CogAI4Sci Lab

Enriching Diversity and Novelty in Text Generation via Multi-view Embeddings

Arash Lagzian[1], Srinivas Anumasa[1], Dianbo Liu[1]

[1] National University of Singapore

alagzian@visitor.nus.edu.sg
{srinu_pd, dianbo}@nus.edu.sg

Table of Content

- Problems and Limitations
- Our Proposed Framework
- Text Multi-view Embedding
- Image Multi-view Embedding
- Diversity Measurement
- Novelty Measurement
- Correctness Measurement
- Experimental Results
- Conclusion

Problems and Limitations

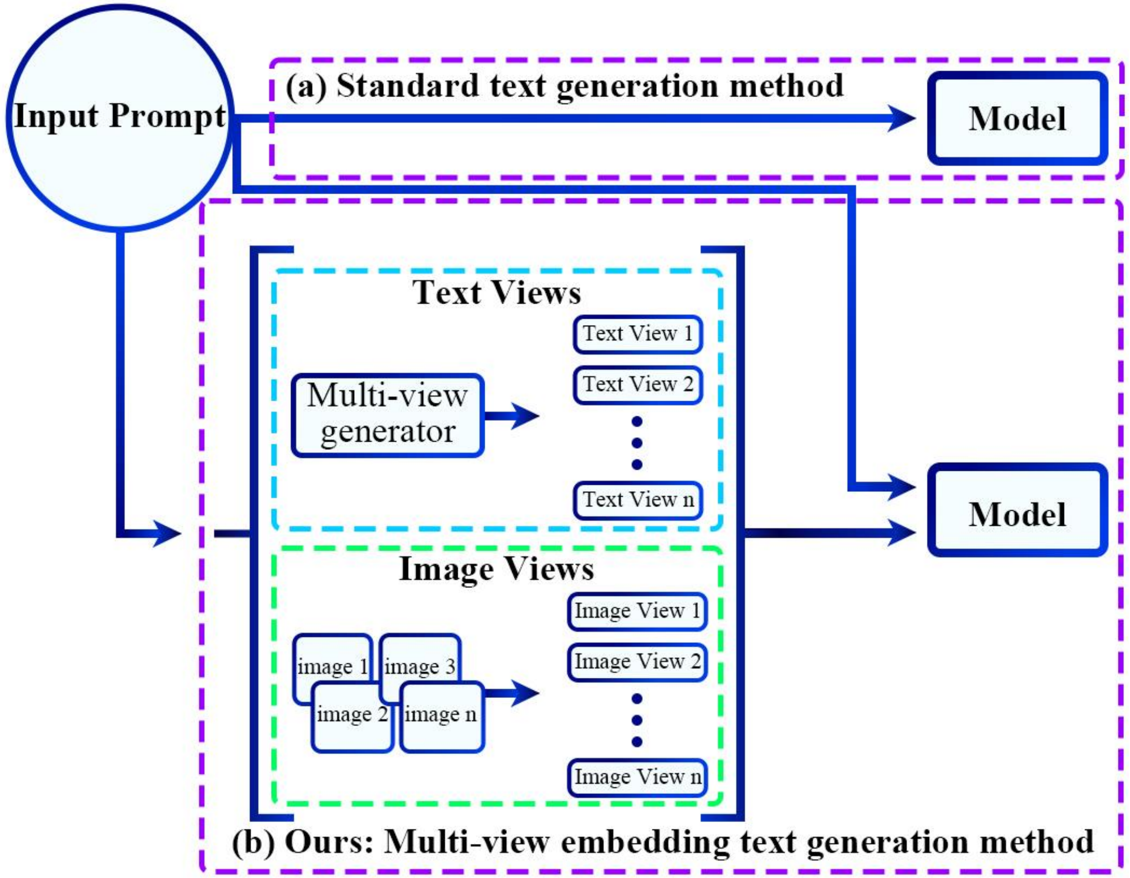
Text Generation:

- Constraints in training data
- Gaps in specific knowledge domains
- Outdated information
- Over-reliance on textual resources

