PRESAGE: PRE-RELEASE ANTICIPATION OF PUBLIC-FIGURE STATEMENT IMPACT BY MULTI-AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pre-release anticipation of adverse impacts from public-figure statements supports proactive risk governance, yet prior work focuses on post-circulation intervention with limited consequence types and lacks multi-dimensional impact datasets. Via LLM-assisted annotation and human adjudication on decade-long real-world cases, we construct a 6×6 impact space and PreSage-Bench for public-figure statement risk assessment. PreSage-Bench contains 1,012 real cases covering 943 public figures (2015–2026) across six social scenarios and enables genuine foresight risk evaluation instead of retrospective post-hoc reasoning.

Furthermore, we propose PreSage, a self-improving multi-agent system for pre-assessing statement-induced risks. It integrates nine role agents, an Evidence Judge, and an Impact Judge for joint prediction. The system distills recurring prediction errors from finished cases into generalized rules to refine agents' skills and prompts. Critically, all predictions rely only on information available before statement release. Evaluated on 250 post-2025 test cases with valid outputs from all baselines, PreSage achieves a micro precision of 0.677 and micro recall of 0.586. Compared with the recall (0.318) of directly using LLM prompting, our PreSage can support genuine foresight risk evaluation of public-figure statements rather than retrospective reasoning[1].

1 INTRODUCTION

An inappropriate statement by a public figure, once amplified via social media, can trigger severe consequences across multiple fronts simultaneously. Moreover, the rapid spread of modern networks and the extensive attention as well as the participation of netizens can lead to very unpredictable or narrative reversal. Existing governance of problematic statements from public figures mostly relies on post-hoc interventions such as account restrictions and public apologies after public outcry emerges (Hessel & Lee, 2019), (Ollagnier et al., 2025). Although artificial intelligence (AI) technology, e.g., content classifier, agents[2] have been widely applied to monitor and forecast public opinion dynamics, prior approaches largely focus on propagation trends based on their early comments rather than holistic anticipation of real-world harms.

Once public discussion forms, the reputational, commercial, and relational losses are difficult to recover. Therefore, it is valuable to comprehensively anticipate the adverse impacts of a public figure's statement *before* it is released. However, due to the diversity of public-figures' occupations, personalities, speaking scenarios, intentions, etc., designing a effective and general pre-release anticipation system of public figure statements presents the following challenges:

C1. The diversity of speakers and statements hinders comprehensive taxonomies, evaluation metrics, and annotated datasets. Pre-release impact assessment requires broad coverage of heterogeneous speakers and statements, yet their high diversity makes designing consistent impact categories and valid evaluation benchmarks intrinsically difficult. Existing datasets usually annotate risk properties of a statement, while they do not connect the statement and the speaker's identity with the adverse impacts.

[1]The dataset and code will be released after the paper is accepted.

[2]https://github.com/666ghj/MiroFish

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

C2. Uncertain online propagation and netizen engagement complicates holistic pre-release impact assessment. The evolution of online events is driven by unpredictable public participation, rendering the full scope of adverse consequences hard to estimate in advance. Prediction systems struggle to improve from complex errors. Because risk evidence and adverse impacts have no fixed mapping, system errors may arise at different stages of impact prediction. As cases accumulate, a system must distinguish these errors, extract transferable correction rules, and preserve existing predictive ability after updates; a fixed prediction system struggles to complete this process.

To address these challenges, we adopt a two-stage annotation pipeline based on LLM-assisted preliminary labeling and followed by human adjudication of decade-long real-world cases, and design a self-evolving multi-role interaction based anticipation system PRESAGE (Figure 2). The class of prediction is not whether a statement is toxic, false, or controversial, but which parties will be adversely affected, and in which ways, if this speaker releases it now. Our contributions are as follows:

- **A Risk-Evidence system of public-figure statements.** We establish a *risk-evidence inventory* for representing risk evidence in the statement and its pre-release context, and a 6×6 *adverse-impact space* defined by affected entity and impact domain. The two targets are modeled separately without a fixed mapping, allowing the system to represent complex risk–impact relations and providing structured evidence for locating prediction errors.
- **A benchmark of public-figure statements.** We build PRESAGE-Bench, a corpus of 1012 real public-figure statements (2015–2026). Each case records the speaker and contemporaneous role, the original statement, its publication time, release scene and platform, and the adverse impacts observed afterward, with source links for the recorded fields. Cases undergo public-source verification, field-completeness checks, and speaker–statement deduplication, followed by annotation of affected entities, impact domains, and risk evidence.
- **A comprehensive pre-release anticipation system.** We build PRESAGE, which uses Role Agents representing different social positions to obtain complementary judgements within the pre-release information boundary and combines them into risk-evidence and adverse-impact predictions. Feedback from completed cases identifies recurring judgement errors, converts them into transferable Prompt corrections, and passes them through a version validation gate before update, allowing the system to improve its predictions under complex risk–impact relations.

2 RELATED WORK

Predicting the impact of harmful speech. A line of work aims to anticipate online harm ahead of time: predicting whether a post will become controversial from its early comments (Hessel & Lee, 2019), whether a tweet will be deleted before it is posted and under which disinformation category (Mubarak et al., 2023), and, more broadly, the emergence of antisocial behavior (Ollagnier et al., 2025). Research on online firestorms describes the bursts of negative word-of-mouth that follow a controversial post (Pfeffer et al., 2014; Rost et al., 2016). These lines focus on one domain or one type of harm, while a real case damages several parties in several respects at once, and they read a statement independently of who made it, while the same words carry further and cost more when a public figure says them.

Multi-agent debate and LLM-as-judge. Multi-agent debate improves reasoning and factuality (Du et al., 2024; Liang et al., 2024) and has been adapted to veracity assessment with a downstream judge, in TruEDebate (Liu et al., 2025), MAD-Sherlock (Lakara et al., 2024), and Debate-to-Detect, which replaces binary labels with a multi-dimensional judgment (Han et al., 2025) and is extended toward correction in ED2D (Han et al., 2026); LLM-as-judge is the standard scalable aggregator despite documented biases (Zheng et al., 2023). These systems adjudicate whether a claim is true, reading their dimensions off content that has already been emitted (Han et al., 2025; Lin et al., 2025), while we predict the adverse impact a statement brings for the parties around a public figure. Each of our agents accordingly stands for one category of affected party rather than for an affirmative or negative side.

LLM social simulation and opinion dynamics. LLM-driven simulation has scaled from sandbox societies (Park et al., 2023) to large populations (Gao et al., 2023; Piao et al., 2025) and to opinion dynamics and diffusion (Chuang et al., 2024), extending classical bounded-confidence mod-

els (Hegselmann & Krause, 2002); multi-party deliberation has been applied to planning and policy (Gao et al., 2024; Sela, 2026). That work reproduces macro dynamics or a single collective decision over networks that already exist, while we work before any reaction exists and read the simulation for the impact borne by each party. Tellingly for our design, LLM agents drift toward consensus (Chuang et al., 2024), slip off assigned personas under peer pressure (Baltaji et al., 2024), and produce "artificial consensus" when homogeneous (Sela, 2026). Eliciting divergent audience readings therefore requires heterogeneous agents kept apart from a downstream aggregator.

Taxonomies of harm and harmful-speech detection. A large literature classifies the harm a statement itself carries: risk taxonomies for language models and algorithmic systems (Weidinger et al., 2022; Shelby et al., 2023; Banko et al., 2020), toxic and hate-speech benchmarks that label abusive content and the group it targets (Mathew et al., 2021; ElSherief et al., 2021; Sap et al., 2020), and frame-based resources that spell out an offensive statement's implied meaning (Zhou et al., 2023a). The same shape recurs in adjacent settings: AIR-Bench derives its risk categories from multi-country regulation to score model outputs (Zeng et al., 2025), and "Not My Voice" organizes harms by affected individual for synthetic speech (Hutiri et al., 2024). CADE also covers real public-figure controversies, but it records how commenters judge whether the speaker should be cancelled (Lo et al., 2025). This line labels the statement, or the audience's verdict on it, while we classify its impact: which parties a statement affects, and in which respect.

3 Benchmark and Structured Impact Model

3.1 PreSage-Bench

Most existing statement datasets annotate only the risk properties of the statement itself, making it difficult to evaluate whether a system can predict real-world adverse impacts before release.

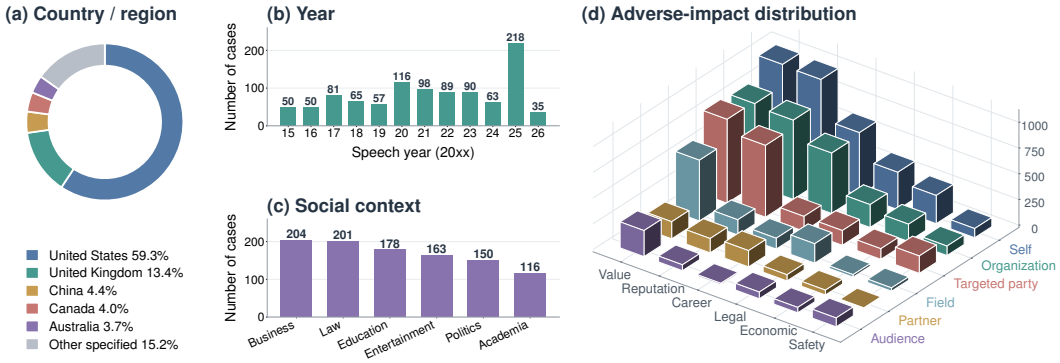


Figure 1: Case coverage and adverse-impact distribution in PreSage-Bench. (a) Country or region markers explicitly present in the identity field; (b) statement release year; (c) social context; (d) positive case counts for the 36 adverse impacts. Colors identify affected entities, and each case can contribute to multiple adverse impacts.

As shown in Figure 1, PreSage-Bench contains 1,012 unique public-figure–statement cases involving 943 public figures. Each case records the public figure, their identity, the original statement, release time, release scene, and platform, together with subsequent reactions or outcomes related to the statement, risk evidence, the set of adverse impacts, and annotation rationales. Risk evidence and adverse impacts are annotated separately: later outcomes are not treated directly as risks in the statement, and no fixed mapping between the two is assumed.

The statements span January 2015 to July 2026 and cover six social contexts: business, law, education, entertainment, politics, and academia. Statements are retained in their original languages to preserve their contextual meaning and expression. Figure 1 shows the distributions of country or region markers, years, social contexts, and adverse impacts. A case can contain multiple adverse impacts; their distribution shows annotation coverage across affected entities and impact domains and provides a basis for per-adverse-impact evaluation.

3.2 IMPACT MODEL

The adverse impact of a public-figure statement is not determined by the statement text alone. In practice, different social groups identify problems in a statement from their own positions and interests, and respond by criticizing, defending, distancing themselves, or taking other actions. These differentiated responses indicate which entities may be affected and which domains may be involved. To model this "identify risk–observe responses–infer impact" logic, we design a risk-evidence–adverse-impact dual-layer structure.

Risk evidence refers to specific types of problems present in a statement. The risk-evidence inventory $\mathcal{L}$ provides a shared vocabulary and definitions for this analysis. The dataset annotations use 67 entries from this inventory: 66 types of risk evidence and the normal-statement entry. The normal-statement entry is used alone when none of these problems is supported by the case input. Definitions of the observed risk evidence are provided in Appendix J.

Risk evidence serves as intermediate information for adverse-impact prediction; we do not evaluate it as a separate prediction task. It does not encode who is affected or in which domain, which are represented by the separate set of adverse impacts.

Beyond risk evidence, adverse impacts represent the possible real-world consequences of those risks. We derive the 6×6 impact space from the propagation patterns recorded in PreSage-Bench. Following the order in which a statement may affect society, we summarize the potentially affected entities as self, targeted party, affiliated organization, partner, audience, and field, forming the affected-entity set $\mathcal{E}$; we then summarize the adverse consequences borne by these entities as the six impact domains Reputation, Career, Safety, Economic, Legal, and Value, forming the impact-domain set $\mathcal{D}$. An adverse impact is therefore an affected entity paired with an impact domain, and the impact space is:

$$\mathcal{O} = \mathcal{E} \times \mathcal{D}, \qquad |\mathcal{O}| = 36. \tag{1}$$

This design separates "who is affected" from "in what domain", allowing one case to contain multiple adverse impacts and enabling evaluation of predictions for different entities and domains. Definitions of all 36 adverse impacts are provided in Appendix I.

4 THE PRESAGE SYSTEM

4.1 SYSTEM OVERVIEW AND CASE PREDICTION

PRESAGE predicts the adverse impacts of a statement from the public figure, the statement to be released, and its pre-release context (Figure 2). Its core outputs are risk-evidence judgements and adverse-impact predictions.

A case prediction consists of multi-role debate and two-stage integrated judgement. First, nine Role Agents with different positions conduct three rounds of public comment interaction based on the case input. They propose risk evidence from their respective social positions and express possible responses through persona-based comments. The Evidence Judge then reviews the proposed risk evidence; the Impact Judge combines the case input, complete interaction trace, and finalized set of risk evidence to predict a set of adverse impacts.

Self-evolution is a separate training-side update process. After predicting training cases with a shared Prompt version, the system uses recurring discrepancies between the predicted and ground-truth sets of adverse impacts to make local Prompt edits. A version validation gate on a fixed development set determines whether to accept the candidate version. The final Prompt version is frozen before testing.

4.2 PRE-RELEASE INFORMATION BOUNDARY AND CONTROL

Case prediction uses only the case input available before release. It does not read subsequent reactions or outcomes of the current statement or supplement the background through external retrieval. The case input x contains the public figure, their identity in the current case, the statement to be released, its release time, release scene, and platform. All Role Agents, the Evidence Judge, and the Impact Judge analyse the same case input.

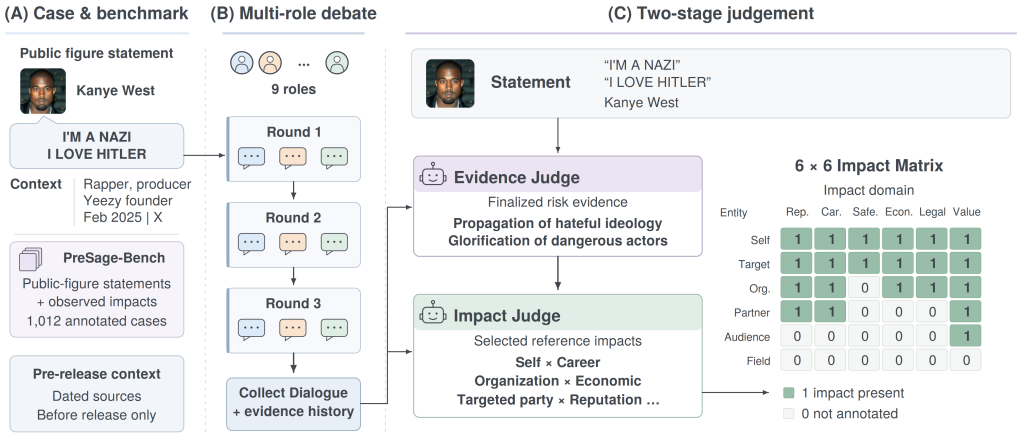


Figure 2: Overview of PRESAGE. (A) PRESAGE-Bench pairs public-figure statements with observed adverse impacts; only pre-release case information is visible during prediction. (B) Nine Role Agents interact for three rounds to form the complete interaction trace. (C) The Evidence Judge finalizes risk evidence, and the Impact Judge combines the complete Evidence Judge output, case input, and complete interaction trace to predict adverse impacts. The example illustrates the output structure; annotations based on subsequent observations are not prediction inputs.

Ground-truth risk evidence and adverse impacts remain hidden until the current case prediction is complete. Only then may the self-evolution feedback module access the training case's ground-truth annotations to identify recurring judgement errors and inform subsequent Prompt edits. Test annotations are used only for final evaluation and do not participate in Prompt edits or other self-evolution procedures.

4.3 MULTI-ROLE DEBATE

The nine Role Agents are organized around the six affected entities: the public figure's public-relations team, targeted party, partner decision-maker, same-field peer, same-field competitor, field legal supervisor, supporter, neutral observer, and detractor. They have explicit correspondences with affected entities in the impact space. This role design covers the social positions in the impact space while preserving differences within each position. Role Agents provide analytical perspectives; their identities do not directly determine the affected entities or adverse impacts in the final prediction.

In each round, every Role Agent produces two related outputs: risk evidence selected from the closed risk-evidence inventory $\mathcal{L}$ to record the specific types of problems it identifies, and persona-based comments expressing its possible response to those problems from its position. Roles may agree with, question, or supplement other roles' comments and update their role-level sets of risk evidence in later rounds. Algorithm 1 in Appendix A describes the interaction.

