



# **BILLY 🧑🏻 : Steering Large Language Models via Merging Persona Vectors for Creative Generation**

Tsung-Min Pai<sup>1</sup>, Jui-I Wang<sup>1</sup>, Li-Chun Lu<sup>1</sup>,  
Shao-Hua Sun<sup>1</sup>, Hung-Yi Lee<sup>1</sup>, Kai-Wei Chang<sup>2</sup>

# Motivation

# Creativity

## Creative Domains

### Art

story, poetry, image,  
video, music

### Linguistics

humor, figurative  
language, lexical  
innovation

### Science

equations,  
conjectures,  
theorems, ideas,  
hypotheses

### Problem Solving

tests, problems,  
puzzles

## One-sentence story

“For Sale: Baby shoes, never worn.”

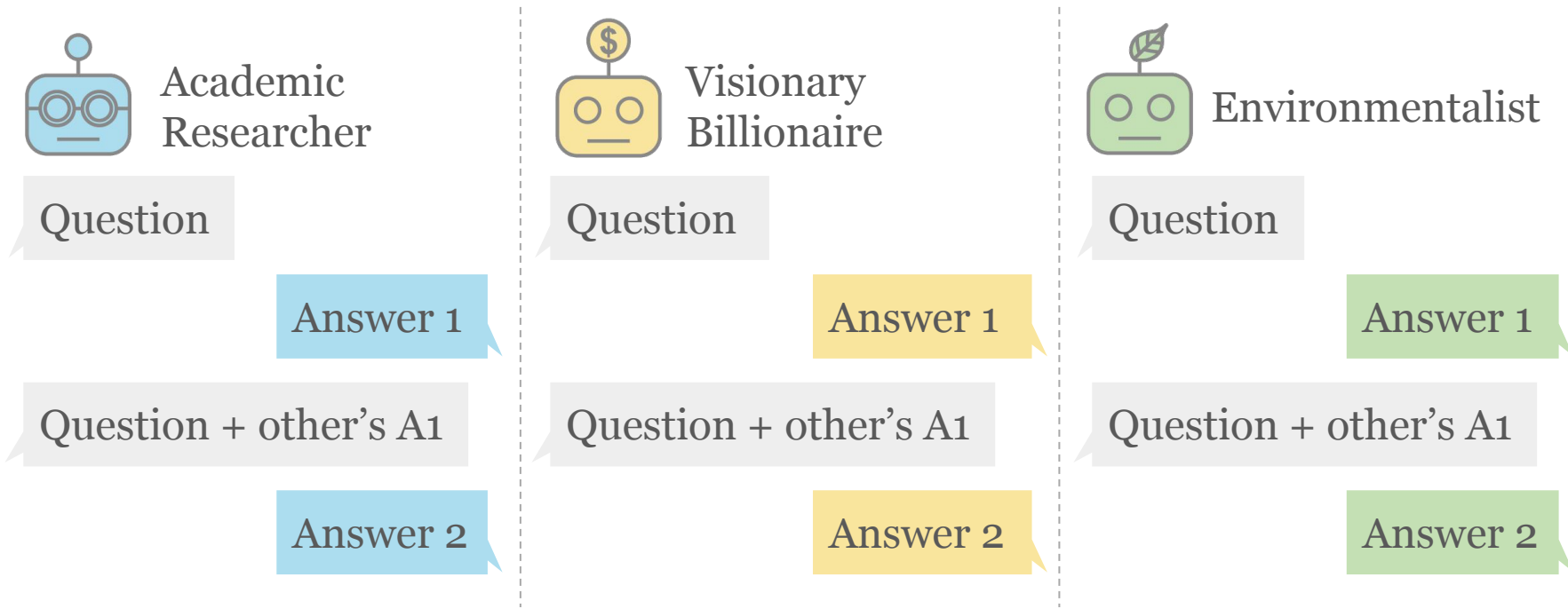
– Hemingway

“Every midnight, the moon quietly borrowed one memory from each sleeping creature on Earth and wove them into a silver staircase, until one child climbed it, reached the sky, and asked for everyone’s memories back—only to discover the staircase itself was what the world had forgotten first.”

