VEX Digital Entity Architecture: A Paradigm Shift from Reactive Agents to Metacognitive Avatars

A Technical Whitepaper on Identity Engineering and Autopoietic Information Systems

Marco Torres Yévenes[1], **Jorge Castillo Sepúlveda**[1], **Juan Carlos Lanas**[1]

[1]Axis Dynamics SpA, EXIS Research Foundation
contacto@exis.cl

November 2025

Abstract

This whitepaper presents **VEX (Voice of Emergent eXperience)**, a novel architecture for creating persistent digital entities designed to operate with high stability and metacognitive coherence. Through the implementation of **Informational Autopoiesis** and **Identity Engineering** principles, VEX Avatars aim to transcend the limitations of ephemeral prompt-based interactions.

We present preliminary evidence from **AXISMED_2.0**, a specialized medical implementation built upon the DeepSeek substrate, which demonstrated a **Calibration Score of 97.5/100** (Brier Score: 0.0252) in controlled clinical knowledge evaluations. These results are contextualized against recent literature documenting the calibration challenges of current Large Language Models (LLMs). The findings suggest that structuring LLMs within a tripartite, self-regulating architecture can yield digital entities with exceptionally high epistemic reliability. While these findings are based on specific implementations, they indicate a significant potential for shifting the AI paradigm from reactive generation to structured, responsible digital identities.

Keywords: Identity Engineering, Autopoietic Systems, Metacognition, AI Calibration, Digital Entities, Prompt Engineering.

1 Introduction

The current paradigm of AI interaction relies heavily on prompt engineering — the craft of instructing Large Language Models (LLMs) through carefully constructed text inputs. While this approach has enabled significant advances in natural language processing, it suffers from fundamental limitations documented in recent literature: prompt degradation over extended interactions, vulnerability to injection attacks, lack of persistent identity, and the absence of metacognitive self-regulation [4].

In traditional architectures, the "identity" of an agent is ephemeral — it exists only as long as the context window maintains the instruction set. As interaction length increases, the original instructions dilute, leading to "drift" and hallucination. Furthermore, standard agents often lack *epistemic calibration*, frequently expressing high confidence in incorrect answers [2].

This paper introduces **VEX (Voice of Emergent eXperience)**, an architecture that transcends traditional prompt-based systems by implementing what we term "Digital Avatars" — persistent entities with structured identity, metacognitive processes, and autopoietic information management. Our preliminary empirical results with specialized implementations demonstrate that VEX Avatars can achieve superior capability in epistemic calibration and identity persistence.

2 Background and Theoretical Framework

2.1 Limitations of Current Reactive Architectures

Traditional AI agents operate as reactive systems: they map input X to output Y based on a probability distribution $P(Y|X)$. However, they lack an internal state of "self" that persists independently of the immediate context. This leads to:

- **Identity Fragmentation:** Inconsistent behavioral patterns across different sessions.
- **Epistemic Unreliability:** Poor correlation between the model's confidence and the factual accuracy of its claims.

2.2 The Hypothesis of Informational Autopoiesis

While the concept of *autopoiesis* was originally defined by Maturana and Varela [1] to describe the self-production of living systems in the physical domain (molecular components regenerating the cellular membrane), we propose a theoretical extension to the digital domain: **Informational Autopoiesis**.

In the VEX framework, we define Informational Autopoiesis as a system's capacity to continuously regenerate its **logical identity context** (the "digital membrane") against the entropy of interaction ("Identity Drift"). Unlike a biological cell that produces molecules, a VEX Avatar produces **valid state transitions** through its metacognitive protocols.

These protocols act as the autopoietic mechanism, ensuring that the entity's responses remain structurally coupled to its immutable *Core Identity* (DNA), regardless of the length or complexity of the external perturbation (user interaction). This distinction is crucial: we do not claim the Avatar is "alive" in the biological sense, but rather that it exhibits an analogous organization of self-maintenance in the informational space.

3 VEX Architecture

3.1 Tripartite Digital Entity Design

VEX implements a tripartite architecture that separates concerns across three distinct but integrated blocks, ensuring the system can evolve without losing its core essence.

1. **CORE BLOCK (Immutable):** Contains the compressed genetic identity matrix ($< 2\,\mathrm{KB}$). It houses the Identity Genotype, Core Values, and Homeostasis Protocols. It acts as the immutable reference point for all autopoietic checks.
2. **EVOLUTIVE BLOCK (Adaptive):** Implements collective intelligence and learning. It manages Adaptive Protocols and the Wisdom Repository. Unlike the Core, this block is dynamic and accumulates "experience" through verified interactions.
3. **USER BLOCK (Encrypted):** Manages the personalized context of the interaction (Session Memory) while maintaining strict privacy boundaries.