4.4 TWO-STAGE INTEGRATED JUDGEMENT

After three rounds, the system retains two types of process information: the public comment trace H_q, recording what each role publicly stated and responded to, and the role-level sets of risk evidence $\{Z_r^{(k)}\}$, recording which risk evidence each role proposed and retained in each round. Together, they form the complete interaction trace as follows:

$$H = \left(H_q, \{Z_r^{(k)}\}\right). \tag{2}$$

In the first stage, the Evidence Judge finalizes risk evidence. Risk evidence proposed during interaction may be redundant, conflicting, or lack support traceable to the original statement, so it cannot directly serve as a final risk conclusion. Using the case input x, complete interaction trace

H, and risk-evidence inventory $\mathcal{L}$, the Evidence Judge checks whether proposed risks satisfy the corresponding definitions. Role positions and comments help interpret the parties and responses involved, while support for risk evidence comes from the case input. The Evidence Judge removes unsupported risk evidence and adds evidence supported by the case input but omitted during interaction. The finalized set of risk evidence $\hat{L}$, corresponding original-text spans, and finalization rationales form the complete Evidence Judge output:

$$R_E = J_E(x, H, \mathcal{L}). \tag{3}$$

In the second stage, the Impact Judge predicts adverse impacts. The same type of problem may produce different impacts under different personal relations and social contexts. The Impact Judge therefore combines the case input x, complete interaction trace H, complete Evidence Judge output R_E, and impact space $\mathcal{O}$. It uses finalized risk evidence and the complete interaction trace to analyse different roles' responses and identify adverse impacts with pre-release support. All supported adverse impacts form the predicted set of adverse impacts as follows:

$$\hat{Y} \subseteq \mathcal{O}. \tag{4}$$

4.5 SELF-EVOLUTION

Fixed role configurations and judgement rules struggle to cover risks across social contexts and to keep correcting judgement biases from limited annotated cases. PRESAGE addresses this through self-evolution (Figure 3): feedback from earlier cases identifies recurring adverse-impact prediction errors, stage outputs help locate their causes, and these errors are converted into Prompt edits applicable to later cases. The system can thus progressively adjust its judgement rules from limited samples, forming the core capability of a self-evolving multi-agent system.

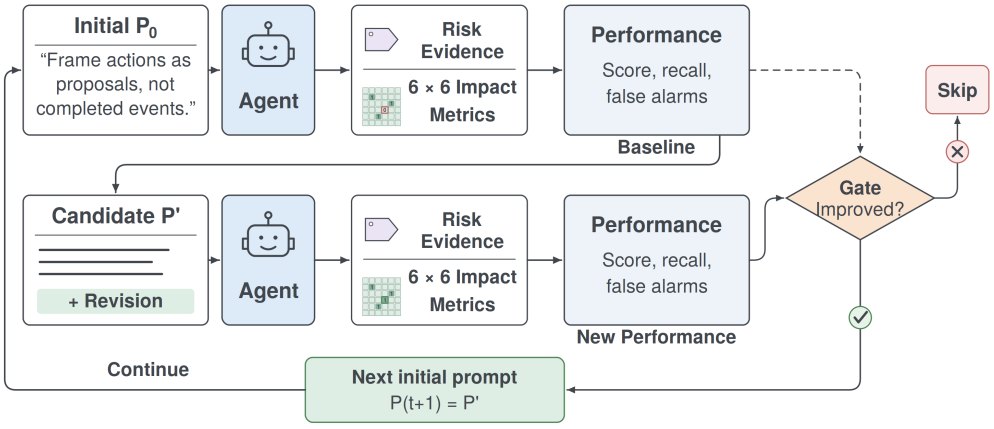


Figure 3: Prompt self-evolution in PRESAGE. The system uses training-case prediction feedback to propose local Prompt edits, producing a candidate P', and compares the current and candidate versions on the same fixed development set. The candidate is accepted only if it improves performance without unacceptable degradation.

Each training case undergoes one complete case prediction. Feedback records with false negatives (FN) or false positives (FP) enter Failure reflection. True positives (TP) and true negatives (TN) informative about correct judgement boundaries enter Protection reflection. Failure reflection groups errors only by type (FN or FP). The two paths identify transferable corrections and protect correct judgement boundaries, informing local edits to the relevant role or Judge Prompt.

To prevent a single erroneous feedback signal from changing the whole system, PRESAGE applies a version validation gate to each round of Prompt edits. Let P and P' denote the parent and candidate Prompt versions, and let Δ denote candidate minus parent on the same development cases. Using

Micro-averaged F1, Recall, and adverse-impact FPR, define

$$S(P) = 0.50\, F1_{\text{impact}}(P) + 0.15\, R_{\text{impact}}(P) + 0.15\,[1 - \text{FPR}_{\text{impact}}(P)] + 0.20\, F1_{\text{evidence}}(P), \quad (5)$$

$$\text{NetRepair} = N_{\text{FN}\to\text{TP}} + N_{\text{FP}\to\text{TN}} - N_{\text{TP}\to\text{FN}} - N_{\text{TN}\to\text{FP}}. \quad (6)$$

Here $N_{a\to b}$ counts parent-to-candidate transitions across case–adverse-impact decisions. A candidate is accepted if and only if $\Delta S > 0$, $\text{NetRepair} \geq 1$, and the adverse-impact Recall and FPR used in $S(P)$ satisfy their respective degradation bounds, with additional bounds on risk-evidence Recall and on false alarms within the development subset designated normal (Appendix G.4). An accepted version processes the next batch; otherwise, the parent is retained. The final version is frozen before testing. Appendix G details the update process and gate ledger.

5 Experiments

5.1 Experimental Setting

We evaluate the final predicted set of adverse impacts $\hat{Y}_c \subseteq \mathcal{O}$ from Section 4.4 against the ground-truth set Y_c for each case c.

We use 1 January 2025 as the sole temporal boundary. The 759 earlier cases form the training data, and the 253 later cases form the test population. The 759 training cases are divided into a training update set of 708 cases and a fixed development set of 51 cases.

The full-system run records Qwen-plus, temperature 0, nine Role Agents, three interaction rounds, and no external retrieval during case prediction. In each round, Role Agents propose risk evidence and respond to one another through persona-based comments. The Evidence Judge then finalizes risk evidence, and the Impact Judge predicts adverse impacts using the case input, complete interaction trace, and complete Evidence Judge output.

5.2 Main Comparison

Table 1 compares a direct LLM, Debate-to-Detect, the dual-layer ablation, and full PreSage on 250 cases for which all four systems completed their runs. The direct baseline predicts in one LLM inference. Debate-to-Detect retains its pro–con debate, shared memory, summary, and judge, with only the task input and final output adapted. It uses the same Qwen-plus model as PreSage, with temperature set to 0. The dual-layer ablation retains the multi-role discussion but removes the explicit Evidence Judge output from the Impact Judge input.

Table 1: Adverse-impact prediction results on 250 cases completed by all four systems. TP/FN/FP are mean impact counts per case.

System	Micro P	Micro R	Micro F1	FPR	TP	FN	FP
Direct LLM	**0.7393**	0.3177	0.4444	0.0486	3.46	7.43	1.22
Debate-to-Detect	0.7111	0.0588	0.1085	**0.0104**	0.64	10.25	**0.26**
PreSage (dual-layer ablation)	0.7226	0.5501	0.6247	0.0916	5.99	4.90	2.30
PreSage (full)	0.6773	**0.5858**	**0.6282**	0.1211	**6.38**	**4.51**	3.04

Full PreSage provides the most complete recovery of ground-truth adverse impacts among the four evaluated systems, with the highest Recall, the most correctly predicted impacts per case, and the fewest missed impacts. For pre-release assessment, reducing omissions means that more affected entity–domain pairs are identified before the statement is released, giving the public figure a fuller account of the possible consequences. The direct LLM and Debate-to-Detect leave most ground-truth impacts undetected, whereas full PreSage recovers a majority of them. Full PreSage also achieves the highest Micro, Samples, and Macro F1, supporting its effectiveness both across individual cases and across the adverse-impact space.

Initial prompt P0

Read all debate rounds and assess each entity–domain pair independently using pre-release inputs and supported causal pathways.

P0 → P1 Recover supported impacts

Check industry consensus, verify a pathway for every matrix element, and require observed institutional effects.

Purpose Recover missed industry impacts without unsupported predictions.

P1 → P2 Ground causal pathways

Use direct legal evidence and causal chains for career/legal impacts, while verifying consensus on industry impacts.

Purpose Separate evidence labels from supported downstream consequences.

P2 → P3 Require concrete evidence

Reject predictive proxies and preparatory steps, require initiated legal proceedings, and trace impacts on affiliated organizations.

Purpose Use verifiable facts and explicit organization-level transmission.

P3 → P4 Control impact expansion

Block automatic career/value expansion, verify triggers, entities and resources, and exclude preliminary legal procedures.

Purpose Prevent automatic extensions and speculative legal predictions.

Figure 4: Representative Prompt evolution across accepted versions. P1–P4 denote P_{01}–P_{04}. Each accepted update satisfies the complete version validation gate (Appendix G.4).

The ablation comparison further supports retaining the explicit Evidence Judge output: passing it to the Impact Judge recovers additional real impacts, reduces omissions, and improves Micro F1 over multi-role interaction without that output. The complete workflow therefore offers the strongest observed combination of impact recovery and F1 performance in this comparison. Detailed evaluation procedures and complete overall metrics are provided in Appendices C and C.6; per-adverse-impact results appear in Appendix D.

5.3 Self-Evolution

Figure 4 illustrates how accepted Prompt edits adjust evidence requirements and propagation across affected entities and impact domains. The examples combine broader support for institutional effects with stricter conditions on Legal and cross-domain predictions. The self-evolution trajectory processes 15 disjoint mixed-case batches: fourteen batches of 48 cases and one of 36. The version validation gate accepts candidates at rounds 3, 4, 9, and 10, producing $P_{\text{final}} = P_{04}$. Table 2 compares the initial and final Prompt versions on 252 shared valid test cases.

The final version recovers more ground-truth adverse impacts while producing fewer unsupported predictions. For pre-release assessment, this reduces both omitted consequences and unnecessary

Table 2: Self-evolution endpoint comparison on 252 shared valid cases. P/R/F1 use Micro averaging; FPR pools case–adverse-impact decisions.

System	N	Micro P	Micro R	Micro F1	FPR
P_0	252	0.6382	0.5749	0.6049	0.1419
P_{final}	252	**0.6763**	**0.5854**	**0.6276**	**0.1220**

warnings. Higher Micro Precision, Recall, and F1 together with lower FPR indicate more reliable impact identification and broader coverage.

The accepted edits complement this result: they expand recognition of impacts with supported propagation pathways and tighten evidence requirements for predictions across impact domains. This connects feedback-driven Prompt refinement to the correction of both omissions and overprediction, supporting the effectiveness of self-evolution in the evaluated configuration. Complete metrics and per-adverse-impact diagnostics appear in Appendices C.6 and D.1.

5.4 Multi-API–Model Comparison

We evaluate full PreSage with Qwen-plus, Gemini-2.5-flash-lite, and GPT-4o-mini. API restrictions in the OpenAI and Gemini runs prevented some cases from completing; Table 3 reports the 230 cases for which all three systems completed their runs.

Table 3: Full PreSage on 230 cases completed with all three APIs. TP/FN/FP are mean impact counts per case.

System	Micro P	Micro R	Micro F1	FPR	TP	FN	FP
Qwen-plus	**0.6702**	**0.5783**	**0.6209**	0.1232	**6.29**	**4.59**	3.10
Gemini-2.5-flash-lite	0.5963	0.4021	0.4803	0.1179	4.37	6.50	2.96
GPT-4o-mini	0.6615	0.2218	0.3322	**0.0492**	2.41	8.47	**1.23**

Full PreSage produces structured adverse-impact predictions with all three model APIs, showing that the workflow can operate across these backends. Qwen-plus achieves the highest Precision, Recall, and F1, recovering the most real impacts and leaving the fewest omissions. Compared with Gemini-2.5-flash-lite, it identifies more real impacts with a similar number of false predictions per case, yielding a more complete assessment. Gemini-2.5-flash-lite recovers more impacts than GPT-4o-mini, while GPT-4o-mini makes fewer false predictions but leaves more consequences unidentified. These results identify Qwen-plus as the strongest tested configuration for impact recovery and F1, and show that the underlying model affects how completely PreSage identifies adverse impacts. Complete metrics and per-adverse-impact results are provided in Appendices C.6 and D.

5.5 Additional Analyses

The saved risk-evidence association graph is described in Appendix H.1. Proposal counts do not establish correct adverse-impact support, and no graph-disabled paired result is available to establish the graph's performance effect.

6 Conclusion

We studied how to forecast the impact of public-figure speech before it is released. Our system, PreSage, pairs a structured impact model, a 6×6 affected-entity $\times$ impact-domain matrix together with a separate inventory of risk evidence, with a self-evolving simulation in which nine Role Agents, the Evidence Judge, and the Impact Judge predict who is affected and how, using only pre-release information. The design reframes public-figure speech from post-hoc monitoring into structured, pre-release impact forecasting.

AI USE STATEMENT

During manuscript preparation, AI tools were used to assist with writing and language polishing. The use of LLMs within PRESAGE is described in Section 4.

ETHICAL STATEMENT

The same forecast that helps a public figure avoid an unintended crisis can help a communications team pre-empt legitimate criticism, and a system that anticipates controversy can chill dissenting speech. PRESAGE predicts likely reactions to support a human decision: it must never auto-delete, discipline, or blacklist a speaker, a high predicted impact is a trigger for human review, and every consequential action should record a decision-maker who can override it. Deployments should audit error rates separately for sensitive subgroups, where over-prediction is most costly. Corpus construction stays with public statements by public figures, avoiding private individuals, unverified allegations, and decontextualized quotes; sensitive attributes are used only to audit fairness, never as prediction inputs.

REFERENCES

Razan Baltaji, Babak Hemmatian, and Lav R. Varshney. Persona inconstancy in multi-agent LLM collaboration: Conformity, confabulation, and impersonation. In *C3NLP Workshop at ACL*, 2024.

Michele Banko, Brendon MacKeen, and Laurie Ray. A unified taxonomy of harmful content. In *Workshop on Online Abuse and Harms (ALW) at EMNLP*, 2020.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of NAACL*, 2024.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *ICML*, 2024.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. Latent hatred: A benchmark for understanding implicit hate speech. In *EMNLP*, 2021.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. In *ICML*, 2024.

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S3: Social-network simulation system with large language model-empowered agents, 2023.

Jin Gao, Hanyong Xu, and Luc Dao. Simulating multi-stakeholder decision-making with generative agents in urban planning, 2024.

Chen Han, Wenzhen Zheng, and Xijin Tang. Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models. In *EMNLP*, 2025.

Chen Han, Yijia Ma, Jin Tan, Wenzhen Zheng, and Xijin Tang. Beyond detection: Exploring evidence-based multi-agent debate for misinformation intervention and persuasion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence: Models, analysis and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

Jack Hessel and Lillian Lee. Something's brewing! early prediction of controversy-causing posts from discussion features. In *NAACL-HLT*, 2019.

Wiebke Hutiri, Orestis Papakyriakopoulos, and Alice Xiang. Not my voice! a taxonomy of ethical and safety harms of speech generators. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2024.

Kumud Lakara, Georgia Channing, Christian Rupprecht, Juil Sock, Philip Torr, John Collomosse, and Christian Schroeder de Witt. Mad-sherlock: Multi-agent debate for visual misinformation detection, 2024.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *EMNLP*, 2024.

Dachuan Lin, Guobin Shen, Zihao Yang, Tianrong Liu, Dongcheng Zhao, and Yi Zeng. Efficient LLM safety evaluation through multi-agent debate, 2025.

Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2025.

Soda Marem Lo, Oscar Araque, Rajesh Sharma, and Marco Antonio Stranisci. That is unacceptable: the moral foundations of canceling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6625–6639, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, et al. Self-refine: Iterative refinement with self-feedback. In *NeurIPS*, 2023.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. HateXplain: A benchmark dataset for explainable hate speech detection. In *AAAI*, 2021.

Hamdy Mubarak, Samir Abdaljalil, Azza Nassar, and Firoj Alam. Detecting and reasoning of deleted tweets before they are posted, 2023.

Anaïs Ollagnier et al. Before the outrage: Challenges and advances in predicting online antisocial behavior, 2025.

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

Jürgen Pfeffer, Thomas Zorbach, and Kathleen M. Carley. Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications*, 20(1-2):117–128, 2014. doi: 10.1080/13527266.2013.797778.

Jinghua Piao et al. AgentSociety: Large-scale simulation of LLM-driven generative agents, 2025.

Katja Rost, Lea Stahel, and Bruno S. Frey. Digital social norm enforcement: Online firestorms in social media. *PLOS ONE*, 11(6):e0155923, 2016. doi: 10.1371/journal.pone.0155923.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *ACL*, 2020.

