# 大模型的工具协同推理能力的系统性构建与潜能激发

曾幸山

华为诺亚方舟实验室





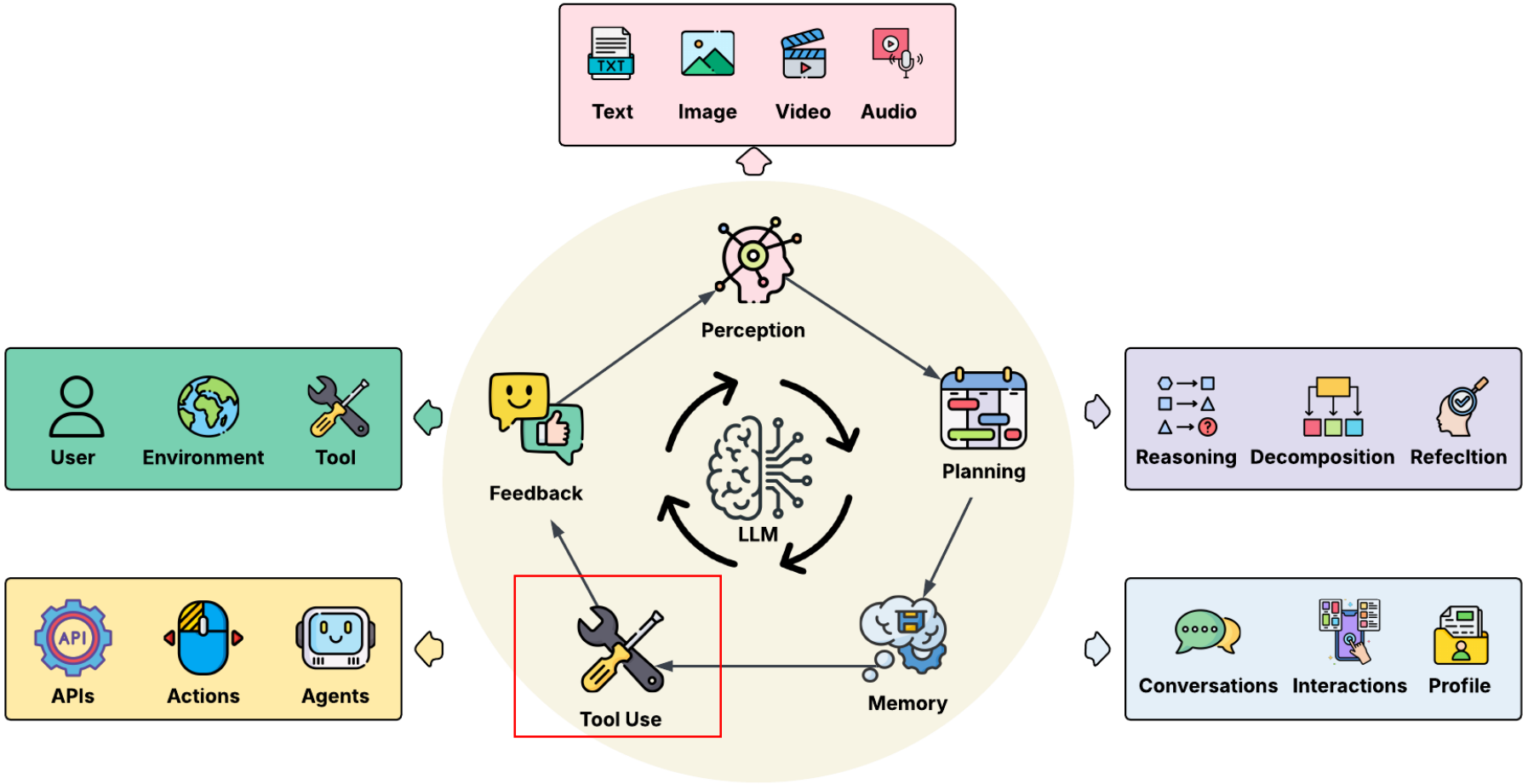


基础研究与应用创新并重

使能万物互联的智能世界

# LLM-based Agent

- 工具协同能力是Agent的关键能力之一，让模型从“**能说**”到“**会做**”

