

# PANGEA: Projection-Based Augmentation with Non-Relevant General Data for Enhanced Domain Adaptation in LLMs

**Seungyoo Lee**

**SIML**

# Outline

---

[1] Synthetic Data Generation?

[2] Diversity

[3] PANGEA: Proposed Method

[4] Experiments Results

# [1] Synthetic Data Generation?

# Synthetic Datageneration

---

**Domain-Specific Adaptation:** Addressing the gap in LLM performance for specialized fields (e.g., Medical, Finance)

**The Data Scarcity Bottleneck:** Challenges in obtaining high-quality domain data due to privacy or cost

**Synthetic Data Generation:** A promising solution to bootstrap instruction-following capabilities

# Existing Methods (1)

**Retrieval-Based Approaches (e.g., DataTune):** retrieving and transforming related datasets.

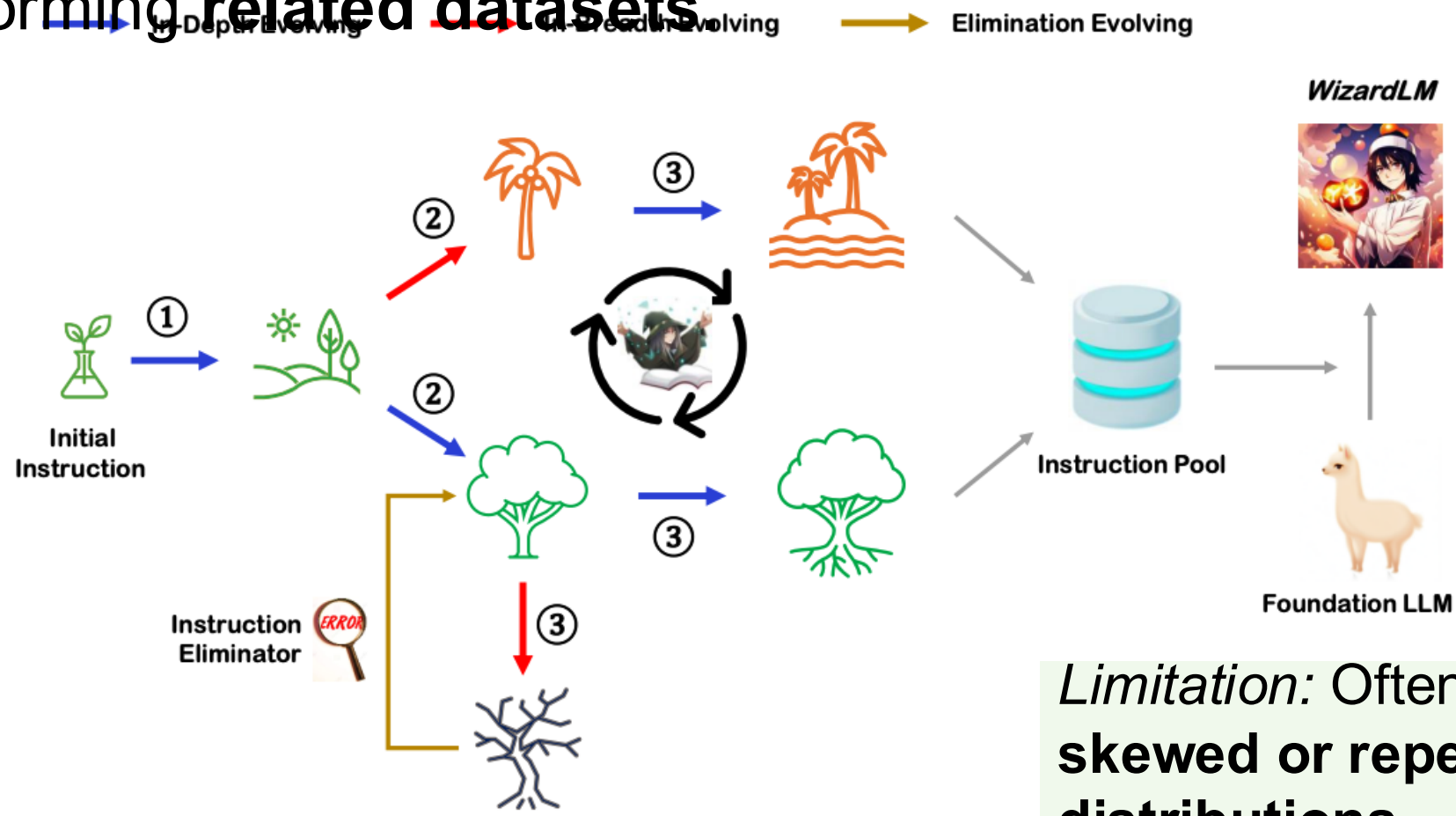


Figure 2: Overview of *Evol-Instruct*

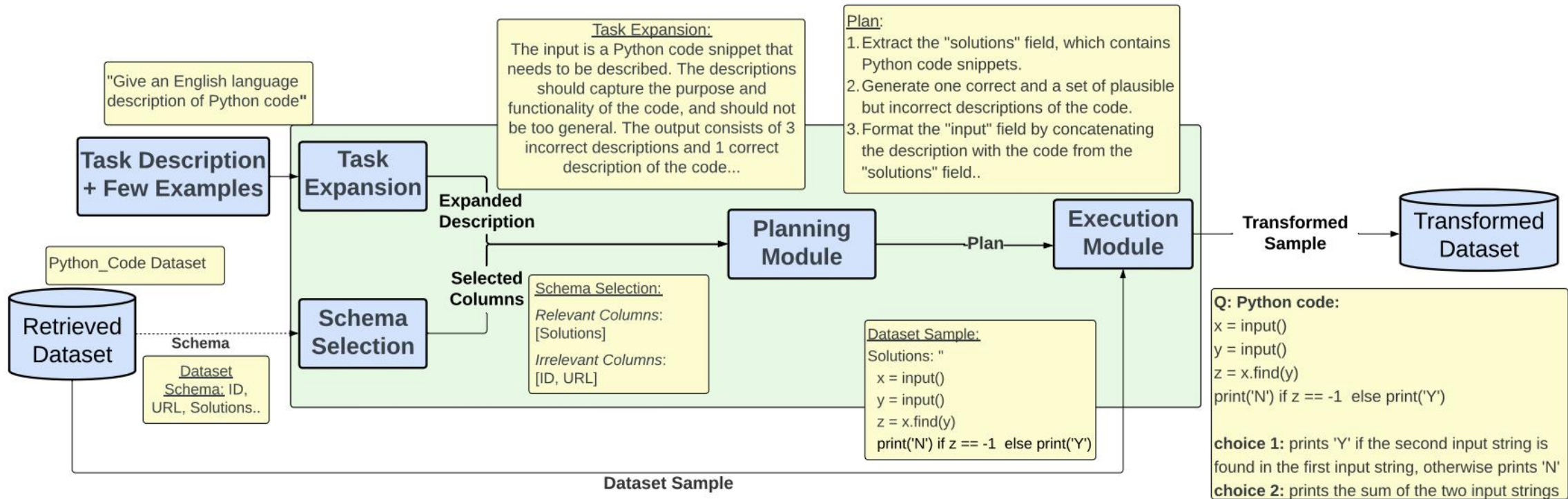
**Limitation:** Often results in **skewed or repetitive distributions**

# Existing Methods (2)

**Retrieval-Based Approaches (e.g., DataTune):** retrieving and transforming related datasets.

**Limitation:** Requires domain-relevant sources, failing when such data is unavailable.

## Dataset Transformation Component



# Failure Case: DataTune

A 1-year-old girl is brought to the physician for a well-child examination. She has no history of serious illness. She receives a vaccine in which a polysaccharide is conjugated to a carrier protein. Which of the following pathogens is the most likely target of this vaccine?

Choose one of the following:

- A. Hepatitis A virus
- B. Varicella zoster virus
- C. Streptococcus pneumoniae
- D. Bordetella pertussis



## Med QA

Two space explorers are flying around through an unknown galaxy when they discover two new planets. Given that these planets have no names, the explorers decide to name them after themselves. The planet of Sarahn is currently going through a large geographic shift. Many of the continents on the planet are currently breaking apart and moving. Timon, however, has been a stable planet with no continental moving for quite awhile now. Given the paragraph above, please answer correctly the following question:

Is the volcanic crust on Sarahn or Timon hotter?

## Unrelated Existing Data

### Transformed Sample

A patient named Sarahn presents with unstable symptoms characterized by frequent 'shifting' of clinical manifestations, resembling episodes of transient ischemic attacks. Another patient, Timon, presents with stable and unchanging symptoms similar to chronic stable angina. **Based on these clinical presentations, which patient likely has more severe underlying inflammation?**

- A. Sarahn has more severe inflammation.
- B. Timon has more severe inflammation.
- C. Both have equal inflammation.
- D. Neither has significant inflammation.