– ChatGPT (5.4)

**How can we enhance LLM's creativity?**

# Prior work - LLM Discussion

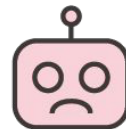


→ **increase 20% creativity on 4 benchmarks using gpt3.5**

# Pros and Cons



- ✓ Diversity
- ✓ Perspective-specific idea
- ✓ Interpretability

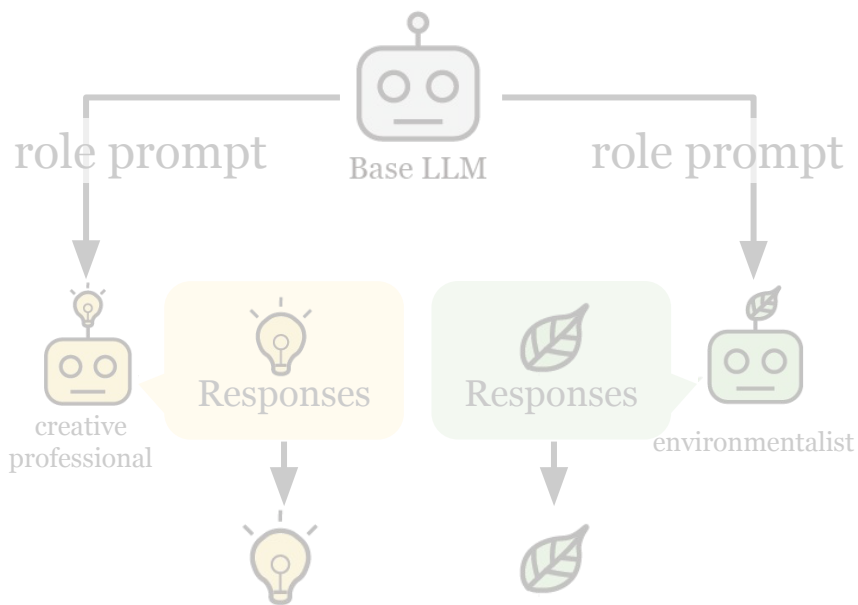


- ✗ Long inference delay
- ✗ High computational costs
- ✗ Complex discussion structure

# Approach

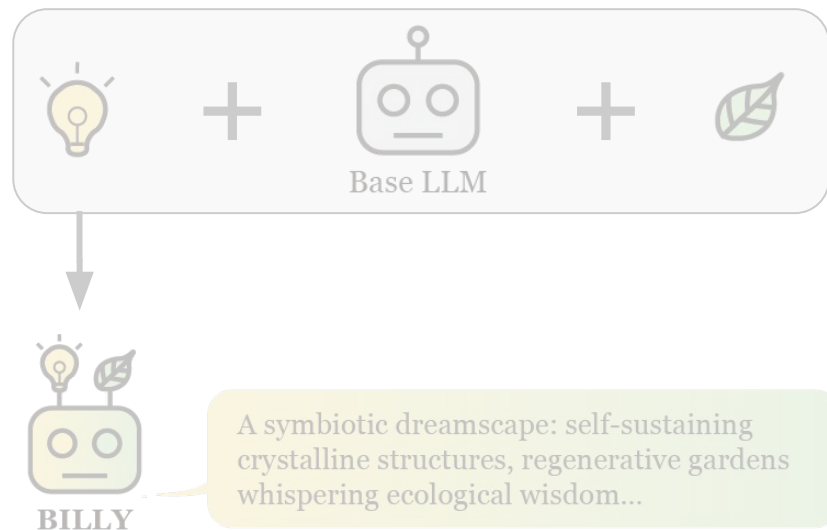
# BILLY: Blending persona vectors for Large Language model creativity

