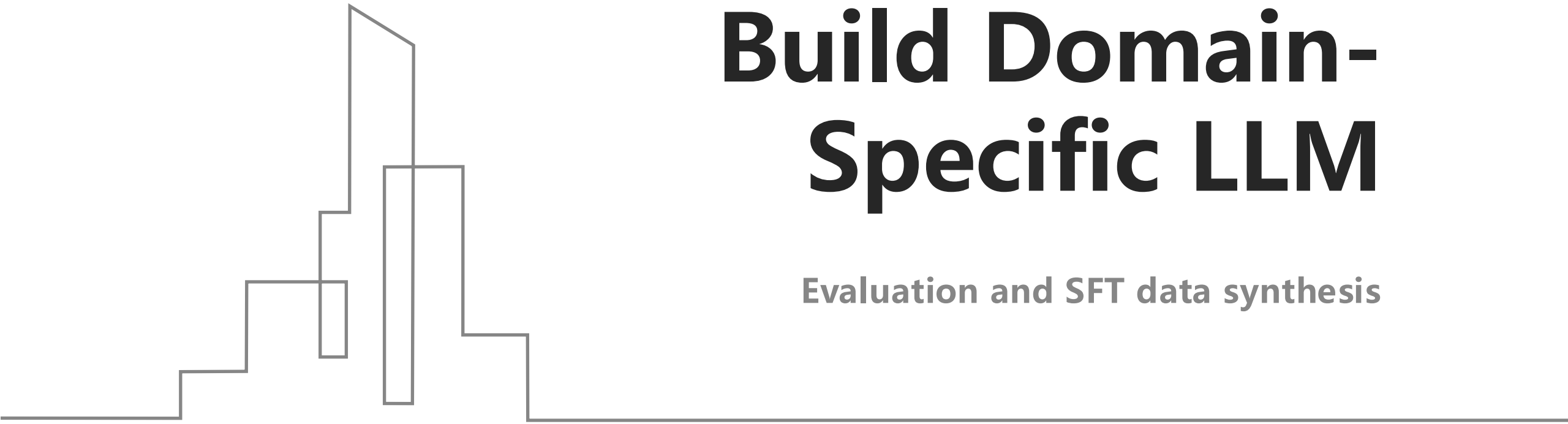


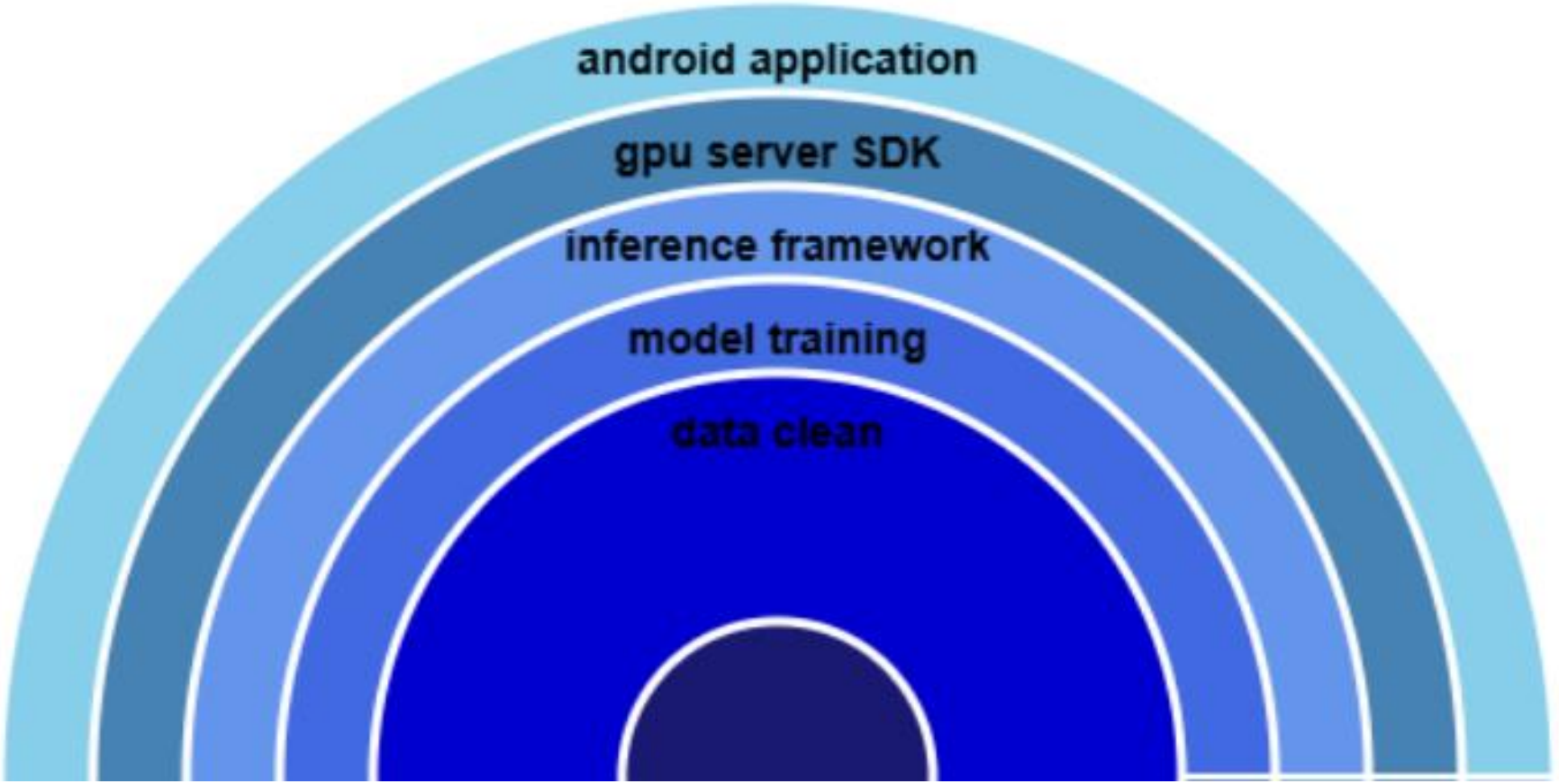
Build Domain-Specific LLM

Evaluation and SFT data synthesis

Kong Huanjun GitHub: **tpoisonooo**

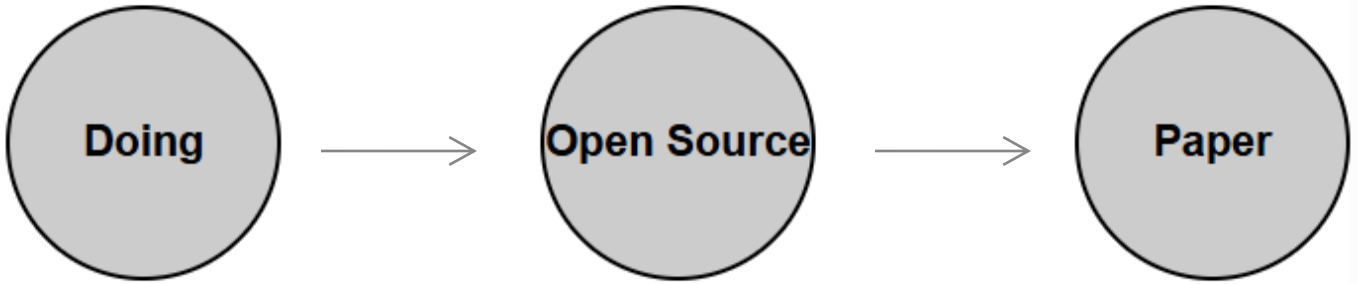
Brief Self-Introduction

As a 10 years’ engineer, I have built..



Engineer to Research

This is how we engineer work:



Our Github repo is by-product.

Our publications are the **by-product of a by-product**.

