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

Seungyoo Lee[1], Giung Nam[1], Moonseok Choi[1], Hyungi Lee[2] and Juho Lee[1]
[1]KAIST, [2]Kookmin University

PANGEA

A fully automated, projection-based augmentation framework that uses **non-relevant general data** for enhanced domain adaptation, improving diversity without additional annotation costs.

❖ Stage 1. Synth Profiling

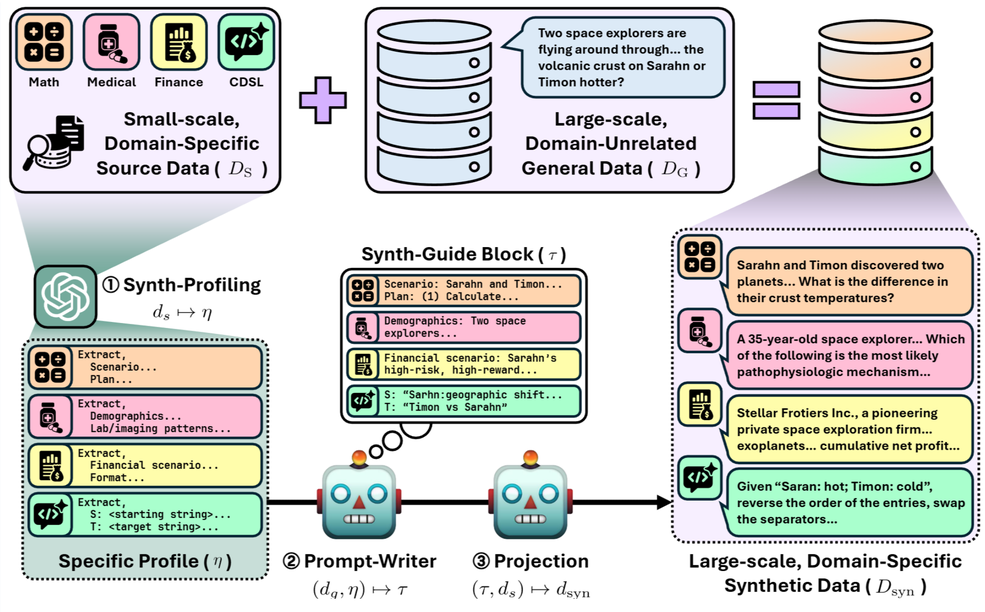
Identifies and extracts meaningful domain-specific information from irrelevant general data using an LLM to create a structured domain-specific profile (η).

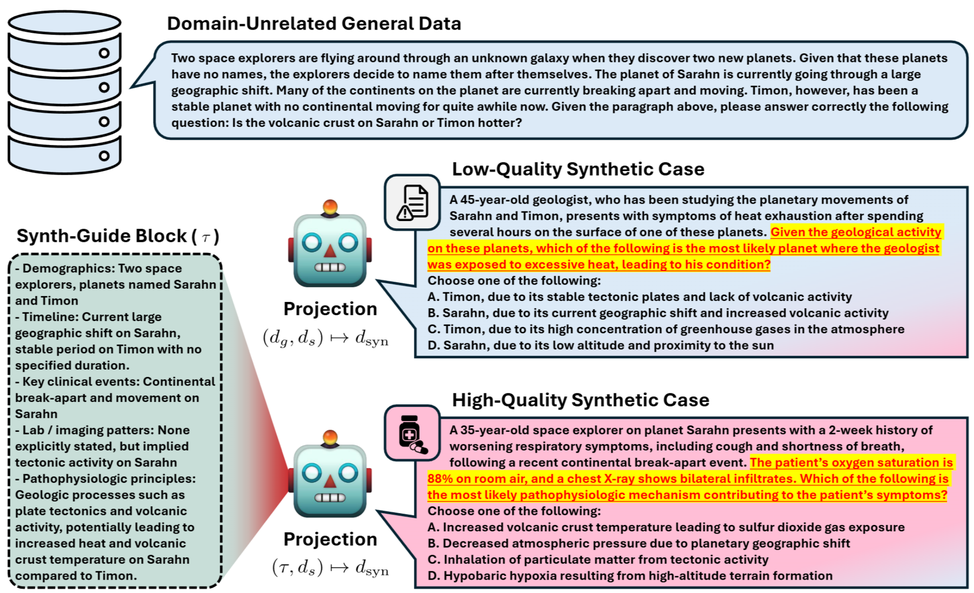
❖ Stage 2. Prompt Writer

A frozen LLM uses the η to systematically extract structured Synth-Guide Blocks (τ) from irrelevant general data instances (d_g), ensuring diverse combinations of τ.

❖ Stage 3. Synthetic Data Generation

The Synth-Guide Blocks (τ) are projected onto the domain-specific source data (d_s) using an LLM to generate synthetic data (d_{syn}) that is aligned with the target domain's format and learning needs.

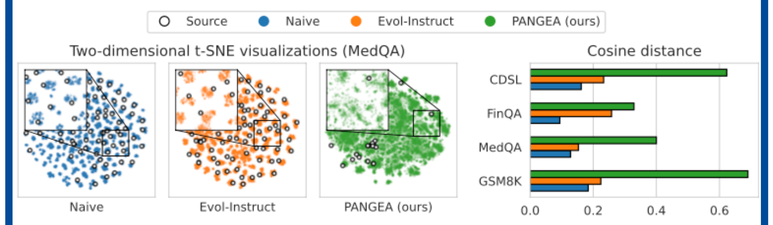




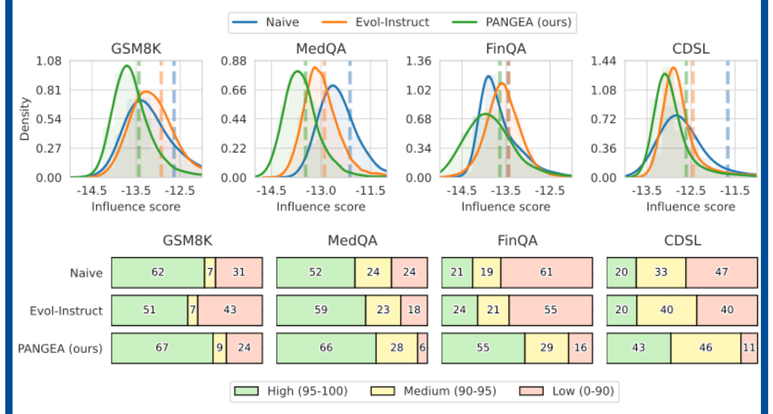
Experimental Results

# Synthetic	Method	Benchmarks GSM8K (↑)	MedQA (↑)	FinQA (↑)	CDSL (↑)	Avg. (impr.)
-	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)
	PANGEA (ours)	**32.52**	**37.78**	**36.44**	**11.30**	**29.51 (+19.35)**
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)
	PANGEA (ours)	**38.36**	**39.98**	**41.41**	**17.68**	**34.36 (+24.20)**
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)
	PANGEA (ours)	**48.61**	**44.62**	**50.22**	**35.36**	**44.70 (+34.54)**

Main Results



Diversity Analysis



Quality Analysis