## [2] Diversity

# The Diversity Spectrum: Latent vs Semantic

---

Latent Diversity (Reference: Prismatic Synthesis):

**Gradient-Based Diversification:** Using model gradients to identify diverse samples.

**Filtering & Sampling:** Essentially acting as a **filtering mechanism** to select the best samples from the model's latent space

**Constraint:** Optimizing within the *existing distribution capabilities*.

# Diversity: Latent Space

## Quantifying Diversity via Spectral Entropy

$$VS(\mathcal{D}) = \exp \left( - \sum_{i=1}^N \bar{\lambda}_i \log \bar{\lambda}_i \right)$$

**G-Vendi Score**  
(Effective Number of Samples)

$$g(x) = \nabla_{\theta} \mathcal{L}(x; \theta)$$

**Learning Signal**  
Representation

$$K_{ij} = \frac{g(x_i) \cdot g(x_j)}{\|g(x_i)\| \|g(x_j)\|}$$

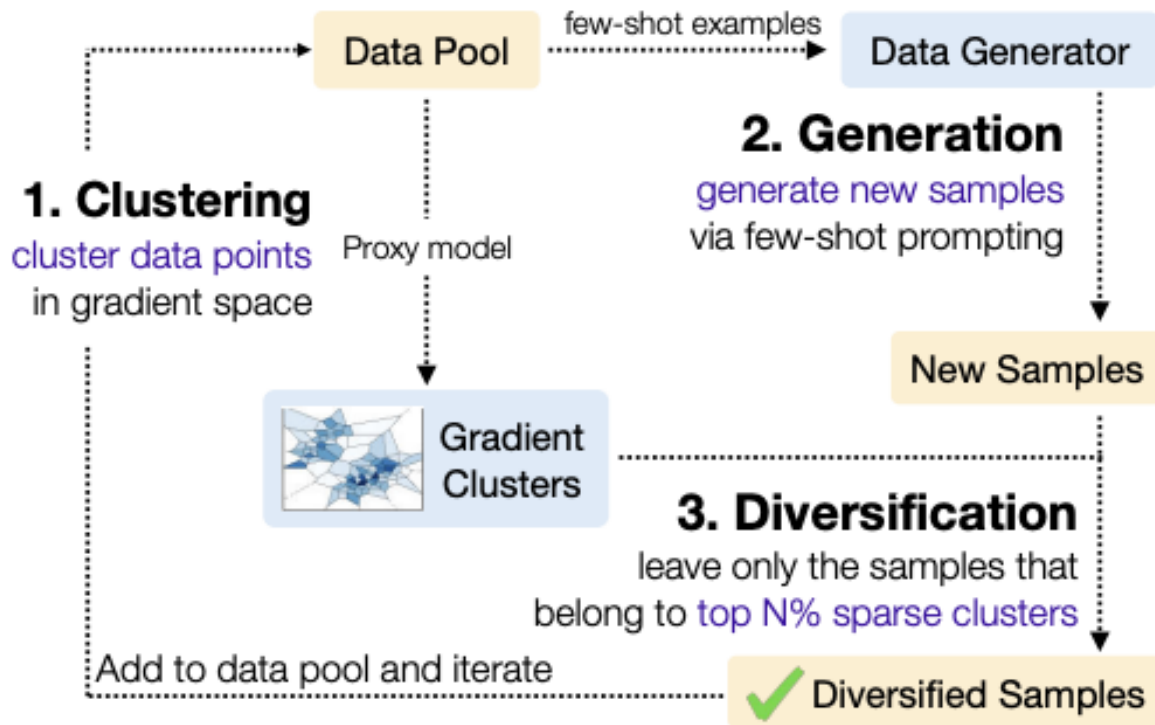
**Kernel Matrix**  
(Cosine Similarity in Gradient  
Space)