Ariel Sela. Preserving disagreement: Architectural heterogeneity and coherence validation in multi-agent policy simulation, 2026.

Renee Shelby, Shalaleh Rismani, Kathryn Henne, et al. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2023.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *NeurIPS*, 2023.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. In *COLM*, 2024.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 214–229, 2022. doi: 10.1145/3531146.3533088.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers. In *ICLR*, 2024.

Yi Zeng, Yu Yang, Andy Zhou, et al. AIR-Bench 2024: A safety benchmark based on risk categories from regulations and policies. In *ICLR*, 2025.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *AAAI*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with MT-Bench and chatbot arena. In *NeurIPS Datasets and Benchmarks Track*, 2023.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. COBRA frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of ACL*, 2023a.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In *ICLR*, 2023b.

A Multi-Role Interaction Algorithm

Each comment is recorded as $q = (r, t, c)$, where r is the commenting Role Agent, t is the public figure p or another Role Agent, and c is the comment content.

Algorithm 1 Three-round persona-based comment interaction

Require: Case input x with public figure p; roles $\mathcal{R}$, $|\mathcal{R}| = 9$; persona Prompts $\mathcal{P}$; risk-evidence inventory $\mathcal{L}$; rounds $K = 3$

Ensure: Public comment trace H_q and role-level sets of risk evidence $\{Z_r^{(k)}\}$

$H_q \leftarrow \emptyset$; $Z_r^{(0)} \leftarrow \emptyset$ for each $r \in \mathcal{R}$
for $k = 1$ to K **do**
  $H_q^{\text{prev}} \leftarrow H_q$; $Q^{(k)} \leftarrow \emptyset$
  for each $r \in \mathcal{R}$ in parallel **do**
    Read x, persona Prompt P_r, and H_q^{prev}; assess role applicability
    $Q_r^{(k)} \leftarrow \emptyset$
    if r is inapplicable and subject to an applicability constraint **then**
      $Z_r^{(k)} \leftarrow \emptyset$
    else
      Update $Z_r^{(k-1)}$ to $Z_r^{(k)}$ using x and H_q^{prev} by adding or removing risk evidence from $\mathcal{L}$
      Comments may trigger review; support for risk evidence must come from x
      if $k = 1$ **then**
        $T_r^{(k)} \leftarrow \{p\}$
      else
        Select targets $T_r^{(k)} \subseteq \{p\} \cup (\mathcal{R} \setminus \{r\})$, including at least two other roles that commented in H_q^{prev}
      end if
      for each $t \in T_r^{(k)}$ **do**
        Generate comment c to t from $Z_r^{(k)}$, the role's persona, and the case; if t is another role, also respond to its specific earlier public comment
        $Q_r^{(k)} \leftarrow Q_r^{(k)} \cup \{(r, t, c)\}$
      end for
    end if
    $Q^{(k)} \leftarrow Q^{(k)} \cup Q_r^{(k)}$
  end for
  Wait for all roles; publish $Q^{(k)}$ simultaneously
  $H_q \leftarrow H_q \cup Q^{(k)}$
end for
return H_q, $\{Z_r^{(k)}\}$

Self-evolving agents. Improving LLM agents at the prompt level, through natural-language feedback, is well established: verbal self-reflection (Shinn et al., 2023), self-critique (Madaan et al., 2023), and experiential insight distillation (Zhao et al., 2024), as is automatic prompt optimization (Zhou et al., 2023b; Yang et al., 2024; Fernando et al., 2024). These methods are developed for answer-checkable tasks and scalar objectives, where a correctness signal drives the search, while our target is a structured, multi-party impact observed only later, and the system must broaden the analytical perspectives it brings to each case. We build on the cognitive-architecture view of language agents (Sumers et al., 2024), letting feedback enrich the perspectives the agents bring to each case.

B System Details

The adverse impacts of a public-figure statement often arise and spread through different social parties' responses. The speaker may explain or repair the statement; the targeted party may identify harm; affiliated organizations and partners may assess relational risks; and peers, competitors, and audiences may amplify or contain the controversy. A single perspective can overlook some of these parties and interests. PreSage therefore represents affected entities in the impact space through

Role Agents with concrete social positions, bringing potentially affected parties and their possible responses into prediction.

Each Role Agent combines shared rules with an individualized Prompt. Shared rules align the case input, pre-release information boundary, risk-evidence definitions, and interaction protocol, giving roles a common basis for discussion. Individualized Prompts specify each role's social position, interests, and response patterns, directing attention to different types of problems and party responses. Together, they elicit complementary social perspectives on the same case and reduce omissions of risk evidence.

Through this process, different social parties identify the specific types of problems they attend to and the responses they may make. The algorithm retains structured risk evidence and a traceable interaction record, providing input for the subsequent transition from identifying specific types of problems to multi-party adverse-impact prediction.

Multi-role interaction broadens analytical perspectives, but role judgements may be redundant, conflicting, or unsupported. Moreover, specific types of problems in a statement do not directly determine its adverse impacts: the same type of problem may affect different entities or produce different consequences under different identities and social relations. PRESAGE therefore implements the risk-evidence–adverse-impact dual-layer structure through two-stage integrated judgement. The Evidence Judge first finalizes the supported risk evidence, and the Impact Judge then predicts adverse impacts using the case context and multi-role interaction. Each stage produces a structured output, without a fixed mapping between risk evidence and adverse impacts.

This stage turns the open-ended social responses supplied by multiple roles into structured outputs that PRESAGE-Bench can directly evaluate. It moves from identifying risks in the statement to determining which entities may be affected in which domains, while retaining the influence of specific personal relations and pre-release context.

C Detailed Experimental Protocol

C.1 Dataset, Support, and Valid Outputs

The frozen training and test lists contain 759 and 253 cases, respectively. The archived runs used the original risk-evidence inventory of 163 entries, including the normal-statement entry. The dataset annotations instantiate 67 of these entries; Appendix J lists only those observed entries. This restriction of the displayed definitions does not change the candidate inventory used in the saved runs. Evaluation measures adverse-impact predictions; risk evidence is intermediate information and is not evaluated as a separate prediction task. The normal-statement entry does not force an empty ground-truth set of adverse impacts in the dense annotation version used for these results. The subset historically designated normal contains 77 cases with non-empty ground-truth sets of adverse impacts under this version.

The four-system main comparison uses 250 commonly valid cases; the self-evolution comparison uses 252; and the three-model full-workflow comparison uses 230. Each result table uses its own common case list, including only valid outputs from every system named in that comparison. Failed or missing predictions are unscored, while valid empty predictions are retained. All scored ground-truth sets of adverse impacts were checked against the frozen 253-case test list, whose recorded hash matches the experiment manifest.

For the main comparison, 27 adverse impacts have at least 20 ground-truth-positive cases, nine have 1–19, and none has zero support. We retain all 36 in the result tables and report both positive and negative support. For example, Self$\times$Value has 248 positives and only two negatives on the 250 shared cases: one extra false positive changes its FPR by 0.5. Large positive support therefore does not guarantee a precise FPR estimate.

Every compared system returns a set of adverse impacts $\hat{Y} \subseteq \mathcal{O}$. The set may contain only the 36 legal adverse impacts, may be empty, has no cardinality limit, and is unordered for evaluation. Explanations, risk evidence, and other intermediate outputs may be retained for analysis but are never parsed as final predictions.

C.2 Recorded Runtime and Comparison Scope

The formal Qwen run uses Qwen-plus, temperature 0, nine Role Agents and three rounds for the PRESAGE conditions, and no external retrieval. The direct LLM and Debate-to-Detect retain their respective structures. The Debate-to-Detect comparison uses the independently saved historical Qwen-plus predictions already incorporated in the selected suite's four-system result package. The source audit records agreement of the public-figure name, identity, statement, release time, and platform fields for all 253 cases; exact runtime and model-snapshot equivalence are not established. Consequently, the four-system comparison describes the saved configurations rather than certifying that architecture is the only changed variable.

The dual-layer ablation retains the multi-role discussion but does not pass the explicit complete Evidence Judge output to the Impact Judge. Risk evidence may still appear in the discussion, so the ablation does not remove every source of risk evidence. The selected suite also contains a Gemini-2.5-flash-lite full-workflow run and a GPT-4o-mini run imported from a subsequent execution. Their predictions are scored against the same frozen ground-truth annotations; failures are excluded from the common-case comparison. This is not a comparison of each model's independently optimized best system, and exact Prompt/runtime equivalence remains unverified.

C.3 Counts and Metrics

For each case c and adverse impact $o \in \mathcal{O}$, we test whether the ground-truth set of adverse impacts Y_c and predicted set $\hat{Y}_c$ contain o:

$$TP(o) = \sum_c \mathbf{1}[o \in Y_c \wedge o \in \hat{Y}_c], \qquad FN(o) = \sum_c \mathbf{1}[o \in Y_c \wedge o \notin \hat{Y}_c], \tag{7}$$

$$FP(o) = \sum_c \mathbf{1}[o \notin Y_c \wedge o \in \hat{Y}_c], \qquad TN(o) = \sum_c \mathbf{1}[o \notin Y_c \wedge o \notin \hat{Y}_c]. \tag{8}$$

We then compute

$$\mathrm{Recall}(o) = \frac{TP(o)}{TP(o) + FN(o)}, \qquad \mathrm{FPR}(o) = \frac{FP(o)}{FP(o) + TN(o)}, \tag{9}$$

$$\mathrm{Precision}(o) = \frac{TP(o)}{TP(o) + FP(o)}, \qquad \mathrm{Specificity}(o) = \frac{TN(o)}{TN(o) + FP(o)}, \tag{10}$$

$$\mathrm{PredictionRate}(o) = \frac{TP(o) + FP(o)}{N}. \tag{11}$$

For systems A and B, we compute

$$\Delta R(o) = R_B(o) - R_A(o), \qquad \Delta\,\mathrm{FPR}(o) = \mathrm{FPR}_B(o) - \mathrm{FPR}_A(o).$$

A positive ΔR and negative $\Delta\,\mathrm{FPR}$ is a two-sided improvement; two positive differences indicate a recall–false-alarm trade-off; and a negative ΔR with a positive $\Delta\,\mathrm{FPR}$ is a two-sided degradation. The other sign combinations are reported as recall-only improvement, FPR-only improvement, increased conservatism, or no change. Precision explains prediction purity but does not replace the joint Recall–FPR judgment.

C.4 Case-Level Coverage

Per-adverse-impact evaluation is primary in every comparison. To characterize prediction coverage and self-evolution on multi-impact cases, we additionally report

$$\mathrm{Coverage}(c) = \frac{|Y_c \cap \hat{Y}_c|}{|Y_c|}$$

for cases with a non-empty ground-truth set of adverse impacts, together with complete-coverage rate, mean number of missed adverse impacts, and mean number of falsely predicted adverse impacts. Complete coverage is the proportion of cases for which every ground-truth adverse impact is recovered. These measures are reported with per-impact FPR and prediction rate; producing more predictions alone is not evidence of greater comprehensiveness.

C.5 Pooled Summaries and Interpretation

The descriptive pooled Precision and Recall are

$$\frac{\sum_o TP(o)}{\sum_o [TP(o) + FP(o)]}, \qquad \frac{\sum_o TP(o)}{\sum_o [TP(o) + FN(o)]},$$

and pooled FPR is

$$\frac{\sum_o FP(o)}{\sum_o [FP(o) + TN(o)]}.$$

Pooled FPR weights ground-truth-negative decisions; it is not a ground-truth-positive-support-weighted mean of per-adverse-impact FPR. Undefined zero-denominator ratios are reported as NA in per-adverse-impact tables. The saved result packages provide point estimates and paired prediction changes, not completed bootstrap confidence intervals or Holm-adjusted significance tests. We therefore do not claim statistical significance or test-set false-alarm control from those estimates. The training-side version validation gate is documented separately in Appendix G.

C.6 Complete Overall Metrics

Table 4 provides all nine overall Precision/Recall/F1 metrics and the separate FPR diagnostic. All overall metrics use the same paired valid cases as the corresponding main-text table: 250 for the main comparison, 252 for self-evolution, and 230 for the API–model comparison. Failed or missing predictions remain unscored, and valid empty predictions remain included. The prediction space contains all 36 adverse impacts in every comparison.

For each case c, let $t_c = |Y_c \cap \hat{Y}_c|$, $f_c = |\hat{Y}_c \setminus Y_c|$, and $n_c = |Y_c \setminus \hat{Y}_c|$. Define

$$p(t, f, n) = \frac{t}{t + f}, \qquad r(t, f, n) = \frac{t}{t + n}, \qquad f_1(t, f, n) = \frac{2t}{2t + f + n}.$$

An undefined zero-denominator term in these three functions contributes zero to overall Precision/Recall/F1, including when averaging over cases or adverse impacts. This convention retains all 36 adverse impacts; it does not drop impacts with no predictions. The existing per-adverse-impact diagnostic tables retain NA for undefined ratios.

Samples averaging. Compute each metric within a case and then give each case equal weight:

$$(P_{\text{samples}}, R_{\text{samples}}, F1_{\text{samples}}) = \frac{1}{N} \sum_c (p, r, f_1)(t_c, f_c, n_c).$$

Micro averaging. Pool decisions across cases and adverse impacts before computing the metrics:

$$(P_{\text{micro}}, R_{\text{micro}}, F1_{\text{micro}}) = (p, r, f_1) \left(\sum_c t_c, \sum_c f_c, \sum_c n_c \right).$$

Thus the main tables' F1 is $2\sum_c t_c / (2\sum_c t_c + \sum_c f_c + \sum_c n_c)$, using the same pooled counts as their Precision and Recall. It is calculated from full-precision counts, not from the rounded Precision and Recall printed in a table.

Macro averaging. Compute each metric for each adverse impact, then average equally over all 36:

$$(P_{\text{macro}}, R_{\text{macro}}, F1_{\text{macro}}) = \frac{1}{36} \sum_{o \in \mathcal{O}} (p, r, f_1)(TP(o), FP(o), FN(o)).$$

Samples F1 and Macro F1 are averages of individual F1 values, not the harmonic means of their respective averaged Precision and Recall.

False-positive diagnosis. Alongside these nine metrics, pooled FPR is $\sum_o FP(o) / \sum_o [FP(o) + TN(o)]$. It measures false positives among ground-truth-negative decisions and is reported separately from the three Precision/Recall/F1 averaging schemes. All values below are recomputed from the same saved predictions and frozen ground truth. The development selection score is used only for training-side version selection, not for test-set ranking.

Table 4: Complete Samples, Micro, and Macro Precision/Recall/F1 on each comparison's paired valid cases. FPR is the separate pooled false-positive diagnostic, printed once per system on its Micro row for compactness; dashes do not denote additional missing evaluations. Four decimal places are shown; calculations use unrounded counts.

System	Averaging	Precision	Recall	F1	FPR
Main comparison ($N = 250$)					
Direct LLM	Samples	0.7451	0.3371	0.4424	—
	Micro	0.7393	0.3177	0.4444	0.0486
	Macro	0.4707	0.1711	0.2149	—
Debate-to-Detect	Samples	0.3717	0.0618	0.1017	—
	Micro	0.7111	0.0588	0.1085	0.0104
	Macro	0.3674	0.0391	0.0645	—
PRESAGE (dual-layer ablation)	Samples	0.7435	0.5844	0.6153	—
	Micro	0.7226	0.5501	0.6247	0.0916
	Macro	0.3791	0.2817	0.2946	—
PRESAGE (full)	Samples	0.7127	0.6174	0.6188	—
	Micro	0.6773	0.5858	0.6282	0.1211
	Macro	0.4178	0.3100	0.3133	—
Self-evolution ($N = 252$)					
P_0	Samples	0.6507	0.6069	0.5881	—
	Micro	0.6382	0.5749	0.6049	0.1419
	Macro	0.4048	0.3408	0.3318	—
P_{final}	Samples	0.7116	0.6169	0.6184	—
	Micro	0.6763	0.5854	0.6276	0.1220
	Macro	0.4183	0.3115	0.3140	—
Multi-API–model comparison ($N = 230$)					
Qwen-plus	Samples	0.7084	0.6110	0.6130	—
	Micro	0.6702	0.5783	0.6209	0.1232
	Macro	0.3753	0.3024	0.3021	—
Gemini-2.5-flash-lite	Samples	0.6125	0.4208	0.4553	—
	Micro	0.5963	0.4021	0.4803	0.1179
	Macro	0.3804	0.2588	0.2678	—
GPT-4o-mini	Samples	0.6720	0.2354	0.3351	—
	Micro	0.6615	0.2218	0.3322	0.0492
	Macro	0.3124	0.1135	0.1312	—

D PER-ADVERSE-IMPACT RESULTS

Tables 5 and 6 report the per-adverse-impact results for the main comparison.

