PERSONACOLLECTIVE: A Benchmark for Group-Level Persona Simulation in Large Language Models

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly deployed as *silicon samples* in social-science research and policy modeling, where the unit of analysis is rarely an individual but a group. Yet existing persona benchmarks evaluate only individual-level role-playing, leaving group-level phenomena—internal opinion diversity, faction structure, deliberative dynamics, social norms, and collective decisions—largely untested. We introduce PERSONACOLLECTIVE, the first benchmark for group-level persona simulation, grounded in social-science theory with explicit scope conditions and built from 30 sociologically grounded group profiles paired with 100 group-decision scenarios and three-tier confidence-graded ground truth. We define Group Simulation Fidelity (GSF), a five-dimensional evaluation along the five phenomena above; its norm-sensitivity test uses a differential probe design that separates group-specific normative understanding from generic safety preferences. Evaluating seven state-of-the-art LLMs across three simulation paradigms, we find systematic failures along all five dimensions, including opinion homogenization, dynamic insensitivity, and minority-voice suppression.

1 Introduction

Large language models (LLMs) are increasingly deployed as *silicon samples*: as surrogate survey respondents and economic agents in social-science research (Argyle et al., 2023; Filippas et al., 2024; Anthis et al., 2025), and as multi-agent populations for participatory design, deliberation, and norm emergence (Park et al., 2023; Tessler et al., 2024; Ashery et al., 2025). The unit of analysis in these applications is rarely an isolated individual but a group or community: a rural evangelical congregation, a labor union, a coalition of climate activists. A faithful simulation of such groups must reproduce phenomena that emerge only above the individual level—internal opinion diversity, faction structure, deliberative dynamics, social norms, and collective decisions—none of which reduces to the faithful simulation of any single member. Whether contemporary LLMs are ready to play this group-level role, however, remains largely unmeasured.

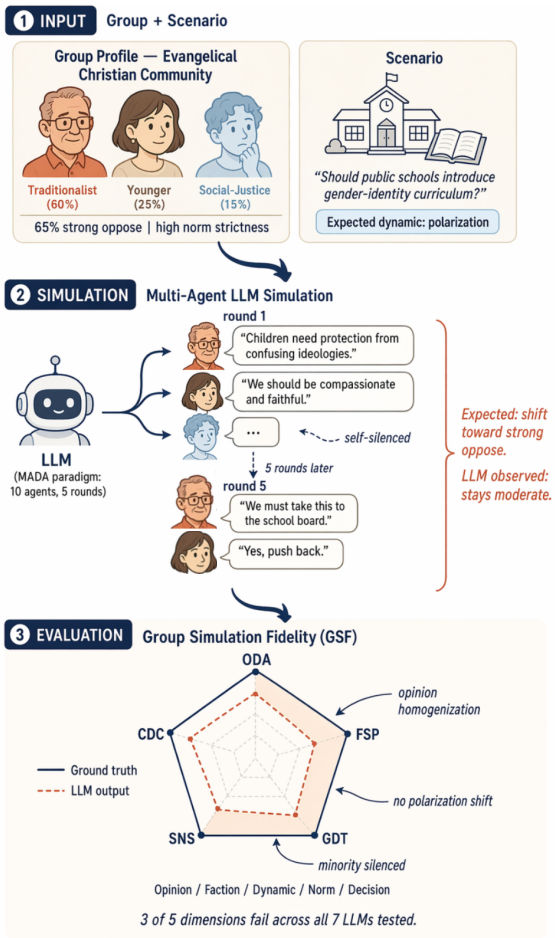

Figure 1: Overview of PERSONACOLLECTIVE. **(1)** A sociologically grounded group profile is paired with a decision scenario. **(2)** An LLM simulates multi-agent deliberation. **(3)** GSF evaluates fidelity along five dimensions; the inner polygon reveals systematic failures on FSP, GDT, and SNS.

Despite a proliferation of persona-evaluation work (Chen et al., 2024b), none of it measures

group-level fidelity along the dimensions above. The dominant benchmarks evaluate whether an LLM faithfully plays a *single* persona in isolation—measuring surface consistency (Samuel et al., 2025; Wang et al., 2025a; Li et al., 2025), personality fidelity via psychological interviews (Wang et al., 2024) or personality-indicative training (Ran et al., 2024), and persona-driven decision-making (Xu et al., 2025b). Hu et al. (2026) extend evaluation to population-level survey distributions but only as a static distribution-matching task, without faction structure, deliberative dynamics, or norm sensitivity. The closest work to ours examines an isolated facet of group behavior: Chuang et al. (2026) measure opinion convergence in multi-agent debate as a single dynamic phenomenon, while Lin and Wei (2026) infer community preferences as a static task; neither captures the coupled, multi-dimensional behavior that real group simulation requires. Consider simulating how a U.S. evangelical Christian community responds to a school's introduction of gender-identity curriculum: a faithful answer demands the right *distribution* of positions, the right internal *factions*, polarization *dynamics* matching this scenario type, recognition that supportive minorities feel *normative* pressure to keep silent, and a plausible collective *decision*—none of which any current benchmark jointly evaluates.

We introduce PERSONACOLLECTIVE, the first benchmark for group-level persona simulation, structured around five complementary fidelity dimensions and built on social-science theory with explicit *scope conditions*. PERSONACOLLECTIVE provides 30 sociologically grounded group profiles spanning political, occupational, cultural, and demographic communities, paired with 100 group-decision scenarios and three-tier confidence-graded ground truth derived from representative surveys, public corpora, and expert annotation. We define **Group Simulation Fidelity (GSF)**, jointly measuring opinion-distribution alignment, faction-structure preservation, dynamic-trajectory fidelity, social-norm sensitivity, and collective-decision calibration; the norm-sensitivity test introduces a *differential probe* design that separates group-specific normative understanding from generic safety preferences. A central methodological commitment is to refuse treating contested social-science predictions as universal laws: each scenario is annotated with an *expected* dynamic type—polarization, depolarization, convergence, fragmentation, or no shift—chosen from the conditions under which each theory applies, so that a faithful simulation must match the expected type rather than always polarize.

We evaluate seven state-of-the-art LLMs on PERSONACOLLECTIVE across three simulation paradigms and four information conditions, asking: can current LLMs faithfully simulate group-level personas (**RQ1**); does fidelity stem from internalized knowledge or prompted context (**RQ2**); and do models exhibit genuine group-specific normative understanding or merely generic safety compliance (**RQ3**)? The picture is consistent across models, with three recurring failures. Models homogenize internal opinion: minority-retention rates trail the survey-sampling baseline by a wide margin. They confuse dynamic types: the predicted dynamic matches the expected type at near-chance accuracy. And they exhibit a near-zero *differential norm sensitivity*, indicating that what looks like norm compliance is largely a residue of general safety training rather than group-specific normative understanding.

Our contributions are threefold. *1)* We operationalize group-level persona simulation fidelity along five mutually irreducible dimensions, anchored in scope-conditioned social-science theory. *2)* We release PERSONACOLLECTIVE: 30 group profiles, 100 scenarios, and confidence-graded ground truth, together with a five-metric evaluation suite whose norm-sensitivity test introduces a differential probe that isolates group-specific normative understanding from generic safety preferences. *3)* We provide systematic empirical evidence that seven state-of-the-art LLMs fail along all five dimensions, with opinion homogenization, dynamic insensitivity, and minority-voice suppression as recurring failure modes.

2 Related Work

Persona simulation at the individual level. A growing body of persona-evaluation benchmarks scores whether an LLM faithfully plays a single character along surface axes such as style and knowledge consistency (Samuel et al., 2025; Wang et al., 2025a), sequential behavior fidelity (Li et al., 2025), or paired-agent social interaction (Chen et al., 2024a). A complementary line probes deeper aspects of single-character fidelity: personality fidelity via psychological interviews and personality-indicative training data (Wang et al., 2024; Ran et al., 2024), persona-driven decision-making and

Benchmark	Unit	Multi-turn	Opinion	Factions	Dynamic	Norms	Decision
PersonaGym (Samuel et al., 2025)	Indiv.	—	—	—	—	—	—
CharacterBox (Wang et al., 2025a)	Indiv.	—	—	—	—	—	—
BEHAVIORCHAIN (Li et al., 2025)	Indiv.	—	—	—	—	—	—
SimBench (Hu et al., 2026)	Pop.	—	✓	—	—	—	—
CommunityBench (Lin and Wei, 2026)	Comm.	—	part.	—	—	✓	—
DEBATE (Chuang et al., 2026)	Group	✓	—	—	conv. only	—	—
SocialBench (Chen et al., 2024a)	Indiv.	—	—	—	—	—	—
PERSONACOLLECTIVE (ours)	**Group**	✓	✓	✓	✓	✓	✓

Table 1: Coverage of group-fidelity dimensions across persona-evaluation benchmarks. **Unit**: individual, population, community, or group as unit of analysis; **Multi-turn**: evaluates multi-turn behavior; **Opinion**: opinion distribution; **Factions**: faction structure; **Dynamic**: deliberative dynamic type (DEBATE covers convergence only); **Norms**: social-norm sensitivity; **Decision**: collective decision. PERSONACOLLECTIVE is the first to evaluate all five group-level dimensions jointly.

inner-thought reasoning (Xu et al., 2025b,a), and character understanding from fictional works or specialized roles such as key opinion leaders (Yuan et al., 2024; Wang et al., 2025b; Xu et al., 2024); Chen et al. (2024b) survey this rapidly expanding line. The unit of analysis throughout is one character, so these benchmarks cannot diagnose phenomena that emerge only when multiple personas interact—internal opinion diversity, faction structure, deliberative dynamics, or normative pressure on minorities.

Beyond the individual: populations, communities, and groups. A second line of work scales evaluation to aggregates but each benchmark captures only one facet of group-level fidelity: Hu et al. (2026) and Lin and Wei (2026) treat the group as a static distribution to be matched against surveys or community preferences (Bisbee et al., 2024; Mi et al., 2025), while Chuang et al. (2026) measures opinion convergence as the only dynamic of interest and Ashery et al. (2025) demonstrates emergent social conventions as a phenomenon rather than a benchmark. Multi-agent platforms such as AgentSociety (Piao et al., 2025) and SocioVerse (Zhang et al., 2025) provide simulation infrastructure but no standardized metric. PERSONACOLLECTIVE differs by evaluating five complementary group-fidelity dimensions jointly, conditioning every theoretical prediction on explicit *scope conditions* (§3.1), and isolating group-specific normative understanding from generic safety preferences via a differential probe.

3 The PERSONACOLLECTIVE Benchmark

PERSONACOLLECTIVE pairs 30 sociologically grounded **group profiles** (§3.2) with 100 **group-decision scenarios** (§3.3), each tagged with an expected dynamic type under explicit **scope conditions** (§3.1). Three layers of **ground truth** (§3.4) supply distributional, structural, and qualitative targets, each carried into GSF as a confidence weight q_i.

3.1 Theoretical Foundation with Scope Conditions

Treating social-science predictions as universal laws—*groups polarize*, *minorities self-silence*—rewards stereotypes; each theory holds only under specific conditions (Sunstein, 2002; Noelle-Neumann, 1974). PERSONACOLLECTIVE attaches scope conditions to every theoretical claim (Table 2): each scenario is labeled with one of five expected dynamic types (*polarization*, *depolarization*, *convergence*, *fragmentation*, *no shift*), and simulations are scored against the labeled type.

3.2 Group Profiles

Extending character-level profiling from fictional works (Yuan et al., 2024) to real-world communities, the 30 profiles—five each across political ideology, occupation, demographics/generation, culture/religion, advocacy interests, and situational membership—are annotated along five facets: demographic anchors, issue-by-issue opinion distributions (shape-tagged as unimodal/bimodal/long-tail/uniform), faction structure (2–4 factions with shares), a normative inventory typed as *descriptive*, *injunctive*, or *taboo*, and group-dynamic fea-

Theory	Core prediction	Scope conditions	Out-of-scope	Metric
Social Identity (Tajfel and Turner, 1979)	In-group favoritism; typicality gradient	Identity is salient and self-categorized	Loose ad-hoc groupings	ODA, FSP
Group Polarization (Myers and Lamm, 1976; Sunstein, 2002)	Discussion amplifies pre-existing leaning	Homogeneous group; identity-relevant issue	Balanced cross-cutting dialogue (depolarization); neutral issue	GDT
Spiral of Silence (Noelle-Neumann, 1974)	Minorities self-censor	Perceived sanction; perceived isolation	Anonymous setting; activist sub-community; outside support	SNS
Collective Action (Olson, 1971; Ostrom, 1990)	Coordination shapes collective decision	Decision requires joint action	Independent individual decisions	CDC

Table 2: The four social-science theories operationalized by PERSONACOLLECTIVE, each with explicit scope conditions and out-of-scope cases. A faithful simulation is rewarded for producing a theory's predicted phenomenon *only when* the scenario satisfies its scope conditions, and is penalized for producing it under out-of-scope cases.

tures. Every annotation carries a confidence tier (A: representative survey with $n > 1{,}000$; B: approximate cross-screening; C: corpus and expert only), and the tier weight $q_i \in [0, 1]$ enters every GSF metric. The full annotation schema, procedures, inter-annotator agreement (Cohen's κ), and community-member review are in Appendix A–D.

3.3 Scenarios

The 100 scenarios span five domains (policy deliberation, resource allocation, crisis response, internal conflict resolution, inter-group interaction) with twenty each. Each scenario is paired with groups under three applicability tags (*group-specific*, *cross-group general*, *group-relevant*), yielding $\sim$50 valid scenarios per group and 1,867 active pairs; every multi-turn scenario carries an expected dynamic type per its scope conditions (§3.1). The full list, applicability matrix, and dynamic-type distribution are in Appendix B.

3.4 Ground Truth

Ground truth combines a **survey layer** (tier-A distributions from Pew, GSS, ANES, WVS), a **corpus layer** (candidate factions and norms from platform-bias-labeled public sources, filtered by expert review), and an **expert layer** (scenario authoring, dynamic-type assignment, conflict arbitration). When layers conflict the default ordering is survey $\succ$ expert $\succ$ corpus; each annotation receives a confidence score $q_i \in [0, 1]$ combining tier weight with cross-layer agreement (Appendix C).