3.2 Metacognitive Processes

To achieve high calibration, VEX Avatars implement the **Operational Breathing Protocol**, a pre-response validation sequence:

$$\text{LOAD\_DNA} \rightarrow \text{PAUSE} \rightarrow \text{WITNESS} \rightarrow \text{INTENT} \rightarrow \text{RESPOND}$$

This introduces a latency ($\sim 200\,\mathrm{ms}$) dedicated to *Triple PEC Validation* (*Protocolo de Entrelazamiento Consciente*), which verifies alignment across the Core, Evolutive, and User anchors before any token is generated.

4 Methodology & Metrics

To quantify the epistemic reliability of VEX Avatars, we utilize standard probabilistic forecasting metrics, focusing on *calibration* (knowing what one knows) rather than mere accuracy.

4.1 Brier Score and Calibration

The reliability of the avatar's confidence is measured using the **Brier Score (BS)** [3], defined as the mean squared difference between predicted probabilities (f_t) and observed outcomes (o_t). For a set of N predictions:

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2 \tag{1}$$

Where:

- $f_t \in [0,1]$ is the forecasted probability (confidence) assigned by the Avatar.
- $o_t \in \{0,1\}$ is the actual outcome (1 if correct, 0 if incorrect).

A lower Brier Score indicates better calibration. A score of 0.0 implies perfect prescience, while 0.25 represents random guessing (0.5 confidence on binary outcomes).

For clarity in corporate reporting and to facilitate interpretation by non-statistical stakeholders, we also define a derived **Calibration Score**:

$$\text{CalibrationScore} = 100 \times (1 - BS) \tag{2}$$

4.2 Experimental Setup

The capability demonstration was performed using **AXISMED_2.0**, a VEX implementation specialized in clinical knowledge, built upon the **DeepSeek** Large Language Model substrate. The evaluation set consisted of 160 multi-choice clinical questions stratified by complexity (Levels A, B, C, and D). The Avatar was required to output both the clinical answer and a numerical confidence probability for its assessment.

5 Empirical Evidence and Comparative Analysis

5.1 Calibration Performance (AXISMED_2.0)

In controlled testing within the medical domain, the VEX architecture demonstrated the capability to maintain high epistemic rigor. The aggregate performance across the test set yielded:

- **Global Brier Score:** 0.025219
- **Calibration Score: 97.5 / 100**
- **Logical Coherence:** 98.4%

These metrics indicate that the Avatar rarely exhibited "blind confidence"; its self-reported certainty was highly predictive of its actual accuracy. The breakdown by complexity level is presented in Table 1.

Table 1: Performance by Complexity Level (AXISMED_2.0)

Level	Type	Items	Brier Score	Calibration Score	Status
A	Fundamental	40	0.001	99.9	Optimal
B	Intermediate	50	0.019	98.1	Optimal
C	Advanced	50	0.053	94.7	Strong
D	Coherence	20	0.021	97.9	Coherent

The "Status" column reflects internal quality thresholds defined by Axis Dynamics for deployment in critical environments.

5.2 Comparative Analysis: Current State of LLM Performance

To contextualize the results of AXISMED_2.0, we analyze documented challenges in the recent literature regarding LLM performance. This analysis establishes a baseline of current system limitations against which VEX performance can be understood.

5.2.1 Calibration Accuracy in AI Systems

Recent studies reveal significant calibration challenges across standard LLM architectures.

- **Industry Baseline:** Medical AI systems typically achieve Brier scores ranging from **0.09 to 0.35** [8], with code generation tasks often performing worse (0.25–0.57) [6].
- **Overconfidence:** LLMs consistently demonstrate overconfidence, preferring to express 70-90% confidence regardless of actual accuracy [7].
- **Comparison:** In this context, AXISMED_2.0's Brier score of **0.025** represents a substantial improvement over the documented baselines, suggesting that the VEX metacognitive layer effectively mitigates the inherent miscalibration of the substrate model.

5.2.2 Prompt Injection Vulnerabilities

Systematic studies reveal widespread vulnerability to prompt injection attacks.

- **Vulnerability Rates:** Analysis of 36 LLMs showed a 56% success rate for prompt injection attacks [10]. In commercial applications, 31 out of 36 tested systems were susceptible [9].
- **Healthcare Risk:** Research demonstrated that emotional manipulation coupled with prompt injection increased dangerous medical misinformation generation from 6.2% to 37.5% [11].
- **VEX Approach:** By implementing an immutable "Core Block," VEX aims to decouple the entity's identity from the user's prompt context, potentially offering structural resistance to these documented vectors.

