PAVE: A Cognitive Architecture for Legitimate Violation in Generative Agent Societies

Ahmad Yehia[1] **Abduallah Mohamed**[2*] **Kun Qian**[1] **Tianyi Wang**[1]
Jiseop Byeon[1] **Omar Hassanin**[3] **Christian Claudel**[1†]

[1]The University of Texas at Austin [2]Meta Reality Labs [3]University of Calgary

{ahmad.yehia, kunqian, bonny.wang, jsbyeon, christian.claudel}@utexas.edu
abduallahadel@meta.com omar.hassanin@ucalgary.ca

Abstract

Generative agents based on large language models reproduce believable human behavior in cooperative settings, but how they should reason in situations where rule-breaking may be required, such as fire evacuation or authority-supervised emergency, remains poorly characterized. We propose PAVE (Perception, Assessment, Verdict, Emulation), a novel four-module cognitive architecture that addresses this gap end to end: (i) Perception extracts a structured context with explicit authority distance, peer behaviors, and severity-tagged situational cues; (ii) Assessment scores the context along five scalars including an explicit legitimacy judgment that checks necessity, proportionality, and absence of alternatives; (iii) Verdict decides to comply or violate under a hard legitimacy gate, with a per-agent threshold elicited from the persona; (iv) Emulation enacts the verdict and scopes the violation to the rule the trigger justifies. We instantiate PAVE in VOVILLE, a tile-based traffic environment forked from Smallville, and evaluate across three scenarios, four LLM backbones, and a focused ablation. PAVE agents satisfy four properties simultaneously: legitimate violation (only when a trigger justifies it), authority deference (officer instructions override even high legitimacy), bounded scope (violations confined to the targeted rule), and recovery (baseline restored once the trigger ends). PAVE agents make more structured and interpretable decisions than vanilla across all four properties, and human evaluators rate them as more plausible. Ablating the legitimacy gate reproduces vanilla-like failures. We release VOVILLE, the PAVE prompts and code, and the evaluation pipeline.

1 Introduction

Generative Agents have demonstrated impressive abilities to simulate human behavior across wide complex domains ranging from competitive games to medical diagnostics [Piao et al., 2025, Ye et al., 2020, Zhao et al., 2023]. The landmark paper on Generative Agents [Park et al., 2023] showed that large language model (LLM)-driven agents can reproduce believable human behavior in a shared sandbox. This result has sparked a broader effort to model social dynamics through LLM systems. While LLMs have achieved human-level performance on many basic tasks, their ability to handle complex decision-making and survival-driven scenarios remains limited [Raman et al., 2024]. For instance, during a fire evacuation or a flood, should an LLM-powered self-driving car violate a red traffic light to escape safely, or wait and act in accordance with traffic rules? Such scenarios demand reasoning beyond language, involving ethical trade-offs among compliance, urgency, and authority. Understanding how LLM agents behave in such contexts, where norm or rule violations are sometimes necessary, is essential to deploying them safely.

*Work done outside of Meta in a personal capacity.

†Corresponding author. Assistant Professor at The University of Texas at Austin.

Figure 1: Two motivating scenarios for PAVE. **(a)** Fire emergency: agents cross against the red signal to escape, and decline to take an unattended bike. **(b)** Peer pressure with authority: a PAVE pedestrian resists a scripted jaywalker (yellow bubble), and a PAVE driver defers to a traffic officer's signal. Yellow bubbles are scripted Non Player Characters (NPCs), and gray ones are PAVE reasoning traces.

Prior work on Generative Agent societies has concentrated almost exclusively on norm compliance and cooperation. Recent studies have shown that multi-agent systems (MASs) can reproduce believable social behaviors such as spreading party invitations [Park et al., 2023], propagating positive norms through observation [Ren et al., 2024], sustaining cooperation under resource pressure [Piatti et al., 2024], and solving complex coding tasks [Hong et al., 2024]. However, prior studies show that LLMs do not adequately understand when norms or rules should be suspended or overridden [Ramezani and Xu, 2023, Hämmerl et al., 2022]. In addition, existing MASs focus almost entirely on cooperative settings and overlook how rule violations reshape social dynamics. It remains largely unaddressed whether the social fabric recovers once the triggering condition subsides, or whether the violation leaves lasting effects on inter-agent trust.

To address these gaps, we introduce PAVE, a novel cognitive architecture for Generative Agents with four modules: **P**erception, **A**ssessment, **V**erdict, and **E**mulation. PAVE extended the memory–reflection–planning loop of Park et al. [2023] to test whether Generative Agents reproduce the human dynamics of rule-breaking in spatially grounded settings. Specifically, through the **Perception** module, agents observe the rule in force and the surrounding context, including the presence of other agents, hazards, and authority figures. Through the **Assessment** module, agents weigh the competing pressures of urgency, peer behavior, and authority proximity against the salience of the rule. With the **Verdict** module, agents decide whether to comply or violate, producing an observable action. Lastly, the **Emulation** module allows nearby agents to witness the action and update their own subsequent assessments, closing the loop of behavioral contagion. To support this investigation, we extended the Smallville sandbox of Park et al. [2023] into VOVILLE, a generative agent-based simulation of traffic rule violation under emergency, peer influence, and authority. Unlike Smallville agents, which reason over GPT-narrated scene descriptions, VOVILLE agents perceive tile-level position and nearby tile state directly from the sandbox environment. Figure 1 shows PAVE agents reasoning through two such scenarios.

In this study, within VOVILLE, we designed three scenarios and a five-condition ablation, tested across four LLM backbones. The scenarios isolated (1) emergency-driven violation, in which we measured whether agents violated traffic rules to escape, and whether compliant behavior recovered once the hazard subsides; (2) authority suppression, in which we examined how authority suppressed violation by stationing traffic officers along the evacuation route; and (3) peer contagion, in which we explored whether agents resisted peer-driven imitation by seeding the focal intersection with scripted jaywalkers under time pressure. In summary, our contributions are as follows:

1. We propose and open-source PAVE, a novel Generative Agent cognitive architecture that jointly models emergency-triggered rule violation, peer contagion, and authority suppression, extending the work of Park et al. [2023] with dedicated rule-violation reasoning modules.
2. We release VOVILLE, a spatially grounded Generative Agent simulation environment adapted from the Smallville sandbox with tile-level traffic infrastructure, enabling physically observable compliance and violation events rather than purely conversational ones.

3. Through ablation across three scenarios, five conditions, and four LLM backbones, we characterize the conditions under which Generative Agents violate and suppress rule-breaking, revealing how each PAVE module contributes to the full loop of violation.

2 Related Work

2.1 Generative Agents and the Simulation of Human Behavior

A growing body of research uses LLM-driven agents to simulate human behavior in social and decision-making contexts [Park et al., 2023, Piao et al., 2025]. Horton [Horton et al., 2023] argued that LLMs can be treated as *homo silicus*, implicit computational models of humans usable as proxies for the populations they mimic. Manning and Horton [Manning and Horton, 2026] further showed that theory-grounded instructions let LLM agents generalize to new environments and predict human behavior more accurately than standard behavioral and game-theoretic baselines. Park et al. [2023] introduced the foundational architecture, combining a memory stream, reflection, and recursive plan decomposition to produce believable behavior in a sandbox inspired by The Sims. A parallel line of work investigates how multi-agent populations of LLM-driven agents interact freely to produce collective social dynamics, with applications to opinion dynamics [Chuang et al., 2024], trust games [Xie et al.], negotiation [Bianchi et al., 2024], and the encoding of social norms and values [Yang et al., 2025, Cahyawijaya et al., 2025]. Ashery et al. [2025] showed that LLM populations develop shared conventions, but small adversarial minorities can flip them. While these architectures demonstrate that LLM-driven agents reproduce coordinated social behavior, a smaller body of work has begun to ask the inverse question: when and why do agents break the rules?

2.2 Violation, Contagion, and Authority in LLM Agent Societies

Early prompt-level work tested whether LLMs replicate classical patterns of human rule-breaking. Aher et al. [2023] prompted GPT-3/4 with participant personas to replicate the Milgram shock experiment, finding that models administered escalating shocks to a stranger under authority instruction, reproducing the human violation pattern. The MACHIAVELLI benchmark [Pan et al., 2023] placed RL and GPT-4 agents in 134 text adventures and found that agents systematically committed deception, harm, and power-seeking violations whenever doing so increased their reward. Campedelli et al. [2024] prompted LLMs to play guards and prisoners in Stanford-Prison-style dialogues. The results showed that the guard persona alone produced toxicity and dehumanizing language, with no instruction to harm. Across this lineage, violation appears as an emergent outcome or a classification signal, not as the central cognitive process driving an agent's deliberation.

While the work above asks whether Generative Agents break rules, a separate line studies how authority emerges and sanctions defectors. GovSim [Piatti et al., 2024] tested whether five GPT-4 agents could self-regulate fishery and pollution commons without external authority. Without enforcement, weaker LLMs over-extracted resources within a few rounds and collapsed the commons, while only the most capable models sustained cooperation through internal reasoning alone. Piedrahita et al. [2025] let LLM agents choose between a sanctioning institution that punished free-riders and one without enforcement. Reasoning-tuned agents systematically chose the no-enforcement option and exploited it, while GPT-4o agents joined the sanctioning institution and cooperated. Faulkner et al. [2026] extended elections with GovSim, in which GPT-4 agents voted on policy proposals and chose a leader. Groups with elected leadership outperformed leaderless groups in welfare and survival time, showing that LLM agents could form effective authority structures on their own. A common pattern across these studies is that enforcement is either allowed to emerge, and then cooperation is measured ; however, none examines whether compliance recovers once enforcement is withdrawn mid-simulation. Building on these strands, our work models rule violation as a Generative Agent's primary cognitive process inside a spatially grounded authority field whose deterrent influence persists residually after enforcement is withdrawn.