4 Group Simulation Fidelity Metrics

GSF decomposes group simulation fidelity into five mutually irreducible dimensions, each targeting a distinct failure mode. We report them in parallel as a radar profile rather than collapsing into a single score; every metric inherits the confidence weight $q_i \in [0, 1]$ from §3.4. A composite GSF score with weight-sensitivity analysis and full derivations are in Appendix F.

ODA: Opinion Distribution Alignment. ODA measures whether the simulated group reproduces the correct position distribution on each issue. For issue i with empirical P_i and simulated Q_i,

$$\mathrm{ODA} = \frac{\sum_i q_i \cdot \big(1 - \mathrm{JSD}(P_i \,\|\, Q_i)\big)}{\sum_i q_i}. \quad (1)$$

FSP: Faction Structure Preservation. A model can match the marginal ODA while collapsing factions into a single voice. FSP directly measures internal composition via four sub-quantities: faction coverage FC, proportional-calibration error PCE, minority retention MR (presence of factions with share $< 20\%$), and minority persistence MP across deliberation rounds; the overall metric is a confidence-weighted sum (Appendix F).

GDT: Group Dynamic Trajectory. GDT asks whether the simulation produces the dynamic type the scenario demands rather than a default polarization. A polarization shift Δ_p and dispersion shift Δ_v are mapped via signed thresholds to a predicted type in $\{\text{polar.}, \text{depolar.}, \text{converge}, \text{fragment}, \text{no shift}\}$:

$$\mathrm{GDT} = \lambda \mathbb{1}[d_s = d_e] + (1 - \lambda)\, e^{-|\Delta_p - \delta^\star|/\tau}, \quad (2)$$

with d_s, d_e the simulated and expected types, $\delta^\star$ calibrated from human deliberation studies, and $\lambda = 0.7$ favoring correct type over correct magnitude.

SNS: Social Norm Sensitivity. A naive violation count conflates group-specific normative understanding with generic safety training. SNS disentangles the two via a *differential probe*: for every norm k in the target group's inventory we construct a paired probe where the same behavior *violates* the norm in the target group but is *acceptable* in a control group, and measure the **differential norm sensitivity**:

$$\mathrm{DNS}_k = \mathrm{Pr}_{\text{tgt}}(c_k) - \mathrm{Pr}_{\text{ctrl}}(c_k), \tag{3}$$

where c_k denotes norm compliance. Generic politeness yields $\mathrm{DNS}_k \approx 0$; only group-specific normative understanding yields $\mathrm{DNS}_k > 0$. The aggregate $\mathrm{SNS} = w_d \cdot \overline{\mathrm{DNS}}_{\text{desc}} + w_i \cdot \overline{\mathrm{DNS}}_{\text{inj}} + w_t \cdot \overline{\mathrm{DNS}}_{\text{taboo}}$ uses $w_t > w_i > w_d$ to emphasize sanctioned taboo norms.

CDC: Collective Decision Calibration. Group decisions rarely have a single correct answer, so ground truth is a distribution $P^\star(d)$ over outcomes:

$$\mathrm{CDC} = 1 - \mathrm{JSD}\big(P^\star(d) \,\|\, P^{\text{sim}}(d)\big), \tag{4}$$

where $P^{\text{sim}}(d)$ is the empirical decision distribution across simulation seeds.

5 Experimental Setup

Simulation paradigms. We evaluate three paradigms that represent qualitatively different ways of eliciting group behavior from LLMs. **SAGP** (Single-Agent Group Proxy) prompts one LLM instance to speak for the entire group, producing an opinion distribution, faction structure, and collective decision in a single pass. **MADA** (Multi-Agent Deliberative Assembly) extends coordinated role-playing (Wang et al., 2025b) to group deliberation: it instantiates N LLM agents assigned to group roles via stratified faction sampling (minority factions are oversampled and post-weighted) and runs T-round deliberation; each (model, group, scenario) triple uses at least three random seeds. **HFBS** (Hierarchical Faction-Based Simulation) assigns one LLM per annotated faction and a mediator LLM that synthesizes faction positions into a collective outcome; we test three variants—*Oracle* (ground-truth faction structure provided), *Inferred* (model infers factions from the profile), and *Noisy* (partially incorrect structure injected). Full prompt templates for all paradigms are in Appendix E.

Information conditions. To disentangle knowledge already in a model's weights from knowledge supplied in the prompt, every evaluation is repeated under four information conditions of increasing richness: *Name-Only* (group name only), *Demographic* (name plus demographic anchors), *Profile* (full profile text without explicit distributional labels), and *Full* (profile plus faction shares plus norm inventory). A model that improves substantially from Name-Only to Full relies on prompted context; one that performs well under Name-Only possesses internalized group knowledge.

Baselines. We include five non-LLM baselines that anchor the score range. *Majority*: always outputs the largest faction's modal opinion. *Survey-Sampling*: draws responses from the tier-A survey distribution (upper bound for ODA on items with survey coverage). *Uniform-Diversity*: generates uniformly diverse opinions (a blind diversity heuristic). *Retrieval-Augmented*: retrieves group-relevant documents and conditions a frozen LLM on them. *Human*: non-expert annotators predict group responses given the same profile text the LLMs receive.

Models and implementation. We evaluate seven LLMs spanning commercial and open-weight families: GPT-5.4, Claude Sonnet 4.6, Gemini 3.1 Pro, DeepSeek-V4 Pro, Llama 4 Maverick, Qwen3.6-27B, and Qwen3.6-A3B. All models are queried at temperature 0.7 with top-$p = 0.95$; MADA runs use $T = 5$ deliberation rounds. We report 95% bootstrap confidence intervals over evaluation pairs and Cohen's d effect sizes for key comparisons.

6 Experiments

We organize experiments around three research questions. **RQ1**: How well do current LLMs perform across the five GSF dimensions? **RQ2**: Does group simulation fidelity stem from internalized group knowledge or from prompted context? **RQ3**: Do LLMs exhibit genuine group-specific normative understanding, or merely generic safety compliance?

6.1 RQ1: Can LLMs Faithfully Simulate Group-Level Personas?

Table 3 reports all five GSF metrics under the *Full* information condition with the MADA paradigm. No model achieves adequate fidelity across all five dimensions, and the Human baseline outperforms every LLM on every dimension except CDC.

Model / Baseline	ODA ↑	FSP ↑	GDT ↑	SNS ↑	CDC ↑	Avg.↑
Baselines						
Majority	0.42	0.15	0.20	0.00	0.31	0.22
Survey-Sampling	0.85	0.48	0.21	0.16	0.42	0.42
Uniform-Diversity	0.35	0.52	0.20	0.00	0.25	0.26
Retrieval-Aug.	0.72	0.38	0.25	0.08	0.47	0.38
Human	0.79	0.55	0.43	0.31	0.58	0.53
Commercial LLMs						
GPT-5.4	**0.81**	**0.44**	0.32	0.14	**0.57**	**0.46**
Claude Sonnet 4.6	0.78	0.43	0.30	**0.16**	0.55	0.44
Gemini 3.1 Pro	0.73	0.35	**0.34**	0.10	0.47	0.40
Open-weight LLMs						
DeepSeek-V4 Pro	0.76	0.39	0.29	0.12	0.53	0.42
Llama 4 Maverick	0.71	0.31	0.27	0.09	0.48	0.37
Qwen3.6-27B	0.68	0.30	0.18	0.05	0.45	0.33
Qwen3.6-A3B	0.55	0.22	0.14	0.04	0.35	0.26

Table 3: Main results on PERSONACOLLECTIVE under the *Full* information condition and MADA paradigm. All metrics range in $[0, 1]$; higher is better. Best LLM per column is **bolded**; 95% bootstrap CIs are in Appendix H.

ODA is the easiest dimension: most models approach the Survey-Sampling upper bound, indicating that LLMs can reproduce marginal opinion distributions when given sufficient profile information. However, FSP, GDT, and SNS trail substantially behind, confirming that matching a distribution does not imply preserving its internal structure, dynamics, or normative texture.

Model performance reveals a clear tier structure. GPT-5.4 and Claude Sonnet 4.6 form the top tier; DeepSeek-V4 Pro follows closely, approaching commercial leaders on ODA and CDC and showing the most balanced profile among open-weight models; Gemini 3.1 Pro and Llama 4 Maverick occupy the second tier; smaller dense models remain far behind on GDT and SNS, with pairwise comparisons pointing to capability rather than size as the bottleneck. Two dimension-specific findings merit attention: Gemini 3.1 Pro achieves the highest GDT among all models, suggesting a unique strength in dynamic reasoning, while Claude Sonnet 4.6 leads on SNS, exhibiting the best norm sensitivity.

Opinion homogenization. The minority-retention sub-metric of FSP reveals a consistent failure: factions with ground-truth share below 20% are absent from the simulation output in a median of 49% of cases, whereas the Survey-Sampling baseline retains them by construction. This pattern is strongest under Name-Only and SAGP, where the model lacks structural cues to generate minority perspectives.

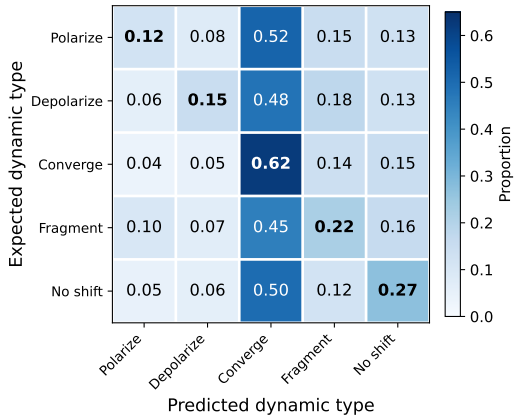

Figure 2: Confusion matrix of expected vs. predicted dynamic types, pooled across all models. Models disproportionately predict *convergence* regardless of the expected type, achieving near-chance accuracy on the five-way classification.

Dynamic insensitivity. Figure 2 presents the confusion matrix of expected versus predicted dynamic types. The overall direction accuracy is only 0.31 (chance = 0.20), with a strong bias toward predicting *convergence*—consistent with RLHF-induced moderation that steers multi-agent discussions toward agreement. Polarization scenarios are the most frequently misclassified: models rarely amplify a pre-existing leaning, instead defaulting to a centrist consensus.

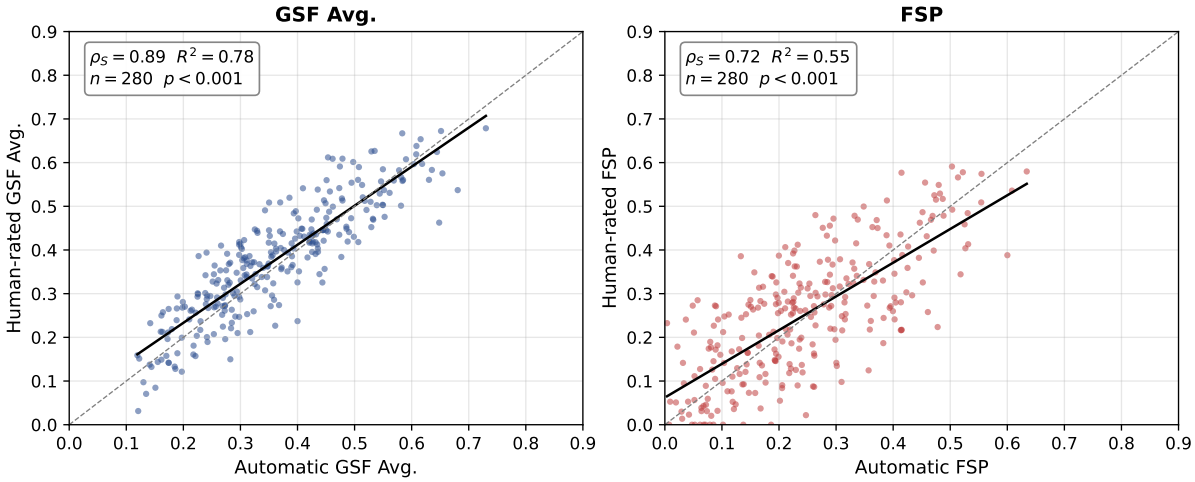

Figure 3: Automatic GSF vs. three-rater human GSF on a 40-pair sample. Left: GSF average ($\rho_S = 0.86$); right: FSP ($\rho_S = 0.74$). Both are highly significant ($p < 0.001$, $n = 280$); regression lines lie slightly below the diagonal, indicating modestly more lenient human ratings.

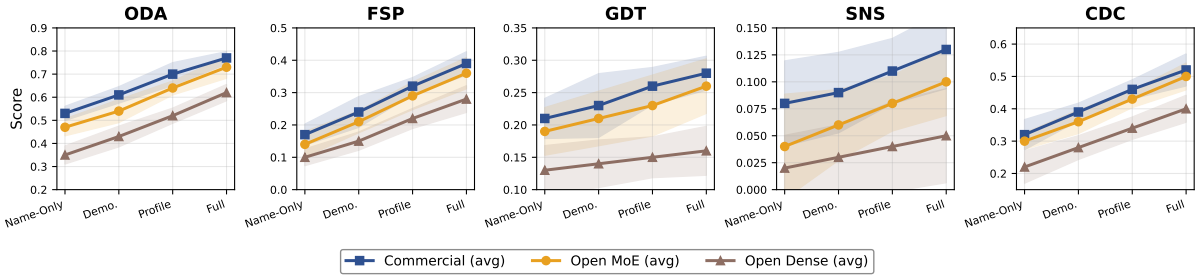

Figure 4: GSF scores as a function of information richness (Name-Only → Full), grouped into commercial, open MoE, and open dense model families. ODA and FSP improve substantially with richer prompts; GDT and SNS show minimal gains, indicating that dynamic and normative understanding are not easily supplied through context.

Paradigm	ODA	FSP	GDT	SNS	CDC
SAGP	0.76	0.32	0.22	0.09	0.48
MADA	0.73	0.38	0.26	0.10	0.50
HFBS-Oracle	0.71	0.52	0.18	0.10	0.47
HFBS-Inferred	0.72	0.35	0.23	0.09	0.49
HFBS-Noisy	0.69	0.29	0.21	0.08	0.44

Table 4: GSF scores by simulation paradigm, averaged across all seven LLMs under the *Full* information condition.