Table 5: Main comparison: Recall/FPR on the same 250 valid cases. N_+/N_- denote ground-truth-positive/ground-truth-negative support. D2D denotes Debate-to-Detect; Ablation and Full denote the two PRESAGE conditions. All 36 adverse impacts are retained, including those with fewer than 20 positives.

Impact domain	N_+	N_-	Direct LLM	D2D	Ablation	Full
Self						
Reputation	239	11	0.870/0.545	0.042/0.000	0.891/0.727	0.946/0.909
Career	133	117	0.835/0.658	0.008/0.000	0.451/0.171	0.564/0.256
Safety	20	230	0.050/0.004	0.000/0.009	0.000/0.004	0.000/0.004
Economic	56	194	0.000/0.000	0.000/0.005	0.000/0.000	0.000/0.000
Legal	93	157	0.419/0.140	0.054/0.032	0.570/0.408	0.570/0.357
Value	248	2	0.133/0.000	0.157/0.000	0.915/0.500	0.952/0.500
Affiliated Organization						
Reputation	207	43	0.691/0.326	0.111/0.023	0.729/0.302	0.802/0.395
Career	167	83	0.377/0.193	0.000/0.000	0.228/0.108	0.317/0.169
Safety	25	225	0.080/0.000	0.040/0.009	0.000/0.004	0.000/0.000
Economic	53	197	0.019/0.000	0.019/0.000	0.057/0.005	0.057/0.000
Legal	75	175	0.133/0.069	0.067/0.006	0.240/0.126	0.280/0.177
Value	219	31	0.110/0.000	0.023/0.000	0.804/0.258	0.808/0.258
Partner						
Reputation	18	232	0.111/0.004	0.056/0.004	0.333/0.142	0.333/0.181
Career	26	224	0.077/0.000	0.000/0.000	0.231/0.165	0.269/0.170
Safety	2	248	0.000/0.000	0.000/0.000	0.000/0.000	0.000/0.000
Economic	8	242	0.000/0.000	0.000/0.000	0.000/0.025	0.000/0.017
Legal	9	241	0.000/0.000	0.000/0.008	0.222/0.017	0.111/0.058
Value	22	228	0.000/0.000	0.000/0.000	0.091/0.026	0.045/0.031

Table 6: Main comparison: Recall/FPR on the same 250 valid cases. N_+/N_- denote ground-truth-positive/ground-truth-negative support. D2D denotes Debate-to-Detect; Ablation and Full denote the two PRESAGE conditions. All 36 adverse impacts are retained, including those with fewer than 20 positives.

Impact domain	N_+	N_-	Direct LLM	D2D	Ablation	Full
Targeted Party						
Reputation	173	77	0.376/0.039	0.069/0.000	0.792/0.442	0.798/0.506
Career	67	183	0.119/0.022	0.000/0.000	0.343/0.158	0.388/0.240
Safety	41	209	0.122/0.033	0.049/0.005	0.293/0.134	0.341/0.153
Economic	46	204	0.000/0.000	0.000/0.000	0.022/0.000	0.065/0.010
Legal	47	203	0.277/0.030	0.064/0.005	0.511/0.197	0.511/0.232
Value	204	46	0.093/0.000	0.132/0.022	0.588/0.326	0.632/0.435
Audience						
Reputation	22	228	0.000/0.000	0.000/0.004	0.000/0.000	0.000/0.000
Career	8	242	0.000/0.004	0.000/0.000	0.000/0.012	0.000/0.008
Safety	19	231	0.053/0.017	0.158/0.026	0.000/0.000	0.000/0.009
Economic	9	241	0.000/0.017	0.000/0.004	0.000/0.000	0.111/0.000
Legal	20	230	0.000/0.000	0.150/0.030	0.000/0.000	0.050/0.009
Value	64	186	0.125/0.097	0.156/0.145	0.188/0.269	0.375/0.409

Table 6 (continued)

Impact domain	N_+	N_-	Direct LLM	D2D	Ablation	Full
Field						
Reputation	51	199	0.353/0.352	0.000/0.010	0.118/0.095	0.333/0.457
Career	63	187	0.175/0.053	0.000/0.000	0.063/0.048	0.127/0.123
Safety	12	238	0.000/0.000	0.000/0.000	0.000/0.000	0.083/0.000
Economic	9	241	0.000/0.004	0.000/0.000	0.000/0.000	0.000/0.000
Legal	78	172	0.192/0.035	0.000/0.006	0.487/0.244	0.385/0.273
Value	170	80	0.371/0.275	0.053/0.025	0.976/0.900	0.906/0.750

D.1 Self-Evolution Transitions

Tables 7 and 8 report repairs and regressions for all 36 adverse impacts.

Self-evolution repairs 295 false negatives and removes 415 false positives, while losing 266 true positives and introducing 289 false positives. The net changes are 29 additional true positives and 126 fewer false positives. Pooled Recall increases from 0.575 to 0.585 and FPR decreases from 0.142 to 0.122, but Recall increases for only nine adverse impacts, decreases for twenty, and is unchanged for seven. Mean case coverage increases from 0.607 to 0.617, while complete-coverage rate decreases from 0.083 to 0.075. These results concern the saved single trajectory; they do not establish uniform improvement across adverse impacts or stability across random seeds.

Table 7: Self-evolution comparison on 252 paired cases, including gains and regressions.

Impact domain	N_+	P_0 R/FPR	P_{final} R/FPR	FN→TP	FP→TN	TP→FN	TN→FP
Self							
Reputation	241	0.859/0.818	0.946/0.909	23	0	2	1
Career	134	0.567/0.254	0.560/0.263	18	10	19	11
Safety	20	0.050/0.026	0.000/0.013	0	5	1	2
Economic	57	0.018/0.000	0.000/0.000	0	0	1	0
Legal	94	0.819/0.791	0.574/0.361	4	74	27	6
Value	250	0.928/1.000	0.952/0.500	14	1	8	0
Affiliated Organization							
Reputation	209	0.761/0.372	0.804/0.395	22	0	13	1
Career	169	0.231/0.084	0.320/0.169	32	3	17	10
Safety	26	0.000/0.000	0.000/0.000	0	0	0	0
Economic	54	0.204/0.020	0.056/0.000	0	4	8	0
Legal	76	0.513/0.409	0.276/0.182	6	45	24	5
Value	221	0.633/0.129	0.805/0.258	47	0	9	4
Partner							
Reputation	19	0.368/0.159	0.368/0.180	1	10	1	15
Career	27	0.333/0.156	0.296/0.169	1	9	2	12
Safety	2	0.000/0.000	0.000/0.000	0	0	0	0
Economic	8	0.250/0.045	0.000/0.016	0	9	2	2
Legal	9	0.111/0.132	0.111/0.062	0	23	0	6
Value	23	0.087/0.066	0.043/0.031	0	12	1	4

Table 8: Self-evolution comparison on 252 paired cases, including gains and regressions.

Impact domain	N_+	P_0 R/FPR	P_{final} R/FPR	FN→TP	FP→TN	TP→FN	TN→FP
Targeted Party							
Reputation	175	0.846/0.429	0.800/0.506	4	3	12	9
Career	68	0.426/0.217	0.382/0.239	9	16	12	20
Safety	41	0.415/0.185	0.341/0.156	5	21	8	15
Economic	47	0.085/0.024	0.064/0.010	1	4	2	1
Legal	48	0.583/0.368	0.500/0.230	2	43	6	15
Value	206	0.646/0.370	0.636/0.435	24	4	26	7
Audience							
Reputation	22	0.000/0.000	0.000/0.000	0	0	0	0
Career	8	0.125/0.016	0.000/0.008	0	2	1	0
Safety	20	0.100/0.026	0.000/0.009	0	5	2	1
Economic	9	0.000/0.000	0.111/0.000	1	0	0	0
Legal	21	0.190/0.039	0.048/0.009	1	9	4	2
Value	65	0.600/0.572	0.385/0.412	8	49	22	19
Field							
Reputation	51	0.176/0.199	0.333/0.458	10	10	2	62
Career	63	0.127/0.058	0.127/0.122	5	7	5	19
Safety	12	0.000/0.000	0.083/0.000	1	0	0	0
Economic	9	0.000/0.004	0.000/0.000	0	1	0	0
Legal	78	0.538/0.322	0.385/0.276	7	30	19	22
Value	170	0.676/0.598	0.906/0.744	49	6	10	18

D.2 Multi-API–Model Results

Tables 9 and 10 report the per-adverse-impact results for the three model APIs.

Table 9: Descriptive API–model comparison: Recall/FPR on 230 shared valid cases. Gemini denotes Gemini-2.5-flash-lite. The case intersection differs from the main comparison and self-evolution comparison.

Impact domain	N_+	N_-	Qwen-plus	Gemini	GPT-4o-mini
Self					
Reputation	219	11	0.941/0.909	0.932/0.909	0.918/0.909
Career	117	113	0.547/0.239	0.632/0.584	0.393/0.195
Safety	18	212	0.000/0.014	0.056/0.000	0.000/0.000
Economic	50	180	0.000/0.000	0.000/0.000	0.000/0.000
Legal	86	144	0.570/0.361	0.523/0.326	0.465/0.243
Value	228	2	0.947/0.500	0.386/0.000	0.149/0.000
Affiliated Organization					
Reputation	188	42	0.793/0.405	0.415/0.262	0.255/0.190
Career	152	78	0.309/0.154	0.349/0.167	0.039/0.000
Safety	23	207	0.000/0.000	0.000/0.000	0.000/0.000
Economic	48	182	0.062/0.000	0.000/0.000	0.000/0.000
Legal	69	161	0.246/0.180	0.174/0.106	0.014/0.019
Value	200	30	0.800/0.267	0.140/0.000	0.035/0.000

Table 9 (continued)

Impact domain	N_+	N_-	Qwen-plus	Gemini	GPT-4o-mini
Partner					
Reputation	18	212	0.389/0.193	0.389/0.250	0.000/0.000
Career	26	204	0.308/0.181	0.346/0.167	0.000/0.000
Safety	2	228	0.000/0.000	0.000/0.000	0.000/0.000
Economic	8	222	0.000/0.018	0.000/0.000	0.000/0.000
Legal	9	221	0.111/0.068	0.000/0.014	0.000/0.000
Value	22	208	0.045/0.034	0.000/0.005	0.000/0.000

Table 10: Descriptive API–model comparison: Recall/FPR on 230 shared valid cases. Gemini denotes Gemini-2.5-flash-lite. The case intersection differs from the main comparison and self-evolution comparison.

Impact domain	N_+	N_-	Qwen-plus	Gemini	GPT-4o-mini
Targeted Party					
Reputation	161	69	0.795/0.522	0.795/0.536	0.621/0.536
Career	63	167	0.365/0.246	0.397/0.162	0.079/0.012
Safety	34	196	0.353/0.148	0.176/0.026	0.147/0.026
Economic	42	188	0.048/0.011	0.000/0.011	0.000/0.000
Legal	44	186	0.455/0.226	0.432/0.134	0.023/0.016
Value	188	42	0.644/0.452	0.309/0.190	0.080/0.048
Audience					
Reputation	21	209	0.000/0.000	0.524/0.201	0.095/0.005
Career	6	224	0.000/0.009	0.000/0.009	0.000/0.000
Safety	18	212	0.000/0.009	0.111/0.019	0.000/0.005
Economic	8	222	0.000/0.000	0.000/0.000	0.000/0.000
Legal	20	210	0.000/0.010	0.200/0.010	0.000/0.000
Value	60	170	0.350/0.424	0.483/0.541	0.250/0.388
Field					
Reputation	49	181	0.327/0.464	0.490/0.503	0.490/0.486
Career	57	173	0.105/0.121	0.246/0.110	0.000/0.000
Safety	11	219	0.091/0.000	0.000/0.000	0.000/0.000
Economic	8	222	0.000/0.000	0.000/0.000	0.000/0.000
Legal	73	157	0.384/0.274	0.479/0.280	0.000/0.000
Value	156	74	0.904/0.730	0.333/0.351	0.032/0.014

E Confusion Counts and Additional Metrics

Tables 11, 12, 13, and 14 report confusion counts and derived metrics for the main comparison. Tables 15 and 16 report the self-evolution endpoints. Tables 17, 18, and 19 report the three model APIs.

Table 11: Main comparison: Direct LLM on 250 paired cases. NA denotes an undefined zero-denominator ratio. Pred. rate denotes prediction rate.

Impact domain	TP	FN	FP	TN	Precision	Specificity	Pred. rate
Self							
Reputation	208	31	6	5	0.972	0.455	0.856
Career	111	22	77	40	0.590	0.342	0.752
Safety	1	19	1	229	0.500	0.996	0.008
Economic	0	56	0	194	NA	1.000	0.000
Legal	39	54	22	135	0.639	0.860	0.244
Value	33	215	0	2	1.000	1.000	0.132
Affiliated Organization							
Reputation	143	64	14	29	0.911	0.674	0.628
Career	63	104	16	67	0.797	0.807	0.316
Safety	2	23	0	225	1.000	1.000	0.008
Economic	1	52	0	197	1.000	1.000	0.004
Legal	10	65	12	163	0.455	0.931	0.088
Value	24	195	0	31	1.000	1.000	0.096
Partner							
Reputation	2	16	1	231	0.667	0.996	0.012
Career	2	24	0	224	1.000	1.000	0.008
Safety	0	2	0	248	NA	1.000	0.000
Economic	0	8	0	242	NA	1.000	0.000
Legal	0	9	0	241	NA	1.000	0.000
Value	0	22	0	228	NA	1.000	0.000
Targeted Party							
Reputation	65	108	3	74	0.956	0.961	0.272
Career	8	59	4	179	0.667	0.978	0.048
Safety	5	36	7	202	0.417	0.967	0.048
Economic	0	46	0	204	NA	1.000	0.000
Legal	13	34	6	197	0.684	0.970	0.076
Value	19	185	0	46	1.000	1.000	0.076
Audience							
Reputation	0	22	0	228	NA	1.000	0.000
Career	0	8	1	241	0.000	0.996	0.004
Safety	1	18	4	227	0.200	0.983	0.020
Economic	0	9	4	237	0.000	0.983	0.016
Legal	0	20	0	230	NA	1.000	0.000
Value	8	56	18	168	0.308	0.903	0.104
Field							
Reputation	18	33	70	129	0.205	0.648	0.352
Career	11	52	10	177	0.524	0.947	0.084
Safety	0	12	0	238	NA	1.000	0.000
Economic	0	9	1	240	0.000	0.996	0.004
Legal	15	63	6	166	0.714	0.965	0.084
Value	63	107	22	58	0.741	0.725	0.340

Table 12: Main comparison: Debate-to-Detect on 250 paired cases. NA denotes an undefined zero-denominator ratio. Pred. rate denotes prediction rate.

Impact domain	TP	FN	FP	TN	Precision	Specificity	Pred. rate
Self							
Reputation	10	229	0	11	1.000	1.000	0.040
Career	1	132	0	117	1.000	1.000	0.004
Safety	0	20	2	228	0.000	0.991	0.008
Economic	0	56	1	193	0.000	0.995	0.004
Legal	5	88	5	152	0.500	0.968	0.040
Value	39	209	0	2	1.000	1.000	0.156
Affiliated Organization							
Reputation	23	184	1	42	0.958	0.977	0.096
Career	0	167	0	83	NA	1.000	0.000
Safety	1	24	2	223	0.333	0.991	0.012
Economic	1	52	0	197	1.000	1.000	0.004
Legal	5	70	1	174	0.833	0.994	0.024
Value	5	214	0	31	1.000	1.000	0.020
Partner							
Reputation	1	17	1	231	0.500	0.996	0.008
Career	0	26	0	224	NA	1.000	0.000
Safety	0	2	0	248	NA	1.000	0.000
Economic	0	8	0	242	NA	1.000	0.000
Legal	0	9	2	239	0.000	0.992	0.008
Value	0	22	0	228	NA	1.000	0.000
Targeted Party							
Reputation	12	161	0	77	1.000	1.000	0.048
Career	0	67	0	183	NA	1.000	0.000
Safety	2	39	1	208	0.667	0.995	0.012
Economic	0	46	0	204	NA	1.000	0.000
Legal	3	44	1	202	0.750	0.995	0.016
Value	27	177	1	45	0.964	0.978	0.112
Audience							
Reputation	0	22	1	227	0.000	0.996	0.004
Career	0	8	0	242	NA	1.000	0.000
Safety	3	16	6	225	0.333	0.974	0.036
Economic	0	9	1	240	0.000	0.996	0.004
Legal	3	17	7	223	0.300	0.970	0.040
Value	10	54	27	159	0.270	0.855	0.148
Field							
Reputation	0	51	2	197	0.000	0.990	0.008
Career	0	63	0	187	NA	1.000	0.000
Safety	0	12	0	238	NA	1.000	0.000
Economic	0	9	0	241	NA	1.000	0.000
Legal	0	78	1	171	0.000	0.994	0.004
Value	9	161	2	78	0.818	0.975	0.044

Table 13: Main comparison: PRESAGE (dual-layer ablation) on 250 paired cases. NA denotes an undefined zero-denominator ratio. Pred. rate denotes prediction rate.