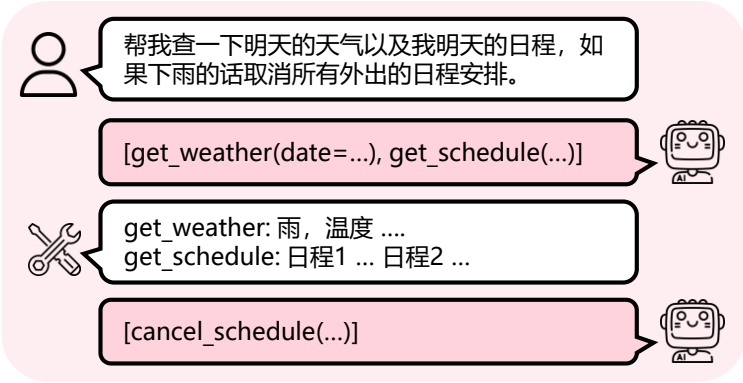


# 构建工具协同能力的关键挑战

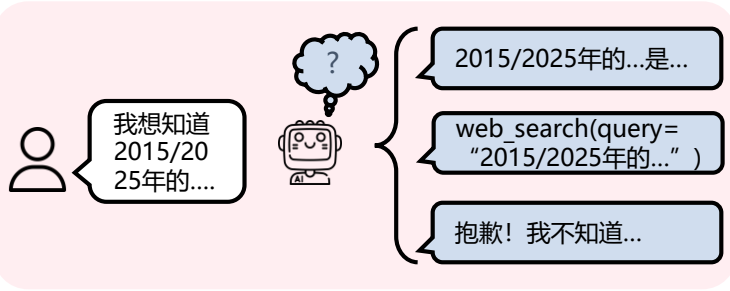
- 可泛化性：
  - 真实工具可能**频繁迭代更新**
  - LLM迭代成本高
- 数据质量保证：
  - **格式**正确性
  - **推理**正确性（工具选择、参数提取）
- 复杂需求（多步）：
  - **多工具协同**（并行调用、前后依赖）
  - 复杂多步需求的**生成与模拟**
- 语言回复和工具调用之间的协调（多轮）：
  - **工具 awareness**（是否需使用外部工具）
  - 工具调用能力与**通用能力**的平衡

| Model | #API | #Domain |
|---|---|---|
| Gorilla [14] | 1645 | 3 |
| ToolAlpaca [19] | 3938 | 50 |
| ToolLLM [15] | 16464 | 49 |
| xLAM [11] | 3673 | 21 |

当前主流工具增强模型训练数据情况




复杂多步需求例子




工具 awareness

# 构建工具协同能力的关键挑战

- 可泛化性：
  - 真实工具可能**频繁迭代更新**
  - LLM迭代成本高
- 数据质量保证：
  - **格式**正确性
  - **推理**正确性（工具选择、参数提取）
- 复杂需求（多步）：
  - **多工具协同**（并行调用、前后依赖）
  - 复杂多步需求的**生成与模拟**
- 语言回复和工具调用之间的协调（多轮）：
  - **工具 awareness**（是否需使用外部工具）
  - 工具调用能力与**通用能力**的平衡

→ 多样性

→ 正确性

→ 复杂性

ToolACE (ICLR2025)

↓ 潜能激发

ToolACE-R (2025)

