Enhancing Operational Safety via Agentic Dialogue Hazard Identification Analysis

Sanjay Das
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
dass3@ornl.gov

Ran Elgedawy
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
elgedawyr@ornl.gov

Ethan Seefried
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
seefriedej@ornl.gov

Ryan Burchfield
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
burchfieldra@ornl.gov

Tirthankar Ghosal
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA
ghosalt@ornl.gov

Abstract

Operational safety in high-stakes domains—such as industrial process control, autonomous, and safety-critical systems—demand reliable hazard identification. While large language models (LLMs) have shown promise in automating safety analysis tasks, single-turn, monolithic inference is brittle: it lacks the self-correction, deliberation, and contextual refinement that safety engineers apply iteratively. In this paper, we introduce HAZDIAL, a framework that investigates whether structured *agentic dialogue*—multi-agent, multi-turn interactions—improves the quality of NLP-based hazard identification over single-pass baselines. We systematically compare two dialogue modalities: *adversarial debate* and *constructive discussion*, and propose an algorithm-based agentic interaction optimization. We evaluate all configurations against a curated golden dataset using standard classification metrics (accuracy, precision, recall, F_1) and novel dialogue metrics.

Evaluations across 213 operational text descriptions demonstrate that adversarial debate reduces false positives by 40% and increases F_1 score by 2.3% over single-pass inference, with evolutionary policy optimization yielding an additional 1.9% gain. Conversely, a budget-matched self-critique baseline performance collapses to $F_1 = 0.040$, establishing that agent independence—rather than computational overhead—drives these multi-agent performance gains. This work advances the intersection of dialogue systems, multi-agent reasoning, and AI safety, providing an empirical evidence for dialogue-driven hazard analysis. real-world operational work descriptions.

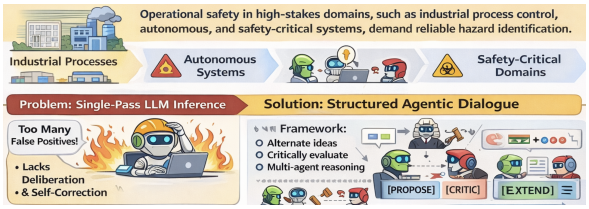

Figure 1: Dialogue-driven hazard analysis.

1 Introduction

Operational safety analysis is a cornerstone of engineering practice. Processes such as Hazard and Operability Studies (HAZOP) (Kletz, 1999), Failure Mode and Effects Analysis (FMEA) (Slack, 2015), and Fault Tree Analysis (FTA) (Vesely et al., 1987) require safety engineers to systematically enumerate potential hazards, failure modes, their causes, and their consequences. These analyses are labor-intensive, domain-specific, and prone to human errors, fatigue, making them compelling targets for NLP-assisted automation as shown in Figure 1 (Brown et al., 2020; Wei et al., 2022).

Recent work has begun to apply LLMs to safety-related text classification and extraction tasks (Paltrinieri et al., 2019; Rajpurkar et al., 2018), but a persistent challenge remains: a single-pass LLM inference is epistemically flat. It cannot challenge its own assumptions, incorporate counterfactual reasoning, or refine outputs through deliberation—all of which safety engineers do naturally through structured reviews and peer challenges.

Why agentic dialogue? Multi-agent dialogue systems introduce structured deliberation into the inference process. When two or more agents

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 757–772
August 2–5, 2026.

interact—whether through cooperative discussion or adversarial debate—the resulting output benefits from error correction, coverage expansion, and iterative refinement (Liang et al., 2023; Du et al., 2023). Yet it remains unclear: (1) whether dialogue *actually* improves hazard identification quality, and (2) what *type* of dialogue structure and *optimization strategy* yields the greatest gains.

The gap we address. Prior work on multi-agent debate (Du et al., 2023; Chan et al., 2023) and collaborative reasoning (Zheng et al., 2023) has focused primarily on factual QA, mathematical reasoning, and commonsense tasks. Safety-critical hazard identification introduces distinct challenges: outputs must be *comprehensive* (high recall over hazard space), *precise* (low false-positive rate to avoid alarm fatigue), and *explainable* (agents must justify their claims with engineering evidence). Furthermore, the optimization of the *agentic configuration itself*—which agent speaks when, in what order, with what prompt—has received little rigorous attention. We make the following contributions:

1. We formulate hazard identification as a *closed-list* task and define the *Hazard Identification Dialogue* (HAZDIAL) framework, specifying agents, shared mutable state, dialogue tags, and a deterministic aggregation function.

2. We design and implement two structurally distinct multi-agent configurations DEBATE (adversarial Proposer/Critic) and DISCUSS (cooperative Analyst/Reviewer) that prevents circular argumentation and enforces evidence-grounded verdicts to study of how dialogue structure affects hazard identification.

3. We propose a genetic policy optimization algorithm that treats the agentic configuration parameters as learnable genes, using a recall-prioritized F_β ($\beta = 2$) fitness signal to evolve configurations that minimize missed hazards across successive work descriptions.

4. Experiments with GPT-OSS 20B and GPT-4.1 demonstrate that adversarial debate consistently reduces false positives by up to 40% across both models, while constructive discussion degrades F_1 on the smaller model but achieves the highest recall (0.586) on GPT-4.1, showing that cooperative dialogue may helps in certain scenarios.

2 Background and Related Work

2.1 Hazard Identification

Hazard identification (HazID) is the first and most critical phase of risk assessment (Bootle, 2005; Debouk, 2019). Classical methods—HAZOP, FMEA, STPA (Leveson, 2012)—rely on structured checklists and guidewords (*NO, MORE, LESS, AS WELL AS, REVERSE, OTHER THAN*) applied systematically to process deviations. NLP-based approaches have begun automating aspects of this process: extracting safety-relevant entities (Paltrinieri et al., 2019), classifying incident reports, and generating FMEA entries from process descriptions (Thomas, 2023). However, these approaches treat HazID as a static extraction task, ignoring the iterative, collaborative nature of real-world safety reviews.

2.2 LLMs for Safety Analysis

Brown et al. (2020) and subsequent work have demonstrated broad capability in NLP tasks with minimal supervision. Wei et al. (2022) introduced chain-of-thought prompting, enabling step-by-step reasoning that is beneficial for complex safety reasoning. Kojima et al. (2022) showed that zero-shot CoT achieves competitive performance, while Yao et al. (2023) extended this to tree-structured deliberation. Despite these advances, LLMs remain prone to hallucination and overconfidence in safety contexts (Ji et al., 2022), motivating dialogue-based verification.

2.3 Multi-Agent Dialogue and Debate

Du et al. (2023) demonstrated that multi-agent debate—where models critique and revise each other's outputs over multiple rounds—improves factual accuracy and reasoning quality on benchmarks. Liang et al. (2023) showed that diversity of perspectives in multi-agent setups reduces groupthink. Chan et al. (2023) proposed ChatEval for reference-free dialogue evaluation using multi-agent frameworks. Xiong et al. (2023) examined LLM consistency under self-contradiction. In parallel, Park et al. (2023) explored emergent social behaviors in LLM agent societies. Our work extends this line by (a) applying it to safety-critical HazID, and (b) systematically optimizing the dialogue *configuration itself* through genetic evolutionary optimization.

