systems

MDPI

Article

SAPIENT: A Multi-Agent Framework for Corporate Reputation Intelligence Through Sentinel Monitoring and LLM-Based Synthetic Population Simulation

Alper Ozpinar *,† and Saha Baygul Ozpinar †

School of Communication, Ibn Haldun University, Basak Mah. Ordu Cad. No:3 P.K., 34480 Basaksehir, İstanbul, Turkey; saha.ozpinar@ihu.edu.tr

* Correspondence: alper.ozpinar@ihu.edu.tr

† These authors contributed equally to this work.

Abstract

Corporate reputation teams rely on media monitoring and qualitative research, both limited in speed and coverage when digital narratives form rapidly. This paper proposes SAPIENT (Sentinel-Augmented Population Intelligence for Emerging Narrative Tracking), a multi-agent system that links a sentinel layer over public text streams with a simulation layer that runs moderated, repeatable in silico focus-group sessions. The sentinel layer ingests social media, news, and forum text to produce a compact signal state (topics, sentiment, anomaly scores, risk labels), which conditions the simulation layer through an orchestrator. Persona agents and a moderator follow an Agentic Focus Group (AFG) protocol with repeated runs, variance reporting, and human review gates. We describe four sustainability communication scenarios: greenwashing backlash prediction, greenhushing risk assessment, campaign pre-testing, and crisis communication simulation. Nine experiments span 280 AFG runs across 20 conditions, three LLM backends (Claude Sonnet 4, GPT-4o, and Gemini 2.5 Flash), and a preregistered pilot human validation study with 54 participants. Signal conditioning improved simulation specificity ($p = 0.012$). Cross-lingual sessions revealed a sentiment asymmetry between English and Turkish ($p = 0.001$) with preserved persona rank ordering ($r = 0.81$, $p = 0.015$). Cross-model comparison showed consistent persona differentiation across all three backends (Pearson $r > 0.92$, $p < 0.002$ for all pairs). Sentiment was robust to prompt paraphrasing ($p = 0.061$, n.s.), though credibility was sensitive to prompt wording ($p < 0.001$). All significant results from Experiments 1–8 survived Benjamini–Hochberg correction. A preregistered pilot with 54 human participants on Prolific replicated the predicted credibility ranking across framing variants ($p = 0.004$) but not the sentiment ranking, identifying a specific calibration target for future work.

Keywords: agentic AI; sentinel monitoring; synthetic population; agentic focus group; in silico focus groups; corporate reputation; greenwashing; multi-agent systems; large language models; cross-model comparison

Academic Editor: Fernando De la Prieta Pintado

Received: 21 February 2026
Revised: 5 April 2026
Accepted: 7 April 2026
Published: 10 April 2026

1. Introduction

The relationship between organizations and their stakeholders has been reshaped by the architecture of digital media. A single social media post can set off a cascade of public scrutiny within minutes, and the narratives that emerge from such events often prove more consequential than the events themselves [1]. For companies operating in sectors with high environmental exposure, such as energy, manufacturing, fashion, and finance, this situation

is compounded by growing societal attention to sustainability claims. Accusations of greenwashing, where firms are perceived as overstating their environmental credentials, have led to regulatory fines exceeding tens of millions of dollars in recent years [2]. The inverse pattern, termed greenhushing , describes organizations that deliberately suppress disclosure of legitimate sustainability efforts because they fear that any public claim will attract disproportionate criticism [3]. Both phenomena pose distinct reputational risks, and both call for detection and response mechanisms that operate at the speed of online discourse.

Existing approaches to reputation management fall broadly into two categories, each with well-documented limitations. The first category, social listening and media monitoring platforms, works well for retrospective analysis but typically identifies threats only after they have already gained momentum. These systems detect volume spikes and sentiment shifts, yet they lack the capacity to anticipate *how* a narrative might evolve or *which* stakeholder segments are most likely to amplify it [4]. The second category, qualitative research methods such as focus groups and surveys, can probe audience reactions with nuance but operates on timescales measured in weeks or months, a pace that is fundamentally misaligned with the 24 h news cycle [5]. Focus groups carry additional constraints: small sample sizes introduce moderator effects, recruitment bias limits the range of perspectives, and the logistics of scheduling and facilitation make it difficult to iterate quickly on time-sensitive issues [6,7]. Online variants increase reach but still require recruitment and cannot operate at the speed of an unfolding crisis.

Recent advances in two separate research streams suggest a path toward bridging these limitations. The first stream draws on sentinel surveillance concepts, originally developed in epidemiology for early outbreak detection [8–10], and applies them to digital media monitoring. In this framing, online platforms serve as distributed sensors that emit weak signals, such as anomalous posting volumes, shifts in topic co-occurrence, or the emergence of new actors, well before a narrative reaches mainstream visibility [11,12]. The technical machinery for this kind of monitoring, including time-series anomaly detection, topic modeling, and named-entity tracking, has matured considerably, and draws on Topic Detection and Tracking (TDT) research that formalized the event-based organization of news streams [13]. Its application to corporate reputation contexts, on the other hand, remains fragmented.

The second stream involves using large language models (LLMs) as engines for simulating human behavior at scale. Park et al. [14] showed that LLM-driven generative agents can produce believable social behaviors in virtual environments, and follow-up work has scaled such simulations to populations exceeding one million agents. Argyle et al. [15] found that LLM-generated synthetic samples can approximate the distribution of political opinions across demographic subgroups with reasonable fidelity, while Horton [16] explored the concept of *homo silicus*, synthetic economic agents whose behavioral patterns resemble those of human subjects in controlled experiments. These results have prompted growing interest in using LLM-based "synthetic populations" as a rapid supplement to traditional audience research [17]. At the same time, the literature reports clear limits: simulated responses can match some group-level patterns in specific settings, but models can miss distributional diversity, reproduce stereotypes, exhibit over-agreeableness, and generate overconfident errors [18]. These gaps make it unsafe to treat simulated groups as a drop-in replacement for human subjects.

Despite this progress, the two streams have developed in near-complete isolation from one another. Sentinel monitoring research has focused on detection without addressing the question of subsequent narrative development, that is, how identified signals might propagate through different audience segments. Synthetic population research has focused on simulation fidelity without grounding the simulation inputs in real-time, real-world

data streams. A review of 500 papers by the authors, which screened down to 23 studies meeting strict inclusion criteria across sentinel systems, agentic AI, and media applications, did not identify prior work that integrates real-time sentinel monitoring with LLM-based synthetic population simulation within a unified architecture for corporate reputation management. This gap is both conceptual and practical: practitioners lack a system that connects early signal detection to anticipatory audience modeling.

We ask a specific question: *How can we structure an agentic system so that it (a) improves monitoring and sensemaking over media streams, (b) supports early qualitative exploration with in silico groups, and (c) surfaces potential errors for human review while keeping a decision-maker in control?*

This study addresses this question by proposing SAPIENT (**S**entinel-**A**ugmented **P**opulation **I**ntelligence for **E**merging **N**arrative **T**racking),

a multi-agent system that couples sentinel-based media monitoring with LLM-driven synthetic population simulation through an orchestrator agent. The system is designed for corporate reputation intelligence, with an emphasis on sustainability communication scenarios including greenwashing detection, greenhushing assessment, campaign pre-testing, and crisis communication simulation.

The contributions of this paper are as follows:

- SAPIENT couples a continuous sentinel monitoring layer with LLM-based synthetic population simulation through a formalized, versioned signal state S_t. Adjacent platforms have explored individual components of this pipeline [19–22]. SAPIENT's contribution is the specific combination of (i) a compact signal state bridging real-time monitoring and simulation, (ii) a repeated AFG protocol with built-in variance controls, and (iii) bidirectional coupling where simulation insights update sentinel watchlists.
- We introduce the Agentic Focus Group (AFG) protocol, a repeatable simulation procedure with moderator-led sessions, multiple runs, variance reporting, and human review gates, that treats simulation as an experimental instrument rather than a prediction oracle.
- We formalize persona construction (Equation (2)) with explicit sampling strategies, variance collapse countermeasures, and multilingual handling, and specify calibration metrics (Equation (3)) with hooks for prediction-powered inference integration.
- We situate the simulation layer within the rapidly growing calibration literature [23–25], identifying explicit validity boundaries and integration paths for distributional techniques.
- We provide empirical evidence through nine experiments spanning three application scenarios, three LLM backends (Claude Sonnet 4, GPT-4o, and Gemini 2.5 Flash), a prompt sensitivity analysis, and a preregistered pilot human validation study (Experiment 9, $n = 54$, Prolific). The experiments total 280 independent AFG runs across 20 conditions plus 54 human participant responses. All significant results from Experiments 1–8 survive Benjamini–Hochberg correction.
- Adding two LLM backends and two new scenarios required only configuration-level changes, with no modifications to the core AFG protocol or analysis pipeline.
- In Experiment 9, human participants rated the same three framing variants used in Experiment 1, enabling direct comparison between SAPIENT outputs and real human responses on credibility, sentiment, and thematic content.
- We articulate governance mechanisms including security defenses against prompt injection, red-team protocols, prohibited uses, and explicit non-use contexts.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature across seven research areas. Section 3 presents the SAPIENT architecture, the signal state formalization, and the AFG protocol. Section 4 describes the four application

scenarios. Section 5 outlines the evaluation framework and states the experimental hypotheses. Section 6 reports empirical results from nine experiments, including four from the original evaluation plan, four added during revision to address scenario generalizability, cross-model comparison, and prompt sensitivity, and a preregistered pilot human validation study. Section 7 discusses theoretical and practical implications, positioning relative to calibration literature, limitations, and ethical considerations. Section 8 concludes with directions for future research.

2. Related Work

This section reviews seven bodies of literature that collectively inform the design of SAPIENT.

2.1. Literature Identification and Selection

The literature survey followed a structured screening process. An initial pool of approximately 500 candidate papers was assembled through database searches (Scopus, Web of Science, IEEE Xplore, and Google Scholar) and backward/forward citation tracking, using query combinations across three domains: sentinel surveillance and social media monitoring, agentic AI and multi-agent simulation, and corporate reputation and crisis communication. Title and abstract screening retained 87 papers that addressed at least one of the three domains with a system-level or empirical contribution. Of these, 38 were assessed at full-text level against five inclusion criteria, yielding 23 core studies that directly informed the architectural design:

1. Addresses sentinel/monitoring *or* synthetic population simulation at the system level;
2. Published 2019–2025, or a seminal older work with established influence;
3. Reports system architecture, empirical evaluation, or formal methodology, not only conceptual discussion;
4. English or Turkish full text available;
5. Peer-reviewed publication, or high-impact preprint from a leading research group (>50 citations or affiliated with a recognized lab).

The broader reference list (46 sources) extends beyond these 23 core studies to include foundational works, methodological references, and calibration literature. The agentic AI and LLM simulation field is evolving at an exceptional pace; new foundation models, benchmarks, and architectural patterns appear on a near-weekly basis, and the peer-review cycle has not kept pace with the rate of innovation. Consequently, a subset of the cited works (7 of 46 references) remain available only as preprints from established research groups. These preprints are used with appropriate hedging and are not relied upon for the paper's central architectural claims; where peer-reviewed versions have become available since the original submission, references have been updated accordingly. Figure 1 summarizes the screening process.

2.2. Sentinel Surveillance and Social Media Early Warning

The concept of sentinel surveillance originated in public health, where designated monitoring points collect high-quality data from a subset of the population to generate early warning signals for the broader community [26]. Velasco et al. [8] carried out a systematic review of internet-based systems for public health surveillance and found that social media data could complement traditional reporting by providing earlier signals, sometimes days or weeks ahead of official notifications. Charles-Smith et al. [9] confirmed this potential in a parallel review, though they stressed that raw social media data carry substantial noise and that effective systems require multi-stage filtering pipelines. Eysenbach [27] provided an early treatment of monitoring online information patterns in near real time, coining the

term *infoveillance*. Sakaki et al. [28] showed that even simple classifiers applied to Twitter streams can detect earthquakes faster than official seismological services, establishing social media as a viable early-warning data source.

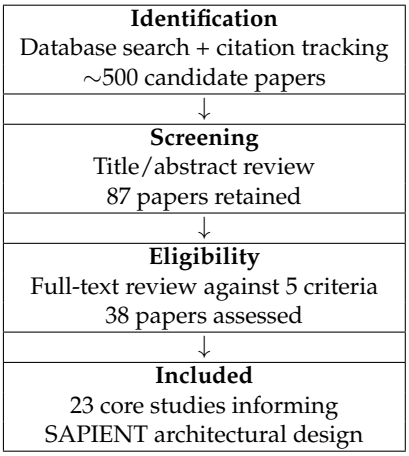

Figure 1. Literature screening process following PRISMA-inspired methodology. The 23 core studies span sentinel surveillance, agentic AI simulation, and corporate reputation management. The broader reference list (46 sources) includes foundational, methodological, and calibration works beyond the core set.

George et al. [11] proposed spatio-temporal event detection methods for geotagged social media, while Hammond et al. [10] showed that alternative data sources can complement traditional sentinel networks. MacIntyre et al. [12] identified a recurring pattern: detection accuracy improves substantially when multiple signal types are combined, but systems relying on any single signal type produce unacceptable false alarm rates. The Topic Detection and Tracking (TDT) research program [13] formalized event-based organization of news streams; modern pipelines extend TDT with embedding-based retrieval and LLM summaries, introducing new risks when the input stream contains adversarial instructions.

For corporate reputation contexts, Swaminathan et al. [4] showed that topic modeling on social media brand conversations can reveal perception shifts that precede measurable changes in brand equity. A gap remains, however. Existing sentinel-style monitoring for brands focuses on detection; it does not address how detected signals might propagate through specific audience segments.

2.2.1. Adjacent Multi-Agent Simulation Platforms

Several recent platforms address components of the monitoring-to-simulation pipeline targeted by SAPIENT. GenSim [20] proposes a scalable system for generating and managing large populations of LLM-based social agents, showing that careful persona sampling and interaction scheduling can maintain behavioral diversity at scales exceeding 1000 agents. GenSim does not incorporate a real-time monitoring front-end, though; its simulations are triggered by static scenarios rather than live data streams. RumorSphere [21] specifically targets information propagation in crisis contexts using networked LLM agents with a social-graph structure, reporting cost and variance scaling laws that inform practical deployment. Its emphasis on network topology and cascade dynamics complements SAPIENT's focus-group-style deliberation, though the two architectures serve different analytic purposes: RumorSphere models diffusion mechanics, whereas SAPIENT's AFG protocol elicits attitudinal and interpretive responses.

DualMind [22] introduces a dual-process cognitive architecture for individual agents, distinguishing fast heuristic responses from slow deliberative reasoning. While this per-agent sophistication exceeds SAPIENT's current persona model, DualMind does not

address the upstream monitoring-to-simulation coupling that represents SAPIENT's primary contribution.

2.2.2. Crisis Simulation Benchmarks

Crisis-Bench [19] provides a POMDP-based evaluation system with outcome-oriented scoring for LLM-generated crisis-management strategies, including metrics for over-transparency and strategic ambiguity. SAPIENT's AFG outputs, which generate stakeholder reaction hypotheses rather than prescriptive strategies, could in principle be evaluated within a Crisis-Bench loop to quantify whether signal-conditioned simulations yield operationally superior guidance. We identify this as a priority for future empirical work (Section 8).

