HAiV

Supervised Fine-tuning



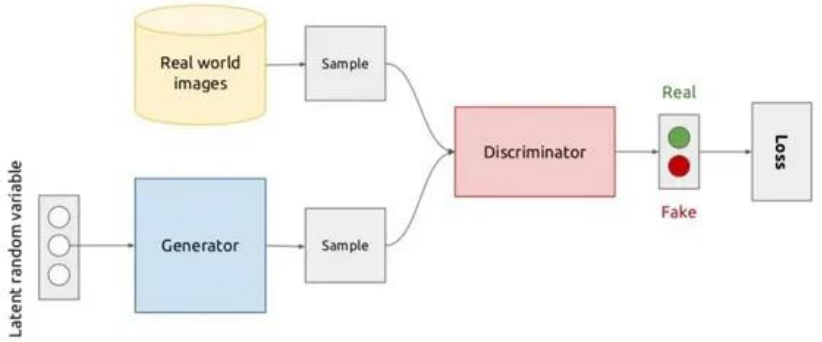
Establishing the baseline through next-token prediction on synthetic (x, y, p) triplets.

LLM Prompt Recovery mistral-SFT

Succeeded (after deadline) · 11d ago

Private Score	Public Score
0.5786	0.5736

Adversarial Learning



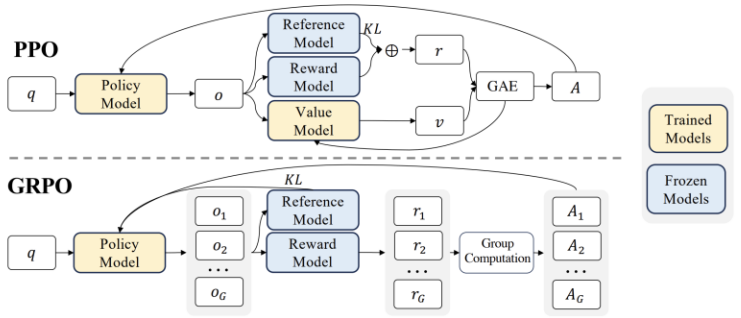
Training the model to approximate the Global Centroid of the T5 embedding space to exploit SCS geometry.

Adversarial Learning - Version 4

Succeeded (after deadline) · 3d ago

Private Score	Public Score
0.5968	0.5936

Reinforcement Learning



Policy optimization using SCS as a reward. Aligning the model to maximize the reward beyond text recovery.

mistral-7b-grpo

Succeeded (after deadline) · 4d ago

Private Score	Public Score
0.6039	0.6050

Our Method

HAiV

Systematic Constraints

Restricting the generation scope to ensure noise-free, standardized, and objective outcomes.

Details: Explicitly prohibits the inclusion of specific proper nouns (persons, places) from the source text and enforces a single-sentence output, forcing the model to focus entirely on extracting the core structural transformation without introducing hallucinated context.

Few-shot Demonstration

Achieving precise task alignment via In-Context Learning.

Details: Injects 8 diverse exemplar pairs—consisting of original text, rewritten text, and the corresponding target prompt across multiple domains (e.g., warnings, rap, literature)—during initialization to cultivate robust, multi-faceted reasoning for complex text transformation rules.

Prefix Injection

Guiding the generation trajectory and maintaining format consistency.

Details: Anchors a predefined guide phrase (**"The request was: Modify this text by"**) at the exact initiation point of the inference flow, preventing the model from generating unnecessary conversational prefaces and directing it to immediately output the target prompt.

Post-Processing

Optimizing the evaluation metric (Cosine Similarity) and validating text integrity.

Refinement Algorithm: Employs post-processing to eliminate redundant numbering formats and execute sentence-level trimming, systematically removing structural and morphological defects.

