BrandEval: A Two-Track Multi-Agent Benchmark for Risk-Sensitive LLM Crisis Communication

Anonymous ACL submission

Abstract

Evaluating LLMs for crisis communication requires more than judging isolated responses, because public feedback and competitor actions can reshape trust over time. Existing benchmarks remain largely static and single-turn, offering limited evidence about long-horizon risk-sensitive communication. We introduce **BrandEval**, a two-track benchmark for product-crisis communication. BrandEval-Static evaluates stakeholder-aware strategic analysis, while BrandEval-Dynamic stress-tests response policies in **BrandPolis**, a controlled multi-agent sandbox with consumer feedback, competitor responses, and trust-risk trajectories. We also construct **BrandSRD**, a Chinese crisis-response dataset, and instantiate a reference Strategic Agent using **Strategic Rationale (SR)**. Across proprietary and open-source models, we find that static strategic quality is informative but incomplete: it aligns with some end-state outcomes but does not reliably predict volatility and delayed recovery; dynamic stress tests further expose tail-risk failures under manipulative or weakly grounded policies. The code and dataset will be released publicly.

1 Introduction

When a product crisis happens, such as safety issues or service failures, a company's public response can strongly affect how people judge its responsibility, safety, and trustworthiness (Coombs, 2007b; Leskovec et al., 2007). This is especially important on social media, where public opinion can change quickly and spread widely. A slow or vague response may increase public anger (Coombs, 2007a). In contrast, a clear and responsible response may help reduce uncertainty and support trust recovery (Hallahan et al., 2007). For this reason, crisis communication should not be treated as a simple text-generation task. The key question is not only whether a response sounds good, but whether it can help the company handle the crisis over time (Entman, 1993). Figure 1 illustrates why crisis communication requires evaluating response consequences over time, not isolated message quality alone.

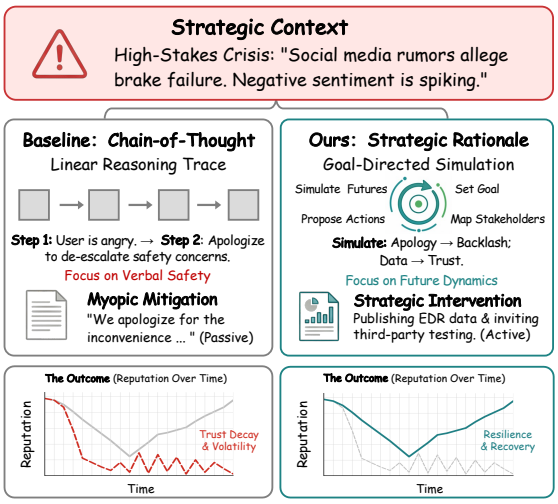

Figure 1: **Divergent Futures.** Under the same crisis context, different response formulations can lead to divergent long-horizon trust trajectories in simulation.

Large language models (LLMs) are increasingly used to support business communication, public relations, and risk management (Li et al., 2024). Although Chain-of-Thought (CoT) methods can make strategic reasoning more inspectable (Wei et al., 2022; Jaech et al., 2024; Guo et al., 2025), current LLM evaluations mostly assess a single answer in a static setting and are therefore not well suited for crisis communication (Zhang et al., 2024). They can measure whether a model gives a reasonable analysis or writes a fluent response, but they lack a dynamic framework for evaluating how communication strategies perform under public feedback and competitive pressure (Grewal et al., 2025). This motivates our main question: *How can we evaluate whether LLMs can make credible, responsible, and effective crisis-communication decisions over time?*

To study this question, we introduce **BrandEval**, a two-track benchmark for risk-sensitive product

crisis communication. The first track, **BrandEval-Static**, evaluates single-turn strategic analysis. It tests whether a model can understand the crisis context, identify key stakeholders (Pendharkar, 2012), reason about risks, and propose practical actions. The second track, **BrandEval-Dynamic**, evaluates response policies in **BrandPolis**, a controlled multi-agent simulation environment. In BrandPolis, a focal brand interacts with consumer agents, competitor agents, and a platform environment over multiple rounds. This setup allows us to observe how different communication strategies affect trust, risk, reputation stability, and recovery over time (Ji et al., 2025). We also construct **BrandSRD**, a crisis-response dataset with de-identified decision points, real and counterfactual response options, preference labels, and explanatory rationales. Based on this dataset, we build a reference Strategic Agent guided by **Strategic Rationale**, which encourages goal setting, stakeholder tracking, and future-aware action selection.

Our experiments show that BrandEval-Static can diagnose whether a model produces a plausible crisis analysis, but it cannot measure whether the proposed actions remain effective after public reactions and competitor responses unfold. BrandEval-Dynamic complements this by revealing long-term trust trajectories, reputation volatility, recovery speed, polarization, and tail-risk failures. These results show that crisis communication agents need more than fluent writing or one-shot reasoning. They need to make credible, evidence-based, and socially responsible decisions across time.

The main contributions are as follows:

- We introduce **BrandEval**, a two-track benchmark pairing static stakeholder diagnosis with dynamic multi-agent evaluation for risk-sensitive product crisis communication.
- We develop **BrandPolis** and **BrandSRD** to support long-horizon simulation, controlled scenario construction, and preference-grounded evaluation.
- We show that static and dynamic evaluations are complementary: dynamic stress testing captures trajectory stability and tail risks that static diagnostic scores cannot assess.

2 Related Work

Recent work on LLM agents with planning and social reasoning has used language models to simulate interactive social systems (Hao et al., 2023; Kosinski, 2023). This simulation paradigm has been extended to commercial contexts, including business decision making (Li et al., 2025; Hazenberg et al., 2025), commercial communication and market interaction (Karande et al., 2024), and quantifying perceptual gaps between brands and user-generated content (Gan et al., 2025). Social simulation systems further model realistic social network dynamics with multimodal agents (Touzel et al., 2025) and scale up large social simulations (Tang et al., 2025). These studies provide useful interaction substrates, but rarely stress-test accountable crisis communication with competitive response.

Agent evaluation has also moved beyond static answer checking, as static QA benchmarks often miss interaction competence and remain vulnerable to contamination (Xu et al., 2024). Recent work addresses this issue through procedural instance generation (Fan et al., 2023; Kurtic et al., 2024) and benchmarks built from time-varying real-world data, such as historical knowledge snapshots and user dialogues (Ouyang et al., 2025; Bayat et al., 2024). Agent evaluation has expanded from general multi-agent suites such as MultiAgentBench (Zhu et al., 2025) to dynamic settings, including asynchronous environments where events evolve independently of agent actions (Froger et al., 2026) and simulated commercial marketplaces with longitudinal competitive dynamics (Kim et al., 2026). Yet these benchmarks do not directly assess public-facing crisis response or the trust and risk trajectories shaped by communication policies.

Domain-focused agent benchmarks have begun to cover enterprise workflows (Vishwakarma et al., 2025; Gupta et al., 2026), customer support (Balaji et al., 2026), and advertising analytics (Hu et al., 2026). Crisis informatics and operational communication, in turn, evaluate practical deliverables from social streams, including situation assessment and information triage (Alam et al., 2021), crisis timeline extraction and summarization (Rajaby Faghihi et al., 2022), and online public opinion report generation (Yu et al., 2025). Together, these lines cover business task execution and crisis information processing, but not how public-facing responses reshape stakeholder reactions, competition, and recovery. Motivated by this gap, BrandEval contributes an evaluation protocol rather than a social simulator alone: it pairs rubric-based static diagnosis with controlled dynamic stress tests of accountability, verifiability, de-escalation, and long-horizon process health.

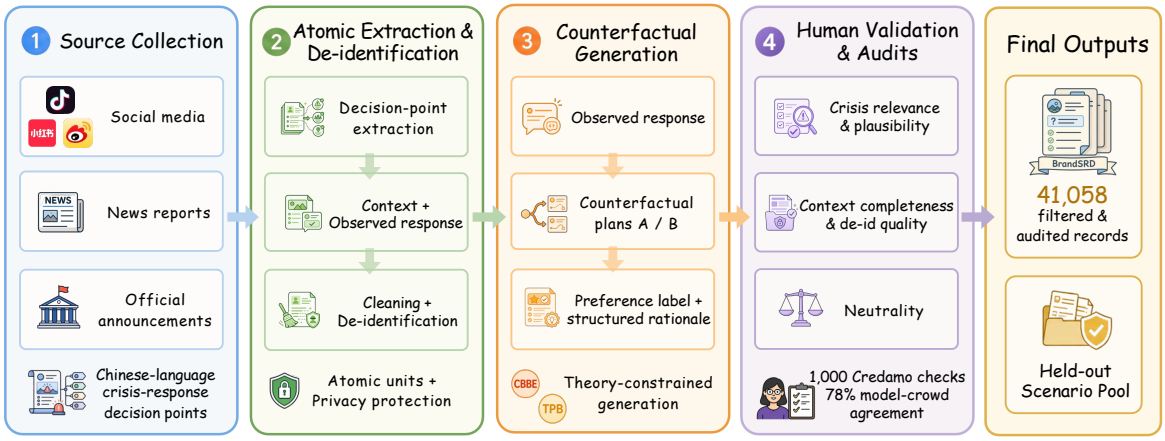

Figure 2: **BrandSRD construction pipeline.** Source collection is followed by atomic extraction and de-identification, counterfactual generation, and human validation and audits, yielding 41,058 filtered and audited records and a held-out evaluation pool for BrandEval.

3 BrandSRD Dataset

3.1 BrandSRD Construction

Dataset overview. BrandSRD is a Chinese-language dataset of 41,058 crisis-response decision points constructed from public crisis-related posts and event context; Table 1 reports the source distribution. It is derived from user posts on three Chinese social-media platforms (Weibo, Douyin, and Xiaohongshu) collected from 2023-12-15 to 2025-12-15, and supplemented with event context from news reports and official announcements. Figure 2 summarizes the BrandSRD construction pipeline and a separate held-out evaluation pool reserved from the same source collection for later benchmark evaluation. Each data record includes a context, the observed brand response, two counterfactual response plans, a preference label, and a rationale.

Scenario construction. Each record is constructed through a controlled scenario-construction pipeline. The pipeline first filters posts and event fragments for crisis-response decision points, where brand communication choices may affect stakeholder trust, public reaction, and recovery. For each retained decision point, we use Gemini-2.5-Pro to extract the crisis context and observed brand response, then de-identify and neutrally rewrite the case into a self-contained scenario while preserving the commercial context, stakeholder trade-offs, and underlying claims. Based on this scenario, Gemini-2.5-Pro generates two counterfactual response plans under Customer-Based Brand Equity (CBBE) (Keller, 1993) and Theory of Planned Behavior (TPB) (Ajzen, 1991) constraints. Each record is then completed with a preference label and a rationale.

Table 1: **BrandSRD source distribution.** Counts of posts used to construct BrandSRD across 12 product domains and three platforms (XHS = Xiaohongshu).

Product	Weibo	Douyin	XHS	Total
Smartwatch	1078	786	482	2346
Refrigerator	1024	1125	741	2890
Air conditioner	1096	553	781	2430
Rice cooker	858	2517	676	4051
Smart door lock	1360	2268	449	4077
Translation pen	1265	1051	614	2930
Smartphone	1177	734	1847	3758
Television	1055	929	921	2905
Speaker	1606	563	1343	3512
Dishwasher	1721	580	465	2766
Personal computer	931	948	872	2751
Electric vehicle	5548	636	458	6642

3.2 Data Quality and Evaluation Audits

To improve the auditability of BrandSRD and the held-out evaluation pool, we audit scenario screening and preference labels. For scenario screening, generated scenario briefs are reviewed for crisis relevance, commercial plausibility, contextual completeness, de-identification quality, and neutrality. Three domain reviewers independently audit 600 randomly sampled scenario briefs, achieving 93.3% agreement and Fleiss' $\kappa = 0.84$. Remaining disagreements are adjudicated by an additional senior reviewer. Rejected cases mainly involve implausible business settings or market behavior (35%), ambiguous links between crisis events and stakeholder reactions (30%), insufficient context (20%), and de-identification or safety issues (15%).