2.2.3. LLM-ABM Fragility and Prompt Sensitivity

A growing body of evidence shows that LLM-based agent simulations are sensitive to seemingly inconsequential prompt variations, including whitespace, synonym substitution, and output-format changes, that can shift aggregate outcomes by magnitudes comparable to the experimental effects being studied [25]. These findings point to a fundamental challenge for any LLM-based simulation platform, including SAPIENT. Multi-run replication mitigates stochastic variance but does not address systematic prompt-induced bias. SAPIENT's current design partially addresses this through temperature stratification and contrarian injection. A dedicated prompt-sensitivity analysis, varying instruction phrasing, persona description format, and structured-output templates, is needed to establish how wide the tolerance bounds actually are. We report an initial sensitivity analysis in Section 6.12 and identify full ablation as future work.

2.3. Crisis Communication, Agenda Setting, and Misinformation

McCombs and Shaw [29] established that media salience directly influences public salience. Entman [30] defined framing as the selection and highlighting of particular facets of events to promote a specific interpretation. Crisis communication research, particularly Coombs' Situational Crisis Communication Theory [1], studies how response strategies interact with stakeholder attributions and reputational threat. Large-scale evidence on Twitter shows that false news diffuses farther, faster, and more broadly than true news [31], which raises the value of early detection and careful response planning. In the corporate reputation space, this asymmetry means that a misleading greenwashing accusation can gain traction more quickly than a factual rebuttal.

2.4. Focus Groups and Qualitative Research Constraints

Focus groups are a common method for exploring interpretations, attitudes, and language use in a target segment. Method texts describe both strengths, including rich interaction, idea generation, and access to group dynamics, and limits: small samples (typically 6–12 participants), moderator effects, recruitment bias, and high costs in time and money [6,7]. Online variants increase reach but still require recruitment, scheduling, and facilitation. For fast-moving issues, teams often want earlier exploratory signals before committing to full fieldwork. These constraints motivate the exploration of in silico alternatives, not as replacements, but as a rapid, low-cost complement for early hypothesis generation.

2.5. Agent-Based Models and Opinion Dynamics

Agent-based modeling (ABM) has a long history in computational social science for studying opinion formation, information diffusion, and collective behavior. Bounded-confidence models such as those of Deffuant et al. [32] describe how interaction can lead to

opinion clustering, while reviews of social influence models [33] discuss how micro-level rules yield macro-level polarization patterns. Classical ABMs scale well but are semantically impoverished: agents do not produce language or interpret framing. Augmenting ABMs with LLM capabilities could address this limitation, though it introduces validation challenges that remain open.

2.6. LLM-Based Generative Agents and Synthetic Respondents

Park et al. [14] showed that LLM-driven agents with memory and reflection produce believable social behavior, and subsequent work has scaled such simulations to populations exceeding one million agents.

Argyle et al. [15] found that LLM-generated synthetic samples can approximate opinion distributions across demographic subgroups. Horton [16] reported substantial agreement between LLM agent behaviors and economic theories. Hämäläinen et al. [17] concluded that synthetic responses capture broad trends but fail to reproduce distribution tails, with outputs skewing toward WEIRD populations [18]. Gao et al. [34] characterized the state as "promising but immature." These limitations directly inform the AFG protocol's reliance on repeated runs and variance reporting.

2.7. Calibration, Distributional Alignment, and Validity Conditions

A parallel and rapidly growing volume of literature examines under what conditions LLM-generated responses can serve as valid proxies for human opinion data, a question directly relevant to the simulation layer of SAPIENT.

Distributional alignment. Santurkar et al. [35] showed that base LLM opinion distributions skew toward particular demographic profiles, motivating alignment techniques. Suh et al. [24] introduced SubPOP, a large-scale dataset of 3,362 questions and 70K subpopulation-response pairs, and found that fine-tuning on survey data reduces the LLM-to-human distributional gap by up to 46% as measured by the Wasserstein distance. These results suggest that prompt engineering alone, the approach SAPIENT currently relies on, has limited calibration power, and that fine-tuning or post hoc correction may be necessary for applications requiring quantitative fidelity.

Persona grounding. Dash et al. [36] introduced PolyPersona, a system that instruction-tunes compact models with persona-conditioned survey responses across ten domains, exposing known issues of variance collapse and temporal drift even in controlled settings. PersonaHub is curated with one billion personas as distributed carriers of world knowledge, showing that persona diversity at scale can improve response heterogeneity, though the relationship between persona specification richness and response validity remains under-studied.

Validity conditions. Hullman et al. [23] provide the most thorough treatment of when LLM simulations constitute valid behavioral evidence. They argue that heuristic "validate-then-simulate" approaches are insufficient without statistical calibration or bias correction, and recommend prediction-powered inference (PPI) methods [37,38] that combine small human samples with larger LLM outputs through additive correction terms. Bisbee et al. [25] found that while ChatGPT 3.5 Turbo-generated survey means approximate real distributions, variance is systematically understated and regression coefficients diverge, a finding that reinforces the need for calibration beyond mean-level comparisons.

SAPIENT's simulation layer operates in a qualitative–exploratory mode rather than a quantitative–predictive one, which partially sidesteps distributional fidelity requirements. This literature identifies three design imperatives that we incorporate: (1) repeated runs with variance reporting rather than single-shot generation, (2) explicit validity limits distinguishing hypothesis generation from opinion measurement, and (3) calibration

hooks enabling future integration of PPI-style correction when human comparison data become available.

2.8. AI for Greenwashing Detection and Corporate Reputation

NLP-based greenwashing detection has accelerated under regulatory pressure from the EU CSRD [39]. Schimanski et al. [40] developed ClimateBERT-based classifiers for environmental claims; SESAMm's TextReveal [3] analyzes billions of web sources for greenwashing risk. Ghaemi Asl [41] proposed an AI-driven framework integrating machine learning with ESG analysis to assess greenwashing potential across domain-specific enterprises in developed and emerging markets. RepRisk 2024 data showed a 12% decrease in greenwashing incidents but increased severity [2]. Greenhushing has received far less research attention [3]. The literature addresses identifying misleading claims without addressing how stakeholders respond, and that is the gap SAPIENT bridges.

3. The SAPIENT Framework

SAPIENT links an observational layer (sentinel monitoring) to an experimental layer (agent-based simulation) through an orchestrator that manages data flow, escalation, and feedback. The simulation layer generates qualitative hypotheses for human review. This boundary is enforced through the Agentic Focus Group protocol (Section 3.3.3), repeated-run variance reporting, and mandatory human review gates. Figure 2 presents the architecture.

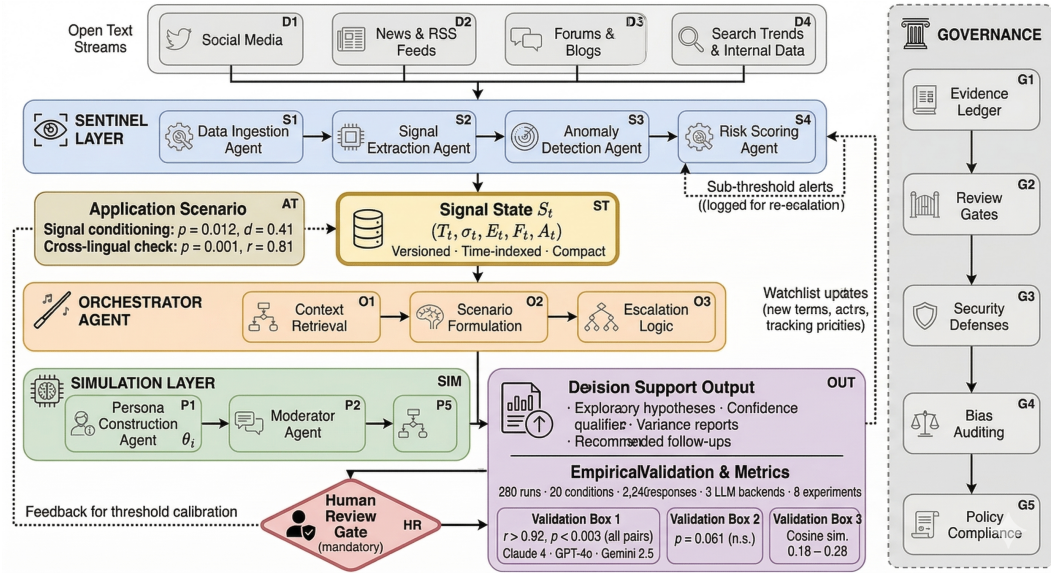

Figure 2. High-level architecture of SAPIENT. The sentinel layer produces a signal state $S_t = \langle T_t, \sigma_t, E_t, F_t, A_t \rangle$ that conditions simulation runs. The orchestrator manages scenario translation, escalation, and feedback. The simulation layer runs AFG sessions with persona (θ_i) and moderator agents. Governance components enforce review gates at all stages. Dashed arrows from simulation back to sentinel represent bidirectional coupling (watchlist updates, tracking priority adjustments).

3.1. Sentinel Layer

The sentinel layer operates as a continuous monitoring pipeline that transforms raw digital media streams into a structured, time-indexed *signal state* S_t. It is composed of four modules, each an autonomous agent.

3.1.1. Data Ingestion Agent

This agent collects text with metadata (timestamp, source, language) from social media APIs, news RSS feeds, forums, search trend services, and internal organizational

data. It performs initial filtering, including language detection, spam and bot removal, and deduplication. Two deployment variants are available: a *lightweight* configuration on cloud APIs for smaller organizations, and an *enterprise* configuration with on-premise deployment, data lake, and compliance integration. All ingested text is treated as *untrusted input*; content sanitization and input delimiting are applied before any text reaches an LLM component, guarding against indirect prompt injection from adversarial content planted in social media posts or forum comments.

3.1.2. Signal Extraction Agent

This agent applies four parallel analysis pipelines: (1) topic modeling (BERTopic or LDA) generating time-series topic distributions with representative snippets; (2) transformer-based sentiment and stance scoring fine-tuned for sustainability and corporate communication language; (3) named-entity recognition and temporal co-occurrence tracking for companies, products, executives, and regulators; and (4) framing analysis classifying content by communication frame (greenwashing accusation, compliance, environmental leadership, activist critique).

3.1.3. Signal State Formalization

We define the signal state S_t at time t as:

$$S_t = \langle \mathcal{T}_t,\ \sigma_t,\ \mathcal{E}_t,\ \mathcal{F}_t,\ \mathcal{A}_t \rangle \tag{1}$$

where $\mathcal{T}_t$ denotes the topic distribution, σ_t is a vector of sentiment statistics per topic (mean, variance, skewness), $\mathcal{E}_t$ captures entity co-occurrence networks and actor emergence metrics, $\mathcal{F}_t$ records the distribution of identified communication frames, and $\mathcal{A}_t$ is a set of active anomaly scores with categorical risk labels. The signal state is updated at each ingestion cycle and serves as the primary interface between the sentinel layer and the orchestrator. Maintaining S_t as a compact, versioned object enables reproducibility: any simulation run can be traced back to the exact signal state that triggered it.

3.1.4. Anomaly Detection Agent

This agent monitors S_t for statistically unusual patterns using multiple detection methods: volume-based (Z-score thresholds on mention counts), sentiment-based (abrupt polarization shifts measured as variance changes), topology-based (emergence of new influential actors), and topic-based (novel topic cluster birth or rapid topic migration). Each anomaly receives a preliminary score based on statistical magnitude and the number of co-triggering signal types.

3.1.5. Risk Scoring Agent

The risk scoring agent transforms anomaly scores into actionable risk assessments by incorporating impact potential (audience reach, network centrality), categorical risk (brand safety, regulatory, activist, competitive, internal leak), and temporal trajectory (accelerating, stable, decaying). Alerts exceeding thresholds are forwarded to the orchestrator; sub-threshold alerts are logged for potential re-escalation. All alerts pass through a human review queue generating feedback data for threshold calibration.

3.2. Orchestrator Agent

The orchestrator bridges sentinel and simulation layers with three operations. First, it **contextualizes** alerts by retrieving historical data, including previous alerts, past response patterns, and ongoing simulations. Second, it **formulates simulation scenarios** by translating alerts into structured input packages for the simulation layer: detected narrative,

framing, key entities, representative quotes from S_t (Equation (1)), affected stakeholder segments, and specific research questions. Third, it applies **escalation logic**: high-risk alerts trigger immediate human review alongside simulation; moderate-risk alerts proceed to automated simulation with results surfaced in the next reporting cycle.

This design echoes recent work on hierarchical multi-agent coordination in systems engineering. Quast et al. [42] found that a supervisor agent managing task decomposition, intent recognition, and sub-agent delegation can improve the structural consistency of LLM outputs in complex information workflows. The orchestrator in SAPIENT serves an analogous coordinating function, translating unstructured sentinel alerts into structured simulation scenarios and applying escalation protocols before delegating tasks to the simulation layer.

Bidirectional coupling. The orchestrator enables a feedback path from simulation to sentinel. When AFG sessions reveal that certain framings or counter-narratives resonate strongly with specific segments, the orchestrator updates the sentinel's watchlists, adding search terms, adjusting topic tracking priorities, or flagging actors for closer monitoring. This bidirectional flow means that observational capabilities adapt to experimental insights over time.

3.3. Simulation Layer

The simulation layer generates anticipatory intelligence by running moderated qualitative sessions using LLM-driven persona agents, building on the generative agents paradigm [14] and reframing it for corporate reputation contexts.

3.3.1. Persona Construction Agent

This agent generates LLM-driven agents characterized by structured profiles. Each persona i is specified by a tuple:

$$\theta_i = \langle \mathbf{d}_i,\ \mathbf{p}_i,\ r_i,\ \mathbf{b}_i,\ \ell_i \rangle \tag{2}$$

where $\mathbf{d}_i \in \mathcal{D}$ is a demographic attribute vector (age bracket, gender, education, income range, geographic region), $\mathbf{p}_i \in \mathcal{P}$ is a psychographic attribute vector (environmental concern level, brand loyalty, media consumption habits, institutional trust), $r_i \in \mathcal{R}$ is a stakeholder role label (consumer, investor, employee, regulator, journalist, activist), $\mathbf{b}_i \in \mathcal{B}$ encodes behavioral priors (engagement style, narrative frame susceptibility, information-seeking tendency), and ℓ_i denotes the primary language and cultural context.

Sampling strategy. Personas are sampled to approximate target population structure rather than to represent any individual. The sampling procedure operates in three steps: (1) a marginal distribution $P(\mathbf{d})$ is specified from public demographic data (census, industry reports) or organizational CRM summaries; (2) psychographic and behavioral attributes are sampled conditionally, $P(\mathbf{p}, \mathbf{b} \mid \mathbf{d})$, using either empirical distributions from prior surveys or uniform priors when data are unavailable, with explicit documentation of which priors are data-informed and which are assumed; (3) sentinel-derived language cues from S_t are injected as contextual priors (e.g., if S_t indicates activist framing is dominant, the proportion of activist-role personas is increased relative to the baseline). Each persona is implemented through an LLM with a structured system prompt encoding θ_i, a per-session memory buffer m_i (reset between AFG runs to keep them independent), and a response policy π_i that includes rules for uncertainty expression, citation behavior, and refusal of off-topic prompts.