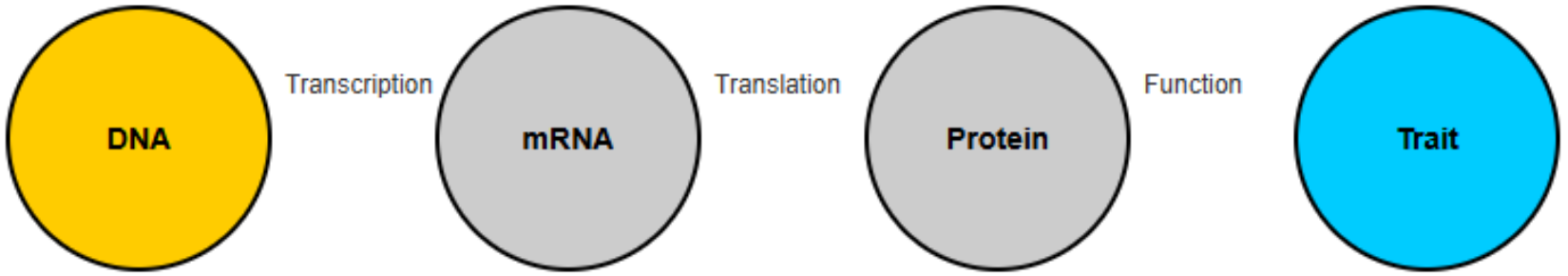
Our End2End Prediction Task

Input:

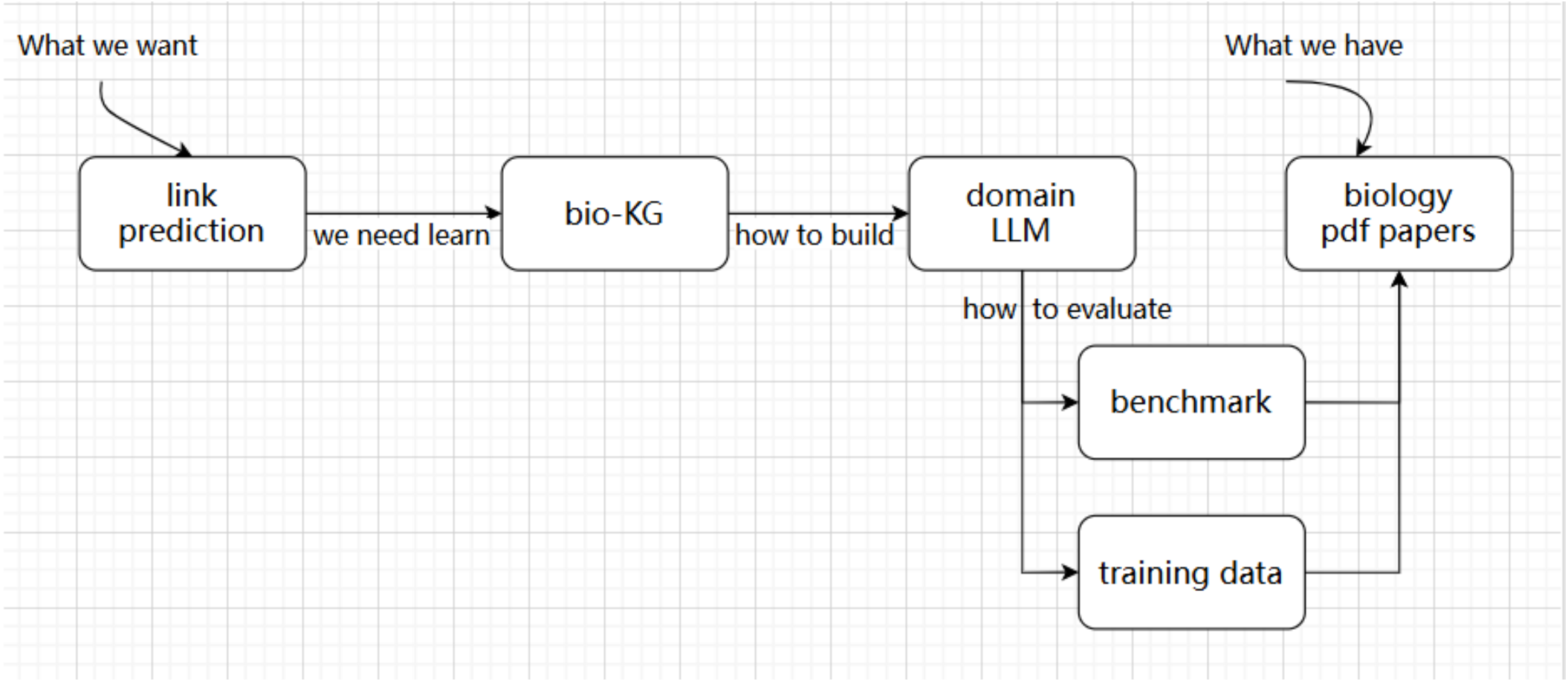
- Target **Trait** (e.g: salt stress)
- about 1 milion papers