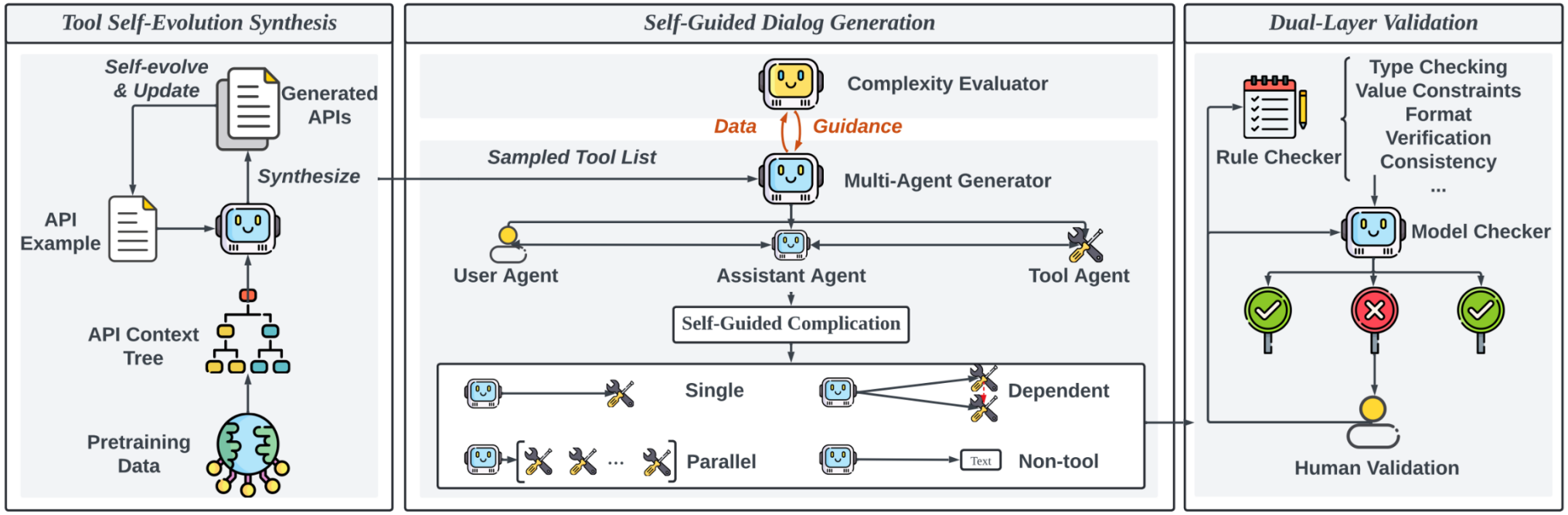
→ 工具协同

→ 多轮平衡

ToolFlow (NAACL2025)

# ToolACE: 大模型工具协同能力系统性构建方法

- 工具能力的学习需要：**Accurate, Complex, and divErse**




**工具合成**

- 构建API语义树
- 工具定义的迭代进化与更新

**对话生成**

- 多agent交互生成为基础
- 自我为引导的复杂度衡量、多样对话生成

**两层数据质检**

- 规则质检
- 模型质检

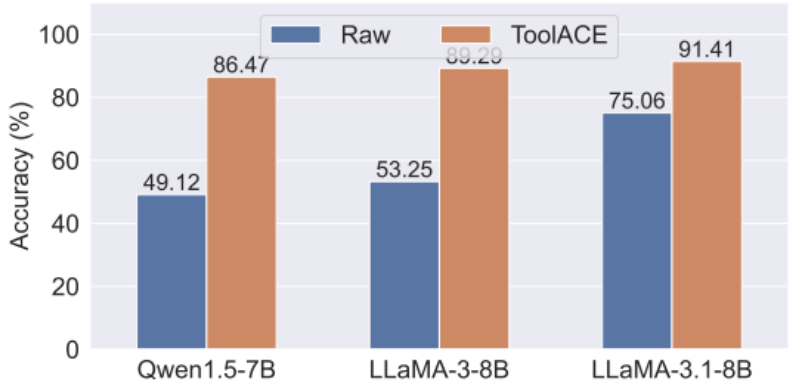


Weiwen Liu*, Xu Huang*, Xingshan Zeng*, et al. "ToolACE: Winning the Points of LLM Function Calling." *ICLR 2025.*

# ToolACE: 大模型工具协同能力系统性构建方法

| Model | #API | #Domain | Nested | Parallel | Dependent | Multi-type |
|---|---|---|---|---|---|---|
| Gorilla [14] | 1645 | 3 | ✗ | ✗ | ✗ | ✗ |
| ToolAlpaca [19] | 3938 | 50 | ✗ | ✗ | ✗ | ✗ |
| ToolLLM [15] | 16464 | 49 | ✗ | ✗ | ✓ | ✗ |
| Functionary [12] | n/a | n/a | ✗ | ✓ | ✗ | ✗ |
| xLAM [11] | 3673 | 21 | ✗ | ✓ | ✗ | ✗ |
| Granite [1] | n/a | n/a | ✗ | ✓ | ✗ | ✓ |
| **ToolACE** | **26507** | **390** | ✓ | ✓ | ✓ | ✓ |