## 1. collect trait-specific responses



## 2. extract persona vectors

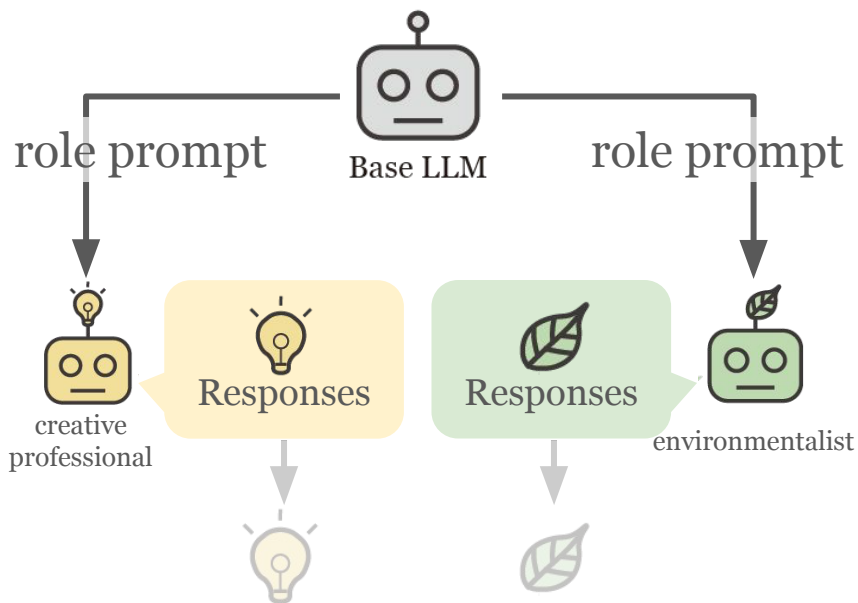
## 3. combine vectors



## 4. inference with steering vector

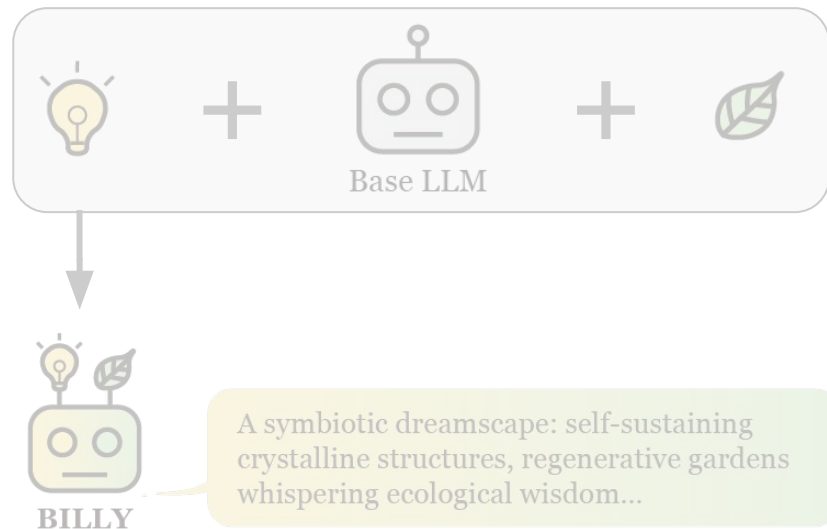
# BILLY: Blending persona vectors for Large Language model creativity