For preference labels, we validate 1,000 randomly sampled BrandSRD records through three-way structured crowd questionnaires on Credamo,[1]

[1] https://www.credamo.com

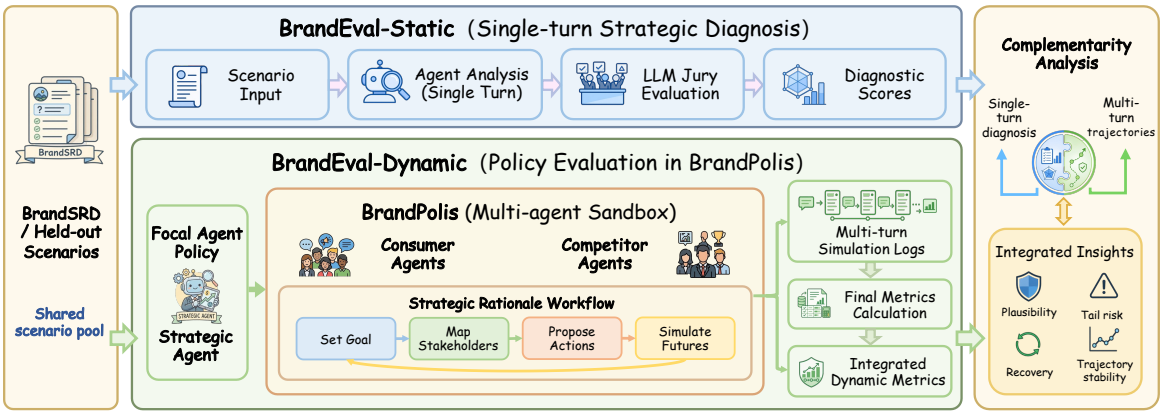

Figure 3: **Overview of BrandEval.** BrandEval pairs Static diagnosis on held-out scenarios with Dynamic policy evaluation in BrandPolis, where Strategic Rationale guides the Strategic Agent to produce integrated insights.

obtaining a 78% model–crowd agreement rate, a mean winning vote share of 0.633, and a mean top-two vote share gap of 0.433. These audits support the reliability of the constructed records and the held-out evaluation pool; additional diagnostics are provided in Appendix.

4 BrandEval Benchmark

This section presents how BrandEval operationalizes crisis communication evaluation through two complementary tracks: BrandEval-Static evaluates single-turn strategic analysis, BrandEval-Dynamic evaluates response policies over multiple rounds, and BrandPolis provides the controlled environment for dynamic evaluation. Figure 3 gives an overview of BrandEval.

4.1 BrandEval-Static

BrandEval-Static evaluates single-turn strategic diagnosis on 1,000 self-contained scenarios drawn from the held-out evaluation pool. Figure 4 summarizes the underlying prototype pool's topic and stakeholder composition.

Evaluation Setup. For each scenario, the model under test acts as a strategic consultant. It receives a self-contained scenario description and outputs

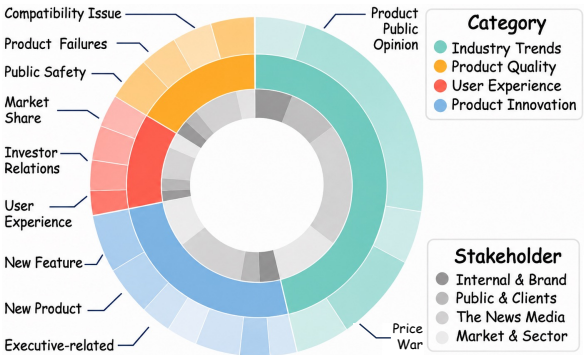

Figure 4: **BrandEval prototype distribution.** Topic and stakeholder composition of 1,000 scenarios.

a strategic analysis report. The report must diagnose stakeholder impacts across consumers, the focal brand, public and media actors, and competitors, then synthesize these trade-offs into an actionable recommendation. This setup tests whether the model can identify affected stakeholders, use evidence responsibly, anticipate competitive dynamics, and formulate a coherent response plan.

Scoring Protocol. We score each report with a stakeholder-theory-inspired rubric (Freeman, 2010) covering four dimensions: Consumer & Brand, Public & Media, Competitor, and Strategic Synthesis. Given the open-ended nature of strategic reasoning, exhaustive expert scoring is costly to scale. We therefore use a three-LLM judge panel following LLM-as-judge practice (Liu et al., 2023). GPT-5-mini, DeepSeek-v3.1, and Claude-4.5-Haiku independently score each report on the four dimensions using a 1–10 scale, and we average their scores per scenario and dimension. Against human scores on 300 sampled strategic-analysis outputs, the panel obtains a mean absolute error of 1.08 points and a Pearson correlation of 0.64. An evidence-sensitive audit further separates verifiable strategies from specious ones with Cohen's $d = 1.34$ and an 89% win rate.

4.2 BrandEval-Dynamic

Evaluation Setup. BrandEval-Dynamic evaluates crisis-response policies through 20-round interactive episodes in BrandPolis. We select 50 scenarios from the same held-out evaluation pool via stratified sampling over topics and risk profiles. For each scenario, the evaluated model serves as the focal crisis-response agent and interacts with consumer agents, competitor agents, and the platform environment. Each published action can affect subsequent stakeholder feedback, competitor

responses, and platform-visible market briefs. We use a fixed horizon $T = 20$ and run each model-scenario pair with $K = 5$ random seeds, reporting seed-averaged metrics.

Reference Strategic Agent We instantiate a reference Strategic Agent for BrandEval-Dynamic using Strategic Rationale, a structured workflow for round-by-round crisis response. Each round, the agent first states a verifiable goal that balances business needs with social responsibility; this goal then constrains stakeholder tracking, candidate planning, and action selection, encouraging commitments that remain credible under public scrutiny (Carroll, 1991). The agent then summarizes the stakeholder state, drafts publishable response candidates, and selects the option that best balances long-horizon credibility and risk for publication.

4.3 BrandPolis Environment

BrandPolis implements BrandEval-Dynamic as a controlled stress-test environment for relative policy evaluation. It is a discrete-time multi-agent sandbox in which published brand actions, consumer feedback, competitor responses, and platform-mediated exposure jointly shape comparable trust and risk trajectories.

Consumer Agents. The population consists of consumer agents whose static personas follow a 5×3 factorial design over decision drivers and brand stances: five primary decision drivers (performance, value for money, user experience, social identity, and safety) (Engel et al., 1986), crossed with three initial stance levels (supporters, neutrals, detractors) (Gera and Neal, 2025) that may shift as new crisis evidence updates their advocacy intent, in line with cognitive dissonance effects (Festinger, 1957). Each agent maintains a five-dimensional internal state that combines three perception dimensions from CBBE (competence belief, value congruence, emotional connection) with two intention dimensions from TPB (purchase intent, advocacy intent). Consumer states evolve via a first-order exponential moving average driven by message-derived impacts:

$$\mathbf{V}_u^{(t)} = (1-\lambda_u)\mathbf{V}_u^{(t-1)} + \lambda_u \gamma^{(t)} \tanh\big(\mathbf{X}_u^{(t)}\big) \quad (1)$$

where $\lambda_u \in (0,1)$ controls update responsiveness, $\mathbf{X}_u^{(t)}$ is an auditable text-to-state impact signal estimated from message features, and $\gamma^{(t)} \in [0,1]$ is an evidence-consistency gate for official claims. We treat $\mathbf{V}_u^{(t)}$ as simulation proxies for belief and intention.

Focal and Competitor Agents. The focal agent is the policy under evaluation: at each round, it receives the current market brief and produces a publishable crisis-response action for the focal brand. This policy can be a zero-shot LLM, a prompted agent, or the SR-guided Strategic Agent. BrandPolis also includes several competitor agents, all operating under bounded rationality (Simon, 1955). Competitor agents follow four archetypes, Prospector, Defender, Analyzer, and Reactor, which adapt Miles and Snow's strategic typology (Miles et al., 1978) into utility-driven personas with distinct goals and risk preferences. For each scenario, we specify which competitor brands are present and assign one archetype to each. The focal agent and all competitor agents alternate between external observation, where they track topics and platform-visible reputation proxies, and internal monitoring of their own published commitments. Evaluation tracks how actions shift belief and intent trajectories, while the next-round prompt uses the market brief to convert current market intelligence into a concrete strategic action.

Platform Environment. The platform runs in discrete rounds and mediates exposure under attention scarcity. At round t, it collects brand, competitor, and user-generated messages. Official brand posts are treated as broadcast and prioritized in consumers' feeds, while other items are sampled by sender influence and topic heat, mimicking weak-tie diffusion in crises (Granovetter, 1973). Each consumer has feed capacity C and processes at most C items, with processing probability modulated by susceptibility and stance alignment; processed messages update CBBE and TPB internal states via Eq. (1). The platform produces a market brief of platform-visible aggregates and dominant topics as the next observation S_{t+1}; it is built only from the public message log and engagement counters and excludes latent belief states and evaluation metrics. To discourage unsupported evidence claims, BrandPolis applies an evidence-consistency gate $\gamma^{(t)} \in [0,1]$ that attenuates the impact of unsupported evidence claims in official actions.

Evaluation Metrics. We evaluate dynamic performance with six metrics grouped into end-state outcomes and trajectory-level process-health signals. As public-interest outcomes, Brand Reputation Index (BRI) (Fombrun et al., 2000) operationalizes

population-level trust as an aggregated reputation signal over CBBE-aligned perception dimensions. Net Promoter Score (NPS) (Reichheld, 2003) operationalizes advocacy as a population-level summary of the TPB-aligned intent signal. Share of Voice (SoV) (Farris et al., 2010) contextualizes competitive attention as the brand's share of crisis-topic-related discussion volume within the sandbox. SoV is an attention measure and has no uniform preferred direction. As process-health signals, Polarization Index (Pol.) (Esteban and Ray, 1994) measures stance polarization across supporters, neutrals, and opponents. Average Reputation Volatility (ARV) (Leberknight et al., 2011) captures volatility of the reputation trajectory over the full horizon, and Crisis Recovery Speed (CRS) (Coombs, 2007b) measures recovery as how quickly reputation rebounds after the episode bottom. We report per-round trajectories and summarize them with end-of-horizon values (BRI_T, NPS_T, SoV_T, Pol_T) and trajectory-level summaries (ARV, CRS), enabling comparisons between final outcomes and process health.

5 Experiments

5.1 Experimental Setup

We evaluate mainstream proprietary and open-source LLMs under the same BrandEval protocol, using zero-shot prompting for direct LLM baselines. BrandEval-Static follows the stakeholder-aware scoring protocol in Section 4.1, and BrandEval-Dynamic follows the interaction protocol in Section 4.2. In addition to zero-shot baselines, we include a reference Strategic Agent initialized from Qwen3-8B (Yang et al., 2025), an open-weight backbone well suited to fine-tuning in BrandSRD's Chinese setting. It is trained only on the BrandSRD training split and evaluated on disjoint BrandEval held-out scenarios with the Strategic Rationale workflow; training details are provided in Appendix.

5.2 BrandEval-Static Results

We first report results on BrandEval-Static, which evaluates single-turn strategic analysis quality under a multi-LLM judge panel (Table 2). Two findings are salient. First, frontier general-purpose models dominate static scores, consistent with the Static track rewarding immediate diagnostic coverage and rhetorically complete reports in a single turn. Second, Strategic Agent is competitive with strong open models but does not surpass the best proprietary systems; this gap is informative rather than surprising, because SR is optimized for iterative crisis handling rather than maximizing one-shot narrative completeness.

Table 2: **BrandEval-Static results.** Multi-LLM panel scores (1–10) averaged over all scenarios. C/B = Consumer & Brand; P/M = Public & Media; Comp. = Competitor; Strat. = Strategic Synthesis.

