



# Bee: A High-Quality Corpus and Full-Stack Suite to Unlock Advanced Fully Open MLLMs

Yi Zhang<sup>1,2</sup>, Bolin Ni<sup>2</sup>, Xin-Sheng Chen<sup>1</sup>, Heng-Rui Zhang<sup>1</sup>, Yongming Rao<sup>2</sup>,  
Houwen Peng<sup>2\*</sup>, Qinglin Lu<sup>2</sup>, Han Hu<sup>2</sup>, Meng-Hao Guo<sup>1\*</sup>, Shi-Min Hu<sup>1†</sup>

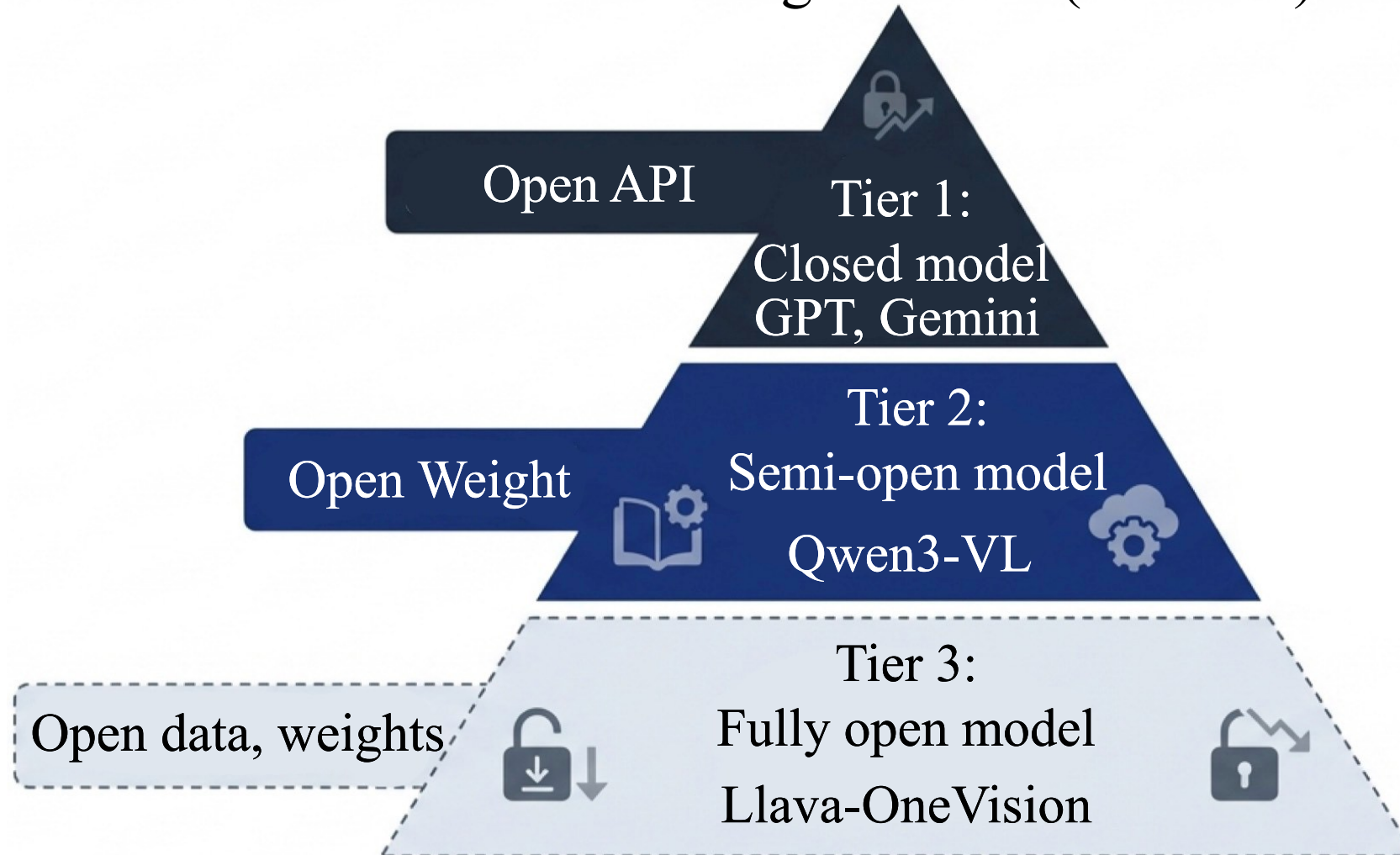
<sup>1</sup>Tsinghua University, <sup>2</sup>Tencent Hunyuan Team

\*Project lead. †Corresponding author.