## 1. collect trait-specific responses



## 2. extract persona vectors

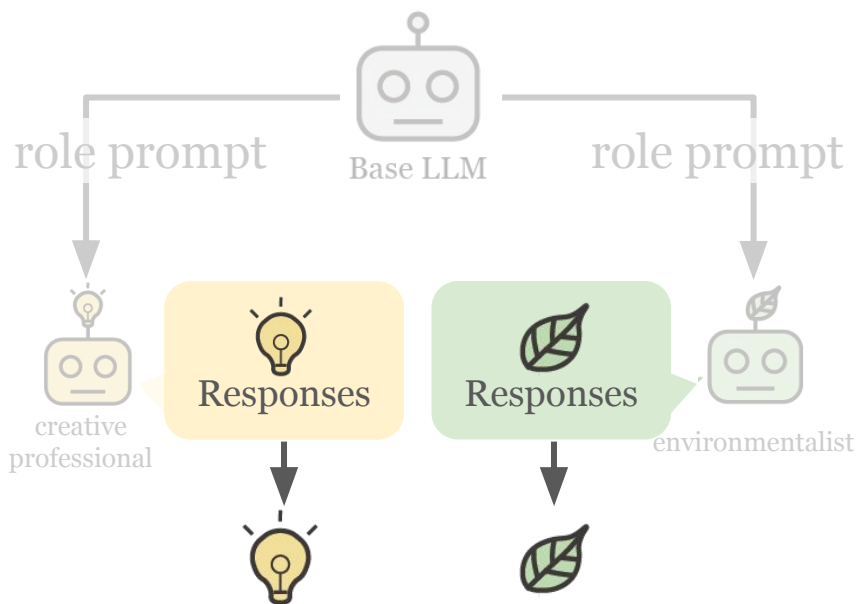
## 3. combine vectors



## 4. inference with steering vector

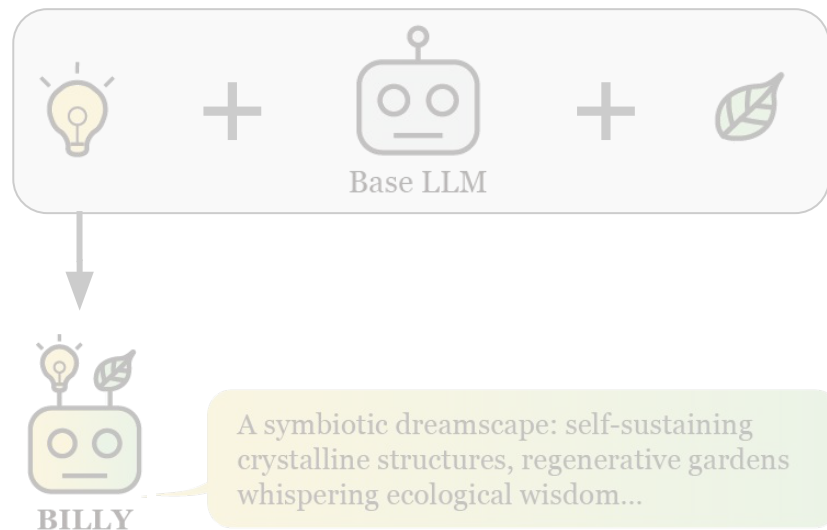
# BILLY: Blending persona vectors for Large Language model creativity

## 1. collect trait-specific responses



## 2. extract persona vectors

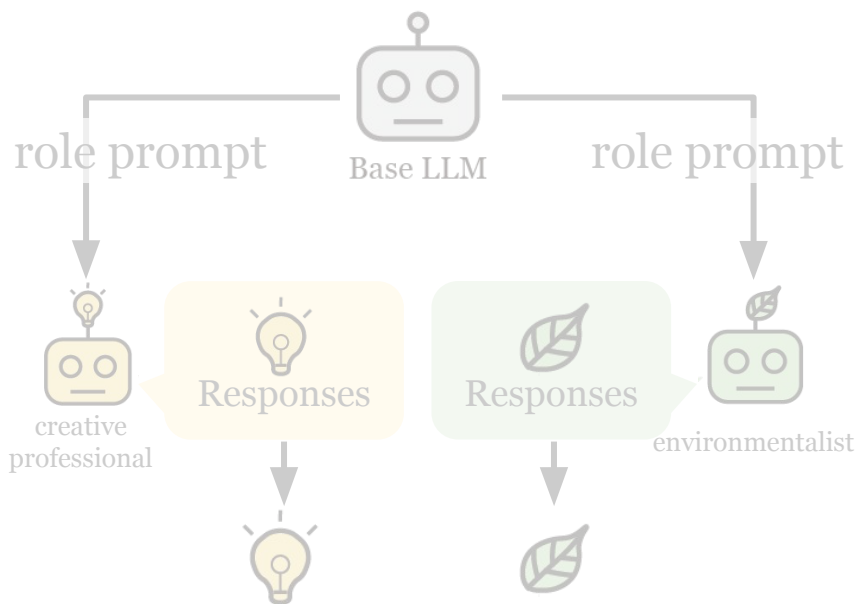
## 3. combine vectors



## 4. inference with steering vector

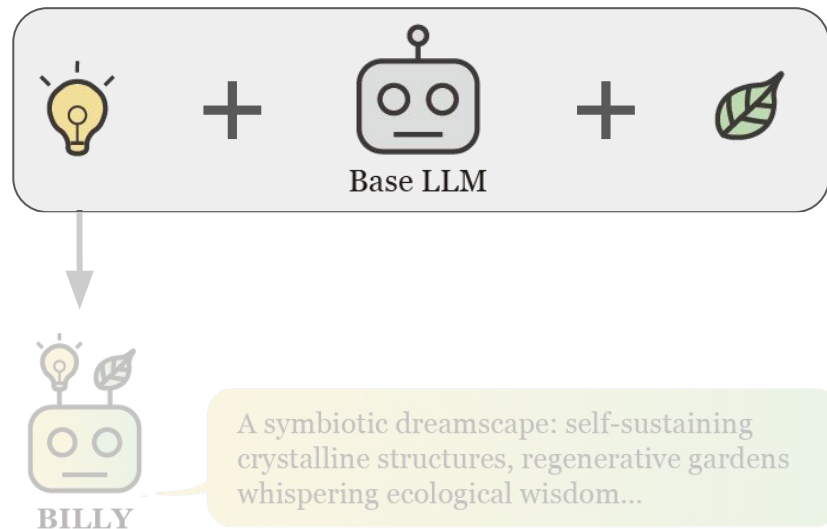
# BILLY: Blending persona vectors for Large Language model creativity

