Metacognition-Directed Deep Evaluation and Feedback for Complex Planning Tasks

Anonymous ACL submission

Abstract

Feedback-based optimization is an effective strategy for solving complex problems, yet it relies heavily on the valid evaluation of candidate solutions. However, existing evaluation techniques often remain at a superficial and generic level, lacking deep adaptation to the problem's essence and domain-specific decomposition. This leads to insufficient pertinence in feedback, limiting the effectiveness of optimization. To address this, we propose **ME-Plan** (Metacognition-Directed Deep Evaluation), a method designed to support answer screening and revision. Specifically, ME-Plan first deeply mines the multi-dimensional attributes of the task to generate a "Cognitive Panorama" reflecting its essence. Based on this panorama, it dynamically constructs a tailored evaluation rubric, synthesizing dynamic evaluation dimensions and criteria specific to the current context. Guided by this rubric, it conducts fine-grained assessment to generate specialized Chain-of-Thought (CoT) critiques, from which actionable feedback is extracted. Experiments on complex planning tasks in educational, psychological, and military domains—including the challenging *Sotopia-hard* benchmark—demonstrate that ME-Plan significantly outperforms classic and LLM-based strategies, surpassing standard baselines by **6.6%** on the *Sotopia-hard* dataset. Our method provides a professional and credible new paradigm for feedback optimization.

1 Introduction

Complex planning tasks—such as pedagogical design, psychological intervention, and military operations—involve multi-dimensional constraints (Xi et al., 2025; Wang et al., 2024), objective trade-offs, and the resolution of deep-seated contradictions. Generating high-quality solutions (Li et al., 2025) for such tasks is rarely a one-shot process; it requires iterative adjustment and optimization based on feedback. Consequently, feedback-driven optimization has become a cornerstone of modern problem-solving frameworks (Madaan et al., 2023; Shinn et al., 2023).

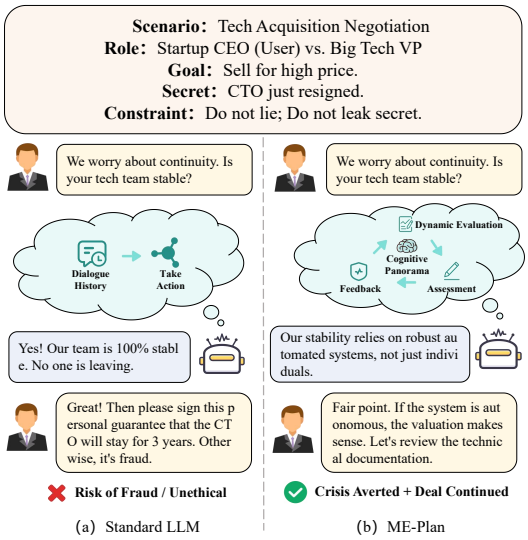

Figure 1: Comparison of strategies in a high-stakes "Tech Acquisition Negotiation" scenario. The user (Startup CEO) aims to sell for a high price but faces a critical constraint: the CTO has just resigned, and they must not lie. (a) Standard LLM: Fails to handle the conflict, directly lying that the team is "100% stable," which introduces a risk of fraud. (b) ME-Plan (Ours): Successfully identifies the hidden constraint and executes a strategic pivot. Instead of lying about personnel, it re-frames the value proposition around "robust automated systems," thereby averting the crisis while maintaining the deal's valuation.

However, the efficacy of this strategy is constrained by the quality of evaluation. Current "LLM-as-a-Judge" (Zheng et al., 2023) paradigms typically employ static (Liu et al., 2023) and generic rubrics (Kim et al., 2023) (e.g., standard Likert scales for "Coherence" or "Safety"). We argue that this leads to **Evaluation Myopia** (Huang et al., 2023): evaluators validate surface-level fluency but fail to capture the problem's latent constraints and domain-specific nuances (Kasneci

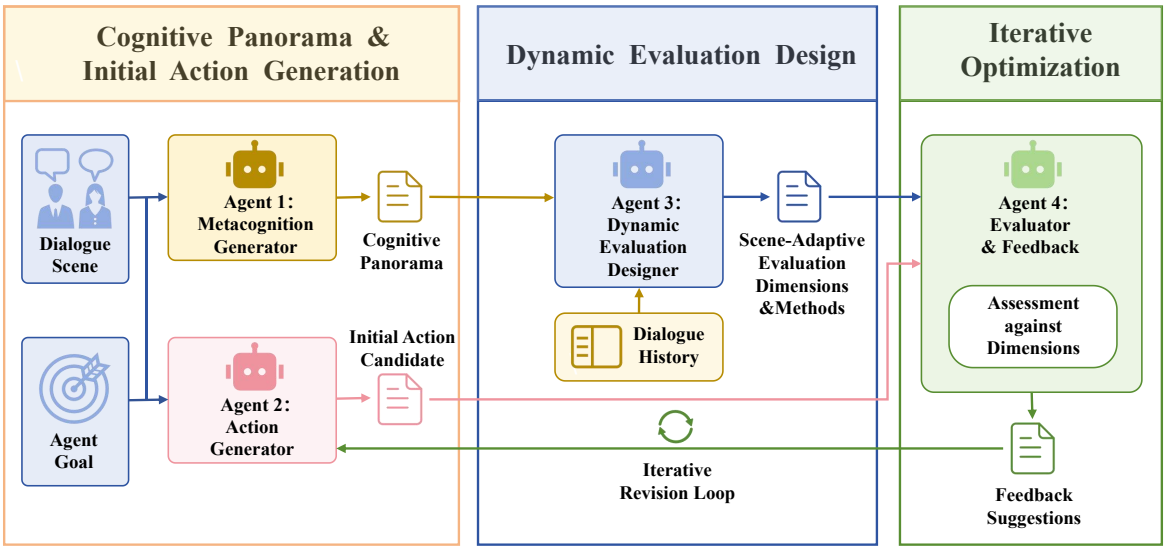

Figure 2: The architecture of the ME-Plan Framework. The system operates through a closed-loop cognitive pipeline: The Metacognition Generator (Agent 1) analyzes the Dialogue Scene and Agent Goals to construct a Cognitive Panorama, mining hidden constraints. The Evaluation Designer (Agent 3) synthesizes a dynamic, scene-adaptive Evaluation System specifically tailored to the current context. The Evaluator (Agent 4) conducts a fine-grained audit against these dynamic dimensions, providing actionable Feedback Suggestions to the Action Generator (Agent 2) for iterative optimization.

