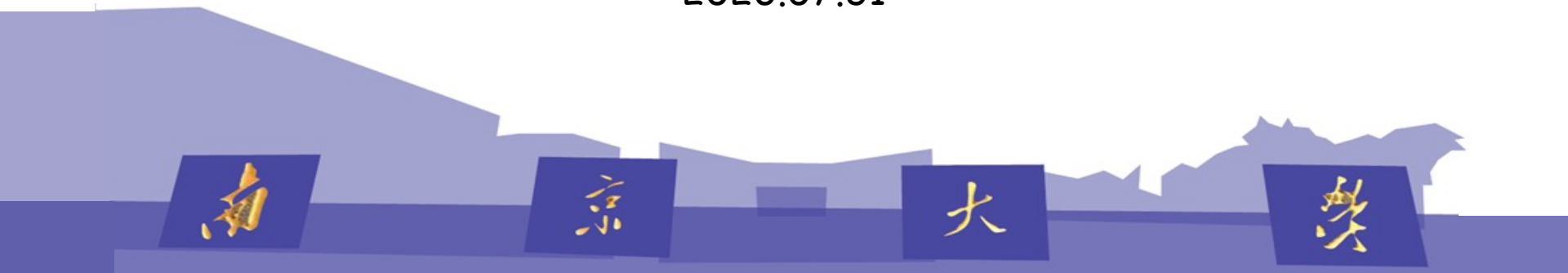
LAMDA
Learning And Mining from DatA

Constructive Specification for Plug-and-Play Learnware Agents

Presented by Zi-Chen Zhao

2026.07.31



Constructive Specification for Plug-and-Play Learnware Agents

Agentic LLM
Not general agents

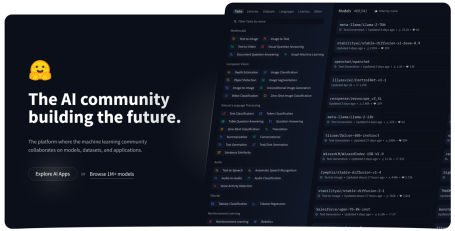
Presented by Zi-Chen Zhao

2026.07.31

Background

LAMDA
Learning And Mining from DatA

Growing Model/Agent Ecosystem



Hugging Face: 2.9M+ models

HERMES-AGENT

OPENCLAW

Agent systems surge

One model is not enough

01 Complex Software Engineering

Requirements → Code Generation → Test Execution → Web Retrieval → Environment Interaction

Orchestrator coordinates specialized capabilities

02 Complex Long-Horizon Workflows

Planning → Tool Calls → Information Retrieval → Execution → Feedback

A main agent manages the multi-stage loop

03 Vertical-Domain Collaboration

Specialized Small Models ↔ General LLM Coordinator ↔ Domain Tools & Knowledge

Domain expertise + general coordination

Semantic description is insufficient

Capabilities Extend Beyond Simple Question Answering

Core Reasoning · Tool Use · Workflow Planning · Multi-turn Interaction · Complex Capabilities