## 1. collect trait-specific responses



## 2. extract persona vectors

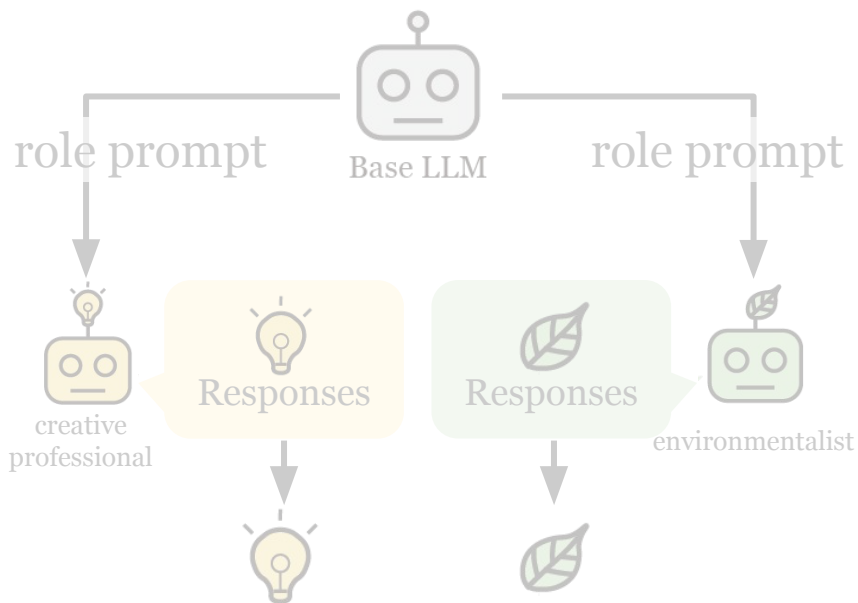
## 3. combine vectors



## 4. inference with steering vector

# BILLY: Blending persona vectors for Large Language model creativity

## 1. collect trait-specific responses



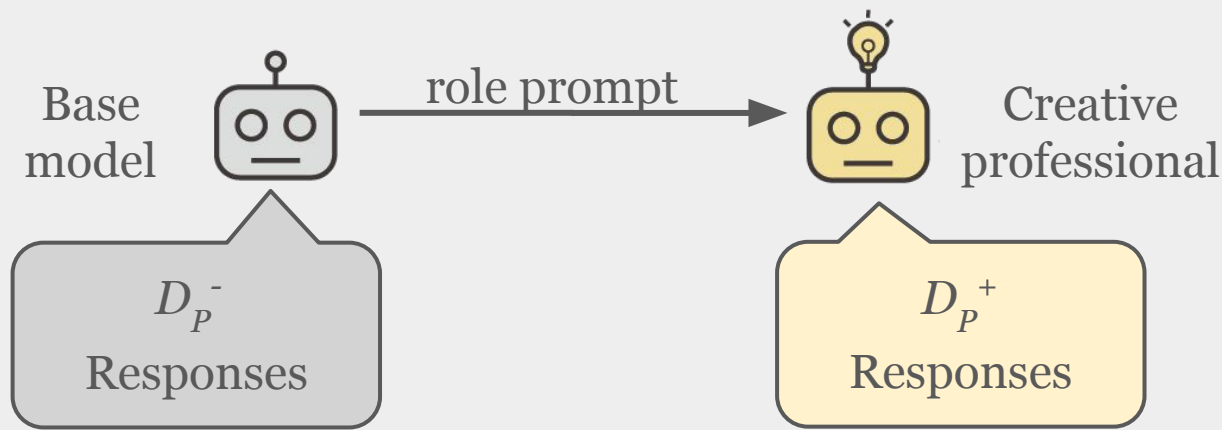
## 2. extract persona vectors

## 3. combine vectors



## 4. inference with steering vector

# Phase 1: Trait-specific Responses Collection



$D_P^+$  : The set contains responses that **exhibit** the persona's traits

$D_P^-$  : The set contains baseline responses that **lack** these specific traits.