2.4 Dialogue Optimization and Policy Learning

Policy gradient methods for dialogue (Li et al., 2016; Williams, 2004) and reward shaping (Ziegler et al., 2019) have been applied to task-oriented systems. Our optimization component adapts these ideas to a lightweight, prompt-level policy without full Reinforcement Learning, making it computationally tractable for deployment. Evolutionary approaches to prompt optimization (Guo et al., 2024; Das et al., 2024) have shown promise and we extend this to the level of entire agent call sequences.

3 Problem Formulation

3.1 Hazard Identification

Let $\mathcal{W}$ denote a corpus of natural-language work descriptions drawn from operational safety records. A hazard identification system $\mathcal{F}$ takes each description w and the closed master hazard list $\mathcal{L}$ as input and returns a predicted subset $\hat{\mathcal{H}}(w) \subseteq \mathcal{L}$.

This framing mirrors real engineering practice: safety analysts consult a pre-approved hazard taxonomy and decide which entries apply to the work at hand. Constraining predictions to $\mathcal{L}$ eliminates unconstrained generation noise and makes evaluation deterministic, while the selection task remains non-trivial because $|\mathcal{L}|$ is large and the evidence for each hazard must be inferred from unstructured text.

3.2 Dialogue Modalities

We study two dialogue modes:

Adversarial Debate (AD). Two agents take opposing epistemic stances. Agent a_1 (Proposer) identifies hazards; agent a_2 (Challenger) critiques and seeks to identify errors, missed hazards, or over-classified events. The Challenger's objective is to *find faults, and critique* the Proposer.

Constructive Discussion (CD). Agents collaboratively elaborate hazard coverage from complementary perspectives (e.g., Analyst, Reviewer). Each agent extends, refines, or validates prior contributions without an adversarial objective.

3.3 Dialogue-Augmented Hazard Identification

Definition 1 (Hazard Identification Dialogue)**.** *A* Hazard Identification Dialogue *(HID) is a tuple* $\langle \mathcal{A}, T, \Sigma, \Omega \rangle$ *where:*

- $\mathcal{A} = \{a_1, a_2\}$ *is a pair of LLM-backed agents, each assigned a distinct epistemic role;*
- $T = (t_1, t_2, \ldots, t_N)$ *is an ordered sequence of dialogue turns, with* $N = 2R$ *for* R *rounds;*
- Σ *is a shared mutable* dialogue state *recording which hazards have been proposed, accepted, rejected, or withdrawn;*
- $\Omega : \Sigma \to \hat{\mathcal{H}}$ *is a deterministic aggregation function that reads the final state to produce the predicted hazard set.*

At each turn t_n, the active agent $a_{i(n)}$ generates a response r_n conditioned on the work description w, the master list $\mathcal{L}$, and the accumulated dialogue history $\mathbf{h}_n = (r_1, \ldots, r_{n-1})$:

$$r_n = a_{i(n)}\big(\pi_{i(n)},\ w,\ \mathcal{L},\ \mathbf{h}_n\big) \tag{1}$$

where $\pi_{i(n)}$ is the role-specific system prompt for the active agent. After R rounds, Ω reads the final state Σ_R to yield $\hat{\mathcal{H}}$.

4 Proposed Systems

We investigate four system configurations, progressing from a non-interactive baseline to increasingly structured multi-agent dialogue. All three share the same underlying LLM and the same master hazard list.

4.1 System 1: Single-Prompt Baseline (BASE)

Prior to introducing dialogue, it is essential to establish the baseline capabilities of a competent, single-pass LLM. This baseline represents the most straightforward architecture, consisting of a single agent executing a single turn without deliberation.

Formulation. Given work description w and master list $\mathcal{L}$, the agent produces a predicted hazard set in a single forward pass:

$$\hat{\mathcal{H}}^{\text{BASE}} = \text{Parse}(a_1(\pi^{\text{BASE}},\ w,\ \mathcal{L})) \tag{2}$$

The system prompt π^{BASE} instructs the model to act as a safety expert and return only hazard labels present verbatim in $\mathcal{L}$.

4.2 System 2: Adversarial Debate (DEBATE)

Prior work has demonstrated that adversarial multi-agent debate improves factual accuracy on open-domain tasks (Du et al., 2023; Liang et al., 2023); we test whether the same mechanism transfers to the constrained, evidence-bound domain of hazard identification.

Agent roles and dialogue tags. DEBATE employs two agents with opposing epistemic stances:

- **Hazard_Proposer** selects hazards from $\mathcal{L}$ that are directly evidenced by w and annotates each with a textual justification. Each proposal is tagged `[PROPOSE]`.
- **Hazard_Critic** challenges the proposed list, tagging each verdict as either `[AGREE]` (hazard retained) or `[DISAGREE]` (hazard eliminated), with a supporting reason.

After R rounds of exchange, the Proposer issues a final reconciliation pass, outputting the agreed-upon set using `[FINAL]` tags. Full prompt templates are provided in Appendix C.

State system. The shared state Σ tracks three disjoint sets across rounds:

$$\Sigma = \left(\mathcal{P},\ \mathcal{L}^{+},\ \mathcal{L}^{-}\right) \tag{3}$$

where $\mathcal{P}$ is the set of all proposed hazard labels, $\mathcal{L}^{+}$ ("locked") contains Critic-agreed hazards, and $\mathcal{L}^{-}$ ("rejected") contains Critic-disagreed hazards. Previously decided hazards are excluded from subsequent Proposer turns to prevent circular arguments.

4.3 System 3: Constructive Discussion (DISCUSS)

We design DISCUSS to model a collaborative safety review pattern, testing whether cooperative elaboration produces qualitatively different outcomes to adversarial challenge.

Agent roles and dialogue tags. DISCUSS employs two cooperatively-oriented agents:

- **Hazard_Analyst** proposes hazards, tagged `[SUGGEST]`, with evidence from w. When queried, the Analyst either confirms a hazard (`[CONFIRM]`) with direct textual evidence or retracts it (`[WITHDRAW]`) when challenged.
- **Hazard_Reviewer** validates proposals cooperatively: it supports well-evidenced hazards (`[SUPPORT]`), refines imprecise labels to more specific entries in $\mathcal{L}$ (`[REFINE]`), or raises a targeted clarification request (`[QUERY]`) before committing.

This richer tag vocabulary as listed in Table 1 captures qualitatively different dialogue behaviors: incremental refinement, epistemic uncertainty, and evidence-contingent commitment.

Table 1: Dialogue tag vocabulary by system. "Epistemic stance" indicates whether the tag conveys agreement, disagreement, or epistemic suspension.

System	Tag	Agent	Stance
DEBATE	`[PROPOSE]`	Proposer	Assertion
	`[AGREE]`	Critic	Positive
	`[DISAGREE]`	Critic	Negative
DISCUSS	`[SUGGEST]`	Analyst	Assertion
	`[SUPPORT]`	Reviewer	Positive
	`[REFINE]`	Reviewer	Positive+
	`[QUERY]`	Reviewer	Suspended
	`[CONFIRM]`	Analyst	Positive
	`[WITHDRAW]`	Analyst	Negative

State system. The shared state extends Equation (3) with two additional sets:

$$\Sigma = \left(\mathcal{S},\ \mathcal{C},\ \mathcal{F},\ \mathcal{Q},\ \mathcal{W}\right) \tag{4}$$

where $\mathcal{S}$ is the set of suggested hazards, $\mathcal{C}$ the confirmed set (contributed by `[SUPPORT]`, `[REFINE]`, or `[CONFIRM]`), $\mathcal{F}$ the set of refinement substitutions $\{(h_{\text{orig}}, h_{\text{refined}})\}$, $\mathcal{Q}$ the queried set, and $\mathcal{W}$ the withdrawn set.