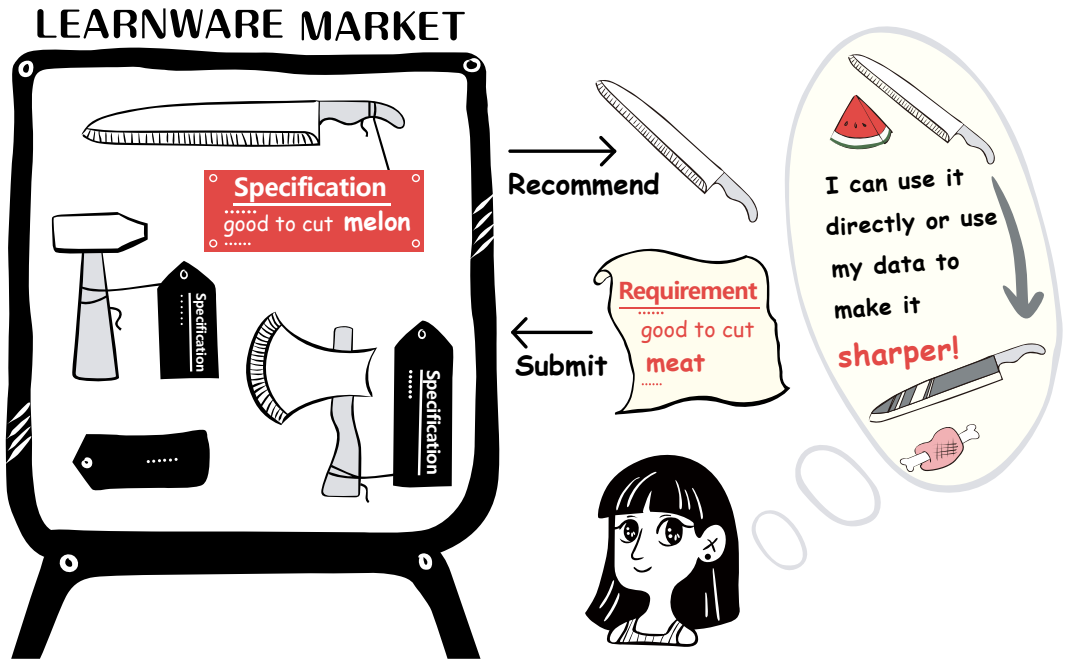
Open Agent Ecosystem

Dynamic Evolution · Heterogeneous Resources · Multidimensional Capabilities

Natural language alone is insufficient
Selecting the right model or agent for a task remains difficult

Key Problem: How to characterize the capabilities of agents?
How to enable large-scale collaboration among massive agents?

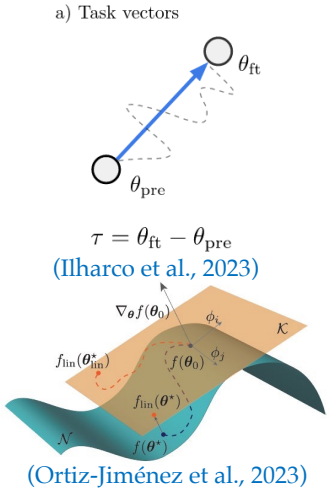
Learnware Paradigm

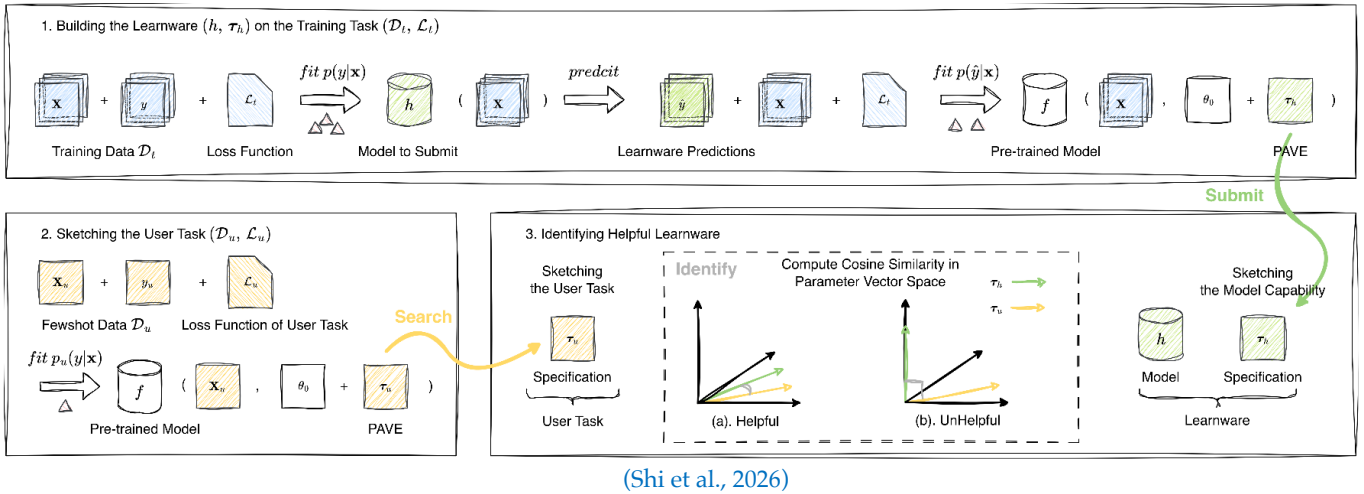


Specification makes agents plug-and-play, enabling **growable** learnware agent system