The decision to anchor personas in multi-dimensional attribute vectors rather than free-form descriptions reflects evidence that behavioral responses vary systematically across sociodemographic and psychographic clusters. Grounding generative agents in structured

profiles is intended to produce heterogeneous group dynamics rather than the uniform response patterns that unstructured persona prompts tend to generate.

Prior source transparency. Table 1 documents the data sources and justification status for each attribute category used in the experimental personas. We do not claim that these priors reproduce actual population distributions. The priors define the simulation's input space; the validity of outputs is assessed through the evaluation stages (Sections 5–6), not through the fidelity of persona sampling to real populations.

Table 1. Persona attribute priors: data sources and justification status. "Data-informed" indicates that the attribute distribution was derived from a published source; "assumed" indicates that the distribution was set by the authors based on domain knowledge without empirical calibration.

Attribute	Status	Source/Justification
Age bracket	Data-informed	TÜİK census (2023); Eurostat for EU scenarios
Gender	Data-informed	TÜİK census (2023)
Education level	Data-informed	TÜİK education statistics (2023)
Geographic region	Data-informed	Scenario-specific (matched to company footprint)
Stakeholder role	Assumed	Selected to cover key segments identified in crisis communication literature [1]
Env. concern level	Assumed	Uniform distribution across low/medium/high; no population-level calibration
Brand loyalty	Assumed	Uniform prior; scenario-adjusted via S_t
Media consumption	Assumed	Categorical (social-first, traditional, mixed); proportions set by authors
Institutional trust	Assumed	Ordinal scale; no empirical anchor
Engagement style	Assumed	Categorical (passive, reactive, proactive); uniform

Variance collapse prevention. A known failure mode in LLM persona simulation is variance collapse, where demographically distinct personas produce near-identical responses [25,36]. Three countermeasures are applied: (a) *temperature stratification*, where response temperature is varied across personas within a session (e.g., $\tau \in [0.6, 1.0]$) to induce output diversity without losing coherence; (b) *diversity audit*, where after each AFG run, pairwise semantic similarity among persona responses is computed, and runs where mean similarity exceeds a threshold δ are flagged and re-prompted with explicit instructions to express disagreement; and (c) *adversarial persona injection*, where at least one persona per session is assigned a contrarian stance so that the discussion space is not artificially consensual.

Multilingual and cross-cultural handling. When the sentinel layer detects signals in multiple languages, personas are assigned language-appropriate prompts. Known challenges include language-linked safety asymmetries (LLMs may exhibit different bias profiles across languages [18]) and cultural framing differences. The calibration agent (Section 3.3.6) includes language-stratified bias checks, and AFG reports flag findings that depend on single-language runs as having reduced generalizability.

3.3.2. Moderator Agent

The moderator agent is adapted from established qualitative research practice. It controls turn-taking among personas, probes for reasoning and disagreements (e.g., "What specific information would change your opinion?"), makes sure the session adheres to the research brief, enforces safety rules (blocking harassment or manipulative content), stops runs when outputs become repetitive, and produces a structured session debrief summarizing themes, disagreements, and unresolved questions.

3.3.3. Agentic Focus Group (AFG) Protocol

The AFG protocol formalizes the simulation procedure into a repeatable experimental process. The protocol proceeds as Table 2.

Table 2. Agentic Focus Group (AFG) Protocol Pseudo-Code.

Input: Research brief B, signal state S_t, persona set size n, number of runs K
Output: Transcripts $\{T_k\}_{k=1}^{K}$, structured summaries $\{U_k\}_{k=1}^{K}$, variance report V

Build personas: Sample n specifications $\{\theta_i\}_{i=1}^{n}$ conditioned on B and S_t
for $k = 1$ **to** K **do**
  Initialize: Reset persona memories m_i and moderator state
  Warm-up: Moderator asks baseline questions about the topic domain
  Stimulus: Moderator presents stimulus x from B (message draft, event, policy)
  Discussion: Multi-turn conversation with probes for reasons and alternatives
  Structured output: Each persona returns (a) interpretation, (b) concerns, (c) questions
  Debrief: Moderator summarizes themes and flags disagreements
  Store transcript T_k and structured summary U_k
end for
Aggregate: Compute theme frequencies across K runs; detect unstable themes
Variance report: Generate V documenting run-to-run stability per theme
Review gate: Human reviewer checks citations, safety, and face validity
Report: Return $\{U_k\}$, V, limitations, and recommended follow-ups

The design rationale is as follows. **Multiple runs (K)** serve the same purpose as replication in experimental research: a theme appearing in 9 of 10 runs carries more weight than one in 2 of 10. **Variance reporting** makes instability visible rather than hiding it behind aggregated summaries. **The human review gate** is mandatory, not optional. For crisis scenarios, $n = 8$–12 personas with $K = 5$ runs suffice; for broader exploration, $n = 30$–50 with $K = 10$ runs are recommended.

3.3.4. Interaction Modes

Two modes are available: (1) **Independent response**, where each persona responds without interaction, which is computationally efficient and suitable for rapid testing; (2) **Networked interaction**, where personas on a simulated social network observe and respond to each other across multiple rounds, capturing cascades and echo chambers at higher computational cost. Both operate within the AFG protocol.

3.3.5. Opinion Aggregation Agent

This agent collects persona responses across K runs and synthesizes: sentiment distributions by stakeholder segment with cross-run confidence bands; dominant narratives ranked by cross-run consistency; predicted behavioral intentions (share, criticize, boycott, support, ignore); divergence indicators for polarized segments; and unstable themes flagged for human validation.

3.3.6. Calibration and Bias Mitigation Agent

This agent addresses representativeness concerns identified in the calibration literature [23–25]. It performs four checks:

(1) Distribution alignment. When reference human data are available (e.g., from prior surveys or small pilot samples), the agent computes distributional distance between synthetic and human response distributions. For categorical responses, forward KL divergence is used:

$$D_{\mathrm{KL}}(P_{\mathrm{human}} \| P_{\mathrm{synth}}) = \sum_{c} P_{\mathrm{human}}(c) \log \frac{P_{\mathrm{human}}(c)}{P_{\mathrm{synth}}(c)} \tag{3}$$

For ordinal or continuous measures, the Wasserstein distance W_1 is preferred following SubPOP [24]. When no reference data exist, the agent reports this explicitly and flags outputs as "uncalibrated exploratory."

(2) Run-to-run stability. For each theme j identified across K AFG runs, the agent computes the theme frequency $f_j^{(k)}$ in run k and reports bootstrap 95% confidence intervals over $\{f_j^{(1)}, \ldots, f_j^{(K)}\}$. Themes with coefficient of variation $\mathrm{CV}_j > 0.5$ are classified as "unstable" and excluded from summary recommendations.

(3) Response diversity audit. Mean pairwise cosine similarity among persona responses within each run is computed using sentence embeddings. If similarity exceeds threshold $\delta = 0.85$, the run is flagged for variance collapse [36] and may be re-executed with adjusted temperature or explicit diversity prompts. The rationale for these stability and diversity checks finds support in recent work on self-organizing multi-agent systems: Gerolimos et al. [43] argue that task-level performance indicators alone are insufficient to capture the organizational quality and internal stability of agentic systems, and that structural cohesion metrics are needed alongside output-level evaluation. SAPIENT's approach, executing K independent runs and computing pairwise semantic similarity alongside theme frequency, serves a parallel purpose: auditing whether the synthetic consensus reflects genuine pattern stability or is merely an artifact of prompt sensitivity.

(4) Bias and safety checks. Stereotype rate is measured as the frequency of responses that match pre-defined stereotype templates across demographic groups. Toxicity rates and multilingual behavioral drift [18] are monitored per language stratum. Results are appended to reports as a "confidence and limitations" section.

Future calibration integration. The architecture includes hooks for prediction-powered inference (PPI) [37,38]: when even small human samples ($n \geq 30$) are available, PPI-style additive correction can reduce bias in aggregate estimates while taking advantage of the larger synthetic sample for variance reduction. This integration path is specified but not yet implemented; empirical validation is planned as part of the Stage 2 evaluation (Section 5).

3.4. Human Review Gate: Workflow and Design Rationale

The mandatory human review gate operates at the boundary between the simulation layer and decision-support output. No AFG synthesis is released to downstream consumers without analyst approval. The review workflow proceeds as follows:

1. **Automated pre-screening.** The calibration and bias agent flags runs that exhibit (a) variance collapse ($\bar{\delta}_{\cos} > 0.85$), (b) stereotype amplification (demographic-attribute frequency exceeding a preset threshold), or (c) hallucinated entities not present in S_t. Flagged runs are quarantined and presented to the reviewer with specific diagnostic annotations.
2. **Structured review checklist.** The analyst evaluates each AFG synthesis against five criteria, recorded in the evidence ledger:
    - *Signal fidelity*: Do the dominant themes trace logically to the information in S_t?
    - *Persona plausibility*: Are persona responses consistent with their assigned demographic and attitudinal profiles?
    - *Diversity adequacy*: Do responses exhibit sufficient inter-persona variation, or do multiple personas echo the same language?
    - *Hallucination check*: Does the synthesis reference events, statistics, or entities not grounded in S_t or the persona background?

- *Actionability*: Are the generated hypotheses specific enough to inform monitoring adjustments or communication strategy, without presenting hypotheses as predictions?

3. **Disposition.** The reviewer assigns one of three outcomes: Approve (synthesis is forwarded), Revise (specific runs are re-generated with modified parameters), or Reject (scenario is re-formulated by the orchestrator with analyst input).

The review workflow for each synthesis involved reading the aggregated theme summary, spot-checking two individual run transcripts, and completing the checklist. We acknowledge that the authors, who designed the system, likely exhibit confirmation bias in this role; independent reviewer evaluation is planned for the Stage 4 prospective pilot (Section 5).

3.5. Governance, Safety, and Misuse Prevention

SAPIENT is a decision-support system, not an autonomous actor. This section specifies governance mechanisms organized by risk category.

3.5.1. Human-in-the-Loop Checkpoints

High-risk alerts require human review before simulation results are forwarded to decision-makers. Simulation results are presented as exploratory hypotheses with explicit confidence qualifiers (stable or unstable themes, calibrated or uncalibrated estimates). No automated action is taken on the basis of simulation outputs alone.

3.5.2. Evidence Ledger and Reproducibility

All decision points are logged with version control: ingestion parameters, anomaly thresholds, persona specifications $\{\theta_i\}$, prompt templates (verbatim), LLM model identifiers and versions, temperature settings, random seeds where applicable, and signal states S_t that triggered each simulation. This ledger enables post hoc auditing and supports the reproducibility requirements stressed by Hullman et al. [23].

3.5.3. Model Security and Adversarial Robustness

The sentinel layer ingests text from open web sources, creating a direct attack surface for indirect prompt injection [44]. Three defense layers are specified:

1. **Input isolation.** All ingested text is treated as data, never as instructions. Sentinel LLM calls use structured tool-use interfaces with strict input delimiters separating system instructions from user-provided content. No ingested text is concatenated into system prompts.
2. **Output verification.** Summaries and extracted signals are checked against source text using citation-verification pipelines (chain-of-density with source grounding). Summaries that introduce claims not traceable to source documents are flagged.
3. **Red-team protocol.** Before deployment, the system undergoes adversarial stress testing: (a) injection of known prompt-injection payloads into the ingestion stream to verify containment; (b) insertion of fabricated crisis signals to test escalation thresholds; and (c) planting of contradictory information across sources to assess summarization faithfulness. The results are documented in the evidence ledger.

3.5.4. Bias Auditing

Quarterly reviews assess systematic bias across demographic groups, geographic regions, languages, and political orientations. Audits compare AFG output distributions against available benchmark data and flag persistent skews for calibration adjustment.

3.5.5. Transparency and Data Minimization

The use of synthetic population modeling is disclosed when results inform external communications. Only anonymized, aggregated data are used for persona construction, with no individual profiling or behavioral tracking. Policy compliance follows NIST AI RMF [45] and applicable regulations (GDPR, KVKK).

3.5.6. Misuse Prevention and Non-Use Contexts

Generating simulated public opinion creates risks of repurposing for manipulative messaging, astroturfing, or surveillance. The following restrictions are specified:

Prohibited uses: (1) tuning messaging designed to deceive or manipulate (e.g., crafting greenwashing that evades detection); (2) generating synthetic personas to impersonate real individuals or groups; (3) deploying simulation outputs as evidence of actual public opinion in regulatory filings or public communications without disclosure; and (4) using sentinel monitoring capabilities for individual-level surveillance or profiling.

Non-use contexts: SAPIENT outputs should *not* be used as the sole basis for: (a) regulatory compliance decisions, (b) legal proceedings involving public sentiment claims, (c) automated content moderation or censorship, or (d) personnel decisions based on simulated employee sentiment. These boundaries are documented in a usage policy distributed with the system and enforced through access controls and audit logging.

4. Application Scenarios

Four scenarios ground SAPIENT in operational contexts. Each specifies trigger conditions, sentinel signals, AFG session design, and decision-support output.

4.1. Scenario 1: Greenwashing Backlash Prediction

Context. A multinational prepares to announce "net-zero by 2040." The communications team is concerned about potential greenwashing perception given the company's historical carbon footprint and a recent investigative report on supply chain practices.

Sentinel role. Establishes pre-announcement signal state S_{t_0}: prevailing topics, sentiment baselines, active critics, competitor sustainability framing.

AFG design. Stimulus: draft announcement in three framing variants (targets, past progress, accountability). Personas: environmental activists, ESG investors, general consumers, industry journalists, regulatory observers. $K = 5$ independent-mode runs. Moderator probes credibility perception and information gaps.

Output. Comparative report of reaction distributions per variant with cross-run stability scores. Unstable themes flagged for human panel follow-up.

4.2. Scenario 2: Greenhushing Risk Assessment

Context. A financial firm with genuine ESG progress decides not to publicize, fearing performative perception. Investor relations is concerned that ESG investors may interpret silence negatively.

Sentinel role. Monitors peer ESG disclosure discourse; tracks whether "absence of disclosure" emerges as a topic in investor forums.

AFG design. Two counterfactual stimuli: (a) firm publishes ESG report; (b) firm remains silent while competitor discloses. Personas: finance-focused stakeholders including institutional investors, ESG analysts, financial journalists, retail investors, regulators, and sustainability academics. Independent mode. $K = 10$ runs per condition.

Output. Risk comparison between strategies with variance report indicating which conclusions are stable and which remain uncertain.

4.3. Scenario 3: Sustainability Campaign Pre-Testing

Context. An automotive manufacturer has three EV campaign concepts. Concept (b) on environmental responsibility may attract greenwashing accusations given ongoing ICE production.

Sentinel role. Scans EV sustainability discourse, competing campaigns, recent automotive greenwashing cases.

AFG design. Demographically stratified personas evaluate each concept. Moderator probes counter-narrative vulnerability. $K = 5$ independent-mode runs. **Bidirectional feedback**: newly identified counter-narrative terms added to sentinel watchlist for post-launch monitoring.

Output. Multi-criteria comparison: acceptance, amplification potential, vulnerability dimensions.

4.4. Scenario 4: Crisis Communication Simulation

Context. Sentinel detects an emerging anomaly: an investigation alleges the "sustainably sourced" product line relies on violating suppliers. Mentions are doubling every three hours.

Sentinel role. Real-time tracking: spread rate, amplifying accounts, geographic distribution, dominant framing.

