ToolACE-MT: Non-Autoregressive Generation for Agentic Multi-Turn Interaction

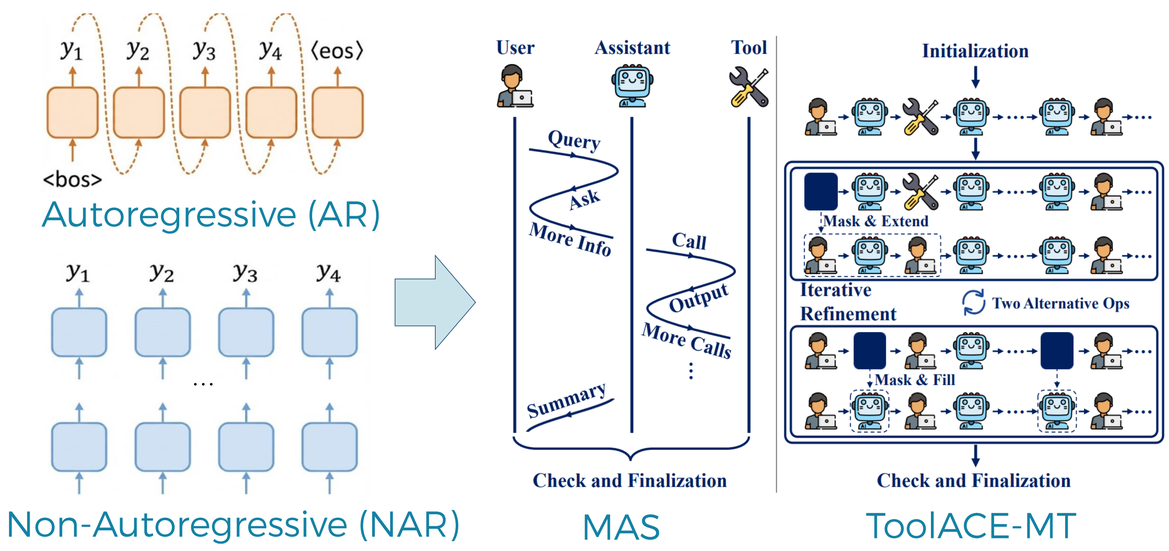
Xingshan Zeng, Weiwen Liu°, Lingzhi Wang, Liangyou Li, Fei Mi
Yasheng Wang, Lifeng Shang, Xin Jiang, Qun Liu
zeng.xingshan@huawei.com; wwliu@sjtu.edu.cn

Introduction

LLM-based agentic task solving requires multi-turn multi-step interactions. Existing agentic data generation relies on **Multi-Agent Simulation (MAS)**, which is autoregressive.

Problems: 1) High cost; 2) Limited control; 3) Accumulated error.

Motivation: Non-autoregressive generation

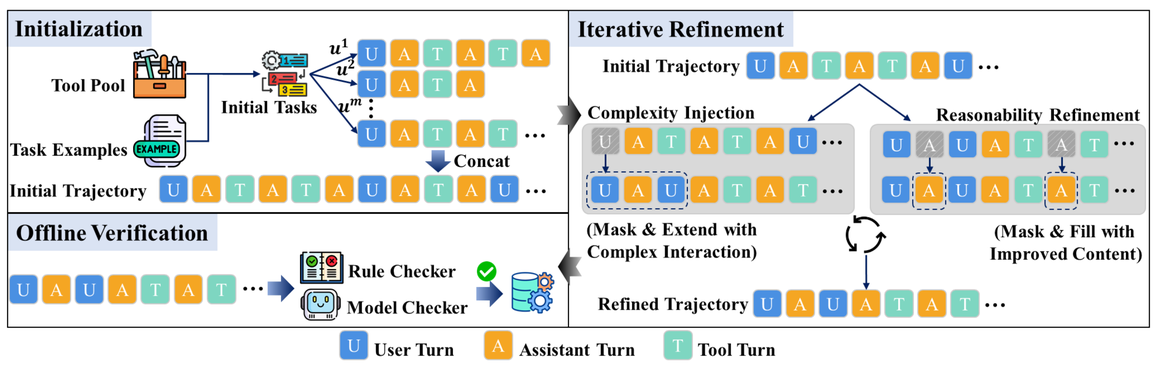


ToolACE-MT

ToolACE-MT pipeline includes 3 stages:

① Coarse-grained Initialization;
② Iterative Refinement;
③ Offline Verification.