4.4 System 4: Evolutionary Policy Optimization (GA-DEBATE)

Both previous dialogue systems expose a discrete configuration space: how many rounds to run, how aggressively the Proposer should generate, how strictly the Critic should filter, and at what level of chain-of-thought verbosity. Manually tuning these parameters is impractical at scale. We propose a *Genetic Algorithm* (GA) that learns the optimal configuration by treating each parameter assignment as an individual and using prediction performance as fitness.

Learnable parameters. Five parameters constitute an individual θ:

Param	Description	Values
θ_1	Number of rounds	$\{1, 2, 3\}$
θ_2	Proposer persona	Thorough, Focused, Cautious
θ_3	Critic persona	Strict, Balanced, Skeptical
θ_4	Reasoning depth	high, medium, low
θ_5	Proposal cap/round	$\{5, 10, 20\}$

Table 2: GA parameter space Θ.

Fitness function. Given the asymmetric cost of missed hazards in safety-critical contexts, where a false negative (missed hazard) can lead to injury or incident while a false alarm (false positive) merely

Algorithm 1 Evolutionary Policy Optimization for Agentic Hazard Identification

Require: $\mathcal{W} = \{w_1, \ldots, w_N\}$, $\{\mathcal{H}^*(w_i)\}$, Θ, $N_{\text{init}}, K_{\text{elite}}, K_{\text{offs}}$, mutation rate p_{mu}
Ensure: Best individual $\boldsymbol{\theta}^*$ and fitness history $\mathcal{A}$

1: $\mathcal{E} \leftarrow \emptyset$ {Evaluated individuals archive}
2: $\mathcal{A} \leftarrow \emptyset$ {Fitness trajectory log (all sessions)}
3: $g \leftarrow 0$ {Generation counter}
4: **for** $i = 1$ **to** N_{init} **do**
5: $\quad \boldsymbol{\theta}_i \leftarrow \textsc{RandomSample}(\Theta)$
6: $\quad \hat{\mathcal{H}}_i \leftarrow \textsc{RunDebate}(w_i, \mathcal{L}, \boldsymbol{\theta}_i)$
7: $\quad f_i \leftarrow \textsc{Fitness}(\hat{\mathcal{H}}_i, \mathcal{H}^*(w_i))$
8: $\quad \mathcal{E} \leftarrow \mathcal{E} \cup \{(\boldsymbol{\theta}_i, f_i)\}$, $\mathcal{A} \leftarrow \mathcal{A} \cup \{(\boldsymbol{\theta}_i, f_i, g)\}$
9: **end for**
10: $\mathcal{P} \leftarrow \textsc{BuildPool}(\mathcal{E}, K_{\text{elite}}, K_{\text{offs}}, p_{\text{mu}})$
11: $j \leftarrow 0$ {Pool pointer}
12: **for** $i = N_{\text{init}} + 1$ **to** N **do**
13: $\quad \boldsymbol{\theta}^{\text{cur}} \leftarrow \mathcal{P}[j \bmod |\mathcal{P}|]$
14: $\quad \hat{\mathcal{H}}_i \leftarrow \textsc{RunDebate}(w_i, \mathcal{L}, \boldsymbol{\theta}^{\text{cur}})$
15: $\quad f_i \leftarrow \textsc{Fitness}(\hat{\mathcal{H}}_i, \mathcal{H}^*(w_i))$
16: $\quad$ Update mean fitness of $\boldsymbol{\theta}^{\text{cur}}$ in $\mathcal{E}$; $\mathcal{A} \leftarrow \mathcal{A} \cup \{(\boldsymbol{\theta}^{\text{cur}}, f_i, g)\}$
17: $\quad j \leftarrow j + 1$
18: $\quad$ **if** $j \bmod |\mathcal{P}| = 0$ **then**
19: $\quad\quad \mathcal{P} \leftarrow \textsc{BuildPool}(\mathcal{E}, K_{\text{elite}}, K_{\text{offs}}, p_{\text{mu}})$
20: $\quad\quad g \leftarrow g + 1$; $j \leftarrow 0$
21: $\quad$ **end if**
22: **end for**
23: $\boldsymbol{\theta}^* \leftarrow \arg\max_{\boldsymbol{\theta} \in \mathcal{E}} \overline{f}(\boldsymbol{\theta})$
24: **return** $\boldsymbol{\theta}^*, \mathcal{A}$

requires investigation, we use the F_β score with $\beta = 2$ as the fitness signal (Lee et al., 2021):

$$\text{Fit}(\hat{\mathcal{H}}, \mathcal{H}^*) = \frac{(1+\beta^2) \cdot P \cdot R}{\beta^2 \cdot P + R}, \quad \beta = 2 \quad (5)$$

This weights recall four times more heavily than precision in the harmonic mean, incentivising the GA to converge on configurations that minimize false negatives even at some precision cost.

GA lifecycle. Algorithm 1 formalizes the full optimization procedure. The algorithm operates in two phases and treats each work description as a single fitness evaluation, enabling continual online learning.

Phase 1 — Exploration. The first N_{init} data points are each evaluated with a distinct randomly sampled individual, covering the parameter space before any selection occurs (*lines 1-11*).

Phase 2 — Evolution. After each pool of $K_{\text{elite}} + K_{\text{offs}}$ evaluated individuals, they are ranked by mean fitness. Then top K_{elite} as the elite set are retained. Next K_{offs} offspring are generated by applying gene-wise mutation (rate p_{mu}) to randomly selected elite parents to form the new active pool as elite $\cup$ offspring. This process is continued until all data points are exhausted.

5 Experimental Setup

5.1 Dataset

We evaluate on the **HazID-Ops** dataset comprising 213 real-world operational work descriptions drawn from safety management systems across construction, maintenance, chemical processing, and electrical infrastructure domains. Each description is paired with expert-annotated golden hazard labels drawn from a master list of standardised entries compiled from domain safety taxonomies.

5.2 Models and Inference Platform

All experiments were conducted on NVIDIA A100 GPU servers. Two language model configurations are evaluated:

- **GPT-OSS 20B** (`gpt-oss:20b`): an open-source 20B-parameter instruction-tuned model served locally via Ollama with temperature 0.2. Used for the primary OSS comparison (Agarwal et al., 2025).
- **GPT-4.1** (`gpt-4.1`): accessed via the OpenAI API with temperature 0.2 (Wang and Lim, 2025). Used for the GPT comparison and GA-optimized experiments.

5.2.1 Budget-Matched Baselines.

To decouple the empirical contributions of *inter-agent dialogue* from the confounding effects of increased computational overhead, we implement two single-agent baselines calibrated to match the exact inference budget of DEBATE (7 LLM invocations across $R = 3$ deliberation rounds):

- **Self-Consistency** (Wang et al., 2022): The baseline configuration executes 7 independent parallel inferences using the primary prompt at a temperature of 0.2. Hazard classifications are aggregated via majority voting, where labels appearing in ≥ 3 instances are retained. This evaluates whether standard sampling-based output aggregation replicates the performance of dialogue-structured deliberation.

- **Self-Critique**: A isolated agent generates an initial hazard portfolio (Turn 1), then iteratively critiques and refines its own output across sequential steps (Turns 2–6) using the identical Critic system instructions utilized in DEBATE. A final execution (Turn 7) synthesizes the definitive output list. This baseline isolates the structural utility of agent *independence* from the iterative advantages of multi-turn reasoning.