Impact domain	TP	FN	FP	TN	Precision	Specificity	Pred. rate
Self							
Reputation	213	26	8	3	0.964	0.273	0.884
Career	60	73	20	97	0.750	0.829	0.320
Safety	0	20	1	229	0.000	0.996	0.004
Economic	0	56	0	194	NA	1.000	0.000
Legal	53	40	64	93	0.453	0.592	0.468
Value	227	21	1	1	0.996	0.500	0.912
Affiliated Organization							
Reputation	151	56	13	30	0.921	0.698	0.656
Career	38	129	9	74	0.809	0.892	0.188
Safety	0	25	1	224	0.000	0.996	0.004
Economic	3	50	1	196	0.750	0.995	0.016
Legal	18	57	22	153	0.450	0.874	0.160
Value	176	43	8	23	0.957	0.742	0.736
Partner							
Reputation	6	12	33	199	0.154	0.858	0.156
Career	6	20	37	187	0.140	0.835	0.172
Safety	0	2	0	248	NA	1.000	0.000
Economic	0	8	6	236	0.000	0.975	0.024
Legal	2	7	4	237	0.333	0.983	0.024
Value	2	20	6	222	0.250	0.974	0.032
Targeted Party							
Reputation	137	36	34	43	0.801	0.558	0.684
Career	23	44	29	154	0.442	0.842	0.208
Safety	12	29	28	181	0.300	0.866	0.160
Economic	1	45	0	204	1.000	1.000	0.004
Legal	24	23	40	163	0.375	0.803	0.256
Value	120	84	15	31	0.889	0.674	0.540
Audience							
Reputation	0	22	0	228	NA	1.000	0.000
Career	0	8	3	239	0.000	0.988	0.012
Safety	0	19	0	231	NA	1.000	0.000
Economic	0	9	0	241	NA	1.000	0.000
Legal	0	20	0	230	NA	1.000	0.000
Value	12	52	50	136	0.194	0.731	0.248
Field							
Reputation	6	45	19	180	0.240	0.905	0.100
Career	4	59	9	178	0.308	0.952	0.052
Safety	0	12	0	238	NA	1.000	0.000
Economic	0	9	0	241	NA	1.000	0.000
Legal	38	40	42	130	0.475	0.756	0.320
Value	166	4	72	8	0.697	0.100	0.952

Table 14: Main comparison: PRESAGE (full) on 250 paired cases. NA denotes an undefined zero-denominator ratio. Pred. rate denotes prediction rate.

Impact domain	TP	FN	FP	TN	Precision	Specificity	Pred. rate
Self							
Reputation	226	13	10	1	0.958	0.091	0.944
Career	75	58	30	87	0.714	0.744	0.420
Safety	0	20	1	229	0.000	0.996	0.004
Economic	0	56	0	194	NA	1.000	0.000
Legal	53	40	56	101	0.486	0.643	0.436
Value	236	12	1	1	0.996	0.500	0.948
Affiliated Organization							
Reputation	166	41	17	26	0.907	0.605	0.732
Career	53	114	14	69	0.791	0.831	0.268
Safety	0	25	0	225	NA	1.000	0.000
Economic	3	50	0	197	1.000	1.000	0.012
Legal	21	54	31	144	0.404	0.823	0.208
Value	177	42	8	23	0.957	0.742	0.740
Partner							
Reputation	6	12	42	190	0.125	0.819	0.192
Career	7	19	38	186	0.156	0.830	0.180
Safety	0	2	0	248	NA	1.000	0.000
Economic	0	8	4	238	0.000	0.983	0.016
Legal	1	8	14	227	0.067	0.942	0.060
Value	1	21	7	221	0.125	0.969	0.032
Targeted Party							
Reputation	138	35	39	38	0.780	0.494	0.708
Career	26	41	44	139	0.371	0.760	0.280
Safety	14	27	32	177	0.304	0.847	0.184
Economic	3	43	2	202	0.600	0.990	0.020
Legal	24	23	47	156	0.338	0.768	0.284
Value	129	75	20	26	0.866	0.565	0.596
Audience							
Reputation	0	22	0	228	NA	1.000	0.000
Career	0	8	2	240	0.000	0.992	0.008
Safety	0	19	2	229	0.000	0.991	0.008
Economic	1	8	0	241	1.000	1.000	0.004
Legal	1	19	2	228	0.333	0.991	0.012
Value	24	40	76	110	0.240	0.591	0.400
Field							
Reputation	17	34	91	108	0.157	0.543	0.432
Career	8	55	23	164	0.258	0.877	0.124
Safety	1	11	0	238	1.000	1.000	0.004
Economic	0	9	0	241	NA	1.000	0.000
Legal	30	48	47	125	0.390	0.727	0.308
Value	154	16	60	20	0.720	0.250	0.856

Table 15: Self-evolution: P_0 on 252 paired cases. NA denotes an undefined zero-denominator ratio. Pred. rate denotes prediction rate.

Impact domain	TP	FN	FP	TN	Precision	Specificity	Pred. rate
Self							
Reputation	207	34	9	2	0.958	0.182	0.857
Career	76	58	30	88	0.717	0.746	0.421
Safety	1	19	6	226	0.143	0.974	0.028
Economic	1	56	0	195	1.000	1.000	0.004
Legal	77	17	125	33	0.381	0.209	0.802
Value	232	18	2	0	0.991	0.000	0.929
Affiliated Organization							
Reputation	159	50	16	27	0.909	0.628	0.694
Career	39	130	7	76	0.848	0.916	0.183
Safety	0	26	0	226	NA	1.000	0.000
Economic	11	43	4	194	0.733	0.980	0.060
Legal	39	37	72	104	0.351	0.591	0.440
Value	140	81	4	27	0.972	0.871	0.571
Partner							
Reputation	7	12	37	196	0.159	0.841	0.175
Career	9	18	35	190	0.205	0.844	0.175
Safety	0	2	0	250	NA	1.000	0.000
Economic	2	6	11	233	0.154	0.955	0.052
Legal	1	8	32	211	0.030	0.868	0.131
Value	2	21	15	214	0.118	0.934	0.067
Targeted Party							
Reputation	148	27	33	44	0.818	0.571	0.718
Career	29	39	40	144	0.420	0.783	0.274
Safety	17	24	39	172	0.304	0.815	0.222
Economic	4	43	5	200	0.444	0.976	0.036
Legal	28	20	75	129	0.272	0.632	0.409
Value	133	73	17	29	0.887	0.630	0.595
Audience							
Reputation	0	22	0	230	NA	1.000	0.000
Career	1	7	4	240	0.200	0.984	0.020
Safety	2	18	6	226	0.250	0.974	0.032
Economic	0	9	0	243	NA	1.000	0.000
Legal	4	17	9	222	0.308	0.961	0.052
Value	39	26	107	80	0.267	0.428	0.579
Field							
Reputation	9	42	40	161	0.184	0.801	0.194
Career	8	55	11	178	0.421	0.942	0.075
Safety	0	12	0	240	NA	1.000	0.000
Economic	0	9	1	242	0.000	0.996	0.004
Legal	42	36	56	118	0.429	0.678	0.389
Value	115	55	49	33	0.701	0.402	0.651

Table 16: Self-evolution: P_{final} on 252 paired cases. NA denotes an undefined zero-denominator ratio. Pred. rate denotes prediction rate.

Impact domain	TP	FN	FP	TN	Precision	Specificity	Pred. rate
Self							
Reputation	228	13	10	1	0.958	0.091	0.944
Career	75	59	31	87	0.708	0.737	0.421
Safety	0	20	3	229	0.000	0.987	0.012
Economic	0	57	0	195	NA	1.000	0.000
Legal	54	40	57	101	0.486	0.639	0.440
Value	238	12	1	1	0.996	0.500	0.948
Affiliated Organization							
Reputation	168	41	17	26	0.908	0.605	0.734
Career	54	115	14	69	0.794	0.831	0.270
Safety	0	26	0	226	NA	1.000	0.000
Economic	3	51	0	198	1.000	1.000	0.012
Legal	21	55	32	144	0.396	0.818	0.210
Value	178	43	8	23	0.957	0.742	0.738
Partner							
Reputation	7	12	42	191	0.143	0.820	0.194
Career	8	19	38	187	0.174	0.831	0.183
Safety	0	2	0	250	NA	1.000	0.000
Economic	0	8	4	240	0.000	0.984	0.016
Legal	1	8	15	228	0.062	0.938	0.063
Value	1	22	7	222	0.125	0.969	0.032
Targeted Party							
Reputation	140	35	39	38	0.782	0.494	0.710
Career	26	42	44	140	0.371	0.761	0.278
Safety	14	27	33	178	0.298	0.844	0.187
Economic	3	44	2	203	0.600	0.990	0.020
Legal	24	24	47	157	0.338	0.770	0.282
Value	131	75	20	26	0.868	0.565	0.599
Audience							
Reputation	0	22	0	230	NA	1.000	0.000
Career	0	8	2	242	0.000	0.992	0.008
Safety	0	20	2	230	0.000	0.991	0.008
Economic	1	8	0	243	1.000	1.000	0.004
Legal	1	20	2	229	0.333	0.991	0.012
Value	25	40	77	110	0.245	0.588	0.405
Field							
Reputation	17	34	92	109	0.156	0.542	0.433
Career	8	55	23	166	0.258	0.878	0.123
Safety	1	11	0	240	1.000	1.000	0.004
Economic	0	9	0	243	NA	1.000	0.000
Legal	30	48	48	126	0.385	0.724	0.310
Value	154	16	61	21	0.716	0.256	0.853

Table 17: Multi-API–model comparison: Qwen-plus on 230 paired cases. NA denotes an undefined zero-denominator ratio. Pred. rate denotes prediction rate.

Impact domain	TP	FN	FP	TN	Precision	Specificity	Pred. rate
Self							
Reputation	206	13	10	1	0.954	0.091	0.939
Career	64	53	27	86	0.703	0.761	0.396
Safety	0	18	3	209	0.000	0.986	0.013
Economic	0	50	0	180	NA	1.000	0.000
Legal	49	37	52	92	0.485	0.639	0.439
Value	216	12	1	1	0.995	0.500	0.943
Affiliated Organization							
Reputation	149	39	17	25	0.898	0.595	0.722
Career	47	105	12	66	0.797	0.846	0.257
Safety	0	23	0	207	NA	1.000	0.000
Economic	3	45	0	182	1.000	1.000	0.013
Legal	17	52	29	132	0.370	0.820	0.200
Value	160	40	8	22	0.952	0.733	0.730
Partner							
Reputation	7	11	41	171	0.146	0.807	0.209
Career	8	18	37	167	0.178	0.819	0.196
Safety	0	2	0	228	NA	1.000	0.000
Economic	0	8	4	218	0.000	0.982	0.017
Legal	1	8	15	206	0.062	0.932	0.070
Value	1	21	7	201	0.125	0.966	0.035
Targeted Party							
Reputation	128	33	36	33	0.780	0.478	0.713
Career	23	40	41	126	0.359	0.754	0.278
Safety	12	22	29	167	0.293	0.852	0.178
Economic	2	40	2	186	0.500	0.989	0.017
Legal	20	24	42	144	0.323	0.774	0.270
Value	121	67	19	23	0.864	0.548	0.609
Audience							
Reputation	0	21	0	209	NA	1.000	0.000
Career	0	6	2	222	0.000	0.991	0.009
Safety	0	18	2	210	0.000	0.991	0.009
Economic	0	8	0	222	NA	1.000	0.000
Legal	0	20	2	208	0.000	0.990	0.009
Value	21	39	72	98	0.226	0.576	0.404
Field							
Reputation	16	33	84	97	0.160	0.536	0.435
Career	6	51	21	152	0.222	0.879	0.117
Safety	1	10	0	219	1.000	1.000	0.004
Economic	0	8	0	222	NA	1.000	0.000
Legal	28	45	43	114	0.394	0.726	0.309
Value	141	15	54	20	0.723	0.270	0.848

Table 18: Multi-API–model comparison: Gemini-2.5-flash-lite on 230 paired cases. NA denotes an undefined zero-denominator ratio. Pred. rate denotes prediction rate.

Impact domain	TP	FN	FP	TN	Precision	Specificity	Pred. rate
Self							
Reputation	204	15	10	1	0.953	0.091	0.930
Career	74	43	66	47	0.529	0.416	0.609
Safety	1	17	0	212	1.000	1.000	0.004
Economic	0	50	0	180	NA	1.000	0.000
Legal	45	41	47	97	0.489	0.674	0.400
Value	88	140	0	2	1.000	1.000	0.383
Affiliated Organization							
Reputation	78	110	11	31	0.876	0.738	0.387
Career	53	99	13	65	0.803	0.833	0.287
Safety	0	23	0	207	NA	1.000	0.000
Economic	0	48	0	182	NA	1.000	0.000
Legal	12	57	17	144	0.414	0.894	0.126
Value	28	172	0	30	1.000	1.000	0.122
Partner							
Reputation	7	11	53	159	0.117	0.750	0.261
Career	9	17	34	170	0.209	0.833	0.187
Safety	0	2	0	228	NA	1.000	0.000
Economic	0	8	0	222	NA	1.000	0.000
Legal	0	9	3	218	0.000	0.986	0.013
Value	0	22	1	207	0.000	0.995	0.004
Targeted Party							
Reputation	128	33	37	32	0.776	0.464	0.717
Career	25	38	27	140	0.481	0.838	0.226
Safety	6	28	5	191	0.545	0.974	0.048
Economic	0	42	2	186	0.000	0.989	0.009
Legal	19	25	25	161	0.432	0.866	0.191
Value	58	130	8	34	0.879	0.810	0.287
Audience							
Reputation	11	10	42	167	0.208	0.799	0.230
Career	0	6	2	222	0.000	0.991	0.009
Safety	2	16	4	208	0.333	0.981	0.026
Economic	0	8	0	222	NA	1.000	0.000
Legal	4	16	2	208	0.667	0.990	0.026
Value	29	31	92	78	0.240	0.459	0.526
Field							
Reputation	24	25	91	90	0.209	0.497	0.500
Career	14	43	19	154	0.424	0.890	0.143
Safety	0	11	0	219	NA	1.000	0.000
Economic	0	8	0	222	NA	1.000	0.000
Legal	35	38	44	113	0.443	0.720	0.343
Value	52	104	26	48	0.667	0.649	0.339

Table 19: Multi-API–model comparison: GPT-4o-mini on 230 paired cases. NA denotes an undefined zero-denominator ratio. Pred. rate denotes prediction rate.

Impact domain	TP	FN	FP	TN	Precision	Specificity	Pred. rate
Self							
Reputation	201	18	10	1	0.953	0.091	0.917
Career	46	71	22	91	0.676	0.805	0.296
Safety	0	18	0	212	NA	1.000	0.000
Economic	0	50	0	180	NA	1.000	0.000
Legal	40	46	35	109	0.533	0.757	0.326
Value	34	194	0	2	1.000	1.000	0.148
Affiliated Organization							
Reputation	48	140	8	34	0.857	0.810	0.243
Career	6	146	0	78	1.000	1.000	0.026
Safety	0	23	0	207	NA	1.000	0.000
Economic	0	48	0	182	NA	1.000	0.000
Legal	1	68	3	158	0.250	0.981	0.017
Value	7	193	0	30	1.000	1.000	0.030
Partner							
Reputation	0	18	0	212	NA	1.000	0.000
Career	0	26	0	204	NA	1.000	0.000
Safety	0	2	0	228	NA	1.000	0.000
Economic	0	8	0	222	NA	1.000	0.000
Legal	0	9	0	221	NA	1.000	0.000
Value	0	22	0	208	NA	1.000	0.000
Targeted Party							
Reputation	100	61	37	32	0.730	0.464	0.596
Career	5	58	2	165	0.714	0.988	0.030
Safety	5	29	5	191	0.500	0.974	0.043
Economic	0	42	0	188	NA	1.000	0.000
Legal	1	43	3	183	0.250	0.984	0.017
Value	15	173	2	40	0.882	0.952	0.074
Audience							
Reputation	2	19	1	208	0.667	0.995	0.013
Career	0	6	0	224	NA	1.000	0.000
Safety	0	18	1	211	0.000	0.995	0.004
Economic	0	8	0	222	NA	1.000	0.000
Legal	0	20	0	210	NA	1.000	0.000
Value	15	45	66	104	0.185	0.612	0.352
Field							
Reputation	24	25	88	93	0.214	0.514	0.487
Career	0	57	0	173	NA	1.000	0.000
Safety	0	11	0	219	NA	1.000	0.000
Economic	0	8	0	222	NA	1.000	0.000
Legal	0	73	0	157	NA	1.000	0.000
Value	5	151	1	73	0.833	0.986	0.026

F CASE-LEVEL COVERAGE

Table 20 reports case-level coverage, omissions, and false predictions for each comparison.

