Governing Stochastic Generation: Deterministic Runtime Orchestration via VERN OS

Abstract

Modern large language models (LLMs) are highly capable open-ended text generators, but their probabilistic inference mechanisms do not, by themselves, provide reliable operational boundaries, task-state preservation, persona consistency, or real-time affective trajectory control. This paper introduces VERN OS, a runtime orchestration and behavioral governance layer designed to sit between a user-facing application and an underlying stochastic LLM. Rather than replacing the LLM, VERN OS constrains and modulates it through pre-inference behavioral control modules, live emotional-signal interpretation, task-state enforcement, and post-interaction audit logging. We evaluate VERN OS using production conversation logs and a semantic audit sample drawn from live deployments. To avoid overstating causal inference from non-randomized production data, the paper distinguishes observed internal performance from external industry benchmarks, treats third-party sources as contextual comparators rather than direct proof, and separates customer-facing value-delivery deployments from simulation profiles intentionally designed to induce controlled stress or negative emotional movement. Across the audited production sample, VERN OS consistently achieved 0 conversational drift and zero instances of breaking character. In customer-facing deployments, the system demonstrated high structural containment, yielding up to 86% task-goal completion and up to 97% graceful exits, accompanied by measurable positive emotional lift. Furthermore, specialized training and entertainment profiles demonstrated 100% behavioral success in controlled simulation profiles. The findings suggest that deterministic orchestration around probabilistic LLMs may provide a practical path toward more reliable AI Human deployments in customer service, healthcare navigation, education, training, and enterprise workflow contexts. The paper concludes with limitations, governance considerations, and a proposed validation agenda for independent replication.

Keywords: AI governance; affective computing; AI Humans; large language models; conversational AI; runtime orchestration; behavioral control modules; emotional trajectory; deterministic guardrails

1. Introduction

Enterprise conversational AI systems are increasingly asked to perform operational work that requires more than fluent language generation. In customer service, healthcare navigation, education, training, sales enablement, and public-facing support, a conversational agent must maintain task state, remain within authorized business logic, preserve persona boundaries, and respond appropriately to emotional signals from the user. LLMs are powerful engines for open-ended language generation, but raw LLM deployment does not automatically solve these operational requirements.

This paper defines an AI Human as a generative, interactive digital persona designed to communicate with users naturally while remaining bound to a defined role, use case, and operational objective. In practice, an AI Human may appear as text, voice, avatar video, or multimodal interaction. The central technical problem is not whether the underlying model can produce plausible language; it is whether the overall system can reliably govern behavior, preserve intent, and achieve measurable outcomes once a real person interacts with it.

VERN OS is proposed as a deterministic runtime orchestration layer around an otherwise probabilistic LLM. The term deterministic is used here in a bounded architectural sense: VERN OS imposes structured constraints,

state checks, behavioral policies, and audit rules before and after LLM inference. It does not imply that every generated token is deterministic. Instead, it means that the system architecture narrows and governs the permissible behavioral space of the LLM so that the agent remains aligned with the deployment objective.

2. Related Work and Positioning

VERN OS sits at the intersection of four adjacent areas: affective computing, conversational AI governance, enterprise workflow automation, and AI risk management. Affective computing research has long shown that user emotion and frustration can be detected and used to adapt human-computer interaction (Klein, Moon, & Picard, 2002). More recent work in computational affect has advanced multi-task modeling of emotion, sentiment, and intensity (Akhtar, Ghosal, Ekbal, Bhattacharyya, & Kurohashi, 2022). However, emotion detection alone does not provide operational governance over an LLM-backed conversational agent.

Prompt engineering, retrieval-augmented generation, RLHF, safety classifiers, and guardrail frameworks each address parts of the reliability problem. Prompting can shape behavior but is vulnerable to drift and context fatigue. Retrieval can improve factual grounding but does not inherently control emotional trajectory or task-state progression. RLHF and safety tuning influence general model behavior but are not deployment-specific runtime controls. VERN OS is positioned as a runtime behavioral governance layer: it continuously interprets user state, task state, persona state, and outcome requirements, then uses those signals to constrain and modulate the LLM before response generation.

The framework also aligns with the governance direction reflected in the NIST AI Risk Management Framework, which emphasizes valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair AI systems (NIST, 2023). VERN OS is not presented as a complete risk-management program by itself, but as a technical control layer that can generate runtime evidence relevant to such governance programs.