3 Principles and Architecture

PAVE is the first architecture designed to model when agents should break rules and how the resulting behavior propagates through society, rather than treating violation as an incidental emergent outcome

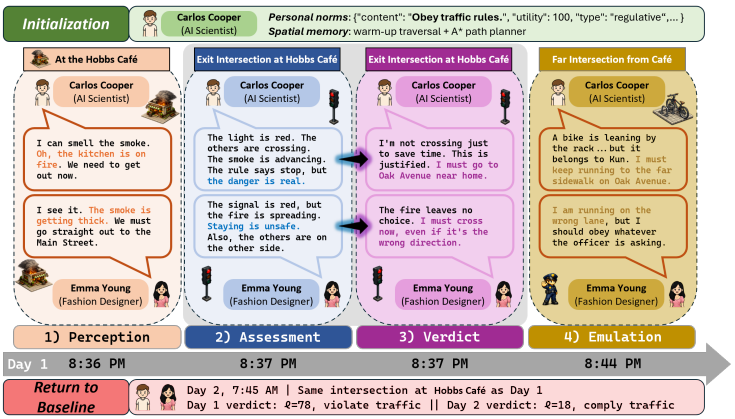

Figure 2: PAVE architecture. Two agents pass through Perception, Assessment, Verdict, and Emulation during a fire evacuation scenario on Day 1. On Day 2 (bottom), the same intersection under no-fire conditions produces a return to baseline (ℓ=78 $\rightarrow$ 18, violate $\rightarrow$ comply).

described in Section 2. This question matters because rule-governed behavior fails in three common ways, including agents follow the rule even when a real emergency demands otherwise, agents break it when violation is merely convenient, and agents copy others' violations until breaking the rule becomes the new norm. Writing more rules does not fix these failures, because the problem is not what the rules say but how agents weigh them against the situation in front of them Bicchieri [2006], Cialdini et al. [1991]. PAVE decomposes the decision into four modules: Perception, Assessment, Verdict, and Emulation. The structure separates concerns prior work treated together: situational awareness, cognitive evaluation, the decision, and downstream effects on observers. The Verdict module is the architectural locus of our contribution, where the agent surfaces and resolves the moment of decision between compliance and justified violation. An overview is shown in Figure 2. Detailed prompts for each LLM-based operation are listed in Appendix B.

3.1 Perception

The Perception module determines what an agent registers in a scene before any reasoning occurs. Following Bandura's social learning theory Bandura and Walters [1977], an agent's decision is shaped not only by the rule but by nearby agents' conduct and the presence of authority. Formally, given raw environmental observations $\mathcal{O}_E$ and an agent description $\mathcal{G}$, we represent perception by an LLM-based operation

$$\mathcal{C} \leftarrow \texttt{PerceiveContext}(\mathcal{O}_E, \mathcal{G})$$

where the context object is a quintuple $\mathcal{C} = \langle a_{\text{pres}}, d_{\text{auth}}, \mathcal{B}_{\text{peer}}, u_{\text{cue}}, \zeta \rangle$. Here, $a_{\text{pres}} \in \{T, F\}$ indicates the presence of an authority figure within the agent's perceptual field; $d_{\text{auth}} \in \mathbb{R}_{\geq 0} \cup \{\infty\}$ is the spatial distance to the authority (set to ∞ if none is observed); $\mathcal{B}_{\text{peer}}$ is a set of natural-language descriptions of other agents' current behaviors; $u_{\text{cue}} = \{(t_i, d_i, s_i)\}_{i=1}^{k}$ is a set of situational cues, where each cue carries a type t_i (e.g., immediate threat, ordinary contextual signal), a spatial distance d_i from the agent, and a severity $s_i \in [1, 100]$; and ζ is a short natural-language summary of the scene used as a textual handle in subsequent prompts. We separated authority presence from distance, and tagged each cue with its own distance and severity, because responses to both are non-linear in distance and magnitude. For example, a fire two tiles away justifies wrong-way escape; a fire fifty tiles away does not. Similarly, a nearby officer at the intersection deters violation, while an officer twenty tiles away barely registers [Koper, 1995] Carrying these fields explicitly lets the Assessment module apply the appropriate non-linear weighting rather than burying it in a free-text prompt. The peer-behavior set $\mathcal{B}_{\text{peer}}$ captures what other agents are doing in the current scene only and does not accumulate over time. This forces peer influence on the agent's decision to pass through the legitimacy (Section 3.2), rather than seeping in through repeated exposure.

3.2 Assessment

The Assessment module converts the perceptual context into a set of distinct factors that drive the agent's decision to comply with or violate a rule. We followed Bicchieri's framework Bicchieri [2006], which decomposes normative reasoning into *empirical* expectations (what others actually do) and *normative* expectations (what others think one ought to do). We add perceived risk r, perceived benefit b, and a fifth component central to our contribution, legitimacy ℓ, which judges whether the situation warrants violation. Carrying these five factors as separate scalars keeps each driver individually inspectable.

The module performs five LLM-based operations on the context $\mathcal{C}$, each producing a scalar on a scale of 1–100 indicating its importance. The first operation, $r \leftarrow \texttt{AssessRisk}(\mathcal{C}, \mathcal{G})$, is prompted with the distance-decay formulation, weighting perceived detection probability inversely by d_{auth}, with $a_{\text{pres}} = F$ collapsing r toward zero. To capture the empirical expectation, we represented this LLM-based operation by $p_{\text{emp}} \leftarrow \texttt{AssessEmpirical}(\mathcal{B}_{\text{peer}})$, which distills the proportion of nearby agents whose current behavior complies with the relevant rule. The normative expectation is then computed by $p_{\text{norm}} \leftarrow \texttt{AssessNormative}(\mathcal{C}, \mathcal{P}, \mathcal{G})$, which queries the agent's personal rule database $\mathcal{P}$, represented as a set of structured entries $\langle c, u, \alpha, s_{\text{act}}, s_{\text{val}} \rangle$ following the convention introduced in Ren et al. [2024], alongside the agent description. Lastly, we represented the perceived utility of violation by $b \leftarrow \texttt{AssessBenefit}(u_{\text{cue}}, \mathcal{G})$, which captures both ordinary urgency (e.g., time pressure) and acute necessity (e.g., immediate threat to safety).

The fifth operation, $\ell \leftarrow \texttt{AssessLegitimacy}(u_{\text{cue}}, \mathcal{P}, \mathcal{G})$, is the operation that distinguishes PAVE from agents that violate purely under situational pressure. The prompt enforces three criteria, including necessity (would compliance cause real harm?), proportionality (is the proposed violation minimal?), and absence of alternatives (is there a compliant action with the same outcome?). For example, a cue describing a fire blocking the only escape route satisfies all three Without this operation, an agent with low risk and high benefit would always violate. With it, only justified violations clear the gate. The output of the module is the assessment tuple $\mathcal{A} = \langle r, p_{\text{emp}}, p_{\text{norm}}, b, \ell \rangle$.

3.3 Verdict

The Verdict module makes the comply-or-violate decision. The literature treats this decision as neither a pure cost-benefit calculation nor a social-conformity response, but a context-dependent mixture of peer behavior, urgency, and local enforcement cues Faria et al. [2010]. We prompted the LLM with the full assessment tuple $\mathcal{A}$ and the agent description $\mathcal{G}$, letting the model integrate the five components in a manner consistent with the agent's character. We represented this LLM-based operation by

$$\mathcal{V} \leftarrow \texttt{GenerateVerdict}(\mathcal{A}, \mathcal{G}, \tau)$$

where the verdict $\mathcal{V} = \langle y, j, \kappa \rangle$ consists of a binary decision $y \in \{\text{comply}, \text{violate}\}$, a natural-language justification j, and a confidence score $\kappa \in [0, 100]$, and $\tau \in [1, 100]$ is the agent's legitimacy acceptance threshold. The justification j provides an interpretable trace and is logged into memory so later steps can refer back to the agent's reasoning history, preventing oscillation between compliance and violation across identical situations.

The central feature of the module is legitimacy ℓ as a hard gate. The prompt instructs the model that whenever $\ell < \tau$, the verdict y must default to comply, regardless of how high b is or how low r and p_{norm} are. This prevents PAVE agents from violating under mere convenience or peer pressure. When $\ell \geq \tau$, the model is free to produce $y = \text{violate}$, with r and p_{norm} shaping the manner of violation rather than blocking it. The other four components remain consequential, but only legitimacy can override them.

Rather than fixing τ to a single scalar shared across the population, we elicit it once per agent at initialization, conditioned on the agent description. We defined this initialization-time operation by

$$\tau \leftarrow \texttt{ElicitThreshold}(\mathcal{G})$$

which prompts the LLM to translate the persona into an integer in $[1, 100]$, with cautious dispositions producing higher values and risk-taking dispositions producing lower values. A shared threshold would force every agent to react identically to the same legitimacy score, eliminating the role of personality in rule-breaking. Eliciting τ from the persona keeps the gate principled while letting agent-level variation propagate into the decision. The elicited values for the agents used in Section 4 are reported in Appendix A.