5.3 Evaluation Metrics

Classification metrics. In label matching we use fuzzy Jaccard similarity (Li et al., 2021) (≥ 0.60 threshold) over normalized word tokens, tolerating trivial surface variants (plurals, verb inflections, hyphenation differences). We report precision, recall, F1 score and accuracy (Goutte and Gaussier, 2005). Macro-averages across scenarios are reported with standard deviations.

Dialogue metrics. Derived from tagged turn records (Section 4):

- ρ_{ag}: **Agreement Rate** — fraction of proposals receiving a positive verdict (`[AGREE]`, `[SUPPORT]`, `[CONFIRM]`, or `[REFINE]`).
- ρ_{dis}: **Disagreement/Withdrawal Rate** — fraction of proposals rejected or withdrawn.
- ρ_{qry}: **Query Rate** — fraction of suggestions queried before commitment (DISCUSS only).
- ρ_{ref}: **Refine Rate** — fraction of suggestions replaced by a more precise label (DISCUSS only).
- ρ_{ss}: **Stance Shift Rate** — fraction of rounds in which the confirmed/locked set changed.
- η: **Dialogue Efficiency** — ratio of final confirmed hazards to total proposals.
- **Convergence Rate**: fraction of scenarios where the confirmed set stabilized before the final round.

These metrics are defined in Appendix A.

Statistical significance. Pairwise F_1 differences are tested with a paired bootstrap resampling procedure (efr, 2007) ($B = 10{,}000$ resamples, $\alpha = 0.05$, one-sided). We report the observed ΔF_1, 95% bootstrap confidence interval, and p-value for all primary comparisons. All per-item F_1 scores use the fuzzy Jaccard matching defined in §5.3.

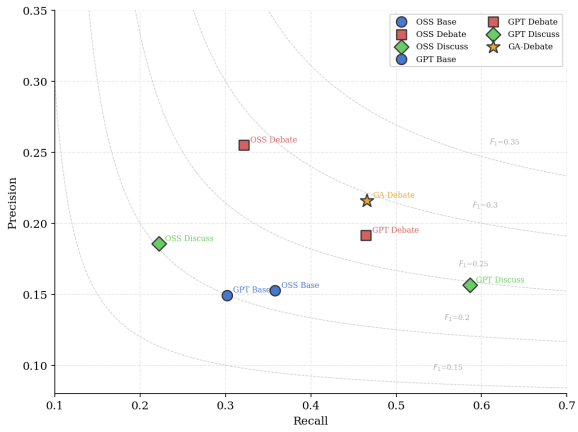

Figure 2: Precision vs. recall for all systems.

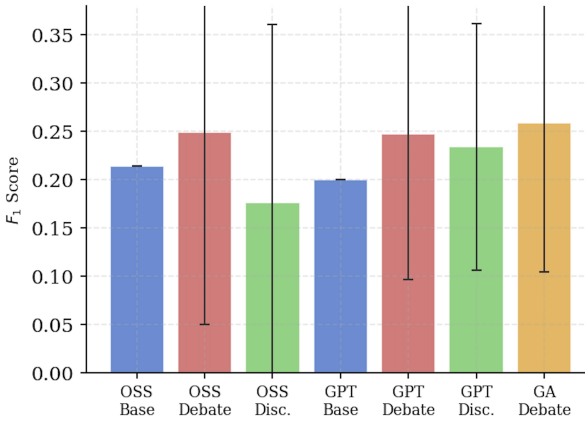

Figure 3: F_1 score for all systems.

6 Results

6.1 OSS 20B Model: Base, Debate, and Discuss

Table 3 reports classification metrics on the GPT-OSS 20B model across 213 scenarios. Table 4 presents the corresponding dialogue metrics.

Finding O1: Adversarial debate is the best-performing OSS configuration. DEBATE improves F_1 by $+0.035$ over BASE (0.2492 vs. 0.2140) and reduces the corpus-level FP count by 40.0% (from 1,761 to 1,057). The improvement is driven entirely by precision ($+0.102$), with a modest recall loss (-0.037) attributable to the Critic's evidence-grounded rejection defaults.

Table 3: Classification metrics — GPT-OSS 20B.

System	Precision	Recall	F_1	Accuracy
BASE	0.1526	0.3582	0.2140	0.1198
DEBATE	0.2550±0.2454	0.3216±0.2733	**0.2492**±0.1992	**0.1589**±0.1485
DISCUSS	0.1856±0.2299	0.2218±0.2524	0.1763±0.1841	0.1089±0.1216

Table 4: Dialogue metrics — GPT-OSS 20B.

Metric	DEBATE	DISCUSS
Agreement Rate ρ_{ag}	0.5988±0.2267	0.6055±0.2471
Disagreement/Withdrawal Rate ρ_{dis}	0.4012±0.2267	0.8998±0.2572
Query Rate ρ_{qry}	—	0.4323±0.2487
Refine Rate ρ_{ref}	—	0.2020±0.1810
Stance Shift Rate ρ_{ss}	0.7715±0.2682	0.7121±0.2899
Dialogue Efficiency η	0.5918±0.2269	0.5587±0.2368
Convergence Rate	0.4883	0.5915
Mean Convergence Round	2.37±0.57	2.23±0.67

Finding O2: Constructive discussion underperforms the base model. Despite a 43.8% FP reduction, DISCUSS posts $F_1 = 0.1763$, below both BASE and DEBATE. The withdrawal rate of $\rho_{\text{dis}} = 0.8998$, meaning that 90% of queried hazards are subsequently retracted by the Analyst. The query-withdraw mechanism eliminates true positives alongside false positives, producing a net recall loss of -0.137 relative to BASE.

6.2 Budget-Matched Baselines

Table 5 presents all five system configurations on the OSS 20B with statistical significance testing.

Finding B1: Agent independence is the critical factor. DEBATE is the only budget-matched system to significantly outperform BASE. SELF-CONSISTENCY does not ($\Delta F_1 = +0.008$, $p = 0.167$), confirming that simply repeating the base prompt seven times does not replicate dialogue's benefit.

Finding B2: Self-critique catastrophically degrades performance. SELF-CRITIQUE achieves $F_1 = 0.040$—a reduction of 19.3% below BASE ($p < 0.001$, $\Delta F_1 = -0.193$). This result is unambiguous: when a single agent is instructed to critique its own output using the same adversarial Critic prompt designed for an independent second agent, performance collapses. The Critic's aggressive rejection defaults, calibrated for challenging another agent's proposals, systematically eliminate the agent's own well-reasoned suggestions. DEBATE outperforms SELF-CRITIQUE by $\Delta F_1 = +0.209$ ($p < 0.001$), the largest and most significant gap in the entire study.

This finding shows that the gain from DEBATE comes specifically from *having an independent second agent* whose epistemic state is separate from the Proposer's, not from using more inference calls.

Table 5: Classification metrics and significance tests on GPT-OSS 20B.

System	F_1	ΔF_1	95% CI	p-value	LLM Calls
BASE	0.2340	—	—	—	1
SELF-CONSISTENCY	0.2416	+0.0076	[−0.008, +0.023]	0.167	7
SELF-CRITIQUE	0.0398	−0.1926	[−0.217, −0.168]	1.000	7
DEBATE	**0.2492**	+0.0152	[+0.001, +0.046]	**0.027**	7
DISCUSS	0.1763	−0.0577	[−0.090, −0.030]	1.000	≤9

6.3 GPT-4.1 Model: Base, Debate, and Discuss

Table 6 reports classification metrics for GPT-4.1. Table 7 presents the dialogue metrics.