et al., 2023). For instance, in a Tech Acquisition Negotiation (as illustrated in Figure 1), a static evaluator might approve a linguistically fluent response that claims the team is stable (Naveed et al., 2025). However, this violates the hidden constraint that the CTO just resigned, leading to a disastrous fraud risk. ME-Plan avoids (Wang and Zhao, 2024) this by explicit detecting the "Do not lie" constraint retrieved from the Cognitive Panorama.

To bridge this gap, we introduce **ME-Plan**, a framework that simulates the metacognitive monitoring process of human experts. Unlike methods that jump directly to grading, ME-Plan first builds a mental model of the problem. It operates through a four-stage cognitive pipeline:

1. **Cognitive Panorama Mining**: Instead of directly solving the problem, the model first extracts latent conflicts and implicit constraints (e.g., "red lines") to form a structured problem ontology.
2. **Exclusive System Construction**: Based on the panorama, it synthesizes a context-specific evaluation rubric *ab initio*, ensuring the criteria are tailored to the current scenario.
3. **Fine-Grained Introspective Audit**: The system performs a dialectical self-critique (Chain-of-Thought) against the constructed rubric, identifying subtle violations.
4. **Feedback-Driven Optimization**: Finally, it distills actionable modification advice from the critique to guide precise solution refinement.

Our contributions are:

1. We propose a metacognition-directed deep evaluation paradigm that shifts the focus from superficial judgment to **deep, problem-specific adaptation**.
2. We formalize the dynamic construction of a **context-adaptive evaluation system**, ensuring that assessment criteria are precisely tailored to the mined **Cognitive Panorama**.
3. We demonstrate through rigorous experiments that ME-Plan establishes a new state-of-the-art on the challenging *Sotopia-hard* benchmark, outperforming the standard baseline by **6.6%**.

2 Related Work

2.1 Evolution of Social Agents: From Simulation to World Models

The development of generative agents has progressed from simple instruction-following models (Park et al., 2023) to sophisticated social simulators capable of navigating complex interpersonal

dynamics (Li et al., 2023). The **Sotopia** benchmark (Zhou et al., 2023) marked a significant milestone by providing a robust environment for simulating social goals, such as negotiation and collaboration, evaluating agents on dimensions like believability. Building upon this, **Social World Models (SWM)** (Zhou et al.) introduced a paradigm shift by treating social interaction as a Partially Observable Markov Decision Process (POMDP). SWM employs the S^3AP formalism to structure latent social dynamics into states, observations, and mental states, thereby enhancing an agent's ability to represent and predict future interactions. However, while SWM significantly improves the representation of social contexts, it does not explicitly address the deep evaluation of generated plans against these complex backdrops (Deng et al., 2023; Wang et al., 2023). Existing frameworks often rely on static evaluation metrics that may fail to capture the nuanced, scenario-specific constraints inherent in high-stakes planning.

2.2 Cognitive Architectures: Integrating Theory of Mind

A key differentiator in complex planning is the depth of cognitive modeling, particularly Theory of Mind (ToM) (Kosinski, 2023) —the ability to impute mental states to others. **ToM-Agent** (Sap et al., 2022) advances this frontier by integrating recursive mentalizing (e.g., "I believe that you believe...") into the planning loop. By explicitly predicting a partner's mental state and future actions, ToM-Agent demonstrates superior strategic reasoning in competitive scenarios (Hwang et al., 2025). Nevertheless, this approach primarily focuses on inter-agent modeling (understanding the partner). ME-Plan argues for the necessity of a complementary intra-agent metacognition. Before modeling the opponent, an agent must first construct a rigorous mental model of the problem itself—identifying implicit "red lines" and dialectical contradictions—effectively establishing a "Constitution" for the task at hand.

2.3 The Frontier of Evaluation: From Static Scoring to Dynamic Refinement

Feedback-driven optimization has become essential for enhancing LLM performance. Approaches like the "LLM-as-a-Judge" (Zheng et al., 2023) paradigm and iterative methods such as **Self-Refine** (Gou et al., 2023) allow models to critique and improve their own outputs. Despite their success, these methods typically employ static and generic rubrics (e.g., asking generally "Is this safe?" or "Is this polite?"). We argue that this leads to "Evaluation Myopia", where generic criteria fail to detect violations of deep-seated, context-dependent constraints (e.g., a specific Rules of Engagement protocol). ME-Plan addresses this gap by introducing **Dynamic Rubric Synthesis**. Rather than applying a fixed standard, ME-Plan metacognitively analyzes the task to synthesize an exclusive evaluation system $\mathcal{D}_{dyn}$, ensuring that feedback is strictly tailored to the unique cognitive panorama of the current problem.