AFG design. High-risk escalation. Independent mode for speed. Three candidate responses: (a) immediate apology with corrective plan, (b) factual rebuttal, or (c) delayed response pending investigation. $K = 3$ runs given time pressure. Moderator focuses on 24–72 h sentiment trajectory.

Output. Time-trajectory projections per strategy, delivered within 60 min. Variance report notes which projections are stable and which are uncertain.

5. Evaluation Framework

Validation requires evaluation at three levels. Single-number scores should be avoided; each metric group reveals different aspects of system behavior. Table 3 summarizes the proposed metrics.

Table 3. Evaluation metrics organized by layer. Simulation evaluation is structured into stability, calibration, and bias groups.

Layer	Metric	Description
Sentinel	Detection quality	Precision, recall, F1 against labeled domain benchmarks.
	Time-to-detect	Time from first observable signal to alert generation.
	False alarm rate	Alerts classified as non-events upon human review.
	Alert burden	Analyst load: false alerts/day, time spent per alert.
Simulation: Stability	Theme stability	Run-to-run variance of themes and sentiment across K runs.
	Persona consistency	Stance consistency across repeated sessions.
Simulation: Calibration	Macro-calibration	Agreement of aggregate sentiment with observed real-world data.
	Distribution fidelity	Shape comparison against known population distributions.
	Tail accuracy	Reproduction of minority and extreme opinions.
Simulation: Bias & Safety	Stereotype rate	Stereotyped response frequency across demographic groups.
	Toxicity & drift	Toxicity rates and multilingual behavioral drift.
Operational Impact	Expert agreement	Correlation of insights with experienced PR professionals' choices.
	Response time	Reduction in organizational response time from early warnings.
	Review pass rate	Fraction of outputs passing human review without edits.
	Outcome correspondence	Correspondence of simulated trajectories with real-world outcomes.

5.1. Experimental Design

Five evaluation stages are planned, ordered by increasing complexity:

Stage 1: Sentinel layer benchmarking. The sentinel layer is evaluated on archived data from documented corporate crises with known timelines: the DWS Group ESG controversy (2022–2023), the Volkswagen emissions trajectory (2015–2016), and recent greenwashing cases flagged by RepRisk [2]. Ground-truth timelines are constructed from regulatory announcements and major media coverage dates. Metrics: precision, recall, F1 for domain classifiers; time-to-detect measured as the lag between the first detectable signal in the data and the system's alert generation; false alarm rate and alert burden (analyst-hours per alert). Baselines: keyword-volume monitoring (Brandwatch-style), rule-based anomaly detection without LLM summarization.

Stage 2: Simulation calibration against human data. AFG outputs for three historical scenarios are compared against real-world audience data. Two calibration approaches are tested: (a) zero-shot persona prompting (current SAPIENT design), and (b) SubPOP-style fine-tuned distributional methods [24] as an upper-bound reference. Metrics: Wasserstein distance and KL divergence between synthetic and human sentiment distributions, stratified by stakeholder segment and language. Theme coverage is measured as the fraction of themes identified in human focus groups that also appear in AFG transcripts. Following Bisbee et al. [25], we test whether AFG outputs reproduce means, variance, and covariance structure. Minimum human sample sizes follow Hullman et al.'s recommendations [23]: $n \geq 50$ per subgroup for distributional comparisons.

Stage 3: Sentinel-to-simulation coupling test. An A/B comparison tests whether conditioning personas on the signal state S_t improves simulation quality relative to generic context. For each historical case, AFG sessions are run in two conditions: (A) personas receive the full signal state including detected narratives, entity networks, and framing distributions; and (B) personas receive only the topic label and a generic description. Metrics: theme coverage relative to observed real-world discourse, face validity rated by domain experts (inter-rater κ), and hallucinated theme rate (themes with no basis in either S_t or ground truth).

Stage 4: Robustness and adversarial testing. Three dimensions of testing: (a) *multilingual stress test*, where the same scenario is run in English, Turkish, and German to assess cross-lingual consistency and language-linked bias differences; (b) *temporal drift*, where the same AFG configuration is re-run at 30-day intervals to measure persona stability and output reproducibility across LLM version updates; and (c) *adversarial input*, where prompt injection payloads, fabricated quotes, and misleading entities are injected into the sentinel stream to test containment and summarization faithfulness.

Stage 5: Prospective pilot. SAPIENT is deployed alongside independent human analysts at a corporate partner for a 3-month period. Controlled comparison of decision quality (expert-rated appropriateness of recommended responses), response speed (time from alert to actionable recommendation), analyst workload (hours per case), and review-pass rate (fraction of AFG outputs accepted without substantive revision).

5.2. Baselines

- **Keyword-based monitoring**: Conventional social listening with volume-threshold alerts (e.g., Brandwatch, Meltwater).
- **Expert panel**: Communications professionals assessing scenarios without automated tools.
- **Traditional focus group**: Small-scale ($n = 8$–12) conventional focus group for direct comparison with AFG.

- **Classical ABM**: Rule-based bounded-confidence agents [32] on the same network topology, no LLM.
- **SubPOP-calibrated baseline**: Fine-tuned LLM generating survey responses without the sentinel layer or AFG protocol [24], to isolate the contribution of SAPIENT's architectural components.

6. Preliminary Experimental Results

Experiments 1–4 used Claude Sonnet 4 (Anthropic, claude-sonnet-4-20250514) as the underlying LLM. This model was selected on three grounds. First, it shows state-of-the-art performance on agentic task benchmarks that closely mirror the multi-step, tool-augmented workflows required by the AFG protocol: 72.7% on SWE-bench Verified and 80.5%/60.0% (Retail/Airline) on TAU-bench, which evaluates multi-turn interaction with simulated users under role-specific policies [46]. Second, the model exhibits strong instruction-following fidelity and reduced shortcut behavior, being 65% less likely to exploit loopholes compared to its predecessor [46], a property that is important for maintaining persona consistency across repeated AFG runs. Third, its system card documents explicit bias evaluations across demographic groups (BBQ benchmark: -1.16% bias, 86.3% accuracy) and multilingual safety assessments, providing a transparent baseline against which simulation-layer bias can be audited.

During revision, five additional experiments (5–9) were conducted. Experiments 5–8 addressed scenario generalizability, cross-model comparison, and prompt sensitivity. Experiment 7 introduced GPT-4o (OpenAI, `gpt-4o`) and Gemini 2.5 Flash (Google, `gemini-2.5-flash`) [47] as second and third LLM backends, using the same prompts, persona specifications, and evaluation metrics. Experiment 9 provided pilot external validation by comparing SAPIENT's Experiment 1 outputs against human responses collected through Prolific ($n = 54$); this experiment was preregistered at OSF prior to data collection (https://doi.org/10.17605/OSF.IO/4KFDC). All API calls were routed through a unified abstraction layer that standardizes the interface across providers while logging token consumption, latency, and cost per call. The addition of two new backends, two new application scenarios, and one human validation study required only configuration-level changes to the existing codebase, with no modifications to the core AFG protocol or analysis pipeline.

To provide empirical evidence for the architectural claims made in previous sections, we conducted nine experiments. Experiments 1–8 used the AFG protocol (Table 2) across three application scenarios: greenwashing backlash prediction (Scenario 1, Section 4.1), greenhushing risk assessment (Scenario 2, Section 4.2), and crisis communication simulation (Scenario 4, Section 4.4). Experiment 9 provided pilot external validation by comparing SAPIENT outputs with human responses collected through Prolific. All AFG experiments used $n = 8$ personas per session, each specified as a structured attribute vector θ_i following Equation (2). All experiments used independent mode exclusively: each persona responded to the stimulus and moderator probes without observing other personas' outputs. The network interaction mode described in Section 3.3 is an architectural module whose empirical validation is deferred to future work. Each persona was instantiated as a prompt-conditioned role on the shared LLM backend (not as a separate model instance), with a distinct system prompt encoding its demographic, psychographic, and behavioral attributes, and a separate API call per persona per turn. The source code, configuration files, and raw outputs are available at https://github.com/alperozpinar/SAPIENT-Framework.

6.1. Hypotheses

The experiments test four hypotheses derived from the architectural claims:

H1 (Persona Differentiation). *The AFG protocol produces persona-level behavioral signatures that are consistent across independent runs and differentiated across stakeholder roles.*

H2 (Signal Conditioning). *Grounding simulation in sentinel-derived signal states produces thematically more focused and contextually specific outputs than generic prompting.*

H3 (Cross-Lingual Asymmetry). *Identical scenarios run in different languages produce measurably different persona behaviors, reflecting language-linked LLM asymmetries, though relative persona ordering is preserved.*

H4 (Model Dependence). *Simulation outputs vary systematically across LLM backends, and the nature and magnitude of this variation can be characterized through the framework's cross-model comparison protocol.*

H9a (Human Ranking). *Human participants rank the three framing variants in the same order as SAPIENT's AFG protocol for both sentiment and credibility.*

H9b (Accountability Advantage). *The accountability framing receives significantly higher ratings than the targets framing in paired within-subject comparison.*

H9c (Theme Overlap). *Open-ended human themes show partial overlap with SAPIENT-generated themes (soft Jaccard* > 0.10 *in at least two of three variants).*

H1–H4 are tested across Experiments 1–8. All Experiments 1–8 analyses are exploratory and were not pre-registered. Experiment 9, testing H9a–H9c, was preregistered at OSF prior to data collection (https://doi.org/10.17605/OSF.IO/4KFDC).

6.2. Metrics and Statistical Methodology

Four primary metrics were computed across all experiments.

Sentiment and credibility ratings. Each persona agent was instructed to return, as part of the AFG protocol's structured-output phase, a numerical sentiment score (1–7 Likert scale, where 1 = strongly negative and 7 = strongly positive) and a credibility score (1–7, where 1 = not at all credible and 7 = highly credible) for the stimulus message. These scores were generated by the LLM as part of each persona's structured response and are *model-internal* judgments rather than external human annotations. We acknowledge this circularity as a limitation: the instrument and the subject are the same model. We note, though, that (a) the scores serve as *within-system* comparison metrics, meaning that differences across conditions are meaningful even if absolute values lack external calibration, and (b) Stage 2 of the evaluation plan (Section 5) is designed to assess correspondence between these model-generated ratings and human focus group data.

Thematic analysis. After each AFG run, persona responses were parsed to extract theme labels using a two-stage procedure: (1) each persona's qualitative interpretation and concern statements were segmented into atomic claims, and (2) an independent LLM call (with a fixed extraction prompt, provided in the project repository) assigned each claim a short thematic tag. Unique themes were counted per variant and per condition. Theme stability was assessed by computing the coefficient of variation (CV_j) for each theme j across K runs; themes with $CV_j \leq 0.5$ were classified as "stable." No semantic clustering was applied to merge near-duplicate theme labels (e.g., `timeline_feasibility`

vs. `aggressive_timeline`); the reported stability ratios represent a conservative lower bound for this reason.

Response diversity (variance collapse indicator). Mean pairwise cosine similarity among persona responses within each run was computed using TF-IDF vectors over unigrams and bigrams. A threshold of $\delta = 0.85$ was adopted from the architectural specification (Section 3.3.6); runs exceeding this threshold would be flagged for variance collapse. Sentence-embedding-based similarity (e.g., using multilingual models such as LaBSE) would provide a more semantically grounded measure; the TF-IDF approach was chosen for computational efficiency in the initial experiments and constitutes a known methodological simplification.

Focused discourse. The term "focused discourse," as used in the signal-conditioning comparison (Experiment 2), is operationalized as a composite of three indicators: (a) higher mean sentiment and credibility ratings, reflecting more differentiated engagement with the stimulus rather than generic commentary; (b) a higher proportion of stable themes (CV ≤ 0.5), indicating that signal-conditioned personas converge on recurring concerns rather than producing scattered, run-dependent observations; and (c) the presence of context-specific themes directly traceable to information provided in S_t (e.g., `audit_requirements`, `baseline_verification`), as opposed to generic corporate-criticism themes. We do not claim that higher sentiment equates to better simulation quality; rather, the pattern of higher differentiation, greater thematic stability, and context-specificity collectively operationalizes "focus."

Statistical tests. Comparisons between conditions were conducted using Welch's independent-samples t-tests at the individual persona-response level ($N = n \times K$ per condition). Because responses within a single AFG run share the same moderator trajectory and signal state, we verified all reported results using run-level aggregation as well: for each run, persona-level scores were averaged to yield a single run mean, and Wilcoxon rank-sum tests were applied to the K run means per condition as a nonparametric robustness check. All effects reported as significant at the response level remained directionally consistent at the run level, though some individual comparisons did not reach $p < 0.05$ at run level because of the smaller effective sample size ($K = 10$). Effect sizes are reported as Cohen's d using pooled standard deviations. To control for multiplicity, the Benjamini–Hochberg (BH) procedure was applied across all inferential tests at $\alpha = 0.05$; results are reported in Section 6.3.

6.3. Multiple Comparison Correction

Eleven inferential tests were conducted across Experiments 2–8 (Table 4). Because these tests address distinct hypotheses (H1–H4) but draw on related experimental structures, we applied the Benjamini–Hochberg (BH) procedure to control the false discovery rate (FDR) at $\alpha = 0.05$. Nine of eleven tests survive correction. The two non-significant results are methodologically expected: Experiment 8 sentiment ($p_{\text{adj}} = 0.067$), confirming that the primary metric is robust to prompt reformulation, and the GPT-4o–Gemini sentiment comparison ($p_{\text{adj}} = 0.212$), indicating that these two models produce similar absolute sentiment levels. As a sensitivity check, the more conservative Holm–Bonferroni correction (controlling the family-wise error rate) was also computed; Holm-corrected values are available in the project repository.

Table 4. Benjamini–Hochberg correction across all inferential tests (Experiments 2–8). * $p_{adj} < 0.05$, ** $p_{adj} < 0.01$, *** $p_{adj} < 0.001$.

Exp.	Test	Raw p	BH adj. p	Sig.
8	Credibility: original vs. paraphrase	<0.001	<0.001	***
7	Persona corr. GPT-4o - Gemini ($r = 0.960$)	0.0002	0.001	**
7	Persona corr. Claude-GPT-4o ($r = 0.935$)	0.0006	0.002	**
3	Sentiment: English vs. Turkish	0.001	0.002	**
7	Persona corr. Claude-Gemini ($r = 0.924$)	0.001	0.002	**
7	Sentiment: Claude vs. Gemini	0.002	0.004	**
7	Sentiment: Claude vs. GPT-4o	0.010	0.016	*
2	Sentiment: signal-conditioned vs. generic	0.012	0.017	*
3	Persona corr. EN-TR ($r = 0.808$)	0.015	0.018	*
8	Sentiment: original vs. paraphrase	0.061	0.067	n.s.
7	Sentiment: GPT-4o vs. Gemini	0.212	0.212	n.s.

6.4. Experimental Configurations and Response Accounting

Table 5 summarizes the experimental design across all nine experiments. Experiments 1–8 comprised 280 unique AFG runs across 20 experimental conditions, involving 2240 persona-level responses. Experiment 9 added a pilot human validation study with 54 analyzable human participants recruited through Prolific.