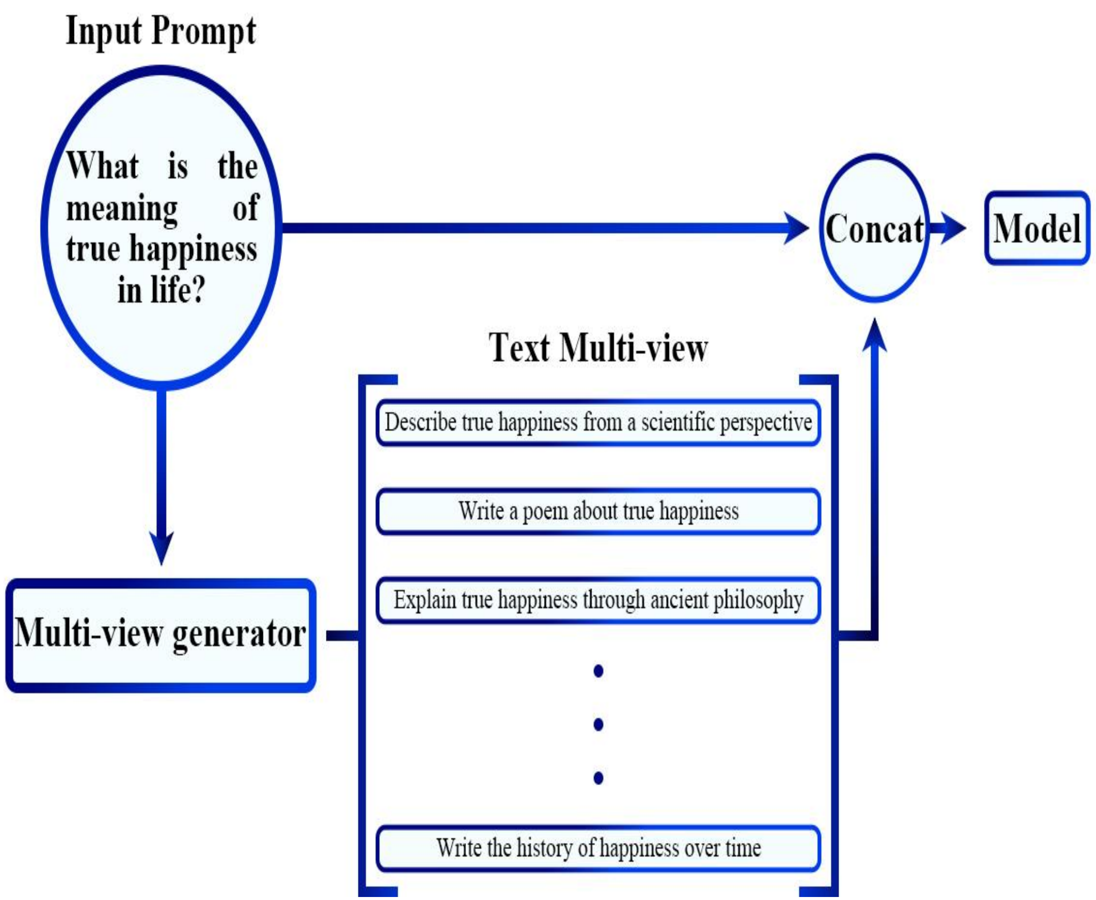
Generated Text Evaluation:

