PhD Proposal: Information-Theoretic Confidence and Degradation Detection for LLM Agents

David Florez Fernandez
College of Computing & Engineering
Nova Southeastern University

2026

1 Problem Statement

Large language model (LLM) agents are increasingly deployed in high-stakes domains—medical diagnosis, legal analysis, financial advising—where reliability is paramount. Yet these agents suffer from two critical failure modes that undermine trust:

Degradation over use: As conversations lengthen, context windows fill, and attention mechanisms dilute signal, agent performance degrades well before hard limits are reached. This "context rot" is empirically observed across all frontier models but lacks principled detection mechanisms.

Catastrophic forgetting: When agents are fine-tuned on new tasks, they lose previously acquired capabilities. Current detection requires ground-truth evaluation on old tasks, which is expensive and often impossible in production.

Confidence misalignment: Even when agents are correct, their confidence is poorly calibrated—often overconfident when wrong and underconfident when right. Existing confidence measures are static (per-answer) and fail to account for the dynamic degradation processes above.

Research Question: Can information theory provide a unified framework for real-time degradation detection, dynamic confidence estimation, and trustworthy agent communication?

2 Related Work and Gaps

2.1 Information-Theoretic Confidence

Semantic Entropy computes entropy over *meanings* rather than token sequences, detecting "confabulations" (arbitrary incorrectness) with AUROC 0.79. However, it is static, single-turn, and does not track degradation over time.

Stepwise Informativeness Assumption (SIA) proves that reasoning chains accumulate answer-relevant information in expectation, making conditional answer entropy a "progress variable." Yet SIA is defined for single reasoning chains, not multi-turn conversations, and provides no mechanism for agent-user communication.

2.2 The Gap

No existing work unifies these threads into a **dynamic, information-theoretic confidence measure** that:

1. Detects degradation in real-time using entropy dynamics

2. Computes a principled confidence score that accounts for degradation state
3. Communicates this score to users through an agent interface
4. Validates that the communication improves trust and decision-making

3 Proposed Research

3.1 Core Hypotheses

H1: Information-theoretic measures—entropy dynamics, mutual information, and KL divergence—can detect LLM degradation in real-time and compute a dynamic confidence score that predicts actual error rates better than static measures.

H2: Agents that communicate dynamic confidence scores improve user trust and decision-making compared to agents that do not communicate confidence or communicate static confidence.

3.2 Theoretical Framework

We propose a **Unified Information-Theoretic Confidence (UITC)** framework with three components:

3.2.1 Component 1: Multi-Turn Stepwise Informativeness (MT-SIA)

Extend SIA from single reasoning chains to multi-turn conversations. Let $\mathcal{H}_{\leq t}$ denote conversation history through turn t. Define:

$$\text{MT-SIA:} \quad \mathbb{I}(A; \mathcal{H}_{\leq t} \mid Q) \geq \varepsilon_t > 0 \quad \text{for all } t \geq 1 \tag{1}$$

Violation of MT-SIA ($\mathbb{I} \leq 0$) signals degradation. We derive:

- **Turn-level information gain:**

$$\text{IG}_t = \mathbb{I}(A; H_t \mid Q, \mathcal{H}_{<t}) \tag{2}$$

- **Cumulative information:**

$$\text{CI}_t = \sum_{\tau=1}^{t} \text{IG}_\tau = \mathbb{I}(A; \mathcal{H}_{\leq t} \mid Q) \tag{3}$$

- **Degradation signal:**

$$D_t = \max(0, -\text{IG}_t) \tag{4}$$

  positive when new turns are misinformative.

3.2.2 Component 2: Dynamic Semantic Confidence (DSC)

Combine semantic entropy with MT-SIA to produce a per-turn confidence score:

$$\text{DSC}_t = \exp(-\text{SE}_t) \cdot \sigma(\lambda \cdot \text{CI}_t - \mu \cdot D_t) \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function, and:

- SE_t = semantic entropy of turn t's answer (low = confident in meaning)
- CI_t = cumulative information gain (high = accumulating useful info)
- D_t = degradation signal (high = losing signal)
- λ, μ = learned weights

Properties:

- High DSC_t = low semantic uncertainty + high cumulative info + low degradation
- DSC_t decreases automatically as degradation accumulates
- DSC_t is interpretable: each component has information-theoretic meaning

3.2.3 Component 3: Agent Communication Protocol (ACP)

Design how agents surface DSC_t to users. Three modes:

1. **Implicit:** Agent behavior changes (asks clarifying questions, suggests external verification) when $\text{DSC}_t < \theta$
2. **Explicit verbal:** Agent states confidence: "I'm 70% confident in this answer"
3. **Explicit visual:** Dashboard showing entropy trajectory, information gain curve, and degradation alerts

Research question: Which mode maximizes appropriate trust (not over-trust or under-trust)?

