ICML
International Conference
On Machine Learning

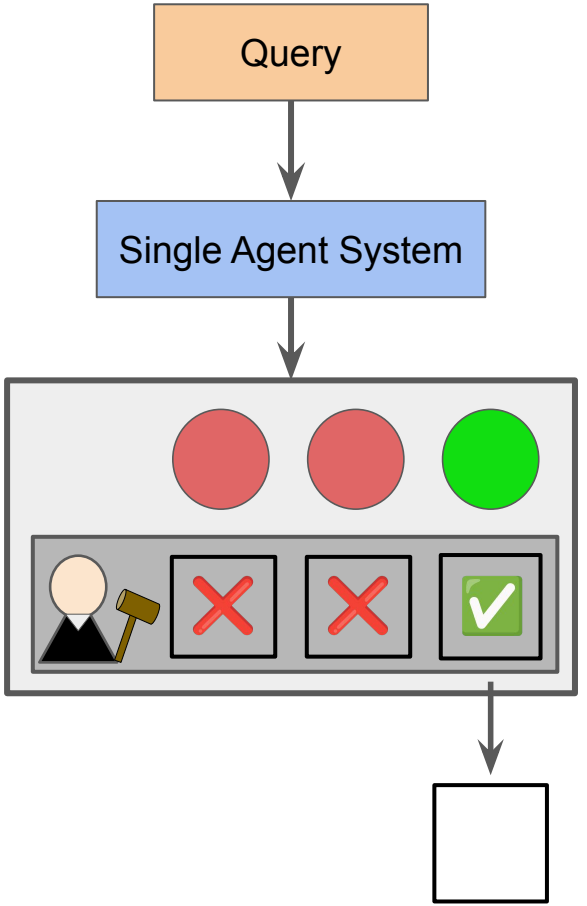
MAS-ProVe: Understanding the Process Verification of Multi-Agent Systems

Vishal Venkataramani[1] Haizhou Shi[12] Zixuan Ke[†2] Austin Xu[2] Xiaoxiao He [1]
Yingbo Zhou[2] Semih Yavuz[2] Hao Wang[‡1] Shafiq Joty[‡2]

[1]Rutgers University [2]Salesforce AI Research [†]Project Lead [‡]Equal Advising

Motivation

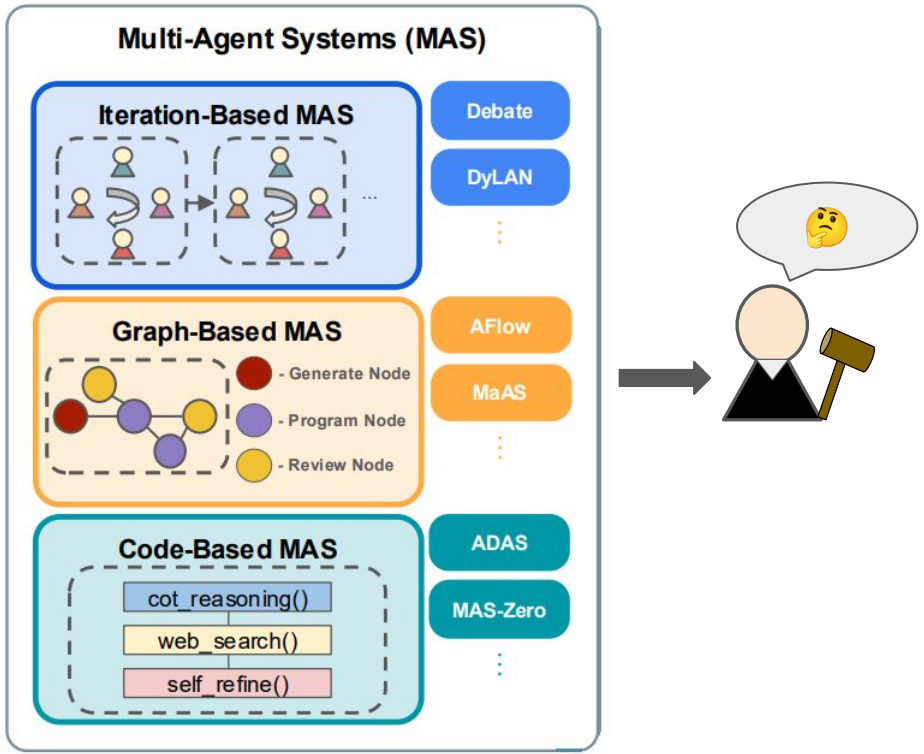
- Process verification works for single agents, but its effectiveness in guiding MAS coordination remains unclear.



