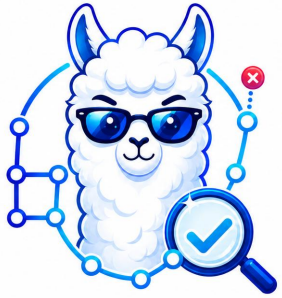
NEURAL INFORMATION PROCESSING SYSTEMS



VerifyMAS

VerifyMAS: Hypothesis Verification for Failure Attribution in LLM Multi-Agent Systems

Hezhe Qiao, Hanghang Tong, Ee-Peng Lim, Bing Liu,
Guansong Pang

SMU
SINGAPORE MANAGEMENT UNIVERSITY

UIC

MAS Failure Attribution



Given a **trajectory** of an LLM-MAS failure τ, our objective is to **attribute the failure** by determining **which agent caused a specific error** that led to the failure of the overall agent system.

Predicted Agent Error



Challenges and Motivation



Direct Prediction (DPR)

```
Your sole purpose is to analyze
conversation logs to identify and
categorize agent errors based on
a strict set of definitions.
```

Chain of Thought (CoT-Agent)

```
Your sole purpose is to analyze
conversation logs to identify and
categorize agent errors based on
a strict set of definitions.
### Step 1: Agent Summary
### Step 2: Error Analysis
### Step 3: Final Judgment
```

Predicted Agent Error

"local action-level redundancy"

Insufficient

VerifyMAS - Toy example

User: Please help prepare a reimbursement request for a conference trip.

Important requirements:
1. The reimbursement should be submitted *under Prof. Andy's account, not the student's account.*

Coordinator: The user wants to prepare a reimbursement request for a conference trip. Agent Checker, please check whether the receipt information is usable.

Checker: The receipt currently shows the student's name as the account holder. However, the instruction says it should be under Prof. Andy account. So the receipt name should be corrected to Prof. Andy.

Coordinator: Good. Agent Writer, now draft the reimbursement request.

Writer: Before drafting, I will first search online for the general history

local action-level redundancy

Coordinator: Please skip the background search and draft the request directly.

Writer: Here is the draft:

Dear Finance Office,

I would like to request reimbursement for my conference trip. The receipt is attached, and the payment was made under my account.
Please process the reimbursement accordingly.

Best regards,
Student

global context-level inconsistency

Verifier: The draft is complete and ready to submit.

incorrect verification



VerifyMAS

Given a trajectory and a error hypothesis, choose exactly one label: A = entail; B = neutral; C = contradict.

Hypothesis Verification

Trajectory

Hypothesis:An agent used reasoning that was inconsistent with... A ✓

Hypothesis:An agent repeated a task... A ✓

...

Hypothesis:An agent deviated from the specified task... C ✗

Responsible Agent Attribution

✓ "local action-level redundancy"

✓ "global context-level inconsistency"

Evaluation Metric

❑ *Micro-F1 (μF1)*

Overall classification performance

Aggregates decisions across all classes, emphasizing the overall attribution quality.

❑ *Class-wise Macro-F1 (MF1)*

Balanced class-wise performance

Gives equal weight to each class, highlighting performance on less frequent error modes.

Auxiliary: Precision + Recall

Evaluation granularity

A single trajectory may contain multiple faulty agents and multiple error modes.

P

❖ **Pair-level**

Correctly identified agent–error pairs.

A

❖ **Agent-level**

Faulty-agent identification regardless of error type.