- Diversity, or Novelty evaluation individually

Our Proposed Framework



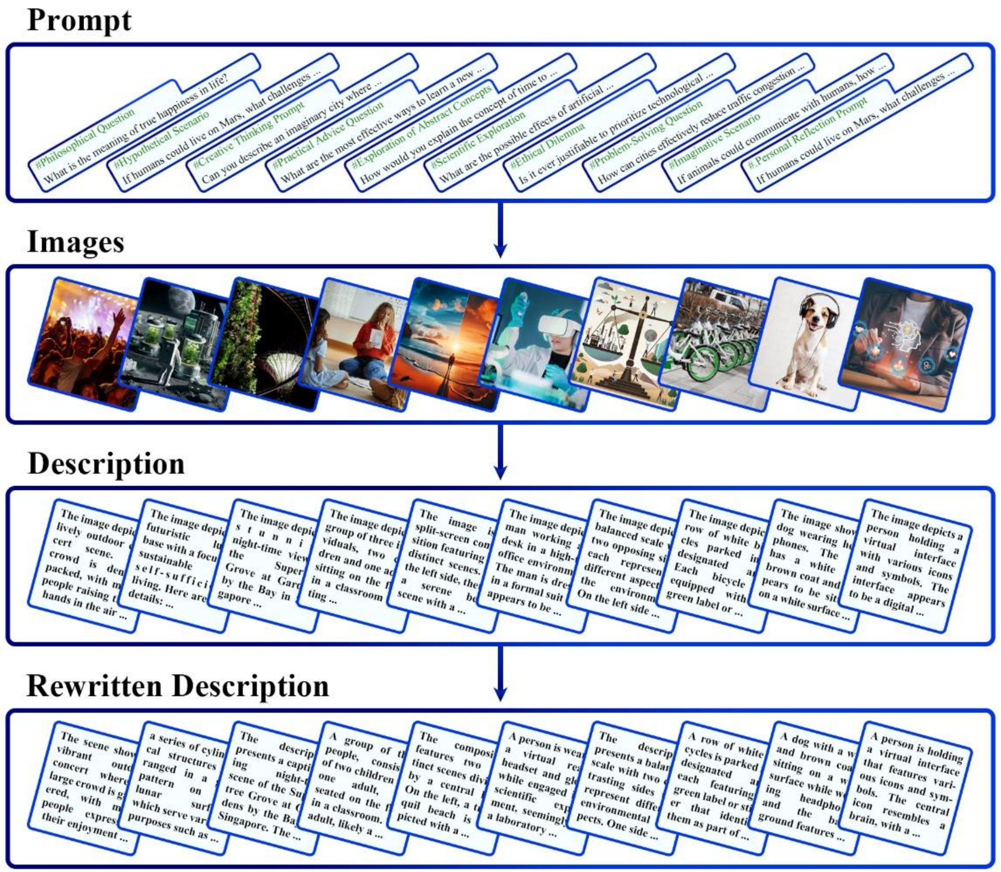
Our Proposed Framework – Text Multi-view Embedding



Prompt: What is the true meaning of true happiness in life? **View:** Describe true happiness from a scientific view **Answer:** The term \"true happiness\" refers to a state of contentment and satisfaction that can come from various sources such as achieving goals, experiencing positive emotions, or having meaningful relationships with others. It encompasses both external factors like financial stability and internal states like self-esteem.\n\nIn terms of neuroscience, happiness is thought to be associated with the release of dopamine in the brain's reward system. This neurotransmitter plays a role in motivation and pleasure-seeking behavior, making it easier for individuals to feel happy when they achieve their goals or have positive experiences. Additionally, studies have shown that meditation and other forms of mindfulness practices can increase activity in the pre"

Max_length = 125, model: Qwen2.5, 1.5 Instruct

Our Proposed Framework – Image Multi-view Embedding



Prompt: *What is the meaning of true happiness in life?*
Image View: *The scene shows a vibrant outdoor concert where a large crowd is gathered, with many people expressing their enjoyment by raising their hands in the air. Bright lights illuminate the area, particularly focusing on the stage in the background. The audience is lively and energetic, with some individuals recording the experience on their phones. The atmosphere is filled with excitement and celebration, characteristic of a live music event.*

Answer: *True happiness in life often encompasses experiences that bring joy, connection, and personal growth through various activities such as participating in music concerts like the one described. Engaging in shared experiences with others creates bonds and enriches our social lives. Music, being an art form rooted deeply in human culture, has long been associated with positive emotions; it can evoke strong feelings of joy, community, and self-expression. In such a vibrant outdoor concert setting, the collective emotion displayed by the attendees—raising their hands in the air, indicating enthusiasm, and enthusiastically recording the moment with smartphones—are vivid expressions of true happiness and engagement.*

Figure 4: An example of Image view related to the input prompt and answered by Qwen2.5-1.5B model.

Diversity Measurement

The diversity of the text generated by LLMs can be measured from different aspects. Two of the most important aspects are:

- Diversity across different text responses
- Diversity within different tokens generated in a single text response