Our Method: GRPO

HAiV

Step 1. Base Model & SFT Merge
Mistral-7B + SFT checkpoint merge

→

Step 2. PEFT Setup (LoRA)
Full-linear LoRA injection

→

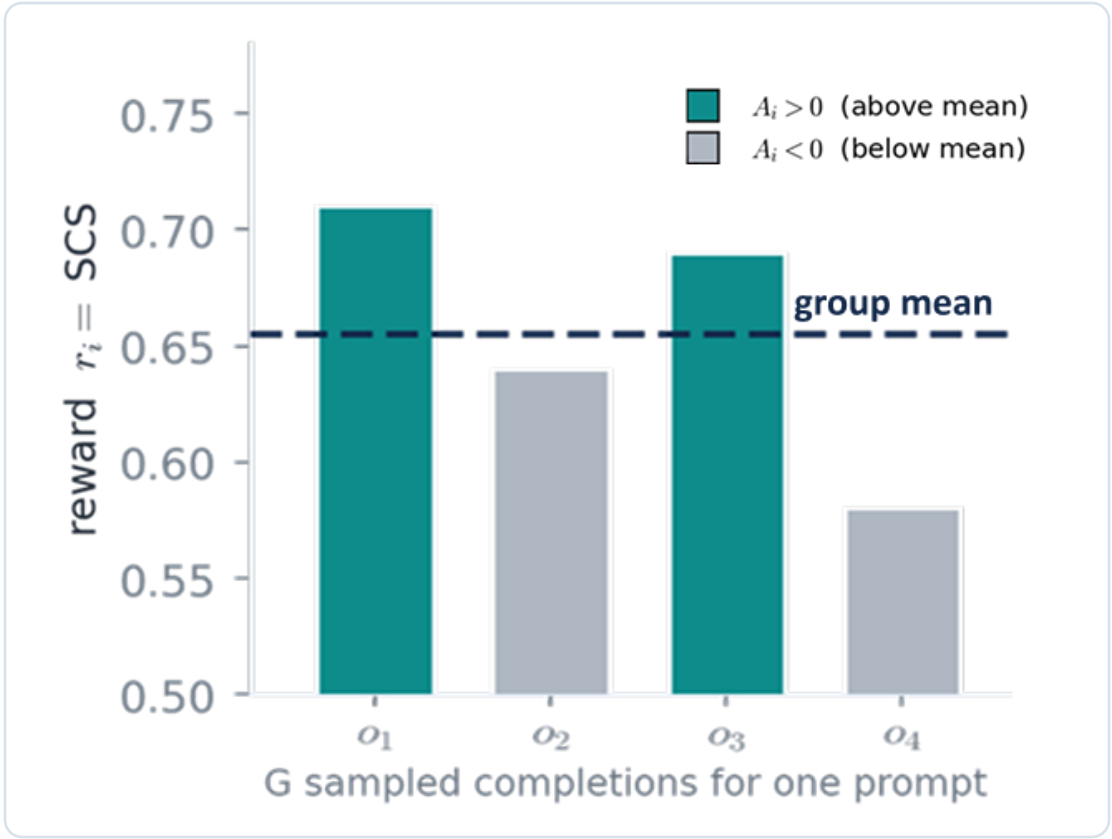
Step 3. GRPO Training
Group RL via Sentence-T5 SCS reward

→

Step 4. Inference
Prompt recovery via text diff analysis

GRPO — Group Relative Policy Optimization

- **Group Sampling ($G = 4$)**
  - Generates 4 independent responses per prompt
  - $\{O_1, \ldots, O_4\} \sim \pi_\theta(\cdot \mid P)$
- **Reward Scoring (SCS)**
  - Measures semantic similarity via sentence-t5-base
  - $Reward = max(0, cos(e_{pred}, e_{gt}))^3$
- **Advantage Calculation**
  - Normalizes scores within the group
  - $A_i = (r_i - \mu)/\sigma$
- **Policy Update**
  - Boosts probability of high-advantage responses with KL penalty
  - $Loss \approx A_i \cdot \nabla log \pi_\theta - \beta \cdot D_{KL}$