Motivation

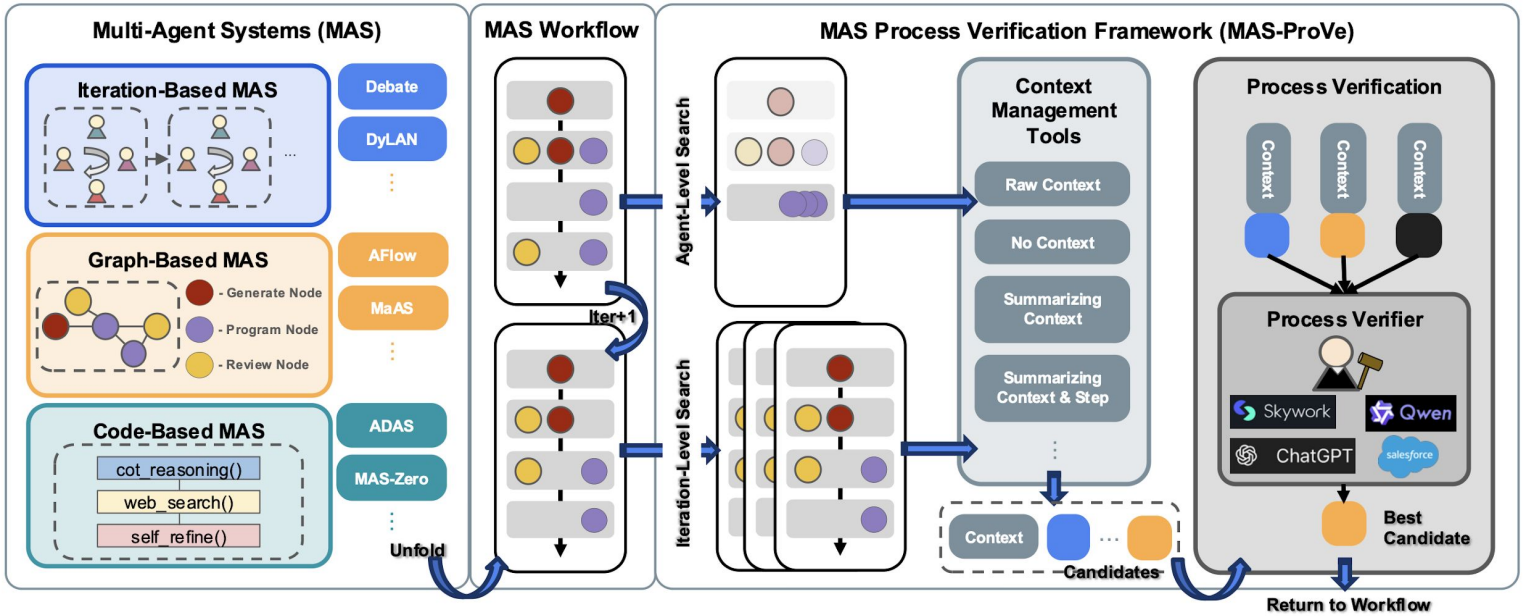
- Process verification works for single agents, but its effectiveness in guiding MAS coordination remains unclear.
- Multi-Agent Systems (MAS) offer parallelism but introduce high variance in reasoning trajectories and multiple failure points.

Do multi-agent systems actually benefit from automatic process-level verification?"



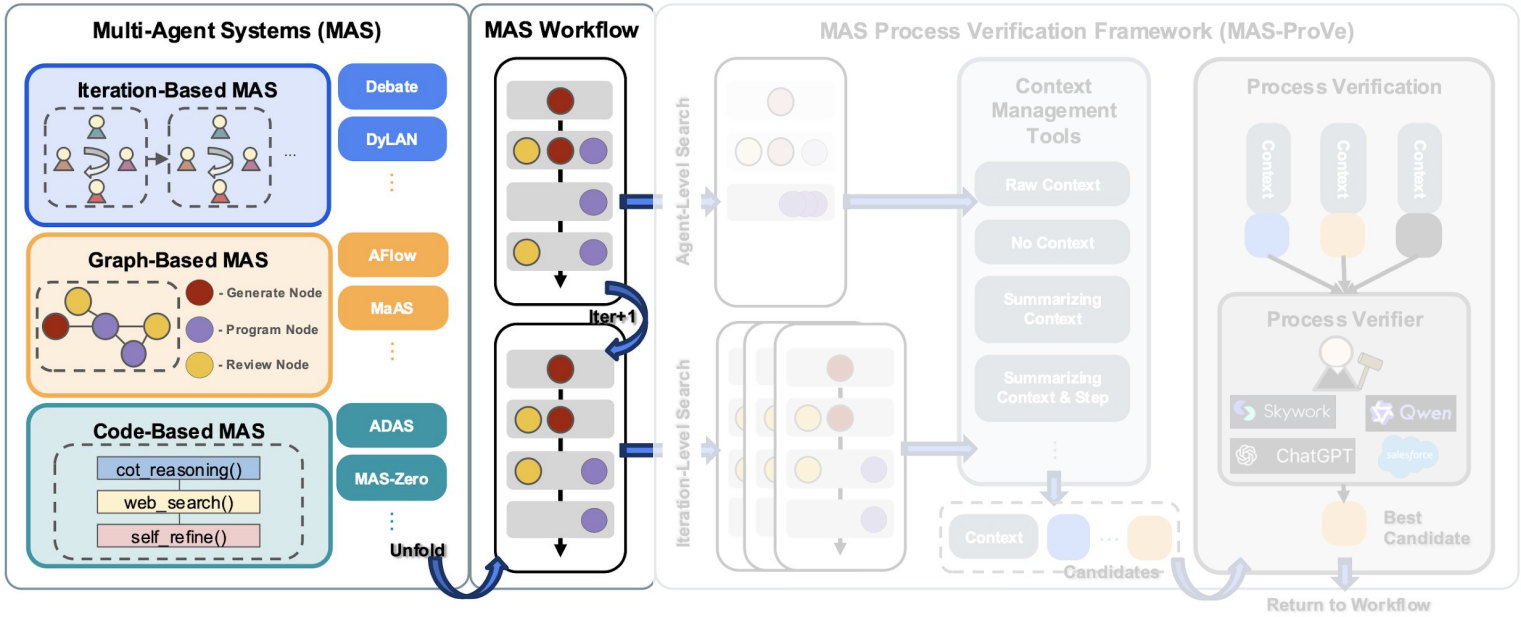
MAS-ProVe: Design Philosophy

- We designed MAS Process Verification framework (MAS-ProVe)
  - ***No (minor) modification*** of original MAS logic
  - Works with ***any verifier***
  - Works with ***any MAS*** framework
  - Clean separation of MAS execution and process verification.