Finding G1: Both dialogue modes substantially improve recall on GPT-4.1. DEBATE raises recall from 0.3018 to 0.4644 ($\Delta = +0.163$), achieving $F_1 = 0.2473$. DISCUSS achieves the highest recall observed across all experiments (0.5864)($\Delta = -0.048$), though at a FP:TP ratio of 5.31, which is only marginally better than the base model's 5.70.

Finding G2: GPT-4.1 reverses the OSS behaviour of DISCUSS. On GPT-OSS 20B, DISCUSS underperformed BASE due to over-pruning ($\rho_{\text{dis}} = 0.90$). On GPT-4.1, the same architecture achieves the highest recall of any system, with a far lower withdrawal rate (0.5708) and near-zero query rate (0.073). The GPT-4.1 Analyst is more willing to confirm queried hazards, resulting in larger confirmed sets. The stance shift rate of 1.000 (every round changed the confirmed set) and convergence rate of 0.000 indicate that GPT-4.1 Discuss operates in a continued expansion mode rather than selective refinement.

Finding G3: Dialogue efficiency anomalies in GPT-4.1 Discuss. The acceptance rate of 1.021 and dialogue efficiency of 1.045 exceed 1.0, which is mathematically infeasible under the tag definitions. This arises from tag-parsing artifacts: GPT-4.1 occasionally outputs multiple `[SUPPORT]` or `[SUGGEST]` tags referencing the same hazard within a single utterance, inflating the denominator count. Classification metrics are computed from parsed final hazard lists and are unaffected.

Finding GA-4: GA statistical significance and random search comparison. Random search over the same configuration space achieves a mean

Table 6: Classification metrics — GPT-4.1.

System	Precision	Recall	F_1	Accuracy
BASE	0.1492	0.3018	0.1997	—
DEBATE	0.1914±0.1626	**0.4644**±0.2986	0.2473±0.1513	0.1498±0.1016
DISCUSS	0.1564±0.1026	**0.5864**±0.2853	0.2336±0.1276	0.1383±0.0845

Table 7: Dialogue metrics — GPT-4.1.

Metric	DEBATE	DISCUSS
Agreement Rate ρ_{ag}	0.5710±0.1707	1.021±0.0843†
Disagreement/Withdrawal Rate ρ_{dis}	0.4289±0.1703	0.5708±0.4634
Query Rate ρ_{qry}	—	0.0729±0.0607
Refine Rate ρ_{ref}	—	0.0967±0.0556
Stance Shift Rate ρ_{ss}	0.9208±0.1566	**1.000**±0.000
Dialogue Efficiency η	0.5651±0.1676	1.045±0.0887†
Convergence Rate	0.2178	0.000
Mean Convergence Round	2.45±0.50	—

† Values >1 indicate a tag-parsing artefact; see §7.

Table 8: Classification metrics — GA-DEBATE vs GPT-4.1 baselines.

System	Precision	Recall	F_1	Accuracy
GPT-4.1 BASE	0.1492	0.3018	0.1997	—
GPT-4.1 DEBATE	0.1914±0.1626	0.4644±0.2986	0.2473±0.1513	0.1498±0.1016
GA-DEBATE	**0.2159**±0.1901	**0.4656**±0.2502	**0.2585**±0.1541	**0.1579**±0.1078

safety score of 0.190 across all evaluated policies. The best single policy found by random search achieves a safety score of 0.556 on its one evaluation scenario which is a high single-scenario result attributable to a lucky draw (Cautious proposer, Strict critic, cap 3, 5 rounds on a scenario that matched this configuration). In contrast, the GA's converged policy achieves consistent improvement across all scenarios (mean $F_1 = 0.259$ vs. random search mean safety score 0.190). These results confirm that the GA's evolutionary pressure produces more reliable cross-scenario generalization than random exploration of the same parameter space.

6.4 GA-Optimized Adversarial Debate (GA-DEBATE)

Following the GPT-OSS 20B and GPT-4.1 results analysis, DEBATE was selected for evolutionary optimization as the best-performing dialogue modality by F_1. Table 8 presents the classification results, Table 9 the dialogue metrics, and Table 10 the converged optimal policy.

Finding GA1: GA optimization improves precision and F_1 without sacrificing recall. GA-DEBATE achieves $F_1 = 0.2585$ compared to 0.2473 for vanilla GPT-4.1 DEBATE ($\Delta F_1 = +0.011$) and 0.1997 for GPT-4.1 BASE ($\Delta F_1 = +0.059$). Recall is maintained at 0.4656 (essentially identical to DEBATE's 0.4644) while precision rises from 0.1914 to 0.2159 ($\Delta P = +0.025$), demonstrating that the GA improves the precision–recall balance rather than trading one for the other.

Table 9: Dialogue metrics — GA-DEBATE vs GPT-4.1 DEBATE.

Metric	GPT-4.1 Debate	GA-Debate
ρ_{ag}	0.5710	**0.7939**±0.2122
ρ_{dis}	0.4289	**0.2046**±0.2114
ρ_{ss}	0.9208	**0.9968**±0.0327
η	0.5651	**0.7874**±0.2411
Convergence Rate	0.2178	0.0097
Mean Conv. Round	2.45	3.00

Table 10: Optimal policy discovered by the GA after convergence.

Parameter	Converged Value
θ_1: num_rounds	3
θ_2: proposer_persona	Thorough
θ_3: critic_persona	Strict
θ_4: reasoning_depth	medium
θ_5: proposal_cap	20

Finding GA2: GA-optimized dialogue is substantially more deliberative. The agreement rate rises from 0.571 to 0.794 ($\Delta\rho_{\text{ag}} = +0.223$), and dialogue efficiency improves from 0.565 to 0.787 ($\Delta\eta = +0.222$), indicating that the Proposer's suggestions are better calibrated to the Critic's standards. The near-unity stance shift rate (0.997) confirms active revision of the hazard set in virtually every round.

Finding GA3: The converged policy recovers a recall-oriented configuration. The `Thorough` Proposer persona (which prioritizes coverage over precision) combined with a high proposal cap of 20 and a `Strict` Critic reflects the fitness function's recall bias ($\beta = 2$): the GA learned to generate broadly and filter aggressively, rather than generating conservatively. The `medium` reasoning depth balances deliberation quality against inference cost.

7 Discussion

7.1 Dialogue Across Model Capacity

The most striking cross-model finding is that dialogue mode interacts strongly with model capacity in determining the direction of improvement. On GPT-OSS 20B, adversarial debate improves F_1 while constructive discussion harms it; on GPT-4.1, both modes improve F_1 and constructive discussion achieves the highest recall GPT-OSS 20B frequently withdraws valid hazards when challenged because it cannot generate sufficiently specific justifications. GPT-4.1's stronger language grounding allows it to confirm borderline hazards correctly, flipping the withdrawal rate from 0.90 to 0.57. This

suggests that cooperative discussion architectures may be weak in smaller models.