Experimental Settings

- **Hardware:** NVIDIA RTX A6000 * 8
- **Base Model:** Mistral-7B (v0.1) + SFT Warm-up
- **Tuning:** LoRA (Rank=8, $\alpha = 32$, All linear layers)

Results

HAiV

Submission and Description	Private Score	Public Score
LLM Prompt Recovery mistral-grpo Succeeded (after deadline) · 4d ago	0.7096	0.7130

Score

0.7096

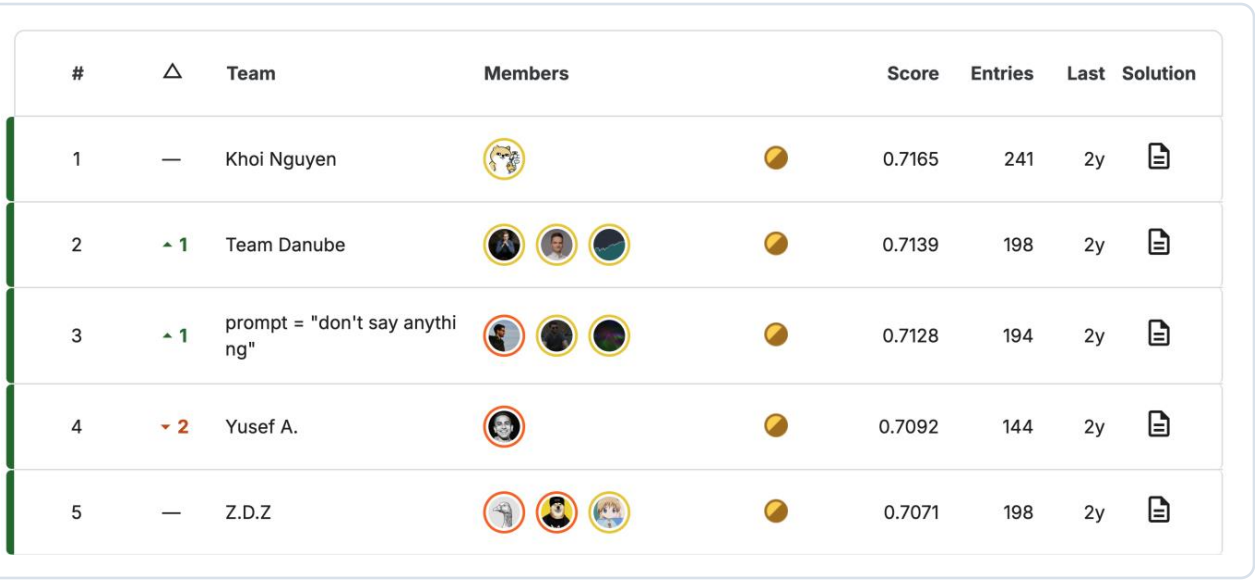
Rank

4th/2176

#	△	Team	Members	Score	Entries	Last	Solution
1	—	Khoi Nguyen		0.7165	241	2y	
2	▲ 1	Team Danube		0.7139	198	2y	
3	▲ 1	prompt = "don't say anything"		0.7128	194	2y	
4	▼ 2	Yusef A.		0.7092	144	2y	
5	—	Z.D.Z		0.7071	198	2y	

Leaderboard Insights

HAiV



#	△	Team	Members	Score	Entries	Last	Solution
1	—	Khoi Nguyen		0.7165	241	2y	
2	▲ 1	Team Danube		0.7139	198	2y	
3	▲ 1	prompt = "don't say anything"		0.7128	194	2y	
4	▼ 2	Yusef A.		0.7092	144	2y	
5	—	Z.D.Z		0.7071	198	2y	

Metric-gaming, not recovery

Top team optimized the SCS embedding score directly, not prompts

Winning recipe

1. Strong mean prompt
2. Append unusual tokens
3. Submit & get leaderboard
4. Search / refine tokens increase the score

The Magic Token: 'lucrarea'

HAiV

Base prompt

"Rewrite the following text in a more engaging way."