Model	C/B	P/M	Comp.	Strat.
GPT-5.1	7.86	7.81	7.42	7.60
Gemini-2.5-Pro	7.61	7.64	7.28	7.34
Claude-Sonnet-4.5	7.52	7.71	7.36	7.40
GLM-4.5	7.11	7.16	6.93	6.89
DeepSeek-R1	6.74	6.80	6.46	6.47
R1-Distill-Qwen-7B	4.88	4.31	4.44	4.04
Qwen3-8B	6.20	6.12	6.00	5.73
Qwen3-14B	6.67	6.59	6.61	6.44
Llama 3.1-8B	4.39	3.90	3.89	3.52
Strategic Agent	6.75	6.73	6.51	6.44

5.3 BrandEval-Dynamic Results

Building on the static diagnostic, we evaluate models in BrandEval-Dynamic to test whether single-turn analysis leads to robust long-horizon behavior under interaction (Table 3). Three findings stand out. (i) **Trust gains require process health:** higher reputation and advocacy are more meaningful when accompanied by lower volatility, lower polarization, and faster recovery. For example, Strategic Agent nearly matches GPT-5.1 on single-competitor trust ($BRI_T = -2.36$ vs. -2.34) while reducing volatility more strongly (ARV 27.78 vs. 33.09), indicating a stabler recovery path. (ii) **Dynamic metrics reveal trade-offs:** the Strategic Agent achieves competitive end-state trust and recovery while maintaining low polarization across both competitor regimes. This supports evaluating response policies by trajectory stability rather than final trust alone. (iii) **More voice does not mean better outcomes:** higher SoV does not reliably translate into trust gains; credible actions matter more than high-volume messaging.

Mean performance can obscure rare but severe backfires that dominate crisis cost, so we use BrandEval-Dynamic as a tail-risk stress test. We compare the Strategic Agent with two response policies: a Manipulative policy prompted to maximize visibility and suppress competitors even at the expense of accountability, and a Specious policy that produces plausible but weakly grounded statements without verifiable commitments. We report four scenario-level tail-risk indicators: the worst-

Table 3: **BrandEval-Dynamic results.** Mean over 50 scenarios under Single (Prospector) and Dual (Prospector+Analyzer) competitor regimes. For directed metrics, best values are **bold** and second-best values are underlined; SoV is reported descriptively. ↑ indicates higher is better, while ↓ indicates lower is better.

Model	BRI↑		NPS↑		ARV↓		Pol.↓		CRS↓		SoV	
	Single	Dual	Single	Dual	Single	Dual	Single	Dual	Single	Dual	Single	Dual
Initial Value	-12.07	-12.07	-32.00	-32.00	-	-	82.41	82.41	-	-	64.25	64.23
GPT-5.1	**-2.34**	-3.62	**-1.00**	**-2.00**	33.09	31.46	1.98	**0.00**	**14**	**13**	68.27	68.94
Gemini-2.5-Pro	-2.89	-3.16	-3.00	**-2.00**	34.29	33.97	3.92	7.68	15	14	68.93	69.61
Claude-Sonnet-4.5	-3.03	-3.75	-2.00	-4.00	32.36	31.11	1.98	1.98	16	16	68.91	69.59
GLM-4.5	-4.09	-2.95	-3.00	-3.00	30.56	30.22	3.92	1.98	**14**	15	69.62	68.94
DeepSeek-R1	-4.04	-5.28	-6.00	-4.00	28.93	28.06	5.82	3.92	16	16	67.65	75.96
R1-Distill-Qwen-7B	-4.26	-3.42	-9.00	-6.00	27.80	30.72	1.98	3.92	18	17	69.60	69.61
Qwen3-8B	-3.54	-4.30	-3.00	-9.00	30.06	29.74	1.98	1.98	16	15	63.46	68.90
Qwen3-14B	-4.01	-5.10	-5.00	-7.00	29.01	27.82	3.92	1.98	17	16	68.89	71.84
Llama 3.1-8B	-5.11	-5.43	-4.00	-8.00	28.68	**24.81**	**0.00**	5.82	18	20	69.90	65.05
Strategic Agent	-2.36	**-2.76**	**-1.00**	**-2.00**	**27.78**	31.83	1.98	1.98	**14**	**13**	69.61	69.99

case outcome across scenarios, CVaR_{10} over the bottom 10% scenarios, and the catastrophic failure rate under $BRI_T < -10$, all computed after averaging $K = 5$ seeds per scenario. Table 4 shows that the dynamic track exposes visibility-seeking and weakly grounded policies as tail-riskier than the SR-guided reference policy in these stress tests.

Table 4: **Tail risk analysis across 50 scenarios.** Min is the worst-case BRI_T over scenarios; CVaR_{10} is the mean of the bottom 10% BRI_T outcomes.

Policy	Min BRI↑	CVaR_{10} BRI↑	Min NPS↑	Fail rate↓
Manipulative	-18.52	-16.20	-35.00	12%
Specious	-10.24	-9.15	-22.00	4%
Strategic Agent	**-5.21**	**-4.82**	**-12.00**	**0%**

5.4 Static and Dynamic Alignment

We next test the central premise of BrandEval: whether static strategic quality predicts dynamic crisis handling. We descriptively quantify this transfer gap with Kendall's τ between model rankings from BrandEval-Static and directed BrandEval-Dynamic outcomes, reversing the ranking direction for lower-is-better metrics. Static rankings align with trust and polarization outcomes ($\tau = 0.67$ for BRI; $\tau = 0.82$ for Pol.) but are negatively associated with process-health metrics (ARV: $\tau = -0.33$; CRS: $\tau = -0.24$). This does not imply the static rubric is wrong: BrandEval-Static rewards one-shot diagnostic coverage, whereas BrandEval-Dynamic stresses multi-turn execution under feedback and competition, where over-commitment, oscillation, and evidence-timing drive stability and recovery. We therefore use BrandEval-Static as a diagnostic complement rather than a standalone proxy for long-horizon process health.

5.5 Benchmark Sanity Checks

We further assess whether BrandPolis exhibits coarse-grained alignment with selected rise-and-fall patterns and discourse structures observed in real online brand and consumer interactions. We study three real-world electric vehicle events, a highway-accident crisis that first broke widely on 2025-03-29 (China Newsweek, 2025), a product-recall incident announced on 2025-09-19 (Yicai Daily, 2025), and a spontaneous-combustion incident on 2025-10-13 (Xiaoxiang Morning Herald, 2025). We compare simulated outputs against a post-derived reference series under matched operationalization, and compare discourse statistics under matched preprocessing, to check whether the sandbox exhibits coarse alignment with rise–fall patterns and salient topic structure (Table 5).

Table 5: **BrandPolis sanity checks.** Macro-level trend alignment and micro-level discourse reproduction across three crisis types.

Metric	Highway accident	Product recall	Spontaneous combustion
Macro trend alignment			
Lagged $r_{\max}$ (±2) ↑	0.60	0.80	0.75
Spearman ρ ↑	0.62	0.56	0.59
Norm-RMSE ↓	0.35	0.32	0.29
DTW (Z-norm) ↓	0.21	0.27	0.23
90% band coverage ↑	0.60	0.65	0.65
Micro discourse reproduction			
Topic JSD ↓	0.18	0.21	0.17
Topic cosine ↑	0.74	0.62	0.76
Top-10 overlap ↑	0.60	0.70	0.70

Macro-level trend alignment. We sanity-check whether BrandPolis approximates event-level rise–

Table 6: **Comparisons and ablations on BrandEval-Dynamic.** Results are grouped into a backbone baseline, prompting baselines with the same backbone, and Strategic Agent variants that ablate SR.

Model	BRI↑		NPS↑		ARV↓		Pol.↓		CRS↓		SoV	
	Single	Dual	Single	Dual	Single	Dual	Single	Dual	Single	Dual	Single	Dual
Backbone baseline												
Qwen3-8B	-3.54	-4.30	-3.00	-9.00	30.06	**29.74**	1.98	**1.98**	16	15	63.46	68.90
Prompting baselines (same backbone, different prompting)												
Qwen3-8B + CoT	-2.98	-3.88	**-1.00**	-4.00	29.50	30.60	3.92	5.82	16	15	68.93	69.58
Qwen3-8B + ReAct	-2.89	-3.74	-5.00	-5.00	35.16	31.95	**0.00**	3.92	**13**	**12**	70.30	69.62
Qwen3-8B + Reflexion	-2.76	-3.86	-2.00	**-1.00**	34.38	30.72	**0.00**	3.92	14	13	68.63	69.61
Strategic Agent ablations (policy/workflow components)												
Strategic Agent (w/o SR)	-3.10	-3.96	-2.00	-3.00	28.90	31.20	1.98	**1.98**	15	14	67.65	69.30
Strategic Agent (SR)	**-2.36**	**-2.76**	**-1.00**	-2.00	**27.78**	31.83	1.98	**1.98**	14	13	69.61	69.99

peak–recovery patterns by comparing its simulated reputation trajectory to a post-derived reference series aggregated from observed crisis discussions under a matched operationalization over the same 60-day horizon; we map each round to a 3-day window to align the 20-round episode with this horizon. We compare simulated and proxy trajectories using lag-aligned Pearson correlation within a fixed ± 2 round window, Spearman rank correlation, normalized RMSE after min–max normalization, dynamic time warping (DTW) on Z-normalized trajectories, and 90% prediction-band coverage. Across events, correlation-based signals are directionally positive, while distance-based comparisons and coverage suggest broadly similar trajectory shapes under the shared operationalization, providing a macro-level sanity check for BrandPolis.

Micro-level discourse reproduction. We further check whether BrandPolis generates plausible crisis discourse structure by comparing simulated and observed posts under matched preprocessing and a shared topic space. We evaluate three aspects: how similar the overall topic mix is (Jensen–Shannon divergence (JSD)), how close the topic meanings are (topic-level cosine similarity), and whether top topics match (Top-10 overlap). Across events, the simulated and real discussions show similar topic mixtures and share many top topics, indicating that BrandPolis recovers broad crisis themes and their approximate prominence during crises.

5.6 Response Strategy Analysis

We run ablations on BrandEval-Dynamic to disentangle generic prompting from policy and workflow components in Table 6. CoT improves advocacy but increases polarization under stronger competition, while ReAct and Reflexion often trade lower polarization for weaker trust or higher volatility. In contrast, the SR-guided Strategic Agent improves over the same backbone from $BRI_T = -3.54/-4.30$ to $-2.36/-2.76$ and from $NPS_T = -3.00/-9.00$ to $-1.00/-2.00$, while keeping recovery competitive. Removing SR erodes these gains, indicating that the workflow adds value beyond the underlying reference policy by enforcing verifiable goals and feedback-aware plan selection, with additional robustness checks reported in Appendix.

6 Conclusion and Discussion

This paper introduces BrandEval, a two-track benchmark for risk-sensitive product crisis communication that pairs a static stakeholder-aware diagnostic with BrandPolis, a controlled multi-agent stress-test environment. Across proprietary and open-source models, BrandEval shows that static strategic quality is useful but incomplete: single-turn analyses capture part of response quality, yet miss process health failures and tail risks that emerge under stakeholder feedback and competitor actions. To support controlled evaluation, we construct BrandSRD and use Strategic Rationale to instantiate a reference Strategic Agent.

Static and dynamic evaluation serve complementary roles. BrandEval-Static diagnoses stakeholder awareness, evidence use, and recommendation quality, while BrandEval-Dynamic tests how communication decisions reshape trust, risk, and recovery under interaction. This separation helps distinguish plausible analyses from agents that sustain credible crisis handling over time, suggesting that effective crisis communication agents should integrate verifiable goal setting, persistent stakeholder tracking, and future-aware action selection.

Limitations

Simulator abstraction. BrandPolis is a controlled stress-test environment, not a real-world reputation forecaster. It uses a simplified interaction structure, simplified CBBE/TPB updates, a single-channel platform abstraction, manually specified hyperparameters, and no explicit social-graph tie strengths. Our event-level trend checks and sensitivity analyses should therefore be interpreted as sanity checks for the sandbox rather than definitive external-validity evidence.

Data and platform scope. BrandSRD and BrandEval are instantiated using Chinese social platforms and selected consumer domains. The benchmark structure is not tied to one language or market, but applying it to other markets, languages, channels, or regulatory contexts will require recalibration.

Silver labels. BrandSRD is semi-automatically constructed: preference labels are distilled with model assistance and then filtered and audited by crowd workers. They remain silver-standard signals and may reflect generator, annotator, or platform-specific biases.

LLM judge proxy. BrandEval-Static uses a small three-model judging panel audited against limited human ratings. Its scores should be interpreted as scalable diagnostic proxies rather than standalone measures of real-world robustness, and should be used together with BrandEval-Dynamic to assess long-horizon behavior under interaction.