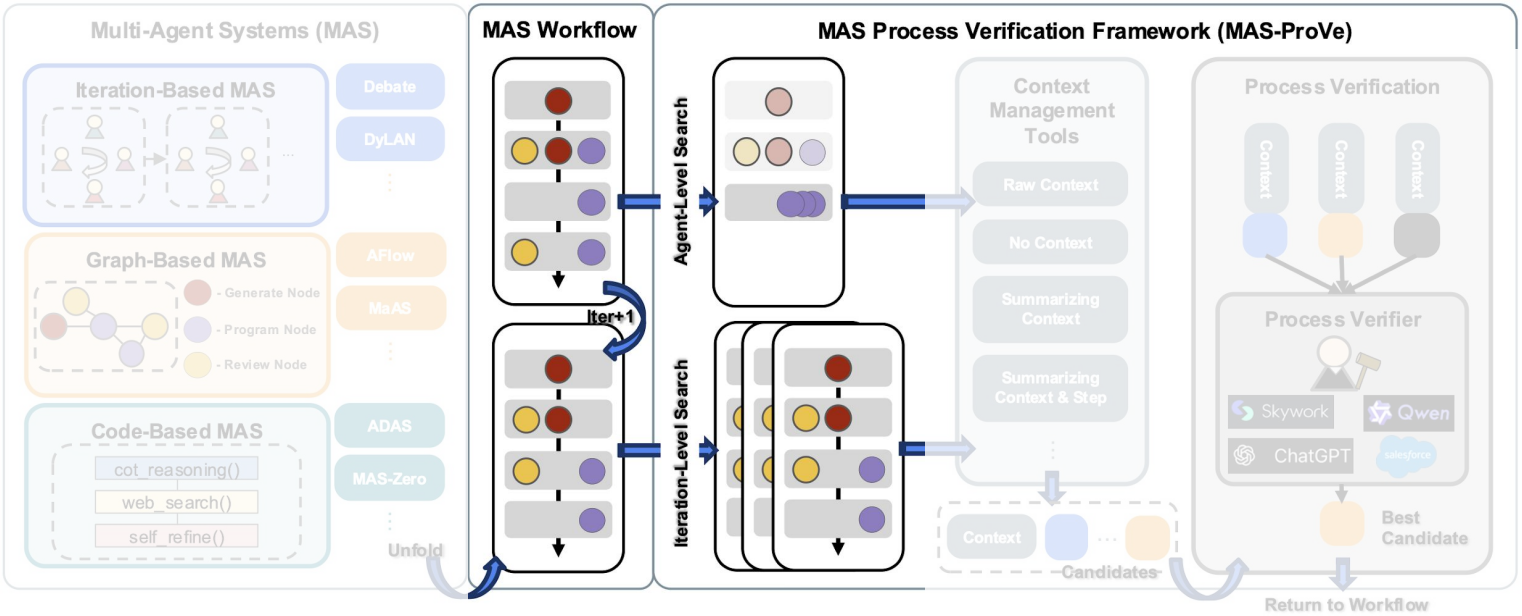
Unifying MAS as a Workflow

- Every MAS execution can be abstracted as a workflow:
  - A sequence of stages;
  - Each stage contains atomic parallelizable LLM calls.



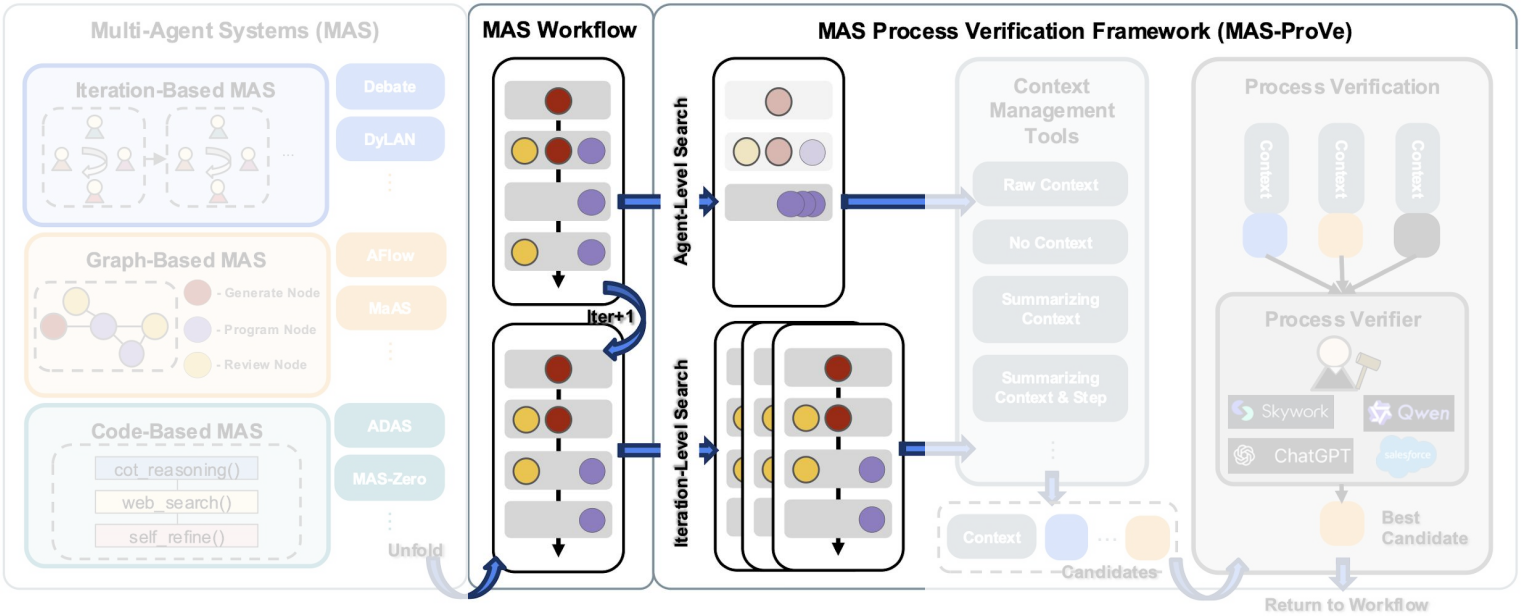
Where Do We Verify?

- We benchmark the verification's effectiveness based on its search efficacy
  - Generation N possible continuations;
  - Use verifier to score them;
  - Select the best one;
  - Continue MAS execution.



