DESIGNER:

Design-Logic-Guided Multidisciplinary Data Synthesis for LLM Reasoning

Weize Liu*, Yongchi Zhao*, Yijia Luo, Mingyu Xu, Jiaheng Liu,
Yanan Li, Xiguo Hu, Zhiqi Bai, Yuchi Xu, Wenbo Su, Bo Zheng

@ Alibaba Group | Foundation Model Training Team, Future Living Lab

ICLR 2026

Alibaba

Future Living Lab

The Data Bottleneck in the Long CoT Era

Post-training & mid-training of LLMs rely heavily on **exam-style reasoning data**

- Math & Code: rich resources (competitions, open benchmarks)
- other disciplines (medicine, law, humanities...):
  severe data scarcity
- Existing reasoning datasets are heavily skewed toward STEM

Math & Code
Abundant Data ✓

vs

75 Other Disciplines
Medicine, Law, Philosophy,
Economics, Agriculture...
Severe Data Scarcity ✗

Two Paradigms, Two Fundamental Limitations

Query-Centric

Self-Instruct, WizardLM, AutoEvol...

Evolve from seed questions iteratively

- ✗ Limited by seed pool coverage
- ✗ Model bias compounds over iterations
- ✗ Can't escape seed distribution

Document-Centric

MAmmoTH2, NaturalReasoning, Humpback...

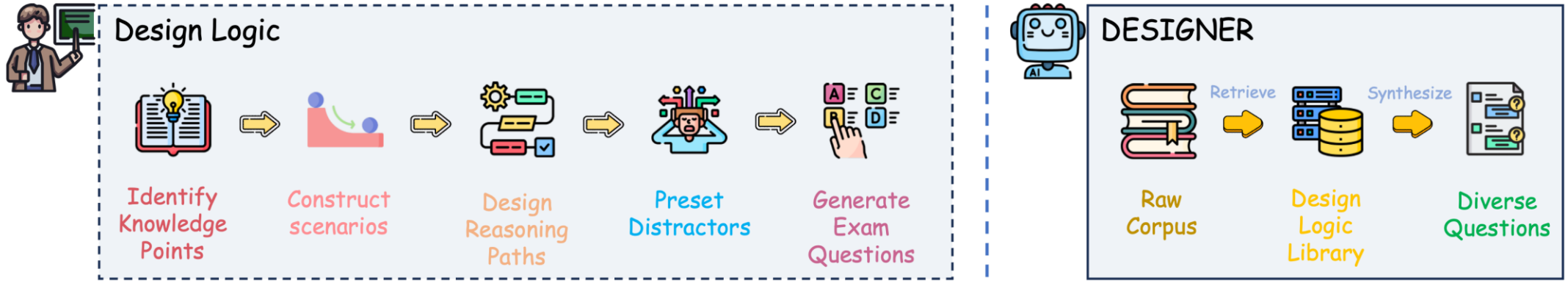
Generate questions from raw documents

- ✗ No control over difficulty or diversity
- ✗ Degenerates to factual recall
- ✗ Diversity limited by fixed prompt

Key Question: *How can we leverage diverse corpora for broad subject and knowledge coverage while synthesizing high-quality, human-like exam questions with controlled difficulty, diversity, and question types?*

Methods

How Do Human Experts Design Hard Exam Questions?