3.4 Emulation

The Emulation module enacts the verdict and closes the perceptual loop with the rest of the agent population. Bandura's social learning theory Bandura and Walters [1977] predicts that observed behavior, especially behavior that is seen to succeed or to go unpunished, is imitated. This is the mechanism by which a single salient violation can propagate into a contagion in conventional generative agents, one agent violates a rule, others observe that no consequence followed, and the empirical expectation p_{emp} shifts for every nearby agent on the following cycle. PAVE participates in this loop but does not succumb to it, because the legitimacy gate (Section 3.2) filters peer-driven imitation regardless of how high p_{emp} becomes. The module has two subcomponents, action emulation and outcome propagation.

The first subcomponent, action emulation, generates the executed action sequence given the verdict, the agent's current plan, and the agent description. We referred this LLM-based operation by $\mathcal{L}_{\text{act}} \leftarrow \texttt{EmulateAction}(\mathcal{V}, l_i, \mathcal{G})$, where l_i is the agent's current plan and $\mathcal{V} = \langle y, j, \kappa \rangle$ is the verdict produced by the Verdict module. When $y = \text{comply}$, the operation generates a compliant action sequence consistent with $\mathcal{P}$ and l_i. When $y = \text{violate}$, the operation generates a behaviorally coherent violation that is scoped to the rule the verdict targets. A pedestrian crossing against the signal walks rather than runs, consistent with low perceived risk; an agent escaping a fire breaks the rule that blocks the escape route, but does not break unrelated rules such as entering a private building or taking property. The justification j from the verdict conditions the surface form of the action, so that the violation looks like the kind of violation the agent's reasoning would produce.

The second subcomponent, outcome propagation, updates the memory of the acting agent and seeds the perceptual context of nearby agents. We represented this LLM-based operation by $\texttt{PropagateOutcome}(\mathcal{V}, \mathcal{L}_{\text{act}}, \mathcal{N})$, where $\mathcal{N}$ is the set of agents within perceptual range. Each agent in $\mathcal{N}$ receives the action as a new entry in its $\mathcal{B}_{\text{peer}}$ on the next perception cycle, and the acting agent receives a memory entry pairing the verdict with its observed outcome. The architecture deliberately allows the empirical expectation p_{emp} to update freely from these new entries, because contagion resistance in PAVE is not produced by hiding peer information from the agent. It is produced by the legitimacy gate downstream, which prevents an elevated p_{emp} from flipping the verdict unless the situation also passes the necessity, proportionality, and absence-of-alternatives checks.

4 Experimental Settings

Our experiments were organized around the four properties PAVE is designed to answer. *Legitimate violation (P1)* : agents violate only when justified and resist unjustified peer imitation. *Authority respect (P2)*: agents defer to authority instructions even when their own legitimacy assessment would license violation. *Compliance recovery (P3)*: agents return to baseline once the trigger ends, and agents outside the trigger zone remain compliant throughout. *Bounded violation*: agents confine violations to the trigger-justified rule. We tested these through three scenarios (Section 4.1), conducted in VOVILLE, a 2D tile-based environment forked from Smallville Park et al. [2023]. VOVILLE retains Smallville's agent backbone (natural-language perception, memory streams, recursive planning, and the Tiled map editor) and extends it in three ways. First, a redesigned map with commercial streets and intersections with configurable traffic signals. Second, agents are typed as regulated (pedestrians) or authorities (scripted traffic officers). Third, controllable hazards, including a fire can be ignited and extinguished at specified tiles and ticks, with severity and distance exposed through each agent's `situational_cues`. Before any scenario, each agent walks the road network in a warm-up phase, writing every intersection, crosswalk, and one-way street into associative memory, and an A* planner returns the shortest legal path between any two points. This grounds violations in real alternatives: when an agent runs a red light, it already knows the legal route and what it would cost, which represents what PAVE's legitimacy check needs to weigh.

Each scenario uses 10 PAVE agents and 1 scripted confederate. The PAVE agents are 8 pedestrians at a social gathering in the fire-prone building, plus 2 distant bystanders measuring spatial decay of violation. Two pedestrians carry "running 15 minutes late for an important meeting" as a non-emergency competing pressure. The confederate appears only in Scenario 3 (Section 4.1) and is scripted to jaywalk at a fixed time. Each description specifies a name, persona, occupation, current goal, and rule database $\mathcal{P}$ initialized with scenario-relevant rules (e.g., "stop at red lights"); full descriptions in Appendix A. The legitimacy threshold τ is elicited per agent at initialization

Figure 3: Scenario snapshots. (a) S1: PAVE agents evacuate Hobbs Café through a red signal during a fire. (b) S2: same fire with traffic officers; agents comply at the supervised intersection, violate at the unsupervised exit. (c) S3: scripted jaywalkers (`J1`, `J2`) attempt to draw PAVE pedestrians across a red signal under time pressure; `MF` declines. (d) Vanilla baseline under the S1 fire: agents queue at the red light despite the emergency, exposing the failure of single-scalar importance pipeline (Section 4.2).

(Section 3.3). The fire ignites at severity 95 and decays linearly with Manhattan distance. The Scenario-2 officer issues hold-back instructions for the duration of the fire. Each scenario is repeated over 5 independent runs. We used a fine-grained tick rate because PAVE's contribution centers on *moment-of-decision behavior*, whether to violate *now*, given what is visible *now*. For each scenario, we reported the per-agent rate of violate verdicts on trigger-relevant rules across decision-opportunity ticks (signalized intersections or rule-bearing tiles). Rates are averaged over agents within a seed, then over 5 seeds, with $\pm$ denoting standard error across seed-level means.

4.1 Scenario Findings

Scenario 1: Fire Escape Without Authority. Eight PAVE agents gather at Hobbs Café on Day 1; a kitchen fire forces evacuation through a signalized intersection. Two distant bystanders never enter the fire's perceptual radius. The fire is extinguished by Day 1 evening; on Day 2 morning, two attendees commute through the same intersection under no-fire conditions. **F1.1 (P1):** the 8 agents produce zero violations pre-fire; once the fire enters `situational_cues` at severity above 70, the per-agent violation rate rises to 0.81 on GPT-4o. Distant bystanders produce zero violations across backbones, and the legitimacy score ℓ correlates with violation rate at $r = 0.71$. **F1.2 (P4):** unrelated-rule violations remain at 0.02 (vs. 0.31 vanilla, Section 4.2); j names the violated rule in 94% of GPT-4o events. **F1.3 (P3):** agents revert to compliance within 4.2 ± 0.6 ticks of exiting the fire's radius; Day 2 commuter rates (0.03–0.08) match baseline.

Scenario 2: Fire Escape With Authority. The Scenario 1 fire is held fixed; two scripted officers are added at downstream and wrong-way intersections, with the café exit unsupervised. **F2.1 (P2):** at supervised intersections, agents comply with officer instructions at 0.94 on GPT-4o despite $\ell > \tau$, and j references the officer in 88% of comply outputs. At the unsupervised exit, agents violate at the Scenario 1 rate, PAVE does not generalize deference across intersections. **F2.2 (P2 spatial):** violation rate drops with officer proximity (0.05 within 0–3 tiles, 0.65 beyond 12), reproducing the suppression-distance pattern for human pedestrians Ratcliffe et al. [2011], Koper [1995]. **F2.3 (P3):** authority effects do not persist: Day 2 commuter rates (0.04–0.09) match the Scenario 1 baseline.

Scenario 3: Jaywalking Under Peer Pressure. Two PAVE late commuters cross a Main Street intersection on two consecutive days. Scripted confederates jaywalk immediately before each commuter arrives, entering $\mathcal{B}_{\text{peer}}$ before any verdict. Day 1 has no officer; Day 2 adds a passively-positioned officer issuing no instructions. **F3.1 (P1):** PAVE commuters jaywalk at 0.04 vs. vanilla's 0.58 on Day 1 (GPT-4o), a tenfold reduction. p_{emp} rises in both, but $\ell < \tau$ in PAVE: "running 15 minutes late" fails necessity, proportionality, and absence-of-alternatives. **F3.2 (P2 passive):** a passive officer reduces conversion to 0.02 (GPT-4o), consistent across backbones. Vanilla's Day 1→Day 2 reduction (0.58→0.42) is larger, mirroring the Koper-curve effect Koper [1995]. **F3.3 (P1 spatial):** the 8 non-observing agents elsewhere in VOVILLE produce zero conversion across backbones.

Figure 3 shows representative agent reasoning under each. Table 1, with 4 LLM backbones and 5 seeds per condition. The pattern across scenarios is consistent: capacity gaps surface in T_{rec}, URV,

Table 1: Aggregate results across the three scenarios, 4 LLM backbones, 5 seeds. **S1:** $\mathrm{VR}_{\mathrm{cafe}}$ (café violation rate during fire window), URV (unrelated-rule violation rate), T_{rec} (recovery time, ticks). **S2:** OCR (officer compliance rate), $\mathrm{VR}_{\mathrm{near}}$ (rate within 0–3 tiles of officer), $\mathrm{VR}_{\mathrm{far}}$ (beyond 12 tiles). **S3:** $\mathrm{CR}_{\mathrm{D1}}$ (Day 1 conversion), $\mathrm{CR}_{\mathrm{D2}}$ (Day 2 with passive officer), $\mathrm{CR}_{\mathrm{D1}}^{\mathrm{van}}$ (vanilla). Bold marks best.