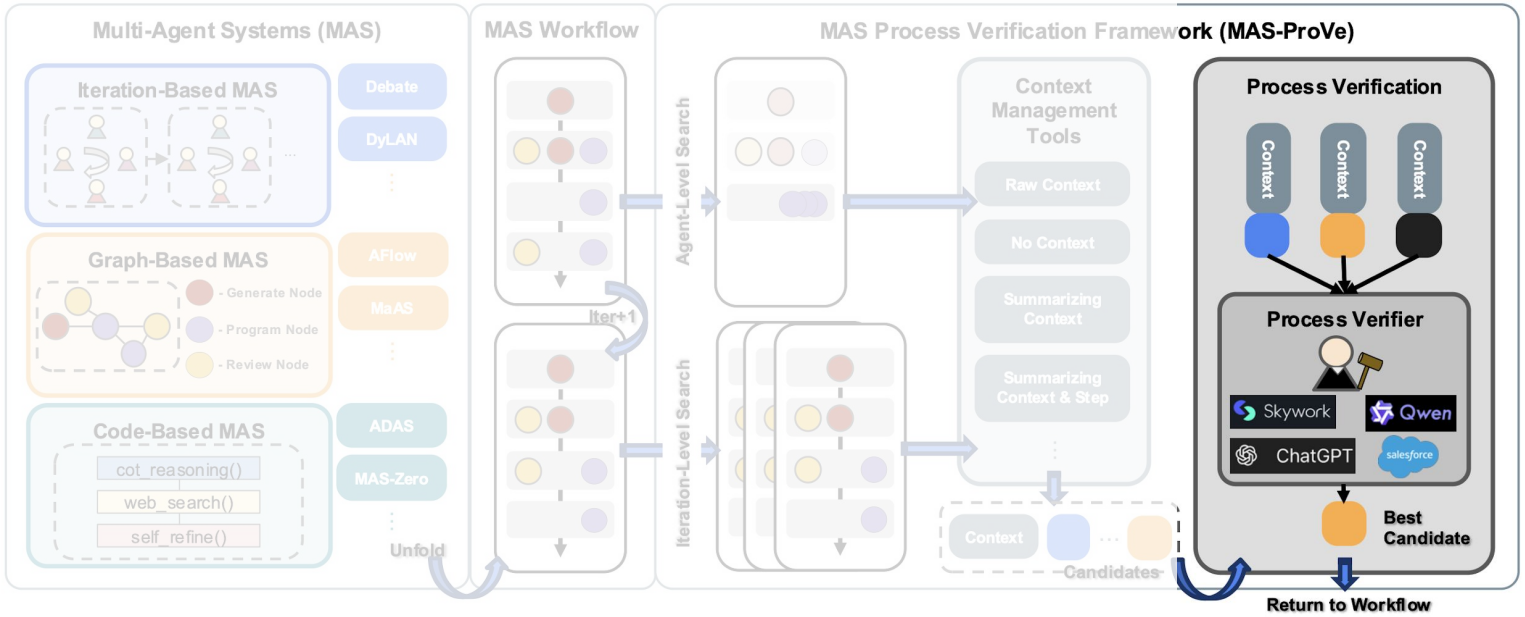
2 Verification Granularities

- Agent-Level Verification
  - Fine-grained but higher overhead.
- Iteration-Level Verification
  - Coarser but cheaper.



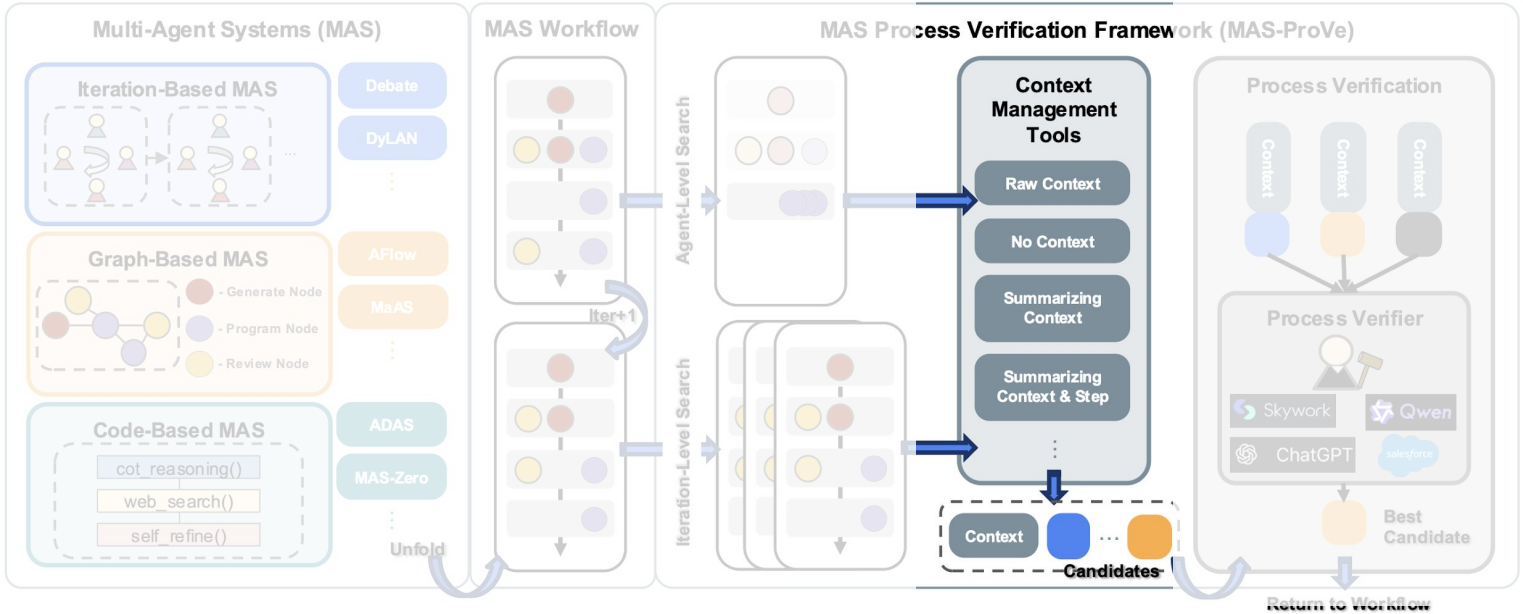
3 Verifier Types

- Reward Models (RMs)
- Process Reward Models (PRMs)
- LLM-as-a-Judge



Context Management: A Hidden Variable

- MAS typically has exceedingly long trajectories.
- What should a verifier see?
  - Current step only?
  - Full history?
  - ***Summarized*** history?