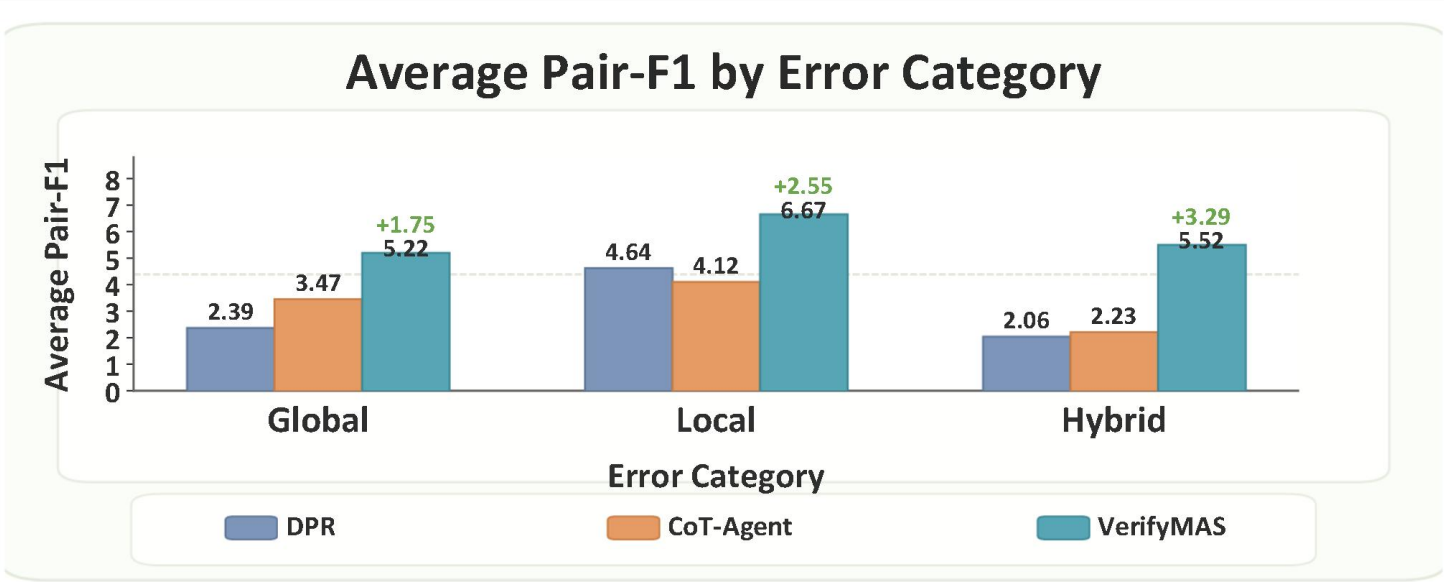
E

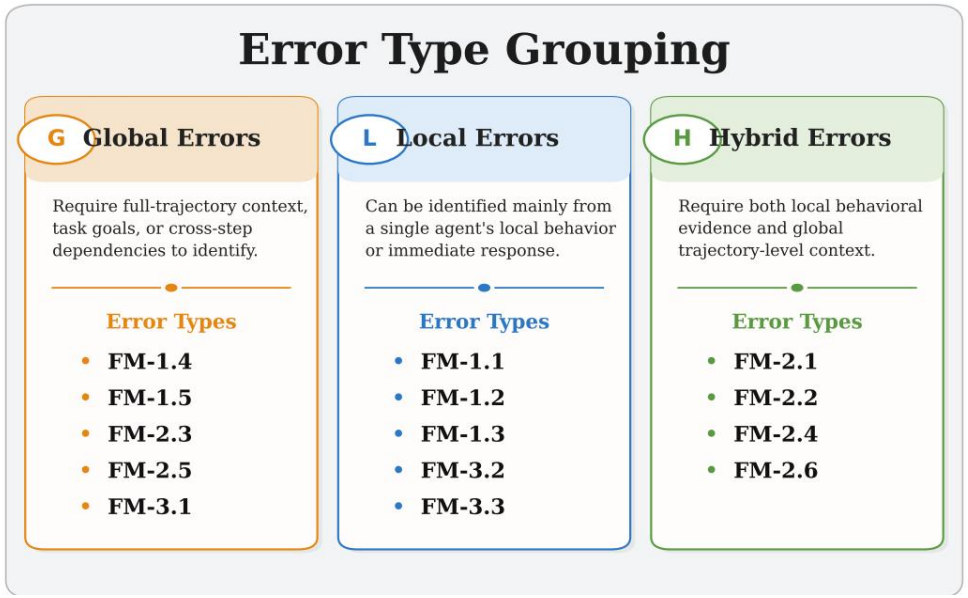
❖ **Error-level**