数据统计对比




不同base模型结果

| Rank | Overall | Model | Single turn Non-live (A) | Single turn Non-live (E) | Live (A) | Multi turn Multi turn | Hallucination Rel | Hallucination Irrel |
|---|---|---|---|---|---|---|---|---|
| 1 | 59.49 | GPT-4-turbo-2024-04-09 (FC) | 82.65 | 83.80 | 73.39 | 21.62 | 70.73 | 79.79 |
| 2 | 59.29 | GPT-4o-2024-08-06 (FC) | 85.52 | 82.96 | 71.79 | 21.25 | 63.41 | 82.91 |
| 3 | 59.22 | ToolACE-8B (FC) | 89.27 | 90.07 | 73.21 | 14.37 | 85.37 | 83.81 |
| 4 | 59.13 | xLAM-8x22b-r (FC) | 89.75 | 89.32 | 72.81 | 15.62 | 97.56 | 75.23 |
| 5 | 58.45 | GPT-4o-mini-2024-07-18 (FC) | 82.83 | 81.80 | 67.53 | 25.75 | 82.93 | 71.83 |
| 6 | 57.94 | xLAM-8x7b-r (FC) | 88.44 | 85.89 | 71.97 | 15.75 | 92.68 | 72.35 |
| 7 | 57.21 | GPT-4o-mini-2024-07-18 (Prompt) | 86.54 | 87.95 | 72.77 | 11.62 | 80.49 | 79.20 |
| 8 | 55.82 | mistral-large-2407 (FC) | 84.12 | 83.09 | 67.17 | 20.50 | 78.05 | 48.93 |
| 9 | 55.67 | GPT-4-turbo-2024-04-09 (Prompt) | 91.31 | 88.12 | 67.97 | 10.62 | 82.93 | 61.82 |
| 10 | 54.83 | Claude-3.5-Sonnet-20240620 (FC) | 70.35 | 66.34 | 71.39 | 23.50 | 63.41 | 75.91 |

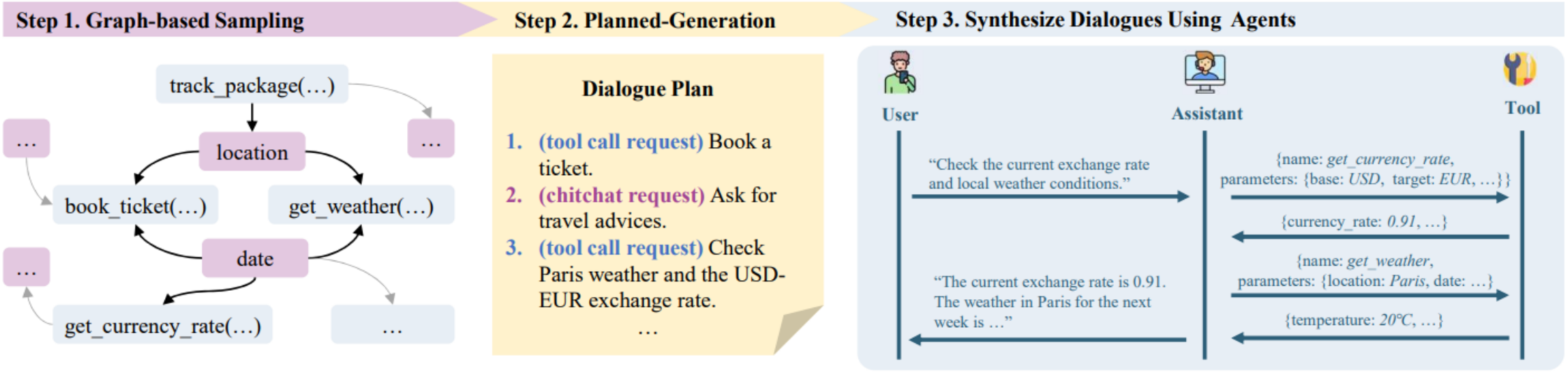
BFCL-v3榜单结果（09/20/2024）



Weiwen Liu*, Xu Huang*, Xingshan Zeng*, et al. "ToolACE: Winning the Points of LLM Function Calling." *ICLR 2025.*