Homepage: <https://open-bee.github.io>

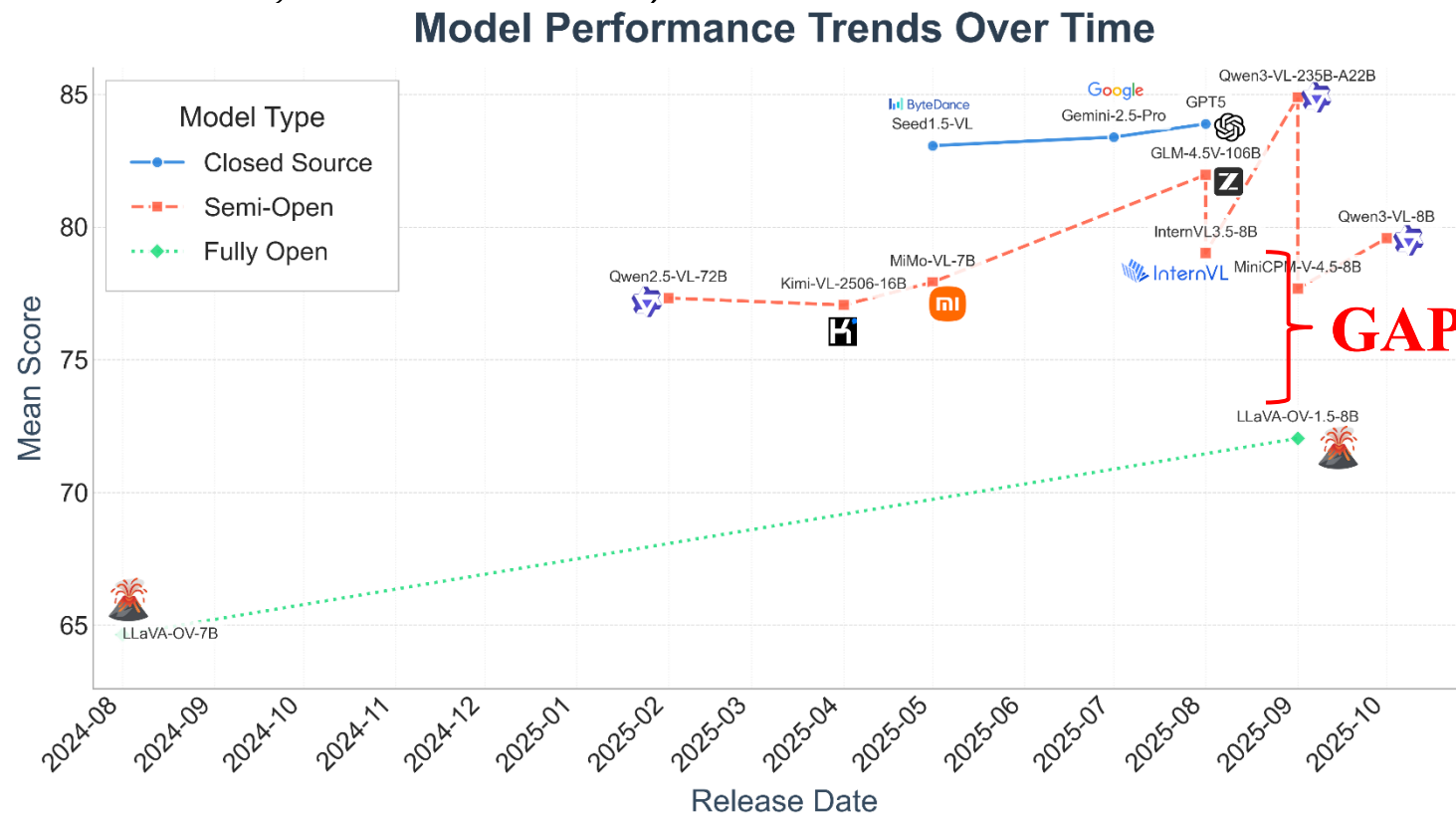
# Current Levels of MLLM

- Three Levels of Multimodal Large Models (MLLMs)



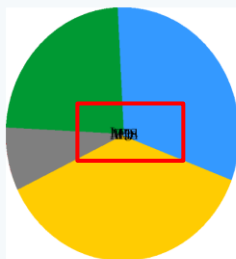
# Motivation and Challenges

- The fully open MLLMs are significantly lagging behind the top proprietary models (such as Gemini, GPT-5) and semi-open models (such as Qwen3-VL, GLM-4.5V).



# Motivation and Challenges

- There is a huge gap in data quality during the Training stage.
  - ◆ Common data noise: Existing open-source datasets are filled with factual errors, mismatches between text and images, and format defects.



**User Question:**

Based on the pie chart representing the distribution of parliamentary seats in Italy, which party has the second-largest share, and what percentage of seats does it hold?

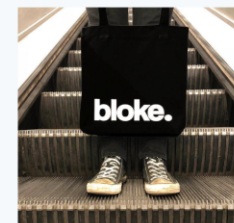
**Response:**

M5S, 25%

**Filtering Reason:**

The image is a pie chart, but it lacks labels or a legend to identify which colors correspond to which parties. Without this information, it is impossible to determine which party has the second-largest share or the percentage of seats it holds.

Image contains an error



**User Question:**

What natural elements contribute to the sense of tranquility in this landscape, and how does the play of light enhance the overall atmosphere?

**Response:**

The natural elements that contribute to the tranquility include the lush green meadow, vibrant wildflowers, and the gently flowing stream. These elements evoke a sense of peace and harmony with nature. The play of light, with sunlight filtering through the tree branches, creates a dynamic interplay of shadows and highlights, further enhancing the serene atmosphere by adding depth and visual interest to the landscape.

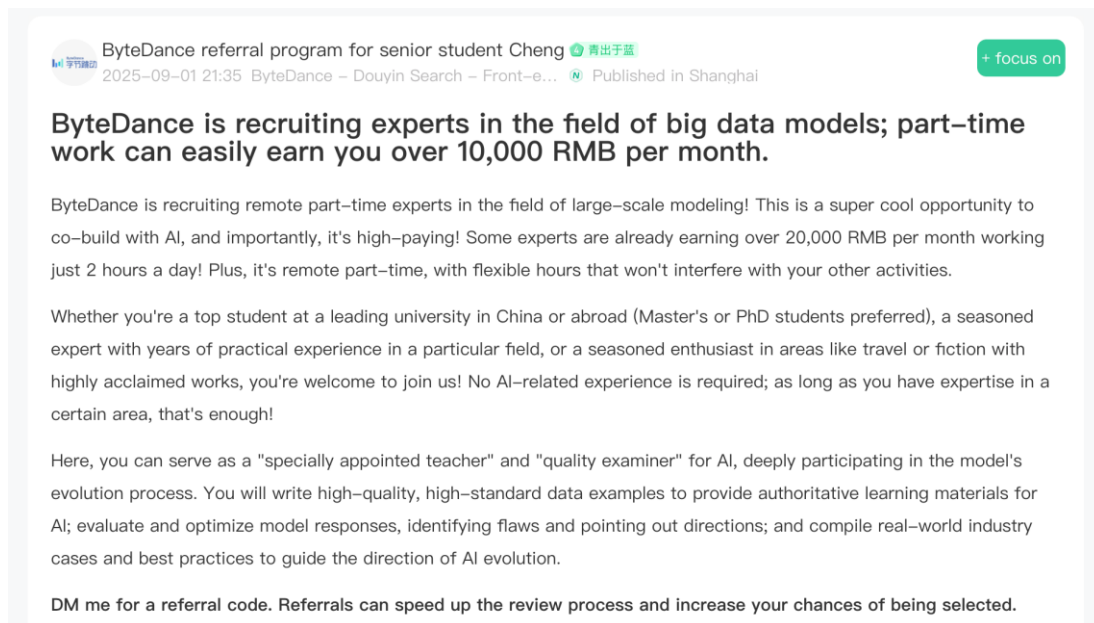
**Filtering Reason:**

The image shows a person on an escalator with a bag labeled 'bloke,' which is unrelated to natural elements or tranquility in a landscape. The question and image are not relevant to each other.