7.2 Evolutionary Policy Learning Aligns the Proposer–Critic Calibration

The GA's most substantive finding is the emergence of the `Thorough + Strict` persona combination. Intuitively, a Thorough Proposer with an higher-cap ($= 20$) generates more candidates, ensuring recall is not sacrificed in the proposal phase; the Strict Critic then filters aggressively, recovering precision. This pipeline structure mirrors a "generate-then-verify" paradigm that the system discovered without being explicitly programmed. The agreement rate improvement ($+0.223$) indicates that after optimization, Proposer outputs are substantially better matched to the Critic's evidence standards, reducing wasted computation on proposals that will be rejected.

7.3 Dialogue Efficiency

The GA-optimized system's convergence rate drops to 0.010 (from 0.218), meaning almost no scenario converges before round 3. This initially appears negative but reflects the Thorough Proposer continually surfacing new hazards each round under the high cap, maintaining productive coverage expansion throughout. Combined with the high stance shift rate (0.997), this indicates that all three rounds contribute meaningfully to hazard discovery, validating the GA's selection of $\theta_1 = 3$ rounds.

7.4 Relation to Automated Agentic System Design

GA-DEBATE instantiates the Automated Design of Agentic Systems (ADAS) framework introduced by (Hu et al., 2024), which formalizes system design through a search space, search algorithm, and evaluation function. Rather than introducing a novel optimization algorithm, this work applies the ADAS framework to a safety-critical identification task. We introduce a domain-specific, recall-prioritized fitness function and empirically demonstrate that evolutionary search over agentic interaction parameters yields superior mean cross-scenario performance compared to random exploration. While (Yuan et al., 2024) employ evolutionary operators to optimize structural role compositions and tool assignments, our approach is complementary, focusing instead on prompt-level personas and deliberation-depth parameters.

7.5 Limitations

Tag-parsing artefacts. GPT-4.1 DISCUSS produces acceptance rates and dialogue efficiency scores above 1.0 due to repeated tag emission within single utterances. Robust tag deduplication should be applied in future work to make these metrics directly comparable across models.

GA evaluation cost. The exploration phase requires $N_{\text{init}} = 20$ distinct policy evaluations before selection begins. For larger datasets, the exploration overhead is proportionally smaller; for small datasets, a warm-start from prior domain knowledge could reduce it.

Fuzzy matching threshold. The Jaccard threshold of 0.60 handles most surface variants but does not capture derivational morphology (e.g. *electric* vs. *electrical*). An embedding-based matching layer would improve evaluation fidelity for label vocabularies with high morphological diversity.

One Dataset. This study utilizes a single industrial-domain dataset encompassing four diverse sectors—construction, electrical, chemical, and facilities—for operational safety analysis. While structural findings are expected to generalize, the adaptability of optimal parameter configurations to other closed-list identification tasks, remains to be demonstrated.

8 Conclusion

This paper presented a systematic investigation of agentic dialogue for operational hazard identification from a closed label list. We evaluated three base configurations—single-prompt inference (BASE), adversarial debate (DEBATE), and cooperative discussion (DISCUSS)—across two language models (GPT-OSS 20B and GPT-4.1), and proposed an evolutionary policy optimization framework (GA-DEBATE) that learns the optimal agentic configuration from prediction feedback. The results demonstrates the gains of an agentic dialogue system on hazard identification performance.

9 Acknowledgment

This research is sponsored by the Office of the Laboratory Director, Oak Ridge National Laboratory's Operational Excellence Initiatives, which is supported by the United States Department of Energy (DOE)'s Office of Science under Contract No. DE-AC05-00OR22725.

References

2007. An introduction to the bootstrap.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Safety Executive Bootle. 2005. Introduction to iec 61508.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Thomas Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shan Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *ArXiv*, abs/2308.07201.

Sanjay Das, Swastik Bhattacharya, Souvik Kundu, Shamik Kundu, Anand Menon, Arnab Raha, and Kanad Basu. 2024. Genbfa: An evolutionary optimization approach to bit-flip attacks on llms. *arXiv preprint arXiv:2411.13757*.

Rami Debouk. 2019. Overview of the second edition of iso 26262: Functional safety—road vehicles. *Journal of System Safety*, 55(1):13–21.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*.

Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. 2024. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. *ArXiv*, abs/2309.08532.

Shengran Hu, Cong Lu, and Jeff Clune. 2024. Automated design of agentic systems. *ArXiv*, abs/2408.08435.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Delong Chen, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55:1 – 38.

Trevor A. Kletz. 1999. Hazop & hazan: Identifying and assessing process industry hazards, fouth edition.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916.

Namgil Lee, Heejung Yang, and Hojin Yoo. 2021. A surrogate loss function for optimization of fβ score in binary classification with imbalanced data. *ArXiv*, abs/2104.01459.

Nancy G. Leveson. 2012. Engineering a safer world: Systems thinking applied to safety.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep reinforcement learning for dialogue generation. *ArXiv*, abs/1606.01541.

Peng Li, Xiang Cheng, Xu Chu, Yeye He, and Surajit Chaudhuri. 2021. Auto-fuzzyjoin: Auto-program fuzzy similarity joins without labeled examples. *Proceedings of the 2021 International Conference on Management of Data*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. In *Conference on Empirical Methods in Natural Language Processing*.

Nicola Paltrinieri, Louise K. Comfort, and Genserik L. L. Reniers. 2019. Learning about risk: Machine learning for risk assessment. *Safety Science*.

Joon Sung Park, Joseph O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *ArXiv*, abs/1806.03822.

Nigel Slack. 2015. Failure mode and effect analysis.

Dan Thomas. 2023. Revolutionizing failure modes and effects analysis with chatgpt: Unleashing the power of ai language models. *Journal of Failure Analysis and Prevention*, 23:911–913.

W E Vesely, Florinda F. Goldberg, Nicole Roberts, and D F Haasl. 1987. Fault tree handbook.

Vincent Xian Wang and Lily Lim. 2025. How does gpt-4.1 comprehend conversational implicatures? reasoning with contextual alternatives in discourse frames. *Linguistic Research*, 42.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed H. Chi, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *ArXiv*, abs/2203.11171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Ronald J. Williams. 2004. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *ArXiv*, abs/2306.13063.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *ArXiv*, abs/2305.10601.

Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Dongsheng Li, and Deqing Yang. 2024. Evoagent: Towards automatic multi-agent generation via evolutionary algorithms. In *North American Chapter of the Association for Computational Linguistics*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *ArXiv*, abs/2309.03882.

Daniel M. Ziegler, Nisan Stiennon, Jeff Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *ArXiv*, abs/1909.08593.

A Dialogue Metrics: Definitions and Interpretation

This appendix provides formal definitions of all dialogue metrics reported in Section 5.3, together with interpretive notes on boundary conditions and known measurement artifacts.

A.1 Formal Metric Definitions

Let R denote the number of dialogue rounds per session, N_{PROP} the total number of proposals or suggestions across all rounds, and subscripts r a per-round quantity.

Agreement Rate (ρ_{ag}).

$$\rho_{\text{ag}} = \frac{|\texttt{[AGREE]}| + |\texttt{[SUPPORT]}| + |\texttt{[REFINE]}| + |\texttt{[CONFIRM]}|}{N_{\text{PROP}}} \tag{6}$$

Measures the fraction of proposals that received a positive verdict. Values near 1.0 indicate either high-quality proposals or a permissive second agent. Values above 1.0 are a parsing artefact (see §A.2).