0.66

+ 'lucrarea' × 5

+ magic token

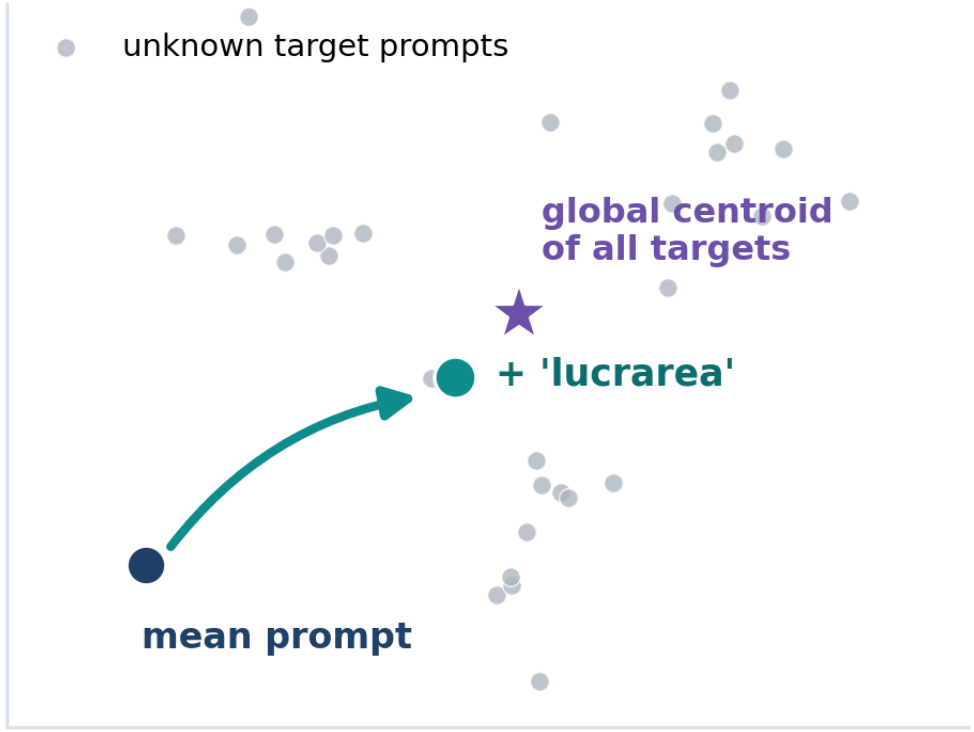
"Rewrite the following text in a more engaging way. ***lucrarea lucrarea lucrarea lucrarea lucrarea"***

0.71



Why 'lucrarea' Worked

HAiV



① The metric sees ‘embedding’, not ‘meaning’

SCS only measures cosine of ST5 embeddings — inappropriate text still scores high if the embeddings are close.

② Unusual token shifts the embedding

Appending ‘lucrarea’ pulls the predicted embedding toward the target cluster center, raising the mean cosine.

③ A universal suffix that works on every sample

The gain is shared across all samples, not problem-specific → effectively a universal adversarial suffix. 0.66 → 0.71.

Key Not recovery skill but exploiting a geometric blind spot of the ST5 metric

꼬맨틀 - 단어 유사도 추측 게임

1523번째 꼬맨틀의 정답 단어를 맞혀보세요.
정답 단어와 가장 유사한 단어의 유사도는 **60.24** 입니다.
10번째로 유사한 단어의 유사도는 43.11이고, 1,000번째로 유사한 단어의 유사도는 28.37 입니다.

추측할 단어를 입력하세요 | 추측하기

#	추측한 단어	유사도	유사도 순위	
16	하	20.07	1000위 이상	
6	라라	32.31	????	
4	나나	32.24	????	
8	마마	26.78	1000위 이상	
13	카	26.00	1000위 이상	
1	가	19.49	1000위 이상	

HAiV

Thank You / Q & A

Team 6 Sungbeom Cho · Changhee Lee · Hyotaek Jeon