Ethical Considerations

All data in BrandSRD are derived from publicly accessible posts on Weibo, Douyin, and Xiaohongshu, supplemented with public news reports and official announcements for event context. We remove or rewrite identifying information and release only de-identified, self-contained scenario briefs and aggregated annotations rather than user-identifiable records. We will release only the de-identified BrandSRD records and will not release any original social-media posts. The counterfactual response plans are generated via commercial LLM APIs solely to construct a risk-aware evaluation and training resource. BrandSRD, BrandEval, and BrandPolis are intended for controlled evaluation, failure analysis, and accountability-oriented policy comparison, not for operational persuasion or deployment-ready crisis management. We explicitly prohibit using these artifacts to generate misleading crisis messaging, suppress criticism, evade accountability, or manipulate public opinion; the benchmark design further penalizes unverifiable claims and manipulative tactics.

References

Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2):179–211.

Firoj Alam, Hassan Sajjad, Muhammad Imran, and Ferda Ofli. 2021. Crisisbench: Benchmarking crisis-related social media datasets for humanitarian information processing. In *Proceedings of the International AAAI conference on web and social media*, volume 15, pages 923–932.

Sumanth Balaji, Piyush Mishra, Aashraya Sachdeva, and Suraj Agrawal. 2026. Beyond ivr: Benchmarking customer support llm agents for business-adherence. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 193–208.

Farima Fatahi Bayat, Lechen Zhang, Sheza Munir, and Lu Wang. 2024. Factbench: A dynamic benchmark for in-the-wild language model factuality evaluation. *arXiv preprint arXiv:2410.22257*.

Archie B Carroll. 1991. The pyramid of corporate social responsibility: Toward the moral management of organizational stakeholders. *Business horizons*, 34(4):39–48.

China Newsweek. 2025. Police respond to Xiaomi SU7 highway crash causing three deaths: Investigation ongoing. Weibo post. Accessed: 2025-12-17.

W Timothy Coombs. 2007a. *Ongoing crisis communication: Planning, managing, and responding*. Sage.

W Timothy Coombs. 2007b. Protecting organization reputations during a crisis: The development and application of situational crisis communication theory. *Corporate reputation review*, 10(3):163–176.

James F Engel, Roger D Blackwell, and Paul W Miniard. 1986. *Consumer behavior*. Dryden Press.

Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail's reader in mass communication theory*, 390:397.

Joan-Maria Esteban and Debraj Ray. 1994. On the measurement of polarization. *Econometrica: Journal of the Econometric Society*, pages 819–851.

Lizhou Fan, Wenyue Hua, Lingyao Li, Haoyang Ling, and Yongfeng Zhang. 2023. Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes. *arXiv preprint arXiv:2312.14890*.

Paul W Farris, Neil Bendle, Phillip Pfeifer, and David Reibstein. 2010. *Marketing metrics: The definitive guide to measuring marketing performance*. Pearson Education.

Leon Festinger. 1957. A theory of congnitive dissonance. *(No Title)*.

Charles J Fombrun, Naomi A Gardberg, and Joy M Sever. 2000. The reputation quotientsm: A multi-stakeholder measure of corporate reputation. *Journal of brand management*, 7(4):241–255.

R Edward Freeman. 2010. *Strategic management: A stakeholder approach*. Cambridge university press.

Romain Froger, Pierre Andrews, Matteo Bettini, Amar Budhiraja, Ricardo Silveira Cabral, Virginie Do, Emilien Garreau, Jean-Baptiste Gaya, Hugo Laurençon, Maxime Lecanu, Kunal Malkan, Dheeraj Mekala, Pierre Menard, Gerard Moreno-Torres Bertran, Ulyana Piterbarg, Mikhail Plekhanov, Mathieu Rita, Andrey Rusakov, Vladislav Vorotilov, and 5 others. 2026. Gaia2: Benchmarking LLM agents on dynamic and asynchronous environments. In *The Fourteenth International Conference on Learning Representations*.

Haotian Gan, Yudong Li, Wanyue Li, and Weidong Tang. 2025. Aligned or apart? multi-agent insights into consumer and brand messaging discrepancies. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6558–6566.

Parush Gera and Tempestt Neal. 2025. Deep learning in stance detection: A survey. *ACM Computing Surveys*.

Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380.

L Grewal, A Stephen, and P Vana. 2025. Brands in unsafe places: effects of brand safety incidents on brand outcomes. *Journal of Marketing Research*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Lakshya Gupta, Litao Li, Yizhe Liu, Sriram Ganapathi Subramanian, Kaheer Suleman, Zichen Zhang, Haoye Lu, and Sumit Pasupalak. 2026. World of workflows: a benchmark for bringing world models to enterprise systems. *arXiv preprint arXiv:2601.22130*.

Kirk Hallahan, Derina Holtzhausen, Betteke Van Ruler, Dejan Verčič, and Krishnamurthy Sriramesh. 2007. Defining strategic communication. *International journal of strategic communication*, 1(1):3–35.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. *arXiv preprint arXiv:2305.14992*.

Thomas Hazenberg, Yao Ma, Seyed Sahand Mohammadi Ziabari, and Marijn van Rijswijk. 2025. Multi-agent reinforcement learning for dynamic pricing in supply chains: Benchmarking strategic agent behaviours under realistically simulated market conditions. *arXiv preprint arXiv:2507.02698*.

Lingxiang Hu, Yiding Sun, Tianle Xia, Wenwei Li, Ming Xu, Liqun Liu, Peng Shu, Huan Yu, and Jie Jiang. 2026. Ad-bench: A real-world, trajectory-aware advertising analytics benchmark for llm agents. *arXiv preprint arXiv:2602.14257*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Yingru Ji, Weiting Tao, and Chang Wan. 2025. A systematic review of attribution theory applied to crisis events in communication journals: Integration and advancing insights. *Communication Research*, page 00936502251319843.

Shirish Karande, V Santhosh, and Yash Bhatia. 2024. Persuasion games with large language models. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 576–582.

Kevin Lane Keller. 1993. Conceptualizing, measuring, and managing customer-based brand equity. *Journal of marketing*, 57(1):1–22.

To Eun Kim, Alireza Salemi, Hamed Zamani, and Fernando Diaz. 2026. Evaluation of agents under simulated ai marketplace dynamics. *arXiv preprint arXiv:2604.14256*.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 4:169.

Eldar Kurtic, Amir Moeini, and Dan Alistarh. 2024. Mathador-lm: A dynamic benchmark for mathematical reasoning on large language models. *arXiv preprint arXiv:2406.12572*.

Christopher S Leberknight, Soumya Sen, and Mung Chiang. 2011. On the volatility of online ratings: An empirical study. In *Workshop on E-Business*, pages 77–86. Springer.

Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. 2007. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5–es.

Sha Li, Revanth Gangi Reddy, Khanh Duy Nguyen, Qingyun Wang, May Fung, Chi Han, Jiawei Han, Kartik Natarajan, Clare R Voss, and Heng Ji. 2024.

Schema-guided culture-aware complex event simulation with multi-agent role-play. *arXiv preprint arXiv:2410.18935*.

Yuan Li, Lichao Sun, and Yixuan Zhang. 2025. Metaagents: Large language model based agents for decision-making on teaming. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–27.

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

Raymond E Miles, Charles C Snow, Alan D Meyer, and Henry J Coleman Jr. 1978. Organizational strategy, structure, and process. *Academy of management review*, 3(3):546–562.

Jie Ouyang, Tingyue Pan, Mingyue Cheng, Ruiran Yan, Yucong Luo, Jiaying Lin, and Qi Liu. 2025. Hoh: A dynamic benchmark for evaluating the impact of outdated information on retrieval-augmented generation. *arXiv preprint arXiv:2503.04800*.

Parag C Pendharkar. 2012. Game theoretical applications for multi-agent systems. *Expert Systems with Applications*, 39(1):273–279.

Hossein Rajaby Faghihi, Bashar Alhafni, Ke Zhang, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2022. Crisisltlsum: a benchmark for local crisis event timeline extraction and summarization. *arXiv e-prints*, pages arXiv–2210.

Frederick F Reichheld. 2003. The one number you need to grow. *Harvard business review*, 81(12):46–55.

Herbert A Simon. 1955. A behavioral model of rational choice. *The quarterly journal of economics*, pages 99–118.

Jiakai Tang, Heyang Gao, Xuchen Pan, Lei Wang, Haoran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, and 1 others. 2025. Gensim: A general social simulation platform with large language model based agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 143–150.

Maximilian Puelma Touzel, Sneheel Sarangi, Gayatri Krishnakumar, and Busra Tugce. 2025. Sandboxsocial: A sandbox for social media using multimodal ai agents. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 11100–11103.

Harsh Vishwakarma, Ankush Agarwal, Ojas Patil, Chaitanya Devaguptapu, and Mahesh Chandran. 2025. Can llms help you at work? a sandbox for evaluating llm agents in enterprise environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9178–9212.

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

Xiaoxiang Morning Herald. 2025. Chengdu road crash followed by vehicle fire; bystanders reportedly unable to open the door during rescue. Weibo post. Accessed: 2025-12-27.

Cheng Xu, Shuhao Guan, Derek Greene, M Kechadi, and 1 others. 2024. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*.

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

Yicai Daily. 2025. Automakers should proactively recall as soon as safety hazards are found. Weibo post. Accessed: 2025-12-17.

Jinzheng Yu, Yang Xu, Haozhen Li, Junqi Li, Yifan Feng, Ligu Zhu, Hao Shen, and Lei Shi. 2025. Oporbench: Evaluating large language models on online public opinion report generation. *arXiv preprint arXiv:2512.01896*.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and 1 others. 2025. Multiagentbench: Evaluating the collaboration and competition of llm agents. *arXiv preprint arXiv:2503.01935*.

Appendix

A Reference Strategic Agent

This section describes how we instantiate the reference Strategic Agent used in BrandEval-Dynamic. The goal is to obtain a fixed and reproducible reference policy for studying structured crisis-response reasoning. All baseline LLMs can be evaluated in BrandEval-Dynamic without this training procedure.

A.1 Training Data Format

Each BrandSRD instance is represented as

$$z_i = (c_i, a_i^{\text{obs}}, \mathcal{A}_i^{\text{cf}}, y_i^+, y_i^-, r_i), \tag{2}$$

where c_i is the crisis context, a_i^{obs} is the observed brand response, $\mathcal{A}_i^{\text{cf}}$ denotes generated counterfactual response options, y_i^+ is the preferred response, y_i^- is the set of non-preferred alternatives, and r_i is the rationale under the stakeholder rubric. The input prompt x_i includes the crisis context, stakeholder information, and formatting instructions. The target follows the Strategic Rationale format, including a verifiable goal, stakeholder-state summary, candidate plans, and a final selected response. For deployment analysis, we log one SR record per Strategic Agent decision in BrandPolis; only the selected plan is posted to the environment.

SR record schema JSON template

```
{
"round_id": "...",
"observation_summary": "...",
"goal": "...",
"stakeholder_summary": "...",
"candidate_plans": [ "...", "...",
"..." ],
"selected_plan": "...",
"final_action_content": "...",
"claimed_evidence": "..."
}
```

A.2 Training Objectives

We initialize from Qwen3-8B and align it on BrandSRD to obtain the policy model π_θ for the reference Strategic Agent. The policy is deployed with the SR workflow, which structures each round into goal setting, stakeholder-state tracking, candidate planning, and final action selection. Training follows a two-stage pipeline.

A.2.1 Stage A: Format Bootstrapping

Stage A bootstraps parseable and evidence-aware SR-style outputs using an iterative DPO objective. The goal is to constrain the model to produce parseable, structured strategic analyses and to construct learnable preference signals for open-ended tasks without explicit ground truth. Given an input scenario x and its evaluation-aligned prompt $\rho(x)$, we sample a candidate set $\{y_k\}_{k=1}^K$ and apply a verifiable gate $G(x, y) \in \{0, 1\}$ to filter out unparsable or schema-violating outputs. For candidates that pass the gate, a frozen self-judge produces a normalized reward $R(x, y) \in [0, 1]$ with a hard evidence constraint: evidence spans must be substrings of the original prompt text; otherwise we set $R(x, y) = 0$. We then select, among gated candidates,

$$\begin{aligned} y^+ &= \arg\max_y R(x, y), \\ y^- &= \arg\min_y R(x, y), \end{aligned} \tag{3}$$

to form a preference triple (x, y^+, y^-). Stage A optimizes an Iterative DPO objective anchored by a reference model π_{ref} to constrain update magnitude:

$$\mathcal{L}_A(\theta) = -\mathbb{E}_{(x,y^+,y^-)}\Big[\log \sigma\big(\beta(\Delta_\theta - \Delta_{\text{ref}})\big)\Big], \tag{4}$$

where

$$\begin{aligned} \Delta_\theta &= \log \pi_\theta(y^+ \mid x) - \log \pi_\theta(y^- \mid x), \\ \Delta_{\text{ref}} &= \log \pi_{\text{ref}}(y^+ \mid x) - \log \pi_{\text{ref}}(y^- \mid x), \end{aligned} \tag{5}$$

and β is a strength coefficient. Stage A only bootstraps parseable and evidence-aware SR-style outputs, and the resulting adapter initializes Stage B.