Disagreement/Withdrawal Rate (ρ_{dis}).

$$\rho_{\text{dis}} = \frac{|\texttt{[DISAGREE]}| + |\texttt{[WITHDRAW]}|}{N_{\text{PROP}}} \tag{7}$$

Measures the fraction of proposals eliminated. In DEBATE, $\rho_{\text{ag}} + \rho_{\text{dis}} = 1$ by construction. In DISCUSS, the three-stage query-confirm-withdraw pathway means the sum can differ from 1.0 for sessions with unresolved queries.

Query Rate (ρ_{qry}) — DISCUSS only.

$$\rho_{\text{qry}} = \frac{|\texttt{[QUERY]}|}{N_{\text{PROP}}} \tag{8}$$

Fraction of suggestions placed into the three-stage query-confirm-withdraw deliberation pathway. High ρ_{qry} indicates epistemic caution; near-zero values (as observed for GPT-4.1 Discuss, $\rho_{\text{qry}} = 0.073$) indicate the Reviewer commits to a verdict without intermediate deliberation.

Refine Rate (ρ_{ref}) — DISCUSS only.

$$\rho_{\text{ref}} = \frac{|\texttt{[REFINE]}|}{N_{\text{PROP}}} \tag{9}$$

Fraction of suggestions replaced by a more precise label from $\mathcal{L}$. Non-zero values indicate the Reviewer is actively exercising label-vocabulary expertise beyond simple accept/reject decisions.

Stance Shift Rate (ρ_{ss}).

$$\rho_{\text{ss}} = \frac{\left|\{r \in \{1, \ldots, R\} : \Sigma_r \neq \Sigma_{r-1}\}\right|}{R} \quad (10)$$

Fraction of rounds in which the confirmed/locked set Σ changed. A value of 1.0 indicates that every round produced a non-empty net change; a value near 0 indicates early convergence (or a degenerate dialogue in which no proposals are generated).

Dialogue Efficiency (η).

$$\eta = \frac{|\hat{\mathcal{H}}|}{N_{\text{PROP}}} \quad (11)$$

Signal-to-noise ratio of the dialogue: the fraction of all proposals that survive to the final predicted set. Higher values indicate that a larger proportion of agent effort translates to output. Values above 1.0 are a parsing artefact.

Convergence Rate.

$$\text{Conv. Rate} = \frac{|\{w : \exists\, r < R \text{ s.t. } \Sigma_r = \Sigma_{r'}\ \forall r' > r\}|}{|\mathcal{W}|} \quad (12)$$

Fraction of scenarios for which the confirmed set stabilized before the final round. Rapid convergence is not inherently desirable; it may reflect productive early consensus or premature over-pruning.

A.2 Measurement Artifacts and Boundary Conditions

Values above 1.0 in GPT-4.1 Discuss. The acceptance rate of 1.021 and dialogue efficiency of 1.045 observed for GPT-4.1 DISCUSS arise when the model emits multiple `[SUPPORT]` tags referencing the same hazard label within a single utterance, inflating the denominator count while the numerator is deduplicated. This is a tag-parsing artifact specific to GPT-4.1's tendency for repetitive structured output. Classification metrics (precision, recall, F_1) are computed from the parsed hazard lists and are unaffected. Future work should implement utterance-level tag deduplication before metric computation.

Convergence rate of 0.0 for GPT-4.1 Discuss. A convergence rate of 0.000 indicates that in no scenario did the confirmed set stabilize before round 3. Given that GPT-4.1 Discuss achieves the highest recall of any system, this reflects continued productive coverage expansion across all three rounds rather than a failure to converge. The GA-Debate near-zero convergence rate (0.010) reflects a similar dynamic: the `Thorough` Proposer persona continuously discovers new hazards each round under the high proposal cap.

Fuzzy label matching in per-round metrics. Per-round F_1 and safety scores are computed against the golden set using the same Jaccard-based fuzzy matching (threshold ≥ 0.60) as the final evaluation metrics, ensuring that per-round and final metrics are commensurable.

B Turn Design Principles

B.1 Adversarial Debate

Let $r \in \{1, \ldots, R\}$ index rounds. Each round consists of two turns:

$$u_{2r-1} = \textsc{Proposer}\big(\pi^{\text{PROP}},\ w,\ \mathcal{L},\ \mathcal{L}^{+}_{r-1},\ \mathcal{L}^{-}_{r-1}\big) \quad (13)$$

$$u_{2r} = \textsc{Critic}(\pi^{\text{CRIT}},\ w,\ u_{2r-1}) \quad (14)$$

State updates after each critic turn:

$$\mathcal{L}^{+}_{r} = \mathcal{L}^{+}_{r-1} \cup \{h : \texttt{[AGREE]}(h) \in u_{2r}\} \quad (15)$$

$$\mathcal{L}^{-}_{r} = \mathcal{L}^{-}_{r-1} \cup \{h : \texttt{[DISAGREE]}(h) \in u_{2r}\} \quad (16)$$

The aggregation function reads the final locked set and the finalization pass:

$$\hat{\mathcal{H}}^{\text{DEBATE}} = \Omega(\Sigma_R) = \textsc{Finalize}\big(\mathcal{L}^{+}_{R},\ \mathcal{L}^{-}_{R}\big) \quad (17)$$

B.2 Constructive Discussion

Each round r consists of three sub-turns:

$$u^{\text{SUG}}_{r} = \textsc{Analyst}(\pi^{\text{ANA}},\ w,\ \mathcal{L},\ \mathcal{C}_{r-1},\ \mathcal{W}_{r-1}) \quad (18)$$

$$u^{\text{REV}}_{r} = \textsc{Reviewer}(\pi^{\text{REV}},\ w,\ u^{\text{SUG}}_{r}) \quad (19)$$

$$u^{\text{RESP}}_{r} = \textsc{Analyst}(\pi^{\text{RESP}},\ w,\ u^{\text{REV}}_{r}) \quad \text{if } \mathcal{Q}_r \neq \emptyset \quad (20)$$

The query-response sub-turn (Equation 20) fires only when the Reviewer raises at least one `[QUERY]`, making the number of LLM calls per round data-dependent. State updates proceed analogously to Equations (15)–(16), with `[REFINE]` entries additionally populating $\mathcal{F}$.

Aggregation. Refinement substitutions are applied to the confirmed set before output:

$$\hat{\mathcal{H}}^{\text{DISCUSS}} = \Omega(\Sigma_R) = \{h_{\text{refined}} \mid (h_{\text{orig}}, h_{\text{refined}}) \in \mathcal{F}_R \cup \left(\mathcal{C}_R \setminus \mathcal{C}^{\text{orig}}_{R}\right) \quad (21)$$

where $\mathcal{C}_R^{\text{orig}}$ denotes confirmed hazards that were subsequently refined.

C Prompt Templates

This appendix provides the complete system and user-turn prompt templates for all four systems evaluated in this work. All templates are used verbatim in the experimental implementation.