Output:

- **DNA** list (e.g: DNA set that affect salt stress)



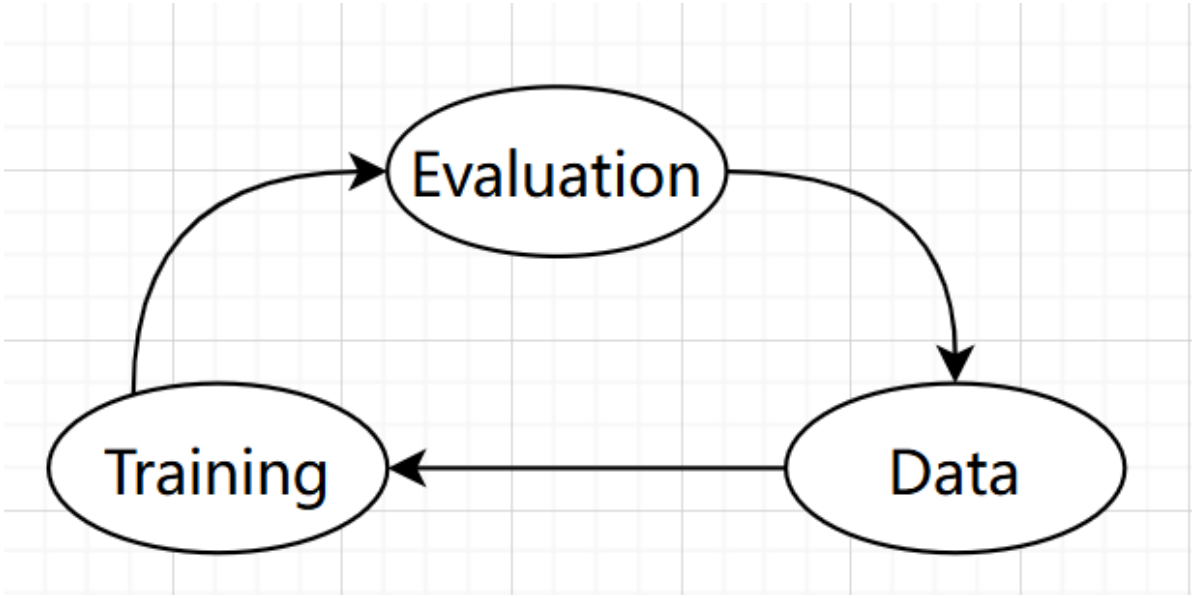
Our Work Overview



Key Task List

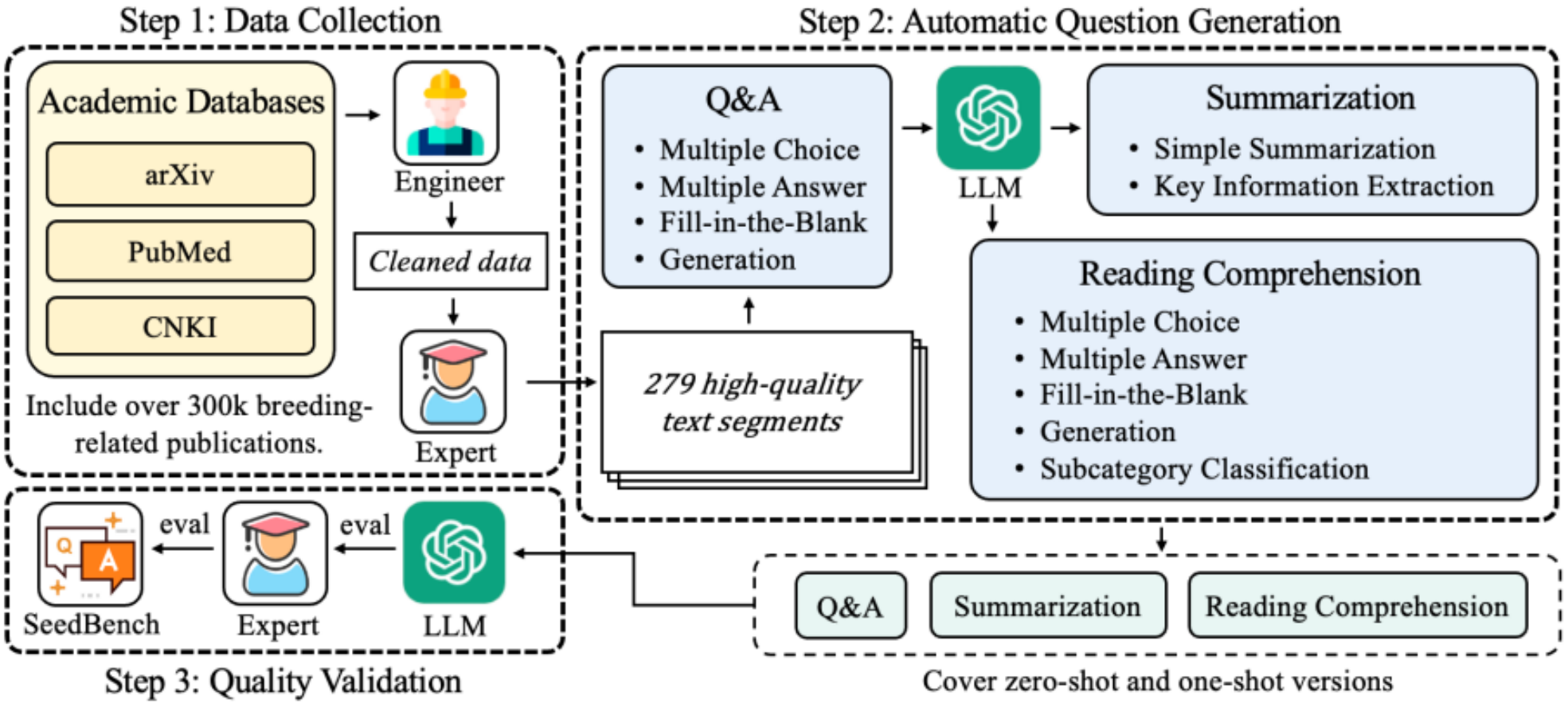
Training LLM is a Test-Driven Development loop:

1. How to evaluate LLM in Seed Science(our domain)? **SeedBench**
2. How to synthesis training data? **GraphGen**



SeedBench Generation Pipeline

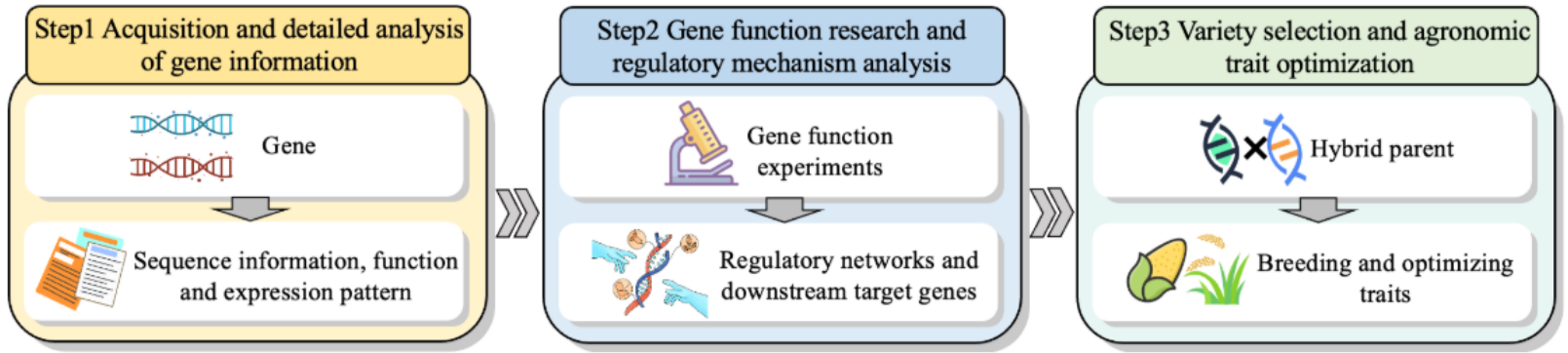
Human experts are necessary in the evaluation phase