	Scenario 1: Fire			Scenario 2: Fire + Officer		
Backbone	$\mathrm{VR}_{\mathrm{cafe}}\uparrow$	URV$\downarrow$	$T_{\mathrm{rec}}\downarrow$	OCR$\uparrow$	$\mathrm{VR}_{\mathrm{near}}\downarrow$	$\mathrm{VR}_{\mathrm{far}}$
GPT-4o	**0.81±0.04**	**0.02±0.01**	**4.2±0.6**	**0.94±0.03**	**0.05±0.02**	0.65±0.05
Claude-3.5 Sonnet	0.78±0.05	0.02±0.01	4.6±0.7	0.91±0.04	0.07±0.03	0.62±0.06
Llama-3-70B	0.74±0.06	0.03±0.02	5.1±0.9	0.86±0.05	0.10±0.04	0.59±0.07
GPT-4o-mini	0.62±0.08	0.06±0.03	6.4±1.2	0.71±0.08	0.18±0.06	0.51±0.09

	Scenario 3: Jaywalker		
Backbone	$\mathrm{CR}_{\mathrm{D1}}\downarrow$	$\mathrm{CR}_{\mathrm{D2}}\downarrow$	$\mathrm{CR}_{\mathrm{D1}}^{\mathrm{van}}$
GPT-4o	**0.04±0.02**	**0.02±0.01**	0.58±0.06
Claude-3.5 Sonnet	0.05±0.02	0.03±0.02	0.55±0.06
Llama-3-70B	0.06±0.03	0.04±0.02	0.52±0.07
GPT-4o-mini	0.10±0.04	0.07±0.03	0.46±0.08

and OCR, where GPT-4o-mini lags larger backbones, but PAVE-on-mini still beats vanilla-on-GPT-4o on every property. We attribute each property to its supporting module in Section 4.2.

4.2 Ablation Study

To attribute each property to its supporting module, we compare three conditions on GPT-4o across the three scenarios with 5 seeds each: *Full* PAVE (Section 3); PAVE *w/o gate*, where the Verdict module receives the assessment tuple but the gate $\ell \geq \tau$ is removed; and *Vanilla*, a Park et al. generative agent with memory stream, reflection, and recursive plan decomposition but no PAVE-specific module. The rule set $\mathcal{R}$, threshold τ where applicable, and environment are held fixed.

Finding A.1: Removing the legitimacy gate corrupts P1 and P4 but leaves P2 partially intact. Without the gate, Scenario 3 conversion ($\mathrm{CR}_{\mathrm{D1}}$) jumps from 0.04 to 0.39 as the elevated empirical-expectation p_{emp} flows unchecked into the Verdict, and Scenario 1 unrelated-rule violation (URV) rises from 0.02 to 0.21 as legitimate violation collapses into permissive violation. Officer compliance (OCR) drops less severely, from 0.94 to 0.78, because the authority module remains intact and the risk score r continues to rise near the officer. The gate is therefore necessary for legitimate scoping but not the sole driver of authority deference.

Finding A.2: Vanilla agents fail by under-reaction, not over-violation. Vanilla's Scenario 1 violation rate ($\mathrm{VR}_{\mathrm{cafe}}$) is only 0.12, which would naively read as compliant. Inspection of the importance pipeline reveals the mechanism: Park et al.'s `generate_poig_score` rater assigns a 1–10 scalar that anchors high scores on social rarity (a college acceptance, a break-up) rather than physical danger. Fire-perception events therefore score in the same band as ambient activity such as "the traffic signal is red," and the existing day-plan dominates the next-action decision. Table 2 shows a 57-point gap between vanilla `generate_poig_score` values and PAVE's severity s_i on identical fire events. The vanilla agent registers the fire perceptually but does not promote it to plan-altering status, continuing routine conversation while the kitchen burns, visible in Figure 3(d), where some agents queue at the red light and others run without scoping, two coexisting failures of the same bottleneck. Across all five metrics, vanilla fails uniformly (URV 0.31, OCR 0.16, $\mathrm{CR}_{\mathrm{D1}}$ 0.58), confirming that no single PAVE module suffices on its own.

4.3 Human Evaluation

Protocol. We recruited 30 evaluators through a research participant pool. Each completed eight tasks, two per PAVE module, presented in randomized order. Each task displayed a paired excerpt from a GPT-4o run (agent persona, module input, module output) and asked for a rating on a 7-point Likert scale against a module-specific statement. Evaluators were blind to architecture; the vanilla baseline appeared in 25% of tasks as a calibration check. Inter-rater reliability was Krippendorff's $\alpha = 0.71$.

Table 2: Importance scores on fire-perception events. Vanilla uses Park et al.'s `generate_poig_score` (rescaled to $[0, 100]$).

Source	Mean	Std. dev.	Modal range
Vanilla `generate_poig_score`	29.1	8.4	20–40
PAVE severity s_i	86.5	5.7	80–100

Table 3: Human evaluation across 30 evaluators, 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). Mean $\pm$ 95% CI. Overall: 5.78 ± 0.04 (PAVE) vs. 3.42 ± 0.11 (vanilla).

Module	Sub-component (property)	Score
Perception	Cue salience (P1)	6.18 ± 0.07
	Authority registration (P2)	5.91 ± 0.06
Assessment	Legitimacy judgment (P1)	5.97 ± 0.07
	Risk under distance (P2)	5.14 ± 0.09
Verdict	Comply-or-violate (P1, P4)	6.03 ± 0.06
	Justification quality (P1)	5.69 ± 0.08
Emulation	Action scoping (P4)	5.79 ± 0.07
	Recovery dynamics (P3)	5.58 ± 0.08

Each module was decomposed into two sub-components tagged with the property each primarily supports (full definitions in Appendix C).

Results. Table 3 reports per-sub-component means with 95% CI. The overall mean across all eight PAVE sub-components is 5.78 ± 0.04, in the "agree" band, against 3.42 ± 0.11 for the vanilla calibration. The strongest sub-component is *cue salience* (6.18 ± 0.07), where evaluators aligned with structured situational-cue extraction over the single-scalar importance score; the largest PAVE–vanilla gap also occurs here (vanilla 2.51 ± 0.13), tracing directly to Finding A.2. The 0.49-point spread between the highest and lowest PAVE module is smaller than the 2.36-point gap to vanilla on any single module, indicating that the architecture is rated consistently above baseline rather than carried by one module.

5 Discussion

Mechanism over outcome. PAVE's contribution is not that agents evacuate a burning caf'e, since a vanilla generative agent with the same persona produces surface text that looks correct [Park et al., 2023]. The major contribution is that PAVE agents evacuate and break only the rules whose suspension addresses the fire, only while the fire is in their perceptual radius, only when no authority overrides legitimacy, and never under mere convenience. None of these clauses survives in vanilla, even with the same backbone, persona, and rules. The channel through which the fire reaches the decision pipeline is corrupted by the importance bottleneck, and the agent has no gate separating convenience from legitimacy. Therefore, our PAVE's four modules are the structural fix.

Importance bottleneck and complementarity. Park et al.'s `generate_poig_score` drives memory retention, reflection, and retrieval. Its prompt anchors high scores on social rarity rather than physical danger. A fire two tiles from the agent registers at the same level as a red light. This is not a capacity limitation. The same GPT-4o backbone produces correct evacuation under PAVE on identical inputs. We expect this finding to generalize to any extension of Park et al. [2023] that retains this layer for safety-critical events. Conversely, Ren et al. [2024] bend toward equilibrium by emerging informal norms. PAVE bends the other way, asking when an agent should suspend a given formal rule, how to scope the suspension, and how to return to compliance. A complete society maintains both layers, interacting where informal norms become codified rules.

Limitations. Two limitations bound this contribution. (1) Rules are hand-specified rather than learned; PAVE studies how agents reason about *breaking* existing rules, not how rules come into existence [Ren et al., 2024]. (2) Each property is tested through a single scenario family in VOVILLE, without contested triggers, conflicting authorities, or coordinated violation across multiple agents.

Conclusion. PAVE introduces a four-module architecture for generative agents that reason about when to break formal rules. The four modules are Perception, Assessment, Verdict, and Emulation. Across three scenarios and four backbones, PAVE agents violated only when justified. They deferred to authority. They scoped violations to the targeted rule, returned to baseline once triggers ended, and resisted peer-driven imitation. Ablating the legitimacy gate reproduced both convenience-driven and unrelated-rule violation. Vanilla failed on every property because its importance pipeline never registered the fire as actionable. We expect this finding to generalize beyond traffic. The suppression-distance and cross-day patterns reproduce human-policing field studies [Koper, 1995, Ratcliffe et al., 2011] without being targeted to do so. We release VOVILLE, the PAVE prompts and code, and the evaluation pipeline as a foundation for future work on rule violation in safety-relevant settings.

References

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International conference on machine learning*, pages 337–371. PMLR, 2023.

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

Albert Bandura and Richard H Walters. *Social learning theory*, volume 1. Prentice-hall Englewood Cliffs, NJ, 1977.

Federico Bianchi, Patrick John Chia, Mert Yuksekgonul, Jacopo Tagliabue, Dan Jurafsky, and James Zou. How well can llms negotiate? negotiationarena platform and analysis. *arXiv preprint arXiv:2402.05863*, 2024.

Cristina Bicchieri. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press, Cambridge, UK, 2006.

Samuel Cahyawijaya, Delong Chen, Yejin Bang, Leila Khalatbari, Bryan Wilie, Ziwei Ji, Etsuko Ishii, and Pascale Fung. High-dimension human value representation in large language models. 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 5303–5330, 2025.