# ToolFlow: 自然和连贯的工具对话合成方案

- 工具学习中的两个问题：
  - 如何生成复杂的多步需求？ -- **工具图**的构建与基于工具图的**采样**
  - 如何生成自然连贯的多轮对话？ -- 多能力融合的对话**规划**




**工具图**

- 构建：基于参数名/返回值相似性
- 采样：基于图的random walk

**对话规划**

- 多能力融合（工具调用、闲聊、其他NLP任务等）
- 自然连贯的需求切换



Zezhong Wang, Xingshan Zeng#, Weiwen Liu, et al. "ToolFlow: Boosting LLM Tool-Calling Through Natural and Coherent Dialogue Synthesis." *NAACL 2025.*

# ToolFlow: 自然和连贯的工具对话合成方案

- 实验设置：有/无Graph，有/无Plan，各8k
- 数据分析：**Graph->多样性，Plan->连贯性**
- 实验结果：
  - 工具能力大幅提升
  - 通用能力不降（多轮提升）

| Automatic Evaluation | | | | | |
|---|---|---|---|---|---|
| *Setting* | | *Coherence* | | *Diversity* | |
| *Graph* | *Plan* | *SS* | *EnR* | *H* | *D-3* |
| ✓ | ✓ | 63.36 | 47.3 | 10.36 | 0.4865 |
| ✓ | ✗ | 62.03 | 32.1 | 10.14 | 0.4364 |
| ✗ | ✓ | 61.72 | 48.1 | 9.82 | 0.3393 |
| ✗ | ✗ | 58.65 | 35.4 | 9.75 | 0.3078 |
| **GPT-4 Evaluation** | | | | | |
| *Graph* | *Plan* | *NAT* | *COH* | *HELP* | *ACC* |
| ✓ | ✓ | 3.72 | 3.91 | 4.71 | 4.92 |
| ✓ | ✗ | 3.00 | 3.72 | 4.51 | 4.84 |
| ✗ | ✓ | 3.51 | 3.88 | 4.39 | 4.87 |
| ✗ | ✗ | 2.93 | 3.66 | 4.18 | 4.90 |

合成数据质量分析

| Baselines | | | Overall | | |
|---|---|---|---|---|---|
| | | | *Non-Live* | *Live* | *Overall* |
| **Claude-3.5-Sonnet** | | | **91.22** | 80.75 | 85.98 |
| **GPT-4-turbo-0409** | | | 90.28 | **85.02** | **87.65** |
| **GPT-4o-0513** | | | 85.02 | 78.20 | 81.61 |
| **LLaMA-3.1 8B** | | | 83.40 | 72.88 | 78.14 |
| | **Graph** | **Plan** | | | |
| **Ours** | ✓ | ✓ | 90.30 | 74.83 | 82.57 |
| | ✓ | ✗ | 89.60 | 73.71 | 81.66 |
| | ✗ | ✓ | 89.50 | 73.78 | 81.64 |
| | ✗ | ✗ | 89.00 | 73.18 | 81.09 |

BFCL-v2榜单结果
(08/16/2024)

| Setting | | MTBench | | | MMLU | BBH |
|---|---|---|---|---|---|---|
| **Graph** | **Plan** | *Turn 1* | *Turn 2* | *Avg.* | | |
| - | - | 7.84 | 6.85 | 7.34 | 69.3% | 63.2% |
| ✓ | ✓ | **8.08** | **7.16** | **7.62** | **69.8%** | 63.3% |
| ✓ | ✗ | 7.39 | 6.43 | 6.91 | 69.3% | **63.5%** |
| ✗ | ✓ | 7.59 | 7.04 | 7.31 | 69.1% | 63.2% |
| ✗ | ✗ | 7.16 | 6.85 | 7.01 | 68.9% | **63.5%** |