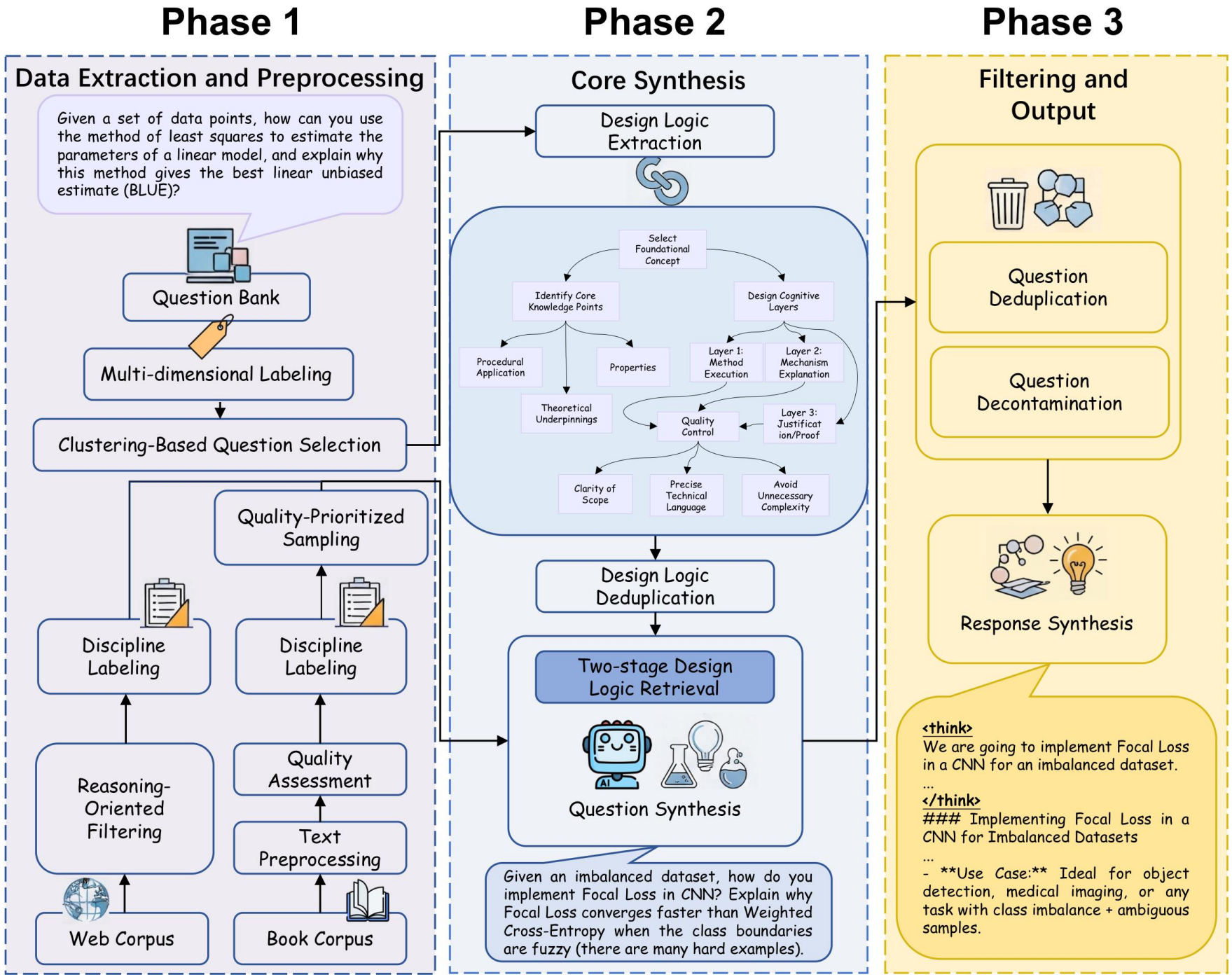


"Design Logic" = Reusable meta-knowledge that encapsulates **the structured design process**

Same Design Logic + **Different Source Text** = **New question with same reasoning complexity**

Best of both: **Query-Centric**'s rich knowledge coverage + **Document-Centric**'s explicit difficulty control — scalable, and without the drawbacks of either

Pipeline


Phase 1
Phase 2
Phase 3
Data Extraction and Preprocessing
Given a set of data points, how can you use the method of least squares to estimate the parameters of a linear model, and explain why this method gives the best linear unbiased estimate (BLUE)?
Question Bank
Multi-dimensional Labeling
Clustering-Based Question Selection
Quality-Prioritized Sampling
Discipline Labeling
Discipline Labeling
Reasoning-Oriented Filtering
Quality Assessment
Text Preprocessing
Web Corpus
Book Corpus
Core Synthesis
Design Logic Extraction
Select Foundational Concept
Identify Core Knowledge Points
Design Cognitive Layers
Procedural Application
Properties
Theoretical Underpinnings
Layer 1: Method Execution
Layer 2: Mechanism Explanation
Layer 3: Justification/Proof
Quality Control
Clarity of Scope
Precise Technical Language
Avoid Unnecessary Complexity
Design Logic Deduplication
Two-stage Design Logic Retrieval
Question Synthesis
Given an imbalanced dataset, how do you implement Focal Loss in CNN? Explain why Focal Loss converges faster than Weighted Cross-Entropy when the class boundaries are fuzzy (there are many hard examples).
Filtering and Output
Question Deduplication
Question Decontamination
Response Synthesis
<think>
We are going to implement Focal Loss in a CNN for an imbalanced dataset.
...
</think>
Implementing Focal Loss in a CNN for Imbalanced Datasets
...
- **Use Case:** Ideal for object detection, medical imaging, or any task with class imbalance + ambiguous samples.