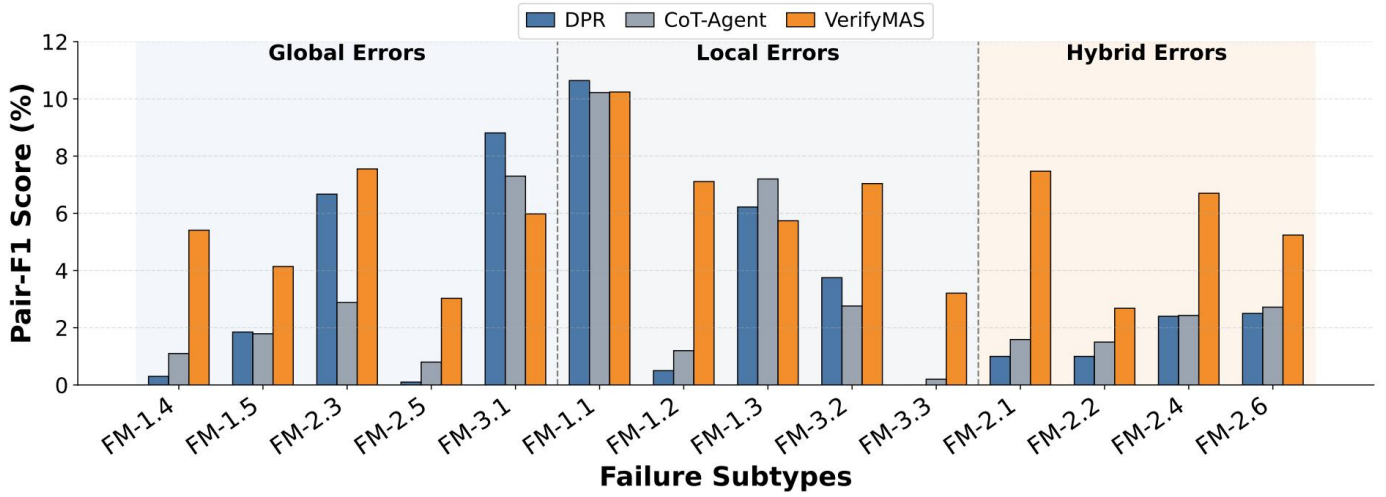
Error-mode identification regardless of responsible agent.