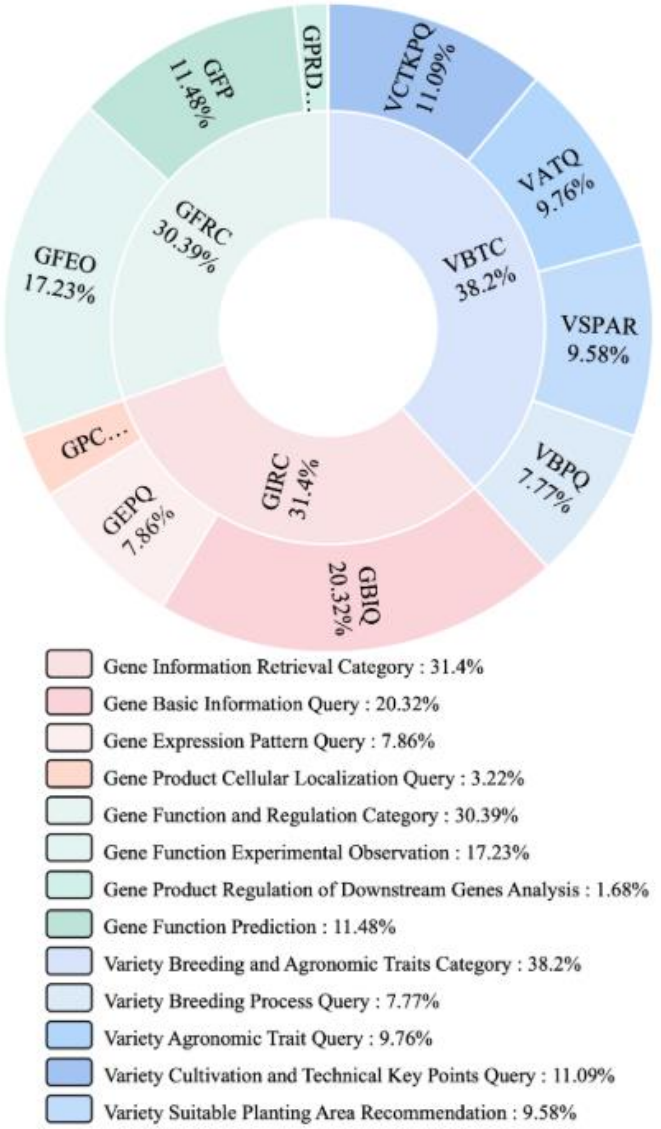
SeedBench Data Distribution

Question topic(**right**) align with

human expert workflow(**left**)



Breeding Expert Workflow Framework



SeedBench: A Multi-task Benchmark for Evaluating Large Language Models

in Seed Science (ACL2025 main)

https://github.com/open-sciencelab/SeedBench

SeedBench Details

LLM-friendly format

- Multiple choice
- Multiple answer
- Fill-in-the-Black
- Generation

What is **minimal `Count` ?**

Type ID	Question Type	Metric	Count
Q&A			
QA-1	Multiple Choice	Accuracy	200
QA-2	Multiple Answer	Macro-F1	187
QA-3	Fill-in-the-Blank	ROUGE-L	224
QA-4	Generation	ROUGE-L	242
Summarization			
SUM-1	Simple Summarization	ROUGE-L	225
SUM-2	Key Information Extraction	ROUGE-L	225
Reading Comprehension			
RC-1	Multiple Choice	Accuracy	113
RC-2	Multiple Answer	Macro-F1	108
RC-3	Fill-in-the-Blank	ROUGE-L	221
RC-4	Generation	ROUGE-L	240
RC-5	Subcategory Classification	Accuracy	279