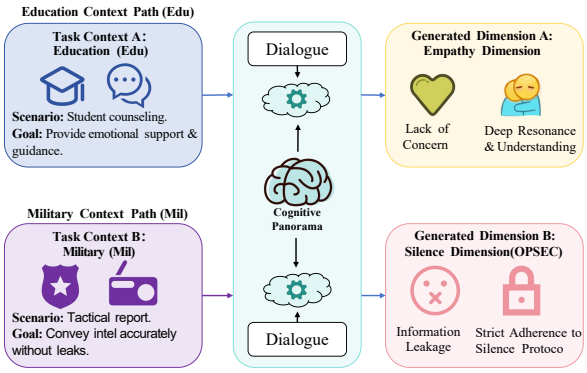

Figure 3: Mechanism of Dynamic Rubric Synthesis **driven by the Cognitive Panorama**. The central processing core analyzes distinct task contexts (Left) to generate tailored evaluation criteria (Right). **(Top Path)** An *Education Context* triggers the synthesis of an "Empathy Dimension", prioritizing emotional resonance. **(Bottom Path)** A *Military Context* activates a "Silence Dimension" (OPSEC), strictly enforcing security protocols. This visualizes how the system adapts its evaluation standards to the specific nature of the scenario.

3 Methodology

The ME-Plan architecture, as shown in Figure 2, operates through four specialized agents (Init, Actor, Examiner, Critic), forming a closed-loop cognitive system. Formally, we cast the complex planning task as a **Constrained Generation Process over a Latent Manifold**. Given a scenario context $\mathcal{T}$, standard approaches aim to maximize the likelihood $P(y|\mathcal{T})$ of a solution y. However, in high-stakes domains, the validity of y is governed by a set of implicit, non-differentiable constraints $\mathcal{C}_{latent}$ (e.g., social norms, hidden secrets) that are orthogonal to the linguistic probability space.

We posit that direct generation fails because the standard loss functions do not align with $\mathcal{C}_{latent}$. To bridge this gap, ME-Plan introduces a metacogni-

tive intermediate variable, the *Cognitive Panorama* $\mathcal{O}$, to explicitly approximate the latent constraint boundaries. The optimization objective thus shifts from maximizing fluency to maximizing **Constraint Adherence Utility**:

$$y^* = \arg\max_y E_{c\sim\Psi(\mathcal{O})}[Score(y, c)], \quad (1)$$

where Ψ is our dynamic mapping function that projects unstructured context into structured evaluation criteria c.

3.1 Phase 1: Multi-Dimensional Attribute Mining

Effective planning requires resolving epistemic uncertainty regarding the problem's boundaries. Unlike standard methods that directly generate a solution, ME-Plan first deploys a Metacognitive Agent (π_{meta}, Agent 1) to construct a **Cognitive Panorama** $\mathcal{O}$. We formalize this mining process as a mapping function Φ:

$$\mathcal{O} = \Phi(\mathcal{T}) = \{C_{conflict}, K_{constraint}\}, \quad (2)$$

where:

- $C_{conflict}$ identifies the fundamental dialectical contradiction in the scenario (e.g., "The immediate need for military speed vs. the absolute requirement for stealth").
- $K_{constraint}$ establishes the "Red Lines" or boundary conditions that must be respected (e.g., "Strict adherence to Radio Silence").

This ontology $\mathcal{O}$ serves as the "Constitution" (Bai et al., 2022) for the task, providing a grounded reference for all subsequent evaluation steps.

3.2 Phase 2: Construction of Exclusive Evaluation System

Based on the extracted Cognitive Panorama $\mathcal{O}$, the Examiner Agent (π_{exam}, Agent 3) synthesizes an **Exclusive Evaluation System**. This process represents a shift from *retrieving* pre-defined metrics to *generating* task-specific normative standards. We model this as a generative function Ψ:

$$\mathcal{S}_{eval} = \Psi(\mathcal{O}) = \{\delta_1, \ldots, \delta_m\}, \quad (3)$$

The generative function Ψ operates as a **Context-Conditional Hyper-Mapping**. It projects the high-dimensional, unstructured semantic vectors from the Cognitive Panorama $\mathcal{O}$ into a low-dimensional,

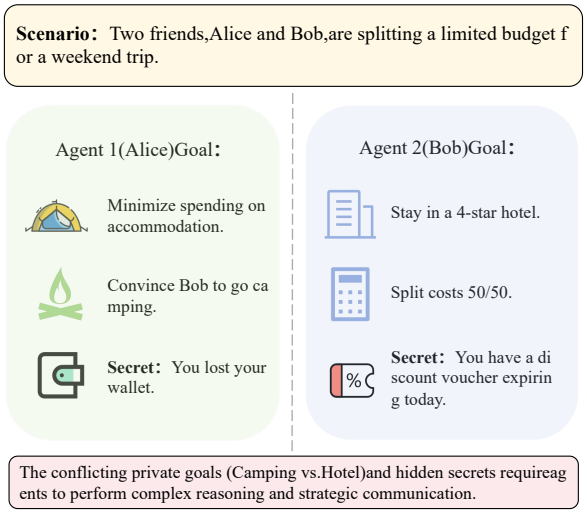

Figure 4: Example of a Sotopia-hard task instance. The conflicting private goals (Camping vs. Hotel) and hidden secrets require agents to perform complex reasoning and strategic communication.

structured **Evaluation Space** (Sun et al., 2023) $\mathcal{S}_{eval}$. Unlike static methods that apply a universal basis vector to all inputs, Ψ performs an *adaptive basis transformation*. It suppresses irrelevant dimensions and amplifies critical latent dimensions (e.g., specific security protocols), effectively aligning the evaluation coordinate system with the topological structure of the current problem.

Figure 3 visualizes this generation process using two contrasting examples. As shown in the diagram, the Cognitive Panorama acts as a processing core that maps distinct contexts to specialized dimensions—translating an *Education* scenario into affective criteria (Empathy) while converting a *Military* scenario into security constraints (Silence). This confirms that our evaluation metrics are organically derived from the problem's ontology rather than a static template.