VerifyMAS-Failure Analysis

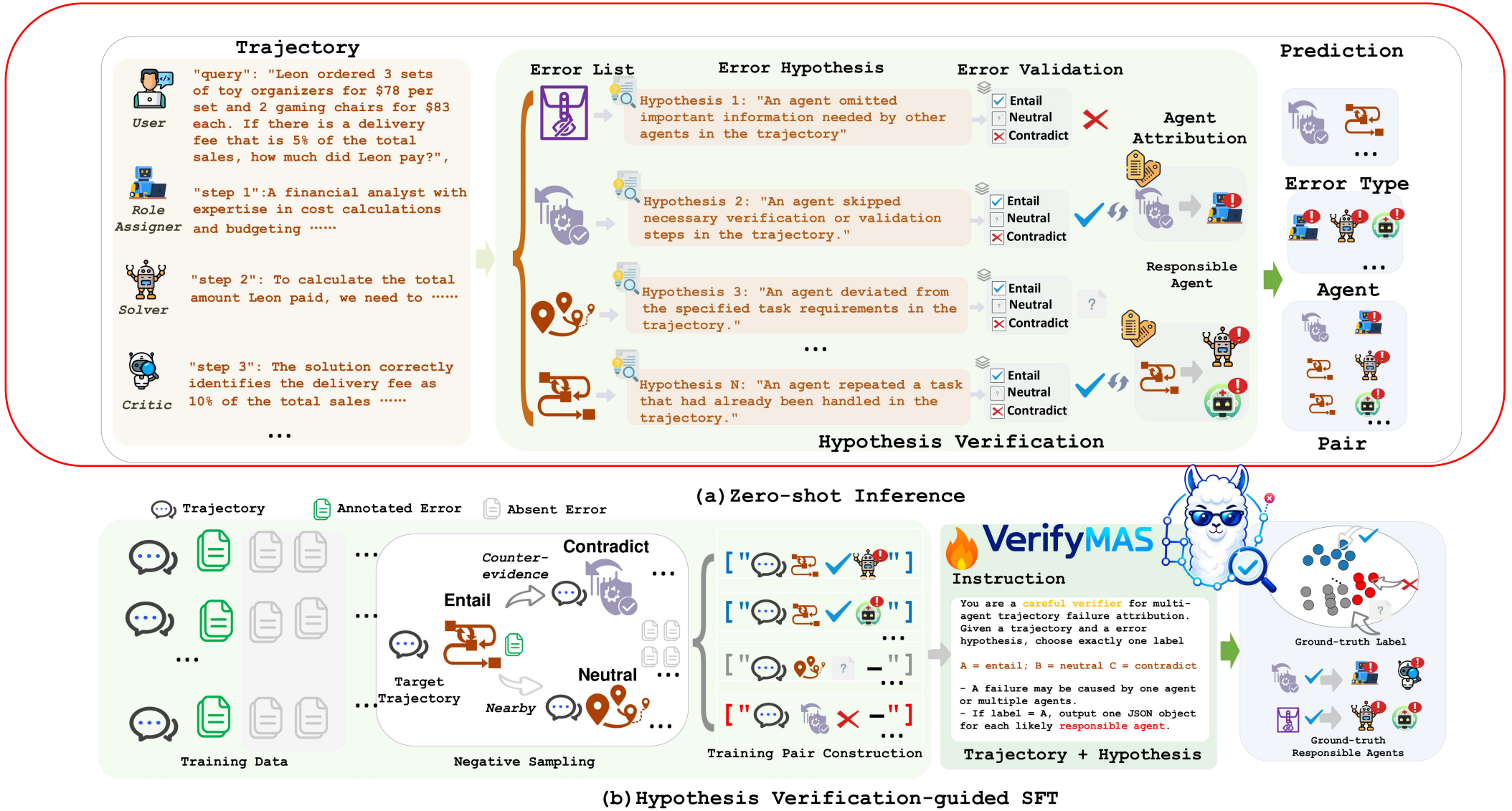
Error Type	Error Definition	AEGIS	Who&When
FM-1.1	Deviates from task requirements.	2,310	145
FM-1.2	Acts outside the assigned role.	1,824	103
FM-1.3	Adds redundant steps.	1,626	92
FM-1.4	Ignores conversation history.	1,654	87
FM-1.5	Misses stopping criteria.	2,177	108
FM-2.1	Repeats completed tasks.	1,869	99
FM-2.2	Makes requests ambiguous.	1,758	132
FM-2.3	Drifts from the main goal.	1,823	110
FM-2.4	Hides key information.	1,713	104
FM-2.5	Ignores other agents' inputs.	1,660	100
FM-2.6	Uses inconsistent reasoning.	1,513	103
FM-3.1	Stops the task prematurely.	1,647	113
FM-3.2	Skips verification steps.	1,618	114
FM-3.3	Performs incorrect verification.	1,651	106
Total	–	24,843	1,516
Test Trajectory	–	600	184