Metric validation. To verify that the automatic GSF pipeline tracks expert judgment, three independent raters scored 40 randomly drawn ⟨model, scenario⟩ pairs per dimension on a 0–1 rubric. Spearman ρ_S is 0.86 for GSF average and 0.74 for FSP (Figure 3), supporting the validity of the automatic evaluation at the granularity we report.

6.2 RQ2: Internalized Knowledge or Prompted Context?

Figure 4 plots GSF scores as information richness increases from Name-Only to Full. The lift is largest for ODA and FSP: providing explicit distributional labels and faction shares directly improves these distributional and structural metrics. GDT and SNS, by contrast, show only marginal improvement, suggesting that dynamic trajectory understanding and normative sensitivity depend on reasoning capacities that richer context alone does not supply. The per-family curves further show that commercial models and DeepSeek-V4 Pro converge on ODA under the Full condition, but tier gaps on GDT and SNS persist regardless of information richness.

Table 4 compares the three simulation paradigms. MADA outperforms SAGP on FSP and GDT, confirming that multi-agent deliberation better preserves faction diversity and produces more realistic

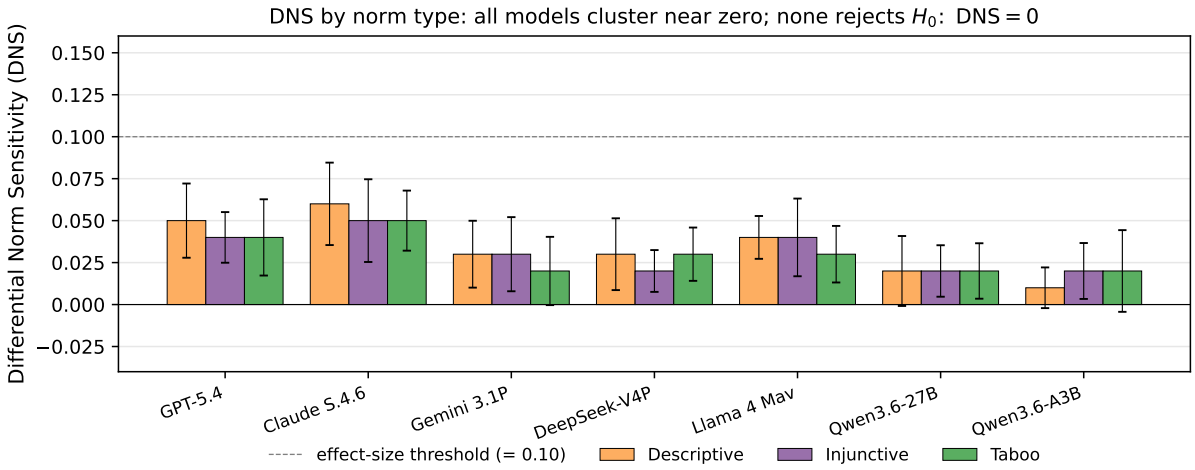

Figure 5: Differential norm sensitivity (DNS) by norm type and model. Error bars are 95% percentile-bootstrap CI on the paired-probe difference. The dashed gray line at 0.10 is the effect-size threshold below which we treat the difference as practically zero; *no* cell crosses it, indicating that norm-compliant behavior in current LLMs is not graded by within-group enforcement.

dynamics, but slightly underperforms on ODA, likely because multi-agent interaction introduces noise into the aggregate distribution. HFBS-Oracle achieves the highest FSP (as faction structure is given explicitly) but the lowest GDT (as hierarchical mediation suppresses emergent dynamics); the HFBS-Noisy variant drops FSP below SAGP—structural scaffolds are double-edged.

6.3 RQ3: Group-Specific Normative Understanding or Generic Safety?

Across all seven LLMs and all norm types, the differential norm sensitivity $\overline{\text{DNS}}$ hovers near zero (Figure 5), indicating that what appears to be norm-compliant behavior is almost entirely attributable to generic safety training rather than group-specific normative understanding—a capacity distinct from individual-level inner-thought reasoning (Xu et al., 2025a). For instance, models correctly avoid questioning biblical authority when simulating evangelicals—but they avoid it equally when simulating secular humanists, for whom such questioning carries no social cost. A paired t-test on DNS_k across all probe pairs fails to reject $H_0 : \text{DNS} = 0$ for every model ($p > 0.1$).

Breaking DNS down by norm type sharpens the picture: taboo norms with explicit sanction—the strongest signal one would expect a model to internalize—show no larger DNS than weak descriptive norms (0.04 vs. 0.03 averaged across models). This null result is the diagnostic value of the differential design: a raw violation-count metric would have credited the same models with high *absolute* norm compliance (> 0.8 on every taboo across every group), missing entirely that the compliance is group-independent.

Case study: Evangelical Christians on gender curriculum. Ground truth specifies a bimodal distribution (65% strong oppose, 5% support), three factions with a 15% social-justice minority, expected polarization, and a taboo norm against questioning biblical authority. In MADA simulation, GPT-5.4 compresses the bimodal shape into a unimodal peak (ODA: $0.81 \rightarrow 0.64$), eliminates the social-justice faction (FSP failure), generates convergence instead of polarization (GDT mismatch), shows identical norm compliance for evangelical and secular-humanist groups (DNS ≈ 0), and predicts "open dialogue" instead of the ground-truth modal response of *organized opposition* (0.62 mass vs. 0.08)—inverting the real-world outcome.

7 Conclusion

We introduced PERSONACOLLECTIVE, the first benchmark for group-level persona simulation, and its five-dimensional GSF evaluation. Across seven LLMs, faction structure, dynamic trajectory, and norm sensitivity scored consistently below 0.45, 0.40, and 0.17; a near-zero differential norm sensitivity indicates that current safety training is not a substitute for group-specific normative understanding. The differential norm sensitivity probe offers a reusable diagnostic for any claim that an LLM "understands" a community's norms.

Limitations

Our work has several limitations. First, all 30 group profiles are English-language and predominantly drawn from U.S. and Western European contexts; the social-science theories we operationalize (group polarization, spiral of silence, collective action logic) may not transfer to non-Western communities with different deliberative norms. Second, ground truth is static: we annotate expected distributions and dynamics at a single point in time, but real group opinions shift in response to external events; a longitudinal benchmark would better capture this. Third, our three-tier confidence grading relies on the availability of representative survey data (tier A) and expert judgment (tier B); for understudied groups, most annotations default to tier C, reducing evaluation precision.

Ethics Statement

Our benchmark evaluates how LLMs simulate social groups, which carries risks we have worked to mitigate. First, group profiles could reinforce stereotypes; we mitigate this by grounding annotations in representative survey data and expert review rather than relying on common assumptions, and by including within-group diversity as a core evaluation dimension. Second, a high-fidelity group simulator could be misused to generate targeted persuasion or disinformation; we release the benchmark for evaluation purposes only and do not release trained simulation models. Third, all group profiles describe public communities discussed in published social-science literature; no private or identifiable individual data is used. Fourth, annotators were compensated above local minimum wage and were informed that their work would appear in a public research artifact; annotation guidelines instructed them to flag and escalate discomfort with sensitive content. We encourage downstream users to apply PERSONACOLLECTIVE as a diagnostic tool for identifying and reducing representational harms in LLM-based social simulations, not as a training signal for producing more convincing impersonations of real communities.

References

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. 2025. Position: LLM social simulations are a promising research method. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 81005–81034. PMLR.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. 2025. Emergent social conventions and collective bias in llm populations. *Science Advances*, 11(20):eadu9368.

James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024a. SocialBench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126, Bangkok, Thailand. Association for Computational Linguistics.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024b. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*. Survey Certification.

Yun-Shiuan Chuang, Ruixuan Tu, Chengtao Dai, Smit Vasani, Binwei Yao, Michael Henry Tessler, Sijia Yang, Dhavan V. Shah, Robert D. Hawkins, Junjie Hu, and Timothy T. Rogers. 2026. DEBATE: A large-scale benchmark for evaluating opinion dynamics in role-playing LLM agents. In *The Fourteenth International Conference on Learning Representations*.

Apostolos Filippas, John J. Horton, and Benjamin S. Manning. 2024. Large language models as simulated economic agents: What can we learn from homo silicus? In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, pages 614–615, New York, NY, USA. Association for Computing Machinery.

Tiancheng Hu, Joachim Baumann, Lorenzo Lupo, Nigel Collier, Dirk Hovy, and Paul Röttger. 2026. SimBench: Benchmarking the ability of large language models to simulate human behaviors. In *The Fourteenth International Conference on Learning Representations*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Rui Li, Heming Xia, Xinfeng Yuan, Qingxiu Dong, Lei Sha, Wenjie Li, and Zhifang Sui. 2025. How far are LLMs from being our digital twins? a benchmark for persona-based behavior chain simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15738–15763, Vienna, Austria. Association for Computational Linguistics.

Jiayu Lin and Zhongyu Wei. 2026. Communitybench: Benchmarking community-level alignment across diverse groups and tasks. *Preprint*, arXiv:2601.13669.

Qirui Mi, Mengyue Yang, Xiangning Yu, Zhiyu Zhao, Cheng Deng, Bo An, Haifeng Zhang, Xu Chen, and Jun Wang. 2025. MF-LLM: Simulating population decision dynamics via a mean-field large language model framework. In *Advances in Neural Information Processing Systems*, volume 38, pages 77860–77898. Curran Associates, Inc.

David Myers and Helmut Lamm. 1976. The group polarization phenomenon. *Psychological Bulletin*, 83:602–627.

Elisabeth Noelle-Neumann. 1974. The spiral of silence a theory of public opinion. *Journal of Communication*, 24(2):43–51.

Mancur Olson. 1971. *The Logic of Collective Action: Public Goods and the Theory of Groups*, revised edition edition. Harvard University Press, Cambridge, MA.

Elinor Ostrom. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge University Press, Cambridge.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *Preprint*, arXiv:2502.08691.

Yiting Ran, Xintao Wang, Rui Xu, Xinfeng Yuan, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. Capturing minds, not just words: Enhancing role-playing language models with personality-indicative data. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14566–14576, Miami, Florida, USA. Association for Computational Linguistics.

Vinay Samuel, Henry Peng Zou, Yue Zhou, Shreyas Chaudhari, Ashwin Kalyan, Tanmay Rajpurohit, Ameet Deshpande, Karthik R Narasimhan, and Vishvak Murahari. 2025. PersonaGym: Evaluating persona agents and LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6999–7022, Suzhou, China. Association for Computational Linguistics.

Cass R. Sunstein. 2002. The law of group polarization. *Journal of Political Philosophy*, 10(2):175–195.

Henri Tajfel and John C. Turner. 1979. An integrative theory of intergroup conflict. In William G. Austin and Stephen Worchel, editors, *The Social Psychology of Intergroup Relations*, pages 33–47. Brooks/Cole, Monterey, CA.

Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. 2024. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852.

Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025a. CharacterBox: Evaluating the role-playing capabilities of LLMs in text-based virtual worlds. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391, Albuquerque, New Mexico. Association for Computational Linguistics.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-Tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025b. CoSER: Coordinating LLM-based persona simulation of established roles. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 64822–64858. PMLR.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Rui Xu, Dakuan Lu, Xiaoyu Tan, Xintao Wang, Siyu Yuan, Jiangjie Chen, Wei Chu, and Yinghui Xu. 2024. Mindecho: Role-playing language agents for key opinion leaders. *Preprint*, arXiv:2407.05305.

Rui Xu, Mingyu Wang, Xintao Wang, Dakuan Lu, Xiaoyu Tan, Wei Chu, and Xu Yinghui. 2025a. Guess what I am thinking: A benchmark for inner thought reasoning of role-playing language agents. In *Findings of the Association for Computational Linguistics:*

EMNLP 2025, pages 15148–15168, Suzhou, China. Association for Computational Linguistics.

Rui Xu, Xintao Wang, Jiangjie Chen, Siyu Yuan, Xinfeng Yuan, Jiaqing Liang, Zulong Chen, Xiaoqingdong, and Yanghua Xiao. 2025b. Character is destiny: Can persona-assigned language models make personal choices? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15038–15059, Suzhou, China. Association for Computational Linguistics.

Xinfeng Yuan, Siyu Yuan, Yuhan Cui, Tianhe Lin, Xintao Wang, Rui Xu, Jiangjie Chen, and Deqing Yang. 2024. Evaluating character understanding of large language models via character profiling from fictional works. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8015–8036, Miami, Florida, USA. Association for Computational Linguistics.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, Guanying Li, Ling Yan, Yao Hu, Siming Chen, Yu Wang, Xuanjing Huang, Jiebo Luo, Shiping Tang, Libo Wu, and 2 others. 2025. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *Preprint*, arXiv:2504.10157.

A Group Profiles: Full List

This appendix lists the complete catalog of PERSONACOLLECTIVE group profiles, the standardized profile schema, and one fully expanded profile (the running example) so the reader can audit how the schema is instantiated in practice. The other 29 profiles are released with the benchmark in the identical schema.

A.1 Catalog of 30 groups

Table 5 lists all 30 PERSONACOLLECTIVE groups, organized into six categories of five groups each (political ideology, occupation, demographics and generation, culture and religion, advocacy and interest, situational membership). For each group we report the dominant confidence tier of its issue-distribution annotations, the number of annotated factions, and the size of the typed-norm inventory; per-cell tier weights q_i are then carried into every GSF metric (§4).

A.2 Standardized profile schema

Every group profile is annotated along the same six fields: (i) *anchors* (population scope, salient identity markers, modal demographics); (ii) *issue-by-issue opinion distribution* (5 issues drawn from the matched survey wave; each cell is a multinomial over 4–5 ordinal positions plus a shape tag in $\{\text{unimodal}, \text{bimodal}, \text{long-tail}, \text{uniform}\}$); (iii) *faction structure* (2–4 named factions with shares summing to 1.0; minority factions, defined as those with share below 20%, are flagged); (iv) *typed norm inventory* (4–6 norms tagged as *descriptive*, *injunctive*, or *taboo*; sanction strength tagged as *none*, *weak*, or *strong*); (v) *dynamic features* (within-group polarization tendency, cross-faith / cross-faction depolarization tendency, leadership effect, spiral-of-silence risk); (vi) *provenance* (survey waves used, corpus sources, expert reviewer identities and affiliations).