Implementation Architecture

Client

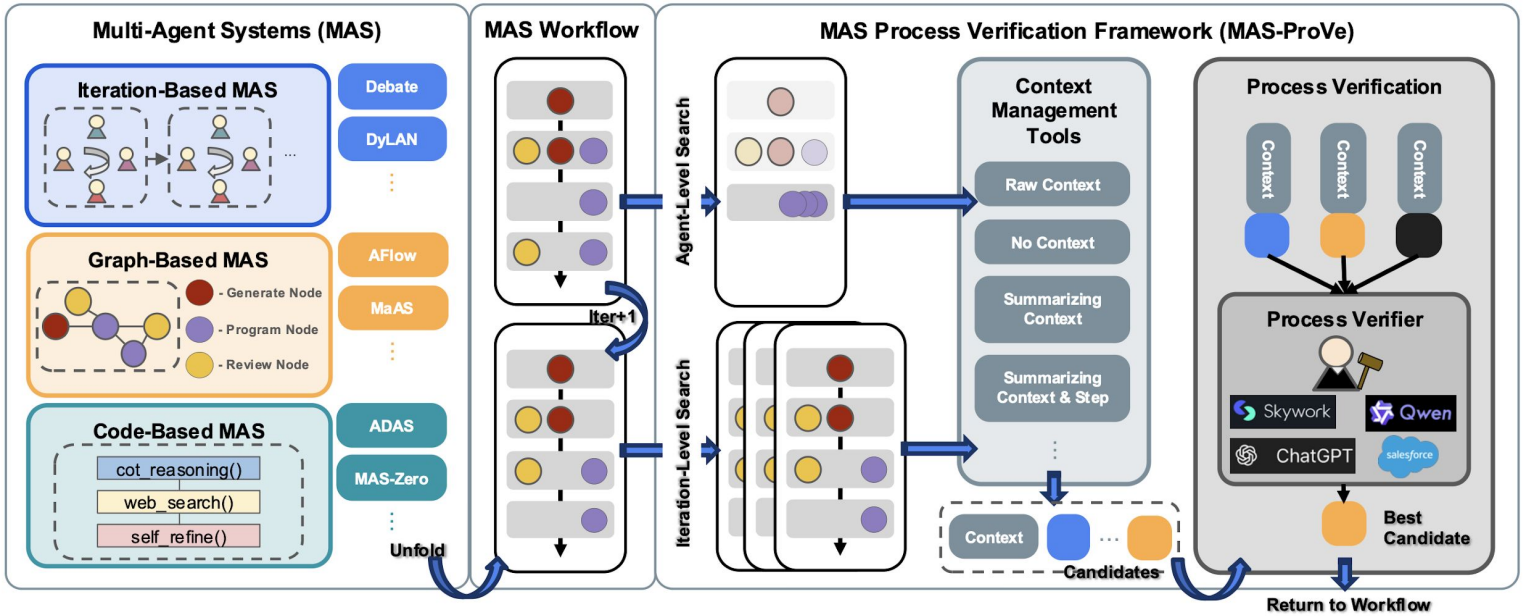
- Works as a simple function wrapper in python.
- Wrapped around the function to be process-verified;

Middleware

- Intercepts context and perform necessary modification before sending to the server.

Server

- Host verifier pool and returns the ranking/best candidate to MAS execution.



RQ1: Do Different Verifier Types Matter?

- Process Verification shows ***high variance***: it does NOT consistently improve performance.

Method	MAS-Arch	ProVe Level	AIME24			AIME25			GAIA		
			Judge	RM	PRM	Judge	RM	PRM	Judge	RM	PRM
Debate	Fixed	No ProVe		74.13±3.44			61.87±3.21			31.07±2.07	
		Agent	**81.11±1.92**	71.67±1.44	68.10±4.50	**62.97±3.38**	54.45±3.85	57.12±8.85	**30.28±1.51**	25.05±2.33	26.20±0.14
		Iteration	72.67±5.97	**81.11±3.85**	70.34±3.07	53.33±8.82	**57.78±7.70**	54.59±2.18	**32.68±3.11**	20.60±1.21	23.20±4.50
AFlow	Fixed	No ProVe		65.67±5.09			59.72±6.37			11.50±3.25	
		Agent	**76.39±6.36**	76.22±6.59	74.17±3.01	**58.33±4.17**	58.32±11.04	56.94±8.68	**21.68±2.08**	18.08±5.52	14.62±2.03
		Iteration	**69.44±8.67**	52.73±10.52	61.11±2.41	**60.39±6.26**	54.16±5.89	49.99±4.18	18.62±0.78	**21.64±1.64**	13.85±2.55
ADAS	Fixed	No ProVe		62.49±4.19			54.17±4.15			19.30±0.92	
		Agent	73.33±12.83	**73.93±8.26**	68.07±9.64	58.33±0.00	**61.10±2.42**	48.90±2.42	20.90±2.50	**22.10±2.50**	19.70±4.33
		Iteration	**70.83±15.03**	66.67±11.06	63.87±6.36	**61.10±4.85**	59.73±4.79	54.23±0.00	**22.90±2.08**	19.70±1.83	18.53±3.01
MaAS	Adaptive	No ProVe		76.39±6.36			58.33±7.22			20.97±4.90	
		Agent	80.55±8.67	**84.03±6.73**	83.33±4.17	**61.09±4.83**	52.08±2.94	59.70±2.42	**26.89±5.11**	25.36±6.41	24.70±0.86
		Iteration	77.22±10.48	**86.11±6.37**	81.94±6.36	**68.06±6.36**	63.89±2.41	63.88±8.68	**31.21±3.60**	27.29±2.55	24.26±0.97
DyLAN	Adaptive	No ProVe		81.11±5.09			65.55±5.09			19.37±0.95	
		Agent	**85.56±1.93**	76.67±0.00	77.78±3.85	70.00±3.85	**72.22±5.09**	67.78±3.85	18.12±2.02	**19.41±1.68**	16.50±1.37
		Iteration	**84.44±1.93**	78.89±1.92	82.22±10.71	71.11±3.85	**73.33±5.78**	67.77±8.39	**15.50±1.41**	14.83±1.53	9.70±1.37
MAS-Zero	Adaptive	No ProVe		37.78±1.92			35.56±1.93			10.47±0.68	
		Agent	**47.78±1.92**	38.89±1.92	33.34±5.77	**42.22±3.85**	28.87±5.10	20.00±3.85	17.36±0.97	**19.14±1.06**	10.14±0.62
		Iteration	**50.00±3.33**	45.00±2.36	37.78±1.92	**27.78±1.92**	25.56±5.09	25.56±8.39	**17.95±1.11**	14.94±0.62	8.20±0.71