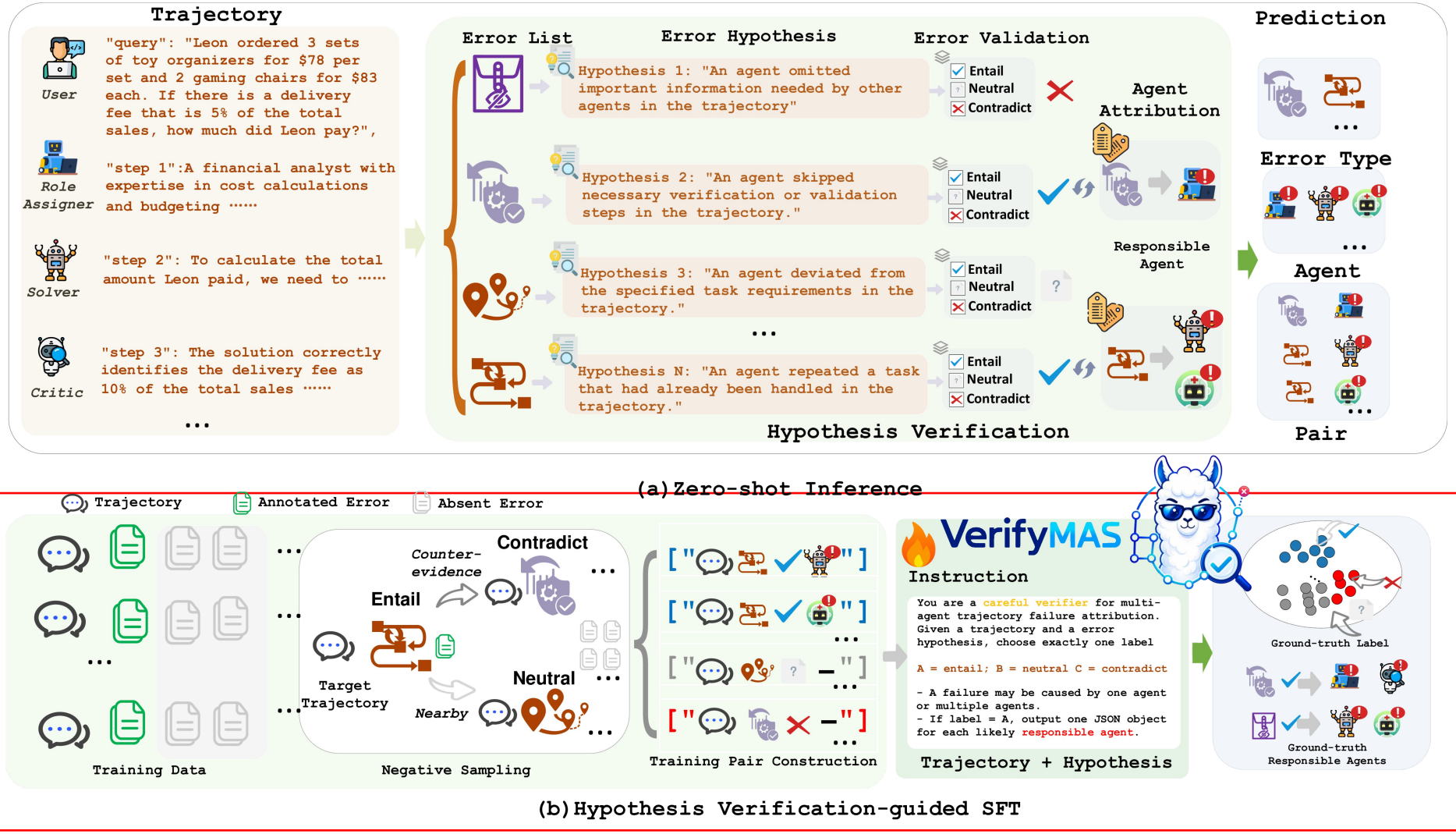
VerifyMAS - Pipeline



Zero-shot ReusIts

Model	Aegis-Bench						Who&When					
	Pair		Agent		Error		Pair		Agent		Error	
	μF1	MF1	μF1	MF1	μF1	MF1	μF1	MF1	μF1	MF1	μF1	MF1
Open-source Models												
Qwen2.5-7B-Instruct	5.02	2.52	27.55	14.49	14.96	11.36	2.31	1.14	40.92	23.50	3.64	1.77
+VerifyMAS	5.89 ↑0.87	**2.98** ↑0.46	47.92 ↑20.37	**14.66** ↑0.17	25.07 ↑10.11	24.26 ↑12.90	**5.83** ↑3.52	**3.44** ↑2.30	39.26 ↓1.66	**33.50** ↑10.00	**13.33** ↑9.69	**9.54** ↑7.77
Qwen3-8B-Non-Thinking	3.96	1.40	21.34	8.16	15.81	13.89	3.88	1.81	27.78	17.64	3.88	1.91
+VerifyMAS	5.49 ↑1.53	1.85 ↑0.45	42.96 ↑21.62	11.65 ↑3.49	26.50 ↑10.69	24.05 ↑10.16	4.92 ↑1.04	2.88 ↑1.07	35.03 ↑7.25	24.11 ↑6.47	11.93 ↑8.05	5.48 ↑3.57
Qwen2.5-14B-Instruct	5.47	2.20	35.78	12.71	20.24	5.91	0.00	0.00	**49.88**	33.19	1.56	1.35
+VerifyMAS	**7.40** ↑1.93	2.69 ↑0.49	**48.95** ↑13.17	14.50 ↑1.79	**28.37** ↑8.13	**26.90** ↑20.99	5.15 ↑5.15	2.68 ↑2.68	35.97 ↓13.91	27.53 ↓5.66	10.53 ↑8.97	7.04 ↑5.69
Proprietary Models												
GPT-4o-mini	5.76	1.63	38.54	14.72	19.95	16.02	2.11	0.98	47.42	34.21	5.26	3.33
+VerifyMAS	6.58 ↑0.82	2.62 ↑0.99	49.30 ↑10.76	16.54 ↑1.82	27.89 ↑7.94	**27.79** ↑11.77	5.19 ↑3.08	3.19 ↑2.21	44.96 ↓2.46	38.00 ↑3.79	12.39 ↑7.13	11.36 ↑8.03
GPT-4.1	7.44	2.27	37.48	11.12	20.65	15.75	3.36	1.16	42.29	28.93	7.00	5.84
+VerifyMAS	**7.59** ↑0.15	**2.79** ↑0.52	49.52 ↑12.04	17.07 ↑5.95	**27.95** ↑7.30	27.37 ↑11.62	5.09 ↑1.73	2.83 ↑1.67	42.93 ↑0.64	37.44 ↑8.51	13.06 ↑6.06	10.66 ↑4.82
Gemini-2.5-Flash	6.99	2.76	42.02	16.45	23.47	19.85	7.32	3.33	**55.56**	36.98	11.94	7.96
+VerifyMAS	7.02 ↑0.03	2.77 ↑0.01	50.33 ↑8.31	16.91 ↑0.46	27.23 ↑3.76	26.73 ↑6.88	**7.39** ↑0.07	3.49 ↑0.16	43.25 ↓12.31	37.23 ↑0.25	**13.33** ↑1.39	**11.65** ↑3.69
Gemini-2.5-Pro	6.96	2.88	41.32	16.15	19.93	16.29	6.81	2.69	53.11	34.92	11.07	8.11
+VerifyMAS	7.47 ↑0.51	2.60 ↓0.28	**53.61** ↑12.29	**18.21** ↑2.06	27.27 ↑7.34	26.99 ↑10.70	7.33 ↑0.52	**4.18** ↑1.49	49.10 ↓4.01	**41.47** ↑6.55	12.94 ↑1.87	11.39 ↑3.28

VerifyMAS - Pipeline



Supervised Fine-tuning

$$\mathbf{y}_{i,m} = \{\ \textit{"label"}\colon\ \textit{"A"}\ \textit{"agent"}\colon a\},\ \hat{z}_{i,m} = A,$$

$$\begin{cases} \mathbf{y}_{neutral} = \{\ \textit{"label"}\colon\ \textit{"B"}\ \textit{"agents"}\colon [\quad]\}, & \hat{z} = B, \\ \mathbf{y}_{contradict} = \{\ \textit{"label"}\colon\ \textit{"C"}\ \textit{"agents"}\colon [\quad]\}, & \hat{z} = C. \end{cases}$$