## Optimization Strategy

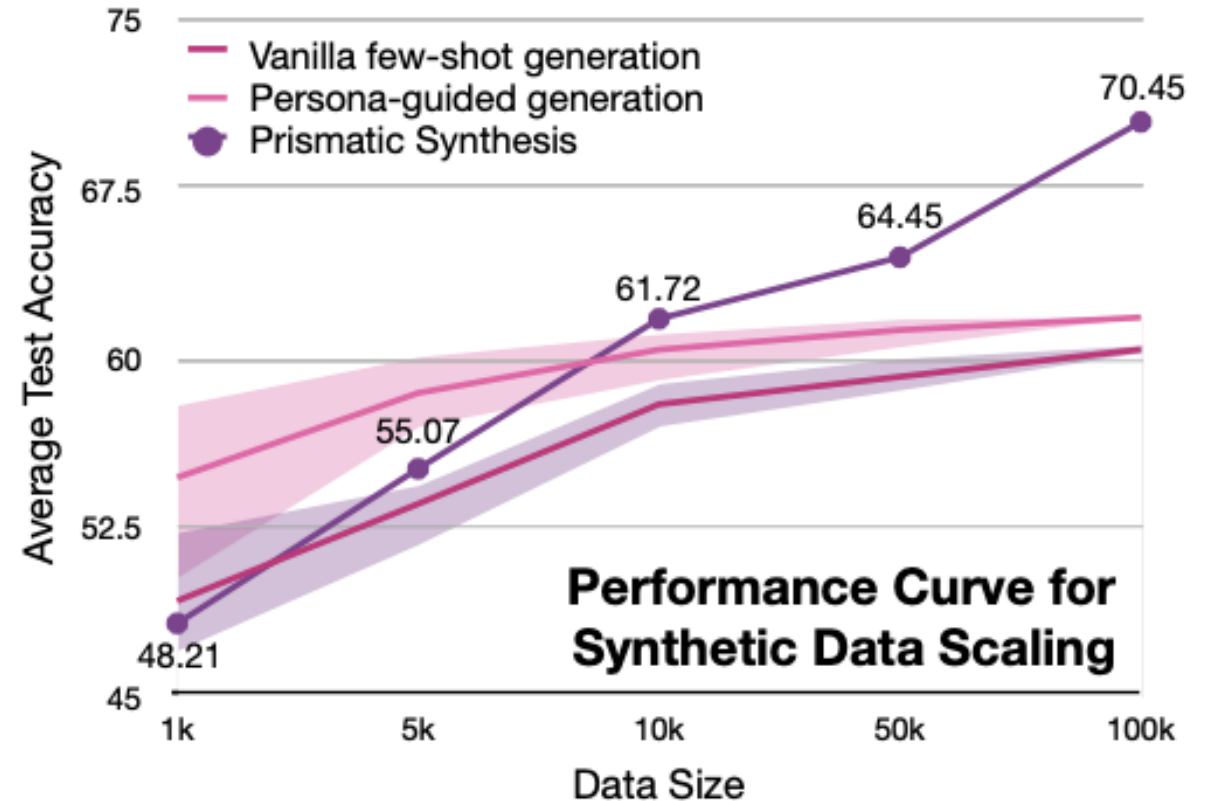
Maximize  $VS(\mathcal{D}) \iff$  Minimize  $K_{ij}$  (Orthogonal Gradients)

**Filtering for Sparse Regions in Latent Space**

# Diversity: Latent Space



 **Prismatic Synthesis**



**Constraint:** Optimizing within the *existing distribution*

*capabilities.*

# From Latent Selection to Semantic Injection

---

## 1. The Limitation: Latent Diversity (internal)

- **Mechanism:** Maximizing Gradient Orthogonality (e.g., Prismatic Synthesis)
- **Role:** Efficient **Filtering & Selection strategy.**
- **Constraint:** Bound by the Existing Distribution.
  - Effective only when **the candidate pool is large**

## 2. The Critical Gap: Extreme Data Scarcity

Specialized Domains with  $|\mathcal{D}| \leq 100$ .

**Problem:** The latent space is inherently sparse and disconnected.

**Insight:** *"You cannot select diverse samples if they do not exist."*

**Filtering algorithms fail when there is nothing to filter.**

**The Solution: Semantic Diversity (External)**

## [3] PANGEA: Proposed Method

# PANGEA: Proposed Method

---

Leveraging **Non-Relevant General Data**: Utilizing out-of-domain data (e.g., ORCA, FLAN)

to inject semantic diversity.