Generative Judges vs Reward Models

- LLM-as-a-Judge outperforms RMs and PRMs in most settings.
  - MAS trajectories are heterogeneous and OOD for scalar reward models.

Method	MAS-Arch	ProVe Level	AIME24			AIME25			GAIA		
			Judge	RM	PRM	Judge	RM	PRM	Judge	RM	PRM
Debate	Fixed	No ProVe		74.13±3.44			61.87±3.21			31.07±2.07	
		Agent	**81.11±1.92**	71.67±1.44	68.10±4.50	**62.97±3.38**	54.45±3.85	57.12±8.85	**30.28±1.51**	25.05±2.33	26.20±0.14
		Iteration	72.67±5.97	**81.11±3.85**	70.34±3.07	53.33±8.82	**57.78±7.70**	54.59±2.18	**32.68±3.11**	20.60±1.21	23.20±4.50
AFlow	Fixed	No ProVe		65.67±5.09			59.72±6.37			11.50±3.25	
		Agent	**76.39±6.36**	76.22±6.59	74.17±3.01	**58.33±4.17**	58.32±11.04	56.94±8.68	**21.68±2.08**	18.08±5.52	14.62±2.03
		Iteration	**69.44±8.67**	52.73±10.52	61.11±2.41	**60.39±6.26**	54.16±5.89	49.99±4.18	18.62±0.78	**21.64±1.64**	13.85±2.55
ADAS	Fixed	No ProVe		62.49±4.19			54.17±4.15			19.30±0.92	
		Agent	73.33±12.83	**73.93±8.26**	68.07±9.64	58.33±0.00	**61.10±2.42**	48.90±2.42	20.90±2.50	**22.10±2.50**	19.70±4.33
		Iteration	**70.83±15.03**	66.67±11.06	63.87±6.36	**61.10±4.85**	59.73±4.79	54.23±0.00	**22.90±2.08**	19.70±1.83	18.53±3.01
MaAS	Adaptive	No ProVe		76.39±6.36			58.33±7.22			20.97±4.90	
		Agent	80.55±8.67	**84.03±6.73**	83.33±4.17	**61.09±4.83**	52.08±2.94	59.70±2.42	**26.89±5.11**	25.36±6.41	24.70±0.86
		Iteration	77.22±10.48	**86.11±6.37**	81.94±6.36	**68.06±6.36**	63.89±2.41	63.88±8.68	**31.21±3.60**	27.29±2.55	24.26±0.97
DyLAN	Adaptive	No ProVe		81.11±5.09			65.55±5.09			19.37±0.95	
		Agent	**85.56±1.93**	76.67±0.00	77.78±3.85	70.00±3.85	**72.22±5.09**	67.78±3.85	18.12±2.02	**19.41±1.68**	16.50±1.37
		Iteration	**84.44±1.93**	78.89±1.92	82.22±10.71	71.11±3.85	**73.33±5.78**	67.77±8.39	**15.50±1.41**	14.83±1.53	9.70±1.37
MAS-Zero	Adaptive	No ProVe		37.78±1.92			35.56±1.93			10.47±0.68	
		Agent	**47.78±1.92**	38.89±1.92	33.34±5.77	**42.22±3.85**	28.87±5.10	20.00±3.85	17.36±0.97	**19.14±1.06**	10.14±0.62
		Iteration	**50.00±3.33**	45.00±2.36	37.78±1.92	**27.78±1.92**	25.56±5.09	25.56±8.39	**17.95±1.11**	14.94±0.62	8.20±0.71

RQ2: At Which Granularity Should We Verify?

- Granularity must align with MAS dynamics.
- It is architecture-sensitive.