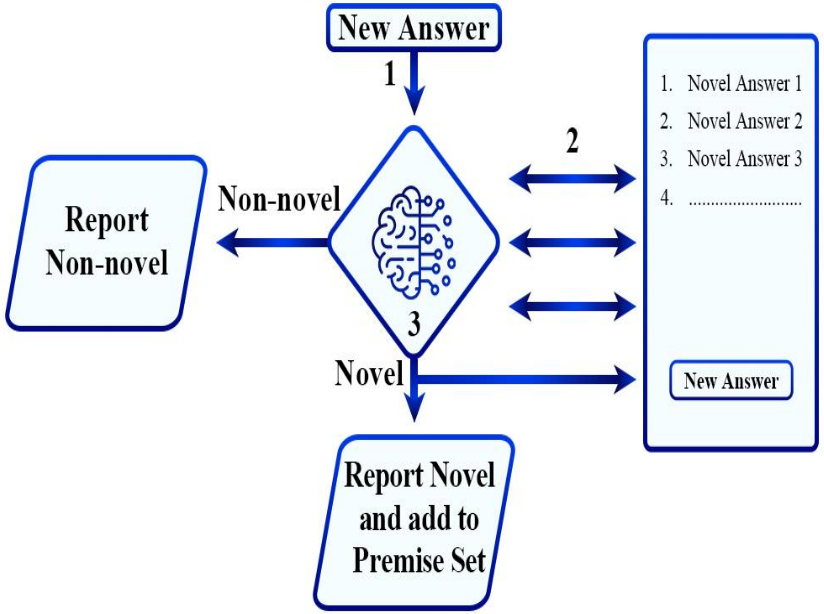
Metrics:
1. Measure of Textual Lexical Diversity (MTLD)
2. Semantic Diversity of Text (SDT)
3. Semantic Diversity of Embeddings (SDE)
4. Lexical Entropy
5. Self-BLEU

Novelty Measurement

Table1: Classification metrics to evaluate novelty detectors GPT-4o and SBERT models on dataset TAP-DNLD 1.0 (2.8k novel and 2.7k non-novel documents). (Ghosal et al. 2018).

Seed	Model	Accuracy	Precision	Recall	F1-score
Seed0	SBERT	0.6700	0.6269	0.8400	0.7179
	GPT-4O	**0.6900**	**0.6415**	**0.8600**	**0.7347**
Seed1	SBERT	**0.7200**	**0.6528**	**0.9400**	**0.7705**
	GPT-4O	0.6600	0.6086	0.9000	0.7260
Seed2	SBERT	0.6900	0.6338	0.9000	0.7438
	GPT-4O	**0.6950**	**0.6368**	**0.9100**	**0.7492**

Novelty Detection Framework



Correctness Measurement

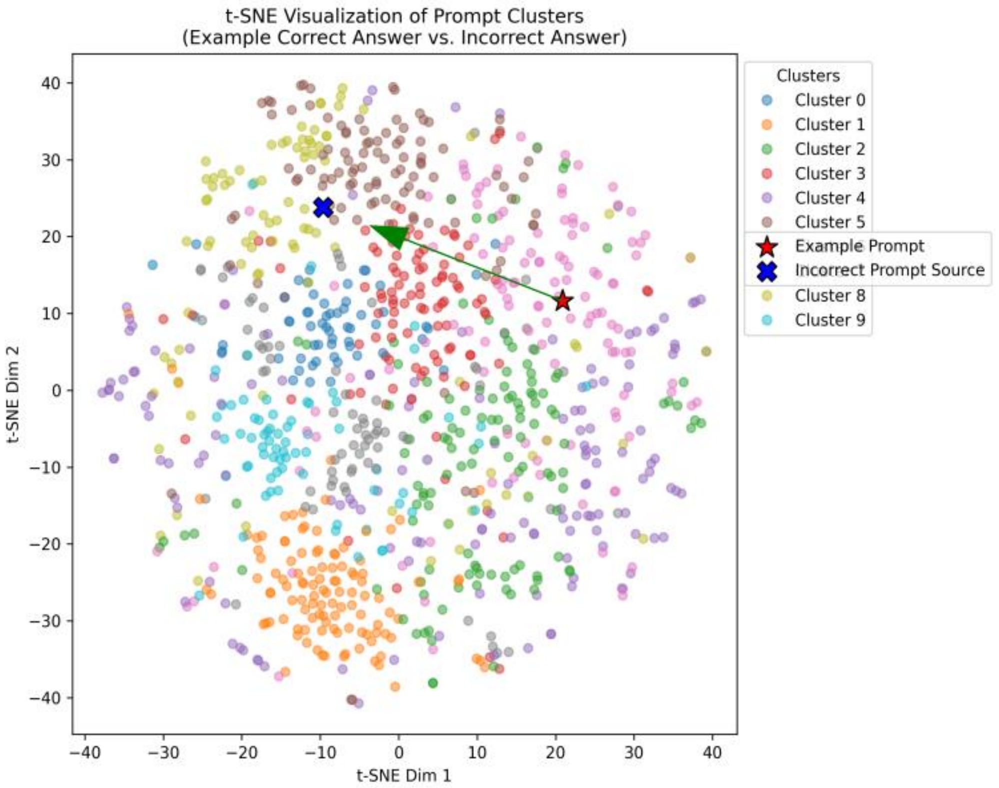
- Q1: Is the generated answer correct and relevant to the given input prompt?
- Q2: How well does the generated answer adhere to proper language structure and grammatical accuracy?