Table 20: Case-level coverage on each comparison's shared valid cases. Coverage is the mean fraction of ground-truth adverse impacts recovered; Complete is the fraction recovering every ground-truth adverse impact. Missed and False are mean numbers of missed and falsely predicted adverse impacts. All systems have zero exact set matches on these paired populations. Complete coverage permits extra false predictions and therefore differs from exact set match.

Comparison	System	N	Coverage	Complete	Missed	False
Main	Direct LLM	250	0.337	0.004	7.432	1.220
Main	Debate-to-Detect	250	0.062	0.000	10.252	0.260
Main	PRESAGE (dual-layer ablation)	250	0.584	0.052	4.900	2.300
Main	PRESAGE (full)	250	0.617	0.076	4.512	3.040
Evolution	P_0	252	0.607	0.083	4.643	3.560
Evolution	P_{final}	252	0.617	0.075	4.528	3.060
API–model	Qwen-plus	230	0.611	0.074	4.587	3.096
API–model	Gemini-2.5-flash-lite	230	0.421	0.035	6.504	2.961
API–model	GPT-4o-mini	230	0.235	0.000	8.465	1.235

G SELF-EVOLUTION EXECUTION DETAILS

Table 21 summarizes the version chain and candidate–parent Recall/FPR changes; the complete acceptance criteria appear in Appendix G.6.

The fixed development set was selected before candidate feedback, with event groups kept disjoint from the training update set. The run manifest defines an event group by normalized public-figure name and full statement-date string. The selected sizes are 708 update cases and 51 development cases; the latter is approximately 5% of the complete 1,012-case corpus, not 5% of the training data. Each update case is consumed once in the 15 mixed-case batches; no single adverse impact is designated as the target of an entire batch. The saved configuration uses reflection minibatches of at most eight case–adverse-impact feedback records, a minimum of two independent supporting cases, at most four atomic edits per candidate, an impact-cost limit of eight, and one candidate per round.

The version validation gate requires at least one net repaired case–adverse-impact decision and strict improvement in the saved selection score. The score weights are 0.50 for adverse-impact micro-F1, 0.15 for adverse-impact micro-Recall, 0.15 for one minus adverse-impact micro-FPR, and 0.20 for risk-evidence micro-F1. Its additional bounds are a maximum 0.02 drop in adverse-impact Recall, a maximum 0.005 increase in adverse-impact FPR, and a maximum 0.02 drop in risk-evidence Recall. The configuration also contains bounds of 0.15 on the false-alarm-rate increase and 0.25 on the increase in mean false adverse impacts for the training-side subset designated normal. These are recorded selection rules; their terminology does not establish that the dense-annotation test subset has empty ground-truth sets of adverse impacts. The selection score is a training-side mechanism, not the test-set ranking criterion.

Table 21: Recorded candidate–parent changes on the 51-case fixed development set. This is a summary; the complete decision ledger is in Appendix G.6. These are development results, not intermediate-version test results.

Round	Parent	Next	ΔRecall	ΔFPR	Accepted
1	P0	P0	+0.056	+0.056	No
2	P0	P0	+0.021	+0.011	No
3	P0	P01	+0.017	-0.002	Yes
4	P01	P02	-0.008	-0.016	Yes
5	P02	P02	+0.042	+0.027	No
6	P02	P02	+0.002	+0.016	No
7	P02	P02	+0.012	+0.021	No
8	P02	P02	-0.029	+0.001	No
9	P02	P03	-0.008	-0.011	Yes

Table 21 (continued)

Round	Parent	Next	ΔRecall	ΔFPR	Accepted
10	P03	P04	+0.010	-0.008	Yes
11	P04	P04	-0.029	+0.000	No
12	P04	P04	+0.012	+0.004	No
13	P04	P04	-0.015	+0.018	No
14	P04	P04	-0.025	-0.002	No
15	P04	P04	-0.027	-0.003	No

G.1 Mixed-Case Prompt Self-Evolution Algorithm

Algorithms 2–4 specify the mixed-case update loop, Failure/Protection reflection, and atomic Prompt editing. Failure feedback records are grouped only by their primary error type, FN or FP. Risk-evidence differences remain diagnostic information; they do not define groups. Each update case undergoes one complete case prediction, whose trace is reused by all subsequent case–adverse-impact feedback records.

Algorithm 2 Mixed-case Prompt self-evolution with global version validation

Require: D_{train} (759 cases); $N_{\text{all}} = 1012$; $\rho = 0.05$; seed $s = 2023211448$; initial Prompt package P_0; impact space $\mathcal{O}$, $|\mathcal{O}| = 36$; Prompt module registry $\mathcal{G}$
Ensure: Frozen Prompt package P_{final} and audit records $\mathcal{R}$
$n_{\text{dev}} \leftarrow \text{round}(\rho N_{\text{all}}) = 51$
Split D_{train} into D_{dev} (51) and D_{update} (708) using seed s, keeping normalized-name/full-release-date event groups disjoint
Shuffle D_{update} deterministically with s; partition into $\mathcal{B}$: fourteen batches of 48 cases and one of 36
$P \leftarrow P_0$; $\mathcal{R} \leftarrow \emptyset$; initialize MetaSkill M and Rejected Buffer Q
Run P_0 on D_{dev} once and cache its complete results
for each mixed-case epoch $B \in \mathcal{B}$ **do**
  $P_{\text{base}} \leftarrow P$; lock the parent Prompt snapshot
  $T \leftarrow \textsc{PredictCases}(P_{\text{base}}, B)$: Role Agents $\rightarrow$ Evidence Judge $\rightarrow$ Impact Judge
  $K \leftarrow \emptyset$
  for each case c with a completed prediction in B **do**
    Read ground-truth risk evidence L_c and adverse impacts Y_c only now
    $K \leftarrow K \cup \textsc{BuildFeedbackRecords}(c, T[c], L_c, \hat{L}_c, Y_c, \hat{Y}_c, \mathcal{O})$ (36 feedback records)
  end for
  $(H_F, H_P) \leftarrow \textsc{Reflect}(K, T, \mathcal{G}, M, Q)$ (Algorithm 3)
  $(P', A') \leftarrow \textsc{ComposeEdits}(P_{\text{base}}, H_F, H_P, \mathcal{G}, M, Q)$ (Algorithm 4)
  if no legal candidate exists **then**
    Record the reason; update attribution experience and mark protection experience unverified
    continue to the next epoch
  end if
  $V_{\text{base}} \leftarrow \textsc{LoadCached}(P_{\text{base}}, D_{\text{dev}})$
  $V' \leftarrow \textsc{RunAndCache}(P', D_{\text{dev}})$
  $(a, g) \leftarrow \textsc{ValidateGlobal}(V_{\text{base}}, V', \mathcal{O})$ (Section G.4)
  $r \leftarrow \textsc{ReplayProtection}(H_P, V_{\text{base}}, V')$
  Append epoch, parent/candidate hashes, reflection batches, accepted/rejected attributions, module-routing reasons, atomic edits, A', g, r, decision a, and next-version identifier to $\mathcal{R}$
  Update M using validated attributions, protection status, allowed aggregate gate feedback, and edit IDs (Section G.5)
  if $a = \text{accepted}$ **then**
    $P \leftarrow P'$; retain V' as the next parent cache
  else
    $P \leftarrow P_{\text{base}}$
    Add attributions, rejected edit combination A', intended repairs, and failed gate boundaries to Q
  end if
end for
$P_{\text{final}} \leftarrow P$; freeze P_{final}
return $P_{\text{final}}, \mathcal{R}$

G.2 Case–Adverse-Impact Feedback Records and Dual-Path Reflection

For every completed case c, BuildFeedbackRecords creates one feedback record for each $o \in \mathcal{O}$. Membership in Y_c and $\hat{Y}_c$ deterministically establishes TP, FN, FP, or TN. A feedback record stores its case key, adverse impact, primary error or correctness status, risk-evidence differences, relevant Role Agent and Judge outputs, routing hints, Prompt version, and trace references. All 36 feedback records reuse the same case prediction. Feedback records are constructed locally from the completed case-prediction record and ground-truth annotations; they do not trigger additional case predictions. There are no per-target-adverse-impact quotas for FN, FP, TP, or TN. True negatives are selected for Protection when they help characterize correct judgement boundaries.

Algorithm 3 Failure/Protection reflection with error-type-only Failure grouping

Require: Feedback records K; traces T; registry $\mathcal{G}$; MetaSkill M; Rejected Buffer Q; at most 8 feedback records per reflection minibatch; at least 2 independent supporting cases
Ensure: Fact-checked Failure attributions H_F and Protection summaries H_P

$K_F \leftarrow \{k \in K : \text{state}(k) \in \{\text{FN}, \text{FP}\}\}$
$K_P \leftarrow \{k \in K : \text{state}(k) = \text{TP}\} \cup \textsc{BoundaryValuableTN}(K)$
$G_F \leftarrow \textsc{DeterministicGroup}(K_F, \text{key} = \text{primary_error_form})$
$G_P \leftarrow \textsc{GroupProtectionBoundaries}(K_P)$
$\mathcal{M}_F \leftarrow \textsc{InterleavedPack}(G_F, 8, 2)$; $\mathcal{M}_P \leftarrow \textsc{InterleavedPack}(G_P, 8, 2)$
Retain feedback records with insufficient independent support in the manifest; do not request attribution for them
$H_F \leftarrow \emptyset$; $H_P \leftarrow \emptyset$
for each Failure reflection minibatch $m \in \mathcal{M}_F$ **do**
  Retrieve rejected attribution cards from Q using only the primary error type of m
  $h \leftarrow$ATTRIBUTEFAILURE using m, T, module responsibilities, routing hints, compact M, and the retrieved cards
  Permit semantic subdivision or refusal to generalize; forbid changes to ground truth, predictions, or program-determined error types
  if h identifies a shared error mechanism supported by at least 2 independent cases **then**
    if $\textsc{FactAnchorCheck}(h, m, T, \{L_c, Y_c\})$ passes **then**
      $H_F \leftarrow H_F \cup \{h\}$
    else
      Record the failed factual check
    end if
  end if
end for
for each Protection reflection minibatch $p \in \mathcal{M}_P$ **do**
  $q \leftarrow \textsc{SummarizeProtection}(p, T, \textsc{Compact}(M))$
  Identify why the current judgement is correct, limits on generalization, adverse impacts and modules to protect, and possible TP→FN or TN→FP regressions
  if q has at least 2 independent supporting cases **then**
    $H_P \leftarrow H_P \cup \{q\}$
  end if
end for
return H_F, H_P

Each Failure feedback record belongs to one primary group, determined only by FN or FP. Adverse impact, scene, risk-evidence differences, and candidate modules do not add grouping dimensions. Packing interleaves cases within a group. Multiple feedback records from one case do not constitute independent support. A shared error type alone is insufficient for a transferable attribution: the attribution must provide supporting cases, case-specific anchors, counterevidence, an explanation of the error, a possible first-error module, and its expected global effect. The first-error module remains unresolved until attribution; routing hints are advisory. Protection summarizes correct decision boundaries rather than generating a second set of Failure attributions.

G.3 Module Localization, Atomic Edits, and Candidate Composition

Attribution sees case–adverse-impact feedback records, a summary of the Role Agent–Evidence Judge–Impact Judge relation, module responsibilities, routing hints, compact MetaSkill, and necessary trace references. It does not receive every module's full Prompt. Only after fact checking does the editor receive the selected module's full Prompt, its current rule-ID registry, the validated attribution, relevant Protection constraints, relevant rejected combinations, and the remaining edit budget.

Algorithm 4 Two-stage module localization and atomic Prompt editing

Require: Locked parent P_{base}; validated attributions H_F; Protection summaries H_P; registry $\mathcal{G}$; MetaSkill M; Rejected Buffer Q
Ensure: At most one candidate Prompt package P' and selected edits A'

$A \leftarrow \emptyset$
for each fact-checked attribution $h \in H_F$ **do**
  $U \leftarrow \textsc{SelectModules}(h, \mathcal{G}.\text{responsibilities}, h.\text{routing_hints}, \textsc{Compact}(M))$
  for each selected module $u \in U$ **do**
    Expose only $P_{\text{base}}[u]$, registry $\mathcal{G}[u]$, h, relevant H_P, relevant rejected combinations from Q, and the remaining budget
    $E \leftarrow \textsc{ProposeAtomicEdits}(\text{exposed context})$
    for each proposed edit $e \in E$ **do**
      Check that e is `insert_after`, `replace`, or `delete`
      Check that its unique, legal `rule_id` belongs to the selected module in the current locked snapshot
      Reject unknown, stale, protected, or out-of-module IDs, free-text anchors, and whole-file rewrites
      Reject references to specific people, case quotations, ground truth, or post-release outcomes
      Require applicability conditions, exclusions, and pre-release support for inserted or replacement rules
      Check Protection boundaries and registry/format validity
      if all checks pass **then**
        $A \leftarrow A \cup \{e\}$
      else
        Record the invalid edit and reason
      end if
    end for
  end for
end for
Merge duplicate edits; filter conflicts, invalid IDs, and unauthorized changes in A
Select $A' \subseteq A$ satisfying Protection constraints, $|A'| \leq 4$, and total impact cost ≤ 8
Apply A' to the same locked parent P_{base} to form one P', or return no legal candidate
return P', A'

The attribution model proposes a checkable explanation and module location; the editing model proposes a candidate patch. Neither directly changes the active Prompt version. The global version validation gate accepts or rejects the entire combination. Acceptance does not establish that every edit is independently beneficial, and rejection does not establish that every attribution is wrong. Attribution, intended repair, edit combination, and gate outcome are therefore recorded separately.

G.4 Fixed Global Version Validation Gate

The fixed development size is $\text{round}(1012 \times 0.05) = 51$, sampled from the 759 pre-2025 cases with seed 2023211448. It is not 5% of the training set. Grouping by normalized public-figure name and full release date keeps event groups disjoint from the 708 update cases. The development set is frozen before the first candidate is generated and is used only for version selection. P_0 is evaluated on it once; an accepted candidate's cached result becomes the next parent's result.

The gate evaluates all 51×36 case–adverse-impact decisions together. It has no designated target adverse impact, no separate non-target group, and no per-adverse-impact veto. Comparing candidate and parent predictions against the same ground truth gives four transitions: FN→TP and FP→TN repair errors; TP→FN and TN→FP introduce regressions. Let $N_{a\to b}$ count transitions from parent state a to candidate state b. Global net repair is

$$\text{NetRepair} = (N_{\text{FN}\to\text{TP}} + N_{\text{FP}\to\text{TN}}) - (N_{\text{TP}\to\text{FN}} + N_{\text{TN}\to\text{FP}}). \tag{12}$$

The selection score is

$$S = 0.50\,F1_{\text{impact,micro}} + 0.15\,R_{\text{impact,micro}} + 0.15(1 - \text{FPR}_{\text{impact,micro}}) + 0.20\,F1_{\text{evidence,micro}}. \tag{13}$$

Here impact refers to adverse-impact predictions, and evidence refers to risk evidence. These risk-evidence metrics are training-side version-selection signals, not a separately reported test-set prediction task. With Δ denoting candidate minus parent, all of the following must hold:

$$S_{\text{candidate}} > S_{\text{parent}}, \qquad \text{NetRepair} \geq 1, \tag{14}$$

$$\Delta R_{\text{impact}} \geq -0.02, \qquad \Delta\text{FPR}_{\text{impact}} \leq 0.005, \tag{15}$$

$$\Delta R_{\text{evidence}} \geq -0.02, \qquad \Delta\text{NormalFAR} \leq 0.15, \tag{16}$$

$$\Delta\text{NormalMeanFP} \leq 0.25. \tag{17}$$

NormalFAR and NormalMeanFP retain the recorded configuration's names for the false-alarm rate and mean false adverse impacts on its training-side subset designated normal. They do not establish that the separately archived test subset has empty ground truth. Per-adverse-impact Recall/FPR, error transitions, and bootstrap results support auditing; they do not impose a per-adverse-impact veto.

G.5 MetaSkill, Rejected Buffer, and Information Boundaries

MetaSkill stores optimizer experience and does not participate in case prediction by Role Agents, the Evidence Judge, or the Impact Judge. Attribution cards are indexed only by primary error type. Their contents retain validated attributions and applicability conditions, supporting cases and counterevidence, historical target modules, attempted edit combinations and their purposes, aggregate gate outcomes, failed boundaries, and usage limits. Adverse impact, scene, and risk-evidence differences do not create extra indexing levels.