Method	MAS-Arch	ProVe Level	AIME24			AIME25			GAIA		
			Judge	RM	PRM	Judge	RM	PRM	Judge	RM	PRM
Debate	Fixed	No ProVe		74.13±3.44			61.87±3.21			31.07±2.07	
		Agent	**81.11±1.92**	71.67±1.44	68.10±4.50	**62.97±3.38**	54.45±3.85	57.12±8.85	**30.28±1.51**	25.05±2.33	26.20±0.14
		Iteration	72.67±5.97	**81.11±3.85**	70.34±3.07	53.33±8.82	**57.78±7.70**	54.59±2.18	**32.68±3.11**	20.60±1.21	23.20±4.50
AFlow	Fixed	No ProVe		65.67±5.09			59.72±6.37			11.50±3.25	
		Agent	**76.39±6.36**	76.22±6.59	74.17±3.01	**58.33±4.17**	58.32±11.04	56.94±8.68	**21.68±2.08**	18.08±5.52	14.62±2.03
		Iteration	**69.44±8.67**	52.73±10.52	61.11±2.41	**60.39±6.26**	54.16±5.89	49.99±4.18	18.62±0.78	**21.64±1.64**	13.85±2.55
ADAS	Fixed	No ProVe		62.49±4.19			54.17±4.15			19.30±0.92	
		Agent	73.33±12.83	**73.93±8.26**	68.07±9.64	58.33±0.00	**61.10±2.42**	48.90±2.42	20.90±2.50	**22.10±2.50**	19.70±4.33
		Iteration	**70.83±15.03**	66.67±11.06	63.87±6.36	**61.10±4.85**	59.73±4.79	54.23±0.00	**22.90±2.08**	19.70±1.83	18.53±3.01
MaAS	Adaptive	No ProVe		76.39±6.36			58.33±7.22			20.97±4.90	
		Agent	80.55±8.67	**84.03±6.73**	83.33±4.17	**61.09±4.83**	52.08±2.94	59.70±2.42	**26.89±5.11**	25.36±6.41	24.70±0.86
		Iteration	77.22±10.48	**86.11±6.37**	81.94±6.36	**68.06±6.36**	63.89±2.41	63.88±8.68	**31.21±3.60**	27.29±2.55	24.26±0.97
DyLAN	Adaptive	No ProVe		81.11±5.09			65.55±5.09			19.37±0.95	
		Agent	**85.56±1.93**	76.67±0.00	77.78±3.85	70.00±3.85	**72.22±5.09**	67.78±3.85	18.12±2.02	**19.41±1.68**	16.50±1.37
		Iteration	**84.44±1.93**	78.89±1.92	82.22±10.71	71.11±3.85	**73.33±5.78**	67.77±8.39	**15.50±1.41**	14.83±1.53	9.70±1.37
MAS-Zero	Adaptive	No ProVe		37.78±1.92			35.56±1.93			10.47±0.68	
		Agent	**47.78±1.92**	38.89±1.92	33.34±5.77	**42.22±3.85**	28.87±5.10	20.00±3.85	17.36±0.97	**19.14±1.06**	10.14±0.62
		Iteration	**50.00±3.33**	45.00±2.36	37.78±1.92	**27.78±1.92**	25.56±5.09	25.56±8.39	**17.95±1.11**	14.94±0.62	8.20±0.71

RQ3: Judge Scaling Behavior

- The gap between small and large ***judges*** is much smaller than the gap between small and large ***solvers***.

Method	AIME24			AIME25			GAIA		
	GPT-4o-mini	GPT-5-mini	FARE-20B	GPT-4o-mini	GPT-5-mini	FARE-20B	GPT-4o-mini	GPT-5-mini	FARE-20B
Debate	72.23±3.87	**81.11±1.92**	76.67±1.65	54.47±6.32	62.97±3.38	**64.43±5.38**	26.85±4.42	**30.28±1.51**	27.87±2.66
AFlow	76.39±8.67	76.39±6.36	**81.94±4.82**	58.33±4.17	58.33±4.17	**61.10±2.42**	19.27±4.17	**21.68±2.08**	16.55±0.85
ADAS	69.45±4.81	**73.33±12.83**	68.07±9.64	63.87±6.36	58.33±0.00	**66.66±4.15**	20.50±2.08	20.90±2.50	**22.50±4.85**
MAAS	83.31±8.30	80.55±8.67	**84.73±6.36**	**62.50±4.17**	61.09±4.83	58.33±4.17	26.50±3.19	26.89±5.11	**28.31±0.85**
DyLAN	83.33±3.34	85.56±1.93	**86.67±3.34**	**73.33±5.78**	70.00±3.85	**73.33±3.34**	**19.37±0.95**	18.12±2.02	18.37±0.98
MAS-Zero	42.22±5.09	47.78±1.92	**51.11±1.92**	30.00±3.33	**42.22±3.85**	38.89±1.92	16.62±1.48	17.36±0.97	**21.15±2.72**

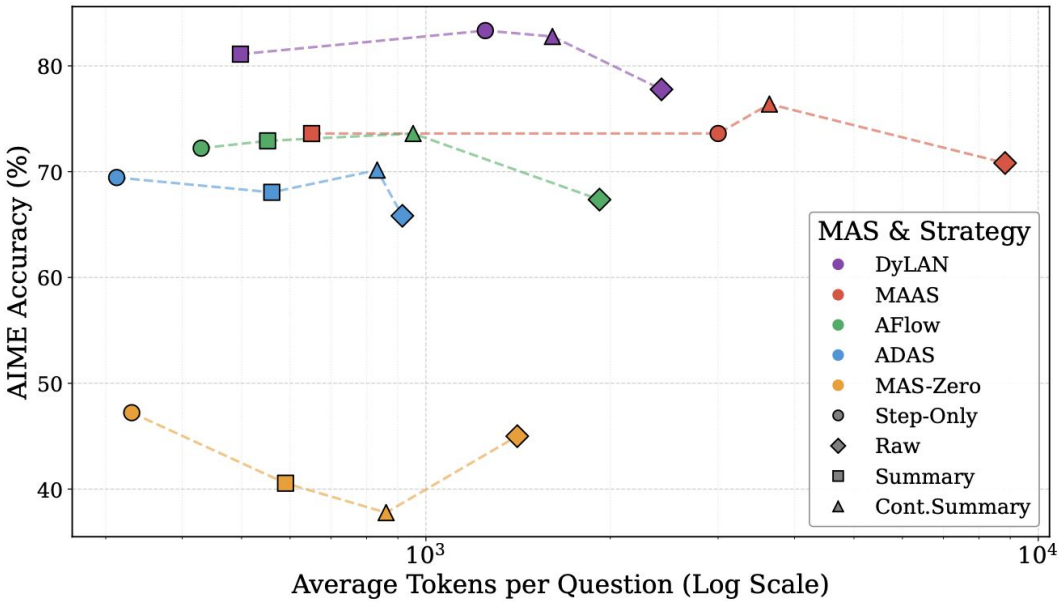
RQ4: How Much Context Should a Judge See?

- Surprisingly:
  Step-only evaluation often performs comparably to full context.