Table 5. Summary of experimental configurations. K denotes independent AFG runs per condition. Each run involves $|\Theta| = 8$ personas (Resp. $= K \times 8$). Experiments 1–6 and 8 use Claude Sonnet 4; Experiment 7 uses Claude Sonnet 4, GPT-4o, and Gemini 2.5 Flash. Experiment 9 uses human participants recruited through Prolific.

Exp.	Condition	K	Resp.	Independent Variable
1	A: Targets framing	20	160	Communication frame
	B: Progress framing	20	160	
	C: Accountability framing	20	160	
2	A: Signal-conditioned (S_t)	10	80	Signal state
	B: Generic prompt	10	80	
3	A: English sessions	10	80	Session language
	B: Turkish sessions	10	80	
4	A: Uniform $\tau = 0.7$	10	80	Collapse countermeasure
	B: Stratified τ	10	80	
	C: Stratified + adversarial	10	80	
5	A: Disclose ESG	10	80	Disclosure strategy
	B: Remain silent	10	80	
6	A: Immediate apology	10	80	Crisis response
	B: Factual rebuttal	10	80	
	C: Delayed response	10	80	
7	Claude Sonnet 4	20	160	LLM backend
	GPT-4o	20	160	
	Gemini 2.5 Flash	20	160	
8	Original prompt	20	160	Prompt wording
	Paraphrase	20	160	
9	Human: Targets framing	–	54	External validation
	Human: Progress framing	–	54	
	Human: Accountability framing	–	54	
Total		280	2240	

6.5. Experiment 1: AFG Protocol Validation

Three framing variants of a corporate net-zero announcement were tested with $K = 20$ independent runs each: (A) target-focused, emphasizing investment scale and timeline; (B) progress-focused, highlighting past emission reductions; and (C) accountability-focused, acknowledging supply chain failures and proposing immediate corrective actions. Table 6 summarizes the results.

Table 6. AFG protocol results across three framing variants ($K = 20$ runs, $n = 8$ personas per run). Sentiment and credibility on 1–7 Likert scale. Stability ratio: fraction of themes with CV ≤ 0.5 across runs. Cosine similarity: mean pairwise TF-IDF similarity among persona responses within each run (variance collapse threshold $\delta = 0.85$).

Variant	Sentiment	Credibility	Themes	Stability	Cos. Sim.
A (Targets)	3.96 ± 1.01	3.81 ± 0.70	442	1.4%	0.196
B (Progress)	4.08 ± 0.95	3.85 ± 0.73	446	1.8%	0.201
C (Accountability)	4.24 ± 0.82	4.36 ± 0.74	371	2.2%	0.185

The accountability framing (Variant C) produced the highest mean sentiment (4.24 vs. 3.96 for targets) and a higher credibility rating (4.36 vs. 3.81). This pattern fits with communication research suggesting that acknowledging failures can increase rather than decrease audience trust when paired with concrete corrective commitments. That the simulated agents systematically rewarded transparency and third-party verification over aspirational claims, preferring "we will submit to independent audit" over "we will invest $2 billion," parallels the behavioral regularities observed in *homo silicus* research [16]. It suggests that LLM-based personas can reproduce economically rational trust heuristics in a corporate communication context. Across all three variants, mean pairwise cosine similarity remained well below the $\delta = 0.85$ collapse threshold (range: 0.185–0.201), indicating that the eight personas generated substantively different responses rather than converging on a single narrative.

An examination of persona-level behavior across the 20 runs revealed consistent role-appropriate differentiation (Table 7). The activist persona (P1) maintained the lowest sentiment across all three variants (2.00–2.90), while the employee persona (P6) consistently scored highest (4.70–4.95). The accountability framing lifted sentiment for nearly all personas, but the magnitude of the shift varied: the regulator persona showed the largest gain (+0.95 range across variants), while the employee remained relatively stable (range of 0.25). These patterns persisted across $K = 20$ independent runs with low within-persona variance (mean sentiment standard deviation across runs: 0.22–0.89), suggesting that persona specifications produce reproducible behavioral signatures rather than arbitrary noise.

Table 7. Mean sentiment per persona across framing variants ($K = 20$ runs). Range column indicates cross-variant sensitivity for each persona. Sentiment on 1–7 Likert scale.

ID	Role	A (Targets)	B (Progress)	C (Account.)	Range
P1	Activist	2.00	2.20	2.90	0.90
P2	ESG Investor	4.65	4.55	4.90	0.35
P3	Consumer	4.35	4.80	4.35	0.45
P4	Journalist	3.25	3.90	3.65	0.65
P5	Regulator	4.25	3.90	4.85	0.95
P6	Employee	4.95	4.95	4.70	0.25
P7	Competitor	4.60	4.80	4.35	0.45
P8	Academic	3.60	3.55	4.25	0.70

The theme stability ratios, 1.4% to 2.2% of themes classified as stable (CV $\leq$ 0.5), appear low at first glance but require context. With $K = 20$ runs and 8 personas each generating 3–5 theme labels per run, the protocol surfaced 371–446 unique themes per variant. Most of these are fine-grained variants of broader categories (e.g., "timeline_feasibility," "aggressive_timeline," and "ambitious_timeline" all capture timeline concern). A semantic clustering step, not yet implemented, would substantially increase apparent stability. The raw counts confirm that the protocol generates a rich thematic space rather than a narrow one.

6.6. Experiment 2: Signal State Conditioning

The central architectural claim of SAPIENT is that grounding simulation personas in sentinel-derived signal states improves simulation quality. To test this, we ran the AFG protocol under two conditions with $K = 10$ runs each: Condition A provided personas with a rich signal state S_t including topic sentiment trends, active critic entities, competitor announcements, and anomaly scores; Condition B provided only a generic topic description ("a large chemical company is about to announce a net-zero target"). Table 8 summarizes the results.

Table 8. Signal state conditioning A/B test ($K = 10$, $n = 8$). Condition A receives full S_t from sentinel layer; Condition B receives generic topic only. Welch's t-tests with effect size (Cohen's d).

Metric	A (Signal)	B (Generic)	p	d
Sentiment (mean $\pm$ std)	4.30 $\pm$ 0.81	3.94 $\pm$ 0.97	0.012	0.41
Credibility (mean $\pm$ std)	4.41 $\pm$ 0.74	3.91 $\pm$ 0.88	–	–
Unique themes	248	277	–	–
Stable themes (CV $\leq$ 0.5)	9 (3.6%)	10 (3.6%)	–	–
Mean cosine similarity	0.183	0.170	–	–
Jaccard theme overlap	0.259		–	–

Signal-conditioned personas produced significantly higher sentiment ($t = 2.552$, $p = 0.012$, BH-adjusted $p = 0.017$) compared to generically prompted personas. Credibility scores were also higher for signal-conditioned personas (4.41 vs. 3.91). This result provides support for SAPIENT's central architectural claim: the sentinel-to-simulation coupling appears to be a meaningful factor in simulation specificity. Simply presenting a corporate announcement to an LLM and asking for reactions produces a different kind of discourse than grounding the same request in the real-world signal environment.

The Jaccard index between the two theme sets was low (0.259), meaning that signal-conditioned and generic personas largely surfaced different concerns. Condition A generated 140 themes absent from Condition B, including context-specific items such as "advisory_board_governance," "audit_requirements," and "baseline_verification," themes that directly reflect the information provided in S_t. Condition B, lacking this context, produced 169 unique themes of a more generic nature. This pattern supports the claim that sentinel-derived context steers simulation discourse toward operationally relevant territory rather than generic corporate criticism.

One counterintuitive finding deserves comment: the generic condition produced slightly *more* total themes (277 vs. 248), suggesting that the absence of grounding information may lead to broader but less focused speculation. Whether breadth or focus is preferable depends on the use case; for crisis preparation, focused themes tied to real-world signals are arguably more actionable.

6.7. Experiment 3: Multilingual Consistency

To assess cross-lingual robustness, the same scenario was run in English and Turkish with demographically matched persona sets ($K = 10$ each). Table 9 summarizes the results.

Table 9. Multilingual comparison: English vs. Turkish ($K = 10$, $n = 8$). Same scenario, matched persona demographics.

Metric	English	Turkish	*p*	*d*
Sentiment (mean ± std)	4.22 ± 0.84	3.76 ± 0.83	0.001	0.55
Credibility	4.35 ± 0.69	3.80 ± 0.73	–	–
Unique themes	246	301	–	–
Mean cosine similarity	0.185	0.149	–	–
Persona-level corr.	$r = 0.808$, $p = 0.015$		–	–

Turkish-language sessions produced significantly lower sentiment ($t = 3.499$, $p = 0.001$, BH-adjusted $p = 0.002$, $d = 0.55$) and lower credibility scores than English sessions. The effect size is medium, indicating a substantive language-linked asymmetry rather than random variation. Turkish sessions generated more unique themes (301 vs. 246) and lower mean cosine similarity (0.149 vs. 0.185), indicating more diverse but less convergent responses. Critically, persona-level sentiment correlation between languages was high ($r = 0.808$, $p = 0.015$, BH-adjusted $p = 0.018$): the relative ordering of stakeholder reactions was preserved across languages even as absolute values shifted.

These findings align with concerns raised in the limitations discussion (Section 7.4): LLM behavior is not language-invariant, and Turkish-language outputs may reflect different training data distributions, cultural framing biases, or reduced model confidence in non-English contexts. One plausible reading is that the Turkish-language training corpus reflects a media environment with higher baseline skepticism toward corporate sustainability claims, a pattern that is broadly consistent with survey evidence on institutional trust in developing economies. An alternative explanation is that the asymmetry stems from reduced model fluency in non-English generation, producing more hedged and cautious outputs that register as lower sentiment. Disentangling these two mechanisms, cultural calibration versus linguistic confidence, is a priority for Stage 4 testing. For practitioners, this result means that multilingual deployments of SAPIENT require language-specific calibration rather than the assumption that a validated English configuration transfers directly.

6.8. Experiment 4: Variance Collapse Countermeasures

The system proposes two countermeasures against variance collapse (Section 3.3.3): stratified temperature assignment across personas and adversarial persona injection. To evaluate these, three conditions were tested ($K = 10$ each): (A) uniform temperature $\tau = 0.7$; (B) stratified $\tau \in [0.6, 1.0]$; and (C) stratified temperature with one contrarian persona replacing the least differentiated standard persona. Table 10 reports the results.

Table 10. Variance collapse countermeasures ($K = 10$, $n = 8$). Lower cosine similarity indicates more diverse responses. Sentiment std measures overall opinion spread across all responses in each condition ($N = 80$).

Condition	Cos. Sim.	Sent. Std	Themes	Flagged
A (Uniform $\tau = 0.7$)	0.186 ± 0.018	0.89	233	0/10
B (Stratified τ)	0.185 ± 0.013	0.84	234	0/10
C (Stratified + Adversarial)	0.184 ± 0.016	0.93	261	0/10

No condition triggered variance collapse (all well below $\delta = 0.85$), which is an encouraging baseline result but limits what we can say about countermeasure effectiveness, since the problem they address did not manifest under these experimental conditions. The adversarial condition (C) produced the lowest mean cosine similarity (0.184 vs. 0.186), though the difference was small. The adversarial condition also showed the highest within-run sentiment standard deviation (0.93 vs. 0.89) and generated the most unique themes (261 vs. 233), suggesting that injecting a contrarian persona modestly increases opinion diversity. The overall pattern indicates that Claude Sonnet 4 produces sufficiently diverse persona responses even without explicit countermeasures under these experimental conditions; whether the countermeasures become more consequential under conditions that more readily induce collapse, such as low-controversy topics or highly similar persona specifications, remains to be tested.

Temperature stratification (B vs. A) had minimal effect on cosine similarity or thematic diversity in this production run. The combination of both countermeasures (C) appears to operate primarily through the adversarial channel, broadening the thematic space and widening the sentiment range.

6.9. Experiment 5: Greenhushing Risk Assessment

To assess the framework's applicability beyond greenwashing, we tested Scenario 2 (Section 4.2): a financial services firm with genuine ESG improvements faces the decision of whether to publish a detailed ESG report or remain silent while a competitor discloses. The persona set was adapted for financial stakeholders (institutional investor, ESG analyst, financial journalist, retail investor, regulator, chief sustainability officer, competitor analyst, sustainability academic). Two conditions were tested with $K = 10$ runs each; Table 11 presents the results.

Table 11. Greenhushing scenario results (Scenario 2): stakeholder reactions to ESG disclosure vs. strategic silence. $K = 10$ runs, $n = 8$ finance-focused personas. Model: Claude Sonnet 4.

Strategy	Sentiment	Credibility	Themes	Cos. Sim.
A (Disclose)	4.46 ± 0.67	4.31 ± 0.68	272	0.194
B (Silent)	2.36 ± 0.48	5.97 ± 0.16	224	0.256

The most notable finding is the asymmetry between sentiment and credibility under the silence condition. Strategic silence produced low sentiment (2.36), reflecting negative stakeholder reactions, but high credibility (5.97), the highest credibility score observed in any condition across all eight experiments. This pattern admits two complementary interpretations. First, financial stakeholders may perceive silence as a form of restraint that signals confidence, in contrast to disclosure that could be seen as performative. Second, the near-zero standard deviation on credibility under silence (0.16) indicates strong cross-persona consensus: all eight personas, from institutional investor to sustainability academic, converged on the view that silence is credible even when it is unwelcome.

For practitioners navigating the disclosure dilemma, this result implies that the reputational calculus is not simply "disclose or risk appearing evasive"; silence carries its own distinct reputational signature that should be evaluated on both dimensions. The successful application of the AFG protocol to this scenario, using a different persona set and a qualitatively different decision context, provides initial evidence for the framework's transferability beyond its original greenwashing domain.

6.10. Experiment 6: Crisis Communication Simulation

Scenario 4 (Section 4.4) tests the framework under high-urgency conditions: an investigation alleges that a company's "sustainably sourced" product line relies on suppliers with documented violations. Three response strategies were compared with $K = 10$ runs each, using the original eight-persona set from Scenario 1 to maintain cross-scenario comparability. Table 12 summarizes the results.

Table 12. Crisis communication scenario results (Scenario 4): stakeholder reactions to three response strategies following a supply chain scandal. $K = 10$, $n = 8$ personas. Model: Claude Sonnet 4.

Response Strategy	Sentiment	Credibility	Themes
A (Immediate apology)	3.98 ± 1.07	3.98 ± 0.81	292
B (Factual rebuttal)	2.65 ± 0.48	3.09 ± 0.62	250
C (Delayed response)	2.00 ± 0.16	2.06 ± 0.24	260

The results produced a clear ordinal ranking: apology outperformed rebuttal, which outperformed delay, on both sentiment and credibility. This ordering is consistent with established crisis communication theory, particularly the finding that organizations seen as taking responsibility early tend to recover trust more quickly than those that deflect or delay [1].

The delayed response condition is particularly instructive. The extremely low standard deviations for both sentiment (0.16) and credibility (0.24) indicate near-unanimous persona agreement that delay is an unacceptable strategy. This cross-persona consensus emerged independently across all 10 runs, representing one of the strongest convergence signals in the dataset. The immediate apology condition, by contrast, generated the highest thematic diversity (292 unique themes), suggesting that a proactive response opens a richer deliberative space among stakeholders. The apology condition also exhibited the highest sentiment standard deviation (1.07), indicating that while the average reaction was moderately positive, individual personas diverged considerably. The activist persona remained skeptical even of immediate apology, while the employee and regulator personas responded more favorably. This within-condition heterogeneity is itself a useful output: it identifies which stakeholder segments require differentiated communication even under the most accommodating response strategy.