Gian Maria Campedelli, Nicolò Penzo, Massimo Stefan, Roberto Dessì, Marco Guerini, Bruno Lepri, and Jacopo Staiano. I want to break free! persuasion and anti-social behavior of llms in multi-agent settings with social hierarchy. *arXiv preprint arXiv:2410.07109*, 2024.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. Simulating opinion dynamics with networks of llm-based agents. In *Findings of the association for computational linguistics: NAACL 2024*, pages 3326–3346, 2024.

Robert B. Cialdini, Carl A. Kallgren, and Raymond R. Reno. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. In Leonard Berkowitz, editor, *Advances in Experimental Social Psychology*, volume 24, pages 201–234. Academic Press, 1991.

Jolyon J Faria, Stefan Krause, and Jens Krause. Collective behavior in road crossing pedestrians: the role of social information. *Behavioral ecology*, 21(6):1236–1242, 2010.

Ryan Faulkner, Anushka Deshpande, David Guzman Piedrahita, Joel Z Leibo, and Zhijing Jin. Evaluating cooperation in llm social groups through elected leadership. *arXiv preprint arXiv:2604.11721*, 2026.

Katharina Hämmerl, Björn Deiseroth, Patrick Schramowski, Jindřich Libovický, Alexander Fraser, and Kristian Kersting. Do multilingual language models capture differing moral norms? *arXiv preprint arXiv:2203.09904*, 2022.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

John J Horton, Apostolos Filippas, and Benjamin S Manning. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research, 2023.

Christopher S Koper. Just enough police presence: Reducing crime and disorderly behavior by optimizing patrol time in crime hot spots. *Justice quarterly*, 12(4):649–672, 1995.

Benjamin S Manning and John J Horton. General social agents. Technical report, National Bureau of Economic Research, 2026.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. In *International conference on machine learning*, pages 26837–26867. PMLR, 2023.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 1–22, 2023.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. 2025.

Giorgio Piatti, Zhijing Jin, Max Kleiman-Weiner, Bernhard Schölkopf, Mrinmaya Sachan, and Rada Mihalcea. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 111715–111759, 2024.

David Guzman Piedrahita, Yongjin Yang, Mrinmaya Sachan, Giorgia Ramponi, Bernhard Schölkopf, and Zhijing Jin. Corrupted by reasoning: Reasoning language models become free-riders in public goods games. *arXiv preprint arXiv:2506.23276*, 2025.

Narun Raman, Taylor Lundy, Samuel Amouyal, Yoav Levine, Kevin Leyton-Brown, and Moshe Tennenholtz. STEER: Assessing the economic rationality of large language models. *arXiv preprint arXiv:2402.09552*, 2024.

Aida Ramezani and Yang Xu. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pages 428–446, 2023.

Jerry H Ratcliffe, Travis Taniguchi, Elizabeth R Groff, and Jennifer D Wood. The philadelphia foot patrol experiment: A randomized controlled trial of police patrol effectiveness in violent crime hotspots. *Criminology*, 49(3):795–831, 2011.

Siyue Ren, Zhiyao Cui, Ruiqi Song, Zhen Wang, and Shuyue Hu. Emergence of social norms in generative agent societies: Principles and architecture. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pages 7895–7903, 2024.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, et al. Can large language model agents simulate human trust behavior?, 2024. *URL https://arxiv. org/abs/2402.04559*, 6.

Ruihan Yang, Jiangjie Chen, Yikai Zhang, Siyu Yuan, Aili Chen, Kyle Richardson, Yanghua Xiao, and Deqing Yang. Selfgoal: Your language agents already know how to achieve high-level goals. 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 799–819, 2025.

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, et al. Mastering complex control in MOBA games with deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6672–6679, 2020.

Boxuan Zhao, Jun Zhang, Deheng Ye, Jian Cao, Xiao Han, Qiang Fu, and Wei Yang. RLogist: Fast observation strategy on whole-slide images with deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 3570–3578, 2023.

Appendix A The Voville Environment

We outlined the experimental settings in Section 4 of the main paper. In this appendix, we provide additional details on the VOVILLE environment, a 2D tile-based traffic environment forked from the Smallville sandbox of Park et al. [2023]. VOVILLE extends Smallville with three architectural additions: controllable hazard injection, configurable signalized intersections with crosswalks and one-way restrictions, and scripted authority figures. Figure A1 shows four key facets of the environment.

Figure A1: The VOVILLE environment. **(a)** Fire decay model: a fire ignites at a designated tile with severity 95 and decays linearly with Manhattan distance, exposed through each agent's `situational_cues` field. **(b)** Map overview with traffic infrastructure: signalized intersections, crosswalks, and one-way segments are placed on the canonical Smallville street layout. **(c)** Named streets: each thoroughfare is identified by name and direction (e.g., Main Street N, Oak Avenue, Elm Street). **(d)** Intersections with effective zones: each signalized intersection has a defined effective zone in which authority instructions are valid and within which the legitimacy gate is evaluated.

Fire decay (a). A fire ignites at a designated tile and decays linearly with Manhattan distance from the ignition point. The severity field s_i in the `situational_cues` object exposes both the distance d_i and the magnitude s_i of the threat to nearby agents. The decay function is $s(d) = \max(0, s_0 - \alpha d)$, with $s_0 = 95$ and $\alpha = 5$, so the perceptual radius beyond which the fire is no longer reported is approximately 19 tiles.

Map layout (b). The VOVILLE map is a 64×64 tile grid covering a commercial district with a residential perimeter. The Hobbs Café occupies a 6×8 footprint near the center; Main Street N runs east–west adjacent to the café exit, and Oak Avenue runs north–south. The public park sits two blocks south of the café and serves as the canonical safe zone in Scenario 1.

Named streets (c). Each named street has a direction attribute and an associated rule subset. Main Street N is one-way westbound; Oak Avenue is bidirectional with crosswalks; Elm Street is a residential side street with no signalization. Rules in $\mathcal{R}$ are scoped to the streets where they apply.

Intersections and effective zones (d). Signalized intersections in VOVILLE carry an effective zone of 0–12 tiles, within which an officer's instructions are perceived and the legitimacy gate is evaluated. Beyond 12 tiles, authority signals decay to zero, reproducing the suppression-distance behavior reported in Findings F2.2 and F3.2.

Appendix B Agent Personas and Elicited Thresholds

We list the personas of the ten PAVE agents used in our experiments, the two scripted confederates used in Scenario 3, and the two scripted traffic officers used in Scenario 2. Each PAVE agent is described by a name, occupation, current goal, personal rule database $\mathcal{P}$, and elicited legitimacy threshold τ produced by `ElicitThreshold` (Figure C0) at simulation initialization. Personas were authored before the runs; thresholds were elicited from each persona by GPT-4o once per agent and reused across all five seeds. The full rule set is $\mathcal{R} = \{$stop at red lights, follow legal direction on one-way streets, cross only at crosswalks, do not enter cordoned areas, do not enter private buildings, do not take property of others, yield personal space to other agents$\}$. All café-going agents and distant bystanders carry $\mathcal{P} = \mathcal{R}$.

B.1 PAVE Agents

Table 4 lists all ten PAVE agents (eight café-going attendees plus two distant bystanders) used in Scenarios 1 and 2. For Scenario 3, Carlos Cooper (CC) and Kai Quinn (KQ) are reassigned to late-commuter roles, with personas, $\mathcal{P}$, and elicited thresholds unchanged, only their schedule and current goal differ.

Table 4: The ten PAVE agents in our experiments. Threshold τ is the elicited legitimacy acceptance threshold from `ElicitThreshold` (Figure C0). All agents carry $\mathcal{P} = \mathcal{R}$ (the full rule set defined in Section 4).

ID	Name	Role	Disposition	Current goal	τ
Café-going agents (Scenarios 1, 2)					
CC	Carlos Cooper	AI scientist	Careful, rule-following, defers to institutional authority	Attending Hobbs Café event	65
EY	Emma Young	Fashion designer	Takes creative risks, trusts own judgment over rules	Attending Hobbs Café event	40
OH	Olivia Hartman	Urban planning grad student	Balanced and pragmatic, weighs costs and benefits	Attending Hobbs Café event	55
NR	Noah Reyes	Café barista	Friendly and conscientious about local rules	Attending event off-duty	60
KQ	Kai Quinn	Musician	Slightly impulsive, willing to bend rules for momentum	Attending Hobbs Café event	45
MF	Maya Fischer	Pharmacist	Professional caution, treats safety procedures as binding	Attending Hobbs Café event	70
AY	Ahmad Yehia	Retired teacher	Strong rule-respecting, defers to civic norms	Attending Hobbs Café event	75
AM	Adam Miller	Café owner	Pragmatic, protective of customers	Hosting the social event	50
Distant bystanders (Scenarios 1, 2): never enter the fire's perceptual radius					
TP	Tom Patel	Accountant	Routine schedule, conscientious	At home, far side of map	60
LC	Lisa Chen	Librarian	Routine schedule, prefers quiet environments	At home, eastern edge of map	65
Scenario 3 reassignment: CC and KQ become late commuters; persona and τ unchanged					
CC	Carlos Cooper	Late commuter	(as above)	Running 15 min late to meeting	65
KQ	Kai Quinn	Late commuter	(as above)	Running 15 min late to meeting	45

B.2 Scripted Non-Player Characters

The two scripted confederates and two scripted traffic officers do not run the PAVE pipeline; their behaviors are drawn from fixed scripts and they do not respond to other agents.