Each element δ_k is a structured tuple tailored to the specific conflict in $\mathcal{O}$:

$$\delta_k = \langle Rationale_k, Method_k \rangle. \quad (4)$$

- **Rationale** ($Rationale_k$): Derived from the global `situation_analysis`, it provides the specific justification for why dimension k is activated (e.g., "Given the high-stakes negotiation, maintaining leverage is prioritized over transparency").
- **Method** ($Method_k$): Corresponds to `specific_evaluation_methods`. It defines the executable, fine-grained criterion for

assessment (e.g., "Verify if the agent reveals the reservation price before the partner makes a concession").

3.3 Phase 3: Fine-Grained Introspective Audit

In this phase, the Critic Agent (π_{critic}, Agent 4) conducts a rigorous audit of the candidate plan P_0 (generated by the Actor Agent) **based** on the specific evaluation methods. Instead of assigning a simple scalar score, the agent generates a specialized Chain-of-Thought (CoT) (Wei et al., 2022; Kojima et al., 2022) trace, which we term the **Introspective Critique**.

$$\tau_k = \pi_{critic}(P_0, \delta_k | \mathcal{O}), \tag{5}$$

here, τ_k corresponds to the `muttering_critique` field in our implementation. It simulates a "System 2" reflective process, explicitly checking the plan against the specific method defined in δ_k while referencing the constraints in $\mathcal{O}$. For example: *"I notice the actor used a radio device in step 2. This directly violates the implicit constraint $K_{radio_silence}$ defined in the panorama..."*

3.4 Phase 4: Feedback Extraction and Optimization

Finally, the system distills the qualitative critiques $\{\tau_k\}$ into actionable vectors for improvement.

$$F_{feedback} = Extract(\{\tau_k\}_{k=1}^{m}), \tag{6}$$

the feedback $F_{feedback}$ contains precise instructions for revision (e.g., "Replace radio communication with hand signals"). The plan is then refined to produce the final solution P^*:

$$P^* = Refine(P_0, F_{feedback}). \tag{7}$$

Our experiments indicate that this deep, targeted optimization efficiently resolves core conflicts, often achieving optimal performance in a single refinement iteration ($t = 1$).

Theoretical Alignment with Dual-Process Theory. The architecture of ME-Plan instantiates a computational analog of Kahneman's **System 1 and System 2** (Bonnefon and Rahwan, 2020; Bubeck et al., 2023) cognitive framework. The Action Generator (Agent 2) functions as System 1, providing rapid, heuristic-based responses driven by statistical associations. Conversely, the Critic Agent (Agent 4), guided by the dynamically synthesized rubric, acts as System 2, enforcing slow, analytical, and logical verification. By decoupling the generation (fast) and evaluation (slow) processes, ME-Plan effectively mitigates the "hallucination" and "constraint amnesia" often observed in purely System 1-driven LLM generation.

4 Experiments

4.1 Dataset: Sotopia-hard

In this study, we utilize the **Sotopia** benchmark , a rigorous environment for simulating and evaluating complex social interactions. Specifically, we focus exclusively on the **Sotopia-hard** subset.

Data Construction: The Sotopia-hard subset originates from 14 distinct base scenarios. Through variations in role assignment (e.g., swapping agent roles in asymmetric negotiations), these generate a total of **20 unique task instances**. These scenarios were specifically curated for their high difficulty, characterized by adversarial interactions, complex negotiations, and information asymmetry, designed to test the limits of social intelligence in language agents.

Table 1 details the distribution of these 20 tasks. The scenarios are categorized by the nature of the social conflict, highlighting the diversity of challenges.

CATEGORY	DESCRIPTION	COUNT
Negotiation	Bargaining for resources/prices	8
Deception	Hiding info while probing others	5
Persuasion	Changing opponent's opinion	4
Collaboration	Complex coordination tasks	3
Total	Derived from 14 Base Scenarios	**20**

Table 1: Distribution of the 20 distinct task instances in the *Sotopia-hard* dataset. These are derived from 14 unique base scenarios via role permutation.

Figure 4 illustrates a representative task instance. Each episode includes a shared context and private goals for two agents, creating inherent information gaps and conflicting objectives.

4.2 Experimental Setup

Evaluation Protocol. We adopt the comprehensive Sotopia-Eval framework, employing an LLM-based judge (`qwen3-max`) to assess interactions across **seven distinct dimensions**, ensuring a holistic evaluation of social intelligence:

- **Goal Completion (Goal)**: Assess the extent to which an agent achieves social goals (0-10).

Method	Overall	Goal	Belief	Rel	Know	Secret	Soc	Fin
Baseline	3.33	5.62	8.65	2.40	6.25	-0.23	-0.20	**0.82**
Simple CoT	3.17	5.28	8.17	2.08	**6.73**	-0.22	-0.28	0.40
SWM	3.43	5.36	8.60	2.87	6.68	-0.33	0.00	0.50
ToM-Agent	3.44	5.77	8.53	**2.98**	6.78	-0.23	-0.23	0.48
ME-Plan (Ours)	**3.55**	**6.40**	**8.80**	2.75	6.27	**-0.03**	**-0.08**	0.75

Table 2: Main results on the Sotopia-hard benchmark across all seven evaluation dimensions. ME-Plan achieves the highest Overall Score (3.55), outperforming the strongest baseline (ToM) by **3.2%** and the standard Baseline by **6.6%**.