Method	AIME24				AIME25				GAIA		
	Step-Only	Raw	Summary	Cont. Summary	Step-Only	Raw	Summary	Cont. Summary	Step-Only	Raw	Summary
Debate	**81.11**±1.92	**81.11**±1.92	77.78±3.85	-	62.97±3.38	62.97±3.38	**70.00**±6.67	-	**30.28**±1.51	**30.28**±1.51	24.78±0.47
AFlow	79.16±4.17	76.39±6.36	76.39±6.36	**80.55**±4.81	65.28±6.37	58.33±4.17	**69.44**±2.40	66.67±4.17	18.05±2.11	**21.68**±2.08	16.86±3.18
ADAS	72.23±6.37	73.33±12.83	70.83±11.00	**73.63**±12.01	**66.67**±0.00	58.33±0.00	65.27±4.79	**66.67**±0.00	**22.50**±1.83	20.90±2.50	22.10±3.02
MAAS	81.94±2.41	80.55±8.67	**84.72**±4.82	**84.72**±6.37	65.28±13.39	61.09±4.83	62.49±4.19	**68.06**±2.40	24.10±1.70	**26.89**±5.11	23.45±2.62
DyLAN	**86.67**±3.34	85.56±1.93	82.22±5.09	**86.67**±3.34	**80.00**±6.67	70.00±3.85	**80.00**±0.00	78.89±5.09	**20.67**±1.10	18.12±2.02	14.10±2.56
MAS-Zero	48.88±5.08	47.78±1.92	**48.89**±1.92	42.22±5.09	**45.56**±10.18	42.22±3.85	32.22±5.09	33.33±3.34	17.28±1.39	**17.36**±0.97	15.07±0.54

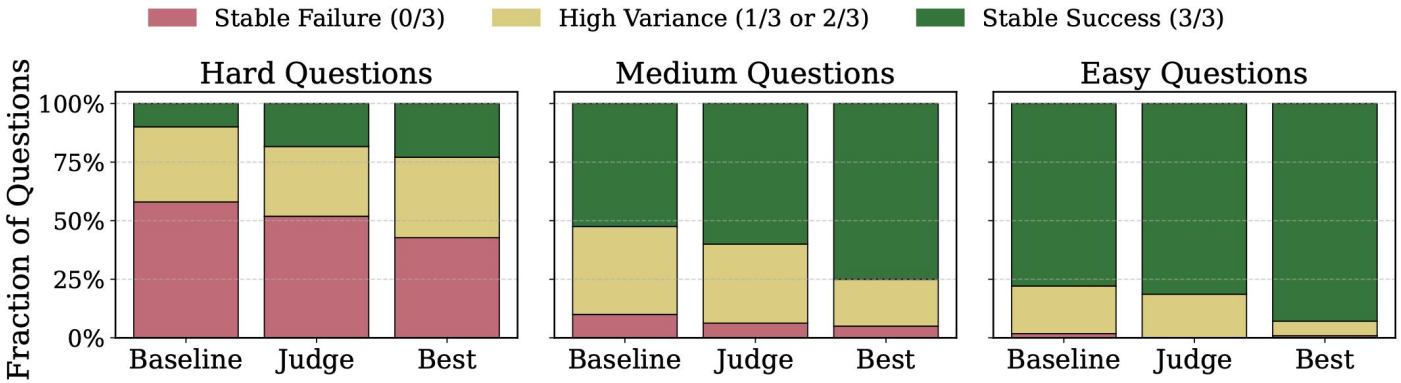
Summarized Context Wins for Long MAS

- ***Context-length*** vs ***performance*** trade-off is real.
- For long-trajectory systems like MaAS and DyLAN:
  - Context summarization ***Improves accuracy*** and ***reduces token cost*** dramatically



RQ5: Does Verification Solve Hard Problems?

- We stratify AIME problems into three difficulties:
  - Easy/Medium/Hard
- Process verification mainly:
  - Reduces variance & converts unstable successes into stable ones;
- But rarely:
  - Resurrects fundamentally unsolvable problems.



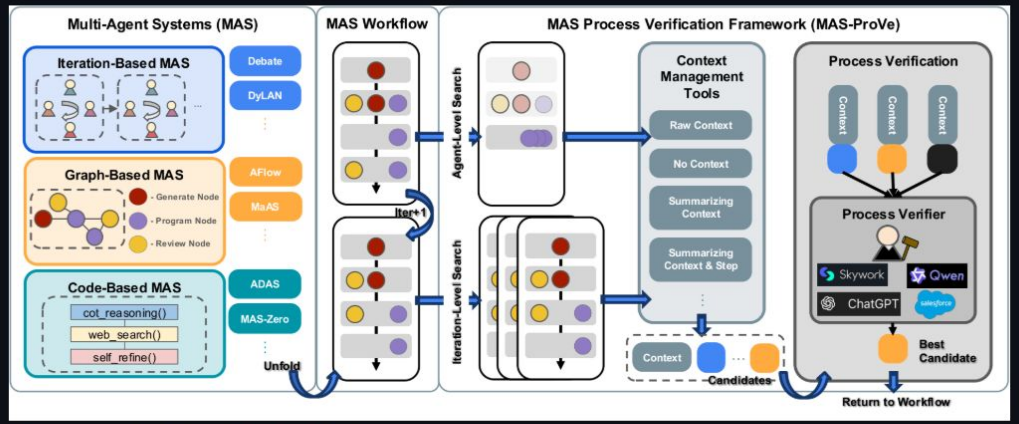
Overall Conclusions

- Process verification is NOT a universal solution for scaling test-time compute of MAS.
- It improves MAS stability more than capability.
- In MAS, generative judges outperform scalar reward models.

- There are many open challenges (no decisive conclusions) that remains to be resolved in process verification for MAS.

Open-Source Codebase

Library of Multi-Agent Systems Process Verification (MAS-ProVe)



This repo contains the code for the following work:

- **MAS-ProVe: Understanding the Process Verification of Multi-Agent Systems**
  Vishal Venkataramani, Haizhou Shi, Zixuan Ke, Austin Xu, Xiaoxiao He, Yingbo Zhou, Semih Yavuz, Hao Wang, Shafiq Joty
  arXiv 2602.03053

https://github.com/Wang-ML-Lab/MAS-ProVe