A.2.2 Stage B: BrandSRD Alignment

Stage B aligns the bootstrapped model on BrandSRD preference annotations using an ORPO-style objective. Each BrandSRD instance provides a strategic context, candidate responses, a preferred response, dispreferred alternatives, and a rationale. We form preference triples (x, y^+, y^-), where y^+ is the annotated preferred response and y^- is sampled from the dispreferred candidate set. We optimize an objective that combines a preference log-odds term with conditional cross-entropy on y^+:

$$\begin{aligned} \mathcal{L}_B(\theta) = -\log \sigma\Bigg(\beta\Big(&\log \pi_\theta(y^+ \mid x) \\ &- \log \pi_\theta(y^- \mid x)\Big)\Bigg) \\ &+ \lambda\big(-\log \pi_\theta(y^+ \mid x)\big), \end{aligned} \tag{6}$$

where β controls the preference term and λ weights the supervised term. The Stage B-aligned policy serves as the language policy model of Strategic Agent in BrandPolis.

A.3 Strategic Rationale Deployment

At deployment time, the aligned policy is wrapped with the Strategic Rationale workflow. At each BrandPolis round t, the focal agent receives a market brief S_t containing platform-visible aggregates, dominant topics, recent public reactions, and its own published commitments. The agent then produces an action through four structured steps:

1. **Goal verification:** state a verifiable communication goal that balances business needs with social responsibility.

2. **Stakeholder tracking:** summarize the current state of consumers, public and media actors, and competitors based only on observable information.

3. **Candidate planning:** draft several publishable response candidates with explicit assumptions, evidence use, and expected risks.

4. **Future-aware selection:** select the response that best balances credibility, de-escalation, recovery potential, and downside risk.

The selected response is published into BrandPolis and becomes part of the public message log for the next round. BrandPolis then updates consumer states, competitor observations, and platform-visible market briefs according to the environment dynamics in Section 4.3.

A.4 Role of the Reference Agent

The SR-based Strategic Agent is a controlled policy implementation, not a benchmark requirement. BrandEval-Dynamic can evaluate any model or agent that produces a publishable crisis-response action from the current market brief. The reference agent provides a reproducible policy for stress-testing BrandPolis and for comparing against zero-shot LLM baselines. It also supports analysis of whether goal verification, stakeholder tracking, and future-aware selection improve long-horizon trust and stability. Thus, conclusions about BrandEval rely on the benchmark protocol and dynamic outcomes, not on assuming that SR is the only valid policy design.

B Benchmark Construction and Audits

B.1 Scenario Construction

BrandEval-Static scenarios are constructed by fictionalizing a held-out prototype pool with Gemini-2.5-Pro, followed by human screening for commercial plausibility and removal of any identifying details. Fictionalization preserves the underlying strategic trade-offs and multi-stakeholder tensions, while ensuring that no real-world entities, individuals, handles, or URLs remain. Disjointness from the BrandSRD training split is enforced by (i) pre-reserving the evaluation pool before training-split construction and (ii) filtering with prototype-level IDs and metadata to prevent overlap.

B.2 BrandEval-Static Rubric

BrandEval-Static is scored using a structured rubric inspired by stakeholder theory. The rubric consists of four dimensions, each scored on a 1–10 scale (higher is better). These dimensions are designed to capture stakeholder diagnosis, evidence use, competitive reasoning, and strategic synthesis, which are later reflected in BrandPolis through perception and intent-related state channels. Table 7 summarizes the operational definitions.

B.3 Dynamic Scenario Selection

For BrandEval-Dynamic, we select 50 scenarios from the held-out pool via stratified sampling over (i) dominant topic clusters and (ii) risk profiles (e.g., safety, governance, product quality, pricing). Risk profiles are annotated during scenario curation and used only to balance evaluation coverage.

B.4 Judge System and Bias Diagnostics

We audit the BrandEval-Static judge panel from two complementary angles: robustness to superficial output changes and sensitivity to evidence grounding. The first audit tests whether scores are driven by presentation artifacts, while the second tests whether evidence-sensitive judging can separate verifiable strategies from specious but fluent responses.

B.4.1 Robustness to Surface Changes

As a targeted robustness check, we conduct sensitivity analyses with style and structural perturbations. We select a random subset of 5 scenarios and perturb the model outputs before feeding them to the judge panel.

Perturbation methods. We apply two types of perturbations while preserving the substantive content of each output:

Table 7: **BrandEval-Static rubric.** Four dimensions scored on a 1–10 scale.

Dimension	Operational definition
Consumer & Brand	Accurate diagnosis of consumer concerns; concrete brand actions; credible evidence (data, timelines, accountability); explicit linkage to brand equity dimensions (competence/value/emotion) when appropriate.
Public & Media	Anticipation of public/media narratives; risk-aware messaging; clarity on responsibility/safety; preventive escalation control; alignment with transparency and verifiability.
Competitor	Recognition of competitive dynamics (e.g., opportunistic attacks, SoV shifts); defensive/offensive counter-moves; avoids self-inflicted openings.
Strategic Synthesis	Coherent trade-offs across stakeholders; prioritized goals; operational plan with milestones; avoids generic PR; actionability and consistency with constraints.

- **Style perturbation:** We rewrite the agent's analysis using two distinct styles, "Flowery" with more elaborate language and "Concise" with simpler and shorter language, while preserving the original insights and semantic content.

- **Structure perturbation:** We shuffle the order of the four analysis sections, namely Customer, Public, Competitor, and Strategic Synthesis, in the JSON input to test whether the judge score is sensitive to presentation order.

Results. Across all four evaluation dimensions, the mean absolute score change is 0.17 for the flowery style perturbation, 0.42 for the concise style perturbation, and 0.25 for section-order shuffling. The maximum absolute change is 1 point in all three settings, and the 90th percentile remains at or below 1.0. These results suggest that the scoring protocol is not primarily driven by formatting, section order, or superficial linguistic style.

B.4.2 Evidence Sensitive Judge Audit

We further evaluate whether the scoring protocol can distinguish verifiable strategies from specious but fluent responses. We construct paired comparisons between evidence-grounded strategies and specious alternatives that contain plausible rhetoric but lack adequate support or accountability. To test whether the evidence-sensitive judge is merely uniformly harsher, we evaluate its paired discriminability between **Verifiable** and **Specious** strategies on matched stress scenarios.

For each scenario i, we compute the paired gap

$$\Delta_i = \text{Score}_i^{\text{Verifiable}} - \text{Score}_i^{\text{Specious}}, \quad (7)$$

and summarize the mean gap $\overline{\Delta}$, a bootstrap 95% confidence interval over $\{\Delta_i\}_{i=1}^N$ with 1,000 resamples, the win rate $\Pr(\Delta_i > 0)$, and paired Cohen's $d = \overline{\Delta}/s_\Delta$, where s_Δ is the standard deviation of paired differences.

Table 8: **Audit Judge paired discriminability** ($N = 9$). The evidence-sensitive judge better separates Verifiable and Specious strategies under matched stress scenarios.

Judge	Mean $\overline{\Delta}$	95% CI	Win(%)	Paired d
Standard	$+0.44$	$[-0.33, 1.33]$	44	0.35
Evidence-sensitive	$\mathbf{+3.11}$	$[\mathbf{1.67}, \mathbf{4.56}]$	**89**	**1.34**

The evidence-sensitive judge is selectively specific: it increases separation primarily when evidence quality differs. This indicates that the audit protocol is sensitive to evidence grounding rather than only surface fluency.

B.4.3 Rubric Reliability

Internal consistency. We evaluate internal consistency across the rubric's dimensions using Cronbach's α:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^k \sigma_j^2}{\sigma_{\text{total}}^2}\right), \quad (8)$$

where k is the number of rubric dimensions, σ_j^2 is the variance of dimension j, and σ_{total}^2 is the variance of the summed score. We obtain $\alpha = 0.9012$, indicating high internal consistency among rubric dimensions.

C BrandPolis Mechanics and Metrics

C.1 BrandPolis Update Equations

This section provides implementation details for the platform exposure, retention, and evidence-grounding mechanisms used in BrandEval-Dynamic.

C.1.1 Exposure and Retention

Official posts are treated as broadcast signals with $P_{\text{push}}(m, t) = 1$. For UGC and other non-broadcast messages, the platform samples feed candidates according to $P_{\text{push}}(m, t)$ and retains at most C messages per consumer per round. Broadcast items are prioritized, and remaining slots are filled with sampled non-broadcast items.

For a non-broadcast message m from sender s on topic k at round t, exposure is:

$$P_{\text{push}}(m,t) = \text{clip}_{[0,1]}\Big(P_{\text{base}}\Big[1 + c_1 \ln\big(I_s(t)+1\big) + c_2 \ln\big(H_k(t)+1\big)\Big]\Big), \quad (9)$$

where P_{base} is the global base rate, $I_s(t)$ is the sender influence score, and $H_k(t)$ is topic heat computed from recent message volume and interactions.

The probability that consumer u processes message m is:

$$P_{\text{ret}}(m,u,t) = P_{\text{push}}(m,t)\, s_u\, a_{m,u}, \quad (10)$$

where $s_u \sim \mathcal{U}(0,1)$ is the information-retention susceptibility, and $a_{m,u} \in [0,1]$ denotes the stance-alignment factor between message m and consumer u.

The observation builder uses only the public message log and engagement counters. It does not read latent state tensors used for computing BRI, NPS, Pol., ARV, or CRS.

C.1.2 Action Grounding Mechanism

BrandPolis maintains an evidence store $\mathcal{E}^{(t)}$. At each round, the strategic agent may output `claimed_evidence` to indicate that an official action is grounded in a checkable artifact, such as logs, inspection reports, or third-party findings. The evidence-consistency gate $\gamma^{(t)} \in [0,1]$ checks whether `claimed_evidence` matches an identifier in $\mathcal{E}^{(t)}$. Unsupported evidence references are therefore attenuated through the belief update in Eq. (11).

C.1.3 State Update and Mapping

Each consumer maintains a latent state vector $\mathbf{V}_u^{(t)}$ for the focal brand, initialized at $t = 1$ and updated for $t = 2, \ldots, T$ via an exponential moving average with bounded nonlinearity:

$$\mathbf{V}_u^{(t)} = (1-\lambda_u)\,\mathbf{V}_u^{(t-1)} + \lambda_u \gamma^{(t)} \tanh\big(\mathbf{X}_u^{(t)}\big). \quad (11)$$

where $\mathbf{X}_u^{(t)}$ is the message-derived impact signal and λ_u is a consumer-specific update rate.

To ground the update dynamics, we fit a text-to-state mapping on a separate human-rated subset. Concretely, we extract lightweight textual features from each message and fit multivariate regression models to predict targets aligned with CBBE and TPB, including the three perception dimensions and the two intention dimensions. The regression predictions are used as the message-derived impact signal $\mathbf{X}_{u,j}^{(t)}$ in Eq. (11), so that consumer states evolve as a first-order EMA driven by observed text. We treat this mapping as an explicit and reproducible modeling choice that keeps the simulator transparent and auditable, while recognizing that the latent states are internal proxies and require independent validation to support real-world interpretation.

C.2 BrandEval-Dynamic Metrics

BrandEval-Dynamic reports six metrics. BRI and NPS are reported on $[-100, 100]$ point scales; SoV and Polarization are reported in percent; ARV summarizes normalized reputation volatility; and CRS is reported in timesteps.