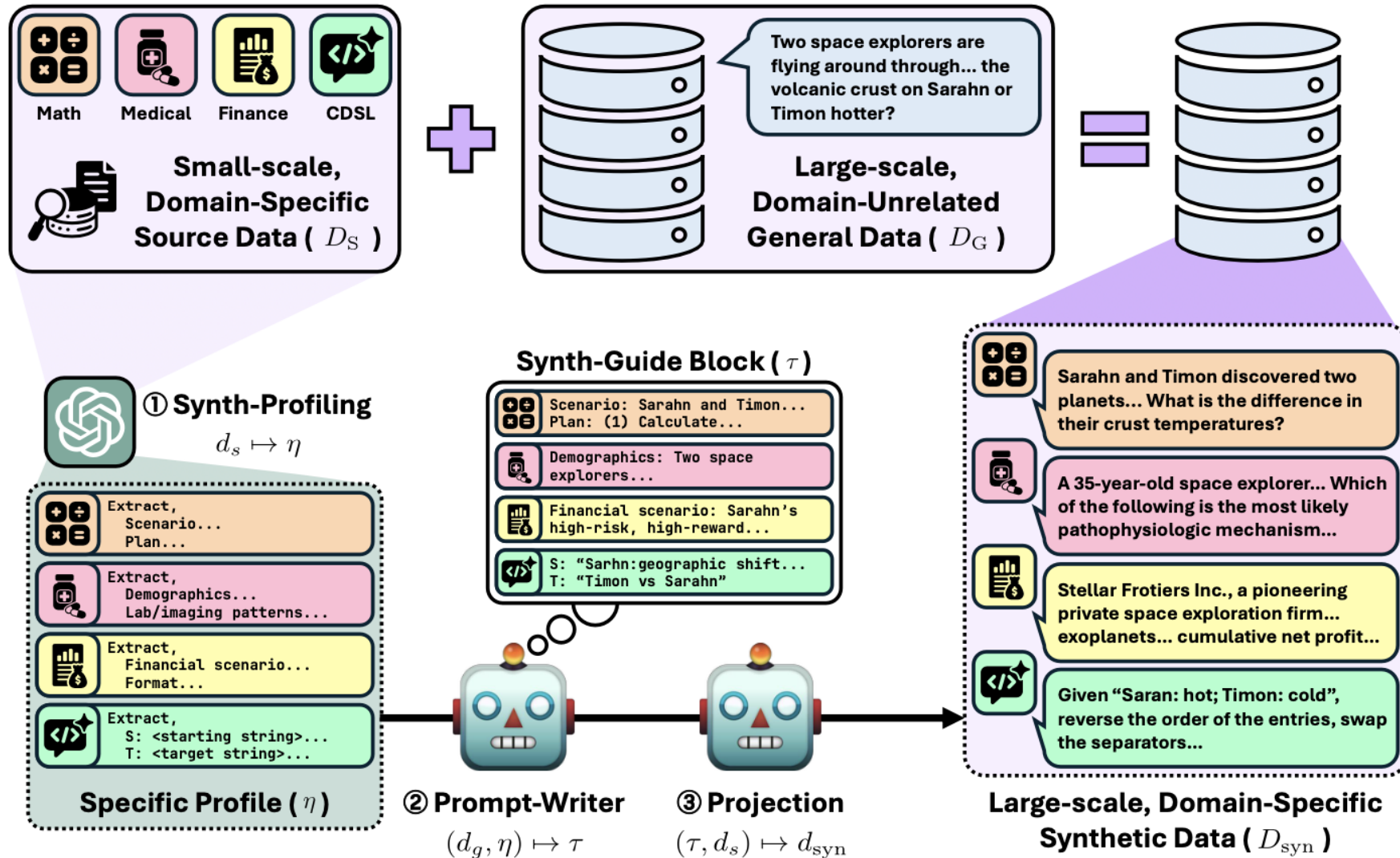
**Projection-Based Augmentation**: Transforming general scenarios into domain-specific

format.

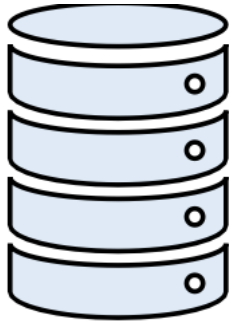
1. Synth-Profiling: Extracting domain blueprints from scarce seed.
2. Prompt-Writer: Distilling general data into structured guide blocks.
3. Projection: Mapping diverse semantics onto domain structures.

**Result:** Achieving **superior Semantic Diversity** (verified via Cosine

# Overview: PANGEA



# Why 3 stage generation Pipeline?

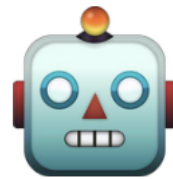


## Domain-Unrelated General Data

Two space explorers are flying around through an unknown galaxy when they discover two new planets. Given that these planets have no names, the explorers decide to name them after themselves. The planet of Sarahn is currently going through a large geographic shift. Many of the continents on the planet are currently breaking apart and moving. Timon, however, has been a stable planet with no continental moving for quite awhile now. Given the paragraph above, please answer correctly the following question: Is the volcanic crust on Sarahn or Timon hotter?

