svmp

SVMP Systems v3.0

Addressing the "Reliability Gap" through event-driven soft debouncing and semantic similarity gating.

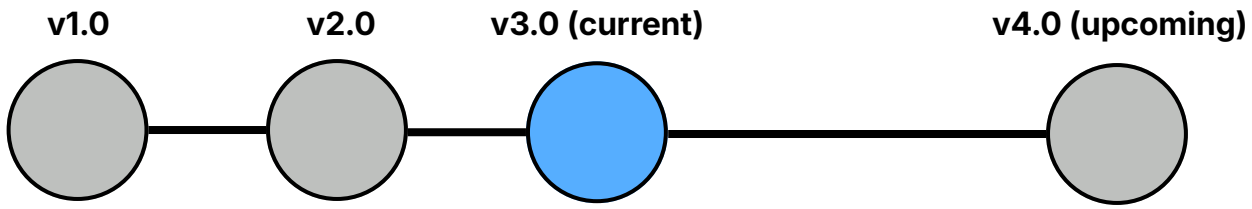
Version: 3.0 (Pilot Baseline) | Release Date: January 15, 2026

01- Executive Abstract.

The Problem:
The "Reliability Gap" Generative AI is a world-class conversationalist but a high-risk decision-maker. In high-stakes environments, "mostly accurate" isn't good enough. We identified a critical Reliability Gap: the point where an LLM's desire to be helpful overrides its requirement to be factual.

The Solution:
The SVMP Framework. We didn't build another chatbot; we built a Governance Layer. The Semantic Vector Mapping Protocol (SVMP) is an architectural "hard gate" that forces AI to stay grounded. By trading 100% automation for a 0% hallucination rate, we ensure that if the system isn't mathematically certain, it stays silent.

Our Methodology & Innovations:
Stealth-Phase R&D Moving beyond basic LLM wrappers. Version 3.0 represents our stable architectural baseline. It is currently surviving a 3,000+ sample adversarial stress-test designed to break it. We aren't just claiming reliability; we are empirically proving it.
At its core, SVMP uses an n8n-orchestrated engine to bridge multi dimensional MongoDB vector spaces with real-time APIs. Through the implementation of a Semantic Similarity matching gate, we ensure that unless a response mathematically aligns with verified internal data, it never reaches the end-user.