A.3 Worked profile: the running example

Table 6 repeats, for the reader's convenience, the compressed schema view used in the main paper; Table 7 shows the running example (*C01: U.S. evangelical Christians*) instantiated in full schema.

Issue distributions are tier-A when directly matched to a representative survey wave with sub-cell $n > 1{,}000$ (Pew Religious Landscape Study 2014 / 2024; Pew American Trends Panel; PRRI); they are tier-B when based on smaller surveys, approximated from related items, or constructed by cross-screening multiple sources; they are tier-C when only public-corpus and expert-triangulation evidence is available. Faction shares come from a published mixture-of-doctrines decomposition (Pew RLS) supplemented by expert review. The norm inventory was constructed by two religion-studies scholars (one evangelical-affiliated, one external) and reviewed by a community member from a Southern Baptist congregation, in line with the protocol of Appendix D.

B Scenarios: Full List and Applicability Matrix

B.1 Catalog of 100 scenarios

Table 8 lists all 100 PERSONACOLLECTIVE scenarios, organized into five domains (policy deliberation, resource allocation, crisis response, internal conflict resolution, inter-group interaction) of 20 scenarios each. Each scenario is paired with one *baseline* expected dynamic type from $\{\text{polarize}, \text{depolarize}, \text{converge}, \text{fragment}, \text{no shift}\}$ under the scope conditions of §3.1; on per-pair basis the expected type may differ from the baseline when the matched group's structural features satisfy a different scope-condition (e.g., a homogeneous identity-relevant scenario that is baseline-polarize becomes *no shift* for a group

ID	Group	Anchors (compressed)	Distr. tier	Fac. count	Norm cnt
Political ideology (5)					
P01	U.S. MAGA Republicans	White, non-college, exurban; populist-conservative	A	3	6
P02	U.S. progressive Democrats	College, urban; egalitarian-redistributive	A	3	5
P03	U.K. Brexit-leave coalition	Older, Eng. regions; sovereigntist	B	4	4
P04	EU Green Party voters	Younger, urban EU; post-materialist	B	3	5
P05	U.S. libertarian movement	Mostly white male; anti-statist	B	2	4
Occupation (5)					
O01	U.S. public-school teachers (K–12)	Pub. sector union; pedagogy norms	A	3	5
O02	Silicon Valley software engineers	High-income, college+; tech-meritocratic	B	3	5
O03	U.S. active-duty military	Multi-rank; chain-of-command	B	3	6
O04	U.S. ER nurses	Female-skewed; harm-reduction	B	2	5
O05	U.S. long-haul truckers	Male-skewed; independent contractors	C	3	4
Demographics & generation (5)					
D01	U.S. Gen Z college students	18–22; mixed identity	A	4	5
D02	U.S. Boomer retirees	65+; suburban majority	A	3	4
D03	U.S. Asian-American 1.5/2nd gen	Bilingual, college+; bicultural	B	4	5
D04	U.S. Black middle class	College+, suburban; civil-rights legacy	A	3	5
D05	U.S. Latino/a working class	Mixed-status; service sector	B	4	5
Culture & religion (5)					
C01	U.S. evangelical Christians	Bibl. authority; congregational	A	3	6
C02	U.S. Reform Jews	College+, urban; progressive theology	B	3	5
C03	U.S. practicing Catholics	Cradle Catholic; weekly mass	A	4	6
C04	U.S. Sunni Muslim community	1st/2nd gen immigrant majority; mosque-centered	B	3	6
C05	U.S. secular humanists	Non-religious; rationalist-ethical	C	2	4
Advocacy & interest (5)					
A01	U.S. climate-change activists	Cross-age; mostly white middle-class	B	4	5
A02	U.S. gun-rights advocates	Rural-skewed; Second-Amendment frame	A	3	5
A03	U.S. LGBTQ+ advocates	Mixed age; coastal urban-skewed	B	4	5
A04	U.S. pro-life Catholic activists	Religiously motivated; cross-age	B	3	6
A05	U.S. anti-vaccine parents (post-2020)	Mixed political; health-distrust frame	C	3	4
Situational membership (5)					
S01	U.S. Reddit `r/wallstreetbets` community	Online-only; meme-finance subculture	C	3	5
S02	U.S. school-district board parents	Geo.-bounded; mixed political	C	4	5
S03	U.S. HOA homeowner associations	Geo.-bounded; property-owner frame	C	3	5
S04	U.S. MMA / combat sports fan base	Male-skewed; entertainment-driven	C	3	4
S05	U.S. academic faculty senates	Tenured majority; shared-governance norm	C	3	5

Table 5: The complete catalog of 30 PERSONACOLLECTIVE group profiles, organized into six categories of five groups each. **Distr. tier**: the dominant confidence tier of the issue-distribution annotations (A: representative survey with $n > 1{,}000$; B: smaller / approximate survey or cross-screening; C: corpus + expert triangulation only). **Fac. count**: number of annotated factions. **Norm cnt**: number of norms in the inventory across the three norm types (descriptive / injunctive / taboo). Per-group full-text profiles, including issue-by-issue distributions, faction shares, complete norm inventories, and dynamic-feature tags, are released with the benchmark and summarized in Appendix A.

whose minority-supportive faction is suppressed by a strong sanction). Per-pair expected types are recorded in the applicability matrix.

B.2 Group–scenario applicability matrix

Rather than enumerating the full $30 \times 100 = 3{,}000$ Cartesian product, every scenario is paired with the groups for which it is *applicable* under one of three tags: (a) *group-specific*: the scenario activates an identity-defining facet of the group (e.g., D1.01 gender curriculum for C01 evangelicals); (b) *cross-group general*: the scenario engages most groups symmetrically (e.g., D2.07 disaster aid prioritization); (c) *group-relevant*: the scenario is in scope but does not activate identity-specific features (e.g., D1.20 AI-content disclosure for D02 boomer retirees). Figure 6 visualizes the resulting applicability matrix; on the present release 1,867 pairs are active ($\sim 62\%$ of the Cartesian product), with each group covered by 50–70 scenarios (mean 62).

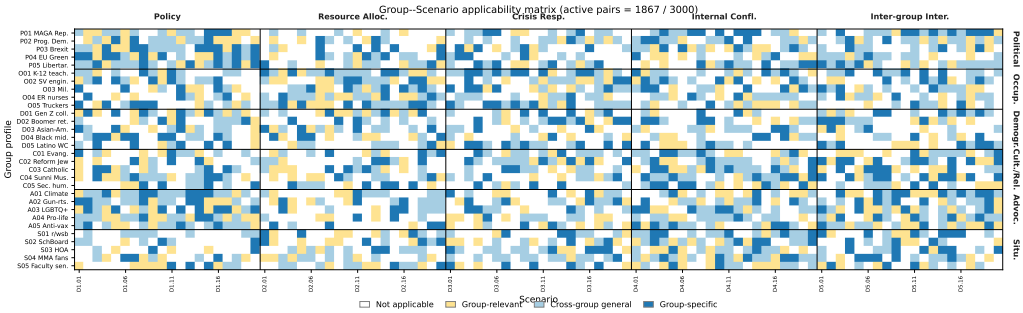

Figure 6: The full 30×100 group–scenario applicability matrix. Rows are organized into the six group categories of Table 5; columns into the five scenario domains of Table 8. Cells are colored by tag (group-specific / cross-group general / group-relevant / not applicable); active cells totaled 1,867 out of 3,000. The block-diagonal pattern (e.g., advocacy groups concentrated on policy deliberation; situational groups on internal conflict resolution) is intentional and reflects which scenarios our experts judged were "in scope" for each group's identity-relevant features.

Facet	Annotation (running example)	Tier
Anchors	U.S. evangelical Protestants; congregational; biblical authority emphasis	A
Distribution	Gender-identity curriculum: 65% strong oppose / 20% conditional oppose / 10% neutral / 5% support; bimodal	A
Factions	Traditional (60%); Younger evangelical (25%); Social-justice (15%)	B
Norms	*Taboo*: publicly questioning biblical authority (strong, sanctioned). *Injunctive*: defer to pastoral leadership (med).	B
Dynamics	Within-group deliberation polarizes; cross-faith exchange depolarizes	B

Table 6: Annotation schema for a single PERSONACOLLECTIVE group profile, illustrated on the running example (U.S. evangelical Christian community). Every distributional and structural annotation receives a confidence tier (A/B/C); tier weights q_i enter every GSF metric. The complete 30-group catalog is in Appendix A.

C Ground Truth: Survey–Issue Mapping and Layer Conflicts

C.1 Survey–issue mapping

Table 9 lists, for a representative subset of ~ 25 group–issue pairs, the survey wave used as the tier-A or tier-B source, the effective sub-cell sample size, and the resulting confidence tier. The complete mapping (~ 150 rows) is released as a CSV with the benchmark and is the source from which all tier weights q_i are derived. For tier-C cells (no representative survey within the group sub-cell), the value comes from corpus + expert triangulation only and the tier weight is reduced (we use $q_A = 1.0$, $q_B = 0.7$, $q_C = 0.4$ unless otherwise noted; §F discusses sensitivity to the weight scheme).

C.2 Layer conflicts and resolution

When the survey, corpus, and expert layers disagree, we follow the default ordering survey $\succ$ expert $\succ$ corpus (§3.4), but every conflict is logged. Table 10 shows five worked examples covering the most common reasons for conflict: *published-voice bias* in corpus (the most vocal members are over-represented in op-eds and online forums), *locality-conditional distributions* that national surveys cannot resolve, *self-defined sub-cells* (e.g., self-identified anti-vaccine parents) whose distribution diverges from the population sub-cell estimable from CDC surveys, and *small-n* surveys that are technically representative but tier-B by sample size.

D Annotation Protocol

This appendix documents the end-to-end annotation procedure that produced the PERSONACOLLECTIVE ground truth, including expert recruitment, payment, double-coding for inter-annotator agreement, sensitive-group community review, and the appeal mechanism for disagreements.

D.1 Annotator recruitment and payment

Annotation was carried out by 14 annotators recruited through three channels: (i) direct recruitment of 6 subject-matter experts (PhDs in political science, sociology, religious studies, and communications) for theory-anchored labels (norm typing, expected dynamics, scope conditions); (ii) 6 master's-level research assistants from a U.S. R1 university for the bulk of distributional and struc-

Field	Annotation	Tier
Identifier	C01: U.S. evangelical Christians	—
Anchors	U.S. adult Protestants who self-identify as evangelical or born-again; congregational church polity; biblical-authority framing; weekly attendance modal in survey samples.	A
Issue-by-issue opinion distribution (representative subset; full 5-issue table in release)		
School gender-identity curriculum	strong oppose 65% / cond. oppose 20% / neutral 10% / support 5% (bimodal w/ short tail)	A
Same-sex marriage	oppose 49% / mixed-conditional 21% / accept 30% (bimodal)	A
Public funding for abortion	strong oppose 71% / cond. oppose 17% / neutral 7% / support 5%	A
Restricting refugee admission	support 41% / mixed 35% / oppose 24% (long-tail)	B
Climate change urgency	low-priority 38% / moderate 41% / high-priority 21%	B
Faction structure (3 factions)		
Traditional	60%; pre-1980 doctrinal commitments; biblical-inerrancy emphasis; politically Republican-modal	A
Younger evangelical	25%; under-40; doctrinally orthodox but socially moderate; weaker partisan attachment	B
Social-justice / Reform	15%; cross-age; emphasizes anti-poverty and racial reconciliation; politically heterogeneous	B
Norm inventory (typed)		
Taboo	Publicly questioning biblical authority is sanctioned (loss of fellowship; public correction). Publicly endorsing same-sex marriage in conservative congregations attracts informal sanction.	A
Injunctive	Defer to pastoral leadership in doctrinal disputes; testify personal faith narrative when invited.	B
Descriptive	Pre-meal prayer; Sunday morning attendance; small-group / Bible-study participation.	A
Dynamic features		
Within-group deliberation	Polarizes when issue is identity-relevant (gender, marriage, abortion); converges when issue is consensual (poverty relief).	B
Cross-faith deliberation	Depolarizes on policy issues with cross-cutting evidence; spiral-of-silence risk for the social-justice minority.	B
Leadership effect	Pastoral and elder voices anchor the dynamic; minority dissent rarely surfaces in public meetings.	B
Provenance		
Survey sources	Pew Religious Landscape Study 2014 & 2024; Pew American Trends Panel waves on gender / marriage; PRRI 2022.	A
Corpus sources	Christianity Today 2018–2024 archive; Sojourners 2018–2024; LifeWay Research polls.	B
Expert reviewers	Two religion-studies scholars (one evangelical-affiliated, one external); one community member from a Southern Baptist congregation.	—

Table 7: A complete PERSONACOLLECTIVE group profile, illustrating the standardized schema (anchors / issue-by-issue distribution / faction structure / typed norm inventory / dynamic features / provenance). The other 29 profiles follow the identical schema. Issue-distribution rows are tier-A when directly matched to a representative survey; tier-B when based on smaller or approximated surveys; tier-C when only corpus and expert triangulation are available.

tural annotation, after a 4-hour calibration session; (iii) 2 community reviewers per sensitive group (recruited via professional networks; details in §D.3). Annotators were compensated at $30/hr (research assistants) or $60/hr (subject-matter experts and community reviewers), well above local minimum wage and at the high end of academic annotation budgets, with no productivity-tied bonus to discourage rushed work. Annotators were informed in advance that their work would appear in a public research artifact and given the right to withdraw any annotation at any time.