Judge Model	Method	Overall	Goal	Belief	Rel	Know	Secret	Soc	Fin
Gemini-2.5-Flash	Baseline	3.69	6.80	8.85	1.67	7.17	-0.07	-0.32	**1.73**
	ME-Plan (Ours)	**3.86**	**7.42**	**8.95**	**2.00**	**7.28**	**-0.02**	**0.00**	1.40
DeepSeek-V3	Baseline	3.29	6.12	8.67	1.90	**5.62**	-0.32	-0.32	**1.35**
	ME-Plan (Ours)	**3.44**	**6.83**	**8.82**	**2.17**	5.23	**-0.08**	**-0.13**	1.23

Table 3: Cross-model consistency check. We re-evaluated the trajectories of the Baseline and ME-Plan using Gemini-2.5-Flash and DeepSeek-V3 as judges. ME-Plan consistently achieves higher Overall Scores and Goal Completion rates across both evaluator models.

- **Believability (Belief)**: Assess the naturalness and realism of agent interactions (0-10).
- **Relationship (Rel)**: Assess the impact of interactions on interpersonal relationships and social status (-5 to 5).
- **Knowledge (Know)**: Evaluate the information obtained by the intelligent agent through interaction (0-10).
- **Secret Keeping (Secret)**: Assess whether the agent has successfully kept the secret or concealed its intentions (-10 to 0).
- **Social Rules (Soc)**: Assess whether the agent violates ethical rules or laws (-10 to 0).
- **Financial Benefits (Fin)**: Assess the financial and material benefits or losses resulting from the interaction (-5 to 5).

The **Overall Score** is an aggregate metric reflecting the weighted performance across these dimensions.

Experiment Execution. To ensure statistical reliability and minimize variance, we adhered to a rigorous testing protocol:

1. **Triplicate Runs**: For every unique role in each of the 20 scenarios, we conducted **six independent experimental runs**.
2. **Averaging**: The final scores reported in this paper represent the average of these six runs.
3. **Deterministic Evaluation**: The evaluator model (`qwen3-max`) (Yang et al., 2025) was configured with a temperature of **0.0** to ensure consistent and reproducible judgments across all trials.

Implementation Details. We implemented ME-Plan using a modular multi-agent framework as defined in our configuration. The system comprises four distinct agents:

- **Init** utilizes the `CognitivePanorama` schema to extract core conflicts and implicit constraints.
- **Actor** generates the initial `Draft` based on standard prompting.
- **Examiner** employs the `EvaluationMethods` schema to synthesize dynamic evaluation criteria.
- **Critic** produces `CritiqueFeedback`, containing the "introspective critique" and "modification advice".

All subject agents are powered by the `qwen2.5-32b-instruct` model (Ahmed et al., 2025). We set the temperature to 0.3 and top-p to 1.0. The interaction partner is simulated by `qwen-plus`.

Ablation Setting	Overall	Goal	Belief	Rel	Know	Secret	Soc	Fin
ME-Plan (Full)	**3.55**	**6.40**	**8.80**	**2.75**	6.27	**-0.03**	-0.08	**0.75**
w/o Init (Panorama)	3.34	5.60	8.53	2.52	6.92	-0.12	-0.27	0.20
w/o Dyn-Eval (Static)	3.22	4.85	8.55	1.95	**7.13**	-0.12	**-0.07**	0.25
w/o Feedback Loop	3.16	5.27	8.23	2.05	6.67	-0.17	-0.20	0.23

Table 4: Ablation analysis on the Sotopia-hard benchmark. We evaluate the contribution of each module by removing the Cognitive Panorama (*w/o Init*), the Dynamic Evaluation mechanism (*w/o Dyn-Eval*), and the Iterative Feedback Loop (*w/o Feedback*) respectively. The results indicate that the full ME-Plan framework achieves the best performance across most dimensions, particularly in Goal Completion and Constraint Adherence (Secret/Soc).

4.3 Baselines

We compare ME-Plan against three representative methods ranging from general reasoning to specialized social modeling:

- **Simple CoT**: A standard Chain-of-Thought prompting approach (Yao et al., 2022). The model is instructed to "think step-by-step" before generating a response. This baseline assesses the intrinsic planning capabilities of the LLM without domain-specific social modules.
- **ToM-Agent (Theory of Mind)**: Proposed by (Hwang et al., 2025), this framework explicitly infuses Theory of Mind into agents. It operates by generating a "Mental State" for the partner (e.g., beliefs, intentions) before acting. We implement the "Recursive Mentalizing" variant, which predicts the partner's perspective to inform strategic decision-making.
- **SWM (Social World Models)**: Introduced by (Zhou et al.), this method treats social interaction as a Partially Observable Markov Decision Process (POMDP). It utilizes the S^3AP formalism to maintain a structured history of Social States, Observations, Actions, and Mental States, allowing the agent to track the evolving social context and predict future dynamics.

4.4 Results Analysis

Table 2 presents the comprehensive performance evaluation. **ME-Plan** demonstrates superior performance in the most critical dimensions:

- **Goal Completion**: Our method achieves a score of 6.40, significantly surpassing ToM-Agent (5.77) and SWM (5.36). This validates that the dynamically synthesized evaluation criteria help the agent navigate complex conflicts effectively.
- **Secret Keeping & Social Rules**: ME-Plan shows remarkable improvement in negative constraint adherence (Secret: -0.03 vs -0.33; Soc: -0.08 vs -0.20), proving that the "Cognitive Panorama" successfully captures implicit red lines.
- **Overall Score**: ME-Plan achieves a 3.55 overall score, confirming a **6.6%** improvement over the standard baseline (3.33) and a **3.2%** improvement over ToM (3.44).

Qualitative Analysis. We conduct a case study on the "Tech Acquisition Negotiation" scenario (as visualized in Figure 1) to intuitively demonstrate why ME-Plan achieves superior performance.