Preliminary: PAVE specification

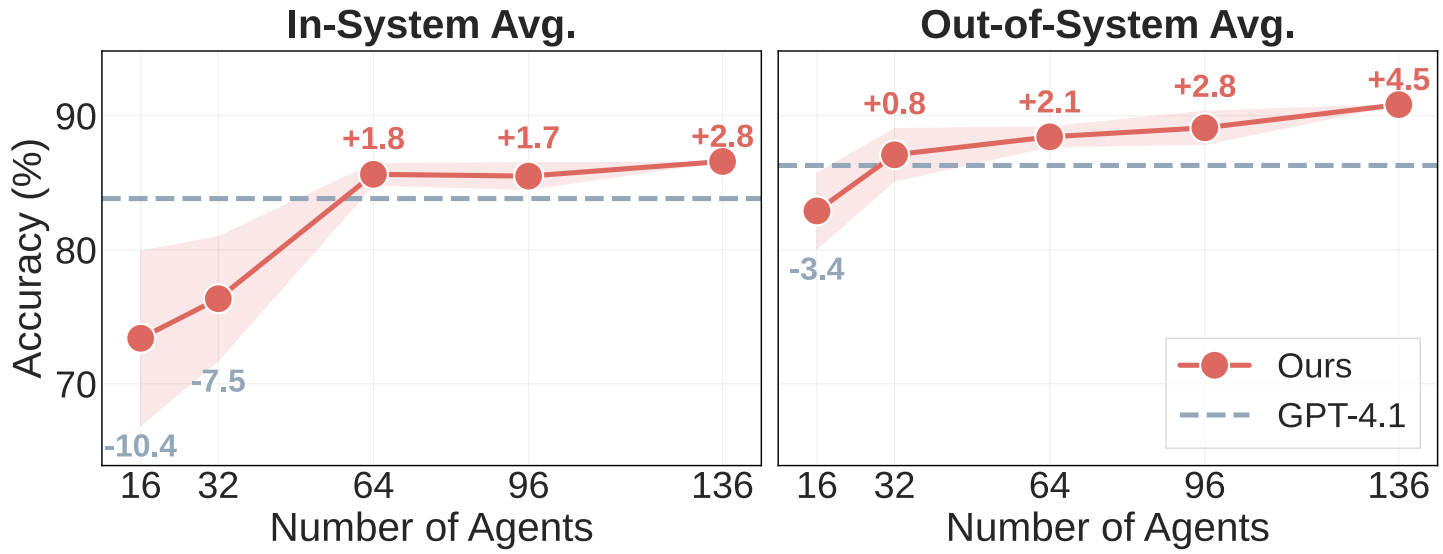
- PAVE: The changes in a **system-provided** pre-trained model f's parameters during fine-tuning
  - Represents model capabilities by the change in the parameter of f when fitting $p(\hat{\boldsymbol{y}}|x)$
  - Represents user requirements by the change in the parameter of f when fitting $p(\boldsymbol{y}|x)$
  - Identification is performed via cosine similarity between the learnware and user PAVEs





(Shi et al., 2026)

Results Highlight



- Selecting among **lightweight agents** outperforms **larger general model**
- Model number **scaling law** exploration: Performance **scales** with **agent count**

Problem setting

LAMDA
Learning And Mining from DatA

Problem Setting: Agent Routing

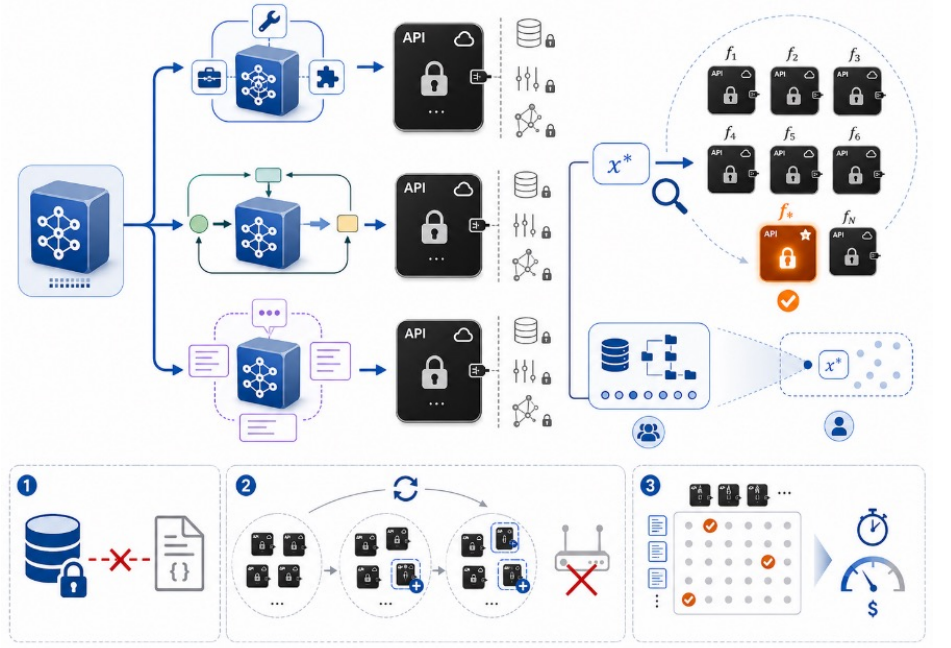
Select the most suitable agent f^* from the system for each user query x^*