The issue is unrelated to the image.

# Motivation and Challenges

- There is a huge gap in data quality during the Training stage.
  - ◆ High-quality dataset acquisition is costly: institutions with closed-source models and semi-open models spend a large amount of money to obtain massive high-quality datasets.



ByteDance referral program for senior student Cheng 青出于蓝

2025-09-01 21:35 ByteDance - Douyin Search - Front-e... Published in Shanghai

[+ focus on](#)

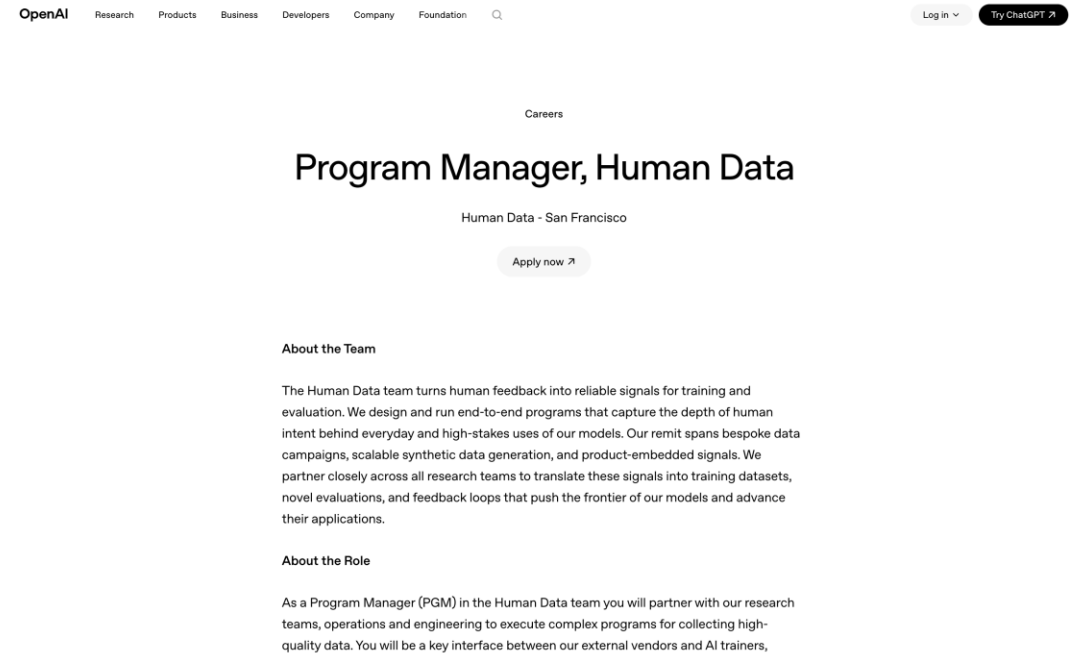
### ByteDance is recruiting experts in the field of big data models; part-time work can easily earn you over 10,000 RMB per month.

ByteDance is recruiting remote part-time experts in the field of large-scale modeling! This is a super cool opportunity to co-build with AI, and importantly, it's high-paying! Some experts are already earning over 20,000 RMB per month working just 2 hours a day! Plus, it's remote part-time, with flexible hours that won't interfere with your other activities.

Whether you're a top student at a leading university in China or abroad (Master's or PhD students preferred), a seasoned expert with years of practical experience in a particular field, or a seasoned enthusiast in areas like travel or fiction with highly acclaimed works, you're welcome to join us! No AI-related experience is required; as long as you have expertise in a certain area, that's enough!

Here, you can serve as a "specially appointed teacher" and "quality examiner" for AI, deeply participating in the model's evolution process. You will write high-quality, high-standard data examples to provide authoritative learning materials for AI; evaluate and optimize model responses, identifying flaws and pointing out directions; and compile real-world industry cases and best practices to guide the direction of AI evolution.

DM me for a referral code. Referrals can speed up the review process and increase your chances of being selected.



OpenAI Research Products Business Developers Company Foundation

Log in Try ChatGPT

Careers

## Program Manager, Human Data

Human Data - San Francisco

[Apply now ↗](#)

### About the Team

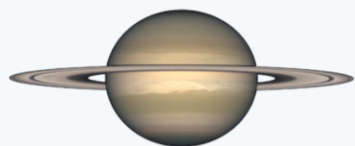
The Human Data team turns human feedback into reliable signals for training and evaluation. We design and run end-to-end programs that capture the depth of human intent behind everyday and high-stakes uses of our models. Our remit spans bespoke data campaigns, scalable synthetic data generation, and product-embedded signals. We partner closely across all research teams to translate these signals into training datasets, novel evaluations, and feedback loops that push the frontier of our models and advance their applications.

### About the Role

As a Program Manager (PGM) in the Human Data team you will partner with our research teams, operations and engineering to execute complex programs for collecting high-quality data. You will be a key interface between our external vendors and AI trainers,

# Motivation and Challenges

- There is a huge gap in data quality during the Training stage.
  - ◆ Lack of complex reasoning data: A serious lack of complex chain-of-thought (CoT) data hinders the development of the model's advanced capabilities.



## User Question:

Who discovered the moon of Saturn called Titan in 1655? Answer the question using a single word or phrase.

## Response:

Huygens

## Filtering Reason:

The image shows Saturn, but the question is about the discovery of Titan, which is not depicted in the image. The image does not provide any information relevant to answering the question.

Current training data, the answer is a single word.



Who discovered the moon of Saturn called Titan in 1655?  
Answer the question using a  
single word or phrase.

已完成思考

用户现在需要回答的问题是，谁在 1655 年发现了土星的卫星泰坦（Titan）。首先，回忆一下天文学史。克里斯蒂安·惠更斯（Christiaan Huygens）是荷兰天文学家，他在 1655 年发现了泰坦。所以应该是 Christiaan Huygens。需要确认一下，没错，惠更斯发现了土卫六泰坦。所以回答的时候用单个词或短语，就是 Christiaan Huygens。

Christiaan Huygens

Closed-source models usually think about the question  
before answering.

# Motivation and Challenges

---

- There is a huge gap in data quality during the Training stage.
  - ◆ Common data noise: Existing open-source datasets are filled with factual errors, mismatches between text and images, and format defects.
  - ◆ High-quality dataset acquisition is costly: institutions with closed-source models and semi-open models spend a large amount of money to obtain massive high-quality datasets.
  - ◆ Lack of complex reasoning data: A serious lack of complex chain-of-thought (CoT) data hinders the development of the model's advanced capabilities.