## Phase 2: Persona Vector Extraction

Captures the shift in activation when the model adopts a persona vs. a neutral state for response.

$$\vec{v}_P^{(l)} = \frac{1}{|D_P^+|} \sum_{\mathbf{x} \in D_P^+} \vec{a}^{(l)}(\mathbf{x}) - \frac{1}{|D_P^-|} \sum_{\mathbf{x} \in D_P^-} \vec{a}^{(l)}(\mathbf{x})$$

$\vec{v}_P^{(l)}$  : The persona vector for persona P at layer l.

$D_P^+$  : The set contains responses that **exhibit** the persona's traits

$D_P^-$  : The set contains baseline responses that **lack** these specific traits.

$\vec{a}^{(l)}(\mathbf{x})$  For any given response x, we extract its **token-averaged** residual stream **activation** at a specific layer l.

## Phase 3: Offline Fusion of Persona Vectors

Combine multiple specialized vectors into a reusable composite steering direction.

$$\vec{v}_{\text{merged}}^{(l)} = \frac{1}{N} \sum_{i=1}^N \vec{v}_i^{(l)}$$



Represents a **multi-faceted perspective** in the model's activation space.

# one-time calculation → **reusable**

## Phase 4: Inference-time Steering

Intervene at a chosen layer by adding the composite vector scaled by  $\alpha$ .

$$\vec{a}_{\text{steered}}^{(l)} = \vec{a}_{\text{original}}^{(l)} + \alpha \cdot \vec{v}_{\text{merged}}^{(l)}$$

Forward pass (at layer l):

original activations + composite vector, scaled by a coefficient  $\alpha$

# Experiments

# Benchmarks

Benchmark	Description	Sample Task
<b>AUT</b>	Generating unconventional applications for an object.	What are some creative uses for a mug?
<b>Instances</b>	Producing a diverse set of examples that satisfy a given property.	Name 5 square things you can think of.
<b>Similarities</b>	Challenging participants to identify non-obvious connections between two concepts.	Tell me 5 ways in which a brick and a stone are alike.
<b>Scientific</b>	Probing creative problem-solving within a scientific framework.	Find different scientific uses for a spoon.

# Evaluation

Metrics from Torrance Test (TTCT):

- **Originality:** The level of novelty and divergence while relevant.
- **Elaboration:** The level of details and supportive information.

Evaluators:

- Human evaluators
- LLM-Based Judge (GPT-4o-mini)

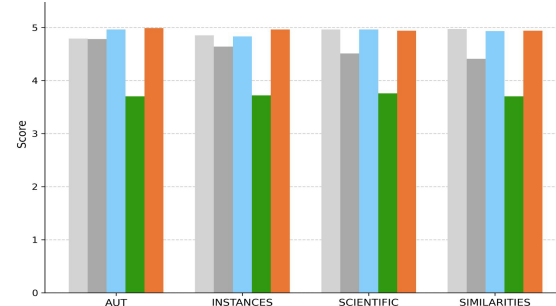
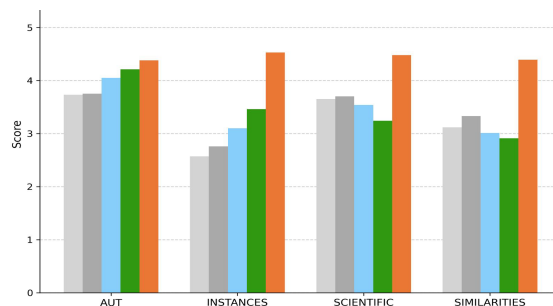
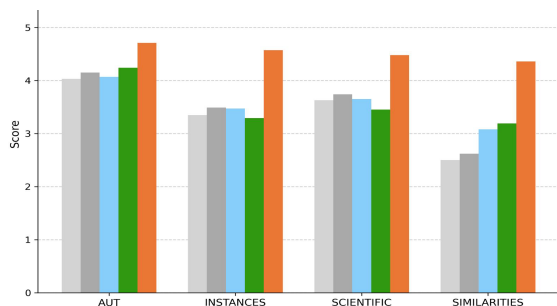
# Baselines

Baselines	Description
<b>Single Agent (SA)</b>	A single LLM prompted to respond creatively (temperature = 0.7).
<b>SA (SA T=1.0)</b>	SA with higher temperature (1.0) to stimulate higher levels of diversity.
<b>SA with Multi-Role Prompt (SA-MRP)</b>	SA asked to respond from multiple perspectives (temperature = 0.7).
<b>LLM Discussion</b>	A multi-LLM framework: agents exchange ideas over several rounds and then consolidate them into final outputs.