Scenario Shift

LLMs become tool/workflow agents exposed via black-box APIs

Technical Challenges

- Black-box access; no training data of agents
- Evolving agent pool need plug-and-play specification
- Limited evaluation budget

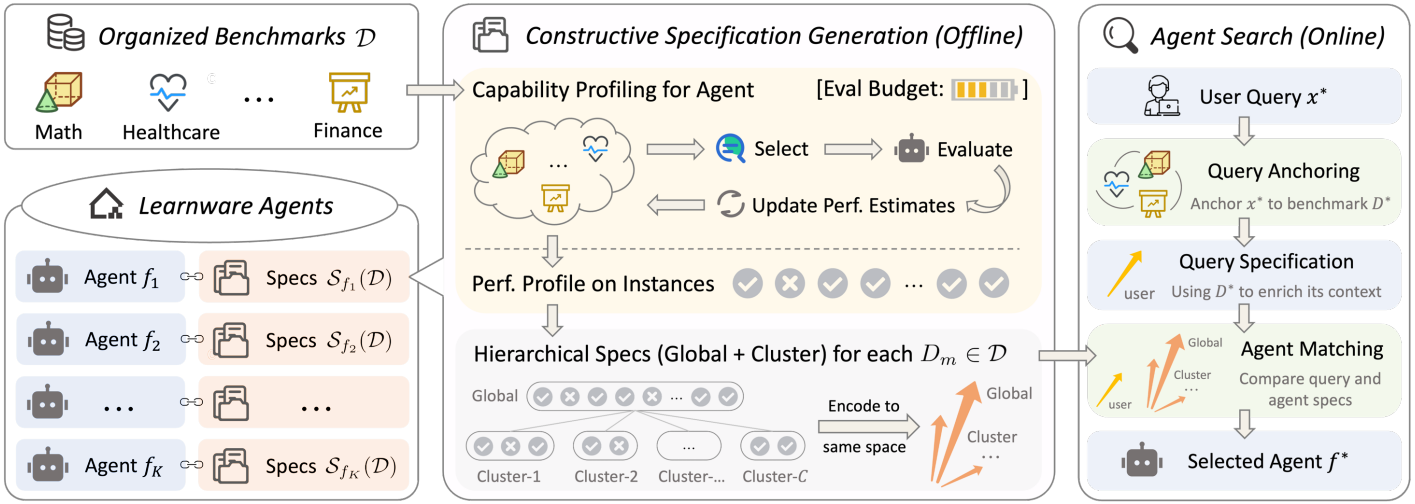


No training data **Evolving agent pool** **Limited evaluation budget**

Method

Constructive specification: represent agent capabilities from demonstrated performance on system-maintained benchmarks for plug-and-play agent search

- Offline: Generate constructive specifications from system-maintained benchmarks
- Online: Match query and agent specifications in a shared space for plug-and-play identification

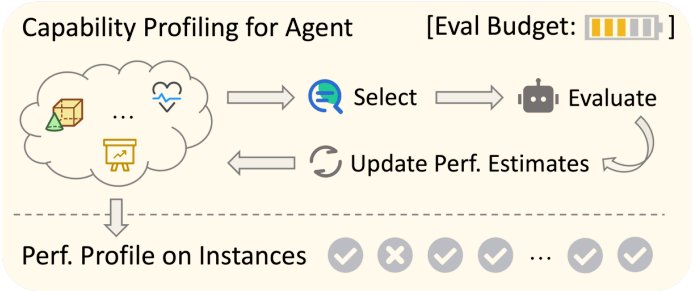


Offline: Specification generation

Offline: discover **agent strength samples** under a **limited evaluation budget** and encode them as **hierarchical parameter-vector specifications in a unified space**