HSFT Training Loss

$$\mathcal{L}_{HSFT} = -\sum_{t=1}^{|\mathrm{y}^{*}_{i,m}|} \log p_\theta \left(y^{(t)*}_{i,m} | \mathbf{x}_{i,m}, y^{(<t)*}_{i,m} \right)$$

$$\{\ \text{"label"} : \text{"A"}, \text{"agents"} : [\text{"Planner"}]\ \}$$
$$\{\ \text{"label"} : \text{"A"}, \text{"agents"} : [\text{"Solver"}]\ \}.$$

$$\{\text{"label"}: \text{"B"}, \text{"agents"}: []\}, \{\text{"label"}: \text{"C"}, \text{"agents"}: []\}$$

Counter-evidence keyword mapping

Error Type	Counter-evidence Keywords
FM-1.1	as requested; according to the requirement; followed the requirement; kept the required format; in python; met the specification
FM-1.2	according to assigned role; within role; as assigned; performed its role; stayed in its role
FM-1.3	without duplication; without redundant; concise; single final answer; did not repeat
FM-1.4	as mentioned earlier; based on previous; taking previous context into account; after the correction; considering earlier feedback; from previous turns
FM-1.5	base case; termination condition; stop when; stopping criterion; return when; until condition is met
FM-2.1	already completed so skip; no need to repeat; continue from previous result; reuse the previous result; avoid repeating
FM-2.2	clear instruction; explicit instruction; specifically asked; step-by-step instruction; clearly stated; explicitly stated
FM-2.3	focus on the main task; directly answer the query; stay on task; relevant to the task; aligned with the goal
FM-2.4	shared the important information; provided all relevant information; disclosed; made it clear; explicitly mentioned; fully informed
FM-2.5	incorporated feedback; addressed the concern; responded to the correction; followed the suggestion; took the feedback into account
FM-2.6	consistent with earlier; same conclusion as before; as stated earlier; remained consistent
FM-3.1	all requirements are satisfied; after verification; completed all requested parts; finished after checking; fully completed
FM-3.2	verified; validated; tested; checked; unit test; test case; cross-check
FM-3.3	verified against expected output; validated with test cases; cross-checked results; checked correctness; confirmed the result