For the open-source community, focusing on data quality is the most feasible path to narrowing the gap.

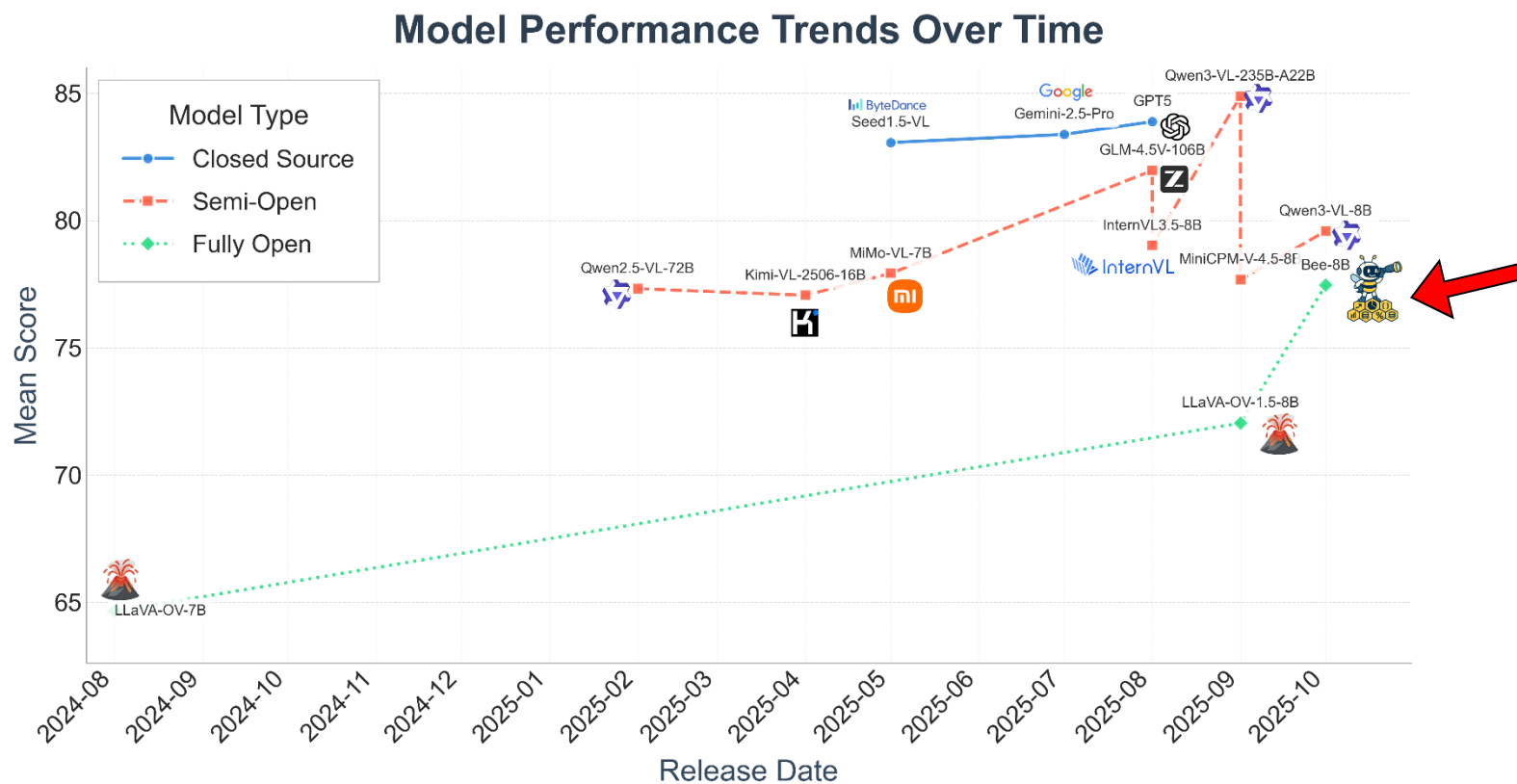
# Bee: Full-stack suit for MLLM

---

- In response to the above challenges, we have launched a full-stack suit aimed at systematically addressing data quality issues.
  - ◆ **Honey-Data-15M (Training Dataset):** A novel SFT dataset containing 15 million QA pairs. It has been meticulously cleaned in multiple stages and employs an innovative dual (long-short) CoT enhancement strategy.
  - ◆ **HoneyPipe & DataStudio (Data Pipeline):** An automated, transparent, and reproducible data pipeline for creating the dataset. Provides the community with an evolutionary data curation methodology beyond static datasets.
  - ◆ **Bee-8B (8B parameters MLLM):** An 8B model trained on Honey-Data-15M. Used to validate the effectiveness of our dataset and data pipeline.