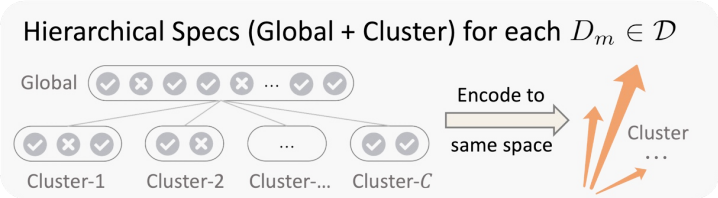
Budget-Efficient Profiling

- Target: find the subset of samples $\mathcal{P}_f(D)$ that agent f handles well on benchmark D.
- Data Preprocess: only keep hard samples
- Active profiling: test promising samples within budget



Hierarchical PAVE Specification

- PAVE: encode capability via parameter changes induced by strength samples
- Weighting: emphasize hard samples to distinguish agents
- Hierarchy: combine global and cluster-level specifications for robust, fine-grained characterization



Online: Agent search

Online: map the user query into the agents' specification space and identify the best agent through multi-granularity similarity matching

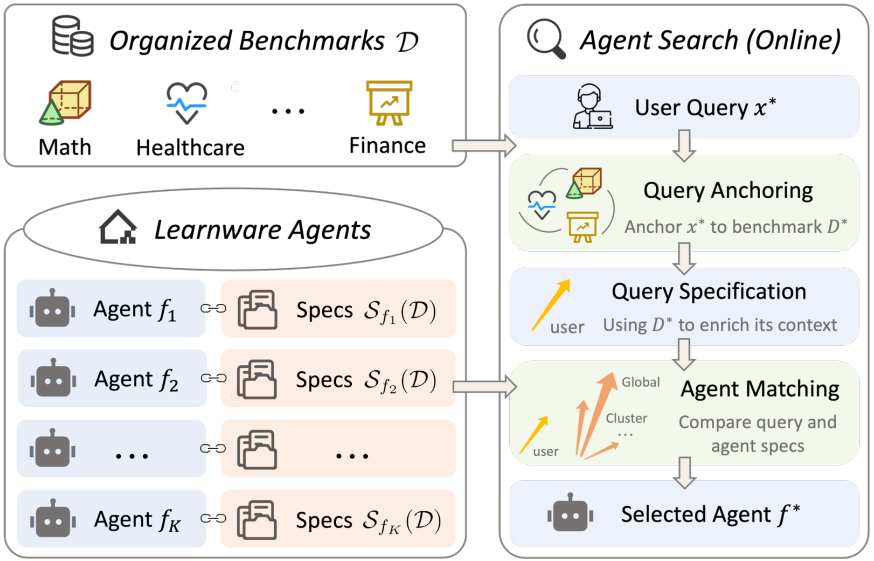
Agent Search Workflow

1. Benchmark Anchoring
Find the benchmark D^* most relevant to query x^* using embedding similarity.

2. Query Specification
Retrieve a set of local neighboring samples from the anchored benchmark D^* to enrich the query with task-specific context and difficulty cues.

3. Multi-Granularity Matching
Match the query and hierarchical agent specifications by aggregating the top-k cosine similarity scores.

4. Score Fusion and Ranking
Fuse similarity and hard-sample capability scores



Performance

LAMDA
Learning And Mining from DatA

Best on in-system and out-of-system benchmarks

- System: 136 lightweight agents (40 APIs + 96 tool-augmented agents)
- Domains: mathematics, healthcare, finance
- No global selector training; our specification-based method beats semantic matching, nearest-neighbor search, and learned routing