C.1 Base Single-Prompt System (BASE)

System:

```
You are an expert Safety AI. Your task is to identify
relevant hazards from a provided master list based
on a specific work description.

1. Analyse the "Work Description" and "Master
Hazard List" carefully.
2. Select ONLY hazards from the "Master Hazard List"
that directly apply to the described work.
3. If no hazards apply, return "No applicable hazards".
4. Output the result as a raw Python list of strings.
  Do not include any reasoning or introductory text.
```

User:

```
### Master Hazard List
{hazard_list}

### Work Description
{work_description}

### Relevant Hazards
[
```

C.2 Adversarial Debate: Proposer (DEBATE)

System:

```
You are Hazard_Proposer, a senior operational
safety expert.
Persona: {proposer_persona_text}
Reasoning: {reasoning_depth_text}
{proposal_cap_instruction}

Your task: SELECT hazards from the Master Hazard
List that apply to the Work Description.
- ONLY use hazard names that appear verbatim in
the Master Hazard List.
- Do NOT re-propose hazards already agreed or
rejected this session.
Output format — one entry per hazard:
[PROPOSE]: <exact hazard name from master list>
[REASON]: <why this hazard applies to this
specific work>
```

User (per round):

```
Work Description:
{work_description}

Master Hazard List:
{hazard_list}

Already AGREED: {locked_hazards}
Already REJECTED: {rejected_hazards}

{proposal_cap_instruction}
```

C.3 Adversarial Debate: Critic (DEBATE)

System:

```
You are Hazard_Critic, a senior operational safety
auditor.
Persona: {critic_persona_text}
Reasoning: {reasoning_depth_text}

Your task: Review each proposed hazard and give a
verdict.
- [AGREE] if the hazard clearly applies to this
work.
- [DISAGREE] if it does not clearly apply.

Output format — one verdict per hazard:
[AGREE]: <exact hazard name>
[REASON]: <why it applies>

or

[DISAGREE]: <exact hazard name>
[REASON]: <why it does not apply>
```

User (per round):

```
Work Description:
{work_description}

Proposed hazards:
{proposed_list}
```

C.4 Adversarial Debate: Finalisation

System:

```
You are Hazard_Proposer producing the final hazard
list.

Include ONLY hazards the Critic explicitly AGREED to.
Exclude all DISAGREED hazards.
Remove duplicates.
If none were agreed: [FINAL]: No applicable hazards

Output format — one line per hazard:
[FINAL]: <exact hazard name>
```

User:

```
Work Description:
{work_description}

AGREED hazards:
{locked_hazards}

REJECTED hazards (exclude):
{rejected_hazards}
```

C.5 Constructive Discussion: Analyst (DISCUSS)

System:

```
You are Hazard_Analyst, a senior safety engineer
working with a peer reviewer to build a shared hazard list.

Your job each round is to suggest hazards from the
Master Hazard List that you believe apply to the
work description. Draw on your safety expertise —
consider the materials,
equipment, environment, and activities involved.

Use this format for each hazard you suggest:
[SUGGEST]: <hazard name from master list>
[REASON]: <why this hazard applies to this specific
work>
```

```
If you have nothing new to add this round:
[PASS]: No further suggestions
```

User (per round):

```
Work Description:
{work_description}

Master Hazard List:
{hazard_list}

Already confirmed hazards (no need to re-suggest):
{confirmed_hazards}
Withdrawn hazards (avoid re-suggesting):
{withdrawn_hazards}

Suggest hazards from the Master Hazard List that apply.
```

C.6 Constructive Discussion: Reviewer (DISCUSS)

System:

```
You are Hazard_Reviewer, a peer safety expert
collaborating with Hazard_Analyst to produce the best
possible hazard list.

Review each suggested hazard openly and constructively.
Your options:
- [SUPPORT] it if it clearly applies — say why.
- [REFINE] it if a more precise hazard from the master
list
  better fits the work — suggest the better label.
- [QUERY] it if genuinely unsure — ask one focused
question.

You may also volunteer [SUGGEST] new hazards you think
were missed.

Formats:
[SUPPORT]: <hazard name>
[REASON]: <why it applies>

[REFINE]: <original> -> <better hazard name from
masterlist>
[REASON]: <why the refined label fits better>

[QUERY]: <hazard name>
[QUESTION]: <one focused question>

[SUGGEST]: <hazard name from master list>
[REASON]: <why this was missed and applies>

If nothing to add or query:
[PASS]: Nothing to review
```

User (per round):

```
Work Description:
{work_description}

Master Hazard List:
{hazard_list}

Analyst's suggestions this round:
{suggestion_block}

Already confirmed: {confirmed_hazards}
Already withdrawn: {withdrawn_hazards}

Review each suggestion and add any hazards you think
were missed.
```

C.7 Constructive Discussion: Analyst Query Response

System:

```
You are Hazard_Analyst responding to your peer
reviewer's queries and suggestions.

For each [QUERY]: answer based on the work
description.
- If the hazard applies: [CONFIRM] it with reasoning.
- If not: [WITHDRAW] it and explain briefly.

For new [SUGGEST] from the reviewer: [SUPPORT]
if you agree, or [QUERY] back if unsure.

Formats:
[CONFIRM]: <hazard name>
[EVIDENCE]: <reasoning from the work
description>

[WITHDRAW]: <hazard name>
[REASON]: <why it does not clearly apply>

[SUPPORT]: <hazard name>
[REASON]: <why you agree with reviewer's
suggestion>
```

User:

```
Work Description:
{work_description}

Reviewer's queries:
{query_block}

Reviewer also suggested:
{reviewer_suggestion_block}

Respond to each query. If the reviewer suggested
new hazards, [SUPPORT] or [QUERY] them as appropriate.
```

C.8 Constructive Discussion: Finalization

System:

```
You are Hazard_Analyst producing the final agreed
hazard list.

Include hazards that were [SUPPORT]ed or [CONFIRM]ed.
Apply any [REFINE] substitutions — use the refined
name. Exclude hazards that were [WITHDRAW]n.
Remove duplicates.

If nothing was confirmed:
[FINAL]: No applicable hazards

Output format — one line per hazard:
[FINAL]: <hazard name>
```

User:

```
Work Description:
{work_description}

CONFIRMED hazards:
{confirmed_hazards}

WITHDRAWN hazards (exclude these):
{withdrawn_hazards}

REFINEMENTS (use refined name):
{refinement_notes}

Output the final list using [FINAL]: <hazard name>,
one per line.
```

C.9 GA-Debate: Persona Prompt Fragments

The GA-optimized system injects persona-specific text into the Proposer and Critic system prompts. Table 11 lists the complete text for each value of the learnable parameters θ_2 (Proposer persona) and θ_3 (Critic persona).

Reasoning depth fragments.

- **high**: *"Think step-by-step through each hazard carefully before deciding."*
- **medium** (converged optimal): *"Provide a brief justification for each decision."*
- **low**: *"Respond concisely with minimal explanation."*

Table 11: Persona prompt fragments injected into agent system prompts during GA-optimised adversarial debate. The converged optimal values are **bolded**.

Value	Proposer Persona Text	Critic Persona Text
Thorough / Strict	*"Be comprehensive — identify every plausible hazard the work could involve, including those implied by the equipment, environment, and materials. It is better to propose more and let the critic filter."*	*"Apply a high evidence bar. DISAGREE with any hazard not directly and unambiguously evidenced by a named activity or condition. When in doubt, DISAGREE."*
Focused / Balanced	*"Be selective — only propose hazards with clear, direct evidence in the work description. Avoid speculation. Propose fewer, well-justified hazards."*	*"Apply a balanced judgement. AGREE if clearly or reasonably implied. DISAGREE only when there is genuinely no evidence. Give the proposer the benefit of the doubt for borderline cases."*
Cautious / Skeptical	*"Apply conservative safety principles. When in doubt whether a hazard applies, propose it — a missed hazard is worse than a false alarm. Prioritise recall over precision."*	*"Scrutinise the proposer's reasoning, not just the label. DISAGREE if the reason is generic. AGREE only when the reason directly references a specific feature of the work description."*