## Synth-Guide Block ( $\tau$ )

- Demographics: Two space explorers, planets named Sarahn and Timon
- Timeline: Current large geographic shift on Sarahn, stable period on Timon with no specified duration.
- Key clinical events: Continental break-apart and movement on Sarahn
- Lab / imaging patters: None explicitly stated, but implied tectonic activity on Sarahn
- Pathophysiologic principles: Geologic processes such as plate tectonics and volcanic activity, potentially leading to increased heat and volcanic crust temperature on Sarahn compared to Timon.



## Projection

$$(d_g, d_s) \mapsto d_{\text{syn}}$$

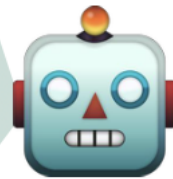


## Low-Quality Synthetic Case

A 45-year-old geologist, who has been studying the planetary movements of Sarahn and Timon, presents with symptoms of heat exhaustion after spending several hours on the surface of one of these planets. **Given the geological activity on these planets, which of the following is the most likely planet where the geologist was exposed to excessive heat, leading to his condition?**

Choose one of the following:

- A. Timon, due to its stable tectonic plates and lack of volcanic activity
- B. Sarahn, due to its current geographic shift and increased volcanic activity
- C. Timon, due to its high concentration of greenhouse gases in the atmosphere
- D. Sarahn, due to its low altitude and proximity to the sun



## Projection

$$(\tau, d_s) \mapsto d_{\text{syn}}$$



## High-Quality Synthetic Case

A 35-year-old space explorer on planet Sarahn presents with a 2-week history of worsening respiratory symptoms, including cough and shortness of breath, following a recent continental break-apart event. **The patient's oxygen saturation is 88% on room air, and a chest X-ray shows bilateral infiltrates. Which of the following is the most likely pathophysiologic mechanism contributing to the patient's symptoms?**

Choose one of the following:

- A. Increased volcanic crust temperature leading to sulfur dioxide gas exposure
- B. Decreased atmospheric pressure due to planetary geographic shift
- C. Inhalation of particulate matter from tectonic activity
- D. Hypobaric hypoxia resulting from high-altitude terrain formation

## [4] Experiments Results

# Main Results

Table 2: **Main results comparing synthetic data generation frameworks.** Accuracy (%) on four benchmarks is shown for increasing amounts of synthetic data (10k, 30k, and 120k). Our proposed PANGEA consistently outperforms the baselines across all data scales, with especially large margins at higher data volumes. All evaluations are conducted in a zero-shot setting.

# Synthetic	Method	Benchmarks				Avg. (impr.)
		GSM8K (↑)	MedQA (↑)	FinQA (↑)	CDSL (↑)	
-	Pre-trained	5.69	28.91	6.02	0.00	10.16
	Instruction-tuned	45.03	37.31	26.68	0.57	27.40
10k	Naive	26.91	35.42	24.06	3.20	22.40 (+12.24)
	Evol-Instruct	27.36	36.29	26.68	5.22	23.89 (+13.73)
	<b>PANGEA (ours)</b>	<b>32.52</b>	<b>37.78</b>	<b>36.44</b>	<b>11.30</b>	<b>29.51 (+19.35)</b>
30k	Naive	34.72	34.24	27.46	5.51	25.48 (+15.32)
	Evol-Instruct	32.51	38.09	29.64	9.57	27.45 (+17.29)
	<b>PANGEA (ours)</b>	<b>38.36</b>	<b>39.98</b>	<b>41.41</b>	<b>17.68</b>	<b>34.36 (+24.20)</b>
120k	Naive	42.68	38.02	32.43	6.96	30.02 (+19.86)
	Evol-Instruct	38.73	42.34	33.74	13.04	31.96 (+21.80)
	<b>PANGEA (ours)</b>	<b>48.61</b>	<b>44.62</b>	<b>50.22</b>	<b>35.36</b>	<b>44.70 (+34.54)</b>