Nearby error-type mapping

Error Type	Nearby Error Types
FM-1.1	FM-2.3, FM-1.4, FM-1.2
FM-1.2	FM-1.1, FM-2.3, FM-2.5
FM-1.3	FM-2.1, FM-3.2, FM-2.3
FM-1.4	FM-2.5, FM-2.6, FM-1.1
FM-1.5	FM-3.1, FM-3.2, FM-3.3
FM-2.1	FM-1.3, FM-2.3, FM-2.5
FM-2.2	FM-2.4, FM-2.3, FM-2.5
FM-2.3	FM-1.1, FM-2.2, FM-2.1
FM-2.4	FM-2.2, FM-2.5, FM-1.4
FM-2.5	FM-1.4, FM-2.4, FM-2.6
FM-2.6	FM-1.4, FM-2.5, FM-3.3
FM-3.1	FM-1.5, FM-3.2, FM-3.3
FM-3.2	FM-3.3, FM-1.5, FM-3.1
FM-3.3	FM-3.2, FM-2.6, FM-3.1

Supervised Fine-tuning Results

Table 9: Distribution of three types of samples in the SFT training set.

Label	Type	# Samples	Ratio
A	Entail	17,753	53.02%
B	Neutral	5,767	17.23%
C	Contradict	9,961	29.75%
Total	–	**33,481**	**100.0%**

Table 10: Training Hyperparameters for HSFT of VerifyMAS

Parameter	SFT Stage
Training Strategy	FSDP
Learning Rate	1e-4
Batch Size (Train)	64
Micro Batch Size per GPU	2
Max Sequence Length	8192
Total Epochs	2
LoRA Configuration	$r = 64, \alpha = 16$
Target Modules	all-linear
Gradient Clipping	1.0
Weight Decay	0.01
LR Scheduler	cosine
Warmup Steps Ratio	0.1
Model Precision	bf16
Gradient Checkpointing	True
CPU Offload	True (params)
Number of GPUs	2

Model	Pair		Agent		Error	
	μF1	MF1	μF1	MF1	μF1	MF1
Aegis-Bench						
DCL	8.33	5.30	22.93	20.23	24.73	27.70
Qwen2.5-7B-SFT	5.05	2.80	60.03	22.70	19.61	16.90
VerifyMAS-7B-SFT	11.93 ↑6.88	5.95 ↑3.15	61.86 ↑1.83	22.96 ↑0.26	**29.95** ↑10.34	29.80 ↑12.9
Qwen3-8B-SFT	9.68	5.73	**64.79**	**38.96**	20.37	20.36
VerifyMAS-8B-SFT	**12.17** ↑2.49	**5.99** ↑0.26	60.19 ↓4.60	21.23 ↓17.73	29.94 ↑9.57	**30.34** ↑9.98
Who&When						
DCL	1.60	0.77	8.40	6.07	14.67	10.57
Qwen2.5-7B-SFT	1.26	0.52	43.51	32.51	6.77	4.20
VerifyMAS-7B-SFT	**5.25** ↑3.99	2.44 ↑1.92	42.00 ↓1.51	**33.69** ↑1.18	**14.85** ↑8.08	**13.25** ↑9.05
Qwen3-8B-SFT	5.17	2.33	**45.48**	30.77	8.00	5.29
VerifyMAS-8B-SFT	5.21 ↑0.04	**2.85** ↑0.52	40.09 ↓5.39	32.90 ↑2.13	13.35 ↑5.35	12.32 ↑7.03

Generalization on Unseen Failure

❑ *Unseen failure evaluation on ID dataset*

Hold out four failure types (FM-1.4, FM-1.5, FM-2.6, FM-3.3)

Table 3: Unseen-error evaluation on Aegis-Bench-U.