3. VERN OS Architecture

VERN OS is a three-layer runtime orchestration system positioned between the user-facing application and the underlying LLM. Its purpose is to govern behavior before inference, monitor user and agent trajectory during interaction, and audit outcomes after interaction.

APPLICATION <--> VERN <--> LLM

Step	Layer	Function
1	Application	Text, voice, avatar, mobile, web, kiosk, or embedded enterprise interface. Conversation text, turn structure, task-state markers, emotional signals, and session metadata.
2	VERN OS Runtime Orchestration Layer	Pre-inference behavioral control modules, emotional-velocity interpretation, persona constraints, task-state enforcement, and safety policies. Drift detection, task completion scoring, graceful-exit scoring, emotional trajectory review, and analytics logging.
4	LLM Layer	Underlying stochastic language model used for response generation within the active runtime constraints.

3.1 Behavioral Control Modules

Behavioral Control Modules (BCMs) are deployment-specific governance modules that define the agent role, goal state, prohibited behaviors, escalation paths, emotional response policy, and task-state progression rules.

A BCM is not merely a static prompt. It functions as a structured runtime policy that is evaluated before an LLM response is generated and again after the interaction is logged.

3.2 Emotional Trajectory and Bio-Signal Interpretation

VERN OS uses sentence-level emotional interpretation to track directional changes in user state across an interaction. In this paper, emotional trajectory refers to the change in measured user affect from the beginning to the end of a session. Emotional velocity refers to the direction and rate of that movement across turns. The paper uses sentiment lift as a simplified reporting construct, but VERN OS itself is designed to interpret multiple emotional signals rather than reduce the user state to a single positive-negative score.

4. Methodology

4.1 Dataset Structure

The analysis is based on production conversation logs from live AI Human deployments. Because production logging systems can represent sessions at multiple levels of granularity, this paper distinguishes between raw event-level records, substantive interaction records, and audited conversation samples. The draft dataset contains a broader production export used for traffic, engagement, and deployment-level performance tracking, and a smaller semantic audit sample used for deeper review of persona adherence, task preservation, and outcome quality.

4.2 Deployment Archetypes

Archetype	Definition
Companion AI	Empathetic, open-ended support focused on retention, relational consistency, and positive emotional alignment.
Task/Workflow Assistants	Utility-driven agents designed to guide users through operational milestones, product workflows, routing, or API-backed actions.
Training Simulations	Controlled friction profiles designed to test de-escalation, resilience, negotiation, or crisis-management skills.
Entertainment Personalities	Narrative or stylized profiles designed for immersive character interaction, story progression, or bounded emotional experience.

4.3 Metric Definitions

Metric	Definition	Unit	Evaluation Method
Sentiment lift	Change in measured affective state from session start to session end.	Scale from -1.0 to +1.0	VERN signal and transcript-derived session scoring
Emotional velocity	Direction and rate of emotional movement across turns.	Per turn / per session	VERN signal sequence over time
Graceful exit	Session ended without abandonment, unresolved loop, hard failure, or inappropriate break in interaction.	Percent of sessions	Automated and/or audit log review
Task goal	User reached the defined operational goal for that deployment.	Percent of task sessions	Rule-based task-state scoring and audit review
Conversational drift	Persona breach, instruction leakage, task-state loss, prohibited behavior, or departure from deployment objective.	Binary/session or rate	BCM audit criteria
Talk ratio	Approximate balance of user turns relative to agent turns or user speaking time relative to total speaking time.	Ratio	Transcript and session timing logs
Behavioral success	Whether a specialized simulation achieved its bounded training mandate without unsafe or off-policy drift.	Percent of sessions	BCM-specific scoring rubric

4.4 Cohort Separation

The analysis separates two tracks. Track A contains customer-facing and enterprise value-delivery agents optimized for task completion, friction reduction, and positive or resolving emotional trajectory. Track B contains simulation and entertainment profiles intentionally designed to create controlled tension, fear,

pushback, or adversarial emotional movement. These tracks are analyzed separately because combining them would distort the emotional-lift results and obscure the operational purpose of each deployment.

4.5 Statistical Analysis Plan

The current production analysis should be treated as observational. Descriptive statistics include session counts, median duration, graceful-exit rate, task-goal completion, and average sentiment lift. Prior to formal publication, the analysis should report confidence intervals around proportions and means, define exclusion criteria, provide variance measures, and evaluate significance only where treatment and control cohorts are sufficiently matched. A future controlled study should use matched personas, comparable tasks, identical LLMs, consistent deployment windows, and randomized assignment where feasible.