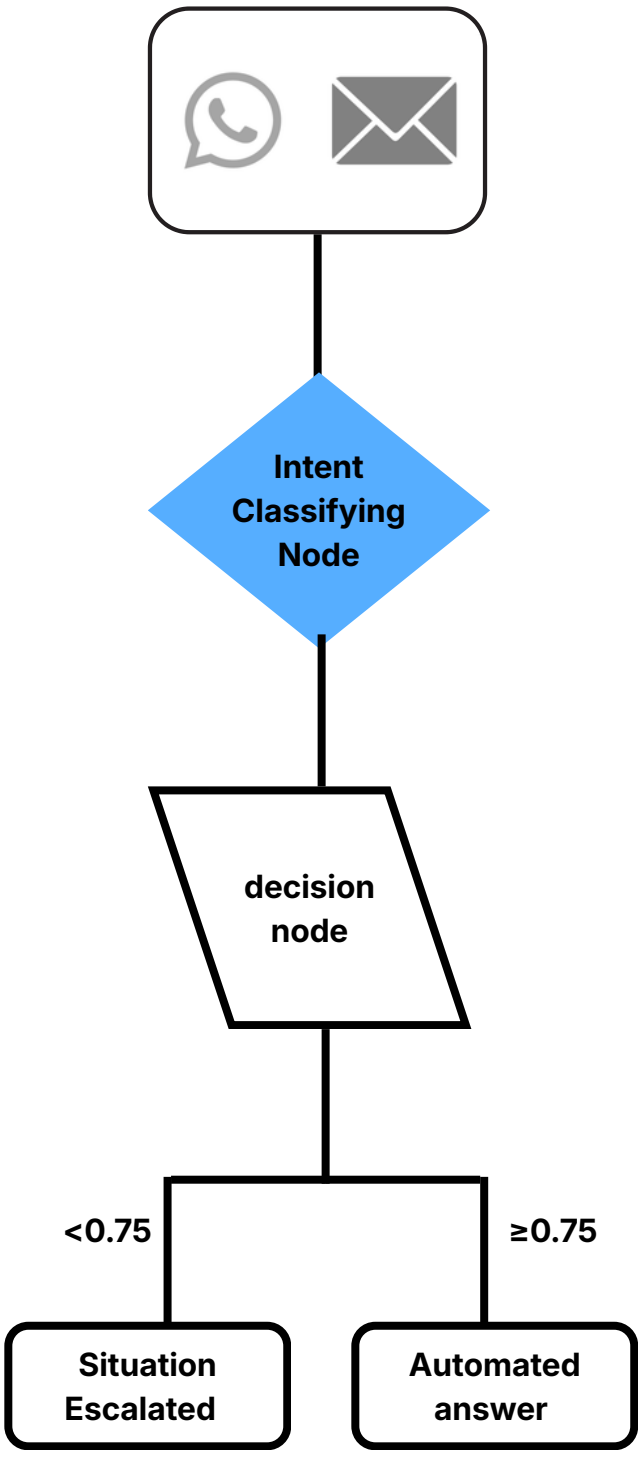
© 2026 SVMP

02- Current Architecture

The Core Orchestrator (n8n)

The system utilizes n8n as a deterministic backbone to manage data flow from the user to the LLM and back. By using a node-based architecture, we maintain 100% observability over the decision-making process.

- Intent Triage: Incoming queries are captured via persistent webhooks and routed by an Intent Classifier.



The Governance Gate (≥0.75)

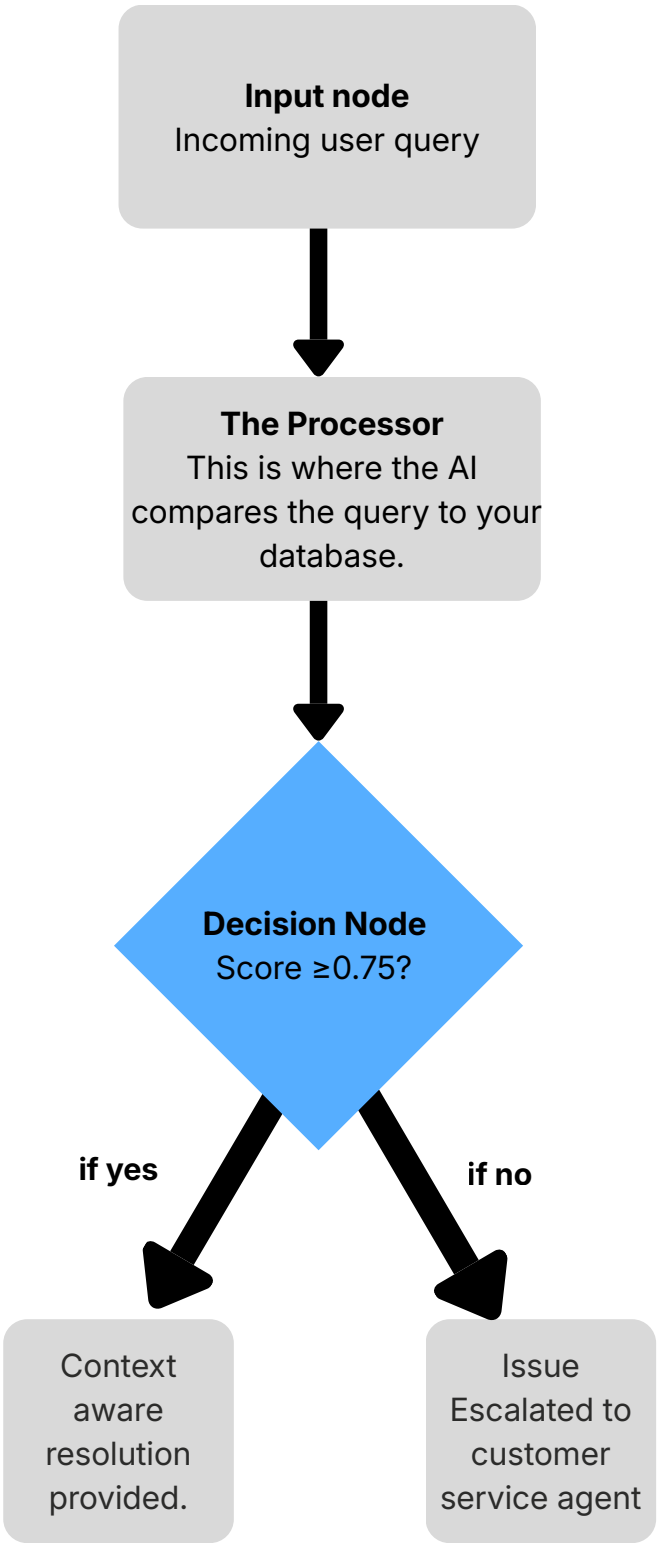
To eliminate hallucinations, every query is converted into a mathematical vector and compared against a verified knowledge base in MongoDB Atlas.