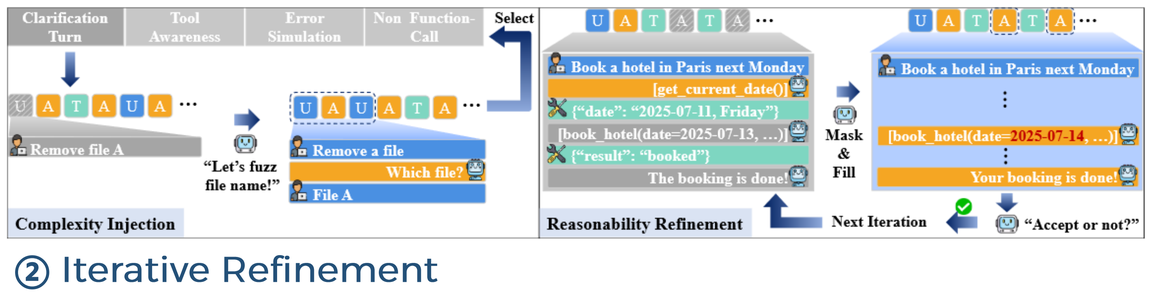
Gradually improving quality (Complexity and Accuracy)



Overview of ToolACE-MT Pipeline

① Coarse-grained Initialization

- **Task Initialization**: high-level planning
- **Trajectory Initialization**: each task is initialized with complete trajectory (without user interaction) and then concatenated



② Iterative Refinement

- Target: improve **complexity and accuracy**
- **Complexity Injection** (mask and extend):
  - Clarification
  - Tool Awareness
  - Error Simulation
  - Non-FC turn

  } 4 kinds of complex interaction
- **Reasonability Refinement** (mask and fill):
  - Function: logical consistency and coherence validation

③ Offline Verification

- **Rule-based**:
  - format compliance, executability, identifiable hallucination...
- **Model-based**:
  - semantic coherence, complex hallucinations...

Experiments

- Settings:
  - **Base model**: LLaMA-3.1-8B-Instruct (trained with LoRA)
  - **Data**: 8000 instances with both ToolACE-MT and MAS
- Data quality:
  - More focus on call turns; improved **coherence and diversity**

Method	Statistics					Quality			
	Tool # /D	Turn #/D	UT #/D	CT #/D	Call #/UT	SS	EnR	H	Dist-3
MAS	8.8	28.0	5.8	8.7	2.3	65.23	43.60	7.92	0.319
ToolACE-MT	8.4	23.7	3.4	8.5	3.7	68.34	50.71	9.28	0.357

- ToolACE-MT consistently outperform MAS, especially in MT settings, across three benchmarks.

Models	Multi-Turn					Single-Turn		Hallucination		Overall
	Overall	*Base*	*Miss Func*	*Miss Param*	*Long Context*	*Non-Live*	*Live*	*Rel*	*Irrel*	*Overall*
GPT-4o-2024-11-20	50.00	61.00	45.50	35.50	58.00	86.81	78.85	83.33	81.31	71.71
Gemini-2.5-Pro-Preview-05-06	34.62	39.50	29.50	31.50	38.00	65.35	74.59	33.33	90.67	59.94
DeepSeek-V3-0324	29.87	41.00	21.00	23.00	34.50	88.54	77.34	83.33	76.49	64.71
Llama3.1-70B-Inst	12.50	17.00	13.00	10.50	9.50	89.98	62.24	100	54.78	53.57
Llama3.1-8B-Inst	9.25	12.00	10.00	7.00	8.00	84.21	61.08	77.78	48.82	49.57
Multi-Agent Simulation	31.38	46.50	19.00	31.00	29.00	80.29	**78.05**	72.22	**90.11**	64.17
ToolACE-MT	**40.25**	**57.50**	**31.50**	**34.00**	**38.00**	**84.94**	71.52	77.78	72.83	**65.41**
- Offline Verification	32.50	48.00	25.50	25.50	31.00	79.71	75.52	**83.33**	80.65	63.01
- Iterative Refinement	20.88	39.00	12.00	10.50	22.00	75.92	61.57	72.22	46.25	52.10