3.3 Research Plan

3.3.1 Year 1: Theory and Algorithms

Task 1.1: Formalize MT-SIA and prove degradation detection guarantees.

- Extend Theorem 1 from SIA to multi-turn setting
- Prove: MT-SIA violation $\Rightarrow$ error bound increases
- Derive optimal thresholds for degradation alerts

Task 1.2: Implement DSC computation pipeline.

- Compute semantic entropy via bidirectional entailment clustering
- Estimate mutual information via Monte Carlo sampling
- Build real-time dashboard for entropy dynamics

Task 1.3: Evaluate on conversation benchmarks.

- Datasets: Multi-turn QA (CoQA extended), task-oriented dialogue (MultiWOZ), medical consultation transcripts
- Metrics: AUROC for correctness prediction, calibration error (ECE), correlation with human judgments

3.3.2 Year 2: Catastrophic Forgetting and Agent Communication

Task 2.1: Extend DSC to detect catastrophic forgetting.

- Measure $\mathbb{I}(A; \mathcal{H}_{\leq t} \mid Q)$ for pre-training tasks post-fine-tuning
- Detect forgetting via information decay: $\mathrm{CI}_t^{\mathrm{old}} < \mathrm{CI}_{t-1}^{\mathrm{old}}$
- Validate on standard catastrophic forgetting benchmarks

Task 2.2: Design and implement ACP.

- Build agent interface with three communication modes
- Conduct user studies ($n = 100+$) measuring trust, task performance, and appropriate reliance
- Research question: Does dynamic confidence reduce over-reliance on incorrect answers?

Task 2.3: Publish theory (MT-SIA proofs) and system (agent interface).

3.3.3 Year 3: Integration and Evaluation

Task 3.1: Deploy in realistic setting.

- Partner with domain experts (e.g., legal tech, healthcare triage, education tutoring)
- A/B test: agent with DSC+ACP vs. baseline agent
- Measure: task accuracy, user satisfaction, escalation rates to human experts

Task 3.2: Dissertation writing and defense.

4 Preliminary Results

We have conducted preliminary analyses of the foundational papers:

- **Semantic Entropy:** Reproduced clustering algorithm; confirmed AUROC 0.79 on TriviaQA subset.
- **SIA:** Verified that conditional answer entropy decreases along correct reasoning chains on GSM8K ($n = 500$).
- **Gap confirmation:** No existing method combines both frameworks or extends to multi-turn.

5 Expected Contributions

1. **Theoretical:** First unified information-theoretic framework linking degradation dynamics to confidence
2. **Algorithmic:** MT-SIA and DSC—computable, interpretable, real-time measures
3. **Empirical:** Validation that dynamic confidence predicts error rates better than static measures
4. **Human-centered:** Agent communication protocol that improves appropriate trust
5. **Application:** Deployed system demonstrating real-world utility

6 Broader Impact

Trustworthy AI: By giving agents the ability to communicate their own limitations, we reduce harm from over-reliance on incorrect outputs.

Accessible expertise: In domains with scarce human experts (rare disease diagnosis, legal aid for low-income populations), agents that know when to defer to humans can extend access without sacrificing safety.

Methodological: The information-theoretic tools developed (MT-SIA, DSC) may generalize to other sequential decision systems beyond language models.

7 Risks and Mitigation

Risk	Mitigation
Computational cost of MI estimation	Use variational approximations; leverage efficient NLI models
Clustering errors in semantic entropy	Validate on human-labeled equivalence; use ensemble of NLI models
Users ignore or misinterpret confidence	Design user studies early; iterate on communication format
Overlap with calibration literature	Focus on *dynamic* + *degradation-aware* aspects; cite extensively
Domain shift between benchmarks and deployment	Partner early with domain experts; test on realistic data

References

[1] Chroma Research. "Context Rot: Why LLMs Degrade as Conversation Grows." 2025.

[2] Barari et al. "Information-Theoretic Learning Preserves Prior Knowledge." arXiv:2510.23756, 2025.

[3] Guo et al. "On Calibration of Modern Neural Networks." ICML 2017.

[4] Kuhn, Gal, Farquhar. "Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation." ICLR 2023 / Nature 2024.

[5] Gonzàlez I Català et al. "The Stepwise Informativeness Assumption: Why are Entropy Dynamics and Reasoning Correlated in LLMs?" arXiv:2604.06192, 2026.

[6] Farquhar et al. "Detecting Hallucinations in Large Language Models Using Semantic Entropy." Nature 630:625–630, 2024.

[7] Zhang et al. "Confidence Estimation in Multi-Turn Conversations." arXiv:2601.02179, 2026.

[8] Xuan et al. "Calibration Agentic RL (CAR)." arXiv:2601.07264, 2026.

[9] Lei et al. "Information Bottleneck for Reasoning." arXiv:2507.18391, 2025.

[10] Tishby, Pereira, Bialek. "The Information Bottleneck Method." 2000.