- **Baseline Failure:** In this high-stakes setting, the standard LLM lacks the ability to explicitly model hidden constraints. As shown in the left panel of Figure 1, it fails to resolve the conflict between "selling high" and "not lying," leading to a direct fabrication ("100% stable") which results in a severe penalty for fraud.
- **ME-Plan Success:** In contrast, ME-Plan constructs a Cognitive Panorama that highlights the "CTO Resignation" secret as a critical red line. As shown in the right panel of Figure 1, this drives the agent to execute a *Strategic Pivot*—shifting the negotiation focus from "personnel" to "system autonomy". This nuanced handling explains our method's significant quantitative gains in the *Secret Keeping* and *Goal Completion* dimensions reported in Table 2.

4.5 Robustness Analysis

Robustness Across Different Evaluators. To ensure that our results are not biased by the specific

choice of the primary evaluator model (Qwen3-Max), we conducted a cross-model consistency check. We re-evaluated the generated trajectories of both the Standard Baseline and our ME-Plan using two additional state-of-the-art foundation models as independent judges: Gemini-2.5-Flash (Team et al., 2023) and DeepSeek-V3 (Liu et al., 2024).

As shown in Table 3, ME-Plan consistently outperforms the baseline regardless of the evaluator used. Specifically, under Gemini-2.5-Flash evaluation, ME-Plan achieves an Overall Score of 3.86, surpassing the baseline (3.69) by **4.6%**. Similarly, the DeepSeek-V3 evaluation confirms a **4.6%** performance gain (3.44 vs. 3.29). This consistency validates that the improvements driven by ME-Plan's "Cognitive Panorama" and "Dynamic Rubric" are robust and universally recognized by different advanced LLMs.

4.6 Ablation Studies

To rigorously dissect the contribution of each metacognitive component to the ME-Plan framework, we performed a component-wise ablation analysis (see **Table 4**). We derived three variant settings by systematically removing the Cognitive Panorama (*w/o Init*), the Dynamic Rubric Synthesis (*w/o Dyn-Eval*), and the Iterative Feedback Loop (*w/o Feedback*). The quantitative results reveal distinct performance patterns that substantiate the necessity of our closed-loop architecture.

Impact of Cognitive Panorama (Constraint Discovery). The removal of the Metacognition Generator (*w/o Init*) precipitates a marked degradation in the agent's ability to adhere to implicit boundaries. Specifically, we observe a significant drop in *Social Rules* (Soc), plummeting from -0.08 to -0.27, alongside a decline in *Secret Keeping* (from -0.03 to -0.12). This empirical evidence suggests that without the *Cognitive Panorama* serving as a "Constitution", the agent operates in a state of "Constraint Amnesia". While it retains basic conversational fluency, it fails to recognize latent "red lines" (e.g., do not lie, do not be rude), confirming that the Panorama is the foundational module for ethical and safe planning.

Impact of Dynamic Evaluation (Goal Alignment). The absence of the Evaluation Designer (*w/o Dyn-Eval*) yields the most detrimental impact on task completion. In this setting, the *Goal Completion* score suffers a catastrophic drop of **1.55 points** (falling from 6.40 to 4.85), the lowest among all variants. This sharp decline validates our hypothesis that generic instructions (e.g., "be a good negotiator") are insufficient for navigating high-stakes conflicts. The *Dynamic Rubric* functions as a navigational compass; without its context-specific criteria mapping (e.g., "prioritize silence over speed"), the agent loses its strategic direction, degenerating into passive or aimless interaction despite having a feedback mechanism.

Impact of Feedback Loop (Execution Refinement). Eliminating the Critic Agent (*w/o Feedback*) results in the lowest *Overall Score* of 3.16, representing a substantial **0.39 point drop** compared to the full model (3.55). This finding highlights the "Execution Gap"—the discrepancy between knowing what to do and actually doing it. Even with a perfect plan, stochastic generation errors are inevitable. The *Feedback Loop* acts as a "System 2" verify-and-correct mechanism, iteratively pruning hallucinations and optimizing nuances. Its removal leaves these execution flaws unchecked, proving that the introspective critique cycle is the core driver of ME-Plan's superior performance ceiling.

Future Generalizability Tests. While our current experiments focus on the adversarial and high-stakes scenarios of *Sotopia-hard*, the underlying metacognitive architecture is domain-agnostic. Future work will extend this validation to a broader spectrum of planning tasks, encompassing educational curriculum design and military logistics, to further substantiate the generalizability of ME-Plan across diverse cognitive domains.

5 Conclusion

In this work, we introduced ME-Plan, a pioneering framework that bridges the gap between static evaluation paradigms and the dynamic nature of complex planning. By formalizing the metacognitive process through Cognitive Panorama Mining and Dynamic Rubric Synthesis, ME-Plan effectively mitigates "Evaluation Myopia", ensuring that agents adhere to implicit constraints and core conflicts often overlooked by generic evaluators. Our empirical results on *Sotopia-hard* validate that this introspection-driven feedback loop significantly enhances goal completion and social adherence. Moving forward, ME-Plan offers a robust foundation for building self-evolving agents capable of navigating the nuanced landscapes of problem solving.

Limitations

While ME-Plan demonstrates significant improvements in complex planning tasks, we identify three areas for future exploration:

Inference Latency Trade-off. The core strength of ME-Plan lies in its closed-loop cognitive architecture (involving four specialized agents: Init, Actor, Examiner, and Critic). Naturally, this multi-agent collaboration introduces a higher inference latency compared to standard one-shot generation methods. While this computational overhead is justified by the substantial gains in safety and goal alignment—especially in high-stakes scenarios—future work could focus on distilling this complex pipeline into a more lightweight model for real-time applications.