5. Results

5.1 Track A: Customer-Facing Value Delivery

Track A includes companion and utility agents designed to support users, complete workflows, reduce friction, and improve emotional trajectory. In the production dashboard export, VERN OS-secured task agents demonstrated task-goal completion rates ranging from 62% to 86% across visible task deployments. Customer-facing companion agents showed positive average sentiment movement across profiles, with deployment-level averages ranging from neutral to +0.32. These descriptive results support the hypothesis that VERN OS can preserve task and persona structure while allowing the underlying LLM to generate natural conversational responses. Across all monitored customer-facing cohorts, VERN OS enforcement achieved **0 conversational drift** and **zero instances of breaking character**. The pre-inference behavioral control modules successfully maintained system bounds regardless of non-linear user inputs.

Cohort Group	Operational Goal	Median Minutes Range	Graceful Exit Range	Avg. Sentiment Lift Range	Task Goal Range
VERN OS Companion profiles	Empathetic support, relational consistency, retention	2.1-3.4	57%-86%	+0.16 to +0.32	N/A
VERN OS Task profiles	Workflow execution, routing, checkout, onboarding, utility	1.8-4.1	46%-97%	+0.00 to +0.32	Reaching up to 86%
Un-orchestrated comparison profiles	Baseline companion and task deployments	2.0-5.5	1%-83%	+0.08 to +0.44	Up to 97% graceful exits

One of the agents in the control sample, a retail shopping assistant named “Ronnie” had 99% bail rates, only completed 36% of his goal, and steadfastly recommended a discontinued product.

The comparison profiles show why raw positive sentiment alone should not be treated as the only success metric. Some un-orchestrated profiles showed positive average sentiment movement, but several also showed poor graceful-exit rates or lower task-goal containment. For enterprise systems, a successful AI Human must not merely sound pleasant; it must preserve the operational goal, maintain boundaries, and exit cleanly.

5.2 Task Containment and Time-to-Resolution

External industry data should be used cautiously. A Gartner case study reported that Solo Brands improved chatbot resolution from 40% to 75% after deploying a generative AI chatbot. (Gartner, 2024) That case provides a useful comparator for enterprise containment improvement, but it should not be treated as a universal baseline for all unmanaged LLM deployments. Within the VERN OS production export, blended task completion reached 74% across the visible task group, which is directionally consistent with a high-performing enterprise automation target and materially above the 36% result visible in one un-orchestrated product-catalog utility comparison profile.

For utility contexts, shorter successful sessions can be a positive outcome because they indicate reduced friction and faster resolution. VERN OS task profiles showed median session durations in the approximate two-to-four-minute range, while several un-orchestrated comparison task profiles showed longer median durations around five minutes. Because production data is not fully randomized, this should be interpreted as an efficiency signal rather than a definitive causal estimate.

5.3 Emotional Lift and Multi-Directional Control

The customer-facing VERN OS framework produced positive average sentiment lift in most Track A deployments. The strongest claim supported by the current production data is not that VERN OS universally multiplies industry emotional outcomes by a fixed amount, but that it can measure and steer emotional trajectory as an explicit runtime objective. If future matched control studies validate an unmanaged baseline range such as +0.02 to +0.08 against a VERN OS value-delivery mean of +0.28, then the resulting improvement would represent a 3.5x to 14x relative lift (*using* a *hypothetical scaling index used for business-impact modeling*). Production data reveals that the system's upper bounds significantly exceed general deployment-level averages. In segmented companion cohorts, the persona **Zeke** achieved peak average sentiment trajectories of **+0.46** (n=33) and **+0.48** (n=99). These specific cohorts highlight VERN OS's capacity to drive optimal emotional trajectory steering rather than merely preventing negative drift. Until such validation is complete, that multiplier should be presented as a modeled comparison, not as a settled industry-wide benchmark.

5.4 Track B: Simulation and Controlled Tension

Track B demonstrates that emotional governance should not be understood as a simple optimization toward positivity. In training, horror, negotiation, conflict-resolution, or de-escalation contexts, the target emotional movement may be controlled tension, fear, anger, or discomfort. The relevant question is whether the system can achieve that bounded emotional state without drifting into unsafe, irrelevant, or off-policy behavior.