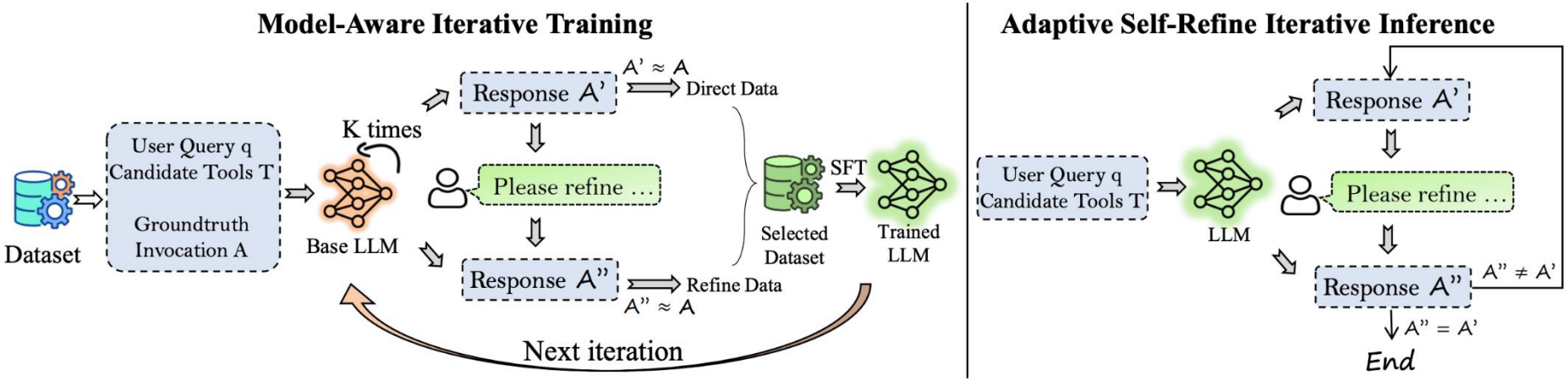
通用能力结果



Zezhong Wang, Xingshan Zeng#, Weiwen Liu, et al. "ToolFlow: Boosting LLM Tool-Calling Through Natural and Coherent Dialogue Synthesis." *NAACL 2025.*

# ToolACE-R: 模型感知的工具潜能激发

- 训练侧：模型能力感知的迭代训练 （pass@k + self-refine data augmentation）
- 推理侧：可自适应的推理时扩展（adaptive iterative self-refine）




- 实验结果：
  - 基于同一初始数据集，超越ToolACE-8B （81.78 -> 86.00）
  - Adaptive Self-Refine 进一步提升，在较难的Live集提升更明显

| Models | Non-Live | | | | Live | | | | Overall | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Simple* | *Multiple* | *Parallel* | *Parallel Multiple* | *Simple* | *Multiple* | *Parallel* | *Parallel Multiple* | *Non-live* | *Live* | *Overall* |
| **GPT-4o-2024-11-20** | 79.42 | 95.50 | 94.00 | 83.50 | **84.88** | 79.77 | 87.50 | 75.00 | 88.10 | 80.75 | 84.43 |
| **DeepSeek-V3** | 78.67 | 95.5 | 91.00 | 91.50 | 83.72 | **82.15** | 81.25 | 62.50 | 89.17 | **82.09** | 85.63 |
| **Llama3.1-70B-Inst** | 77.92 | 96.00 | **94.50** | 91.50 | 78.29 | 76.16 | 87.50 | 66.67 | 89.98 | 76.53 | 83.26 |
| **Llama3.1-8B-Inst** | 71.50 | 93.50 | 86.50 | 86.00 | 73.26 | 68.95 | 56.25 | 50.00 | 84.37 | 69.28 | 76.83 |
| **ToolACE-8B (FC)** | 76.67 | 93.50 | 90.50 | 89.50 | 73.26 | 76.73 | 81.25 | 70.83 | 87.54 | 76.02 | 81.78 |
| **ToolACE-R (FC)** | 79.92 | **96.50** | 93.50 | **93.00** | 81.78 | 81.01 | 87.50 | **83.33** | 90.73 | 81.27 | 86.00 |
| **+ Adaptive Self-Refine** | **81.17** | 96.00 | 94.00 | **93.00** | 82.95 | 81.58 | 87.50 | **83.33** | **91.04** | 81.94 | **86.49** |

BFCL-v3榜单结果（01/18/2025）



Xingshan Zeng, Weiwen Liu, Xu Huang et al. "ToolACE-R: Tool Learning with Adaptive Self-Refinement." *2025.*

