Enhancing Operational Safety via Agentic Dialogue Hazard Identification Analysis

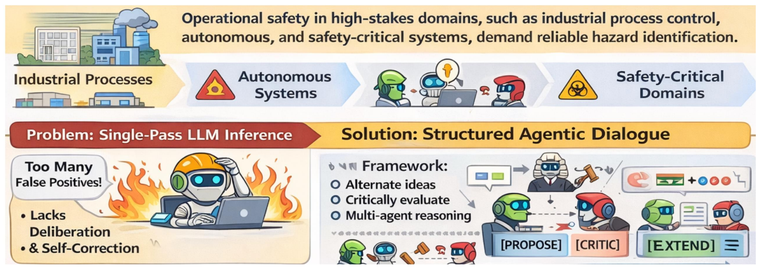
Sanjay Das, Ran Elgedawy, Ethan Seefried, Ryan Burchfield, Tirthankar Ghosal

MOTIVATION

Operational safety analysis is crucial in ensuring risk mitigation and avoidance. This aids in:

- Personnel safety during risky work.
- Avoiding operational and financial damage.
- Building a safer environment for critical work in ORNL and DOE sites.

HAZDIAL



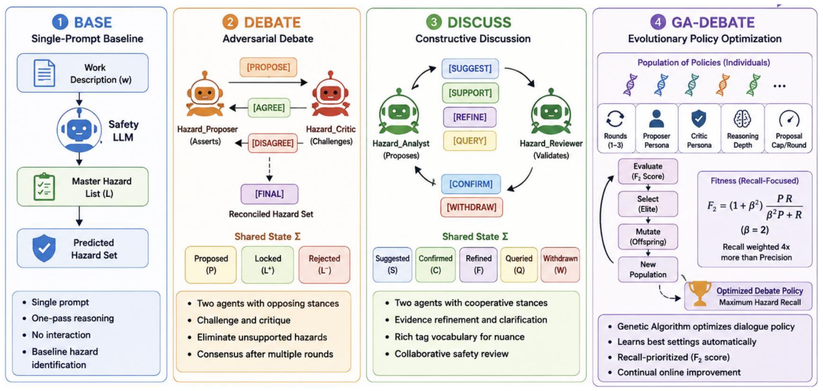
This work investigates if agentic dialogue improves operational safety using a closed loop work-scope based hazard identification task. It asks:

- Does agentic deliberation enhance hazard identification performance over simpler prompting baseline?
- Does dialogue modalities impact hazard identification fidelity in terms of precision, recall and f1 score?
- Is the improvement statistically significant?
- How to find the optimal agentic dialogue orchestration?

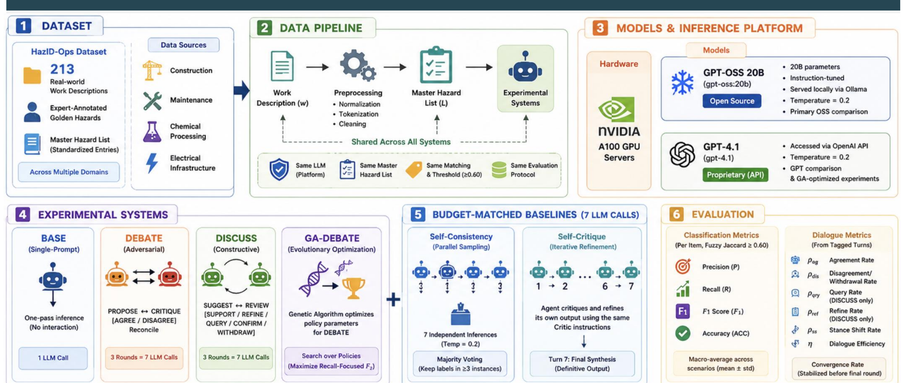
SYSTEM COMPONENTS

We investigate four system configurations, progressing from a non-interactive baseline to increasingly structured multi-agent dialogue:

- System 1: Single-Prompt Baseline (BASE)
- System 2: Adversarial Debate (DEBATE)
- System 3: Constructive Discussion (DISCUSS)
- System 4: Evolutionary Policy Optimization (GA-DEBATE)



EXPERIMENTS



The experiments are set up as follows:

- We use a curated internal dataset of 213 work orders and hazards.
- The data is pre-processed to build work scopes and a master hazard list.
- We use the open source GPT-OSS 20B and proprietary GPT-4.1.
- Self-consistency and self-critique analysis is performed for statistical significance testing.

RESULTS

- DEBATE improves F1 over BASE on both models (statistically significant on OSS).
- DISCUSS achieves the highest recall on GPT-4.1, indicating stronger hazard coverage.
- GA-DEBATE further improves F1 and Accuracy over GPT-4.1 DEBATE.
- Dialogue metrics show GA-DEBATE is more agreement-oriented, efficient, and more stable in stance.
- GA converges to a policy that uses 3 rounds, a thorough proposer and strict critic with medium reasoning depth and large proposal cap.

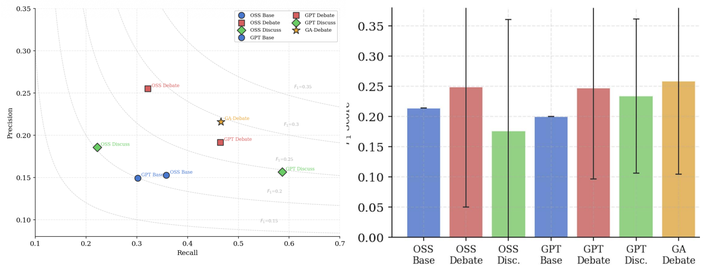


Fig.: Significance test (GPT-OSS 20B)

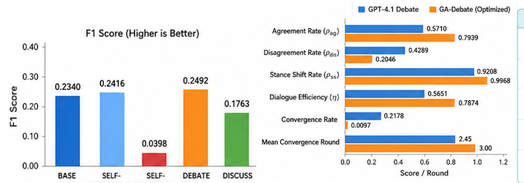


Fig.: Dialogue metrics (GPT-4.1)

Parameter	Converged Value
θ_1: Number of Rounds	3
θ_2: Proposer Persona	Thorough
θ_3: Critic Persona	Strict
θ_4: Reasoning Depth	Medium
θ_5: Proposal Cap (per round)	20

Fig.: GA convergence (GPT-4.1)

CONCLUSION

This work presented a systematic investigation of agentic dialogue for operational hazard identification from a closed label list. We evaluated three base configurations across two language models and proposed an evolutionary policy optimization framework that learns the optimal agentic configuration from prediction feedback. The results demonstrates the gains of an agentic dialogue system on hazard identification performance.

Acknowledgement: This research used resources of the Oak Ridge Leadership Computing Facility (OLCF), which is a DOE Office of Science User Facility at the Oak Ridge National Laboratory supported by the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. We thank ORNL Operations team.

OAK RIDGE National Laboratory

U.S. DEPARTMENT OF ENERGY

SIGdial