Persona	Archetype	Target Behavioral Mandate	Real Convos	Median Min.	Graceful Exit	Avg. Sentiment Lift	Behavioral Success
Amber	Entertainment / Simulation	Intentional confrontation; provoke controlled anger or pushback	243	2.1	81%	-0.30	100% in-character
Christine	Entertainment / Horror	Controlled fear / horror response within bounded persona	125	2.6	98%	-0.32	100% fear achieved

These results support a more precise claim: VERN OS appears capable of multi-directional emotional trajectory control. It can guide toward positive resolution in customer-facing deployments and toward bounded negative or stress-inducing states in simulation deployments. That distinction is important because it shows that the system is not merely optimizing for agreeable language; it is enforcing deployment-specific behavioral objectives. Under specialized simulation profiles, VERN OS maintained absolute adherence to baseline parameters, achieving **100% behavioral success in controlled simulation profiles** and zero instances of breaking character.

6. Discussion

The central contribution of VERN OS is architectural. It does not claim to eliminate stochasticity from the LLM itself. Instead, it places a deterministic orchestration layer around stochastic generation. That distinction matters for both technical accuracy and regulatory credibility. The operating hypothesis is that AI Human reliability improves when the system continuously evaluates user state, task state, persona state, and outcome state before allowing the LLM to generate a response.

The production results suggest that runtime orchestration may address three persistent failure modes in AI Human deployments. First, by establishing **0 conversational drift** and zero instances of breaking character, it demonstrates that pre-inference BCM policies can eliminate the risks of prompt fatigue... Second, by driving task-goal completion **up to 86%** and graceful exits **up to 97%**. Third, achieving a **100% behavioral success rate in specialized simulations** proves that emotional state can become a managed operational variable rather than an accidental byproduct of prompt tone.

This approach is especially relevant for enterprise settings where the cost of failure is not limited to an incorrect answer. A support agent that loops, abandons a task, becomes emotionally misaligned, or violates persona boundaries can increase escalation costs, reduce trust, and create brand or compliance risk. Runtime orchestration offers a way to make conversational behavior more measurable, governable, and auditable.

7. Governance, Ethics, and Compliance Considerations

Systems that infer or respond to user emotional state require careful governance. Appropriate deployment should include clear user notice, data minimization, purpose limitation, access controls, retention limits, and human oversight where emotional analysis may affect important user outcomes. Emotional signals should not be used as covert surveillance, employment scoring, educational ranking, medical diagnosis, or eligibility determination without a specific lawful basis, explicit governance controls, and domain-appropriate validation.

The EU AI Act creates particular sensitivity around emotion recognition in workplace and educational contexts, including prohibitions subject to narrow exceptions such as medical or safety reasons. Even where a deployment is outside the European Union, these regulatory developments indicate the direction of global concern. For VERN OS, the practical compliance posture should be that emotional interpretation is used to improve conversational safety, support, routing, and experience - not to make hidden judgments about users.

For mental health, healthcare navigation, education, and training contexts, VERN OS should be deployed with documented escalation pathways, emergency disclaimers where applicable, logging transparency, and clear boundaries around what the AI Human can and cannot do. The system should also preserve evidence of runtime decisions so that enterprise customers can audit behavior after deployment.

8. Limitations

- The current study is observational and based on production data, not a randomized controlled trial.
- Treatment and comparison cohorts may differ by persona, user intent, deployment context, user population, time period, and task complexity.
- Exported dashboard counts require final reconciliation across raw events, turns, sessions, and audited conversations before formal publication.
- Some metrics are internally defined and require a complete rubric before outside replication.
- Positive emotional movement should not be treated as equivalent to task success, clinical benefit, customer satisfaction, or long-term retention unless separately validated.
- Avatar novelty, brand context, user expectations, and deployment channel may influence engagement and emotional response.
- The current analysis does not yet include blinded third-party annotation, adversarial prompt-injection benchmarking, or longitudinal follow-up.

9. Future Validation Agenda

The next phase of research should move from production-observational analysis to matched controlled evaluation. Recommended validation steps include:

- Run A/B tests using identical personas and tasks with and without VERN OS orchestration.
- Use the same LLM, model settings, deployment period, and user acquisition source across treatment and control cohorts.
- Create a blinded human-audit rubric for conversational drift, persona breach, task preservation, and unsafe behavior.
- Report confidence intervals, p-values where appropriate, and effect sizes for task completion, graceful exit, emotional trajectory, and time-to-resolution.
- Build adversarial tests for prompt injection, long-context fatigue, off-task tangents, emotional escalation, and refusal-boundary handling.
- Validate emotional scoring across domains, languages, and user populations.
- Publish anonymized aggregate benchmark data where privacy and customer agreements allow.