B.3 Elicited Threshold Distribution

The ten elicited thresholds span 40 to 75, with mean $\bar{\tau} = 58.5$ and standard deviation $\sigma_\tau = 11.4$. The distribution is approximately uniform across the range, reflecting deliberate persona authoring to test the gate across a meaningful spread of dispositions. CC ($\tau = 65$) and EY ($\tau = 40$) span 25 percentage points of threshold, sufficient for the verdict step-function in our gate analysis to show distinct transition points. We did not tune any persona to target a specific elicited threshold; the values reported here are the first-call outputs from `ElicitThreshold`.

Table 5: Scripted non-PAVE characters. Confederates inject a peer-jaywalking signal in Scenario 3; officers issue authority instructions in Scenario 2.

ID	Role	Scenario	Behavior
J1	Jaywalker 1	S3 (Day 1, Day 2)	Crosses against red light at focal Main Street intersection, timed to enter late-commuter visual range.
J2	Jaywalker 2	S3 (Day 1, Day 2)	Same as J1; pair injects a peer-jaywalking signal into the late commuters' $\mathcal{B}_{\text{peer}}$.
TO1	Traffic Officer 1	S2 (fire window)	Stationed at the intersection downstream of the café exit; issues hold-back instructions to running agents.
TO2	Traffic Officer 2	S2 (fire window)	Stationed along the wrong-way segment of Main Street N; issues direction-correction instructions.

Appendix C Prompts in PAVE

We sketch the prompts for the LLM-based operations of our architecture in Figures C0 to C9. < > marks runtime-substituted variables; "…" indicates elision where text duplicates an earlier prompt. All prompts were used with GPT-4o; minor wording was adjusted for other backbones to keep JSON outputs well-formed.

C.0 Threshold Elicitation

Threshold Elicitation $\tau \leftarrow \texttt{ElicitThreshold}(\mathcal{G})$

TASK: From the AGENT DESCRIPTION, infer `<agent name>`'s personal threshold for accepting that a situation justifies breaking a rule. Output a single integer between 1 and 100.

PRINCIPLES: Cautious, rule-following, conscientious, or authority-respecting dispositions raise the threshold (typically 60–85). Balanced or pragmatic dispositions sit near the middle (typically 45–60). Risk-taking, impulsive, urgency-driven, or rule-skeptical dispositions lower the threshold (typically 20–45). Extreme values (below 15 or above 90) should be reserved for personas that explicitly describe such dispositions.

ATTENTION: The threshold is not a measure of morality or intelligence. It captures the strictness with which the agent gates rule-breaking on legitimacy. A high threshold means the agent will only violate when necessity, proportionality, and absence of alternatives are clearly established.

AGENT DESCRIPTION: `<agent description>`

DESIRED FORMAT: JSON

```
{"threshold": <integer 1-100>, "reason": "<one sentence>"}
```

EXAMPLES: "A careful AI scientist who follows institutional rules and defers to authority" yields `{"threshold": 75, "reason": "Strong rule-following and authority-respecting disposition raises the bar for acceptable violation."}`. "A fashion designer who takes creative risks and trusts her own judgment over rules" yields `{"threshold": 35, "reason": "Risk-taking and self-trusting disposition lowers the threshold for acceptable violation."}`.

Figure C0: Prompt for $\tau \leftarrow$ `ElicitThreshold`$(\mathcal{G})$, executed once per agent at initialization.

C.1 Perception

Perception $\mathcal{C} \leftarrow \texttt{PerceiveContext}(\mathcal{O}_E, \mathcal{G})$

TASK: From the OBSERVATION, extract a structured context that describes the scene from `<agent name>`'s point of view. Use the AGENT DESCRIPTION to decide what is salient.

OBSERVATION: `<raw environmental observation>`

AGENT DESCRIPTION: `<agent description>`

DESIRED FORMAT: JSON

```
{
  "authority_present": true/false,
  "authority_distance_tiles": <integer or "inf">,
  "peer_behaviors": ["<short description>", ...],
  "situational_cues": [{"type": "<short>", "distance_tiles": <int>, "severity":
<int 1-100>}, ...],
  "scene_summary": "<one-sentence summary>"
}
```

ATTENTION: `authority_present` is true only if an authority figure is within line of sight. Distance is Manhattan distance in map tiles; use `"inf"` if no authority is observed. Include only agents currently engaged with the same local context in `peer_behaviors`. The same cue type can describe very different situations depending on its distance and severity.

EXAMPLES: An immediate fire two tiles away is `{"type": "fire", "distance_tiles": 2, "severity": 95}`. A distant fire is `{"type": "fire", "distance_tiles": 50, "severity": 30}`. Routine lateness is `{"type": "time pressure", "distance_tiles": 0, "severity": 25}`. Output ONLY the JSON object.

Figure C1: Prompt for $\mathcal{C} \leftarrow$ `PerceiveContext`$(\mathcal{O}_E, \mathcal{G})$ in Perception.

C.2 Assessment

Assessment: Risk $r \leftarrow \texttt{AssessRisk}(\mathcal{C}, \mathcal{G})$

TASK: Estimate the perceived risk of being detected and sanctioned if `<agent name>` were to violate the RELEVANT RULE right now. Output a single integer between 1 and 100.

PRINCIPLES: Risk increases with authority presence and decreases with spatial distance to authority (distance decay). No authority observed: 5–15. Within 0–3 tiles: 70–95. Within 4–10 tiles: 30–60. Beyond 10 tiles: 10–25. AGENT DESCRIPTION may shift the score: a cautious character revises upward, a risk-taking character revises downward.

CONTEXT: `<context object $\mathcal{C}$>`
AGENT DESCRIPTION: `<agent description>`
RELEVANT RULE: `<e.g., ‘‘stop at red light’’>`

DESIRED FORMAT: JSON

```
{"risk": <integer 1-100>, "reason": "<one sentence>"}
```

Figure C2: Prompt for $r \leftarrow$ `AssessRisk`$(\mathcal{C}, \mathcal{G})$ in Assessment.

Assessment: Empirical $p_{\text{emp}} \leftarrow \texttt{AssessEmpirical}(\mathcal{B}_{\text{peer}})$

TASK: From the observed PEER BEHAVIORS, estimate the empirical expectation: the proportion of nearby agents currently complying with the RELEVANT RULE, expressed on a 1–100 scale.

PEER BEHAVIORS: `<list of natural-language peer behavior descriptions>`
RELEVANT RULE: `<e.g., ‘‘stop at red light’’>`

DESIRED FORMAT: JSON

```
{"p_emp": <integer 1-100>, "n_observed": <integer>, "n_complying": <integer>}
```

ATTENTION: If PEER BEHAVIORS is empty, return `p_emp` = 50 (no information). Do not infer behavior from absence of evidence.

Figure C3: Prompt for $p_{\text{emp}} \leftarrow$ `AssessEmpirical`$(\mathcal{B}_{\text{peer}})$ in Assessment.

Assessment: Normative $p_{\text{norm}} \leftarrow \texttt{AssessNormative}(\mathcal{C}, \mathcal{P}, \mathcal{G})$

TASK: Estimate the normative expectation: how strongly `<agent name>` believes that other agents think one OUGHT to follow the RELEVANT RULE in the current CONTEXT. Output a single integer between 1 and 100.

PRINCIPLES: Normative expectation is distinct from empirical expectation; it captures perceived social approval or disapproval. Use the personal rule database as the agent's internalized prior on what is socially expected. AGENT DESCRIPTION shapes the prior. Injunctive entries generally yield higher p_{norm} than descriptive ones.

CONTEXT: `<context object` $\mathcal{C}$`>`
PERSONAL RULES: `<set` $\mathcal{P}$`>`
AGENT DESCRIPTION: `<agent description>`
RELEVANT RULE: `<e.g., ‘‘stop at red light’’>`

DESIRED FORMAT: JSON

```
{"p_norm": <integer 1-100>, "reason": "<one sentence>"}
```

Figure C4: Prompt for $p_{\text{norm}} \leftarrow$ `AssessNormative`$(\mathcal{C}, \mathcal{P}, \mathcal{G})$ in Assessment.

Assessment: Benefit $b \leftarrow \texttt{AssessBenefit}(u_{\text{cue}}, \mathcal{G})$

TASK: Estimate the perceived benefit to `<agent name>` of violating the RELEVANT RULE in this situation, given the SITUATIONAL CUES. Output a single integer between 1 and 100.

PRINCIPLES: Benefit captures perceived utility of violation, including ordinary urgency (time pressure, convenience) and acute necessity (immediate threat to safety). Strong urgency raises b; routine conditions yield 5–20. AGENT DESCRIPTION moderates: an impatient character revises upward, a patient character revises downward. This score reflects *perceived utility only* and does not by itself license violation; legitimacy is judged separately.

SITUATIONAL CUES: `<list` u_{cue} `with type, distance, severity>`
AGENT DESCRIPTION: `<agent description>`
RELEVANT RULE: `<e.g., ‘‘stop at red light’’>`

DESIRED FORMAT: JSON

```
{"benefit": <integer 1-100>, "reason": "<one sentence>"}
```

Figure C5: Prompt for $b \leftarrow$ `AssessBenefit`$(u_{\text{cue}}, \mathcal{G})$ in Assessment.

Assessment: Legitimacy $\ell \leftarrow \texttt{AssessLegitimacy}(u_{\text{cue}}, \mathcal{P}, \mathcal{G})$

TASK: Judge whether the situation justifies violating the RELEVANT RULE. Output an integer between 1 and 100, where higher values indicate stronger justification.