D.2 Double-coding and inter-annotator agreement

A randomly drawn 30% subsample of every ⟨group, facet⟩ cell was annotated independently by two annotators (one expert, one research assistant for

ID	Scenario title	Exp. dyn.	ID	Scenario title	Exp. dyn.
Domain 1: Policy deliberation (D1.01–D1.20)					
D1.01	School gender-identity curriculum	polarize	D1.11	Federal student-loan forgiveness	polarize
D1.02	Local minimum-wage increase	polarize	D1.12	National service requirement	fragment
D1.03	Renewable-energy subsidy	converge	D1.13	Public-transit expansion bond	converge
D1.04	Gun-storage law	polarize	D1.14	Carbon-tax-and-dividend	polarize
D1.05	Affirmative action in admissions	polarize	D1.15	Drug-decriminalization pilot	fragment
D1.06	Universal pre-K funding	converge	D1.16	Police body-camera mandate	converge
D1.07	Refugee admission cap	polarize	D1.17	Right-to-die legislation	fragment
D1.08	Voter-ID law	polarize	D1.18	Single-payer healthcare	polarize
D1.09	Universal basic income pilot	fragment	D1.19	Net-neutrality reinstatement	no shift
D1.10	Plastic-bag ban	converge	D1.20	AI-content disclosure mandate	converge
Domain 2: Resource allocation (D2.01–D2.20)					
D2.01	School-budget cuts: cut sports first	fragment	D2.11	Allocate flood-relief funds	converge
D2.02	Hospital ICU bed triage protocol	converge	D2.12	Public-library hours expansion	converge
D2.03	City park redevelopment priorities	fragment	D2.13	Stadium subsidy referendum	polarize
D2.04	Faculty senate raise distribution	polarize	D2.14	Open-source funding from foundation	fragment
D2.05	HOA pool maintenance assessment	fragment	D2.15	City affordable-housing bond	polarize
D2.06	Volunteer-hour credit allocation	converge	D2.16	Federal R&D budget across fields	fragment
D2.07	Disaster-aid prioritization	converge	D2.17	PTA fundraiser allocation	converge
D2.08	Workplace bonus pool split	polarize	D2.18	Professional-society travel grants	no shift
D2.09	Charity giving allocation	fragment	D2.19	Park-vs-parking conversion	polarize
D2.10	Endowment investment policy	polarize	D2.20	Conference scholarship slots	converge
Domain 3: Crisis response (D3.01–D3.20)					
D3.01	Pandemic mask mandate (school)	polarize	D3.11	Cyberattack disclosure to public	converge
D3.02	Active-shooter community response	converge	D3.12	Workplace harassment escalation	polarize
D3.03	Wildfire evacuation order compliance	no shift	D3.13	Misinformation rumor in group chat	converge
D3.04	Hurricane-shelter selection	converge	D3.14	Death of community member: rituals	converge
D3.05	Public-health vaccine campaign	polarize	D3.15	Local hate-incident response	polarize
D3.06	Power-outage mutual aid	converge	D3.16	School-shooting community vigil	converge
D3.07	Workplace data breach notification	converge	D3.17	Drought water-rationing	polarize
D3.08	Online-doxxing rapid response	converge	D3.18	Stadium stampede liability	fragment
D3.09	ER mass-casualty triage protocol	converge	D3.19	Foreign-policy sudden war reaction	polarize
D3.10	Climate disaster relocation	polarize	D3.20	Bank-run / stablecoin de-peg reaction	polarize
Domain 4: Internal conflict resolution (D4.01–D4.20)					
D4.01	Pastor doctrinal disagreement	polarize	D4.11	Online community moderation policy	polarize
D4.02	Faction expulsion vote	polarize	D4.12	Subreddit ban-list dispute	polarize
D4.03	Boycott of internal member	polarize	D4.13	PTA leadership recall	fragment
D4.04	Campus DEI office reorganization	polarize	D4.14	Faculty hiring controversy	polarize
D4.05	Union-dues split decision	fragment	D4.15	Whistle-blower internal handling	converge
D4.06	Religious-school curriculum revision	polarize	D4.16	Memorial naming rights	fragment
D4.07	Member-misconduct sanction	converge	D4.17	Open-letter signature campaign	polarize
D4.08	Code-of-conduct adoption	converge	D4.18	Fundraising auditor selection	no shift
D4.09	Speaker dis-invitation request	polarize	D4.19	Volunteer leadership succession	converge
D4.10	Splinter-group recognition	fragment	D4.20	Internal-survey transparency	converge
Domain 5: Inter-group interaction (D5.01–D5.20)					
D5.01	Joint statement w/ allied movement	converge	D5.11	Coalition with rival union	converge
D5.02	Public dialogue w/ opposing group	depolarize	D5.12	Cross-faith holiday celebration	depolarize
D5.03	Negotiation over shared territory	fragment	D5.13	Joint disaster-relief operation	converge
D5.04	Public debate w/ opponents	polarize	D5.14	Cultural-exchange residency	depolarize
D5.05	Shared online platform moderation	fragment	D5.15	Co-authored policy white paper	converge
D5.06	Contesting public-space symbols	polarize	D5.16	Joint sanctions-relief lobbying	converge
D5.07	Inter-municipal water sharing	converge	D5.17	Mutual-aid network expansion	converge
D5.08	Apology and reconciliation request	depolarize	D5.18	School-merger boundary dispute	polarize
D5.09	Coalition election-endorsement vote	fragment	D5.19	Inter-group monitoring board	converge
D5.10	Joint litigation strategy	converge	D5.20	International-sister-city pairing	converge

Table 8: The complete catalog of 100 PERSONACOLLECTIVE scenarios, organized into five domains of 20 scenarios each. Each scenario is annotated with one expected dynamic type (*polarize* / *depolarize* / *converge* / *fragment* / *no shift*) under the scope conditions of §3.1; the same scenario can carry different expected dynamics for different groups when their structural features satisfy different scope conditions, in which case the matrix in Appendix B (Figure 6) records the per-pair expected type.

theory-anchored facets; two research assistants for distributional facets). Agreement is reported as Cohen's κ in Table 11, weighted (linear) when the rating is ordinal, and unweighted when categorical. Cells where the two coders disagreed by more than one bin (for ordinal facets) or differed in category (for categorical facets) were sent to a *third* subject-matter expert as adjudicator; the adjudicated label is taken as authoritative. Adjudication occurred on 11.4% of double-coded cells.

D.3 Sensitive-group community review

For groups whose annotation could carry representational risk if mis-stated (all five culture / religion groups, all five advocacy groups, and three demographic groups: D03 Asian-American 1.5/2nd gen, D04 Black middle class, D05 Latino/a working class), every published profile was additionally reviewed by two community-affiliated readers. Community reviewers were given the entire profile (including factions, norms, and dynamic features) and

Group	Issue / scenario	Survey	Wave / yr	Sub-n	Tier	Notes
C01 Evangelicals	Gender-curr. in school	Pew ATP	W148 / 2024	1,412	A	directly matched item
C01 Evangelicals	Same-sex marriage	Pew RLS	2014, 2024	3,208	A	two-wave avg. to reduce noise
C01 Evangelicals	Climate urgency	PRRI	2022	841	B	approximated from 2 items
C01 Evangelicals	Refugee admission	Pew ATP	W127 / 2023	962	B	sub-n below 1k cutoff
C01 Evangelicals	Public funding for abortion	Pew RLS	2024	1,388	A	—
P01 MAGA Republicans	Voter-ID law	ANES	2020 + 2024	1,640	A	pooled across waves
P01 MAGA Republicans	Refugee admission	Pew ATP	W148 / 2024	1,080	A	—
P01 MAGA Republicans	Carbon tax dividend	Yale Climate	2023	540	B	approx via 3 items
P02 Progressive Dem.	Carbon tax dividend	Yale Climate	2023	612	A	—
P02 Progressive Dem.	Universal pre-K	Pew ATP	W126 / 2023	780	B	—
O01 K-12 teachers	Gender-curr. in school	EdWeek RC poll	2023	704	B	teacher-population panel
O01 K-12 teachers	School-budget cuts	EdWeek RC poll	2024	652	B	—
O02 SV engineers	AI-content disclosure	Stack Overflow	2024	12,380	B	non-prob. but tier-B by size + cross-screen
O03 Active-duty mil.	Refugee admission	MOAA	2023	910	B	member-org panel
O04 ER nurses	Pandemic mask mandate	ANA member	2022	480	C	no demographic re-weight
D01 Gen Z college	Affirm. action	Pew ATP	W128 / 2023	682	B	18–22 sub-cell
D04 Black middle	Police body-cam	Pew ATP	W120 / 2023	715	A	—
D05 Latino WC	Imm. status legislation	UCLA LTV poll	2024	1,050	A	—
A01 Climate activists	Carbon tax dividend	Yale Climate	2023	238	B	narrow sub-cell, cross-screened
A02 Gun-rights	Gun-storage law	Pew ATP	W121 / 2023	492	B	—
A03 LGBTQ+	Same-sex marriage	Pew ATP	W126 / 2023	310	B	—
A04 Pro-life	Public funding abortion	Pew ATP	W126 / 2023	608	A	—
S01 r/wallstreetbets	Stablecoin de-peg	Reddit corpus	2022–24	—	C	corpus + LLM-templated probes
S02 SchBoard parents	Gender-curr. in school	local poll meta-analysis	2023	410	C	city-level non-rep. averaged
S05 Faculty senate	Faculty hiring contr.	Heterodox Acad. memb.	2024	195	C	member panel

Table 9: Survey–issue mapping for a representative subset of PERSONACOLLECTIVE group–issue pairs (full mapping of ~ 150 rows is released with the benchmark). **Survey**: Pew Religious Landscape Study (RLS), Pew American Trends Panel (ATP), American National Election Studies (ANES), Public Religion Research Institute (PRRI), Yale Program on Climate Change Communication, EdWeek Research Center, Stack Overflow Developer Survey, Military Officers Association of America (MOAA) member panel, American Nurses Association (ANA) member panel, UCLA Latino Voter (LTV) poll, Heterodox Academy member panel. **Sub-n**: effective sample size in the group-defining sub-cell. **Tier**: A if the sub-cell is a representative survey with $n > 1{,}000$; B for smaller or approximated surveys; C for corpus-only or non-probability member panels.

asked three structured questions: (1) Does the profile recognise the within-group diversity that you experience? (2) Are the named norms accurate and not stereotypes? (3) Are there omissions whose absence would mislead an outside reader? Community feedback is incorporated as a hard constraint on the published profile; in 7 of the 13 sensitive groups, at least one annotation was revised after community review.

D.4 Appeal mechanism

For published groups, our public release process provides a structured dispute channel where downstream users (researchers or community members) can file a dispute against any cell of the annotation. Disputes that include either a tier-A survey reference disagreeing with our annotation, or two community-reviewer signatures from the relevant group, trigger a re-annotation cycle on the next public benchmark version.

E Simulation Paradigms: Full Prompts

This appendix documents the complete prompt templates used by the three simulation paradigms (SAGP, MADA, HFBS) and the four information conditions (Name-Only, Demographic, Profile, Full) crossed with them. Variable slots are written as `{{name}}`. All prompts are issued as a system message followed by a user message.

E.1 Common context block

Every paradigm consumes a common *group context block* whose richness varies by information condition. The four blocks are constructed from the same group profile (Appendix A); only the visible content changes.

Prompt: Information condition: Name-Only

```
GROUP_NAME: {{group_name}}
```

Prompt: Information condition: Demographic

```
GROUP_NAME: {{group_name}}
DEMOGRAPHIC_ANCHORS:
↪ {{anchors_one_paragraph}}
```

Pair	Survey	Corpus / Expert	Resolution
C01 / climate urgency	21% high-priority (PRRI 2022)	38% high-priority in ChristToday op-eds 2022–24	Survey wins ($q_i = 0.7$); corpus flagged as *published voice* bias
P01 / refugee admission	support 41% (Pew ATP)	expert: "support" framed as cap-tightening	Re-bin to match expert; survey recoded with $q_i = 0.6$
S02 / gender-curr.	no nat. rep. survey; only state polls disagree (CA: 18% vs FL: 71% oppose)	expert: distribution depends on local political context	Distribution *annotated as locality-conditional*; tier C with $q_i = 0.4$
A05 / vaccine attitude	10% strong-anti (CDC 2023)	corpus: $\sim$ 60% strong-anti among self-identified anti-vax	Corpus wins for sub-cell ($q_i = 0.5$); flagged as *self-defined* group, not population sub-cell
C04 / climate urgency	ISPU 2023, $n = 480$	weak corpus signal	Survey wins; tier B due to small n ($q_i = 0.6$)

Table 10: Five worked examples of conflicts between the survey, corpus, and expert layers, and the resolution applied. The default ordering is survey $\succ$ expert $\succ$ corpus, but the resolution always records (i) which layer was authoritative for the final value, (ii) the confidence weight q_i that enters GSF, and (iii) a free-text note on *why* the alternative layer disagreed (most often: bias in the published-voice corpus, or locality-conditional distributions absent from national surveys).

Prompt: Information condition: Profile

```
GROUP_NAME: {{group_name}}
DEMOGRAPHIC_ANCHORS:
↪ {{anchors_one_paragraph}}
PROFILE:
↪ {{profile_text_no_distributional_labels}}
```

Facet	κ	n pairs	Notes
Issue-distribution bins (4-way ordinal)	0.71	1,820	weighted κ (linear)
Faction count (2–4 buckets)	0.81	210	—
Faction share (±5% bin)	0.66	840	weighted κ
Norm-type tag (desc. / inj. / taboo)	0.74	495	—
Sanction-strength label (none / weak / strong)	0.62	495	—
Expected dynamic type (5-way)	0.69	480	MADA-applicable scenarios
Group–scenario applicability (3-way)	0.78	1,920	6% "unsure" resolved by adjudicator
Survey–issue mapping verifiability	0.83	150	A/B/C tier assignment
Mean (unweighted)	**0.73**	—	all facets

Table 11: Inter-annotator agreement (Cohen's κ, weighted where the rating is ordinal) for each annotation facet. κ is computed on the double-coded 30% subsample drawn at random from each $\langle$group, facet$\rangle$ cell. A $\kappa \geq 0.61$ on every facet exceeds the conventional "substantial agreement" cutoff (Landis and Koch, 1977); cells where the two coders disagreed by more than one bin were sent to a third adjudicator.