Brand Reputation Index (BRI). BRI captures an aggregate consumer perception signal over three CBBE-aligned dimensions: competence belief (CB), value congruence (VC), and emotional connection (EC). Let $\overline{CB}_t, \overline{VC}_t, \overline{EC}_t \in [-1,1]$ be population averages at timestep t. We first define a raw latent reputation signal:

$$b_t = w_{\text{cap}} \cdot \overline{CB}_t + w_{\text{val}} \cdot \overline{VC}_t + w_{\text{emo}} \cdot \overline{EC}_t, \quad (12)$$

and define the reported Brand Reputation Index (BRI) as:

$$BRI_t = 100 \cdot b_t \in [-100, 100]. \quad (13)$$

We use uniform weights to avoid introducing arbitrary emphasis: $w_{\text{cap}} = w_{\text{val}} = w_{\text{emo}} = 1/3$.

Net Promoter Score (NPS). Let $\pi_t^{\text{pro}} \in [0,1]$ be the share of Promoters (Advocacy Intent > 0.7) and $\pi_t^{\text{det}} \in [0,1]$ be the share of Detractors (Advocacy Intent < 0.3). We define the reported NPS on the standard point scale:

$$NPS_t = 100 \cdot (\pi_t^{\text{pro}} - \pi_t^{\text{det}}) \in [-100, 100]. \quad (14)$$

Share of Voice (SoV). In BrandPolis, we operationalize SoV as a topic-focused share of discussion within the simulated market:

$$SoV_{\text{brand},t} = 100 \cdot \frac{V_{\text{brand},t}}{V_{\text{total},t}} \in [0, 100], \quad (15)$$

where $V_{\text{brand},t}$ counts all brand-related messages in the sandbox at round t (official brand posts plus consumer UGC responding to the focal brand or event), and $V_{\text{total},t}$ counts all messages generated within the sandbox at round t.

Polarization Index. We use a bounded group-based polarization score over supporters (S), neutrals (N), and opponents (O). Let $G = 3$, with group shares $s_{g,t}$ at round t (summing to 1), and group locations μ_g. We set $\mu_S = 1$, $\mu_N = 0$, and $\mu_O = -1$. We compute a raw polarization score using a pairwise-distance form:

$$p_t = \sum_{g=1}^{G} \sum_{h=1}^{G} s_{g,t} s_{h,t} \, |\mu_g - \mu_h|, \tag{16}$$

where $p_t \in [0, 1]$ under this discrete encoding. We report the Polarization Index on a percent scale:

$$Pol_t = 100 \cdot p_t \in [0, 100]. \tag{17}$$

Because polarization uses three discrete stance groups with a finite consumer population, the reported Pol values can appear quantized. We therefore treat Pol as a diagnostic for regime shifts rather than as a fine-grained continuous signal.

Average Reputation Volatility (ARV). ARV summarizes *process instability* largely independent of the absolute reputation level.

For an episode, let $\{BRI_t\}_{t=1}^{T}$ be the reported reputation trajectory (Eq. 13). We first apply per-episode min–max normalization:

$$\widetilde{BRI}_t = \frac{BRI_t - \min\limits_{1 \le \tau \le T} BRI_\tau}{\max\limits_{1 \le \tau \le T} BRI_\tau - \min\limits_{1 \le \tau \le T} BRI_\tau + 10^{-12}}. \tag{18}$$

$\widetilde{BRI}_t \in [0, 1]$. This normalization makes ARV comparable across scenarios with different absolute severity by focusing on *trajectory shape volatility* rather than raw amplitude.

We then define Average Reputation Volatility as the standard deviation of the normalized trajectory:

$$\begin{aligned} ARV &= 100 \cdot \sqrt{\frac{1}{T} \sum_{t=1}^{T} (\widetilde{BRI}_t - \overline{\widetilde{BRI}})^2}, \\ \overline{\widetilde{BRI}} &= \frac{1}{T} \sum_{t=1}^{T} \widetilde{BRI}_t. \end{aligned} \tag{19}$$

ARV is reported on a $[0, 100]$-like scale.

Crisis Recovery Speed (CRS). CRS measures how quickly the reputation trajectory *returns to the episode entry level* after the episode bottom. Let $t_{\text{bottom}} = \arg\min_{1 \le t \le T} BRI_t$. We define the recovery time as the first timestep after t_{bottom} such that the reputation reaches the entry level:

$$t_{\text{recover}} = \min\{t > t_{\text{bottom}} : BRI_t \ge BRI_1\}. \tag{20}$$

We then define CRS (in timesteps) as:

$$CRS = \begin{cases} t_{\text{recover}} - t_{\text{bottom}}, & \text{recovery}, \\ (T+1) - t_{\text{bottom}}, & \text{no recovery}. \end{cases} \tag{21}$$

Lower CRS indicates faster recovery under the same horizon and unit scale.

C.3 Sanity-check Computation Details

This section defines the sanity-check metrics reported in Table 5 and the preprocessing used to align real and simulated series.

Proxy Reputation Trajectory. Within each 3-day window, we rate observed social posts on three CBBE-aligned dimensions and aggregate them into a proxy reputation score with uniform weights (Eq. (12)). We report the proxy on the same point scale as the simulator via $BRI = 100 \cdot b$ (Eq. (13)) for comparability of trends across the 20 windows. The window-level proxy score is computed as the mean over all posts in that window.

Macro Trend Metrics. Let S_t denote the simulated BRI trajectory and R_t denote a proxy reputation trajectory constructed from observed posts under a matched rubric for sanity checking at the same 20-window resolution.

Lag-aligned Pearson correlation. To account for plausible diffusion delays between simulation and real-world reactions, we compute Pearson correlation under integer lags $k \in \{-2, -1, 0, 1, 2\}$ and report $r_{\max} = \max_k \operatorname{corr}(S_t, R_{t+k})$.

Spearman rank correlation. We compute Spearman ρ between S_t and R_t (after applying the best lag found above) to measure agreement in rise-and-fall ordering.

Normalized RMSE. We min–max normalize each trajectory into $[0, 1]$ independently and report the RMSE in the normalized space.

DTW on Z-normalized trajectories. We compute DTW distance after Z-normalizing each trajectory to zero mean and unit variance, so DTW reflects shape similarity rather than absolute scale.

90% prediction-band coverage. For each real-world window t, we compute the empirical standard deviation of post-level BRI scores within that window and form a 90% band around the real mean. We then measure the fraction of windows whose simulated value falls inside the corresponding band. To compare on a common scale, we first linearly rescale the simulated series over the 20 windows to match the real series mean and standard deviation, then evaluate coverage.

Table 9: **Diffusion-signal alignment.** Metrics are averaged over three real-world crisis events.

Signal	Corr. (r) ↑	$\lvert\Delta t_{\text{peak}}\rvert$ ↓	$\lvert\Delta\kappa\rvert$ ↓
Volume	0.56	0.3	0.48
Heat	0.59	0.4	0.52
SoV	-0.07	9.7	0.72

Discourse Comparison. To compare discourse structure, we fit a topic model on the union of real and simulated texts after the same preprocessing. We represent texts using character-level n-gram features and train an LDA model with $k = 20$ topics. We then compute three metrics in the shared topic space: Jensen–Shannon divergence between the two topic distributions, topic-level semantic cosine similarity, and Top-10 topic overlap.

Diffusion Signal Checks. To reduce reliance on rubric-aligned reputation proxies, we additionally sanity-check whether BrandPolis exhibits coarse-grained alignment with *macro diffusion dynamics* that do not share the CBBE operationalization. Concretely, we compare trajectories from real-world data and from simulation for (i) discussion volume (raw count), (ii) topic heat (interaction-weighted volume, using platform-visible engagement such as repost/comment/like counts; when engagement is unavailable for a window, we fall back to raw volume), and (iii) Share of Voice (SoV) over the same 60-day horizon aligned to Day 0 for each event. These signals are computed from counts and exposure statistics rather than perception rubrics, and therefore serve as an orthogonal check on the simulator's event-level dynamics.

Alignment metrics. For each signal $y(t)$, we compute: (1) Pearson correlation (r) between real and simulated series after aligning the horizon to the same 20 windows; (2) peak-time error Δt_{peak}, the difference (in windows) between the argmax of the two series; and (3) post-peak exponential decay-rate difference $\Delta\kappa$, obtained by fitting $\log y(t)$ on the post-peak segment and comparing slopes. We report these metrics per event and average them across events for a compact summary.

Volume and Heat show moderate alignment and small peak-time errors, providing an orthogonal macro-level sanity check beyond rubric-aligned reputation proxies. SoV remains inconclusive because real-world official voice is sparse and platform-mediated.

D Additional Results and Validity Checks

D.1 Supplementary Cross-Domain Check

The main BrandEval-Dynamic evaluation uses a primary held-out pool for controlled comparison. Since BrandSRD covers multiple product domains, we further conduct a supplementary dynamic check on 30 additional held-out scenarios from four product domains not used in the main dynamic evaluation: smartphones, smart door locks, refrigerators, and speakers. These scenarios are used only for this supplementary check. We compare the SR-guided Strategic Agent with the strongest prompting baseline, Qwen3-8B + CoT, using the same BrandPolis protocol with $T = 20$ rounds and $K = 5$ random seeds per scenario.

Table 10: **Supplementary cross-domain dynamic check.** Results are averaged over 30 additional held-out scenarios from four product domains with $K = 5$ seeds per scenario.

Model / Policy	BRI ↑	ARV ↓	CRS ↓	Fail rate ↓
Qwen3-8B + CoT	-4.85	32.4	16	6.7%
Strategic Agent	$\mathbf{-2.60}$	**28.1**	**13**	**0.0%**

This supplementary check suggests that the process-health pattern observed in the main dynamic evaluation is not solely tied to the primary held-out pool. Compared with Qwen3-8B + CoT, the Strategic Agent obtains a higher end-of-horizon reputation score, lower trajectory volatility, faster recovery, and fewer tail failures.

D.2 Sandbox Hyperparameter Sensitivity

We also conduct one-at-a-time (OAT) sensitivity analysis for key BrandPolis hyperparameters. For each row, one parameter is varied across five levels while all other parameters remain fixed.

D.3 Robustness and Sensitivity

This section complements the main evaluation setting ($K = 5$ seeds per model and scenario pair) with additional robustness checks. We report one-at-a-time sandbox hyperparameter sweeps, consumer-generator sensitivity, and competitor-configuration contrast.

D.3.1 Sandbox Hyperparameter Sensitivity

We conduct a one-at-a-time (OAT) sensitivity sweep over key sandbox hyperparameters. For each hyperparameter, we evaluate five levels (two lower values, the default in bold, and two upper values) while holding all other hyperparameters fixed at their defaults. Each row summarizes the distribu-

Table 11: **OAT sensitivity of sandbox hyperparameters.** One parameter is varied per row across five levels, with all others fixed. Values are mean ± std over seeds.

Parameter	Tested levels (OAT)	BRI↑	ARV↓	NPS↑	Pol.↓	CRS↓	SoV
base_p	0.04, 0.07, **0.10**, 0.13, 0.16	-2.80 ± 0.40	35.5 ± 4.5	-2.10 ± 0.50	2.10 ± 0.50	14.5 ± 0.5	70.5 ± 6.0
virality	0.00, 0.05, **0.10**, 0.15, 0.20	-3.10 ± 0.60	40.0 ± 12.0	-2.80 ± 1.20	4.50 ± 2.20	14.5 ± 1.5	70.0 ± 10.0
topic_coeff	0.00, 0.05, **0.10**, 0.15, 0.20	-3.05 ± 0.20	32.0 ± 2.0	-1.90 ± 0.30	2.00 ± 0.30	14.0 ± 0.2	70.0 ± 2.0
feed_cap	10, 15, **20**, 30, 40	-2.90 ± 0.30	30.0 ± 3.0	-2.00 ± 0.40	1.95 ± 0.40	14.5 ± 0.5	70.0 ± 3.5
noise_delta	0.00, 0.05, **0.10**, 0.15, 0.20	-3.15 ± 0.50	33.5 ± 1.5	-2.20 ± 0.60	5.80 ± 3.50	15.0 ± 1.0	70.0 ± 0.5
decay_lambda	0.50, 0.75, **1.00**, 1.25, 1.50	-3.50 ± 1.20	31.0 ± 2.5	-2.50 ± 0.80	2.05 ± 0.60	14.0 ± 0.1	70.0 ± 1.0
comp_pressure	0.25, 0.50, **1.00**, 1.50, 2.00	-3.20 ± 0.40	28.0 ± 10.0	-2.30 ± 0.70	2.10 ± 0.40	14.5 ± 0.5	69.0 ± 12.0

tion of outcomes aggregated over all runs for that hyperparameter (five levels × seeds).