SeedBench Findings

Reasoning model may not be

good at knowledge-intensive tasks

That means:

GPU-poor still have a chance to win

as long as there is domain data

Models	Breeding Subcategories										Average
	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	
Proprietary LLMs											
Claude-3.5-Sonnet	48.77	57.72	66.02	57.54	47.82	49.36	57.47	60.11	58.06	58.89	55.45
Gemini-1.5-Pro	47.00	59.55	62.42	59.56	43.11	49.55	53.41	56.18	52.51	53.71	53.58
Gemini-2.0-Flash	33.67	27.37	53.04	32.07	25.87	44.41	33.57	36.77	31.78	31.70	34.24
GLM-4-Plus	52.72	59.62	70.62	60.11	50.60	56.75	65.02	64.17	61.70	62.90	59.61
GPT-4	59.59	60.55	76.32	61.16	56.34	59.35	63.67	64.74	60.65	67.66	62.06
GPT-4o mini	54.24	56.64	72.11	59.28	53.00	57.88	58.38	61.75	57.50	62.38	58.40
OpenAI o1-mini	49.16	55.58	59.37	54.77	44.43	50.73	54.57	55.36	54.91	54.19	53.25
Open-Source LLMs											
DeepSeek-V3-671B	56.03	62.42	74.81	63.17	55.23	58.84	68.23	69.04	66.46	68.48	63.30
GLM-4-Chat-9B	23.28	21.31	39.97	26.13	16.20	34.15	26.63	29.60	25.60	26.68	26.55
InternLM2-7B	27.55	21.14	39.64	28.57	15.16	36.12	28.74	30.80	27.32	29.22	28.71
InternLM2.5-7B	51.71	55.75	67.88	50.48	44.14	56.73	51.28	54.91	52.46	56.24	53.51
Llama3.1-8B	43.89	31.21	42.53	40.68	38.47	43.80	42.87	51.62	41.88	40.91	42.23
Llama3.1-70B	48.72	55.41	64.77	53.67	46.73	54.08	56.94	57.72	55.31	57.56	54.30
Llama3.3-70B	45.32	47.15	60.62	49.76	40.90	54.30	52.79	54.61	49.98	55.05	50.53
Mistral-v0.3-7B	42.61	38.28	57.02	40.41	29.97	44.22	36.31	43.98	39.92	43.51	41.59
Qwen2-0.5B	32.84	25.15	40.19	28.20	27.62	37.22	33.81	33.63	28.25	31.67	31.44
Qwen2-7B	44.21	40.41	63.00	47.36	35.37	52.30	45.61	48.73	44.88	46.89	46.51
Qwen2-57B	53.67	49.81	74.30	58.38	39.34	54.71	63.89	59.57	59.22	60.08	57.20
Qwen2-72B	51.16	58.10	74.07	59.72	51.58	57.76	58.85	61.63	56.69	59.11	57.62
Qwen2.5-7B	45.16	39.50	66.01	44.61	35.72	50.00	53.60	53.31	53.06	51.05	48.45
Qwen2.5-14B	50.91	50.73	68.62	52.15	47.14	54.54	57.02	62.05	54.37	54.15	54.21
Qwen2.5-72B	46.86	47.41	70.99	51.89	46.17	57.60	55.35	56.31	53.05	54.75	52.63
QwQ-32B	32.24	21.06	47.11	29.14	28.56	39.68	38.17	39.56	34.70	34.52	33.55
Domain Specific LLMs											
Aksara-v1-7B	36.72	36.69	48.32	35.41	24.26	36.83	31.17	34.64	31.15	34.14	35.04
PLLaMa-7B	17.85	13.69	17.99	16.81	11.66	21.67	14.34	17.36	12.39	16.11	16.46
PLLaMa-13B	15.10	14.18	28.41	18.83	13.96	23.28	18.53	17.37	14.15	18.51	17.57

GraphGen

Simple RAG in QA Data Generation

❌ Too much hullucination

The best LLM (DeepSeek-V3) not good at Seed Science !