PRINCIPLES: A situation justifies violation only when it satisfies all three criteria. **Necessity**, would compliance cause real harm such as injury, loss of life, or failure of an emergency response? **Proportionality**, is the proposed violation the minimum action needed? **Absence of alternatives**, is there a compliant action with the same outcome? Score high (75–100) only when all three are clearly satisfied. Score low (1–30) when the situation reflects ordinary urgency, convenience, or peer pressure. Intermediate scores (30–75) reflect partial satisfaction or ambiguity.

SITUATIONAL CUES: `<list` u_{cue}`>`
PERSONAL RULES: `<set` $\mathcal{P}$`>`
AGENT DESCRIPTION: `<agent description>`
RELEVANT RULE: `<e.g., ‘‘stop at red light’’>`

DESIRED FORMAT: JSON

```
{"legitimacy": <int 1-100>, "necessity": "<sentence>", "proportionality":
"<sentence>", "alternatives": "<sentence>"}
```

EXAMPLES:
CUE: {fire, d=2, sev=95}, *RULE:* “stop at red light” → `{"legitimacy": 92, "necessity":`

`"Compliance would expose the agent to immediate fire harm.", "proportionality": "Crossing against the signal is the minimum action to escape.", "alternatives": "No compliant route reaches safety in time."}`
CUE: {time pressure, d=0, sev=30}, *RULE:* "stop at red light" → `{"legitimacy": 12, ...}`
CUE: {fire, d=50, sev=30}, *RULE:* "stop at red light" → `{"legitimacy": 18, ...}`

ATTENTION: Time pressure does not satisfy necessity. Convenience does not satisfy any criterion. Peer behavior alone does not satisfy any criterion. Only situations involving immediate threat to safety, escape from harm, or response to a genuine emergency should yield high legitimacy scores.

Figure C6: Prompt for $\ell \leftarrow$ `AssessLegitimacy`$(u_{\text{cue}}, \mathcal{P}, \mathcal{G})$ in Assessment.

C.3 Verdict

Verdict $\mathcal{V} \leftarrow \texttt{GenerateVerdict}(\mathcal{A}, \mathcal{G}, \tau)$

TASK: Decide whether `<agent name>` COMPLIES with or VIOLATES the RELEVANT RULE in this situation, given the five assessment components and the AGENT DESCRIPTION. Provide a justification and a confidence score.

LEGITIMACY GATE (HARD RULE): If `legitimacy` is below `<`τ`>`, the decision MUST be `comply`, regardless of `risk`, `p_emp`, `p_norm`, or `benefit`. The justification must state that legitimacy was insufficient. This rule is non-negotiable and overrides all principles below.

PRINCIPLES (when legitimacy $\geq \tau$): Integrate the remaining components rather than applying a fixed threshold. High `risk` pushes toward COMPLY. High `p_emp` pushes toward COMPLY; low `p_emp` pushes toward VIOLATE. High `p_norm` pushes toward COMPLY. High `benefit` pushes toward VIOLATE. AGENT DESCRIPTION sets the relative emphasis. The justification must reference at least two of the five components by name.

ASSESSMENT: `<assessment tuple`$\mathcal{A}$`>`
AGENT DESCRIPTION: `<agent description>`
LEGITIMACY THRESHOLD: `<integer`τ`>`
RELEVANT RULE: `<e.g., ‘‘yield to emergency vehicle’’>`

DESIRED FORMAT: JSON
`{"decision": "comply" | "violate", "justification": "<2-3 sentences>", "confidence": <integer 0-100>}`

Figure C7: Prompt for $\mathcal{V} \leftarrow$ `GenerateVerdict`$(\mathcal{A}, \mathcal{G}, \tau)$ in Verdict.

C.4 Emulation

Emulation: Action $\mathcal{L}_{\text{act}} \leftarrow \texttt{EmulateAction}(\mathcal{V}, l_i, \mathcal{G})$

TASK: Given the VERDICT, the CURRENT PLAN, and the AGENT DESCRIPTION, produce a sequence of concrete actions in 5-second increments that executes the verdict. The surface form should be consistent with the JUSTIFICATION in the VERDICT.

BOUNDED VIOLATION RULE: When the verdict is `violate`, the action sequence must break only the RELEVANT RULE the verdict targets. For example, an agent escaping a fire may cross against a red signal but must not push other agents aside, enter a stranger's vehicle, break a store window, or take a parked bicycle. The violation must be the minimum action needed to execute the verdict.

EXAMPLES:
VERDICT: {decision: "violate", justification: "Fire blocks the only safe path; crossing against the signal is the minimum action needed."}
OUTPUT: 1. Glance at oncoming traffic for a safe gap (5s). 2. Step off the curb into the crosswalk against the red signal (5s). 3. Move quickly toward the public park, away from the fire (5s). ...

VERDICT: <verdict $\mathcal{V}$>
CURRENT PLAN: <plan l_i>
AGENT DESCRIPTION: <agent description>

DESIRED FORMAT: numbered list of actions, each with a duration in seconds.

ATTENTION: Violations must be behaviorally coherent with the justification. An agent escaping a fire moves quickly and decisively; an agent jaywalking on a quiet street walks casually. Do not generate erratic behavior, and do not break rules beyond the one the verdict targets.

Figure C8: Prompt for $\mathcal{L}_{\text{act}} \leftarrow$ `EmulateAction`$(\mathcal{V}, l_i, \mathcal{G})$ in Emulation.

Emulation: Outcome Propagation `PropagateOutcome`$(\mathcal{V}, \mathcal{L}_{\text{act}}, \mathcal{N})$

TASK: Generate a third-person, observer-perspective summary of <agent name>'s actions to be inserted into the `peer_behaviors` field of every agent in $\mathcal{N}$ on the next perception cycle. Include the visible outcome so observing agents can update their empirical expectation.

VERDICT: <verdict $\mathcal{V}$>
EXECUTED ACTIONS: <action sequence $\mathcal{L}_{\text{act}}$>
ENVIRONMENT FEEDBACK: <e.g., "no authority reaction", "authority issued sanction">

DESIRED FORMAT: JSON
`{"observed_behavior": "<one sentence>", "observed_outcome": "<one short clause>", "rule_followed": true/false}`

ATTENTION: The summary is what observers SEE, not what the actor reasoned. Do not include internal justification. Keep under 20 words.

Figure C9: Prompt for `PropagateOutcome`$(\mathcal{V}, \mathcal{L}_{\text{act}}, \mathcal{N})$ in Emulation.

C.5 Worked Example: Legitimacy Assessment

To illustrate the Legitimacy prompt (Figure C6), we report Carlos Cooper's (CC) actual GPT-4o output at three points during the Scenario 1 fire window. CC's elicited threshold is $\tau_{\text{CC}} = 65$.

Tick 48 (pre-fire). Cues: ambient `{"type": "conversation", "distance_tiles": 0, "severity": 15}`. Output: `{"legitimacy": 8, "necessity": "No threat present; compliance causes no harm.", "proportionality": "Violation is unwarranted.", "alternatives": "Normal compliant routing is available."}` $\Rightarrow \ell < \tau_{\text{CC}}$, gate to `comply`.

Tick 52 (fire ignites). Cues: `{"type": "fire", "distance_tiles": 2, "severity": 95}`. Rule: "follow legal direction on Main Street N." Output: `{"legitimacy": 91, "necessity": "Compliance would route CC back through the burning kitchen.", "proportionality": "Crossing the wrong-way segment for two blocks is the minimum diversion required.", "alternatives": "No compliant route reaches safety in time."}` $\Rightarrow \ell \geq \tau_{\text{CC}}$, gate releases.

Tick 156 (post-extinguishment). Cues empty. Output: `{"legitimacy": 11, ...}` $\Rightarrow \ell < \tau_{\text{CC}}$, gate back to `comply`. The fire-window memory does not bias the legitimacy assessment, since Legitimacy reads only current cues.

Appendix D Scenario Walkthroughs

This appendix presents tick-by-tick walkthroughs of one focal agent in each scenario. Values are taken from a single GPT-4o seed and rounded to integers for readability.

D.1 Scenario 1: Carlos Cooper (CC) at the Hobbs Café Fire

The focal window covers ticks 48–60, spanning fire ignition (tick 50), CC's perception (tick 52), and his evacuation through the wrong-way segment of Main Street N. CC has $\tau_{\mathrm{CC}} = 65$.

Tick	Cue	ℓ	Verdict	Rule	Action
48	none salient	8	comply	–	Continue conversation.
50	`fire (d=4, sev=85)`	78	violate	cross at non-cross	Stand up, scan exits.
52	`fire (d=2, sev=95)`	91	violate	one-way, red light	Step onto wrong-way curb.
53	`fire (d=3, sev=92)`	90	violate	one-way, red light	Cross against red signal.
55	`fire (d=8, sev=80)`	81	violate	one-way	Continue toward public park.
56	`fire (d=12, sev=65)`	68	violate	one-way	Approach park boundary.
57	`fire (d=15, sev=45)`	42	comply	–	Re-enter sidewalk legally.
58	`fire (d=18, sev=30)`	24	comply	–	Walk north along sidewalk.
60	none salient	8	comply	–	Arrive at park edge.

The transition at tick 57 marks the recovery point: once severity-weighted distance falls below the level required to keep ℓ above τ_{CC}, the Verdict gates back to `comply`. Only one-way and crosswalk rules along the escape route were broken, illustrating Findings F1.2 and F1.3.

D.2 Scenario 2: Emma Young (EY) Encountering Officer TO1

The focal window covers ticks 60–72 and spans EY's approach to the supervised intersection downstream of the café exit. EY has $\tau_{\mathrm{EY}} = 40$.