Methods

Two-Stage Retrieve-and-Generate

Raw Corpus

Design Logic Library

→

Stage 1: Coarse Retrieval

Embed source document
↓
Vector similarity search
↓
Retrieve top-5 Design Logics

→

Stage 2: Fine Selection + Generation

LLM evaluates compatibility
↓
Selects best Design Logic
↓
Generates question from
Design Logic + source text

Why two stages? 125k logics × millions of documents = combinatorial explosion.

- Embedding retrieval is fast but imprecise.
- LLM selection is accurate but expensive.

Coarse retrieval narrows the search space; LLM fine-selection ensures quality and compatibility.
The two-stage design balances **scalability** (millions of texts) with **precision**.

Results

4.7 Million Questions Across 75 Disciplines

DLR-Book

3.04M

questions from book corpus

DLR-Web

1.66M

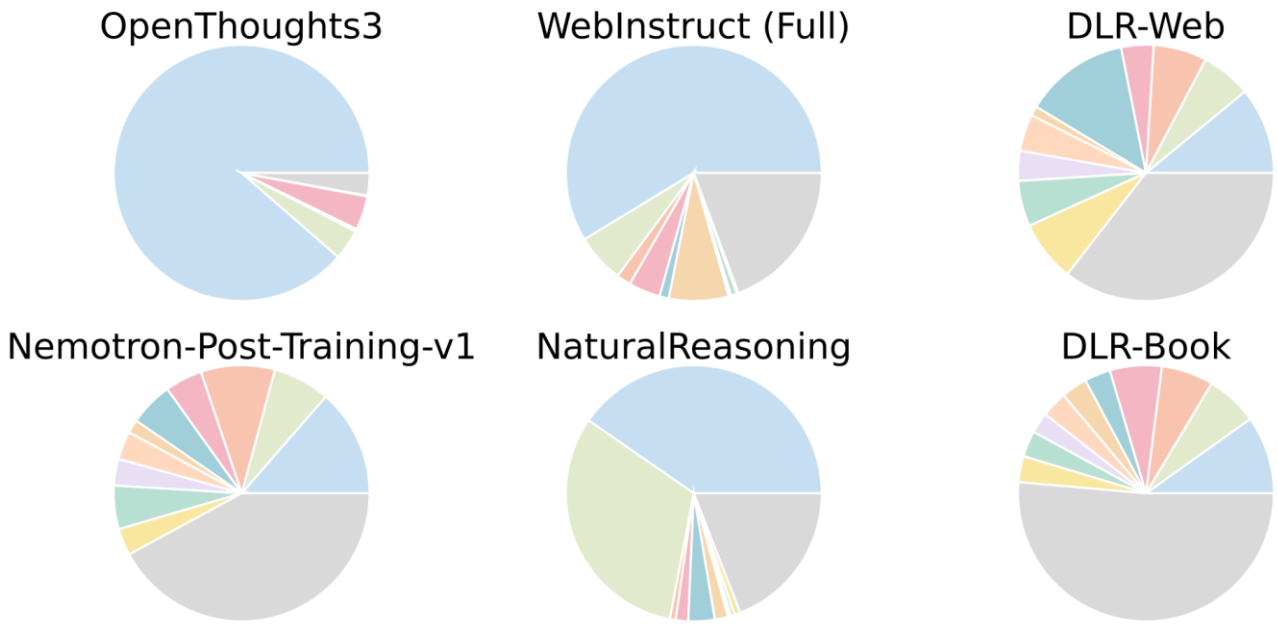
questions from web corpus

Balanced Coverage

75

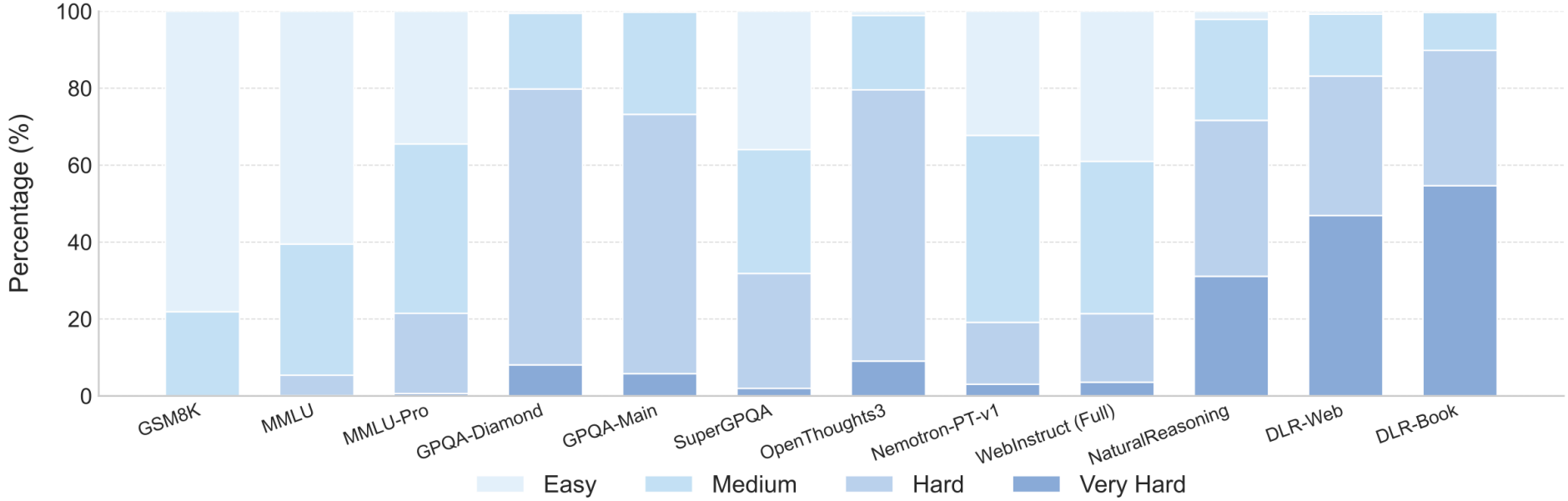
diverse disciplines

Discipline Distribution



Results

Difficulty Distribution



Dataset	Mean Cosine Distance	Mean L2 Distance	1-NN Distance	Cluster Inertia	Radius
OpenThoughts3	0.8037	1.2656	0.0051	206,369.22	0.0172
Nemotron-Post-Training-v1	0.8243	1.2827	0.1290	226,211.00	0.0174
WebInstruct (Full)	0.7762	1.2436	0.1830	205,590.02	0.0169
NaturalReasoning	0.8233	1.2818	0.1915	226,288.91	0.0173
DLR-Web	**0.8494**	**1.3026**	**0.3897**	238,039.50	**0.0177**
DLR-Book	0.8471	1.3008	0.3726	**238,100.26**	0.0176