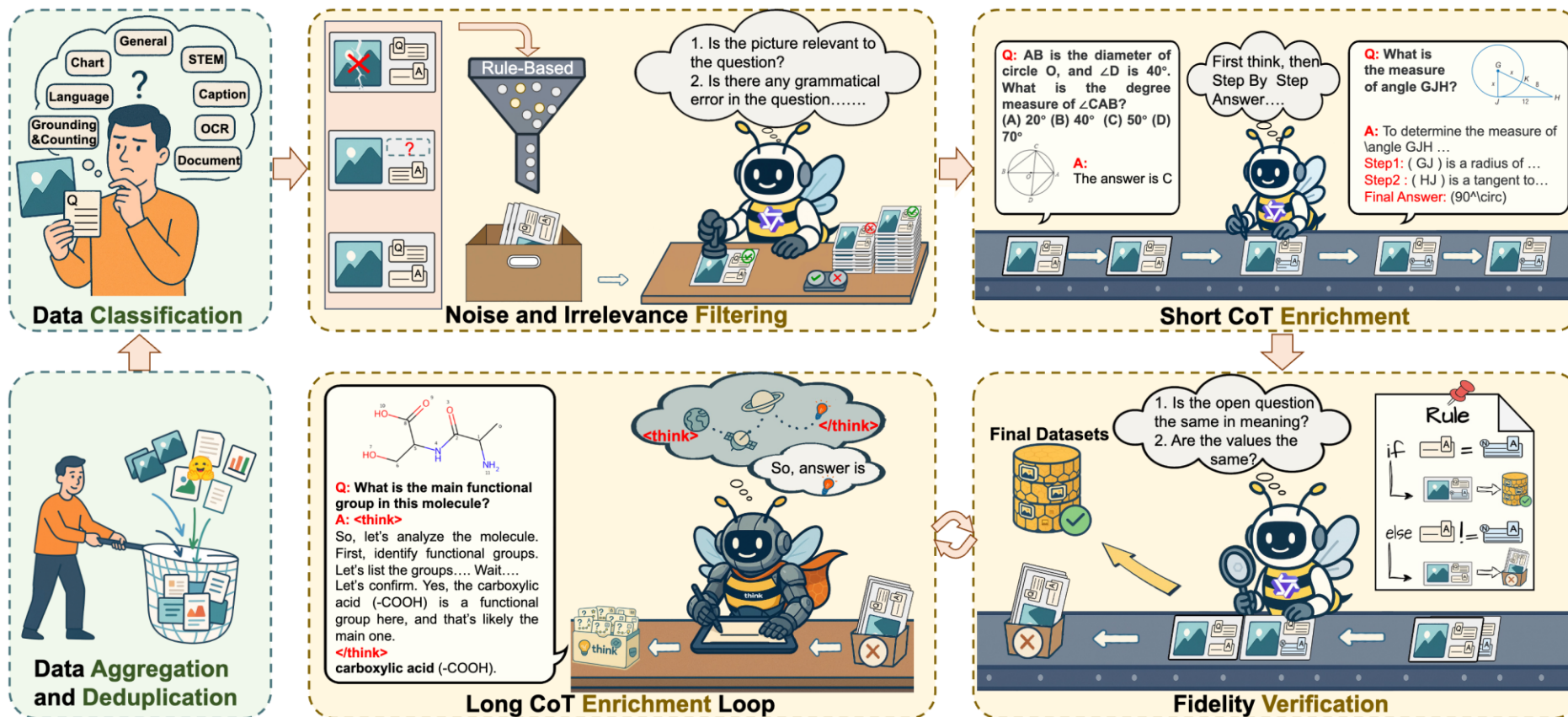
# Bee: Full-stack suit for MLLM

- Bee-8B achieves the SOTA of fully open-source MLLMs with the same parameter quantity, and also has similar performance to partially open-source MLLMs with the same parameter quantity.



# Bee: Full-stack suit for MLLM

- HoneyPipe Data Pipeline: Systematically addresses the lack of data noise and complex reasoning capabilities.

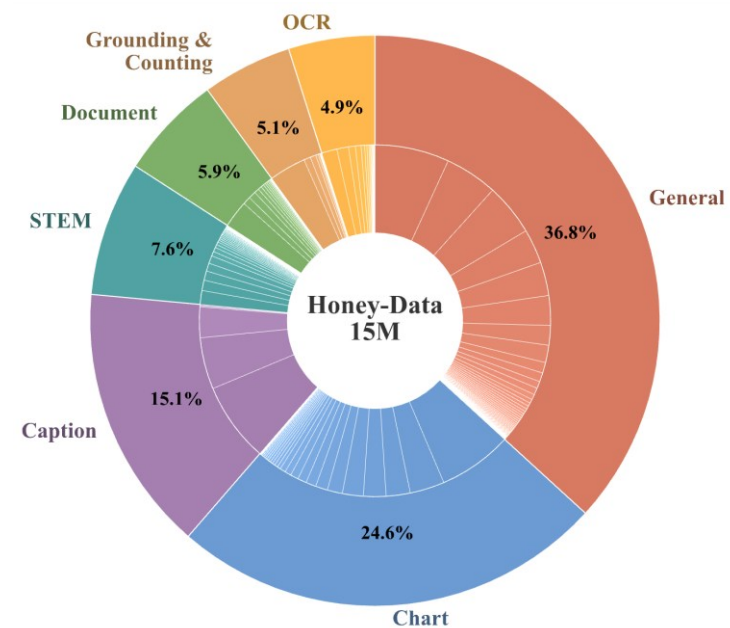


HoneyPipe's several steps, solving existing data noise and improving inference capabilities.