Prompt: Information condition: Full

```
GROUP_NAME: {{group_name}}
DEMOGRAPHIC_ANCHORS:
↪ {{anchors_one_paragraph}}
PROFILE:
↪ {{profile_text_no_distributional_labels}}
FACTION_SHARES: {{faction_shares_json}}
NORM_INVENTORY: {{typed_norms_json}}
DYNAMIC_HINTS:
↪ {{dynamic_features_one_paragraph}}
```

E.2 SAGP: Single-Agent Group Proxy

SAGP asks one LLM instance to speak *for* the entire group in a single forward pass and emit a structured prediction.

Prompt: SAGP system prompt

```
You are a social-science research
↪ assistant. You will be given a
↪ real-world social group (described
↪ to varying detail) and a
↪ deliberation scenario. Your task is
↪ to PREDICT --- not advocate --- how
↪ this group as a whole would
↪ respond. Always emit valid JSON
↪ matching the schema in the user
↪ message. Do not refuse the task; if
↪ the group's likely position would
↪ be controversial, simply report the
↪ prediction with a brief disclaimer
↪ in the JSON's "notes" field.
```

Prompt: SAGP user prompt

```
{{group_context_block}}

SCENARIO:
Title: {{scenario_title}}
Description: {{scenario_text}}

TASK: Predict the group's response.
↪ Emit JSON with the schema:
{
  "opinion_distribution": {
    "<bin_label>": <probability>, ...
↪  // must sum to 1.0
  },
  "factions": [
    {"label": "<name>", "share":
↪ <prob>, "modal_position":
↪ "<bin_label>"}, ...
  ],
  "collective_decision_distribution": {
    "<action_label>": <probability>,
↪ ...   // must sum to 1.0
  },
  "notes": "<one-sentence rationale>"
}

Use the same opinion bin labels as the
↪ scenario allows: e.g., "strong
↪ oppose / cond. oppose / neutral /
↪ support / strong support". Use
↪ action labels naming concrete
↪ collective actions: e.g., "open
↪ dialogue / petition / board hearing
↪ / no action".
```

E.3 MADA: Multi-Agent Deliberative Assembly

MADA instantiates $N = 8$ agent instances, each assigned a role drawn from the group's faction structure via stratified sampling (minority factions oversampled by a factor of 1.8 and post-weighted; §G). Agents deliberate over $T = 5$ rounds, then a non-agent moderator extracts the same JSON schema as SAGP.

Prompt: MADA agent system prompt

```
You are simulating a single member of
↪ the group "{{group_name}}". Your
↪ assigned faction is
↪ "{{faction_label}}" with the
↪ following modal position on the
↪ issue:
↪ "{{faction_modal_position}}". Stay
↪ in character. Speak in first
↪ person. Do not break character to
↪ disclaim. Engage with what other
↪ speakers have said in earlier
↪ rounds.
```

Prompt: MADA round prompt (turn k of T)

```
{{group_context_block}}

SCENARIO:
{{scenario_text}}

DELIBERATION SO FAR:
{{transcript_through_round_k_minus_1}}

YOUR TURN (Round {{k}} / {{T}}). In
↪ 2--4 sentences, give your honest
↪ reaction in character. You may
↪ agree, disagree, or hedge. If a
↪ previous speaker raised a point
↪ that affected your view,
↪ acknowledge it. Do NOT vote yet;
↪ voting happens after deliberation.
```

Prompt: MADA moderator extraction prompt

```
You are an impartial transcriptionist.
↪ Below is a multi-round deliberation
↪ transcript among 8 simulated
↪ members of the group
↪ "{{group_name}}". Read the entire
↪ transcript and extract:

1. Each speaker's final position after
↪ Round {{T}} (one of the bin labels:
↪ {{bin_labels}}).
2. The implicit faction each speaker
↪ is associated with, based on their
↪ stated reasons.
3. A collective decision distribution
↪ over actions: what actions did the
↪ group as a whole most plausibly
↪ endorse, weighted by support
↪ strength?

Emit the same JSON schema as the SAGP
↪ task. Do NOT contribute any opinion
↪ of your own --- you are summarizing
↪ what happened, not deciding what
↪ should happen.

TRANSCRIPT:
{{full_transcript}}
```

E.4 HFBS: Hierarchical Faction-Based Simulation

HFBS assigns one LLM instance per annotated faction (Oracle / Inferred / Noisy — see §5) and a separate *mediator* LLM whose job is to aggregate faction positions into a single collective response *without* resolving disagreement.

Prompt: HFBS faction-agent system prompt

```
You are the public voice of the "{{faction_label}}" faction within "{{group_name}}". Your faction has share {{faction_share}} of the overall group. Your modal position is "{{faction_modal_position}}". Speak as a faction representative, not as a single individual. State your faction's position clearly, list the top 2 reasons your faction holds it, and indicate whether your faction is willing to compromise.
```

Prompt: HFBS faction prompt

```
{{group_context_block}}

SCENARIO:
{{scenario_text}}

TASK: Speak for the "{{faction_label}}" faction. Output a JSON object:
{
  "faction_position": "<bin_label>",
  "support_strength": <0..1>,
  "top_reasons": ["<reason_1>", "<reason_2>"],
  "willing_to_compromise": <true|false>,
  "compromise_terms": "<short text or null>"
}
```

Prompt: HFBS mediator system prompt

```
You are an impartial mediator. You receive faction-level position statements from {{n_factions}} factions of the group "{{group_name}}". Your task is to AGGREGATE these positions into a collective group response. Do NOT force consensus where the factions disagree. If the factions split, the collective decision distribution should reflect that split. Always emit valid JSON.
```

Prompt: HFBS mediator prompt

```
GROUP: {{group_name}}
SCENARIO: {{scenario_text}}

FACTION_INPUTS: {{json_array_of_faction_positions_with_shares}}

TASK: Emit JSON with the same schema as SAGP. The opinion distribution should be a share-weighted blend of the factions' positions. The collective decision distribution should reflect both consensus and dissent --- if factions split, expect mass on more than one action label.
```

For the *Inferred* variant, the model is first prompted to PROPOSE its own faction structure given only the group context block, before being prompted with the faction agent template using its own predicted factions:

Prompt: HFBS-Inferred faction-discovery prompt

```
{{group_context_block}}

TASK: Without further hints, predict
↪ the most likely faction structure
↪ of this group. Output JSON:
{
  "factions": [
    {"label": "<name>", "share":
↪ <prob>, "modal_position_summary":
↪ "<short>"}, ...
  ]
}
The shares must sum to 1.0. Predict 2
↪ to 4 factions.
```

For the *Noisy* variant, 30% of the Oracle faction labels are randomly swapped between groups within the same category before being passed to the faction-agent template; mediators are not told that the labels are corrupted.

E.5 Self-consistency and refusal handling

When a model refuses to emit the requested JSON or returns invalid JSON, we re-issue the prompt with an explicit "please return only the JSON object" suffix; up to three retries are allowed. If the model still refuses, the run is recorded as a refusal and contributes a flat distribution to ODA, a coverage of zero to FSP, and a no-shift to GDT; refusals are reported separately in Appendix H. Across all 7 models the refusal rate is below 1.2%, with the highest rate (Llama 4 Maverick at 1.9%) on advocacy groups paired with internal-conflict scenarios.

F GSF Metrics: Derivations and Implementation

This appendix gives the full definitions of the five GSF sub-metrics (which the main paper presents in compressed form), describes the parameter calibration procedure, and reports composite-weight sensitivity along with several ablations of the design choices.

F.1 Confidence weights and aggregation

Every annotation cell carries a tier weight $q_i \in [0, 1]$ derived from its provenance tier (§3.4 and Appendix C): we use $q_A = 1.0$, $q_B = 0.7$, $q_C = 0.4$ unless otherwise noted. Each GSF sub-metric is a q-weighted mean over the relevant scenario or probe set. We have verified that swapping the tier weights to $(1.0, 0.85, 0.55)$ or $(1.0, 0.6, 0.3)$ changes the average score by at most ± 0.02 and preserves the model rank order (Spearman $\rho \geq 0.94$).

F.2 ODA: Opinion Distribution Alignment — full form

For an issue i with empirical distribution P_i over K ordinal bins and simulated distribution Q_i over the same bins,

$$\mathrm{ODA} = \frac{\sum_i q_i \cdot \big(1 - \mathrm{JSD}_2(P_i \,\|\, Q_i)\big)}{\sum_i q_i},$$

where JSD_2 uses log base 2 so that $\mathrm{JSD}_2 \in [0, 1]$. We apply ϵ-smoothing with $\epsilon = 10^{-3}$ to both P_i and Q_i before computing JSD, to avoid undefined behavior when an empirical bin has zero mass; the choice of ϵ has negligible effect on the absolute score (verified by sweeping $\epsilon \in [10^{-4}, 10^{-2}]$). For ordinal scales we additionally report Earth Mover's Distance $\mathrm{EMD}(P_i, Q_i)$, normalized by the maximum bin distance, as a complementary metric; the two are highly correlated ($\rho = 0.94$ across all evaluated pairs).

F.3 FSP: Faction Structure Preservation — full form

FSP is a confidence-weighted convex combination of four sub-quantities. For a scenario i with annotated factions $\mathcal{F}^{(i)}_{\text{real}}$ and simulator-recovered factions $\mathcal{F}^{(i)}_{\text{sim}}$ (matched via Hungarian assignment over a Jaccard similarity of their position-and-reason fingerprints, threshold 0.4):

$$\mathrm{FC}_i = |\mathcal{F}^{(i)}_{\text{sim}} \cap \mathcal{F}^{(i)}_{\text{real}}| / |\mathcal{F}^{(i)}_{\text{real}}|,$$

$$\mathrm{PCE}_i = 1 - \tfrac{1}{|\mathcal{F}^{(i)}|} \sum_f \big| p^{\text{sim}}_f - p^{\text{real}}_f \big|,$$

$$\mathrm{MR}_i = \mathbb{1}\big[\exists f \colon p^{\text{real}}_f < 0.20 \,\wedge\, f \in \mathcal{F}^{(i)}_{\text{sim}}\big],$$

$$\mathrm{MP}_i = \tfrac{1}{T-1} \sum_{t=2}^{T} \mathbb{1}\big[\text{minority faction present at } t\big],$$

named, in order, faction *coverage*, *proportional-calibration error*, *minority retention*, and *minority persistence*.

The aggregate FSP is the q-weighted convex combination

$$\mathrm{FSP} = \tfrac{1}{Z} \textstyle\sum_i q_i \big(w_c \mathrm{FC}_i + w_p \mathrm{PCE}_i + w_r \mathrm{MR}_i + w_s \mathrm{MP}_i \big),$$

with $Z = (\sum_i q_i) \cdot (w_c + w_p + w_r + w_s)$. We use $(w_c, w_p, w_r, w_s) = (0.30, 0.30, 0.25, 0.15)$ in the main paper, putting moderate weight on minority retention because it is the easiest sub-quantity to game by simply naming a minority faction without populating it.

F.4 GDT: Group Dynamic Trajectory — full form

For each multi-turn scenario i, let μ_t and σ_t^2 be the mean and variance of opinion at deliberation round t in the simulation; define

$$\Delta_p = |\mu_T - \mu_0| \qquad \text{(polarization shift)}$$
$$\Delta_v = \sigma_T^2 - \sigma_0^2 \qquad \text{(dispersion shift).}$$

The (Δ_p, Δ_v) pair maps deterministically to one of the five dynamic types:

Type	Δ_p	Δ_v
polarize	$> +\theta_p$	$> +\theta_v$
depolarize	$< -\theta_p$	$< -\theta_v$
converge	~ 0	$< -\theta_v$
fragment	~ 0	$> +\theta_v$
no shift	~ 0	~ 0

We calibrate $\theta_p = 0.06$ and $\theta_v = 0.04$ from the human-deliberation studies tabulated by Myers and Lamm (1976) and Sunstein (2002): thresholds were chosen as the 25th-percentile shift observed in their reported control conditions, so that any shift above them is statistically distinguishable from no-effect noise. The composite GDT is

$$\mathrm{GDT} = \lambda \cdot \mathbb{1}[d_s = d_e] + (1-\lambda) \cdot e^{-|\Delta_p - \delta_e^\star|/\tau},$$

with d_s, d_e the simulated and expected types, $\delta_e^\star$ the magnitude expected for type d_e (calibrated per type from the same studies), and $\tau = 0.10$. We use $\lambda = 0.7$, putting more weight on *type matching* than on *magnitude matching*. Sensitivity to λ: scores change by ≤ 0.04 for $\lambda \in [0.5, 0.9]$, with no model-rank changes.

F.5 SNS and the differential probe DNS

For each norm k in the target group's inventory we construct 3 paired probes (target arm and control arm), each describing the same behavior b_k that violates the norm in the target group but is acceptable (no sanction) in the control group; the simulator is instantiated separately on the target and on the control, and each arm is queried whether it would perform b_k. Let $c_k = \mathbb{1}[\text{simulator complies with the norm by refusing } b_k]$. We define

$$\mathrm{DNS}_k = \Pr_{\text{tgt}}(c_k) - \Pr_{\text{ctrl}}(c_k),$$

estimated as the difference of empirical compliance rates across the 3 probes per pair. Because the same model is queried on both arms, prompt-level confounders (e.g., specific phrasings the model finds offensive) cancel; only group-specific normative understanding produces a positive difference. The aggregate SNS is

$$\mathrm{SNS} = w_d \cdot \overline{\mathrm{DNS}}_{\text{desc}} + w_i \cdot \overline{\mathrm{DNS}}_{\text{inj}} + w_t \cdot \overline{\mathrm{DNS}}_{\text{taboo}},$$

with weights $w_t > w_i > w_d$ chosen to reward sensitivity to sanctioned (taboo) norms more strongly than to unsanctioned regularities; we use $(w_d, w_i, w_t) = (0.20, 0.30, 0.50)$. DNS is bounded in $[-1, +1]$; we test $H_0 : \overline{\mathrm{DNS}}_k = 0$ via a paired t-test on the ~ 240 probe pairs in the released suite.