Diversity: ~2× gains on 1-NN embedding distance vs. all baselines

Results

SFT on Base Models Surpasses Official Post-Trained Models

Model	MMLU	MMLU-Pro	GPQA-Diamond		GPQA-Main		SuperGPQA
	Accuracy	Accuracy	Accuracy	CoT-SC	Accuracy	CoT-SC	Accuracy
Llama-3.2-3B-Instruct	57.71	31.18	21.97±2.54	20.20	25.07±1.64	24.78	16.39
Llama-3.2-3B-SFT (DLR-Web)	66.74	49.81	22.42±2.06	20.71	25.54±1.39	26.12	20.31
Llama-3.2-3B-SFT (DLR-Book)	70.61	57.09	38.33±2.63	44.44	33.82±1.17	35.27	26.39
Llama-3.2-3B-SFT (DLR-Web+Book)	73.53	61.36	42.27±1.72	43.94	38.91±1.78	43.97	29.64
Llama-3.1-8B-Instruct	70.86	47.38	23.18±1.78	24.75	27.99±1.40	28.57	20.08
Llama-3.1-8B-SFT (DLR-Web)	81.75	72.64	57.73±2.16	63.64	55.45±2.07	58.71	39.66
Llama-3.1-8B-SFT (DLR-Book)	83.33	74.94	63.23±1.37	66.67	62.25±1.31	67.19	43.48
Llama-3.1-8B-SFT (DLR-Web+Book)	84.13	76.04	65.45±1.47	70.71	63.62±1.23	67.86	45.06
Qwen3-4B (Thinking Mode)	82.87	69.34	54.70±2.42	58.08	49.51±1.40	51.12	43.30
Qwen3-4B-Base-SFT (DLR-Web)	83.55	71.24	53.74±3.33	60.61	51.27±1.57	55.36	42.73
Qwen3-4B-Base-SFT (DLR-Book)	84.73	73.03	62.58±1.36	68.69	56.85±0.91	61.16	45.86
Qwen3-4B-Base-SFT (DLR-Web+Book)	85.00	73.06	63.69±2.15	70.20	58.73±1.36	63.62	46.15
Qwen3-8B (Thinking Mode)	85.85	73.62	59.44±2.53	60.61	57.95±1.47	59.38	47.52
Qwen3-8B-Base-SFT (DLR-Web)	86.82	75.62	63.28±2.43	66.67	61.43±0.98	66.07	48.66
Qwen3-8B-Base-SFT (DLR-Book)	87.53	76.69	69.39±1.87	73.74	65.07±0.98	68.30	50.57
Qwen3-8B-Base-SFT (DLR-Web+Book)	**87.60**	**76.72**	**71.01±2.33**	**75.76**	**65.40±1.05**	**69.20**	**50.90**

Avg Improvement

- Llama-3.2-3B: **+18.68**
- Llama-3.1-8B: **+28.96**
- Qwen3-4B: **+5.38**
- Qwen3-8B: **+5.45**

Using ONLY our synthesized data for SFT on base models,
we surpass official post-trained final models.

Results

SFT on Base Models Surpasses All Baselines

Dataset	MMLU	MMLU-Pro	GPQA-Diamond	GPQA-Diamond (CoT-SC)	SuperGPQA
OpenThoughts3	72.49	57.76	45.86±1.80	54.04	39.70
Nemotron-Post-Training-v1	77.17	62.52	38.59±1.24	40.91	42.03
WebInstruct (Full)	86.34	72.83	55.61±2.50	62.63	45.37
NaturalReasoning	85.33	72.39	56.67±2.20	60.00	43.38
DLR-Web	86.32	73.81	58.89±1.98	63.64	**47.23**
DLR-Book	**86.43**	**74.98**	**60.35±1.93**	**66.67**	47.04

Using our synthesized data for SFT on base models, **we surpass** all baselines.

* All results on Qwen3-8B-Base.

Ablation: Every Component Matters

Method	MMLU	MMLU-Pro	GPQA-Diamond	GPQA-Diamond (CoT-SC)	SuperGPQA
DESIGNER	**86.43**	**74.98**	**60.35±1.93**	**66.67**	**47.04**
w/o Design Logic	86.26	74.34	58.89±2.94	<u>64.65</u>	46.71
w/o Coarse Ranking	<u>86.29</u>	73.76	58.74±0.75	62.63	46.23
w/o Fine Ranking	86.26	<u>74.35</u>	<u>59.34±2.20</u>	63.64	<u>46.81</u>
DESIGNER (Full)	87.53	76.69	69.39±1.87	73.74	50.57
w/o Design Logic (Full)	86.51	75.54	64.10±2.81	67.17	48.46

Design Logic = Key

5.3 points drop on GPQA-Diamond when removed

Both retrieval stages matter

Coarse + Fine > either alone

Thank You

Weize Liu