# Bee: Full-stack suit for MLLM

## ■ Honey-Data-15M corpus: A high-quality SFT dataset

<b>General (36.8%)</b>	SVIT-mix-665K(1019.1K)*	ALLaVA(721.1K)*	LLaVA-NeXT-Data(711.5K)	LLaVA-Instruct-300k(470.8K)*	Vision FLAN(458.8K)*
idefics375k(406.1K)*	SVIT-core-150K(269.4K)*	PixMo-CapQA(237.3K)*	LVIS-InstructV4(145.4K)*	ShareGPT4V(SAM)(122.5K)*	Cambrian (Filter)(99.6K)*
PixMo-AskModelAnything(96.3K)	ShareGPT4o(73.9K)*	NLVR2(54.1K)	GQA(45.7K)*	PixMo-Point-Explanations(44.9K)	LRV Normal(39.5K)*
HQ-Edit(36.4K)	ALFWorld(31.9K)*	Visual7W(30.7K)*	Co-Instruct(30.5K)	Cauldron(mulberry)(28.1K)*	A-OKVQA(27.0K)*
IconQA(26.5K)	VIST(21.1K)	KVQA(17.7K)	ContrastiveCaption(17.2K)	FlintstonesSV(17.0K)	InternVL-SA-1B-Caption(16.9K)
IDK(16.6K)*	COCO(16.3K)*	EST-VQA(13.5K)	Birds-to-Words(11.2K)	ART500K(11.0K)	DreamSim(11.0K)
ScanQA(10.2K)	MagicBrush(9.8K)	KonIQ-10k(9.4K)	Hateful Memes(7.3K)*	WebQA(6.6K)	nuScenes(6.5K)
Other(5.7K)*	Objects365(5.2K)	MMChat-Twitter-Post(4.3K)	NextQA(3.1K)	VSR(2.6K)*	New Yorker Caption(2.6K)
ViQuAE(2.4K)*	TQA(2.1K)	ShareGPT4V(Knowledge)(2.0K)*	WildVision(1.9K)		
<b>Chart (24.6%)</b>	TinyChart(1014.1K)*	DVQA(473.8K)*	UniChart(330.0K)	CoSyn(chart)(308.2K)*	ArxivQA(287.0K)*
FigureQA(206.8K)*	MMTab(165.6K)	PlotQA(138.7K)*	UReader QA(121.4K)*	RobuT WikiSQL(110.9K)*	CoSyn(table)(86.5K)*
CoSyn(diagram)(77.1K)*	TabMWP(46.3K)*	RobuT WTQ(44.2K)*	UReader KG(38.2K)*	Chart2Text(34.5K)*	ChartQA(34.4K)*
MMC-Instruction(27.1K)	RobuT SQA(26.6K)*	MAVIS-Function(25.3K)	CoSyn(graphic)(19.4K)*	VisText(15.6K)*	SciTSR(9.0K)
MultiHiert(4.6K)*	SimChart9K(4.5K)	Other(3.9K)*	HiTab(3.6K)*	LRV Chart(3.3K)*	Infographic(2.2K)*
<b>Caption (15.1%)</b>	COYO-Recaption(1091.8K)	PixMo-Cap(706.9K)	WIT(419.9K)*	Sherlock(16.9K)*	ST-VQA(13.5K)*
Other(1.1K)*					
<b>STEM (7.6%)</b>	VisualWebInstruct(filtered)(190.4K)*	MapQA(142.2K)*	VisualWebInstruct(127.1K)*	MetaMathQA(110.0K)	Geo170K(106.6K)*
MAVIS-Metagen(62.5K)*	GeoQA+(60.3K)*	MAVIS-Geo(43.9K)*	CoSyn(math)(37.4K)*	A12D(27.8K)*	PMC-VQA(25.1K)*
RAVEN(23.1K)*	PathVQA(21.8K)*	CoSyn(music)(21.2K)*	CoSyn(chemical)(18.2K)*	CoSyn(circuit)(17.0K)*	MathV360K(TQA)(14.2K)*
ScienceQA(12.6K)*	Geometry3K(11.4K)*	A12D(InternVL)(10.6K)*	MMChem(10.0K)	WebSight(7.6K)*	UniGeo(7.6K)*
Other(7.4K)*	GeomVerse(6.5K)*	A12D(GPT4V)(5.5K)*	VizWiz(5.3K)*	VQA-RAD(1.6K)*	CMM-Math(1.5K)
InterGPS(1.3K)*					
<b>Document (5.9%)</b>	Ureader Chart(371.0K)*	OCR-VQA(107.8K)*	CoSyn(document)(101.5K)*	ScreenQA(64.9K)*	FinTabNet(50.1K)
TextVQA(39.1K)*	EATEN(30.0K)	DocVQA(23.8K)*	LLaVAR GPT4(23.3K)*	Docmatix(19.1K)	CoSyn(nutrition)(13.0K)*
InfoVQA(9.8K)*	UreaderOCR(5.5K)*	Other(4.3K)*	DocReason(4.0K)	InfographicVQA(4.0K)*	VisualMRC(3.3K)*
POIE(2.2K)					
<b>Grounding &amp; Counting (5.1%)</b>	CLEVR(504.2K)*	TallyQA(91.6K)*	VisualGenome(77.4K)	IconQA(35.9K)*	TQA(20.6K)*
MovieNet(7.4K)	CLEVR-Math(7.2K)*	Super-CLEVR(6.8K)*	MathV360K(VQA-AS)(4.1K)*	CLEVR-Change(3.0K)	Other(1.7K)*
<b>OCR (4.9%)</b>	K12 Printing(211.1K)	ArXiv OCR(159.3K)	HME(93.6K)	VCR-Wiki(77.9K)	TextOCR(47.7K)
Sroie(31.2K)	ICDAR-LSVT-zh(28.2K)	ReCTs(17.0K)	CTW(13.6K)	Rendered Text(10.0K)	ICDAR2017(9.9K)
Chrome-Writing(9.2K)	MTWI(zh)(8.3K)	IAM(5.7K)	ICDAR2019(3.4K)	Orand-Car-A(2.0K)	IIIT 5K(2.0K)



Distribution of Honey-Data-15M

**It includes 12.1M short CoT and 2.9M long CoT data**

The data source of Honey-Data-15M covers about 200 data sources.

# Bee: Full-stack suit for MLLM

- Bee-8B: Validity verification of Honey-Data-15M.
  - ◆ Based on the Qwen3-8B (LLM) and SigLIP2-so400m-patch14-384 (Vision Encoder), an MLLM trained from scratch.
  - ◆ Adhering to common training paradigms, it is divided into five training stages: MLP warmup, SFT, GRPO, etc.

Stages	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
Purpose	MLP Warmup	Vision-Language Alignment	Multimodal SFT	Efficient Refinement SFT	Policy Optimization RL (GRPO)
Batch size	512	256	256	256	512
Learning Rate	1e-3	4e-5	5e-5	3e-5,5e-6	2e-6
Dataset Items	1M	14M	15M	1M	50K
Packed Sequence Length	8192	16384	16384	16384	-
Training Epochs	1	1	1	1	-
Trainable Components	MLP	All	All	All	All