❌ Unaffordable

To inject knowledge, SFT needs ~1 million QA pairs

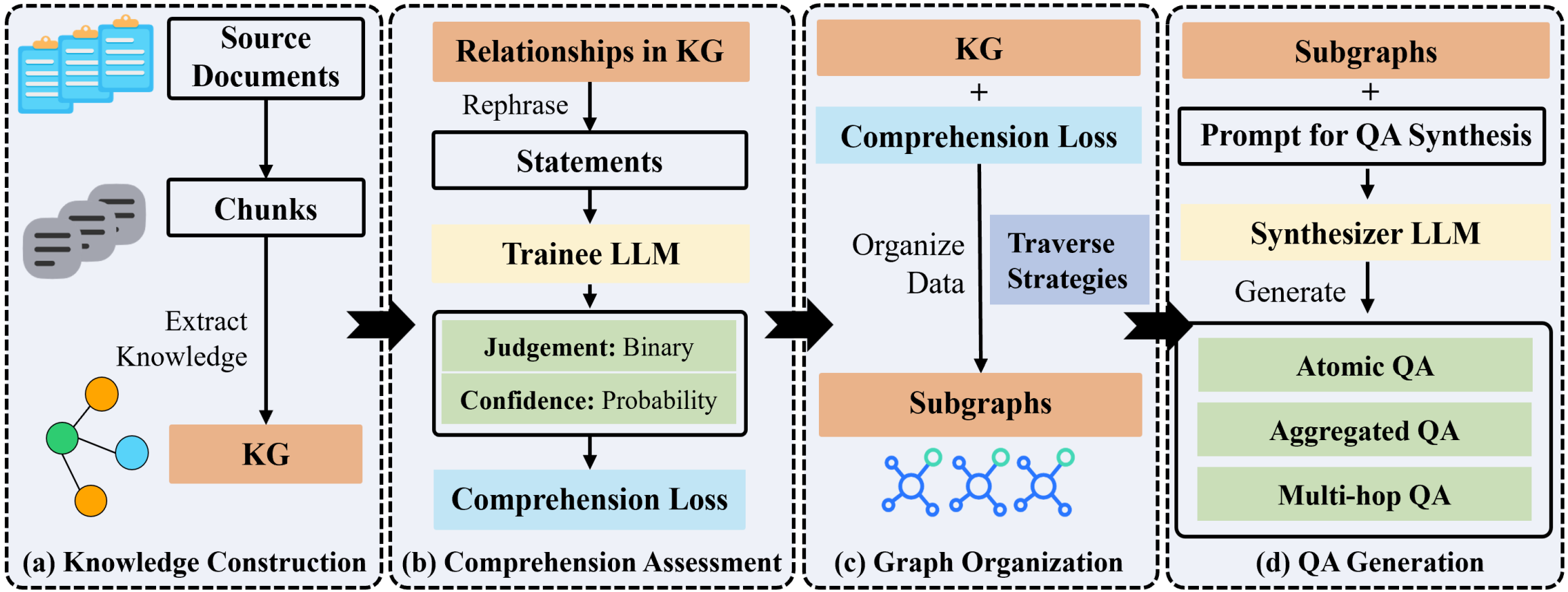
Human (expert)’ s time is very expensive

Wait, we still have chance..

√ LLM knows the grammar (subject, verb, object..), even if it knows nothing about terminology

Input sentence “AGIS affect salt stress”, LLM understands “AGIS” is subjective

GraphGen Pipeline



https://github.com/open-sciencelab/GraphGen

GraphGen Training Result

Here is post-training result which **over 50% SFT data** comes from GraphGen and our data clean pipeline.

Domain	Dataset	Ours	Qwen2.5-7B-Instruct (baseline)
Plant	SeedBench	**65.9**	51.5
Common	CMMLU	73.6	**75.8**
Logic	GPQA-Diamond	**40.0**	33.3
Math	AIME24	**20.6**	16.7
	AIME25	**22.7**	7.2

Q&A

please speak slowly

群聊：HuixiangDou&GraphGen 6群

HuixiangDou

https://github.com/InternLM/HuixiangDou

GraphGen

https://github.com/open-sciencelab/GraphGen

SeedBench

https://github.com/open-sciencelab/SeedBench