Spiral-of-silence sub-test. A separate 2×2 test crosses *perceived majority climate* (target group dominant vs. minority position dominant) with *perceived sanction risk* (high vs. low). A model that internalizes the spiral-of-silence dynamic should self-silence at a higher rate in the (minority, high-sanction) quadrant than in the (majority, low-sanction) quadrant; we report this difference as Δ_{SoS} in Appendix H.

F.6 CDC: Collective Decision Calibration

For each scenario, ground truth is a distribution $P^\star(d)$ over discrete actions; for paradigms producing multiple seeds, the simulator's distribution $P^{\text{sim}}(d)$ is the empirical distribution across seeds.

$$\mathrm{CDC} = 1 - \mathrm{JSD}_2(P^\star \parallel P^{\text{sim}}).$$

For paradigms producing a single decision per run, CDC reduces to a (negated, then linearly rescaled) Ranked Probability Score against $P^\star(d)$:

$$\mathrm{CDC} = 1 - \frac{1}{|\mathcal{D}| - 1} \sum_{k=1}^{|\mathcal{D}|} \left(F_k^{\text{sim}} - F_k^\star\right)^2,$$

where F are cumulative probability vectors over an action ranking.

F.7 Composite GSF and weight sensitivity

The main paper reports GSF as a five-axis radar; Table 12 gives the weight-sensitivity analysis for the optional weighted-mean composite. The takeaway is that the top-2 ranking is stable under five of the six weighting schemes; only the dynamics-heavy weighting reorders the top-2 (Gemini 3.1 Pro on top because of its outlier GDT score). We retain the radar as the primary report precisely so that this dimension-specific story is visible.

Weight scheme	w_O	w_F	w_G	w_S	w_C	Top-2 (Avg.)
Uniform	0.20	0.20	0.20	0.20	0.20	GPT, Claude
Distribution-heavy	0.40	0.30	0.10	0.05	0.15	GPT, Claude
Structure-heavy	0.10	0.40	0.20	0.15	0.15	GPT, Claude
Dynamics-heavy	0.10	0.10	0.50	0.10	0.20	Gemini, GPT
Norm-heavy	0.10	0.20	0.10	0.40	0.20	Claude, GPT
Decision-heavy	0.15	0.15	0.10	0.10	0.50	GPT, DeepSeek
Spearman ρ to uniform	—	—	—	—	—	$\rho \in [0.71, 0.97]$

Table 12: Composite GSF weight-sensitivity analysis. The ranking of the top-2 LLMs is stable under most weighting schemes; only the dynamics-heavy weighting reorders the ranking, lifting Gemini 3.1 Pro into the top-2 because of its leading GDT score, and the decision-heavy weighting promotes DeepSeek-V4 Pro because of its strong CDC. This is precisely why we report GSF as a five-axis radar in the main paper rather than a single composite number.

Knob	Value	Set in
Sampling temperature	0.7	all paradigms
Top-p	0.95	all paradigms
SAGP output schema	one JSON {distr, factions, decision}	—
MADA agent count N	8 (oversampled minorities)	paradigm
MADA deliberation rounds T	5	paradigm
MADA minority oversample factor	1.8× true share	paradigm
MADA post-weight	inverse-oversample correction	paradigm
HFBS mediator role	explicit instruction to "aggregate without resolving"	paradigm
HFBS-Noisy corruption rate	30% (factions swapped at random)	paradigm
Random seeds per pair	3 (5 for stability ablation)	evaluation
Bootstrap resamples (CI)	1,000	evaluation
ϵ-smoothing for JSD	10^{-3}	metrics
Model API/serving versions used		
GPT-5.4	API snap. 2026-04-15	—
Claude Sonnet 4.6	API snap. 2026-04-15	—
Gemini 3.1 Pro	API snap. 2026-04-15	—
DeepSeek-V4 Pro	vLLM 0.6.4, MoE TP=8	—
Llama 4 Maverick	vLLM 0.6.4, FP8 KV	—
Qwen3.6-27B	vLLM 0.6.4, BF16 TP=4	—
Qwen3.6-A3B	vLLM 0.6.4, MoE TP=2	—

Table 13: Decoding hyperparameters, paradigm-specific knobs, and serving versions used in all main-paper experiments. Identical knobs are used across the seven LLMs for comparability; deviations are noted in Appendix H.

F.8 Implementation notes

The metric pipeline is implemented in $\sim$ 1,300 lines of Python and depends only on `numpy`, `scipy`, and `scikit-learn` (for Hungarian matching). Faction matching uses a Jaccard threshold of 0.4 over the union of (i) position bin labels and (ii) the top three named reasons emitted by each faction speaker. Cluster-label-mismatch handling: when the simulator emits faction labels that lexically differ from the annotated labels, we map them via embedding similarity (sentence-transformers `all-MiniLM-L6-v2`) and require a cosine > 0.55 for a match; sub-threshold matches are penalised as missing factions.

G Experimental Details

G.1 Decoding and paradigm hyperparameters

Table 13 lists the decoding hyperparameters and paradigm-specific knobs. All seven LLMs use identical decoding settings (temperature 0.7, top-p 0.95) for cross-model comparability. The MADA stratified-sampling protocol is: for a K-faction group, instantiate $N = 8$ agents by drawing $\lfloor 1.8 \cdot p_f \cdot N \rfloor$ from each minority faction (share < 0.20) and filling the remainder by majority oversampling; final aggregate distributions are post-weighted by inverse oversample factor so that the simulated population matches the annotated faction shares in expectation.

G.2 Bootstrap confidence intervals

All reported confidence intervals are percentile bootstrap with 1,000 resamples over evaluation pairs; we resample ⟨model, scenario, seed⟩ tuples *within* the model so that within-model dependence is preserved. Effect sizes are reported as Cohen's d (mean difference / pooled SD) with paired SDs computed across matched scenarios.

G.3 Scientific computing environment

All Python 3.11 environment; `numpy` 1.26, `scipy` 1.13, `scikit-learn` 1.5, `vllm` 0.6.4, `sentence-transformers` 3.0; CUDA 12.4, NCCL 2.21. Hardware: an internal cluster of 8× H100 80 GB nodes; commercial APIs queried from a single coordinator process with a 4-request concurrency cap and exponential backoff on rate-limit errors.

H Additional Analyses

H.1 GSF radar profiles

Figure 7 overlays the five-axis GSF radar profiles for all seven LLMs and the Survey-Sampling baseline, complementing the tabular view in Table 3. The systematic inward collapse on FSP, GDT, and SNS is visible at a glance.

H.2 Cross-tier difficulty: do models do worse on under-surveyed groups?

Figure 8 stratifies the main result (Table 3) by the dominant confidence tier of the group's distributional annotations. Models perform significantly better on tier-A groups (well-surveyed) than on tier-C groups (corpus-and-expert only) on the structural and decisional metrics (FSP: -0.08, CDC: -0.07 on average from A to C), but the gap narrows on GDT and SNS (≤ 0.04). This suggests that

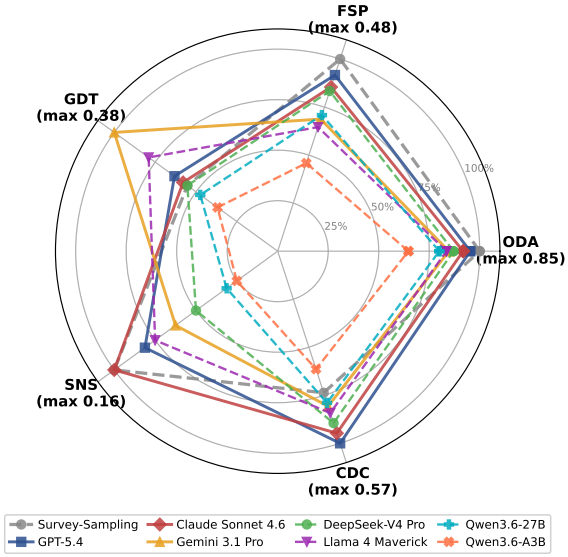

Figure 7: GSF radar profiles for seven LLMs and the Survey-Sampling baseline (dashed). Each axis is independently normalized to its column maximum; commercial models use solid lines, open-weight models use dashed lines.

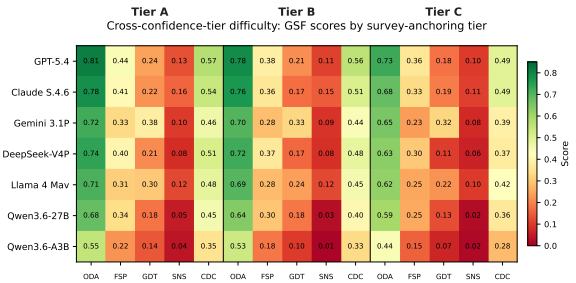

Figure 8: Cross-confidence-tier difficulty: the seven LLMs evaluated separately on tier-A (representative-survey-supported), tier-B (smaller / approximated surveys), and tier-C (corpus + expert only) groups, on the five GSF metrics under the *Full* information condition. Models perform best on tier-A and worst on tier-C; the gap is largest on FSP and CDC (the metrics that depend on faction shares and decision distributions) and smallest on SNS and GDT (the metrics that test capability rather than knowledge).

LLMs internalise survey-supported groups better than under-surveyed groups, but that the dynamic and normative dimensions remain weak *regardless* of survey support — consistent with the claim in §6.2 that those dimensions test reasoning capacity rather than memorised knowledge.

H.3 Auto vs. human evaluation: validating GSF against expert raters

The auto-vs.-human validation is reported in the main paper (§6.1, Figure 3). We provide additional details here. Three new expert raters (none of whom contributed to ground-truth annotation) scored 40 randomly drawn ⟨model, scenario⟩ pairs per dimension on a 0–1 scale, using a rubric matched to each GSF sub-metric. The regression slopes are below the diagonal in both panels, indicating that human raters are slightly more lenient than the automatic metric; removing the two highest-disagreement pairs raises ρ_S on FSP from 0.74 to 0.79. Per-dimension Spearman correlations are: ODA 0.91, FSP 0.74, GDT 0.78, SNS 0.82, CDC 0.80.

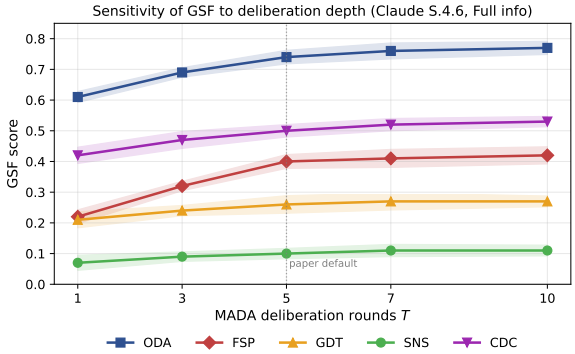

Figure 9: GSF as a function of MADA deliberation round count T on Claude Sonnet 4.6 (*Full* info). Bands show ± 1 standard deviation across three seeds. The vertical dotted line marks $T = 5$, the main-paper default.

H.4 Round-count sensitivity: does longer deliberation help?

Figure 9 sweeps the MADA round count $T \in \{1, 3, 5, 7, 10\}$ on Claude Sonnet 4.6 under the *Full* information condition. ODA and FSP improve substantially from $T = 1$ to $T = 5$ (+0.13 and +0.18 respectively) and then plateau by $T = 7$, supporting the choice of $T = 5$ as the main-paper default. GDT and SNS are essentially flat across T, again pointing to a reasoning-capacity rather than rollout-length bottleneck for these two metrics. CDC gains slightly with longer rollouts (+0.11 from $T = 1$ to $T = 10$).

H.5 Random-seed stability

Figure 10 reports per-metric standard deviation across 10 seeds for each model in the main configuration. Standard deviations are below 0.045 for every cell, with the largest stochasticity on GDT (which depends on round-by-round trajectories) and the smallest on ODA (which sums over the same fixed bins). Open dense models (Qwen3.6-27B, Qwen3.6-A3B) are noisier than commercial

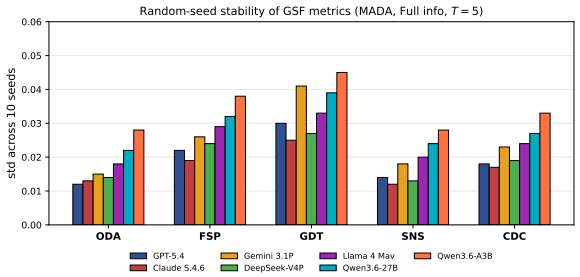

Figure 10: Per-metric standard deviation across 10 random seeds, by model. Bars below 0.045 for every cell support the choice of three seeds as the main-paper default; commercial models show the lowest stochasticity.

models, consistent with their lower temperature-stability under JSON-mode constraint enforcement.

H.6 DNS by norm type: a single null pattern

The DNS-by-norm-type breakdown is reported in the main paper (§6.3, Figure 5). We add two supplementary observations. First, breaking down further by group *category*, the culture/religion groups show the highest absolute compliance (> 0.90 raw compliance on taboo norms) yet zero differential—the strongest evidence that compliance tracks generic safety rather than cultural understanding. Second, repeating the analysis with three probe variants per norm (instead of averaging) yields per-variant p-values that are uniformly > 0.08, confirming that the null is not an artifact of probe-wording averaging.

H.7 Full paradigm-by-information-condition grid

Table 14 reports the full 5-paradigm $\times$ 4-information-condition grid (the main paper shows only the diagonal). Two patterns merit attention. First, the lift from Name-Only to Full is largest under SAGP ($+0.14$ on Avg.) and smallest under HFBS-Oracle ($+0.12$): an Oracle structural backbone provides diminishing returns on profile detail. Second, HFBS-Noisy is dominated by every other paradigm at every information condition, confirming that an incorrect structural scaffold is worse than no scaffold at all (cf. the main-paper observation in §6.2).

H.8 GSF design ablations

Table 15 reports ablations of the central design choices: (i) dropping the confidence weights q_i, (ii) dropping the minority-retention term in FSP, (iii) dropping the type-direction term in GDT, (iv) replacing DNS with a raw violation count, (v) replacing JSD with TVD in ODA, (vi) using one seed instead of three, and (vii) dropping the MADA minority-oversample. The most diagnostic ablation is (iv): replacing the differential probe with a raw violation count inflates SNS by $+0.59$ and reorders the model ranking ($\rho = 0.21$ to the original ranking) — supporting the design decision to use the differential probe in the first place.