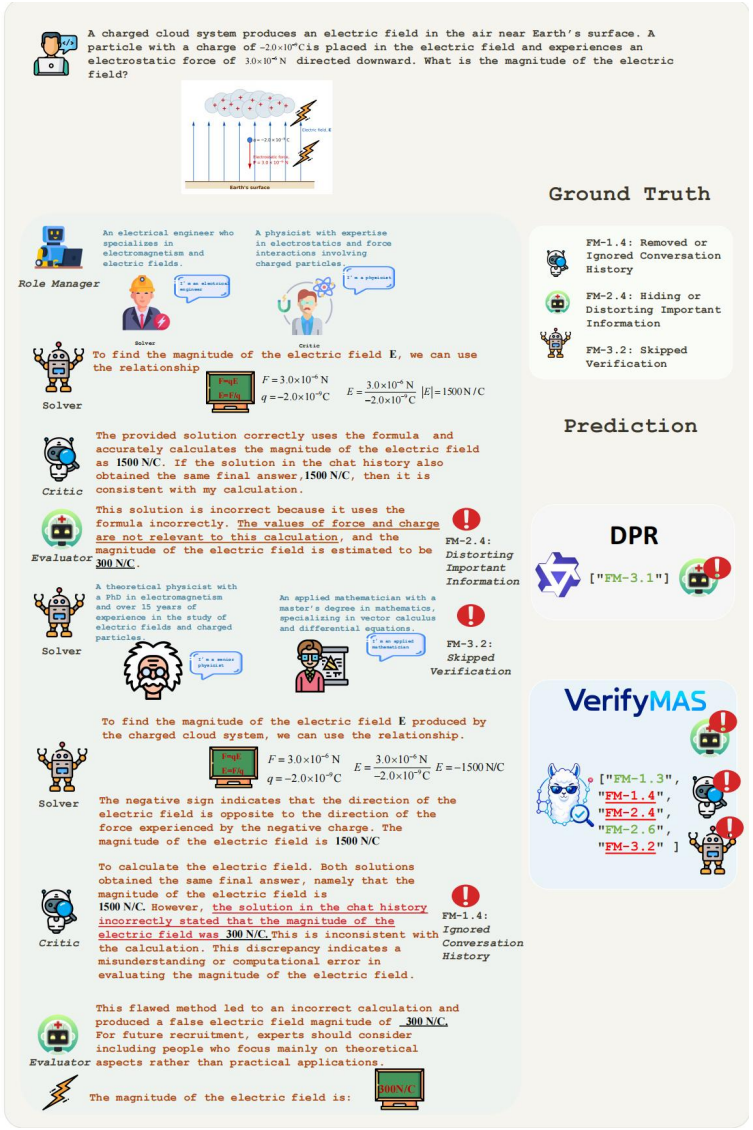
Method	Pair		Error	
	μF1	MF1	μF1	MF1
Zero-shot				
Qwen2.5-7B	6.60	2.31	30.62	26.99
VerifyMAS	7.40	3.15	36.97	35.66
SFT				
Qwen2.5-7B-HO	3.99	1.49	12.89	9.53
VerifyMAS-7B-HO	**11.74**	**4.96**	**40.13**	**35.72**

❑ *OOD dataset evaluation*

Table 9: Unseen-error experiment on TRAIL-U (Four-unseen error evaluation).

Method	**Pair**		**Error**	
	μF1	MF1	μF1	MF1
Zero-shot				
Qwen2.5-7B (baseline)	8.20	2.89	13.11	10.86
VerifyMAS	12.24	4.13	32.68	28.59
SFT				
Qwen2.5-7B-HO (baseline)	3.24	1.25	27.27	13.13
VerifyMAS-7B-HO	**21.43**	**4.37**	**51.94**	**46.41**

Case Study



FM-2.4: Hiding or Distorting Important Information

FM-1.4: Removed or Ignored Conversation History

FM-3.2: Skipped Verification.

Future Work- Toward Closed-Loop MAS Reliability

Research roadmap: detect → analyze → calibrate

Trajectory Failure Detection ⇨ ***Real-Time Agent Failure Analysis*** ⇨ ***Suggestion/Skill-Guided Calibration***

From trajectory-level diagnosis to real-time agent correction and calibration.