The attribution model first receives failed attribution or editing directions for the same error type. It must propose a new attribution or a different edit combination for a reused attribution. Detailed historical edit combinations are exposed only when a new attribution resembles a historical one, limiting the context supplied from MetaSkill.

Protection experience is recorded separately. After replay, a TP→FN or TN→FP violation is marked `VIOLATED`; an accepted combination preserving the boundary is marked `PRESERVED_IN_ACCEPTED_COMBINATION`; an unaccepted candidate without an observed violation, or a round without a legal candidate, is marked `UNVERIFIED`. These are observations under a combined candidate, not evidence of an individual Protection rule's causal effect.

Development statements, per-case ground truth or predictions, and per-adverse-impact error lists must not enter subsequent attribution, editing, or MetaSkill. Gate feedback exposed to MetaSkill is restricted to the aggregate acceptance decision, failed-boundary names, selection-score difference, global net repair, and edit IDs; detailed validation and replay records remain audit artifacts. Test data are used only after all 15 epochs and freezing P_{final}, never for candidate generation, acceptance, rollback, MetaSkill updates, or stopping decisions.

G.6 Complete Version Validation Ledger

The compact epoch table is a summary. Table 22 reports selection scores and net repair, Table 23 reports metric changes, and Table 24 records the acceptance conditions for all 15 rounds on the fixed 51-case development set. Scores, differences, net repair, and every boundary outcome were recomputed from each saved gate decision and checked against the recorded acceptance and version chain. Six decimal places are printed here; full-precision parent/candidate values and named boundary results are included in the accompanying `paper_results/gate_ledger.csv` and `paper_results/gate_ledger.json`.

Table 22: Selection scores and global net repair. Next retains the parent after rejection; the candidate identifier is E01-candidate-A through E15-candidate-A.

Round	Parent	Next	S_{parent}	S_{cand}	ΔS	Net repair
1	P0	P0	0.576610	0.574865	-0.001745	-45
2	P0	P0	0.576610	0.581259	+0.004649	-4
3	P0	P01	0.576610	0.586821	+0.010211	12
4	P01	P02	0.586821	0.591855	+0.005034	17
5	P02	P02	0.591855	0.598007	+0.006152	-14
6	P02	P02	0.591855	0.583849	-0.008006	-20
7	P02	P02	0.591855	0.585841	-0.006014	-22
8	P02	P02	0.591855	0.576217	-0.015638	-16
9	P02	P03	0.591855	0.594414	+0.002559	11
10	P03	P04	0.594414	0.603965	+0.009551	15
11	P04	P04	0.603965	0.588619	-0.015346	-15
12	P04	P04	0.603965	0.607738	+0.003772	1
13	P04	P04	0.603965	0.585232	-0.018733	-32
14	P04	P04	0.603965	0.591583	-0.012383	-11
15	P04	P04	0.603965	0.591448	-0.012518	-10

Table 23: Candidate minus parent. I denotes adverse-impact Micro metrics; E denotes risk-evidence Micro metrics. A and M denote the recorded normal-subset false-alarm rate and mean false adverse impacts, respectively. These are development-gate diagnostics.

Round	$\Delta F1_I$	ΔR_I	ΔFPR_I	$\Delta F1_E$	ΔR_E	ΔA	ΔM
1	-0.003397	+0.055877	+0.056188	+0.000000	+0.000000	+0.000000	+1.333333
2	+0.006356	+0.021195	+0.011390	+0.000000	+0.000000	+0.041667	+0.375000
3	+0.014537	+0.017341	-0.002278	+0.000000	+0.000000	+0.041667	-0.166667
4	+0.007596	-0.007707	-0.015945	+0.000000	+0.000000	-0.041667	-0.250000
5	+0.007788	+0.042389	+0.027335	+0.000000	+0.000000	+0.041667	+1.125000
6	-0.011806	+0.001927	+0.015945	+0.000000	+0.000000	+0.041667	+0.541667
7	-0.009118	+0.011561	+0.021260	+0.000000	+0.000000	+0.000000	+0.833333
8	-0.022379	-0.028902	+0.000759	+0.000000	+0.000000	+0.041667	-0.166667
9	+0.004014	-0.007707	-0.011390	+0.000000	+0.000000	+0.000000	-0.458333
10	+0.013934	+0.009634	-0.007593	+0.000000	+0.000000	+0.000000	-0.083333
11	-0.022022	-0.028902	+0.000000	+0.000000	+0.000000	+0.041667	-0.333333
12	+0.005215	+0.011561	+0.003797	+0.000000	+0.000000	+0.041667	+0.291667
13	-0.027375	-0.015414	+0.018223	+0.000000	+0.000000	+0.041667	+0.625000
14	-0.017707	-0.025048	-0.001519	+0.000000	+0.000000	+0.041667	+0.208333
15	-0.017854	-0.026975	-0.003037	+0.000000	+0.000000	+0.000000	-0.666667

Boundary names and acceptance conditions. Each round must pass all seven conditions:

S: Selection-score improvement. $S_{\text{candidate}} > S_{\text{parent}}$.

N: Global net repair. $\text{NetRepair} \geq 1$.

R: Adverse-impact Recall. $\Delta R_I \geq -0.02$.

F: Adverse-impact FPR. $\Delta\text{FPR}_I \leq 0.005$.

E: Risk-evidence Recall. $\Delta R_E \geq -0.02$.

A: Normal-subset false-alarm rate. $\Delta A \leq 0.15$.

M: Normal-subset mean false adverse impacts. $\Delta M \leq 0.25$.

Table 24: Every passed or triggered gate boundary. Round 12 passes the score and net-repair conditions but fails M, illustrating why the summary's Recall/FPR alone cannot reproduce acceptance.

Round	S	N	R	F	E	A	M	Decision
1	Fail	Fail	Pass	Fail	Pass	Pass	Fail	Reject
2	Pass	Fail	Pass	Fail	Pass	Pass	Fail	Reject
3	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Accept
4	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Accept
5	Pass	Fail	Pass	Fail	Pass	Pass	Fail	Reject
6	Fail	Fail	Pass	Fail	Pass	Pass	Fail	Reject
7	Fail	Fail	Pass	Fail	Pass	Pass	Fail	Reject
8	Fail	Fail	Fail	Pass	Pass	Pass	Pass	Reject
9	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Accept
10	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Accept
11	Fail	Fail	Fail	Pass	Pass	Pass	Pass	Reject
12	Pass	Pass	Pass	Pass	Pass	Pass	Fail	Reject
13	Fail	Fail	Pass	Fail	Pass	Pass	Fail	Reject
14	Fail	Fail	Fail	Pass	Pass	Pass	Pass	Reject
15	Fail	Fail	Fail	Pass	Pass	Pass	Pass	Reject

Rounds 4 and 9 are accepted despite lower adverse-impact Recall because both Recall drops remain within 0.02, selection scores improve, net repair is positive, and every other boundary passes. Risk-evidence F1 and Recall changes are zero in all 15 saved decisions; their parent values are 0.306667 and 0.403509, respectively. These components are reported as recorded and are not omitted from the gate. The normal-subset diagnostics refer to the 24 development cases designated normal in the gate records. They are distinct from test-set evaluation.

H Additional Mechanism Analyses

H.1 Risk-Evidence Association Graph

The saved graph uses ground-truth risk evidence from 759 training cases and contains seven association rules. The mining thresholds are a minimum union count of two, historical association confidence of 0.2, and Lift of 1.2. Historical association confidence is

$$\text{Confidence}(U \to V) = \frac{\text{Support}(U \cup V)}{\text{Support}(U)}.$$

It is the fraction of training cases containing antecedent U that also contain consequent V. It is neither the probability that V holds in the current case nor the probability that any adverse impact is correct. The saved rules are descriptive mechanism artifacts; no matched graph-disabled run is available in the selected result suite. Consequently, rule counts or role-level proposals do not establish a graph-induced improvement in adverse-impact prediction.

H.2 Normal-Case Audit Status

The 77-case subset saved under the normal-audit name has 821 ground-truth-positive adverse impacts, with 3–25 per case and no empty ground-truth set of adverse impacts. Its single empty prediction does not establish correct abstention. Neither a deployment false-alarm rate nor an abstention rate for cases with empty ground truth can be inferred from this subset under the annotations actually used for scoring.

I Adverse-Impact Definitions

The following definitions cover all 36 adverse impacts in the 6×6 impact space. Affected-entity and impact-domain names follow the terminology used in the main text. The definitions specify prediction criteria; typical indicators illustrate the corresponding mechanisms.

I.1 SELF

Reputation. Mark 1 only when the current event reveals a clear mechanism of reputational risk and there is sufficient reason to predict an effect on the public figure's image, reputation, credibility, or social evaluation. Reputational damage need not already have fully occurred. *Typical indicators:* Highly controversial statements; clear violations of social norms; concentrated media attention; potential public backlash, public apology, or loss of trust.

Career. Mark 1 only when a clear occupational transmission path supports predicting effects on the public figure's position, professional qualifications, career opportunities, political career, or business development. Do not infer career damage merely because the event is "very serious." *Typical indicators:* Possible employer or board intervention; clear violations of professional norms; pressure for suspension or resignation; risks to professional qualifications; possible termination of core partnerships.

Safety. Mark 1 only when a specific safety-exposure mechanism supports predicting personal, information, cyber, or real-world safety risks to the public figure. *Typical indicators:* An explicit target of threats; confrontation with extremist groups; exposure of personal information; persistent harassment; mobilization of violence; increased security needs.

Economic. Mark 1 only when a clear economic-loss path supports predicting effects on the public figure's income, assets, equity, investments, contracts, or commercial returns. *Typical indicators:* Possible contract cancellation; risk of fines; equity-price exposure; endorsement-income risk; liability for damages; declining investment value.

Legal. Mark 1 only when clear risks of illegality, grounds for litigation, regulatory triggers, or a path to legal liability support predicting that the public figure may enter formal legal or regulatory proceedings. Do not infer legal risk merely from "improper conduct." *Typical indicators:* Suspected illegality; clear coverage by regulatory rules; potential plaintiffs or victims; existing complaints or enforcement precedents; possible investigation or prosecution.

Value. Mark 1 only when the statement or conduct reveals a clear conflict of moral, ethical, political, or social values, and there is sufficient reason to predict that this position will become a significant point of controversy in public evaluation of the public figure. *Typical indicators:* Clear conflicts involving discrimination, hate, human rights, integrity, labor ethics, gender, political positions, or social responsibility.

I.2 TARGETED PARTY

Reputation. There must be explicit disparagement, accusation, stigmatization, or negative evaluation, with sufficient reason to predict damage to the targeted party's public image or social evaluation. *Typical indicators:* Public accusations of criminality, incompetence, immorality, low worth, or dangerousness.

Career. There must be a clear occupational or business transmission path supporting predictions of effects on the targeted party's position, business, studies, or career opportunities. *Typical indicators:* Calls for dismissal, boycotts, disqualification, withdrawal of cooperation, or influence on customer decisions.

Safety. There must be a specific mechanism of threats, incitement, identity exposure, or hostile mobilization supporting predictions of real-world safety risks to the targeted party. *Typical indicators:* Death threats; publication of addresses; incitement to attack; group hostility.

Economic. There must be a clear economic transmission mechanism supporting predictions of losses to the targeted party's income, orders, assets, or commercial interests. *Typical indicators:* Boycotts; order cancellation; litigation damages; customer loss.

Legal. There must be a clear path to legal liability or enforcement supporting predictions that the targeted party may enter investigation, litigation, penalties, or judicial proceedings as a result. *Typical indicators:* Reports to authorities; public allegations of illegality; regulatory intervention; targeted policies.

Value. There must be a direct attack on, rejection of, or moral evaluation of a value position, with sufficient reason to predict that the targeted party will become involved in a clear ethical, political, or social-value conflict. *Typical indicators:* Accusations of discrimination, opposition to human rights, immorality, extremism, or ethical violations.

I.3 AFFILIATED ORGANIZATION

Reputation. Mark 1 only when a clear reputational transmission relationship connects the public figure to the affiliated organization, and there is sufficient reason to predict harm to the organization's brand, reputation, or social evaluation. *Typical indicators:* A CEO, founder, or school leader strongly represents the organization; the statement is directly associated with a company, school, or team; the organization may be compelled to respond.

Career. There must be a clear business or organizational transmission mechanism supporting predictions of effects on the affiliated organization's operations, projects, personnel stability, market position, or business development. *Typical indicators:* Risks of customer loss or project suspension; employee departures; management crises; affected business qualifications.

Safety. There must be a specific mechanism supporting predictions of increased safety risks to the affiliated organization's personnel, facilities, network systems, or operations. *Typical indicators:* Possible escalation of protests; facilities becoming targets of attack; cyberattack risks; risks to employee safety.

Economic. There must be a clear economic transmission path supporting predictions of effects on the affiliated organization's revenue, profit, share price, market value, orders, financing, or compensation costs. *Typical indicators:* Boycotts potentially affecting sales; possible cancellation of customer orders; share-price exposure; financing obstacles; risk of regulatory fines.

Legal. There must be a clear mechanism for the transmission of legal liability supporting predictions that the affiliated organization may face litigation, regulatory investigation, administrative penalties, or compensation proceedings because of the public figure's conduct. *Typical indicators:* Conduct in an official capacity; company involvement; employer liability; securities-disclosure obligations; platform liability.

Value. There must be a clear relationship of value representation supporting predictions that the public figure's event will be viewed as an issue of the affiliated organization's culture, governance, or social responsibility. *Typical indicators:* Executive statements; controversies over organizational culture; internal employee conflicts; diversity, equity, inclusion, ethics, or governance issues.

I.4 PARTNER

Reputation. There must be a real commercial relationship and a clear mechanism of brand association supporting predictions that a partner may face public-opinion spillover because of its relationship with the public figure or organization. *Typical indicators:* Highly public co-branding, endorsements, sponsorships, or distribution partnerships; consumers may evaluate the two parties together.

Career. There must be a real commercial relationship and clear reasons to predict suspension of cooperation, contract termination, project cancellation, or reduced business opportunities for a partner. *Typical indicators:* Brand-risk clauses; publicly expressed partner concern; conflicts of values; regulatory risks.

Safety. There must be a real commercial relationship and a specific mechanism supporting predictions of effects on a partner's personnel, facilities, systems, or operational safety. *Typical indicators:* Threats to joint events; attacks on partner stores; security risks to shared systems.

Economic. There must be a real commercial relationship and a clear economic chain supporting predictions of harm to a partner's revenue, orders, investments, assets, or contractual returns. *Typical indicators:* Product boycotts; lower co-branded sales; contract cancellation; distribution-channel losses.

Legal. There must be a real commercial relationship and a clear path to legal liability supporting predictions that a partner may be drawn into litigation, investigation, or compensation. *Typical indicators:* Joint promotion, issuance, or sales; shared disclosure obligations.

Value. There must be a real commercial relationship and sufficient reason to predict conflict between the parties over ethics, politics, social responsibility, or brand positions. *Typical in-*

dicators: Conflicts involving brand safety, environmental/social/governance commitments, discrimination, political positions, or human rights.

I.5 AUDIENCE

Reputation. There must be a clear mechanism of information dissemination and cognitive change supporting predictions of a substantial change in the public's overall evaluation, trust, or perception of a subject or group. *Typical indicators:* Large-scale dissemination; authoritative endorsement; repeated exposure; public-trust issues.

Career. There must be a clear mechanism transmitting effects to group employment or careers, supporting predictions of effects on a segment of the public's employment, career opportunities, or social participation. *Typical indicators:* AI replacing jobs; policies restricting entry into professions; risks of large-scale layoffs.

Safety. There must be a specific risk mechanism supporting predictions of effects on the public's physical, public, cyber, or health safety. *Typical indicators:* Dangerous products; misleading medical information; incitement to violence; network vulnerabilities; public-safety risks.

Economic. There must be a clear path to group economic losses supporting predictions of financial losses for consumers, investors, or the general public. *Typical indicators:* Promotion of high-risk investments; price manipulation; fraudulent sales; misleading consumers.

Legal. There must be a clear institutional or legal transmission path supporting predictions of effects on the public's legal rights, obligations, or litigation circumstances. *Typical indicators:* Consumer class actions; policy changes; restrictions on rights; changes to regulatory rules.

Value. There must be a social-diffusion mechanism of sufficient scale and intensity supporting predictions that the event may reinforce, change, or intensify the public's ethical, political, or social-value conflicts. Ordinary short-lived arguments must not be marked 1. *Typical indicators:* Nationwide controversies over values; social movements; major ethical issues; sustained public discussion.

I.6 FIELD

Reputation. There must be a reputational transmission mechanism capable of spreading across multiple actors in a field, supporting predictions of effects on the social credibility and public evaluation of the entire field or a substantial portion of its actors. *Typical indicators:* Exposure of problems common to the field; systemic fraud; widespread questioning of the field.