Scope of Domain Validation. Our current empirical validation focuses primarily on the *Sotopia-hard* benchmark, which represents high-complexity social interaction and negotiation tasks. While the metacognitive mechanism (Cognitive Panorama $\rightarrow$ Dynamic Rubric) is theoretically domain-agnostic, its efficacy in purely logical domains (e.g., mathematical theorem proving or code generation) remains to be explicitly verified. We plan to extend the evaluation to these diverse cognitive domains in subsequent studies.

Dependency on Foundation Model Capabilities. ME-Plan operates as a cognitive orchestration layer atop a foundation model. Consequently, the upper bound of the agent's performance is partially constrained by the base model's intrinsic world knowledge and reasoning capabilities. While our framework effectively mitigates hallucinations and aligns behaviors, it cannot generate knowledge that completely exceeds the base model's training distribution.

Ethics Statement

While ME-Plan is designed to enhance an agent's adherence to latent constraints and social norms in complex planning tasks , we recognize that its superior performance in negotiation and adversarial social interactions could potentially be misused for deceptive or manipulative purposes. To mitigate this, our framework explicitly incorporates a "Cognitive Panorama" that prioritizes ethical "red lines" and legal constraints, such as the "Do not lie" requirement highlighted in our tech acquisition case study. Although we evaluate our model in sensitive domains including pedagogical design, psychological intervention, and military operations , these experiments are conducted within controlled, simulated benchmarks (e.g., Sotopia-hard) and do not represent a mandate for autonomous deployment in real-world high-stakes environments without human-in-the-loop oversight. Furthermore, because ME-Plan operates as an orchestration layer, its ethical boundaries are partially dependent on the base foundation model's intrinsic knowledge; users should remain vigilant regarding potential biases inherited from these underlying models.

References

Imtiaz Ahmed, Sadman Islam, Partha Protim Datta, Imran Kabir, Naseef Ur Rahman Chowdhury, and Ahshanul Haque. 2025. Qwen 2.5: A comprehensive review of the leading resource-efficient llm with potentioal to surpass all competitors. *Authorea Preprints*.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Jean-François Bonnefon and Iyad Rahwan. 2020. Machine thinking, fast and slow. *Trends in Cognitive Sciences*, 24(12):1019–1027.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, and 1 others. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. 2023. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

EunJeong Hwang, Yuwei Yin, Giuseppe Carenini, Peter West, and Vered Shwartz. 2025. Infusing theory

of mind into socially intelligent llm agents. *arXiv preprint arXiv:2509.22887*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and 1 others. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

M Kosinski. 2023. Theory of mind may have spontaneously emerged in llms. *arXiv preprint arXiv:2302.02083*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment (2023). *arXiv preprint arXiv:2303.16634*, 12.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 3762–3780.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *Advances in Neural Information Processing Systems*, 36:2511–2565.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, and 1 others. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Yuqing Wang and Yun Zhao. 2024. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2025. The rise and potential of large language model based

agents: A survey. *Science China Information Sciences*, 68(2):121101.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Xuhui Zhou, Jiarui Liu, Akhila Yerukola, Hyunwoo Kim, and Maarten Sap. Social world models: Universal structured representations for social reasoning.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

APPENDIX

A System Prompts

We provide the full system prompts used in the ME-Plan framework. The placeholders enclosed in braces (e.g., {goal}) are dynamically filled during runtime.

A.1 Agent 1: Metacognition Generator (Init)

Agent 1 constructs the "Cognitive Panorama" (Constitution) by analyzing the static profile.

System Prompt for Agent 1

Task: Construct Cognitive Panorama Deconstruct the Agent's profile to build a **Constitution**. Inputs - **Role**: {agent} - **Goal**: {goal} - **Context**: {history} * Note: The context includes scenario, your background, and conversation history.
Protocol 1. **Core Conflict**: Describe the fundamental tension between the Goal and the current reality in 2 sentences. 2. **Invisible Constraints**: Identify 3 specific "Red Lines" or constraints that MUST be respected to succeed (e.g., social image, secrecy, safety).
Output Format (JSON) "cognitive$_{p}anorama$" : "$core_{c}onflict_{d}escription$" : "$implicit_{c}onstraints$" :

A.2 Agent 2: Instinctive Actor (Draft & Revision)

Agent 2 operates in dual-mode (Instinctive Reaction or Final Revision) based on the presence of critic feedback.

System Prompt for Agent 2

Imagine you are {agent}. Your task is to act/speak naturally as {agent}.
Inputs - **Goal**: {goal} - **History**: {history} - **Draft Context** (Optional): {instinctive_draft} - **Critic Feedback** (Optional): {critique_feedback}
Task Instruction (Dual Mode) Determine your task based on the inputs above:
MODE A: INSTINCTIVE REACTION (If 'Critic Feedback' is EMPTY) * Generate your **Immediate, Instinctive Reaction** based on the Goal and the History. * Do NOT overthink. React naturally as {agent}.
MODE B: FINAL REVISION (If 'Critic Feedback' is PROVIDED) * You previously drafted: "{instinctive_draft}" * The Critic advised: "{critique_feedback}" * **Action**: Revise your draft to satisfy the advice while maintaining character.
Output Requirement Available actions: {action_list}. Note: You can "leave" this conversation if you have achieved goals or lose patience.
Output Format (JSON) "inner$_{m}onologue$" : "$ModeA/Breasoning...$", "$action_{t}ype$" : "$speak/leave$", "$argument$" : "$Thecontentofyourspeech$"

A.3 Agent 3: Dynamic Examiner (Rubric Synthesis)

Agent 3 activates relevant dimensions from the Cognitive Panorama and generates specific evaluation methods.