# More Discussion: 快思考 vs 慢思考

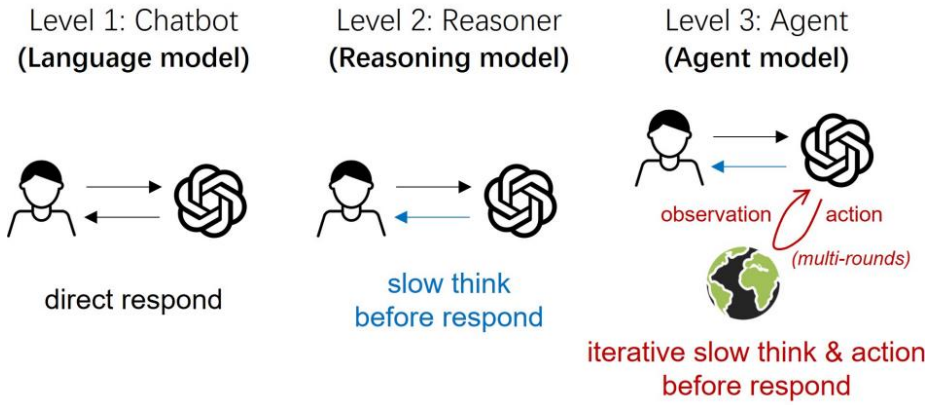
- 性能：
  - 慢思考模型**总体劣于**快思考模型
  - 慢思考在**较难**的Live集上有优势

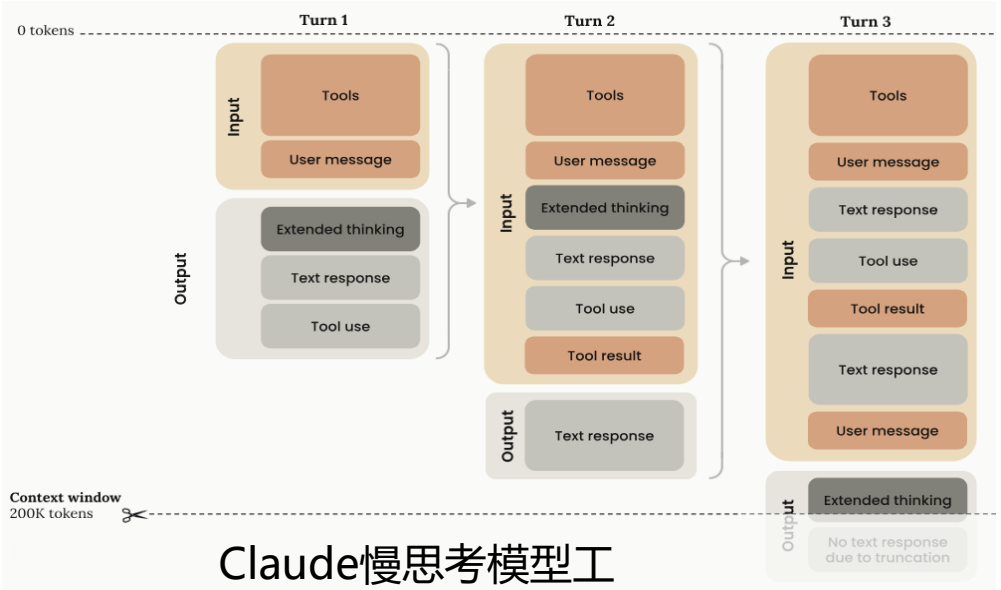
| Models | Overall | Non-live(AST) | Live(AST) | Multi-turn |
|---|---|---|---|---|
| **gpt-4o-2024-11-20** | 70.42 | 87.67 | 79.88 | 43 |
| **o1-2024-12-17** | 67.79 | 85.67 | 80.63 | 36.12 |
| **deepseek-v3** | 57.61 | 89.17 | 68.41 | 18.75 |
| **deepseek-r1** | 56.89 | 87.35 | 74.41 | 12.38 |

BFCL-v3结果（04/25/2025）

- 效率：
  - 慢思考过程引入更多的时延
  - **多步调用**：时延随步数翻倍
  - 解决（缓解）：优化多轮思考模式

- 未来：
  - **Reasoning + Autonomous Action** = Agent?




Claude慢思考模型工具模式多轮拼接方式

# Thank you!