BFCL-v3 results

Models	MT	EA	PA
GPT-4o-2024-11-20	68.0	56.0	77.8
Llama3.1-70B-Inst	61.0	41.0	62.5
Llama3.1-8B-Inst	24.0	6.7	18.3
Multi-Agent Simulation	48.0	6.7	15.0
ToolACE-MT	**51.0**	**8.4**	**34.0**
- Offline Verification	44.0	1.7	28.5
- Iterative Refinement	34.0	1.7	22.8

ACEBench results

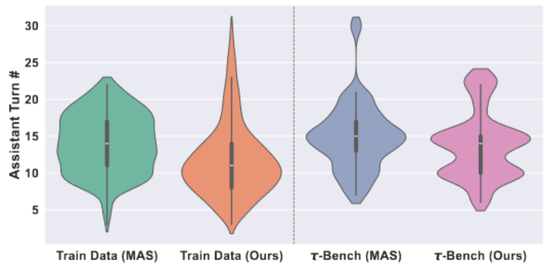
Models	Retail	Airline	Avg.
GPT-4o-2024-11-20	60.4	42.0	51.2
Llama3.1-70B-Inst	50.4	26.0	38.2
Llama3.1-8B-Inst	6.1	26.0*	16.1
Multi-Agent Simulation	21.7	10.0	15.9
ToolACE-MT	**25.2**	**16.0**	**20.6**
- Offline Verification	22.6	6.0	14.3
- Iterative Refinement	9.5	6.0	7.8

Tau-Bench results

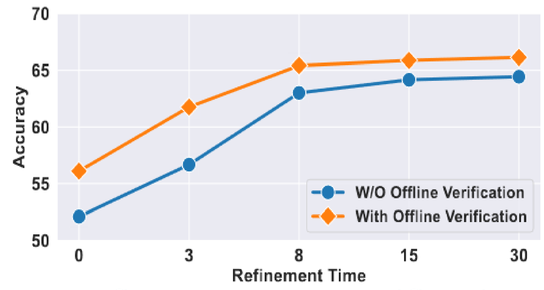
- ToolACE-MT exhibits better generation efficiency.

Method	Cost				Quality Performance	
	API Call times	Input (#)	Output (#)	Pricing (USD)		
MAS with GPT-4o	275k	652,145k	10,712k	1,737	61.1	64.17
ToolACE-MT with GPT-4o	188k	478,115k	18,480k	1,380	72.3	65.41
ToolACE-MT with GPT-4o-mini	394k	855,391k	33,062k	148	48.7	60.13

- More analysis:
  - Better **completion efficiency**: less average assistant turns for completing tasks.
  - Better **effectiveness and generalizability**: complementary roles of Refinement and Verification; robust improvement across different backbones and sizes.



Assistant turn statistics



Refinement v.s. Verification

Conclusion

- We propose ToolACE-MT, an agentic data generation pipeline motivated by non-autoregressive generation. It leverages techniques like mask&fill to produce high-quality multi-turn agentic data.
- Compared to MAS, it generates data with better efficiency and effectiveness.
- Consistent improvement, as well as generalizability and robustness across diverse settings.

Team-ACE

ICLR 2026: The Fourteenth International Conference on Learning Representations, Rio de Janeiro, Brazil, April 23-27, 2026