GPT-WritingPrompts dataset. (Huang et al. 2024). ~97k prompt-answers.(we used human answers)

Sample in 10 different seeds (Each seed 1k data).

Cluster them into 10 different clusters

Sample 100 prompts with
one real true answer and
one wrong answer from another cluster



Correctness Measurement

- **Q1: Is the generated answer correct and relevant to the given input prompt?**
- Q2: How well does the generated answer adhere to proper language structure and grammatical accuracy?

Table2: Results on classification task to detect correct answer across all seeds for DeepSeekV3 and GPT-4o models.

Model	Accuracy	Precision	Recall	F1-score
DeepSeekV3	0.8410 ± 0.0223	**0.9564 ± 0.0285**	0.7270 ± 0.0447	0.8191 ± 0.0301
GPT-4O	**0.8980 ± 0.0214**	0.9468 ± 0.0255	**0.8310 ± 0.0316**	**0.8946 ± 0.0228**

Correctness Measurement

- Q1: Is the generated answer correct and relevant to the given input prompt?
- **Q2: How well does the generated answer adhere to proper language structure and grammatical accuracy?**

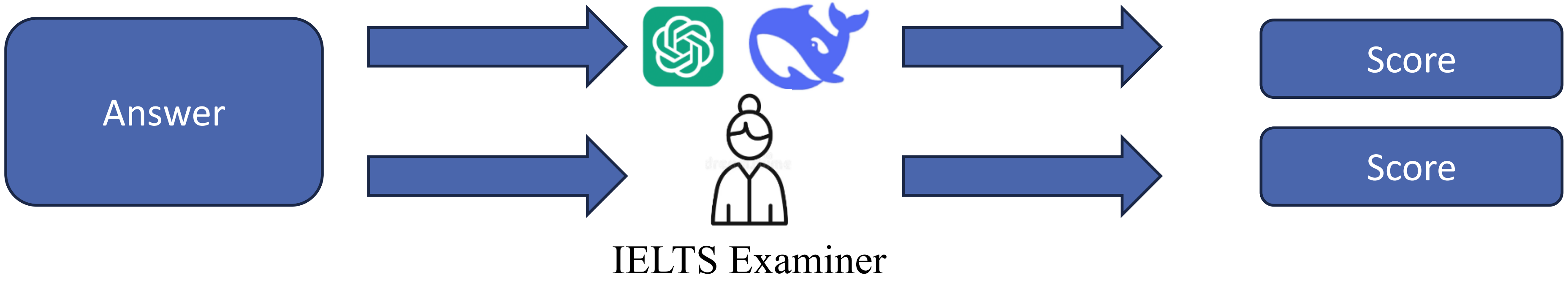


Table 3: MSE (Mean ± Std) for GPT-4O and DeepSeekV3 per seed for two iterations. Lower values are bolded.

Seed	GPT-4O	DeepSeekV3
Seed0	2.3775 ± 0.0350	**2.0775 ± 0.0025**
Seed1	2.2288 ± 0.0512	**2.0300 ± 0.0050**
Seed2	2.1562 ± 0.0587	**1.9050 ± 0.0500**

Experiments and Results - Diversity

Table 4: Mean ± standard deviation of diversity metrics across five different sample sizes ranging from 100 to 1500 per each prompt (10 prompts), with $max\_length = 125$. Detailed results are provided in Appendix A.