The five-stage training process of Bee-8B

# Bee: Full-stack suit for MLLM

## ■ Experimental results of Bee-8B

Task	Benchmark	LLaVA OneVision-7B*	Molmo -7B-D*	Qwen2.5 -VL-7B <sup>†</sup>	Keye-VL -8B <sup>†</sup>	InternVL3.5 -8B <sup>†</sup>	Bee-8B -SFT*	Bee-8B -RL*
General VQA	AI2D	81.4	81.0	84.3	<b>86.7</b>	84.0	83.8	<b>85.3</b>
	BLINK <sub>val</sub>	48.2	49.7	<b>56.4</b>	52.0	<b>59.5</b>	52.5	55.0
	CountBench	–	84.8	74.1	78.0	–	<b>90.5</b>	<b>93.0</b>
	HallusionBench <sub>avg</sub>	31.6	46.4	52.9	<b>67.0</b>	54.5	<b>59.8</b>	58.2
	MMBench-CN <sub>dev</sub>	–	–	81.3	<b>92.0</b>	–	81.2	<b>84.2</b>
	MMBench-EN <sub>dev</sub>	80.8	–	82.1	<b>91.5</b>	–	83.0	<b>85.5</b>
	MMMU <sub>val</sub>	48.8	45.3	58.6	<b>71.4</b>	<b>73.4</b>	66.8	66.1
	MMMU-Pro <sub>standard</sub>	29.5	–	34.7	47.1	–	<b>50.4</b>	<b>50.7</b>
	MMStar	61.7	56.1	63.9	<b>75.5</b>	69.3	69.0	<b>71.4</b>
	MMT-Bench <sub>val</sub>	59.3	56.3	63.6	65.9	<b>66.7</b>	64.6	<b>67.0</b>
	MMVet	57.5	41.5	67.1	79.0	83.1	<b>83.3</b>	<b>83.9</b>
	MMVP	–	–	73.3	79.0	–	<b>80.7</b>	<b>82.0</b>
	POPE <sub>avg</sub>	88.4	<b>89.0</b>	86.4	86.0	<b>88.7</b>	84.0	84.8
	RealWorldQA	66.3	<b>70.7</b>	68.5	67.7	67.5	70.1	<b>73.1</b>
VisuLogic	–	–	20.0	<b>25.6</b>	–	24.4	<b>26.5</b>	
VLMs are Blind	39.2	–	37.4	<b>57.1</b>	–	55.8	<b>56.5</b>	
Table & Chart & OCR	CharXiv <sub>DQ</sub>	–	–	73.9	77.7	72.2	<b>84.7</b>	<b>84.8</b>
	CharXiv <sub>RQ</sub>	–	–	42.5	45.4	44.4	<b>55.3</b>	<b>57.3</b>
	ChartQA <sub>test</sub>	80.0	84.1	<b>87.3</b>	86.3	86.7	<b>86.7</b>	86.1
	DocVQA <sub>val</sub>	–	–	<b>95.5</b>	<b>88.5</b>	–	87.2	87.0
	InfoVQA <sub>val</sub>	–	–	<b>81.4</b>	67.4	–	72.3	<b>72.9</b>
	OCRBench	62.2	65.6	<b>86.4</b>	<b>85.1</b>	84.0	83.1	82.5
	SEED-Bench2-Plus	65.4	67.6	<b>70.4</b>	69.4	<b>70.8</b>	67.7	68.5
Math & Reasoning	DynaMath <sub>worst</sub>	9.0	–	21.0	37.3	37.7	<b>41.3</b>	<b>40.5</b>
	LogicVista	33.3	–	44.1	54.8	<b>57.3</b>	56.8	<b>61.3</b>
	MathVerse <sub>vision_only</sub>	26.2	4.2	25.1	59.8	61.5	<b>61.9</b>	<b>67.0</b>
	MathVision	18.3	16.2	25.4	46.0	<b>56.8</b>	46.8	<b>50.0</b>
	MathVista <sub>mini</sub>	63.2	51.6	68.2	<b>80.7</b>	78.4	78.6	<b>81.4</b>
	WeMath	20.9	–	35.2	<b>60.7</b>	57.0	55.0	<b>59.8</b>

# Bee: Full-stack suit for MLLM

---

- Experimental results of Bee-8B

- ◆ Refresh the SOTA of Fully open MLLM

- ◆ Comparable to top-tier semi-open source models

- ◆ Achieved impressive results on different tasks:

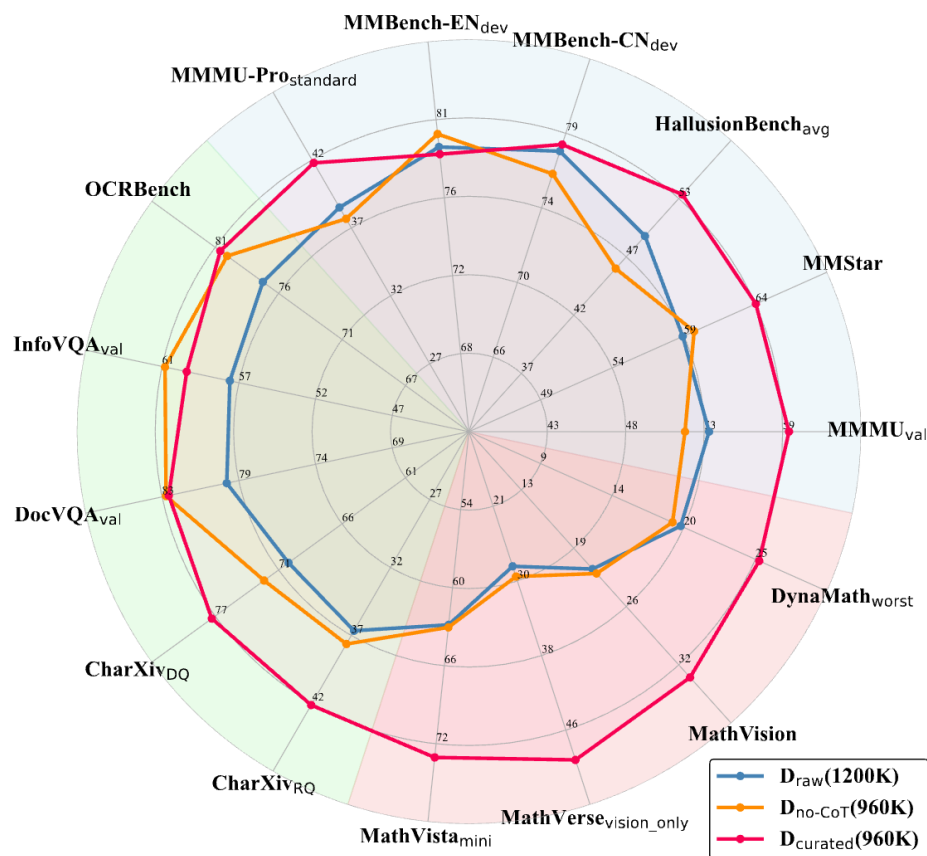
- Mathematics and Reasoning: Ranked first on MathVerse (67.0), significantly ahead of InternVL3.5-8B (+5.5%). Also ranked first on LogicVista (61.3)

- Chart and document: Highest score on CharXiv (RQ), leading the second place by nearly 12%.