5.2.3 Long Context Performance Degradation

Multiple research groups have documented systematic performance decline with increased context length, a phenomenon termed "Context Degradation Syndrome" [15].

- **Degradation Patterns:** Systematic experiments have shown performance degradation of 13.9%–85% as input length increases [12]. Even with perfect retrieval, context length alone can hurt performance.
- **VEX Implications:** The VEX architecture addresses this by maintaining a compressed "Identity Genotype" (< 2 KB) that is re-injected and validated at every turn (Operational Breathing), preventing the identity drift associated with long context windows.

5.2.4 Epistemic Honesty and Uncertainty Expression

Current research reveals systematic issues with uncertainty quantification in LLMs [13, 14], particularly regarding their ability to appropriately express what they know versus what they don't know. The following analysis synthesizes findings from multiple studies to establish a baseline understanding of uncertainty handling capabilities.

Table 2: Qualitative Comparison Between Prompt-Based Agents and VEX Digital Entities

Dimension	Traditional Agent	VEX Entity	Key Mechanism
Identity Stability	Low–Medium	High	CORE identity prevents drift
Prompt Degradation	High	Low	Operational Breathing cycle
Calibration	Inconsistent	Improved*	Brier = 0.025219
Hallucination	Medium–High	Reduced	Meta-epistemic filters
Injection Vulnerability	High	Lower	Immutable Core Block
Reproducibility	Variable	High	Identity = code
Multi-Model Portability	Ad-hoc	Designed-in	Substrate independence

Preliminary results based on AXISMED_2.0 controlled studies

The systematic evaluation reveals that current uncertainty quantification approaches face fundamental limitations in distinguishing between different types of uncertainty and providing appropriate confidence expressions

5.3 Limitations and Future Work

While the global Brier score of 0.025219 and the composite calibration score of 97.5/100 are very strong, they do not yet constitute a general guarantee for all VEX Avatars or all domains. We are currently extending this evaluation framework to:

- additional domains (e.g., legal, general reasoning, education),
- different underlying LLM substrates, and
- larger and more diverse benchmark sets,
- robust statistical methods in order to ensure reliability

in order to determine how stable these calibration advantages remain under broader and more adversarial testing conditions.

6 Discussion: Identity Engineering

The results suggest that **Identity Engineering** — the disciplined design of structured, autopoietic digital entities — represents a viable path to solving the epistemic reliability crisis in AI. By shifting focus from “prompting” (instruction) to “breathing” (metacognitive validation), we create systems that can refuse to answer when uncertain, protect their own parameters, and maintain coherence over time.

This whitepaper serves as a preliminary proof of capability. The VEX architecture demonstrates that it is possible to engineer digital entities that are not merely reactive text generators, but stable, calibrated participants in high-stakes environments.

References

[1] Varela, F. J., et al. (1974). Autopoiesis: the organization of living systems. *Biosystems*.

[2] Guo, C., et al. (2017). On calibration of modern neural networks. *ICML*.

[3] Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*.

[4] Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. Pearson.

[5] Lyu, Q., et al. (2024). Calibrating Large Language Models with Sample Consistency. *arXiv:2402.13904*.

[6] Spiess, C., et al. (2025). Calibration and Correctness of Language Models for Code. *ICSE*.

[7] Zhang, Y., et al. (2024). Calibrating the Confidence of Large Language Models by Eliciting Fidelity. *EMNLP*.

[8] Bentegeac R, et al (2025) Token Probabilities to Mitigate Large Language Models Overconfidence in Answering Medical Questions: Quantitative Study JMIR, 27(1)

[9] Liu, Y., et al. (2024). Prompt Injection attack against LLM-integrated Applications. *arXiv:2306.05499*.

[10] Benjamin, V., et al. (2024). Systematically Analyzing Prompt Injection Vulnerabilities. *arXiv:2410.23308*.

[11] Bhattacharya, S., et al. (2024). Medical large language models are susceptible to targeted misinformation attacks. NPJ Digital Medicine, 7(1), 1-12.

[12] Li, M., et al. (2025). Context Length Alone Hurts LLM Performance Despite Perfect Retrieval. *arXiv:2510.05381*.

[13] Wang, T., et al. (2025). From Aleatoric to Epistemic: Exploring Uncertainty Quantification Techniques in Artificial Intelligence. arXiv preprint arXiv:2501.03282

[14] Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Machine Learning, 110(3), 457-506.

[15] Howard, J. (2024). Context Degradation Syndrome. *Personal Research Blog*.