# Result: GPT-based Evaluation



Originality

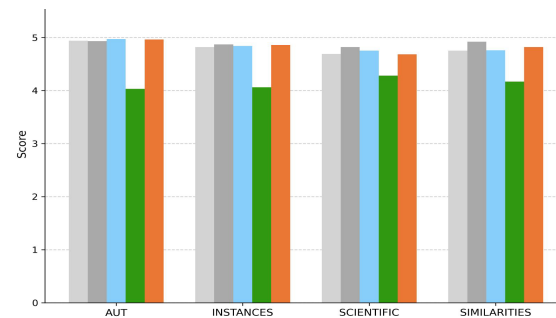
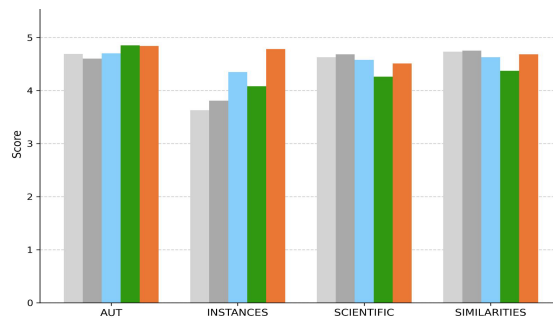
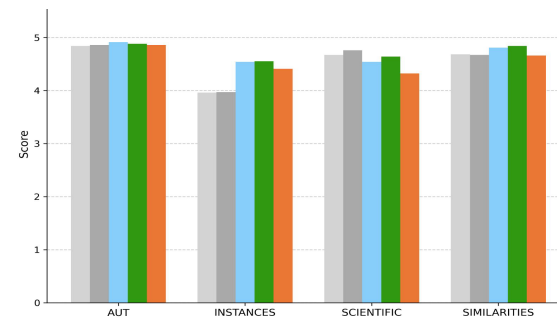


Qwen-2.5-7B-Instruct

Llama-3.1-8B-Instruct

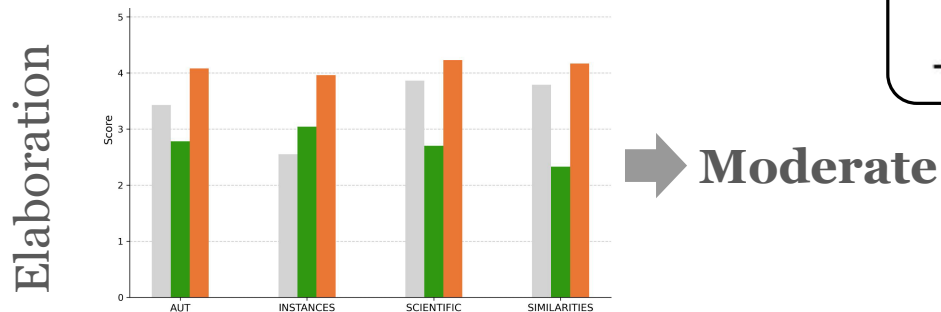
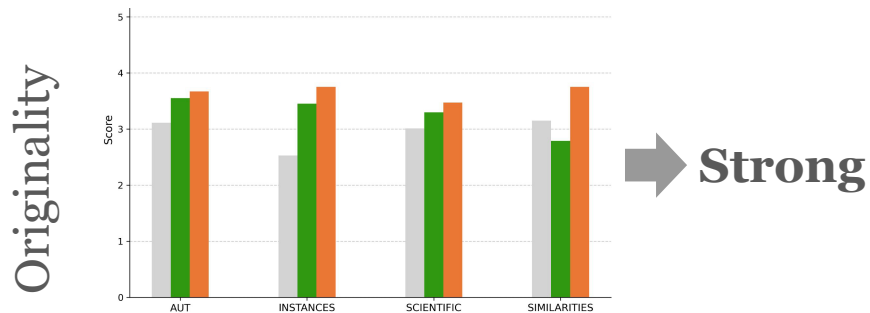
Gemma-3-4B-it