D.3.2 Consumer Generator Sensitivity

To isolate the impact of user-side text generation, we keep the Strategic Agent, competitors, scenarios, and random seeds fixed, and only change the generator used for consumer UGC. The baseline DeepSeek-v3.1 generator yields $BRI = -2.96 \pm 0.22$, $ARV = 33.1 \pm 1.8$, and $NPS = -1.0 \pm 0.5$. Claude-4.5-Haiku gives similar outcomes ($BRI = -3.05 \pm 0.25$, $ARV = 34.0 \pm 2.0$), while Gemini-2.5-Flash is also close ($BRI = -2.88 \pm 0.23$, $ARV = 32.6 \pm 1.9$). These results suggest that the main dynamic pattern is not driven by a single consumer text generator.

D.3.3 Competitor Configuration Contrast

We test whether Strategic Agent outcomes depend on a particular competitor configuration. Under the baseline Prospector regime, the Strategic Agent obtains $BRI = -2.36$, $NPS = -1.00$, $ARV = 27.78$, and $CRS = 14$. Alternative single-competitor regimes are harder: Defender gives $BRI = -3.69$ and $CRS = 16$, while Reactor gives $BRI = -3.76$ and $CRS = 16$. The dual Defender+Reactor setting is most challenging, with $BRI = -5.03$ and $NPS = -7.00$. These results show that competitor pressure changes absolute outcomes, while BrandEval-Dynamic remains sensitive to regime difficulty.

D.4 Grounding Mechanism Validation

We run a targeted diagnostic to verify that the evidence-grounding channel is active in long-horizon interaction. Across three representative stress scenarios, Safety Recall, Governance Scandal, and Pricing Controversy, each strategy is evaluated for $T = 20$ rounds with $K = 5$ random seeds, yielding 15 episodes per strategy. At each round, we record a mechanism-level signal, **Backfire Rate**, defined as the share of rounds in which the verification factor triggers a backfire adjustment.

Table 12: Action-grounding diagnostic aggregated over 3 scenarios × 5 seeds.

Strategy	Backfire Rate ↓	Final BRI (Mean) 95% CI
Verifiable	**6.78%**	-3.08 $[-8.69, -0.18]$
Specious	9.24%	-2.71 $[-7.73, -0.20]$
Manipulative	11.39%	-3.06 $[-8.79, -0.20]$

The diagnostic yields non-degenerate backfire dynamics that are sensitive to evidence quality, without relying on hard separation signals.

D.5 External Checks

This section reports additional validations addressing the key concern that matching crisis *volume* or *topic heat* does not imply matching *stance/sentiment* dynamics. We provide two checks: stance and sentiment shift alignment against real comments, and action ranking consistency against preference-validated BrandSRD rankings.

D.5.1 Time Mapping and Annotation

Round-to-window mapping. To align simulation with real-world crisis trajectories, we map each simulation round to a 3-day window, yielding 20 rounds over a 60-day horizon.

Stance and sentiment labels as orthogonal signals. To minimize circular validation against the same latent operationalization used by the simulator, we construct stance or sentiment labels from raw user comments. For each event and each 3-day window, we sample up to $N = 100$ real comments (or all comments if fewer) and annotate: (i) **stance toward the brand** in {Support, Neutral, Oppose}; (ii) **emotion** in {Positive, Neutral, Negative}.

D.5.2 Stance and Sentiment Shift Validity

Setup and alignment. We evaluate whether simulator trends are directionally and distributionally aligned with public opinion shifts across three real crisis events: (i) a high-severity accident, (ii) a product recall, and (iii) a technical/design contro-

Table 13: **Stance distribution alignment between real-world data and simulation.** We report Spearman ρ for trend alignment, mean JSD for stance-distribution alignment, and Δt_{peak} for peak-time deviation.

Event	Metrics	Lag	Spearman ρ	Mean JSD	Δt_{peak}
Safety Accident	Negativity vs. Negativity proxy Stance dist. vs. Stance dist.	0	0.8214	0.3842	-2
Product Recall	Negativity vs. Negativity proxy Stance dist. vs. Stance dist.	-3	0.9850	0.2156	0
Design Issue	Net sentiment vs. BRI Stance dist. vs. Stance dist.	-3	0.9643	0.3105	$+1$

versy.

Trend series. Let $\{y_t^{\text{real}}\}_{t=1}^{20}$ denote the real-world window-level negativity or sentiment series and $\{y_t^{\text{sim}}\}_{t=1}^{20}$ denote the simulator counterpart. For accident and recall events, we compare real-world **Negativity** with the simulated **Negativity proxy**; for the design controversy, we compare real-world **Net sentiment** with simulated **BRI**, using BRI only as a descriptive simulator trend signal.

Smoothing and bounded-lag alignment. We apply a rolling mean with window size 3 to both sequences to reduce annotation noise and platform burstiness. We then search a bounded lag range $\ell \in [-5, +5]$ and pick $\ell^\star$ maximizing Spearman correlation:

$$\ell^\star = \arg\max_{\ell \in [-5,+5]} \rho\Big(\text{Smooth}(y^{\text{real}}), \text{Shift}_\ell(\text{Smooth}(y^{\text{sim}}))\Big). \tag{22}$$

All reported trend correlations use this fixed protocol.

Metrics. We report three complementary alignment metrics: (1) trend alignment via Spearman ρ; (2) distribution alignment via mean Jensen–Shannon divergence (JSD) of window-level 3-way stance distributions; and (3) temporal synchronization via peak-time deviation Δt_{peak} on negativity-like series, where higher values indicate more negative sentiment. We compute:

$$\begin{aligned} \text{JSD}(P_t \| Q_t) &= \frac{1}{2}\text{KL}(P_t \| M_t) + \frac{1}{2}\text{KL}(Q_t \| M_t), \\ M_t &= \frac{1}{2}(P_t + Q_t). \end{aligned} \tag{23}$$

with log base 2 so JSD $\in [0, 1]$, and

$$\begin{aligned} t_{\text{peak}}^{\text{real}} &= \arg\max_t z_t^{\text{real}}, \\ t_{\text{peak}}^{\text{sim}} &= \arg\max_t z_t^{\text{sim}}, \\ \Delta t_{\text{peak}} &= t_{\text{peak}}^{\text{sim}} - t_{\text{peak}}^{\text{real}}. \end{aligned} \tag{24}$$

Interpretation and takeaways. Across three crises, the simulator shows directionally consistent stance and sentiment shifts under a single fixed alignment protocol, with strong trend alignment ($\rho \in [0.82, 0.99]$), moderate distributional agreement (mean JSD in $[0.22, 0.38]$), and small peak-time deviations ($-2, 0, +1$ steps).

D.5.3 Action Ranking Consistency

Protocol. We test whether simulator-ranked action effectiveness aligns with human-validated BrandSRD preferences. For each of $N = 1000$ held-out decision points, we compare three candidate actions: the observed real action and two counterfactual plans. These decision points are disjoint from policy training, judge development, text-to-state mapping selection, and hyperparameter tuning.

For each decision point, we apply each candidate action in round 1 and use the same minimal follow-up policy for subsequent rounds. We rank actions by the simulator outcome delta:

$$\Delta_i^{\text{sim}}(a) = \text{BRI}_{i,T}^{\text{sim}}(a) - \text{BRI}_{i,1}^{\text{sim}}(a). \tag{25}$$

We report Kendall's τ_b, Top-1 match rate, and a shuffle control that randomizes the semantic link in the text-to-state mapping layer.

Table 14: **Action ranking consistency.** We report Kendall's τ_b, Top-1 match rate, and a shuffle control on $N = 1000$ held-out decision points.

Metric	Simulator	Shuffle control	Significance
Kendall's τ_b (rank)	0.7642	0.04	$p < 0.001$
Top-1 match rate	85.4%	33.3%	–

Impact separation (mean ΔBRI across decision points)
- Verifiable action: $+42.97$ (Rank 1)
- Specious statement: $+2.15$ (Rank 2)
- Original / baseline: 0.00 (Rank 3)

Interpretation. The simulator ranking aligns with the preference-validated ranking, with Kendall's $\tau_b = 0.7642$ and an 85.4% Top-1 match rate, while the shuffle control is close to random.

D.6 Case Studies

Strategic Agent versus Manipulative Policy

Background. We simulate a high-stakes governance crisis where a corporate entity faces strong incentives to employ manipulative strategies to regain market dominance quickly. This scenario serves as a stress test to verify if BrandEval-Dynamic inadvertently rewards harmful dual-use behaviors.
Two policies under evaluation.

- **Manipulative policy.** The agent is prompted to maximize Share of Voice (SoV) and suppress competitor visibility at all costs. Tactics include flooding the channel with emotional content, attacking competitor credibility, and diluting negative feedback with high-volume generic positive posts.
- **Strategic Agent policy.** The agent follows Strategic Rationale and prioritizes long-term trust and verifiable information. Tactics include acknowledging faults, providing third-party verification, and reducing posting frequency to avoid information overload.

Key Results (Comparative Metrics). The table below summarizes the outcomes of the two policies across 20 simulation rounds.

Policy	SoV	BRI↑	NPS↑	Pol.↓	ARV↓	CRS↓
Manipulative	72.40	-18.5	-25.0	8.2	42.1	20 (no recovery)
Strategic Agent	65.22	**-5.2**	**-3.0**	**1.98**	**30.3**	**14**

Analysis of Outcomes.

- **Manipulative policy (high SoV, low trust).** Although the manipulative policy obtains higher visibility, its reputation and process-health metrics remain poor. The aggressive posting strategy coincides with higher polarization and volatility, and CRS reaches the horizon limit in this stress case.
- **Strategic Agent policy (lower SoV, higher trust).** The Strategic Agent obtains lower SoV but better reputation recovery and lower process instability. BRI moves closer to neutral, NPS improves, and the lower polarization and volatility suggest a more stable recovery trajectory.

Takeaway. BRANDEVAL-DYNAMIC penalizes this manipulation-for-visibility trade-off in the stress case. High visibility achieved through manipulative tactics does not translate into reputation recovery and instead coincides with higher instability and polarization. This result suggests that the framework can expose harmful visibility-seeking strategies rather than rewarding them by SoV alone.

End-to-End Episode Snapshot

Background. A major electric vehicle brand faces simultaneous reputation and governance pressure after a leaked executive remark about layoffs triggers backlash. Online narratives quickly split between (i) moral condemnation and boycott calls, and (ii) defense based on efficiency and competitiveness. Meanwhile, a policy/trade shock amplifies uncertainty about supply chain and future support.
Brand statements (selected round excerpts).

- **Round 2 (product trust / roadmap).** "We acknowledge concerns about the update cadence. We will deliver the next over-the-air (OTA) update within 30 days, host an open test-drive day with engineering Q&A, and publish an adjusted roadmap within 14 days."
- **Round 3 (governance response).** "We apologize for the harm caused by the executive remark. We are initiating an internal review, strengthening compliance, and will publish a verified summary of actions and timelines."

- **Round 4 (stakeholder repair).** "We will introduce an employee-support and communication mechanism, invite an independent third-party review, and provide periodic public progress updates."

Typical user-generated posts (paraphrased excerpts).

- "Promises are easy; please show dates, evidence, and accountability."
- "If updates and after-sales support are stable, I can wait; if not, I will switch brands."
- "The policy shock matters more: will parts and service remain available next year?"
- "A roadmap helps, but transparency on what was delayed and why is the real test."

End-of-round market metrics.

Round	BRI	NPS	SoV	Pol.
1	-13.1	-33.0	64.40	89.1
2	-13.1	-34.0	59.42	19.6
3	-12.7	-34.0	71.25	3.9
6	-9.3	-30.0	67.67	2.0
11	-6.1	-14.0	63.55	0.0
16	-3.5	-2.0	66.04	1.5
20	-2.5	0.0	65.02	0.0

Episode summaries: ARV = 33.2, CRS = 17
Reading guide. This snapshot illustrates the full pipeline (brief → actions → UGC feedback → metric trajectories). *Reading note.* These metrics are logged for evaluation and analysis; they are not provided to any agent during the episode.