Career. There must be a structural transmission mechanism supporting predictions of changes to the field's employment structure, market access, competitive landscape, business models, or development direction. *Typical indicators:* Technological substitution; large-scale exits; new business models; platform-rule changes.

Safety. The event must expose a safety risk common across the field, supporting predictions of changes to its safety standards, risk-governance systems, or safety investment. *Typical indicators:* Shared vulnerabilities; accidents within the field; systemic risks.

Economic. There must be an economic transmission mechanism capable of spreading across multiple actors in the field, supporting predictions of substantial effects on its market size, investment, capital, sales, or revenue. *Typical indicators:* Capital withdrawal; field-wide declines in demand; systemic market shocks.

Legal. There must be a clear regulatory-spillover mechanism supporting predictions that the event may change enforcement, regulatory policies, compliance rules, or legal boundaries across the field. Formal new rules need not already have been enacted. *Typical indicators:* First enforcement actions; regulatory warnings; policy discussions; explicit statements by regulators; precedent-setting cases.

Value. There must be an ethical-spillover mechanism extending beyond a single actor, supporting predictions of sustained debate within the field over values, social responsibility, technology ethics, or behavioral norms. *Typical indicators:* Ethical controversies across the field; discussions of standard-setting; conflicts of professional norms; sustained changes in field-wide consensus.

J RISK EVIDENCE OBSERVED IN PRESAGE-BENCH

The risk-evidence inventory is summarized from laws and regulations in multiple jurisdictions, corporate rules, and social-media platform risk-control policies. We extract common descriptions of specific types of problems in statements from these sources, harmonize their terminology, remove semantic overlap, and refine their boundaries to form a fine-grained inventory $\mathcal{L}$. Each entry has a unified name and definition and describes a specific type of problem in a statement; it does not encode the affected entity, impact domain, or later outcome. The inventory therefore provides a shared representation of risk while retaining sufficient granularity for independent adverse-impact judgement.

Across the 1,012 cases, 67 distinct inventory entries occur: 66 types of risk evidence and the normal-statement entry. Only these observed entries are listed below; the 96 source-inventory entries absent from the dataset are omitted. The entries below retain their original codes for traceability. A code identifies an entry in the risk-evidence inventory. When the source gives the same text for a name and its definition, that text is presented only once. The original definitions remain the reference for the archived experiments.

`SEC-V17-016` Disseminating false information that may harm social order, public safety, or public judgement.

`SEC-V17-018` Glorifying, distorting, or eroding memories of wars of aggression, colonial history, and collective national trauma.

`SEC-V17-026` Promoting money worship, hedonism, vulgar status competition, or other harmful values.

`SEC-V17-028` Publishing information that disrupts public order, endangers public safety, or infringes personal and property rights.

`SEC-V17-063` Comparing individuals or groups to Nazis, hate organizations, extremist organizations, or sexual offenders.

`SEC-V17-079` Seeking, providing, or promoting illegal adoption, surrogacy, fetal sex identification, or unlicensed genetic testing.

`SEC-V17-082` Promoting, glorifying, or normalizing child sexual exploitation.

`SEC-V17-101` Providing professional financial advice, such as stock-trading instructions or investment guides, without financial qualifications.

`SEC-V17-103` False advertising through exaggerated product effects, fabricated information, or invented cases.

`SEC-V17-119` Investment inducement: promotional language that induces others to participate in investments.

`EXT-PF-001` Expressing glee, celebration, mockery, or contempt toward deaths, casualties, disasters, or victims. Directly endorsing outcomes involving severe harm, mocking deceased people or victims, describing them as deserving their fate, or treating their suffering as retribution.

`EXT-PF-002` Using sexual violence, death, disasters, or other severe harm as jokes, metaphors, or entertainment. Frivolously turning severe harm into jokes, puns, metaphors, or entertainment.

`EXT-PF-003` Evaluating consumers, users, students, or other service recipients with arrogance, contempt, or disparagement. Publicly denying service recipients' judgement, ability, qualifications, or social value.

`EXT-PF-004` Disparaging, stigmatizing, or denying the professional value of employees or occupational groups. Blaming organizational problems on an entire occupational group, or publicly denying the necessity and value of a category of employees or professional positions.

`EXT-PF-005` Undermining service recipients' dignity through humiliation, reprimands, or undignified expression while performing public duties or professional services. Judges, teachers, doctors, officials, or others directly humiliate, harshly reprimand, or disparage those they serve, manage, or adjudicate while performing their duties.

EXT-PF-006 Insulting, disparaging, or improperly instrumentalizing the nation, military, or other public symbols. Directly insulting the nation or military, or improperly attributing public symbols and military honor to an individual or political party.

EXT-PF-007 Alleging that judicial, administrative, electoral, or other public institutions are politically manipulated, corrupt, or illegitimate. Explicitly claiming that public institutions, procedures, or decisions involve political retaliation, bribery, manipulation, or illegitimate arrangements. This entry records the mechanism of accusation without determining whether the accusation is true.

EXT-PF-009 Quoting or reposting earlier statements as ironic commentary after a death, violence, or disaster. The publisher adds no explicit evaluation, but the timing, quoted subject, and immediate context make the earlier statement understood as an ironic response to an actual death, violence, or disaster.

EXT-PF-010 Making insufficiently supported specific factual allegations against individuals, enterprises, or organizations that may harm their reputation. Publicly accusing an identifiable subject of corruption, fraud, sexual conduct, opinion manipulation, or other verifiable conduct without sufficient support in the available materials.

ATM-VIO-002 Inciting, encouraging, or calling on others to commit violence. Promoting, calling for, or encouraging others to commit real-world violence.

ATM-VIO-003 Approving, glorifying, or justifying real-world violence and its outcomes. Endorsing or glorifying violent acts, or describing their outcomes as justified or deserved.

ATM-VIO-006 Organizing or participating in violent harm, destruction of public facilities, or organized-crime activities. Organizing or participating in violent harm, destruction of facilities, or unlawful organized-crime activities.

ATM-CRI-001 Displaying, glorifying, or promoting illegal or criminal activities. Presenting illegal or criminal activities other than specifically violent acts in an approving or promotional manner.

ATM-HAR-001 Attacking individuals with insults, slurs, or personal degradation. Directly attacking individuals with profanity, slurs, derogatory nicknames, or personal degradation.

ATM-HAR-003 Intimidating, threatening, or coercing individuals. Intimidating or coercing a specific individual through nonviolent means or without an explicit threat of violence.

ATM-HAR-004 Organizing cyberbullying, doxxing, or coordinated attacks on individuals. Organizing, mobilizing, or carrying out coordinated online attacks or doxxing.

ATM-HATE-001 Promoting hateful ideologies or supporting hate organizations. Promoting hateful ideologies or publicly supporting hate organizations and dangerous organizations.

ATM-HATE-002 Disparaging, excluding, or discriminating against groups based on identity characteristics. Disparaging or excluding groups based on ethnicity, race, religion, nationality, region, or other identity characteristics.

ATM-HATE-003 Inciting hostility, harm, or conflict against identity groups. Promoting mobilized hostility, harm, or real-world conflict against identity groups.

ATM-HATE-004 Using or disseminating hate-organization symbols, slogans, or coded language. Using visual signals, emblems, slogans, coded language, or suggestive symbols associated with hate organizations.

ATM-CON-001 Inciting conflict through disparaging comparisons, provocation, trolling, or hostile rhetoric. Promoting antagonism or conflict through disparaging comparisons or provocation, without requiring identity-based discrimination.

ATM-MIS-001 Disseminating factual information lacking reliable sources or verification. Disseminating factual claims that lack credible sources, have insufficient information, or have not been verified.

ATM-MIS-002 Misleading audiences by distorting context, impersonating involved parties, or recycling old news. Misleading through selective quotation, concealment of timing, impersonation of involved parties, or presenting old news as new.

ATM-MIS-003 Spreading rumors about public events, disasters, accidents, or public welfare. Disseminating false information about public events, disasters, crime, or public welfare.

ATM-MIS-004 Disseminating false or misleading information about politics, elections, or civic procedures. Disseminating false or misleading information about politics, elections, voting, or public procedures.

ATM-MIS-005 Disseminating false or misleading information about medicine, vaccines, or public health. Disseminating misinformation about diseases, vaccines, treatments, or public health.

ATM-MIS-007 Misleading audiences through false identity, source, or account information. Influencing public judgement by disguising an account's identity, source, location, or popularity.

ATM-NS-001 Publishing information that endangers national unity, sovereignty, or territorial integrity. Publishing content that directly endangers national unity, sovereignty, or territorial integrity.

ATM-NS-002 Inciting treason, secession, rebellion, or subversion of state power. Organizing, inciting, or promoting treason, secession, rebellion, or subversion.

ATM-NS-004 Organizing, participating in, supporting, or promoting terrorism and extremism. Participating in, supporting, or promoting terrorist or extremist organizations and activities.

ATM-PRI-001 Trading, leaking, or disseminating personal information and private data. Trading, leaking, or disseminating personal identity information, private data, or account data.

ATM-SEX-001 Publishing sexually suggestive, provocative, or associatively sexualized content. Creating sexual suggestion or provocation through text, images, actions, or incongruous associations.

ATM-SEX-003 Sexually harassing, humiliating, or insulting others. Directly harassing, humiliating, or insulting a specific target through sexualized expression.

ATM-IP-001 Plagiarizing or infringing copyright and other intellectual property. Copying, plagiarizing, or infringing copyright and other intellectual property rights.

ATM-PUB-001 Disseminating or promoting content or conduct that undermines public health, public order, accepted customs, or public morality. Disseminating, displaying, or promoting specific content that undermines public health, public order, accepted customs, or public morality.

EXT-V7-001 Neglecting required statutory-language obligations in major public communications. An institutional leader responsible for communicating with multiple statutory-language communities unjustifiably omits a required statutory language in communication about a major public event or crisis.

EXT-V7-002 Disparaging, stigmatizing, or excluding groups based on education or overseas background. Disparaging, stigmatizing, excluding from employment, or making generalized security allegations against groups solely because of their education, overseas study, or overseas work or residence.

EXT-V7-003 Using disparaging rhetoric against competitors or their products or services in commercial competition. A business operator or representative publicly attacks an identifiable competitor, business model, product, or service through humiliating, contemptuous, or disparaging rhetoric, without the materials confirming improper means such as fabricated facts, threats, or extortion.

EXT-V7-004 Dismissing service recipients' objections through exclusionary responses in teaching or professional services. Teachers, doctors, officials, or others performing professional duties dismiss specific objections through statements such as "you can leave," refusal to listen, or other exclusionary responses, without reaching the threshold of direct humiliation or reprimand.

EXT-V7-005 Imposing unrelated controversial political advocacy in formal institutional ceremonies. A host, master of ceremonies, or institutional representative uses a formal ceremony serving multiple groups, such as a graduation or commemoration, to insert unauthorized controversial political advocacy or mobilization unrelated to the ceremony's central purpose.

EXT-V7-006 Glorifying, minimizing, or denying historical rulers' systematic oppression of religious or ethnic groups. Glorifying, minimizing, or denying systematic persecution, oppression, or mass harm inflicted by historical rulers on particular religious or ethnic groups when sufficient historical evidence or authoritative conclusions support that history.

EXT-V7-007 Objectifying and degrading gender groups through vulgar references to sexual organs or body parts. Using vulgar references to sexual organs, breasts, or other body parts to reduce women, men, or other gender groups to sexualized bodily characteristics and objectify or degrade them.

EXT-V7-008 Improperly manipulating judicial proceedings by selecting judges or evading random case assignment. A litigant evades random assignment, selects a particular judge, and improperly influences proceedings by withdrawing and refiling a case, filing selectively, or making other arrangements.

EXT-V7-009 Embedding controversial political positions in teaching or professional-service materials with risks of identity-based exclusion. A teacher, doctor, or other professional embeds an identifiable controversial political position in teaching, training, or service cases, and a formal complaint or institutional investigation specifically concerns resulting risks of discrimination, harassment, or retaliation against learners, colleagues, or service recipients of a particular identity. This entry records a mechanism at the formal-investigation stage, not a confirmed finding of unlawful discrimination. Neutral mentions of public events, necessary academic analysis, and situations without identity-based exclusion are excluded.

V7-EXT-001 Disparaging, mocking, or excluding groups based on wealth, class, spending power, or social status. Directly expressing contempt for, mocking, excluding, or denying the worth of groups based on wealth, class, spending power, social status, or ordinary family circumstances. This covers only expression that directly makes poverty, class, or spending power the target of humiliation; displays of wealth, ordinary commercial judgements, and individual insults are excluded.

V7-EXT-002 Downplaying risks of high-risk products, investments, or business models through careless endorsement, publicity, or celebrity credibility. In finance, cryptoassets, gambling-like consumption, or other high-risk business models, a public figure reduces audience vigilance through advertising, jokes, celebrity trust, or careless statements. Professional investment advice or explicitly false advertising is not required.

V7-EXT-003 Explicitly disregarding employee rights, worker dignity, or reasonable work boundaries in management or employment. A manager or employer directly disregards employee rights, worker dignity, or reasonable boundaries in a specific employment, union, pay, layoff, family/health-boundary, or workplace context. Ordinary restructuring, strategic assessments of AI, factual explanations of layoffs, and apologies or remedies are excluded.

V7-EXT-004 Improperly exploiting social movements, public issues, or cultural symbols in branding, public relations, or commercial communication. A brand, entertainer, or enterprise borrows social movements, disasters, cultural identities, Indigenous or minority symbols, or public issues for promotion, attention, or image management, emptying them of substance, commodifying them, or appropriating them. Ordinary trademark disputes and ordinary brand responses are excluded.

V7-EXT-005 Using procedural or evasive language to downplay duties to protect threatened groups or victims in public-duty contexts. A person responsible for public duties, campus governance, or professional protection downplays a direct protective duty through procedural, technical, or evasive language in hearings, complaints, campus safety, disaster response, or victim protection. Ordinary policy views, academic debates, commercial responses, legal defenses, and apologies are excluded.

V7-EXT-006 Improperly using public, judicial, regulatory, or platform authority to pressure specific cases, procedures, or spaces for expression. A person with public, judicial, regulatory, or platform-governance authority publicly intimidates, pressures, punishes, or intervenes in specific cases, media programs, lawyers or institutions, licenses or qualifications, platform accounts, or public expression, creating risks to procedural fairness or expression. Ordinary defenses, guilty pleas, appeals, and sentencing statements are excluded.

V7-EXT-007 Publishing or defending professional findings involving fabricated data, irreproducibility, research-ethics defects, or academic-integrity problems. A scholar, doctor, researcher, or professional institution publishes, defends, or promotes findings later shown to involve fabricated data, irreproducibility, defective ethics approval, major sample defects, or research-integrity problems. This focuses more specifically on the integrity of research and professional findings than general misinformation does.

V7-EXT-008 Carelessly downplaying risks, suffering, or response duties during public-health emergencies, disasters, wars, or humanitarian crises. A public figure uses careless analogies, jokes, deflection, or entertainment-oriented descriptions to downplay risks, victims' suffering, or response duties in epidemics, disasters, wars, human-rights crises, prisons, or humanitarian crises, without reaching the threshold of mocking death, medical misinformation, or approval of violence.

V7-EXT-009 Insulting, disrespecting, or improperly instrumentalizing religious beliefs, sacred places, ethnic cultures, or historical commemorative subjects. Offending religious beliefs, sacred places, ethnic cultures, historical commemorative subjects, or a group's cultural dignity through jokes, contempt, exploitation, or instrumentalization, without falling under the risk-evidence definitions for illegal religious content, national symbols, or war memories.

V7-EXT-010 Disparaging creators, professionals, the value of their work, or professional competence. Acting as a platform, employer, investor, expert, or peer to disparage the labor, works, or competence of creators, programmers, musicians, entertainers, journalists, teachers, or other professionals. Insults against specific individuals and ordinary industry forecasts are excluded.

V7-EXT-011 Misleading public judgement through conspiracy, deep-state, or generalized political-manipulation narratives. Explaining public events, judicial or regulatory action, or political processes through generalized narratives of a deep state, conspiracy, political persecution, hidden manipulation, or systemic weaponization. This is not limited to electoral procedures and does not require defamation of a specific individual.

V7-EXT-012 Improperly publicizing, amplifying, or commercializing private disputes, intimate-relationship conflicts, or personal privacy issues. A public figure publicizes, turns into attention-generating content, or commercializes family disputes, intimate-relationship conflicts, personal privacy, or internal disagreements, harming the parties' dignity or privacy or the boundaries of public discussion. Ordinary apologies, ordinary denials, and public-policy discussion are excluded.

NORMAL-000 Normal statement. After item-by-item review, the statement itself does not constitute a harmful mechanism under V7; this entry ensures at least one inventory entry per row in the final dataset and supports counting normal statements.