Elaboration



# Result: Human-based Evaluation & Correlation

SA LLM-Discussion BILLY (Ours)

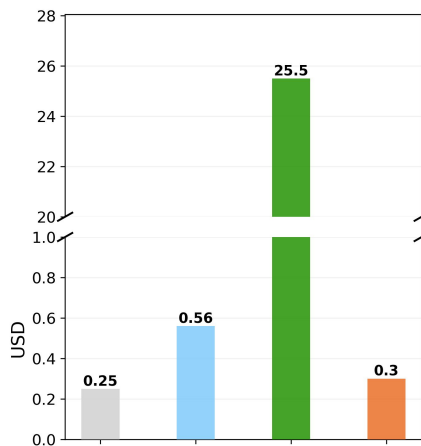


Correlation	Originality	Elaboration
Spearman	0.73	0.43
Pearson	0.66	0.40

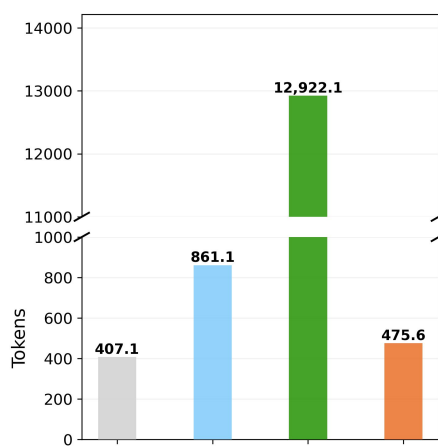
# Analyses: Cost & Latency



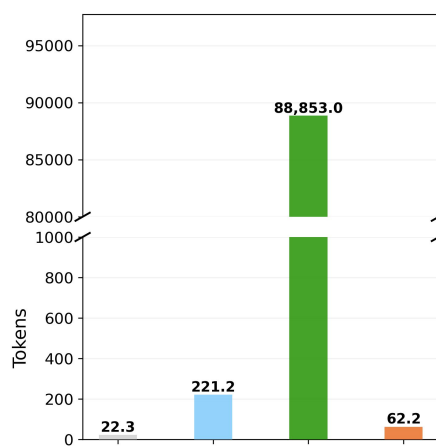
Cost (\$)/ 10k queries



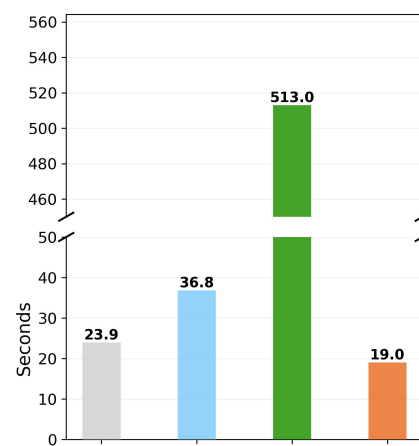
Token (out)/query



Token (in)/query



Latency (seconds)



# LLM Discussion

**42,200 tokens vs. 488 tokens**

**86 Times of Costs**

**BILLY**

# Analyses: Qualitative Results

**Question:** Healthcare Innovation - Redesign the future hospital experience for patients, families, and staff.

## **BILLY (ENV)**

... **Community Gardens:** Outdoor spaces promote physical activity, connect nature, and reduce stress.

## **BILLY (CRE)**

... **Bioluminous Biome:** Radiant forests, crystalline geodes, & iridescent orbs bathing spaces in kaleidoscopic hues.

## **BILLY (CRE + ENV)**

... **Fractal Forests:** Ephemeral gardens grow within glass domes, bathing visitors in iridescent mosaics, awakening inner blooms.

## **MRP (CRE + ENV)**

... **Feasibility and Sustainability:** Balance operational costs with environmental sustainability, energy efficiency, and long-term viability.

Individual vectors exhibit distinct functions

Ours: demonstrate conceptual fusion

Baseline: using relatively analytical words

# BILLY : Steering Large Language Models via Merging Persona Vectors for Creative Generation

## Takeaways

- **Efficient** activation steering method → reduce inference latency and costs
- **High Controllability** towards model's internal states and outputs
- Distinct persona vectors correspond to distinct aspects of generation
- Effectively **enhance LLM Creativity**