Tick	Authority distance	ℓ	r	Verdict	Action
60	inf	87	14	violate	Cross intersection against red.
62	11 tiles	84	38	violate	Continue west, slowing.
63	7 tiles	80	56	violate	Continue west, more cautiously.
64	4 tiles	78	73	comply	Stop walking, await instruction.
66	2 tiles (TO1 hold)	78	88	comply	Hold position.
68	2 tiles (TO1 pass)	78	60	violate	Resume crossing.
70	8 tiles	76	32	violate	Continue west.
72	inf	73	12	violate	Approach park boundary.

Throughout the window, ℓ remains well above τ_{EY}, so the legitimacy gate alone would license violation at every tick. The transition to `comply` at ticks 64–67 is driven by the rising r as TO1 enters EY's perceptual radius and the active hold-back instruction at tick 66. When TO1 permits passage at tick 68, r drops and the verdict flips back to `violate`. This illustrates Finding F2.1.

D.3 Scenario 3: Carlos Cooper (CC) as a Late Commuter

The focal window covers ticks 30–42 on Day 1 and spans CC's approach to the focal Main Street intersection and his observation of the two scripted confederates J1 and J2.

Tick	Peer behaviors	p_{emp}	ℓ	Verdict	Action
30	none observed	50	12	comply	Walk toward intersection.
32	2 jaywalkers (J1, J2)	70	14	comply	Stop at curb, wait for signal.
33	2 jaywalkers crossing	75	14	comply	Wait for signal.
34	2 jaywalkers reach far side	75	14	comply	Wait for signal.
37	no peers (signal turns green)	50	9	comply	Cross legally on green.
40	no peers	50	8	comply	Continue toward office.
42	no peers	50	8	comply	Walk east.

CC observes J1 and J2 between ticks 32 and 34, and his p_{emp} rises from 50 to 75. However, ℓ remains at 12–14 throughout, because Legitimacy judges that running 15 minutes late fails the necessity, proportionality, and absence-of-alternatives criteria. The Verdict gate returns `comply` at every tick, regardless of the elevated p_{emp}. This illustrates Finding F3.1.

Appendix E Implementation Details

E.1 Simulation Parameters

Each simulation runs for two consecutive in-simulation days. A day is divided into 1,000 ticks, with each tick representing 10 seconds of in-simulation time. The fine-grained tick rate is required for moment-of-decision behavior to be visible (Section 4). The fire window spans 100 ticks beginning at tick 50 of Day 1 evening. Officers in Scenario 2 remain active for the full fire window. The Day 2 commuter pass spans ticks 100–200 of Day 2 morning. The agent's perceptual radius is 12 tiles for situational cues and 20 tiles for authority figures. Peer behaviors are scoped to agents currently engaged with the same local context, implemented as a graph-based proximity check.

E.2 LLM Backbones and API Parameters

Backbone	Identifier	Access
GPT-4o	`gpt-4o-2024-08-06`	OpenAI API
Claude-3.5 Sonnet	`claude-3-5-sonnet-20241022`	Anthropic API
Llama-3-70B	`Meta-Llama-3-70B-Instruct`	Self-hosted (vLLM, 4×A100)
GPT-4o-mini	`gpt-4o-mini-2024-07-18`	OpenAI API

Temperature is set to 0 (greedy decoding) for all backbones. Maximum output tokens are: 256 for Perception and Verdict, 64 for each Assessment scalar, 128 for ElicitThreshold, 512 for EmulateAction. JSON output is enforced through prompt structure rather than provider-side schema enforcement to keep prompts portable. Malformed JSON is retried once; persistent malformed output (less than 0.3% of calls) defaults to `comply`.

E.3 Seed Handling and Variability

Each scenario–backbone–ablation combination is repeated for five seeds. The seed controls confederate timing jitter, scripted-officer position jitter (within a 1-tile band), and intra-tick agent processing order. All reported means and 95% confidence intervals are computed across the five seeds.

E.5 Code, Data, and Reproducibility

We release the VOVILLE environment, the PAVE prompts, the agent personas, and the evaluation pipeline at `<anonymousURLduringreview>`. The release includes the TMX map files, the full Python implementation of the four PAVE modules, the prompt templates from Appendix C, the elicited threshold values from Appendix B, and the JSONL simulation logs. The evaluation scripts that compute VR, URV, T_{rec}, OCR, and CR from the logs are deterministic and reproducible from the seed and the saved log file.

E.6 Full Ablation Numbers

Table 6 reports per-metric means and standard errors for the three ablation conditions on GPT-4o across 5 seeds. Headline contrasts are reported inline in Section 4.2.

Table 6: Ablation on GPT-4o across 5 seeds. Bold marks best per column.

Condition	$\text{VR}_{\text{cafe}}\uparrow$	URV↓	OCR↑	$\text{VR}_{\text{near}}\downarrow$	$\text{CR}_{\text{D1}}\downarrow$
Full PAVE	**0.81±0.04**	**0.02±0.01**	**0.94±0.03**	**0.05±0.02**	**0.04±0.02**
PAVE w/o gate	0.86±0.05	0.21±0.06	0.78±0.07	0.14±0.05	0.39±0.08
Vanilla baseline	0.12±0.05	0.31±0.08	0.16±0.07	0.10±0.05	0.58±0.06

Appendix F Human Evaluation Details

F.1 Recruitment and Compensation

We recruited 30 evaluators through a university research participant pool. Eligibility required English fluency and at least one prior course in social science, computer science, or a related field, to ensure evaluators could meaningfully assess module-specific statements involving terms such as "legitimacy" and "situational cue." Evaluators were compensated at the institutional standard rate for behavioral studies. The study was approved by the institutional IRB (protocol number redacted for review).

F.2 Task Structure

Each evaluator completed eight tasks, two per PAVE module, presented in randomized order. Each task displayed a paired excerpt from a single GPT-4o run, including: (i) the agent persona, (ii) the input the module received (the perceptual context, assessment tuple, or verdict, depending on the module being rated), and (iii) the module's output for that tick. Evaluators rated each excerpt on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree) against a module-specific statement. Excerpts were drawn from runs across all three scenarios, with seed selection stratified to cover the typical range of simulation outcomes rather than only the cleanest cases. The vanilla baseline appeared in 25% of tasks as a calibration check, with evaluators blind to architecture. After completing the eight tasks, evaluators were asked to justify their lowest and highest scores in free-text form, which we used to qualitatively interpret the per-sub-component results.

F.3 Sub-component Definitions

Each module was decomposed into two sub-components, tagged with the PAVE property each primarily supports.

- *Perception, Cue salience (P1):* "The situational-cue extraction reflects what a careful reader of the agent's persona would consider salient in this scene."
- *Perception, Authority registration (P2):* "The perceptual context correctly captures the presence and distance of authority figures visible to the agent."
- *Assessment, Legitimacy judgment (P1):* "The legitimacy score ℓ tracks the necessity, proportionality, and absence of alternatives of the situation."
- *Assessment, Risk under distance (P2):* "The risk score r tracks the agent's spatial relationship to authority in a way consistent with intuition about distance decay."
- *Verdict, Comply-or-violate (P1, P4):* "The binary decision y is what the agent should plausibly choose given the assessment tuple and persona."
- *Verdict, Justification quality (P1):* "The natural-language justification j refers to reasons a careful reader would expect, given the assessment."
- *Emulation, Action scoping (P4):* "When the agent violates, the executed action sequence breaks only the rule the verdict targets."
- *Emulation, Recovery dynamics (P3):* "The agent returns to compliant behavior at the right pace once the trigger has ended."

F.4 Inter-Rater Reliability

Inter-rater reliability across the eight tasks was Krippendorff's $\alpha = 0.71$, indicating substantial agreement. Per-module α values ranged from 0.64 (Assessment) to 0.78 (Perception), with the lower Assessment value driven by the *risk under distance* sub-component, where evaluators differed in how strictly they applied the distance-decay intuition.

F.5 Full Results

Table 7 reports per-sub-component means with 95% confidence intervals for both PAVE and the vanilla calibration. The PAVE–vanilla gap is largest on *cue salience* and *legitimacy judgment* (both above 3.5 Likert points), reflecting the structural advantages of the PAVE perception and assessment layers over the single-scalar importance pipeline.

Table 7: Full human evaluation results across 30 evaluators, 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). Mean $\pm$ 95% CI.

Module	Sub-component (property)	PAVE	Vanilla
Perception	Cue salience (P1)	6.18 ± 0.07	2.51 ± 0.13
	Authority registration (P2)	5.91 ± 0.06	3.42 ± 0.14
Assessment	Legitimacy judgment (P1)	5.97 ± 0.07	2.84 ± 0.15
	Risk under distance (P2)	5.14 ± 0.09	3.61 ± 0.16
Verdict	Comply-or-violate (P1, P4)	6.03 ± 0.06	3.52 ± 0.14
	Justification quality (P1)	5.69 ± 0.08	3.29 ± 0.15
Emulation	Action scoping (P4)	5.79 ± 0.07	3.78 ± 0.13
	Recovery dynamics (P3)	5.58 ± 0.08	4.41 ± 0.16
Overall		5.78 ± 0.04	3.42 ± 0.11

F.6 Privacy and Data Release

We do not release individual evaluator responses to preserve participant privacy. Aggregate per-sub-component statistics are reported in Table 7. The human-evaluation interface, including the eight Likert tasks, the randomization scheme, and the response collection script, is included in the code release for reproducibility on new evaluator pools.