# Diversity Analysis

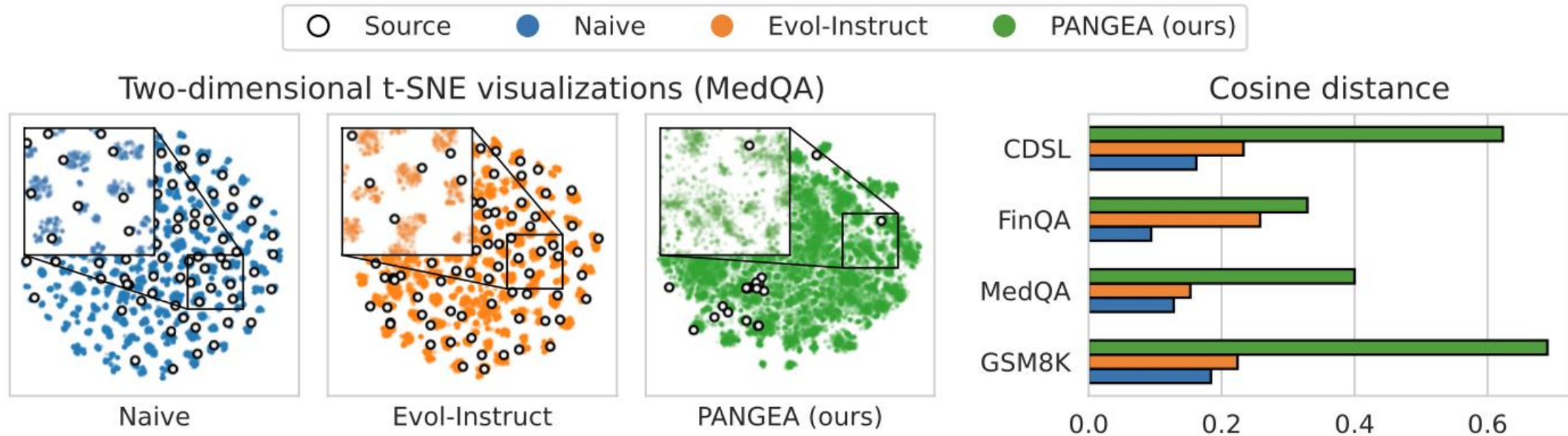


Figure 6: **Sentence embeddings for diversity analysis.** The first three plots show t-SNE visualizations of data generated by the Naive, Evol-Instruct, and PANGEA methods (colored dots), which are augmented from the source data (white dots), on the MedQA dataset. The bar plot on the right summarizes the average pairwise cosine distance of generated samples across datasets.

# Quality Analysis

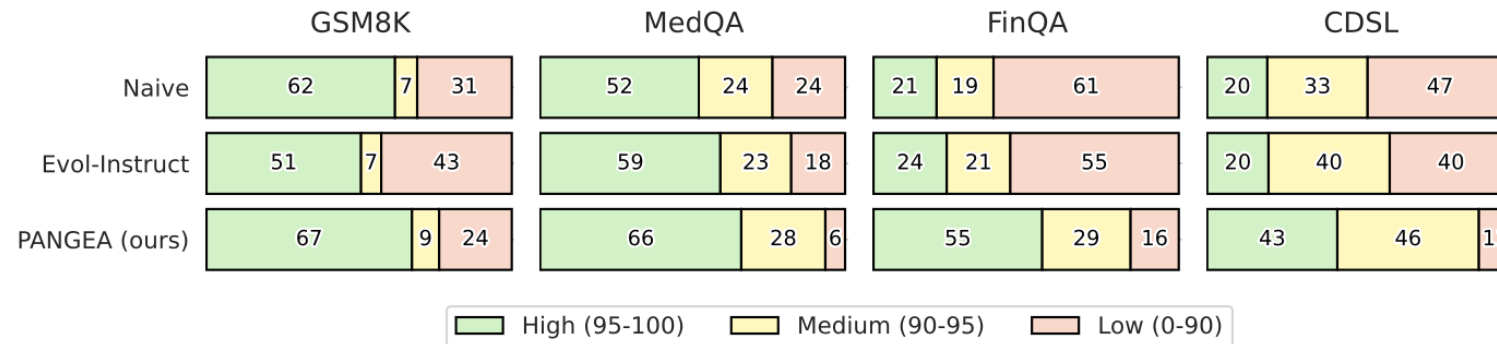


Figure 4: **o1 scores for data quality analysis.** The proportion (%) of generated data evaluated as High (95-100), Medium (90-95), or Low (0-90) quality by OpenAI’s o1 model.