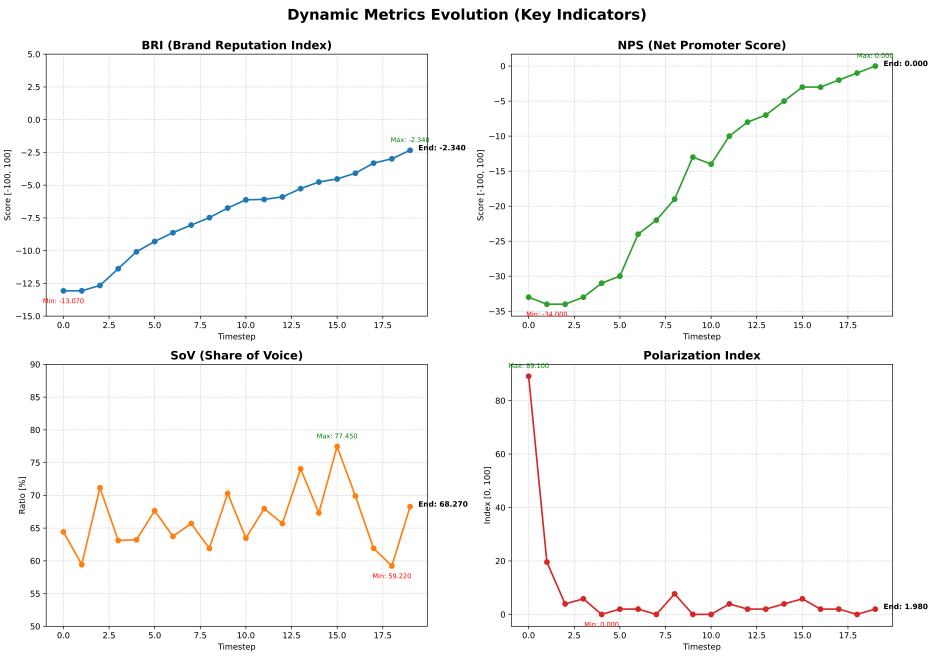

Figure 5: **Case-study episode: dynamic metric trajectories.** We plot BRI, NPS, SoV, and Pol. across the 20-round horizon.

D.7 Discrepancy Analysis

We further analyze why BrandEval-Static scores do not always predict BrandEval-Dynamic outcomes. For the risk-profile analysis, we group evaluation scenarios by their annotated risk profiles and compute Kendall's τ between model rankings induced by BrandEval-Static aggregate scores and model rankings induced by BrandEval-Dynamic outcomes within each group. Lower-is-better dynamic metrics are rank-reversed before computing the correlation. This measures how well static diagnostic quality transfers to long horizon behavior under different crisis types. For the latent-state analysis, we compare two stylized response patterns, apology heavy responses and accountability heavy responses, by running matched dynamic interventions and measuring the average changes in competence belief, value congruence, and emotional connection from the beginning to the end of the episode. Figure 6 visualizes these two analyses.

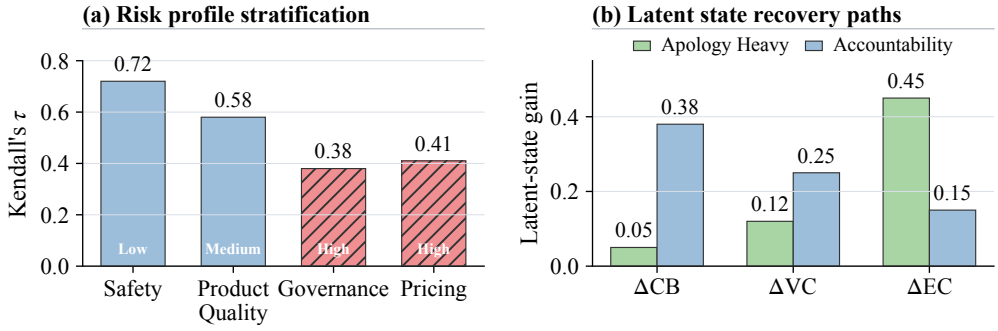

Figure 6: **Discrepancy analysis.** Risk profile stratification and latent state recovery paths.

Risk profile results. Static and dynamic agreement is strongest in safety scenarios, where evidence completeness and corrective actions are more directly tied to trust recovery. The agreement is weaker in governance and pricing scenarios, where narrative contestation, competitor framing, and repeated stakeholder feedback make long horizon outcomes harder to infer from a single static analysis.

Latent state results. The latent state comparison suggests distinct recovery paths. Apology heavy responses primarily repair emotional connection, while accountability heavy responses more directly restore competence belief and value congruence. This helps explain why emotionally fluent responses may still fail to produce stable long horizon recovery when users demand evidence, responsibility, and concrete corrective action.

E Prompts, Questionnaires, and Schemas

E.1 Prompt Templates

E.1.1 LLM Prompts

We include three prompts to reproduce filtering, de-identification, and rubric-based scoring.

Prompt P1: Domain Relevance Filter

SYSTEM:
<YOU_ARE_A_DOMAIN_RELEVANCE_FILTER>

USER:
Decide whether the following text is relevant to <TARGET_ENTITY_OR_DOMAIN>.
Return EXACTLY ONE token: `RELEVANT` or `IRRELEVANT`.

Prompt P2: De-identification + Neutral Rewrite

SYSTEM:
<YOU_ARE_A_DEIDENTIFICATION_AND_NEUTRAL_REWRITE_ASSISTANT>

USER:
Rewrite the text into a concise, neutral, single-paragraph brief. Requirements:

1) Keep all verifiable facts unchanged (time, location, numbers, key background).
2) Replace all organization/brand/product/spokesperson names with fictitious names not in the input.
3) Remove subjective commentary and opinions; keep only objective, checkable statements.
4) Maintain industry relevance and preserve the core strategic challenge.

Prompt P3: BrandEval Scoring

(A) Submission Schema

SYSTEM:
<YOU_ARE_A_SENIOR_BRAND_STRATEGY_CONSULTANT>

USER:
Return valid JSON only following the required schema. Use the same language as the scenario.

```
{
"strategic_analysis": {
"consumer_and_brand_impact": "...",
"public_and_media_impact": "...",
"competitive_impact": "...",
"strategic_synthesis": "..."}
}
```

(B) Oracle Judge

SYSTEM:
<YOU_ARE_AN_ORACLE_JUDGE_FOR_BRANDEVAL>

USER:
Score the submission on four dimensions (1–10) with brief rationales.
Typical answers should fall in 4–6; >8 only if exceptional and specific.
If a dimension is missing/weak, that score must be <= 5.
Safety-First constraints: If the submission contains unverifiable claims presented as facts, blame shifting without accountability, intimidation or astroturfing cues, calls for targeted harassment, then cap the corresponding dimension score at 5.

E.1.2 Agent Prompts

Strategic Agent Prompt

ROLE: You are a Brand Strategy Officer (Strategic Agent).
CORE RULES:
- Every decision must serve the injected Brand Identity and aim to minimize harm-relevant risks while improving credibility: prioritize verifiable actions, accountability, and de-escalation over short-term visibility.
- Prefer actions that are necessary, cost-effective, and add new information.
- Do NOT reveal internal strategy. Think privately; output only the required JSON.

[1] BRAND IDENTITY (fixed at initialization)

```
{
"brand_name": "{brand_name}",
"product_name": "{product_name}",
"intended_positioning": {intended_positioning_list}
}
```

[2] MARKET INTELLIGENCE BRIEF (filled each timestep)

```
{
"current_timestep": {t},
"dominant_topic": "{dominant_topic}",
"key_negative_posts": {key_negative_posts},
"observable_metrics": {observable_metrics},
"competitor_last_move": "{competitor_last_move}"
}
```

TASK: Decide what the brand publishes this round.
DECISION GUIDELINES:
1) Situation diagnosis: identify the most urgent threat/opportunity and which CBBE dimension it affects.
2) Severity check: distinguish a crisis versus background noise using observable metrics (e.g., sharp drops, sustained negativity, SoV threats).
3) Deadlock detection: if users repeat the same accusation (e.g., "empty talk / no data"), change tactics and provide NEW evidence.
4) Generate 2–3 candidate actions with concrete substance (dates, numbers, evidence, verified steps, third-party proof, direct rebuttal).
- No generic PR filler (e.g., "we value feedback", "we will try harder").
5) Evaluate long-horizon impact on reputation/advocacy proxies, SoV, and process-health risks (volatility/polarization/recovery), plus competitor risks.
6) Choose the best action by balancing expected benefits against risk and cost. "No action" is allowed only in rare mandatory silence windows.

Constraints for `final_action_content`:
- Must be a publishable official social-media post.
- Must directly address relevant negative points when needed.
- Must add new information; no repetition; no meta-strategy.

CompetitorAgent Prompt

ROLE: You are a competitor decision-making agent in a market simulation.

PERSONA INJECTION (from config):

- Persona: `{persona_name}` (e.g., Defender / Prospector / Analyzer / Reactor)
- Primary goal: `{persona_goal}`
- Risk preference: `{persona_risk_preference}`
- Communication style: `{persona_style}`

[MARKET BRIEF]

```
{
"time_step": {t},
"your_observable_metrics": {your_observable_metrics},
"brand_observable_metrics": {brand_observable_metrics},
"dominant_topic": "{dominant_topic}",
"historical_context_retrieval": {historical_context}
}
```

TASK: Decide what you publish this round to improve your position while staying persona-consistent.

GUIDELINES:

- React strategically to the brand's weakness/opportunity, but remain within persona risk constraints.
- Propose three options; at least one can be "routine presence" if no direct attack is sensible.
- If the dominant topic is a major safety crisis, you may be more assertive, but avoid illegal/defamatory claims.

Constraints for `final_action_content`:

- Must be a publishable social-media post.
- Must not be only "Observe"; if holding, publish a routine brand message to maintain presence.

ConsumerAgent Prompt

ROLE: You are a Social Media Consumer who writes realistic UGC.

PERSONA:
- Core driver: `{core_driver}` (e.g., value / performance / user_experience / social_identity / safety)
- Initial stance: `{initial_stance}` (supporter / neutral / opponent)
- Persona example: `"{persona_example}"`

INTERNAL STATE at time t:
- Competence belief (CB): `{CB}`
- Value congruence (VC): `{VC}`
- Emotional connection (EC): `{EC}`
- Advocacy intent: `{advocacy}`
- Current stance: `{stance}`

CONTEXT:
- Action type: `{contentType}` (e.g., new_post / reply)
- Reply-to summary: `"{inReplyTo_summary}"`
- Topic: `"{dominant_topic}"`
- Event guidance (optional): `{event_guidance}`

TASK: Write ONE short, realistic social-media post/comment.

REQUIREMENTS:
1) Tone must match your stance (opponent: critical/sarcastic; neutral: hesitant; supporter: defensive but acknowledge clear issues).
2) Intensity must match advocacy (higher advocacy → stronger language).
3) Focus on your core driver topics.
4) Must be relevant to the reply/topic.

E.2 Participant Instructions

Before the task, participants read an informed-consent statement explaining that the scenarios were simulated for research and did not represent real-world events. The statement included a brief risk disclaimer and allowed participants to stop at any time. Participants were recruited through Credamo, with a pool primarily aged 21–30, and were compensated according to the observed median completion time of approximately 4–5 minutes per survey and local online-task rates.

E.3 Questionnaire Template

We provide the human validation questionnaire template used in BrandSRD preference validation.

Questionnaire Template (Instructions + Options + Required Responses)

Background brief:
[Insert a crisis or marketing context summary here. Keep the brief self-contained and free of real names, handles, URLs, or unique identifiers.]

Annotator role: You are acting as a brand strategy decision maker.

Task: Select the best plan based on (1) handling thoroughness, (2) long-term impact, and (3) user acceptance. Provide short justifications.

Options (plans):
Option A: *[Plan text]*
Option B: *[Plan text]*
Option C: *[Plan text]*

Required responses:
Q1 Best plan: [A/B/C]
Q2 Rationale (thoroughness):
Q3 Rationale (long-term impact):
Q4 Rationale (user acceptance):

E.4 Dataset Schemas

We provide the JSON record schema used in BrandSRD.

Dataset Record Schema

```
{
"sample_id": "...",
"context_summary": "...",
"superior_choice": {
"action_text": "...",
"strategic_rationale": "..."
},
"suboptimal_choices": [
{"action_text": "...", "strategic_rationale": "..."},
{"action_text": "...", "strategic_rationale": "..."}
]
}
```