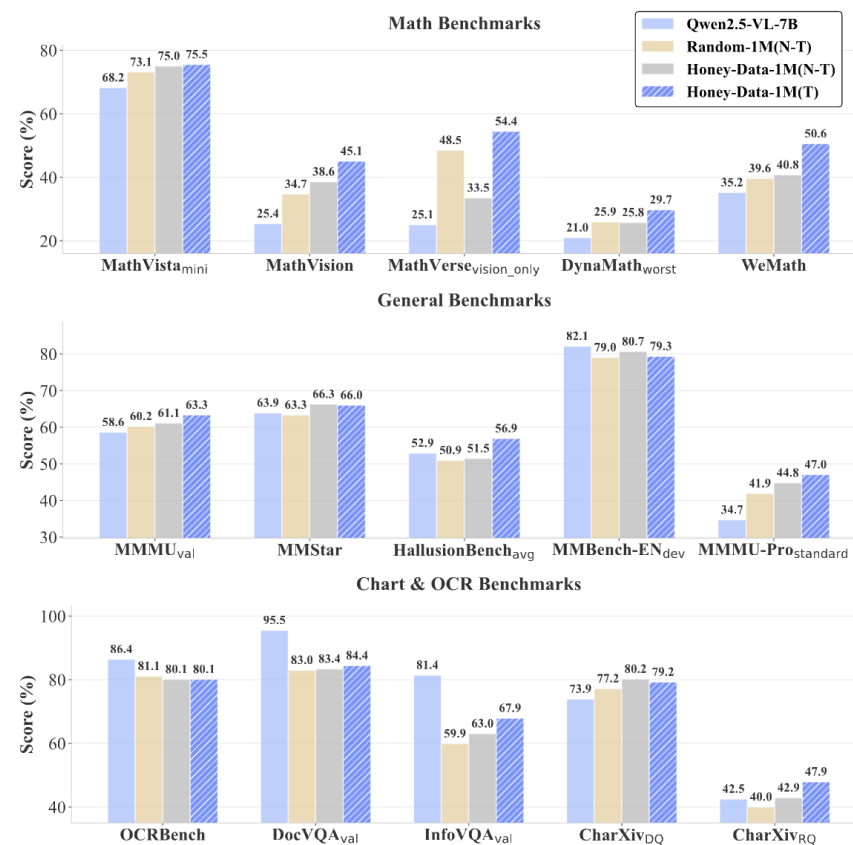
- Top scores achieved on benchmarks such as MMMU-Pro (50.7) and RealWorldQA (73.1)

# Bee: Full-stack suit for MLLM

- Ablation Experiment: Demonstrates the effectiveness of each step of HoneyPipe, and the data subset can achieve good performance.



HoneyPipe's Step Ablation Comparison



Ablation on the subset, some results exceeded Qwen2.5-VL-7B 15

# Bee: Full-stack suit for MLLM

## ■ Community Impact

- ◆ Community influence is significant: The project's open-source ecosystem construction has yielded remarkable results, with the core dataset downloaded over 230,000 times on HuggingFace, and model downloads exceeding 180,000 times; among which, the dataset once ranked in the top five of the HuggingFace dataset trends list.
- ◆ The Bee-8B model has been successfully selected into HuggingFace's official "China Open Source Highlights of October 2025".

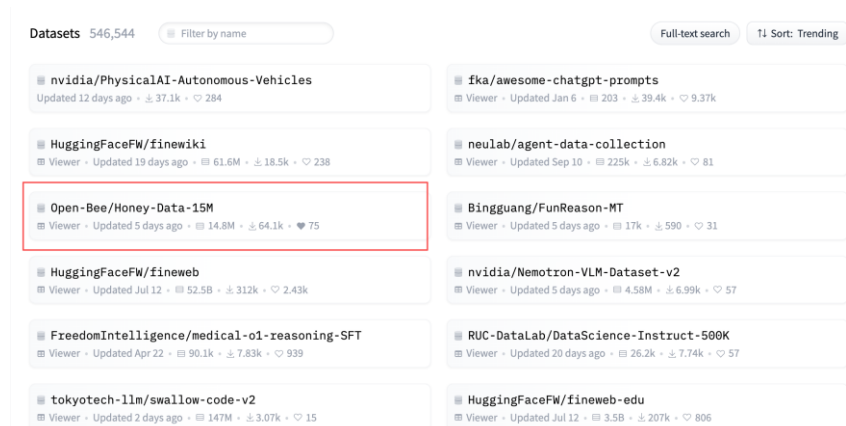


Chinese open source AI in October wasn't about bigger models, it was about real world impact 🧨

[huggingface.co/collections/zh...](https://huggingface.co/collections/zh...)

🌟 Vision-Language & OCR wave 🇨🇳

- DeepSeek-OCR : 3B
- PaddleOCR-VL : 0.9B
- Qwen3-VL : 2B / 4B / 8B / 32B / 30B-A3B
- Open-Bee: Bee-8B-RL
- Z.ai Glyph :10B



# Bee: Full-stack suit for MLLM

- Identified the core obstacles hindering the development of fully open-source MLLM—SFT data quality (noise and CoT absence), and proposed a full-stack solution.

 **Bee: A High-Quality Corpus and Full-Stack Suite to Unlock Advanced Fully Open MLLMs**

ICLR 2026

Yi Zhang<sup>1,2</sup>, Bolin Ni<sup>2</sup>, Xin-Sheng Chen<sup>1</sup>, Heng-Rui Zhang<sup>1</sup>, Yongming Rao<sup>2</sup>, Houwen Peng<sup>2\*</sup>, Qinglin Lu<sup>2</sup>,

Han Hu<sup>2</sup>, Meng-Hao Guo<sup>1†</sup>, Shi-Min Hu<sup>1</sup>

<sup>1</sup>Tsinghua University, <sup>2</sup>Tencent Hunyuan Team

\*Project lead. †Corresponding author.

[Paper](#)

[arXiv](#)

[Models & Datasets\(HuggingFace\)](#)

[Models & Datasets\(ModelScope\)](#)

[Code \(Coming Soon\)](#)

For the open-source community, prioritizing data quality and CoT reasoning is a more effective strategy than simply pursuing the quantity of data.

# Conclusion & Key Takeaways

---

- **Data is the New Bottleneck:** We demonstrated that unlocking complex reasoning in MLLMs relies fundamentally on *data quality*, not just model scaling.
- **Our Contributions**
  - ◆ **Honey-Data-15M:** A dataset with 15M QA pairs, systematically cleaned of noise and enriched with a dual-level CoT reasoning, serving as a new cornerstone for the open community.
  - ◆ **HoneyPipe:** The data curation pipeline and its underlying framework, DataStudio, offering the community a transparent and adaptable methodology that moves beyond static dataset releases.
  - ◆ **Bee-8B:** An 8B model trained on Honey-Data-15M that achieves SOTA among fully open models and competes with semi-open counterparts, validating data quality and pipeline effectiveness.