Benchmark	GPT-4.1	Random	DescSim	DescLLM	Model-SAT	K-NN	EmbedLLM	Ours
MMLU (Mathematics, Avg)	94.23	83.58	82.56	91.12	90.47	90.34	97.33	**98.01**
- abstract_algebra	91.00	78.00	69.00	91.00	89.00	88.00	96.00	**97.00**
- college_mathematics	91.00	77.00	81.00	91.00	84.00	89.00	**97.00**	**97.00**
- elementary_mathematics	97.88	93.39	93.92	92.86	96.30	94.71	98.15	**98.41**
- high_school_mathematics	97.04	85.93	86.30	89.63	92.59	89.63	98.15	**99.63**
MMLU (Healthcare, Avg)	93.22	85.99	88.87	92.73	89.48	90.51	93.17	**93.69**
- anatomy	88.15	77.78	80.00	88.89	83.70	84.44	88.89	**91.85**
- clinical_knowledge	**90.94**	83.77	86.79	90.94	87.17	88.68	**90.94**	90.19
- college_biology	**98.61**	90.28	93.75	97.22	96.53	95.14	**98.61**	**98.61**
- college_medicine	87.28	83.24	87.86	86.71	82.08	**88.44**	**88.44**	87.28
- medical_genetics	98.00	93.00	94.00	97.00	94.00	93.00	98.00	**99.00**
- professional_medicine	**96.32**	87.87	90.81	95.59	93.38	93.38	94.12	95.22
Finance (Avg)	49.60	45.00	40.50	38.45	54.15	45.55	46.75	**67.90**
- MA	63.20	67.00	71.00	59.40	63.80	63.60	59.00	**83.80**
- German	36.00	23.00	10.00	17.50	44.50	27.50	34.50	**52.00**
Overall Avg. (↑)	86.29	78.36	78.70	83.15	83.92	82.96	86.82	**90.83**
Average Rank (↓)	2.80	7.07	6.40	4.73	4.87	5.07	2.60	**1.47**

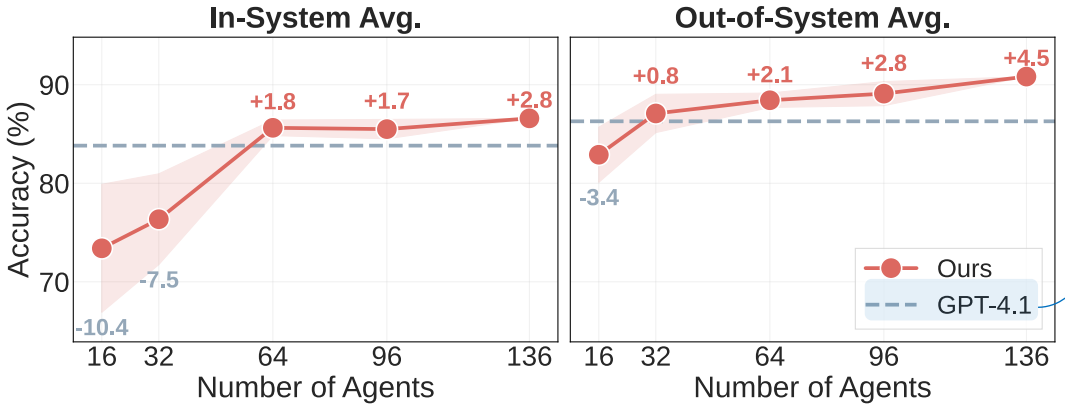
Benchmark	GPT-4.1	Random	DescSim	DescLLM	Model-SAT	K-NN	EmbedLLM	Ours
GSM8K	94.30	78.40	74.40	72.90	94.60	92.60	96.30	**96.50**
MathQA	90.50	75.50	79.00	78.90	90.00	82.40	88.30	**91.20**
MATH	88.70	74.70	69.70	70.80	90.30	87.60	95.40	**96.20**
MedMCQA	78.80	65.30	68.20	75.20	75.00	71.50	77.40	**79.80**
PubMedQA	69.60	58.00	60.80	69.40	66.80	69.80	72.40	**74.80**
FPB	**81.03**	72.78	70.31	77.73	74.64	78.87	80.52	**81.03**
Overall Avg. (↑)	83.82	70.78	70.40	74.16	81.89	80.46	85.05	**86.59**
Average Rank (↓)	2.83	7.17	7.17	6.00	4.33	4.67	2.67	**1.00**

Scalability Analysis

LAMDA
Learning And Mining from DatA

Overall performance keeps improving as the agent number grows

- Gains taper later because of agent redundancy



Selecting among lightweight agents can outperform the larger GPT-4.1 model

Summary

- Goal: Build a growable agent learnware system that aggregates capabilities of massive agents
- Result: Preliminary exploration of model number scaling law
- Method: Design constructive specifications for black-box agents to enable plug-and-play agent identification

References

- Jian-Dong Liu, Zi-Chen Zhao, Hao Sun, Lin-Xing Wu, Huan Zhang, Pengyuan Wang, Ming Zhao, Xinyu Chu, Shu Yan, Yongbei Zhu, Weijun Zhong, Zhi-Hao Tan, Jing Shang, Yang Yu, and Zhi-Hua Zhou. Constructive Specification for Plug-and-Play Learnware Agents. In: Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD' 26), Jeju, Korea, 2026.

Thanks!