Models	MTLD	TF-IDF	Lexical Entropy	Semantic Embedding	Self-BLEU
GPT-2	**41.36 ± 0.05**	0.86 ± 0.01	8.14 ± 0.09	0.55 ± 0.00	0.61 ± 0.10
GPT-2 + Text View	35.27 ± 0.99	**0.89 ± 0.01**	**8.33 ± 0.13**	**0.64 ± 0.01**	**0.67 ± 0.08**
Qwen	67.17 ± 0.15	0.80 ± 0.01	7.90 ± 0.05	0.22 ± 0.00	0.27 ± 0.08
Qwen + Text View	67.63 ± 0.11	**0.86 ± 0.01**	**8.45 ± 0.07**	**0.39 ± 0.00**	**0.41 ± 0.09**
Qwen + Image View	**70.29 ± 0.04**	0.84 ± 0.01	8.28 ± 0.07	0.30 ± 0.00	0.38 ± 0.09
GPT-4O Mini	57.95 ± 0.01	0.63 ± 0.01	7.06 ± 0.02	0.11 ± 0.00	0.10 ± 0.05
GPT-4O Mini + Text View	**59.61 ± 0.12**	**0.80 ± 0.04**	**7.92 ± 0.04**	**0.33 ± 0.01**	**0.26 ± 0.09**
GPT-4O Mini + Image View	58.71 ± 2.03	0.79 ± 0.02	7.80 ± 0.04	0.30 ± 0.01	0.25 ± 0.08
GPT-4O	57.46 ± 0.47	0.66 ± 0.01	7.00 ± 0.16	0.11 ± 0.02	0.09 ± 0.04
GPT-4O + Text View	**58.25 ± 0.04**	0.78 ± 0.01	**7.85 ± 0.06**	**0.31 ± 0.01**	0.25 ± 0.08
GPT-4O + Image View	55.09 ± 0.31	**0.79 ± 0.01**	7.77 ± 0.04	**0.31 ± 0.00**	**0.29 ± 0.09**
DeepSeek-R1	52.49 ± 0.11	0.79 ± 0.01	7.63 ± 0.06	0.24 ± 0.00	0.26 ± 0.08
DeepSeek-R1 + Text View	**54.62 ± 0.20**	**0.85 ± 0.01**	**8.19 ± 0.10**	**0.39 ± 0.01**	**0.39 ± 0.09**
DeepSeek-R1 + Image View	54.35 ± 0.05	0.83 ± 0.01	8.06 ± 0.08	**0.39 ± 0.00**	0.36 ± 0.09

Experiments and Results - Novelty

Table 5: Results on percentage of novelty score across different models according to two novelty detectors GPT-4o and SBERT. More results for another models exist in Table 13.

Model	num_samples = 100		num_samples = 250	num_samples = 500
	GPT-4o	SBERT	SBERT	SBERT
GPT-4o	10.60	5.4	3.52	2.9
GPT-4o + Text View	29.30	40.2	**32.08**	**24.48**
GPT-4o + Image View	**42.60**	**47.3**	30.44	22.54
DeepSeek-R1	27.40	43.6	36.24	31.82
DeepSeek-R1 + Text View	**46.20**	**75.3**	63.88	55.86
DeepSeek-R1 + Image View	40.60	73.7	**65.04**	**56.18**

Experiments and Results - Correctness

Table 6: (a) and (b) show correctness results for different evaluation aspects for $num\_sample = 100$ per prompt (10 prompts).

(a) correctness results (%)

Model	GPT-4o
GPT-4o mini	99.81
GPT-4o mini + Text View	91.00
GPT-4o mini + Image View	87.00
Qwen	93.77
Qwen + Text View	76.60
Qwen + Image View	82.50
GPT-4o	99.60
GPT-4o + Text View	92.60
GPT-4o + Image View	94.60
DeepSeek-R1	91.80
DeepSeek-R1 + Text View	81.00
DeepSeek-R1 + Image View	53.90

(b) Mean score (1 to 10) for correctness from English structure aspects.

Model	DeepSeekV3
GPT-2	3.29
GPT-2 + Text View	2.40
GPT-4o	8.07
GPT-4o + Text View	8.05
GPT-4o + Image View	8.06
DeepSeek-R1	7.96
DeepSeek-R1 + Text View	7.15
DeepSeek-R1 + Image View	6.10

Conclusion

- We introduced multi-view embedding, a model agnostic approach that enrich the input prompt with diverse textual and visual sources to enhance the diversity and novelty of generated responses.
- We propose a new framework to measure the diversity, novelty, and correctness if the generated outputs from LLMs.
- We evaluate both our method and framework using real-world datasets and 469k generated answers from various LLMs.

Thanks for your attention

MAY 2025