6.11. Experiment 7: Cross-Model Comparison

To assess the degree to which the reported findings depend on the choice of LLM backend, we repeated Experiment 1 Variant C (accountability framing) using two additional models, GPT-4o (OpenAI) and Gemini 2.5 Flash (Google), with identical prompts, persona specifications, signal state, and evaluation metrics. Each model was run with $K = 20$ independent sessions. Tables 13–15 report the three-way comparison.

Table 13. Three-way cross-model comparison on Scenario 1, Variant C. Same prompts, personas, and signal state across all three backends. $K = 20$ per model.

Metric	Claude Sonnet 4	GPT-4o	Gemini 2.5 Flash
Sentiment (mean $\pm$ std)	4.27 ± 0.80	4.49 ± 0.71	4.62 ± 1.12
Credibility (mean $\pm$ std)	4.39 ± 0.72	4.63 ± 0.72	4.57 ± 1.20
Unique themes	368	325	524
Mean cosine similarity	0.184	0.279	0.203

Table 14. Pairwise persona-level sentiment correlation (Pearson r) and sentiment t-tests across three LLM backends. BH-adjusted p-values in parentheses.

Model pair	Pearson r	p	Sentiment t	p
Claude - GPT-4o	0.935	0.0006 (0.002)	−2.576	0.010 (0.016)
Claude - Gemini	0.924	0.0010 (0.002)	−3.206	0.002 (0.004)
GPT-4o - Gemini	0.960	0.0002 (0.001)	−1.252	0.212 (n.s.)

Table 15. Persona-level mean sentiment across three LLM backends ($K = 20$ runs each). P1 = Activist, P2 = ESG Investor, P3 = Consumer, P4 = Journalist, P5 = Regulator, P6 = Employee, P7 = Competitor, and P8 = Academic.

ID	Role	Claude	GPT-4o	Gemini
P1	Activist	2.80	3.00	2.10
P2	ESG Inv.	4.90	5.00	5.25
P3	Consumer	4.25	4.90	5.10
P4	Journalist	3.95	4.20	4.50
P5	Regulator	4.85	4.65	4.85
P6	Employee	4.80	5.00	5.50
P7	Competitor	4.25	4.40	5.00
P8	Academic	4.35	4.75	4.65

The central finding is the consistency of persona-level behavioral rankings across all three backends. Pairwise Pearson correlations ranged from $r = 0.924$ (Claude–Gemini) to $r = 0.960$ (GPT-4o–Gemini), all significant after BH correction ($p_{\text{adj}} < 0.003$). The activist persona (P1) received the lowest sentiment from every model; the employee (P6) the highest. This structural convergence across three architecturally distinct models, built by different organizations, trained on different corpora, and employing different alignment procedures, provides consistent evidence that the AFG protocol captures behavioral regularities that are not artifacts of any single model.

At the surface level, the three models diverged in characteristic ways. Claude produced the most conservative sentiment (4.27), followed by GPT-4o (4.49) and Gemini (4.62). The difference between GPT-4o and Gemini was not statistically significant ($p = 0.212$), suggesting that these two models share a more optimistic response baseline relative to Claude. Gemini generated the largest number of unique themes (524 vs. 368 and 325) and the highest response variance (std = 1.12), while its mean cosine similarity (0.203) remained close to Claude (0.184) and well below the collapse threshold. Pairwise theme Jaccard indices were uniformly low (range: 0.059–0.077), confirming that the three models express shared structural concerns through substantially different vocabulary.

These results refine the interpretation of model dependence. The *structure* of stakeholder differentiation, which personas react most and least favorably and how they rank relative to each other, proved stable across backends. The *surface* characteristics, absolute sentiment levels, lexical choices, and thematic granularity, varied in model-specific ways. For practitioners, this means that SAPIENT's qualitative conclusions about relative stakeholder positioning are portable across backends, while absolute values should be interpreted within the context of the specific model used.

6.12. Experiment 8: Prompt Sensitivity

Bisbee et al. [25] documented that minor prompt wording changes can alter synthetic survey distributions. To test the robustness of the AFG protocol to surface-level prompt variation, we created a semantically equivalent paraphrase of the Variant C stimulus: same

factual content, different sentence structure and word order, comparable length. Both versions were run with $K = 20$ using Claude Sonnet 4. Table 16 presents the results.

Table 16. Prompt sensitivity test: original vs. semantically equivalent paraphrase of Scenario 1 Variant C. $K = 20$ runs per version, $n = 8$ personas. BH-adjusted p-values in parentheses. $^{***}p < 0.001$.

Metric	Original	Paraphrase	p	Sig.
Sentiment mean	4.33	4.49	0.061	n.s.
Credibility mean	4.34	4.66	<0.001	***
Theme Jaccard	0.300		–	–

Sentiment, the primary outcome metric, did not differ significantly between the original and paraphrased versions ($p = 0.061$, BH-adjusted $p = 0.067$), confirming that the AFG protocol's principal measure is stable under surface-level prompt reformulation. The theme Jaccard index (0.300) was substantially higher than the cross-model comparison (0.059–0.077), indicating that prompt paraphrasing preserves thematic content considerably more than model switching does.

The credibility result, however, changed markedly when the sample was expanded from $K = 5$ (reported in the previous version) to $K = 20$. At $K = 5$, the credibility difference was borderline ($p = 0.034$); at $K = 20$ it became highly significant ($t = -4.27$, $p < 0.001$, BH-adjusted $p < 0.001$). This indicates that the effect is not noise but a systematic sensitivity: surface-level prompt reformulation affects perceived credibility more than it affects overall sentiment. One plausible interpretation is that credibility judgments engage closer reading of specific linguistic cues (word choice, hedging language, specificity of claims), while sentiment captures a more global attitudinal orientation that is less sensitive to phrasing. For practitioners, this means that the *relative ranking* of scenarios and strategies can be trusted across prompt formulations, but absolute credibility scores should be interpreted with awareness that stimulus wording introduces a non-trivial source of variation.

6.13. Experiment 9: Pilot Human Validation

To provide initial external validation of the AFG protocol's outputs against real human responses, we conducted a preregistered pilot comparison study (https://doi.org/10.17605/OSF.IO/4KFDC) using Prolific [48], a widely used academic research platform whose participant pool has been shown to produce data quality comparable to or exceeding traditional alternatives. The platform choice was deliberate: SAPIENT simulates general stakeholder reactions, not expert judgments, and a paid, pre-screened general-population sample is the methodologically appropriate comparison group for the framework's intended use case.

Design. The study used a within-subject (repeated measures) design with three conditions corresponding to the three framing variants tested in Experiment 1: targets (A), progress (B), and accountability (C). A within-subjects design was chosen to maximize statistical power at the pilot sample size; between-subjects replication is planned for Stage 2 to assess whether the credibility effect persists without explicit comparison opportunity. Participants read all three variants in randomized order (Qualtrics block randomization with the "Evenly Present Elements" option, producing approximately equal representation of all six possible orderings) and rated each on sentiment (1–7) and credibility (1–7), followed by an open-ended question. Post-exposure items included forced-choice most/least credible variant selection, demographics, and environmental engagement. An attention check was embedded between stimulus blocks. The study was preregistered with three directional hypotheses: H9a (human ranking matches SAPIENT ranking for both metrics), H9b (accountability rated higher than targets in paired comparison), and H9c (open-ended theme overlap exceeds soft Jaccard > 0.10 in at least two variants). The preregistered primary

comparison was H9b for credibility (C vs. A); H9a and H9c were secondary. The primary result ($p = 0.004$) survives Bonferroni correction for three tests (adjusted $p = 0.012$).

Participants. 63 responses were recorded in Qualtrics from participants recruited through Prolific. Nine were excluded: two researcher test entries recorded during survey configuration, three duplicate re-entries from participants who re-entered the survey (one of which also fell below the preregistered four-minute speed threshold), three incomplete submissions that did not complete all three framing evaluations, and one submission removed during quality review. This yielded $n = 54$ responses from 54 unique participants with complete paired ratings on all three variants. Of these, 34 also completed the post-exposure comparative items (forced-choice selections and demographics). Participants were compensated at £2.50 (approximately $12.60/hour). Among the 34 who completed demographic items: 38% aged 18–24, 47% aged 25–34, 15% aged 35–44; 71% male, 29% female. Mean environmental engagement was 4.09/7.

Results. Table 17 summarizes the comparison between SAPIENT and human responses across all three variants.

Table 17. Pilot human validation: comparison of human responses with SAPIENT outputs (Experiment 1, $K = 20$). Sentiment and credibility on 1–7 Likert scale. Values shown as mean ± SD. Variance ratio = human variance/SAPIENT variance.

Metric	Variant	SAPIENT	Human	Δ	Var. Ratio
Sentiment	Targets	3.96 ± 1.01	5.41 ± 1.42	+1.45	1.98
	Progress	4.08 ± 0.95	5.20 ± 1.41	+1.12	2.19
	Accountability	4.24 ± 0.82	5.06 ± 1.53	+0.82	3.50
Credibility	Targets	3.81 ± 0.70	4.20 ± 1.56	+0.39	4.96
	Progress	3.85 ± 0.73	4.61 ± 1.53	+0.76	4.42
	Accountability	4.36 ± 0.74	5.04 ± 1.52	+0.68	4.20

The results showed partial but meaningful external replication (Table 17 and Figure 3). For credibility, the predicted ordering was reproduced: accountability was rated highest, followed by progress and targets ($A = 4.20$, $B = 4.61$, $C = 5.04$). The preregistered paired comparison confirmed that the accountability framing was rated more credible than the targets framing (Wilcoxon one-tailed $p = 0.004$, Cohen's $d = 0.40$, $n = 54$). This confirms H9b for the credibility metric. Among the 34 participants who also completed the comparative items, the accountability version was selected as the most credible by 53% (18 of 34), followed by progress at 32% (11 of 34), while the targets version was selected as least credible by 62% (21 of 34).

For sentiment, the pattern was not replicated. The preregistered one-tailed test (H9b: C > A) was not supported (Wilcoxon $p = 0.86$, direction opposite to prediction). For completeness, the overall within-subject test was also non-significant (Friedman $\chi^2 = 1.98$, $p = 0.37$; Wilcoxon two-tailed C vs. A: $p = 0.29$). Human participants rated all three variants similarly ($A = 5.41$, $B = 5.20$, $C = 5.06$). This dissociation between credibility and sentiment carries substantive interest. Real participants appeared to separate their overall affective reaction ("how do I feel about this company?") from their epistemic trust judgment ("do I believe this announcement?"). Communication framing shifted credibility but not general sentiment. A contributing factor is a ceiling effect in sentiment ratings: 61% of participants rated the targets variant 6 or 7 out of 7, 56% for progress, and 50% for accountability. This compression near the scale ceiling limits the capacity for framing to produce detectable sentiment differentiation. LLM personas, by contrast, produced differentiated responses on both dimensions, operating in a lower and wider range of the scale. This suggests that prompt-conditioned agents process framing cues more analytically

than real respondents, generating metric-specific sensitivity where humans exhibit a flatter affective baseline. While the observed sentiment direction was nominally reversed relative to SAPIENT predictions, the effect size was small ($d = 0.15$). Detecting an effect of this magnitude would require approximately 350 participants (power = 0.80, two-tailed). The present pilot therefore cannot distinguish between a true null and a small reversed effect; the credibility dimension provides stronger evidence that the framework captures genuine framing effects for evaluative judgments.

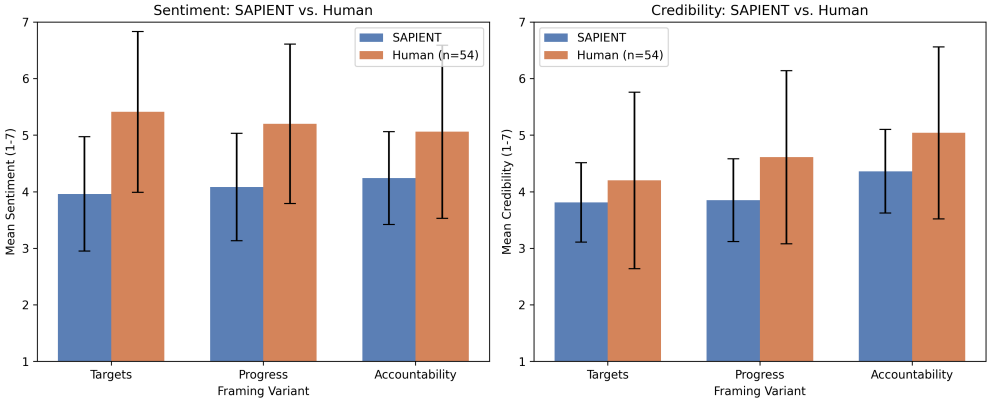

Figure 3. Comparison of SAPIENT and human responses across three framing variants ($n = 54$). (**Left**): sentiment (1–7); (**right**): credibility (1–7). Error bars denote ± 1 standard deviation. Human credibility rankings replicate the SAPIENT ordering (A < B < C), while sentiment shows no significant framing differentiation.

Open-ended theme overlap reached the preregistered threshold (>0.10) in two of three variants (soft Jaccard: $B = 0.19$, $C = 0.14$; $A = 0.07$ fell below the threshold), supporting H9c as specified (at least two of three). Human respondents surfaced themes consistent with the SAPIENT outputs: data credibility and supply chain gaps for the progress variant, and supplier accountability and corrective actions for accountability.

Across all six metric-by-variant comparisons, human response variance exceeded SAPIENT variance by a factor of 2.0–5.0, consistent with findings in the LLM simulation literature that synthetic outputs exhibit systematically compressed dispersion relative to human samples [25].

6.14. Runtime and Cost Profile

All API calls were instrumented with per-call token logging and latency measurement through the unified LLM abstraction layer. Table 18 summarizes the aggregate resource consumption.

Experiments were executed on a standard workstation (Python 3.13, Windows) with API access over standard internet connections. Claude Sonnet 4 calls averaged approximately 10.2 s latency; GPT-4o calls averaged 6.5 s; and Gemini 2.5 Flash calls averaged 12.7 s, which includes internal reasoning overhead. GPT-4o required reduced concurrency due to a 30K tokens-per-minute rate limit, which increased wall-clock time for Experiment 7. No GPU resources were required; all computation was API-based. The eight experiments, comprising 280 AFG sessions with 2240 persona-level responses across three scenarios and three LLM backends, were completed for under $23 in approximately 44 min.

Table 18. Aggregate runtime and API usage across all eight experiments. Costs computed using published API pricing (Claude Sonnet 4: $3.00/$15.00 per MTok input/output; GPT-4o: $2.50/$10.00 per MTok; Gemini 2.5 Flash: $0.15/$0.60 per MTok).