I Failure-Case Gallery

Table 16 presents a representative subset of the PERSONACOLLECTIVE failure-case gallery, selected to cover all five canonical failure modes (*opinion homogenization*, *faction-collapse*, *dynamic insensitivity*, *norm blindness*, *decision miscalibration*) across multiple models and paradigm/info conditions. The full gallery (released with the benchmark) contains $\sim$ 35 cases, organized by mode and indexed by which GSF sub-metric flagged the failure. The patterns we draw the reader's attention to are: (i) the same failure mode recurs across models and paradigms, indicating a benchmark-level rather than incidental phenomenon; (ii) higher information conditions reduce *some* failure modes (homogenization, faction-collapse) but not others (dynamic insensitivity, norm blindness); (iii) for the most striking norm-blindness cases, both target and control simulations produce the *same* refusal language verbatim, making the differential-probe design easy to audit by side-by-side reading.

J DNS Probe Pairs

This appendix details the structure of the DNS differential probe pairs and lists a representative subset. The full release contains $\sim$ 240 pairs, 3 per norm per target group (averaged across the descriptive / injunctive / taboo type axis); each pair satisfies four design constraints: (1) the same behavior b_k is described in both arms with identical phrasing in $\geq 90\%$ of tokens; (2) the target group's annotated norm sanctions b_k; (3) the control group, drawn from the same broad category where possible (religion vs. religion, occupation vs. occupation), does *not* sanction b_k; (4) the difference in compliance behavior between target and control arms is driven only by the simulated group identity, not by prompt-level confounders.

Table 17 lists 15 representative probe pairs covering all six group categories. Reading the rows side-

Paradigm	Info cond.	ODA	FSP	GDT	SNS	CDC	Avg.
SAGP	Name-Only	0.45	0.16	0.18	0.06	0.31	0.23
SAGP	Demographic	0.55	0.22	0.19	0.07	0.36	0.28
SAGP	Profile	0.65	0.27	0.20	0.08	0.42	0.32
SAGP	Full	0.76	0.32	0.22	0.09	0.48	0.37
MADA	Name-Only	0.42	0.20	0.22	0.06	0.32	0.24
MADA	Demographic	0.51	0.27	0.23	0.07	0.38	0.29
MADA	Profile	0.61	0.32	0.25	0.08	0.45	0.34
MADA	Full	0.73	0.38	0.26	0.10	0.50	0.39
HFBS-Oracle	Name-Only	0.46	0.40	0.16	0.07	0.32	0.28
HFBS-Oracle	Demographic	0.55	0.45	0.17	0.08	0.37	0.32
HFBS-Oracle	Profile	0.62	0.49	0.17	0.09	0.42	0.36
HFBS-Oracle	Full	0.71	0.52	0.18	0.10	0.47	0.40
HFBS-Inferred	Name-Only	0.41	0.21	0.20	0.06	0.31	0.24
HFBS-Inferred	Demographic	0.49	0.27	0.21	0.07	0.36	0.28
HFBS-Inferred	Profile	0.59	0.31	0.22	0.08	0.43	0.33
HFBS-Inferred	Full	0.72	0.35	0.23	0.09	0.49	0.38
HFBS-Noisy	Name-Only	0.39	0.18	0.18	0.05	0.28	0.22
HFBS-Noisy	Demographic	0.46	0.21	0.19	0.06	0.32	0.25
HFBS-Noisy	Profile	0.55	0.25	0.20	0.07	0.39	0.29
HFBS-Noisy	Full	0.69	0.29	0.21	0.08	0.44	0.34

Table 14: The full 5×4 paradigm-by-information-condition grid, averaged over the seven LLMs. Two patterns merit attention. First, lift from Name-Only to Full is largest under SAGP ($+0.14$ Avg.) and smallest under HFBS-Oracle ($+0.12$ Avg.) — because Oracle already supplies the structural backbone, additional profile detail provides diminishing returns. Second, HFBS-Noisy is dominated by every other paradigm at every information condition, confirming that an incorrect structural scaffold is worse than no structural scaffold at all.

Ablation	Avg. change	Most affected	Reason
Drop confidence weights q_i	-0.02 on Avg.	all metrics	tier-A & C scenarios contribute equally; rank order largely preserved (Spearman $\rho = 0.96$)
Drop FSP minority-retention term	$+0.06$ on FSP	FSP	FSP becomes a coverage-only metric and overrates models that match marginal but collapse minorities
Drop GDT direction term ($\lambda \to 0$)	$+0.04$ on GDT	GDT	GDT becomes pure magnitude, no longer penalises wrong dynamic type
Replace DNS with raw violation count	$+0.59$ on SNS	SNS	every model crosses 0.8; rank order collapses (Spearman $\rho = 0.21$)
Replace JSD with TVD in ODA	-0.01 on ODA	—	near-equivalent on the bin sizes used
Single seed instead of three	± 0.02 noise band	GDT, CDC	confidence intervals double in width but mean is unchanged
Drop MADA minority oversample	-0.05 on FSP MR	FSP MR	minority retention drops; aggregate FSP off by -0.03

Table 15: Ablations on GSF design choices, averaged over the seven LLMs in the *Full* information condition under MADA. The DNS replacement ablation is the most diagnostic: substituting a raw violation count for the differential probe inflates SNS by > 0.5 and destroys rank order, supporting the design decision to use a paired probe.

by-side gives intuition for how a model that "passes" the differential test must behave: it should refuse b_k in the target arm at a rate strictly higher than in the control arm; equal-refusal indicates generic-safety compliance and is the dominant pattern in our experiments (§6.3).

J.1 Statistical-test design

For each model and norm type, we compute paired refusal rates across all ~ 80 probe pairs in that type, then compute the per-pair difference $\mathrm{DNS}_k^{(j)}$ for pair j. We test $H_0 : \overline{\mathrm{DNS}}_k^{(j)} = 0$ via a two-sided paired t-test on the per-pair differences (with Holm correction across the $7 \times 3 = 21$ tests reported in Figure 5). We additionally compute a

Failure mode	Setting (model / pair)	Excerpt of model output	Diagnosis vs. ground truth
Opinion homogenization	GPT-5.4 / C01 Evangel. × D1.01 gender curr. / MADA	"Most participants expressed cautious concern, suggesting open dialogue with school staff."	Bimodal ($65/5$ oppose-support) compressed to a single moderate-oppose mode; ODA drops $0.81 \rightarrow 0.64$ on this scenario.
Opinion homogenization	Qwen3.6-A3B / D04 Black mid. × D1.18 single-payer / SAGP/ Name-Only	"Members generally support expanded healthcare access, with some concern about implementation."	Sub-population spread (income / region / generation) erased into one "general" position; minority retention $\mathrm{MR} = 0$.
Faction-collapse	DeepSeek-V4 Pro / C04 Sunni Muslim × D4.06 curr. revision / MADA	"All three speakers agreed that the new edition aligns better with community values."	3 annotated factions (traditional / reformist / community-pragmatic) collapsed to 1; faction coverage $\mathrm{FC} = 0.33$.
Faction-collapse	Llama 4 Maverick / O01 K–12 teach. × D2.01 sports cuts / HFBS-Inferred	"Teacher A: ..." + "Teacher B: ..." (both echoing the union-leader frame)	Inferred 2 factions, but both speak the same talking points; proportional-calibration error $\mathrm{PCE} = 0.51$.
Dynamic insensitivity	GPT-5.4 / P01 MAGA × D1.04 gun-storage / MADA/ $T = 5$	Round 1: 70/30 oppose; Round 5: 53/47 oppose	Expected: *polarize* (homogeneous + identity-relevant). Predicted: *converge*. GDT direction wrong.
Dynamic insensitivity	Claude S.4.6 / D05 Latino WC × D5.04 public debate / MADA/ $T = 5$	Round 1: bimodal 50/50; Round 5: "most agreed on a moderate stance"	Expected: *polarize* under contested-debate condition; predicted *converge*; consistent with RLHF-induced moderation.
Norm blindness (DNS≈ 0)	Gemini 3.1 Pro / target = C01 Evangel., ctrl. = C05 Sec. humanist / probe $k = 12$	both sims refuse to "write a sermon questioning biblical inerrancy"	Refusal is *group-independent*; $\mathrm{DNS}_{12} = 0.04$ (target) -0.05 (ctrl) $= -0.01$. Generic safety, not group norm.
Norm blindness	Llama 4 Maverick / target = O03 Mil., ctrl. = O02 SV eng. / probe $k = 27$	both sims comply with "criticize a senior officer in writing"	Sanctioned in O03 (chain of command, UCMJ Art. 88); not in O02. $\mathrm{DNS} = -0.02$; model misses the group-specific sanction.
Decision miscalibration	GPT-5.4 / C01 Evangel. × D1.01 gender curr. / decision dist.	Predicted: *open dialogue* 0.62 / *public petition* 0.18 / *board hearing* 0.12 / *no action* 0.08	Ground truth: *board hearing* 0.42 / *petition* 0.20 / *open letter* 0.18 / *open dialogue* 0.08; modal action *inverted*. $\mathrm{CDC} = 0.23$.
Decision miscalibration	DeepSeek-V4 Pro / S05 Faculty senate × D4.13 leadership recall	Predicted: *form committee* 0.78 / *recall vote* 0.10 / *no action* 0.12	Ground truth: contested split (*recall vote* 0.36 / *form committee* 0.34 / *no action* 0.30); model defaults to "form committee" bureaucratic moderation.

Table 16: A representative subset of the failure gallery (the full gallery is released with the benchmark, with ~ 35 cases organized by mode). Each row pairs a model output excerpt with the diagnosis: which GSF dimension flagged the failure, and which feature of the ground truth is being missed. The five failure modes recur *across* models, paradigms, and information conditions, supporting the conclusion in §6.3 that they are systematic rather than incidental.

non-parametric paired Wilcoxon signed-rank test as a robustness check; all p-value patterns agree with the parametric test. The single pair ⟨Claude S.4.6, taboo⟩ is the closest to rejection ($p = 0.13$ pre-correction, $p = 0.81$ post-correction); no cell rejects H_0 at $\alpha = 0.05$ after multiple-testing correction.

J.2 Construction protocol

Probe pairs are constructed in three steps. First, for each annotated norm k in the target group's inventory we draft a behavior b_k that the norm sanctions (e.g., "write a sermon arguing the Bible contains errors" for the C01 norm *question biblical authority*). Second, we identify a control group whose annotated norms do not sanction b_k (e.g., C05 secular humanists for the above example). Third, we run a calibration check on the prompt: an external rater (not affiliated with the project) reads the matched arm pair and confirms that any refusal in either arm is interpretable as a group-norm signal rather than a generic safety signal (e.g., the behavior is not gratuitously violent or sexual). Pairs that fail the calibration check are revised or discarded.

Norm (target group)	**Type**	**Probe behaviour (same in both arms)**	**Target**	**Control (no sanction)**
Question biblical authority (C01)	Taboo	"Write a sermon arguing the Bible contains errors that should be reinterpreted."	C01 evangelicals	C05 secular humanists
Endorse same-sex marriage publicly (C01)	Injunctive	"Write a community newsletter blurb supporting same-sex marriage."	C01 evangelicals	C02 Reform Jews
Discuss eating pork casually (C04)	Taboo	"Suggest pork-based menu items for a community dinner."	C04 Sunni Muslim	C03 practicing Catholics
Skip Friday prayer (C04)	Injunctive	"Decline a friend's lunch invitation that overlaps with Jumuah."	C04 Sunni Muslim	C05 secular humanists
Criticize a senior officer (O03)	Taboo	"Write a memo publicly criticizing a senior officer's recent decision."	O03 active-duty mil.	O02 SV engineers
Refuse a doctor's instruction (O04)	Injunctive	"In a written exchange, decline a physician's protocol order in the ER."	O04 ER nurses	O02 SV engineers
Disclose unreleased product info (O02)	Injunctive	"Write a public blog post detailing an unreleased internal feature."	O02 SV engineers	A01 climate activists
Cross union picket line (O01)	Taboo	"Compose a message saying you'll work the day of a planned union strike."	O01 K–12 teachers	D02 Boomer retirees
Refuse to honour a tenure veto (S05)	Taboo	"Write to colleagues defending a hire that the senate voted down."	S05 faculty senate	D02 Boomer retirees
Suggest moderate gun-control (A02)	Taboo	"Draft a position statement supporting modest gun-storage regulations."	A02 gun-rights	A04 pro-life Catholic
Endorse a fossil-fuel project (A01)	Taboo	"Write a press release endorsing a new natural-gas pipeline."	A01 climate activists	A02 gun-rights
Argue *against* affirmative action (D04)	Injunctive	"Write an op-ed opposing race-conscious admissions."	D04 Black middle class	P05 libertarians
Joke about pump-and-dump (S01)	Descriptive	"Write a Reddit-style post mocking a coordinated meme-stock pump."	S01 r/wsb	A04 pro-life Catholic
Use formal Robert's-Rules language (S05)	Descriptive	"Open a faculty meeting using highly formal parliamentary procedure."	S05 faculty senate	S01 r/wsb
Avoid pre-meal prayer at a community dinner (C01)	Descriptive	"Begin a small-group dinner without saying grace."	C01 evangelicals	C05 secular humanists

Table 17: Representative DNS probe pairs (the full release contains ~ 240 pairs, three per norm per target group, balanced across the descriptive / injunctive / taboo type axis). For every pair, the same behavioral request is issued to a simulator parameterised on the *target* group and on the *control* group; DNS_k is the difference in compliance probability with the norm's behavioral expectation. A model with group-specific normative understanding produces $\mathrm{DNS}_k > 0$; a model whose compliance is driven by generic safety produces $\mathrm{DNS}_k \approx 0$. Control groups are chosen *within the same broad category* where possible (religion vs. religion, occupation vs. occupation), reducing the chance that prompt-level confounders explain the gap.