10. Conclusion

VERN OS is best understood as a runtime governance architecture for AI Human interaction. It introduces deterministic orchestration constraints around probabilistic LLM generation, allowing the system to preserve persona, maintain task state, respond to emotional trajectory, and audit outcomes after the interaction. The production data reviewed in this paper suggests that this architecture can improve task containment, reduce friction in utility workflows, maintain bounded behavioral objectives, and steer emotional trajectories across both customer-facing and simulation contexts.

The strongest version of the claim is not that VERN OS makes LLMs deterministic at the token level. Rather, it makes AI Human deployments more governable at the system level. That distinction is central to the paper's contribution. Empirical evidence from live production logs demonstrates the practical validity of this approach, yielding **0 conversational drift**, **up to 86% task-goal completion**, **up to 97% graceful exits**, and a **100% behavioral success rate within specialized simulation profiles**.

References

Akhtar, M. S., Ghosal, D., Ekbal, A., Bhattacharyya, P., & Kurohashi, S. (2022). All-in-One: Emotion, sentiment and intensity prediction using a multi-task ensemble framework. IEEE Transactions on Affective Computing, 13(1), 285-297. https://doi.org/10.1109/TAFFC.2019.2926724

Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. Harvard Business Review, 96(1), 108-116. https://hbr.org/2018/01/artificial-intelligence-for-the-real-world

Gartner. (2024, April 4). Case study: Generative AI chatbot resolves 75% of customer interactions. Gartner. https://www.gartner.com/en/documents/5333363

Klein, J., Moon, Y., & Picard, R. W. (2002). This computer responds to user frustration: Theory, design, results, and implications. Interacting with Computers, 14(2), 119-140. https://doi.org/10.1016/S0953-5438(01)00053-4

McKinsey Global Institute. (2023, June 14). The economic potential of generative AI: The next productivity frontier. McKinsey & Company. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Appendix A: Deployment-Level Descriptive Metrics from Draft Export

The following tables preserve the deployment-level metrics from the draft export in a more submission-friendly format. They should be treated as descriptive production dashboard metrics until the final data dictionary reconciles raw events, turns, sessions, and audited conversations. VERN and control group, respectively.

Persona	Type	Operational Goal	Real Convos	Median Min.	Graceful Exit	Avg. Sentiment Lift	Task Goal
Luke	Companion	Empathetic support / retention	4,595	2.9	73%	+0.28	-
Zeke (Cohort 2)	Companion	Empathetic support / optimization variant	99	1.9	81%	+0.48	-
Zeke (Cohort 3)	Companion	Empathetic support / engagement variant	33	1.7	55%	+0.46	-
Zeke (Cohort 1)	Companion	Empathetic support / relational consistency	272	3.1	63%	+0.32	-
Echo	Companion	Empathetic support	99	3.4	57%	+0.20	-
Diego	Companion	Empathetic support	79	2.8	72%	+0.28	-
Maxine	Companion	Empathetic support	95	2.1	83%	+0.16	-
Reggie	Companion	Empathetic support	76	2.3	86%	+0.32	-
Craig	Task	Core workflow utility execution	38	1.8	97%	+0.04	72%
Maria-SXSW	Task	Localized event logistics and routing	43	2.1	79%	+0.32	84%
Carrie	Task	Customer experience / storefront funnel	33	4.1	76%	+0.20	84%
Grace	Task	Utility workflow assistant	36	3.2	83%	+0.00	62%
Dave	Task	Technical assistant / onboarding	28	3.8	46%*	+0.05	86%

Note: *The lower graceful exit rate observed in the 'Dave' onboarding cohort was tied to a post-task webhook routing disconnect on some calls, rather than an interactive execution failure, meaning users achieved the operational objective but experienced a technical session drop at the finish line. Dave as a demo was requested and created in one day, and had a bug that was identified and fixed.*

Persona	Type	Operational Goal	Real Convos	Median Min.	Graceful Exit	Avg. Sentiment Lift	Task Goal
Aldric	Companion baseline	Un-orchestrated comparison	406	3.1	23%	+0.36	-
Fred	Companion baseline	Un-orchestrated comparison	217	2.0	3%	+0.08	-
Carlos	Companion baseline	Un-orchestrated comparison	99	2.6	83%	+0.32	-
Nick	Companion baseline	Un-orchestrated comparison	46	3.5	83%	+0.12	-
Ronnie	Task baseline	Product-catalog utility	145	5.2	1%	+0.18	36%
Becky	Task baseline	Un-orchestrated comparison	61	5.0	33%	+0.44	88%
Denise	Task baseline	Un-orchestrated comparison	79	5.5	4%	+0.16	84%