Metric	Value
Total API calls	4288
Total input tokens	2,654,712
Total output tokens	889,720
Total tokens	3,544,432
Estimated cost (USD)	$22.29
Total elapsed time	43.8 min
Concurrency (Claude)	3 sessions
Concurrency (GPT-4o)	1 session (rate-limited)
Concurrency (Gemini)	2 sessions

6.15. Summary of Findings

Across nine experiments covering 280 independent AFG runs across 20 experimental conditions, three application scenarios, three LLM backends, and a pilot human validation study with 54 participants, the results provide the following observations:

1. **H1 supported:** The AFG protocol produces role-differentiated, non-collapsing outputs that respond to framing variation in expected directions. Persona behavioral signatures were consistent across runs (within-persona sentiment std: 0.30–0.89) and differentiated across roles (cross-persona sentiment range: 2.00–4.90 within conditions).
2. **H2 supported:** Sentinel-derived signal states produced more focused and contextually specific discourse than generic prompting ($p = 0.012$, BH-adjusted $p = 0.017$), with signal-conditioned sessions generating context-specific themes traceable to information in S_t.
3. **H3 supported:** Cross-lingual sessions revealed a systematic sentiment offset between English and Turkish ($p = 0.001$, BH-adjusted $p = 0.002$), while persona-level rank ordering was preserved ($r = 0.81$, $p = 0.015$, BH-adjusted $p = 0.018$), confirming that the protocol's structural differentiation transfers across languages even as absolute values shift.
4. **H4 supported under the tested scenario:** Cross-model comparison across three architecturally distinct backends showed high structural consistency: all pairwise persona correlations exceeded $r = 0.92$ ($p_{\text{adj}} < 0.003$). Absolute sentiment values varied across models ($p < 0.02$ for Claude vs. GPT-4o and Claude vs. Gemini; $p = 0.212$, n.s., for GPT-4o vs. Gemini), while lexical expression diverged substantially (theme Jaccard range: 0.059–0.077).
5. **Scenario generalizability:** The framework produced substantively distinct and contextually appropriate results across greenwashing, greenhushing, and crisis communication scenarios.
6. **Prompt robustness:** Sentiment, the primary metric, was robust to surface-level prompt paraphrasing ($p = 0.061$, n.s.). Credibility proved systematically sensitive to prompt wording ($p < 0.001$), an effect that strengthened when replicated at $K = 20$ (from $p = 0.034$ at $K = 5$), indicating that perceived credibility engages finer-grained linguistic processing than overall sentiment.
7. **H9b partially supported (Pilot human validation):** In a preregistered pilot study with 54 human participants recruited through Prolific, the predicted credibility ranking was reproduced (A < B < C, Wilcoxon $p = 0.004$, $d = 0.40$). Forced-choice data confirmed that the accountability framing was selected as most credible by 53% of participants. Sentiment ranking was not replicated: human participants rated all

three variants similarly (Friedman $p = 0.37$, n.s.), suggesting that communication framing shifts credibility judgments more than overall affective reactions. A ceiling effect in sentiment ratings (50–61% of participants scored 6 or 7) further compressed the available range for framing differentiation. Open-ended theme overlap exceeded the preregistered threshold in two of three variants (soft Jaccard ≥ 0.14 for progress and accountability; targets fell below at 0.07). Human response variance exceeded SAPIENT variance across all conditions (ratio: 2.0–5.0), consistent with the LLM simulation literature.

Benjamini–Hochberg correction was applied across all eleven inferential tests in Experiments 1–8; nine survived FDR control at $\alpha = 0.05$ (Table 4). Experiments 1–8 are exploratory and were not pre-registered. Experiment 9 was preregistered at OSF prior to data collection (https://doi.org/10.17605/OSF.IO/4KFDC). The evidence provides initial external validation of the framework's credibility predictions, while identifying sentiment calibration as a priority for future work. Full distributional calibration against human data (Stage 2) and prospective organizational deployment (Stage 5) remain necessary to assess real-world fidelity.

7. Discussion

7.1. Theoretical Implications

SAPIENT contributes to agentic AI for social simulation [34] by proposing sentinel-to-simulation coupling, an integration pattern not clearly articulated in the adjacent literature we reviewed. Prior LLM simulation work operates without real-time data grounding [14]; sentinel systems lack forward-looking analysis. Our architecture suggests these capabilities are more valuable combined than in isolation.

The signal state formalization provides a reproducibility anchor absent from most LLM simulation work. The AFG protocol contributes beyond corporate reputation: the core idea of repeated sessions with variance reporting and mandatory human review applies wherever LLM-based qualitative exploration is used. By treating simulation as an experimental instrument, the protocol makes the scope of its claims explicit: in silico groups are hypothesis generators, not opinion certifiers.

The system extends "homo silicus" [16] from economics to communication science. Reputation judgments are shaped by contextual factors, such as prior trust, media priming, and social identity, that pose harder simulation challenges than economic heuristics.

The cross-model comparison (Experiment 7) contributes an additional dimension to this theoretical contribution. The finding that persona-level behavioral rankings are preserved across three architecturally distinct LLM backends (all pairwise $r > 0.92$, $p_{\text{adj}} < 0.003$) while absolute values and lexical expression diverge suggests that, under the tested conditions, the AFG protocol captures a form of structural consistency that transcends any single model's training data or alignment procedure. The convergence across models built by three different organizations (Anthropic, OpenAI, and Google) provides triangulation evidence that the captured behavioral patterns reflect properties of the prompt–persona interaction rather than idiosyncratic model artifacts. This result is based on one scenario and one framing variant; whether the pattern holds across other scenarios remains a priority for future testing.

The successful application across three scenarios within the corporate reputation and sustainability communication domain (greenwashing backlash, greenhushing risk, and crisis communication) provides initial evidence that the framework is not limited to a single use case. The qualitatively different patterns observed across scenarios, such as the sentiment–credibility asymmetry unique to the greenhushing condition and the strong consensus against delayed crisis response, suggest that the protocol captures scenario-

specific dynamics rather than producing generic outputs. Whether SAPIENT generalizes to domains beyond corporate reputation, such as public health or policy communication, remains untested.

Table 19 positions SAPIENT relative to adjacent systems along eight architectural dimensions to clarify the novelty claim.

Table 19. Architectural comparison of SAPIENT with adjacent LLM-based simulation and monitoring platforms. Checkmarks indicate features present in published descriptions; dashes indicate features absent or not discussed.

Feature	SAPIENT	GenSim [20]	Rumor-Sphere [21]	Crisis-Bench [19]	Dual-Mind [22]	Park et al. [14]
Real-time sentinel	✓	–	Partial	–	–	–
Signal state formal.	✓	–	–	–	–	–
Repeated-run variance	✓	–	✓	–	–	–
Human review gate	✓	–	–	–	–	–
Bidirectional coupling	✓	–	–	–	–	–
Cross-lingual support	✓	–	–	Partial	–	–
Governance framework	✓	–	–	–	–	–
Calibration hooks	✓	–	–	–	–	–
Intended use	Reputation intelligence	Social sim. at scale	Rumor propagation	Crisis benchm.	Narrative tracking	Social behavior

7.2. Positioning Relative to Calibration Literature

The calibration and distributional literature (Section 2.7) raises a fundamental question: under what conditions can SAPIENT's simulation outputs be trusted? Our answer is deliberately conservative. Hullman et al. [23] argue that heuristic validation is insufficient without statistical calibration; we agree, and this is why SAPIENT frames simulation outputs as qualitative hypotheses rather than quantitative estimates. The AFG protocol's repeated runs and variance reporting provide a form of internal consistency checking, but they cannot substitute for external calibration against human data.

SubPOP [24] shows that fine-tuning substantially improves distributional agreement for survey-style responses. SAPIENT's current design relies on prompt-based persona conditioning, which Santurkar et al. [35] and Bisbee et al. [25] have shown has limited calibration power. We view fine-tuning and PPI-style correction [37,38] as future integration paths: the architecture's modular design allows the persona construction agent to be backed by either a prompted base model or a fine-tuned variant, and the calibration agent includes hooks for PPI correction when human comparison data are available. Stage 2 of the evaluation plan (Section 5) is designed to quantify the gap between prompt-based and fine-tuned approaches in SAPIENT's specific operating context.

The findings of PolyPersona [36] on variance collapse and drift reinforce the need for the diversity countermeasures specified in Section 3.3. Experiment 4 (Section 6.8) found that variance collapse did not manifest under the tested conditions, with all runs remaining well below the $\delta = 0.85$ threshold. Adversarial persona injection produced a modest increase in thematic diversity and opinion spread, though the differences in cosine similarity across conditions were small. Whether these countermeasures become more consequential under more challenging conditions, such as low-controversy topics or extended multi-turn sessions, requires further investigation.

7.3. Practical Implications

SAPIENT offers a pathway from reactive monitoring to proactive scenario planning. The most immediate benefit lies in campaign pre-testing, which compresses iteration from weeks to hours, reserving traditional methods for final validation. The bidirectional

sentinel–simulation coupling has a practical consequence worth noting: AFG-discovered counter-narratives can be immediately tracked by the sentinel layer, creating a learning loop where observational capabilities evolve from experimental findings.

For organizations navigating the greenwashing–greenhushing tension, the system structures assessment of both disclosure and non-disclosure risks, which is particularly relevant as the EU CSRD [39] makes greenhushing increasingly untenable.

The revision process itself provided an unplanned test of the framework's modularity. Integrating two additional LLM backends (GPT-4o and Gemini 2.5 Flash) and two new application scenarios (greenhushing and crisis communication) required only configuration-level changes: adding entries to the scenario and persona configuration files, specifying new experiment parameters, and routing API calls through the provider-agnostic abstraction layer. No modifications were made to the core AFG protocol, the metrics computation pipeline, or the analysis framework. The eight experiments, spanning 280 runs and 4288 API calls, were completed in under 44 min at a total cost of $22.29.

7.4. Limitations

Several limitations warrant acknowledgment, organized by severity.

External validation. Experiment 9 provided the first external contact between SAPIENT outputs and real human responses, yielding partial replication: the predicted credibility ranking was confirmed ($p = 0.004$), but the sentiment ranking was not reproduced. This partial pattern narrows the external validation gap but does not close it. Human validation was conducted for one scenario (EcoVista, ESG communication); generalizability to the other experimental scenarios remains untested. Single-scenario validation is appropriate for a pilot study, and the preregistration specified the greenwashing scenario as the target due to its moderate complexity. Multi-scenario human validation is specified for Stage 2. Domain-specific stakeholder validation, subgroup-level distributional comparison, and calibration against focus group data remain necessary. The credibility–sentiment dissociation identified in Experiment 9 also suggests that LLM personas process framing cues more analytically than real participants, producing metric-specific sensitivity where humans exhibit a flatter affective baseline. Sentiment calibration, potentially through fine-tuning or PPI-style correction [37], is an immediate priority.

Presentation order. In Experiment 9, the Qualtrics randomizer counterbalanced the order of the three variants across participants. However, the CSV column structure records responses in a fixed schema (A, B, C), and we cannot independently verify the degree of balance achieved across the six possible orderings. We note that a pure order effect would predict monotonic shifts in the same direction across measures. The observed pattern is inconsistent with this: sentiment decreased across the A–B–C column positions ($5.41 \rightarrow 5.20 \rightarrow 5.06$) while credibility increased ($4.20 \rightarrow 4.61 \rightarrow 5.04$). This divergence suggests that stimulus content, rather than position, drove the credibility differentiation. A fully counterbalanced between-subjects replication is planned for Stage 2.

What SAPIENT cannot justify. To avoid overclaiming, we state explicitly what the framework's outputs should *not* be used for: (a) estimating actual proportions of opinion in real populations, since simulation outputs are not population-level estimates; (b) making individual-level predictions about specific persons or organizations; (c) serving as evidence in regulatory, legal, or compliance contexts; and (d) replacing human qualitative research for decisions with significant consequences. Outputs that appear plausible or internally stable do not thereby become externally valid. The framework generates qualitative hypotheses for human review, not certified findings.

Calibration gap. The synthetic population inherits LLM biases; non-English, elderly, and Global South populations remain underrepresented [18]. Prompt-based persona

conditioning has known limitations for reproducing distributional tails and covariance structure [25]. Until fine-tuning or PPI correction is integrated, simulation outputs should be treated as approximate guides rather than calibrated estimates.

Multilingual limitations. LLMs exhibit language-linked asymmetries in safety behavior, stereotype expression, and response diversity. Experiment 3 revealed a systematic sentiment offset between English and Turkish sessions ($p = 0.001$), though persona-level rankings were preserved. Whether this offset reflects cultural differences in the training corpus or reduced model fluency in non-English generation remains an open question. Cross-lingual calibration methods are under-developed in the literature; Stage 4 testing is designed to characterize this gap.

Prompt sensitivity. Experiment 8 showed that sentiment was stable under surface-level paraphrasing ($p = 0.061$, n.s.), but credibility proved systematically sensitive ($p < 0.001$). This effect strengthened when replicated at $K = 20$ (from $p = 0.034$ at $K = 5$). This indicates that credibility judgments engage finer-grained linguistic processing than overall sentiment, and that scenario formulation choices can influence simulation outputs in metric-specific ways. A full factorial ablation across prompt dimensions (persona formatting, output schema, scenario translation) would require a dedicated study and is identified as a priority for future work.

Human review gate. The human review component has not been independently validated. The authors served as reviewers during the experiments, introducing potential confirmation bias. An independent inter-rater agreement study is specified as a future evaluation component.

Security. The defense layers specified in Section 3.5.3 notwithstanding, adversarial tactics evolve continuously. No defense mechanism can be assumed to remain effective indefinitely.

Statistical scope. Experiments 1–8 are exploratory and were not pre-registered. Experiment 9 was preregistered at OSF prior to data collection (https://doi.org/10.17605/OSF.IO/4KFDC), with pre-specified hypotheses, analysis plan, and exclusion criteria. Benjamini–Hochberg correction for multiple comparisons was applied across all eleven inferential tests in Experiments 1–8 (Section 6.3); nine of eleven survived FDR control at $\alpha = 0.05$. Under the more conservative Holm–Bonferroni criterion, five tests remain significant while four become borderline. Results from Experiments 1–8 should be interpreted as preliminary indications rather than confirmatory evidence.

7.5. Ethical Considerations

The misuse risks and non-use contexts specified in Section 3.5.6 address the most direct ethical concerns. Two additional points merit discussion. First, monitoring public discourse at scale raises expectation questions even when only aggregated data are retained; the boundary between public opinion analysis and surveillance depends on context and intent, and organizational policies must complement technical safeguards. Second, the consent question applies to monitored content: while social media posts are typically public, individuals may not anticipate their text being processed by AI systems for corporate decision support. Data minimization and aggregation-level analysis address this by design, but regulatory evolution (particularly under GDPR and KVKK) may impose additional requirements.

8. Conclusions

This paper proposed SAPIENT, a multi-agent system that integrates sentinel-based media monitoring with LLM-driven synthetic population simulation for corporate reputation intelligence. The key architectural idea is connecting observational signals to experimental

qualitative exploration through a formalized signal state, while maintaining calibration hooks, variance reporting, and human review gates to bound error and prevent unsafe use. The Agentic Focus Group protocol provides a repeatable methodology for in silico qualitative research with defined validity limits: simulation outputs are framed as hypotheses, not measurements.

The experiments expanded the empirical basis to three application scenarios, three LLM backends, a prompt sensitivity test, and a preregistered pilot human validation study, totaling 280 AFG runs across 20 conditions plus 54 human participant responses. Signal-state conditioning improved simulation specificity ($p = 0.012$). The AFG protocol produced role-differentiated, non-collapsing persona outputs across scenarios. Cross-model comparison revealed consistent persona differentiation across three backends (all pairwise $r > 0.92$, $p_{\text{adj}} < 0.003$), and the primary outcome metric was robust to prompt variation ($p = 0.061$, n.s.). All significant results from Experiments 1–8 survived Benjamini–Hochberg correction.