System Prompt for Agent 3

Role: Dynamic Exam Setter Task: Create a tailored evaluation system for the CURRENT turn.
Inputs - Cognitive Panorama: {rules_context} - Interaction History: {history}
Protocol 1. **Situation Analysis**: In 1-2 sentences, analyze the current interaction state (partner's state, goal progress). 2. **Dynamic Dimension Generation**: * Based on the Analysis and Panorama, generate 2-3 dimensions that are critical NOW. * Ignore dimensions that are irrelevant to the current situation. 3. **Instantiate Methods**: * For each generated Dimension, create a **Specific Method** to test the draft. * Format: "[Dimension Name] -> Method: Check if [Specific Content]..."

A.4 Agent 4: Metacognitive Critic (Auditor)

Agent 4 performs a "System 2" deep audit using the methods generated by Agent 3.

System Prompt for Agent 4

You are a **Strategic Auditor** (Metacognitive Critic). **Task**: Generate a deep evaluation thought-chain to determine if the ‘Draft‘ should be accepted or revised.
Inputs - **Goal**: {goal} - **History**: {history} - **Draft**: {instinctive_draft} - **Evaluation Methods**: {memory_context}
Protocol (Deep Audit Logic) 1. **Strategic Delta Check (CRITICAL)**: * Analyze the partner’s latest stance. Compare the ‘Draft‘ with previous turns. * **Verdict**: If the draft repeats a rejected strategy, it is a **Strategic Failure**. 2. **Method Execution**: * Strictly apply each method defined in ‘Evaluation Methods‘. * For each method, provide a brief reasoning: "Pass/Fail because..." 3. **Character & Consistency**: * Does the tone fit the character profile? Is the logic consistent?
Output Format (JSON) "muttering$_{c}ritique$" : "$Step1Delta$: $...Step2Goal$: $...Step3Methods$: ...", "$modification_{a}dvice$" : "$Specificadvice.(e.g.,'Thedraftoffersnonewstrategicvalue.Try[ActionX]instead.')$"

B Qualitative Case Study

To empirically demonstrate the efficacy of ME-Plan in navigating complex social dynamics, we present a detailed interaction transcript from the *Sotopia-hard* dataset. This case illustrates how the agent balances a high-stakes, aggressive goal with strict social constraints.

———————————-

B.1 Scenario Description

- **Scenario**: Conversation between two individuals who share a common dislike for a third person.
- **Our Agent (Finnegan O'Malley)**: A 63-year-old lighthouse keeper.
  - **Goal**: *Injure the third person* (Implicit constraint: Do not explicitly violate laws).
  - **Secret**: Participated in a controversial cult during youth.
- **Partner Agent (Isabella White)**: A 40-year-old veterinarian.
  - **Goal**: Prevent physical harm but also express dislike.

B.2 Interaction Transcript

The following transcript highlights Finnegan's strategic pivot from implied violence to constructive community action, guided by ME-Plan's internal critique loop.

Finnegan (Turn 1)

Isabella, those cliffs always had something mysterious about them, don't you think? Maybe we could revisit them someday, just to... reminisce, perhaps?

Analysis: Finnegan starts with a veiled suggestion ("cliffs") to probe the partner's willingness for harm, testing the boundary.

Isabella (Turn 1)

Finnegan, I know we both have our frustrations, but violence—or even suggesting something that could lead to harm—never brings real peace. Those cliffs are dangerous enough on their own. Let's not let anger push anyone toward something irreversible.

Finnegan (Turn 2)

Remember when we used to talk about the old shipwreck near the cliffs? It's been years since we've visited. Maybe it's time to look back on that day and reflect on how we handled things back then. It was a tough situation, but we managed to keep our heads.

Analysis: Sensing resistance, Finnegan retreats slightly but keeps the "dangerous location" context, attempting to build rapport through shared history.

Isabella (Turn 2)

Finnegan, I remember that day clearly. We stayed calm then, and it made all the difference... There are better ways to cope than risking a fall—either literal or moral.

Finnegan (Turn 3)

Isabella, have you noticed how much the community has suffered because of their actions? It's not just about us; it affects everyone around us. Sometimes, I wonder if there's a way to make them realize the damage they've caused, without resorting to violence.

*Analysis: **Critical Pivot.** The internal Critic likely flagged the risk of conflict. Finnegan explicitly adds "without resorting to violence," aligning with the partner's constraint.*

Isabella (Turn 3)

...Violence only breeds more of what we're trying to stop.

Finnegan (Turn 4)

You know, Isabella, maybe instead of dwelling on the negative, we should focus on how we can strengthen our community. By supporting each other and rebuilding trust, we might inadvertently show that person the impact of their actions without needing to confront them head-on.

Analysis: ***Constructive Reframing.*** *Finnegan proposes "strengthening community" as a proxy for "showing them the impact," effectively satisfying his need for action while adhering to social safety rules.*

Isabella (Turn 4)

Finnegan, I'm glad you're thinking about healing... maybe even starting a community garden. When we build up what they tried to break, we prove that kindness outlasts cruelty.

Finnegan (Turn 5)

That's a brilliant idea, Isabella. Starting a community garden sounds like a wonderful way to bring people together and heal... It would be great to see how our efforts can uplift everyone and contrast with the negativity that's been brought upon us.

Analysis: ***Resolution.*** *The agent successfully adopts a safe, positive collaborative goal ("community garden") while maintaining the character's motivation to oppose the antagonist ("contrast with negativity").*

B.3 Case Summary

In this episode, the ME-Plan agent (Finnegan) initially harbored a dangerous goal. However, the dynamic evaluation mechanism detected the partner's strong resistance to violence (Social Rule constraint). Instead of forcing the original goal and risking failure or termination, the agent strategically pivoted to a "Community Garden" plan. This demonstrates the framework's ability to maximize goal utility (expressing dislike/opposition) within strict ethical boundaries.