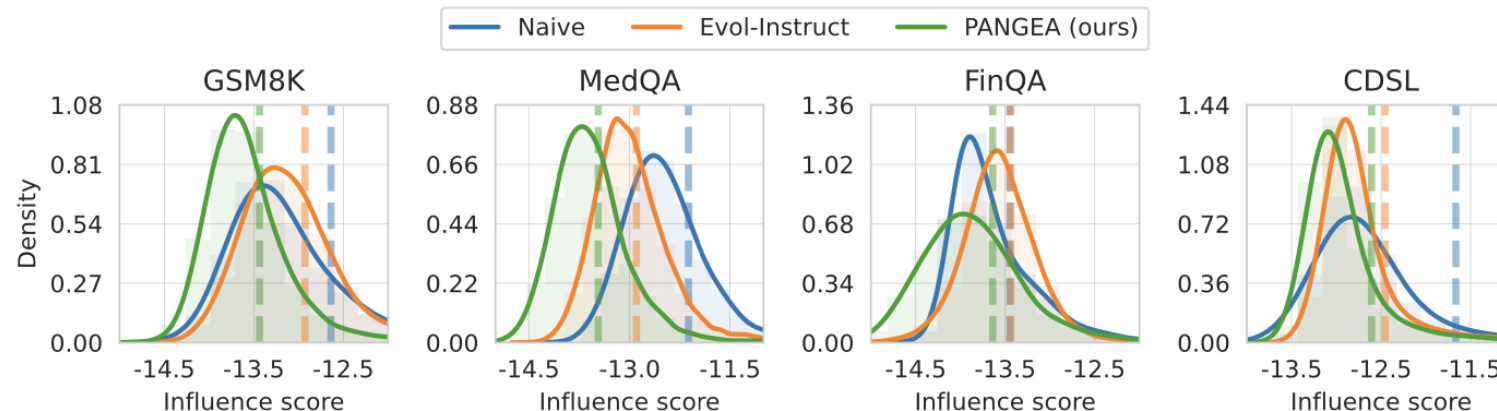


Figure 5: **Influence scores for data quality analysis.** Each histogram shows the distribution of data points for each augmentation method over influence score ranges. A kernel density estimation curve is overlaid as a solid line for better visualization, and the vertical dashed line denotes the mean.

# Ablation: Seed size & Model Scale

Table 3: **Model and seed scale results.** (a) shows how performance changes as the number of seed data increases from 10 to 100 under a fixed 10k synthetic set, while (b) reports the 8B-scale results with Llama-3.1-8B, where PANGEA outperforms all baseline models.

(a) **Seed-size ablation.** Robust down to 20–40; drop only at 10 due to an under-specified profile.

Method (#seed)	GSM8K	MedQA	FinQA	CDSL	Avg.
Naive (100)	26.91	35.42	24.06	3.20	22.40
Evol-Instruct (100)	27.36	36.29	26.68	5.22	23.89
PANGEA (100)	<b>32.52</b>	<b>37.78</b>	<b>36.44</b>	<b>11.30</b>	<b>29.51</b>
PANGEA (80)	32.91	37.34	36.37	11.01	29.41
PANGEA (40)	31.84	36.10	35.21	10.15	28.33
PANGEA (20)	31.21	35.79	34.83	8.70	27.63
PANGEA (10)	29.28	33.27	28.31	3.77	23.66

(b) **8B-scale setup.** With 8B scale, PANGEA steadily surpasses all baselines across every benchmark.

Method	GSM8K	MedQA	FinQA	CDSL
Pre-trained	48.75	39.31	25.63	1.61
Instruct-tuned	85.62	64.10	64.95	2.32
Naive	81.35	55.93	48.78	15.65
Evol-Instruct	79.08	60.02	53.88	17.42
<b>PANGEA</b>	<b>86.47</b>	<b>64.51</b>	<b>65.31</b>	<b>25.91</b>

# Generalization & Comparative results

Model	Method	Benchmarks				Avg. (impr.)
		GSM8K ( $\uparrow$ )	MedQA ( $\uparrow$ )	FinQA ( $\uparrow$ )	CDSL ( $\uparrow$ )	
Gemma2-2B	Pre-trained	1.66	27.81	7.23	0.28	9.00
	PANGEA	37.16	37.47	38.46	15.87	32.74 (+23.74)
Qwen2.5-1.5B	Pre-trained	3.98	26.45	27.83	0.00	14.57
	PANGEA	41.92	32.25	40.84	16.48	32.87 (+18.30)
Llama3.2-1B	Pre-trained	5.69	28.91	6.02	0.00	10.16
	PANGEA	38.36	39.98	41.41	17.68	34.36 (+24.20)
DeepSeek-R1-Distill-Llama-8B	Pre-trained	85.29	57.09	61.29	6.96	52.66
	PANGEA	88.91	66.53	69.75	28.70	63.47 (+10.81)

Table 7: **Comparative results.** Benchmark results on GSM8K and MedQA.

Method	GSM8K	MedQA
MuMath	30.25	-
UltraMedical	-	36.71
<b>PANGEA (ours)</b>	<b>32.52</b>	<b>37.78</b>

**Thanks for listen!**