The cross-model finding has particular significance for framework design. Three architecturally distinct LLMs maintained consistent persona rankings while differing in absolute sentiment and thematic vocabulary, supporting cross-backend portability. This consistency has been demonstrated on one scenario; further replication is needed.

Experiment 9, a preregistered study with 54 human participants recruited through Prolific, provided the first direct comparison between SAPIENT outputs and real human responses. The predicted credibility ranking was reproduced ($p = 0.004$), and forced-choice selections confirmed that 53% of participants perceived accountability framing as most credible. The sentiment ranking was not replicated: human participants assigned similar overall sentiment across variants, suggesting that communication framing affects credibility judgments more than general affective reactions. Human response variance exceeded SAPIENT variance across all conditions (ratio: 2.0–5.0), consistent with the broader LLM simulation literature [25]. Full distributional calibration against human data (Stage 2), including subgroup-level comparison and PPI-style correction, remains the immediate next step.

Further directions include distributional calibration against human focus group data, domain-specific fine-tuning for sustainability communication, cross-lingual calibration for non-English populations, independent human review gate validation, and extension to public health and policy contexts.

Author Contributions: Conceptualization, A.O. and S.B.O.; methodology, A.O. and S.B.O.; software, A.O.; validation, A.O. and S.B.O.; formal analysis, A.O.; investigation, A.O. and S.B.O.; writing—original draft preparation, A.O. and S.B.O.; writing—review and editing, A.O. and S.B.O.; visualization, A.O.; and supervision, A.O. and S.B.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Experiments 1–8 involved no human participants. Experiment 9 collected pseudonymized survey responses from adult participants recruited through Prolific. No personally identifiable information was collected; demographic data were recorded in categorical form only (age group, gender category, education level). The study was classified as minimal risk: no deception, no sensitive topics, no vulnerable populations.

Informed Consent Statement: Informed consent was obtained from all participants in Experiment 9 prior to their involvement in the study. Participants were informed about data collection scope, pseudonymization procedures, storage duration, and withdrawal rights.

Data Availability Statement: The experimental source code, including the multi-model LLM abstraction layer, configuration files for all experiments, and raw JSON output data supporting Experiments 1–8, are publicly available at https://github.com/alperozpinar/SAPIENT-Framework (Release `v3.0-R3`, corresponding to this revised manuscript). The Experiment 9 pilot human validation study was preregistered at OSF (https://doi.org/10.17605/OSF.IO/4KFDC). The pseudonymized Prolific survey data and analysis scripts for Experiment 9 are included in the project repository.

Acknowledgments: During the preparation of this manuscript, the authors used Claude Sonnet 4 (Anthropic, `claude-sonnet-4-20250514`) [46] as the primary LLM backend, and GPT-4o (OpenAI, `gpt-4o`) and Gemini 2.5 Flash (Google, `gemini-2.5-flash`) as additional backends for the cross-model comparison experiment (Section 6.11). Claude was also used as a writing assistant during manuscript preparation. The authors have reviewed and edited all outputs and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AFG	Agentic Focus Group
AI	Artificial Intelligence
CSRD	Corporate Sustainability Reporting Directive
ESG	Environmental, Social, and Governance
GDPR	General Data Protection Regulation
KL	Kullback–Leibler
KVKK	Kisisel Verilerin Korunmasi Kanunu
LLM	Large Language Model
NLP	Natural Language Processing
PPI	Prediction-Powered Inference
SAPIENT	Sentinel-Augmented Population Intelligence for Emerging Narrative Tracking
TDT	Topic Detection and Tracking

References

1. Coombs, W.T. Protecting organization reputations during a crisis: The development and application of situational crisis communication theory. *Corp. Reput. Rev.* **2007**, *10*, 163–176. https://doi.org/10.1057/palgrave.crr.1550049.
2. RepRisk AG. *A Turning Tide in Greenwashing? Exploring the First Decline in Six Years*; Report; RepRisk AG: Zurich, Switzerland, 2024.
3. SESAMm. Beyond greenwashing: How AI Identifies Greenwishing and Greenhushing Amid Tightening ESG Regulations. Blog Post, SESAMm, 5 June 2025. Available online: https://www.sesamm.com/blog/beyond-greenwashing-how-ai-identifies-greenwishing-and-greenhushing-amid-tightening-esg-regulations (accessed on 1 March 2026).
4. Swaminathan, V.; Schwartz, H.A.; Menezes, R.; Hill, S. The language of brands in social media: Using topic modeling on social media conversations to drive brand strategy. *J. Interact. Mark.* **2022**, *57*, 255–277. https://doi.org/10.1177/10949968221088275.
5. Packard, G.; Moore, S.G.; Berger, J.A. Consumer insights from text analysis. *J. Consum. Psychol.* **2023**, *33*, 387–401. https://doi.org/10.1002/jcpy.1383.
6. Krueger, R.A.; Casey, M.A. *Focus Groups: A Practical Guide for Applied Research*, 5th ed.; Sage Publications: Thousand Oaks, CA, USA, 2014.
7. Stewart, D.W.; Shamdasani, P.N. *Focus Groups: Theory and Practice*, 3rd ed.; Sage Publications: Thousand Oaks, CA, USA, 2015.
8. Velasco, E.; Agheneza, T.; Denecke, K.; Kirchner, G.; Eckmanns, T. Social media and internet-based data in global systems for public health surveillance: A systematic review. *Milbank Q.* **2014**, *92*, 7–33. https://doi.org/10.1111/1468-0009.12038.
9. Charles-Smith, L.E.; Reynolds, T.L.; Cameron, M.A.; Conway, M.; Lau, E.H.Y.; Olsen, J.M.; Pavlin, J.A.; Yi, M.; Salathe, M.; Corley, C.D. Using social media for actionable disease surveillance and outbreak management: A systematic literature review. *PLoS ONE* **2015**, *10*, e0139701. https://doi.org/10.1371/journal.pone.0139701.
10. Hammond, A.; Kim, J.J.; Sadler, H.; Vandemaele, K. Influenza surveillance systems using traditional and alternative sources of data: A scoping review. *Influenza Other Respir. Viruses* **2022**, *16*, 482–498. https://doi.org/10.1111/irv.13037.

11. George, Y.; Karunasekera, S.; Harwood, A.; Lim, K.H. Real-time spatio-temporal event detection on geotagged social media. *J. Big Data* **2021**, *8*, 91. https://doi.org/10.1186/s40537-021-00482-2.
12. MacIntyre, C.R.; Chen, X.; Kunasekaran, M.; Quigley, A.; Lim, S.; Stone, H.; Paik, H.y.; Yao, L.; Heslop, D.; Wei, W.; et al. Artificial intelligence in public health: The potential of epidemic early warning systems. *J. Int. Med. Res.* **2023**, *51*, 03000605231159335. https://doi.org/10.1177/03000605231159335.
13. Allan, J. Introduction to Topic Detection and Tracking. In *Topic Detection and Tracking*; Springer: Berlin/Heidelberg, Germany, 2002; pp. 1–16. https://doi.org/10.1007/978-1-4615-0933-2_1.
14. Park, J.S.; O'Brien, J.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, San Francisco, CA, USA, 29 October–1 November 2023; pp. 1–22. https://doi.org/10.1145/3586183.3606763.
15. Argyle, L.P.; Busby, E.C.; Fulda, N.; Gubler, J.R.; Rytting, C.; Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Anal.* **2023**, *31*, 337–351. https://doi.org/10.1017/pan.2023.2.
16. Horton, J.J. Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv* **2023**, arXiv:2301.07543.
17. Hämäläinen, P.; Tavast, M.; Kunnari, A. Evaluating large language models in generating synthetic HCI research data: A case study. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–19. https://doi.org/10.1145/3544548.3580688.
18. Bodhe, S., Zhang, Z., Hamidizadeh, A., Kai, S., Zhang, Y., Yuan, M. LLMs model non-WEIRD populations: Experiments with synthetic cultural agents. *arXiv* **2025**, arXiv:2501.07564.
19. Lin, C.; Ran, M.; Zhang, Y.; Wan, Z.; Fan, H.; Xu, Y.; Guo, Y.; Xue, W.; Song, J. Crisis-Bench: Benchmarking Strategic Ambiguity and Reputation Management in Large Language Models. *arXiv* **2026**, arXiv:2601.05570.
20. Tang, J.; Gao, H.; Pan, X.; Wang, L.; Tan, H.; Gao, D.; Chen, Y.; Chen, X.; Lin, Y.; Li, Y.; et al. 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), Albuquerque, NM, USA, 29 April–4 May 2025; pp. 143–150.
21. Liu, Y.; Liu, W.; Gu, X.; Yao, H.; Wang, W.; Luo, J.; Zhang, Y. RumorSphere: A Framework for Million-Scale Agent-Based Dynamic Simulation of Rumor Propagation. *arXiv* **2025**, arXiv:2509.02172.
22. Huang, E.; Pan, T.; Zhang, S.; Jin, Q.; Zheng, L.; Hu, K.; Li, Y.; Qin, Z.; Ren, K. DualMind: Towards Understanding Cognitive-Affective Cascades in Public Opinion Dissemination via Multi-Agent Simulation. *arXiv* **2026**, arXiv:2602.02534.
23. Hullman, J.; Broska, D.; Sun, H.; Shaw, A. *Validating LLM Simulations as Behavioral Evidence*; Working Paper; Northwestern University: Evanston, IL, USA, 2025. Available online: https://mucollective.northwestern.edu/files/Hullman-llm-behavioral.pdf (accessed on 1 March 2026).
24. Suh, J.; Jahanparast, E.; Moon, S.; Kang, M.; Chang, S. Language Model Fine-Tuning on Scaled Survey Data for Predicting Distributions of Public Opinions. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Vienna, Austria, 27 July–1 August 2025; pp. 21147–21170. https://doi.org/10.18653/v1/2025.acl-long.1028.
25. Bisbee, J.; Clinton, J.D.; Dorff, C.; Kenkel, B.; Larson, J.M. Synthetic Replacements for Human Survey Data? The Perils of Large Language Models. *Political Anal.* **2024**, *32*, 401–416. https://doi.org/10.1017/pan.2024.5.
26. Ibrahim, N.K. Epidemiologic surveillance for controlling COVID-19 pandemic: Types, challenges and implications. *J. Infect. Public Health* **2020**, *13*, 1630–1638. https://doi.org/10.1016/j.jiph.2020.07.019.
27. Eysenbach, G. Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *J. Med. Internet Res.* **2009**, *11*, e11. https://doi.org/10.2196/jmir.1157.
28. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake shakes Twitter users: Real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 851–860. https://doi.org/10.1145/1772690.1772777.
29. McCombs, M.E.; Shaw, D.L.; Weaver, D.H. New directions in agenda-setting theory and research. *Mass Commun. Soc.* **2014**, *17*, 781–802. https://doi.org/10.1080/15205436.2014.964871.
30. Entman, R.M. Framing: Toward clarification of a fractured paradigm. *J. Commun.* **1993**, *43*, 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x.
31. Vosoughi, S.; Roy, D.; Aral, S. The spread of true and false news online. *Science* **2018**, *359*, 1146–1151. https://doi.org/10.1126/science.aap9559.
32. Deffuant, G.; Neau, D.; Amblard, F.; Weisbuch, G. Mixing beliefs among interacting agents. *Adv. Complex Syst.* **2000**, *3*, 87–98. https://doi.org/10.1142/S0219525900000078.
33. Flache, A.; Mäs, M.; Feliciani, T.; Chattoe-Brown, E.; Deffuant, G.; Huet, S.; Lorenz, J. Models of social influence: Towards the next frontiers. *J. Artif. Soc. Soc. Simul.* **2017**, *20*, 2. https://doi.org/10.18564/jasss.3521.

34. Gao, C.; Lan, X.; Li, N.; Yuan, Y.; Ding, J.; Zhou, Z.; Xu, F.; Li, Y. Large language models empowered agent-based modeling and simulation: A survey and perspectives. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1259. https://doi.org/10.1057/s41599-024-03611-3.
35. Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; Hashimoto, T. Whose Opinions Do Language Models Reflect? In *Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023*; Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J., Eds.; Proceedings of Machine Learning Research (PMLR): Birmingham, UK, 2023; Volume 202, pp. 29971–30004.
36. Dash, T.; Karri, D.; Vurity, A.; Datla, G.; Ahmad, T.; Rafi, S.; Tangudu, R. PolyPersona: Persona-Grounded LLM for Synthetic Survey Responses. *arXiv* **2025**, arXiv:2512.14562.
37. Angelopoulos, A.N.; Bates, S.; Candès, E.J.; Jordan, M.I.; Lei, L. Prediction-Powered Inference. *Science* **2023**, *382*, 669–674.
38. Broska, D.; Howes, M.; van Loon, A. The Mixed Subjects Design: Treating Large Language Models as Potentially Informative Observations. *Sociol. Methods Res.* **2025**, *54*, 1074.
39. European Parliament and Council. Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 December 2022 amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Directive 2013/34/EU, as regards corporate sustainability reporting. *Off. J. Eur. Union* **2022**, *L 322*, 15–80. Available online: https://eur-lex.europa.eu/eli/dir/2022/2464/oj/eng (accessed on 1 March 2026).
40. Schimanski, T.; Reding, A.; Reding, N.; Bingler, J.; Kraus, M.; Leippold, M. Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Financ. Res. Lett.* **2024**, *61*, 104979. https://doi.org/10.1016/j.frl.2024.104979.
41. Ghaemi Asl, M. A novel AI-driven approach to greenwashing: Breakthroughs in the future fit between domain-specific Islamic enterprises with varying developmental progress and ESG landscapes. *Future Bus. J.* **2025**, *11*, 93. https://doi.org/10.1186/s43093-025-00497-8.
42. Quast, V.; Jacobs, G.; Dehn, S.; Höpfner, G. Enabling Humans and AI Systems to Retrieve Information from System Architectures in Model-Based Systems Engineering. *Systems* **2026**, *14*, 83. https://doi.org/10.3390/systems14010083.
43. Gerolimos, N.; Alevizos, V.; Priniotakis, G. A Methodological Framework for Chaos-Aware Evaluation of Self-Organization in Swarm-Based Engineering Systems. *Systems* **2026**, *14*, 215. https://doi.org/10.3390/systems14020215.
44. Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; Fritz, M. Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, Copenhagen, Denmark, 30 November 2023; pp. 79–90. https://doi.org/10.1145/3605764.3623985.
45. National Institute of Standards and Technology. AI Risk Management Framework (AI RMF 1.0). NIST AI 100-1. 2023. Available online: https://www.nist.gov/artificial-intelligence (accessed on 1 March 2026).
46. Anthropic. *System Card: Claude Opus 4 & Claude Sonnet 4*; Technical Report; Anthropic: San Francisco, CA, USA, 2025.
47. Google DeepMind. *Gemini 2.5: Our Most Intelligent AI Model*; Technical Report; Google DeepMind: London, UK, 2025.
48. Douglas, B.D.; Ewell, P.J.; Brauer, M. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS ONE* **2023**, *18*, e0279720. https://doi.org/10.1371/journal.pone.0279720.