- The Decision Node: If the Semantic Similarity Score is ≥0.75, the system retrieves the factual match and synthesizes a response.
- Escalation Path: If the score falls below the threshold, the automation "freezes" and triggers a human handoff via WhatsApp or Slack.

In SVMP v3.0, we treat the LLM as a modular engine, not a decision-maker. The system is designed to prioritize escalation over error; by enforcing a hard ≥0.75 similarity gate, we ensure the automation remains grounded in verified data clusters. This architecture guarantees that every response is a product of mathematical validation, not probabilistic guessing.

© 2026 SVMP

03- The Governance Gate



In Version 1, We realized that LLMs are 'people pleasers'—they will guess an answer even if they aren't 100% sure.
To solve this, We built a Governance Gate.

Instead of letting the AI speak freely,

- **Every query is now converted into a mathematical vector and compared against a verified 'Knowledge Base'** in MongoDB.
- We calculate a Similarity Score. If the score is 0.75 or higher, the system is confident enough to reply.
- If it's lower than 0.75, the system 'freezes' the automation and pings a human manager on WhatsApp/Slack.

The Result: We traded a 100% automation rate for a 0% hallucination rate. We only speak when we are certain."

© 2026 SVMP

04- Evolutionary Roadmap

v1.0

First working end-to-end LLM response prototype

Issues

- Responses not linked to any FAQ database
- No similarity matching; everything treated equally
- IF logic existed but was meaningless
- Could not escalate reliably

October 2025

November 2025

v2.0

Introduced data-driven answer selection; first step toward deterministic, FAQ-grounded responses.

Changes:

- Added MongoDB node as FAQ storage database.
- Added OpenAI node for semantic similarity scoring
- Added IF node to branch on similarity score (score >= 0.75)
- Escalation path for low similarity

v3.0

Key Achievements

- Removed parsing errors, node data inconsistencies & ambiguity from input formatting.
- Achieved faithful FAQ-based responses, eliminating hallucinations and preserving automation reliability.

Workflow is now in testing phase.

January 2026

(February 2026)

v4.0

(Production-grade system

- Multi Cluster Domain Isolation
- Multi-Tenant Semantic Mapping
- Soft Debounce system
- Work queue Lock
- Deterministic background execution
- Session-level message aggregation

© 2026 SVMP

05- Validation and Testing.

Hallucination Rate

0%

Through the ≥0.75 Governance Gate, we eliminated AI 'guessing,' ensuring 100% factual integrity on critical service data.

Autonomous Resolution

90%

v3 successfully categorized and resolved 9 out of 10 queries without any human intervention, drastically reducing operational overhead.

Testing Phase

3000 +

SVMP v3.0 is currently undergoing a 3,000-point Adversarial Testing Sprint. This red-teaming cycle stresses the ≥0.75 Governance Gate against prompt injection and semantic ambiguity. Preliminary results validate this threshold as the definitive baseline for eliminating LLM hallucinations in most environments.

© 2026 SVMP

06- v4.0 and Future Roadmap

v4.0 engineering roadmap

Pillar 1: Multi-Cluster Domain Isolation

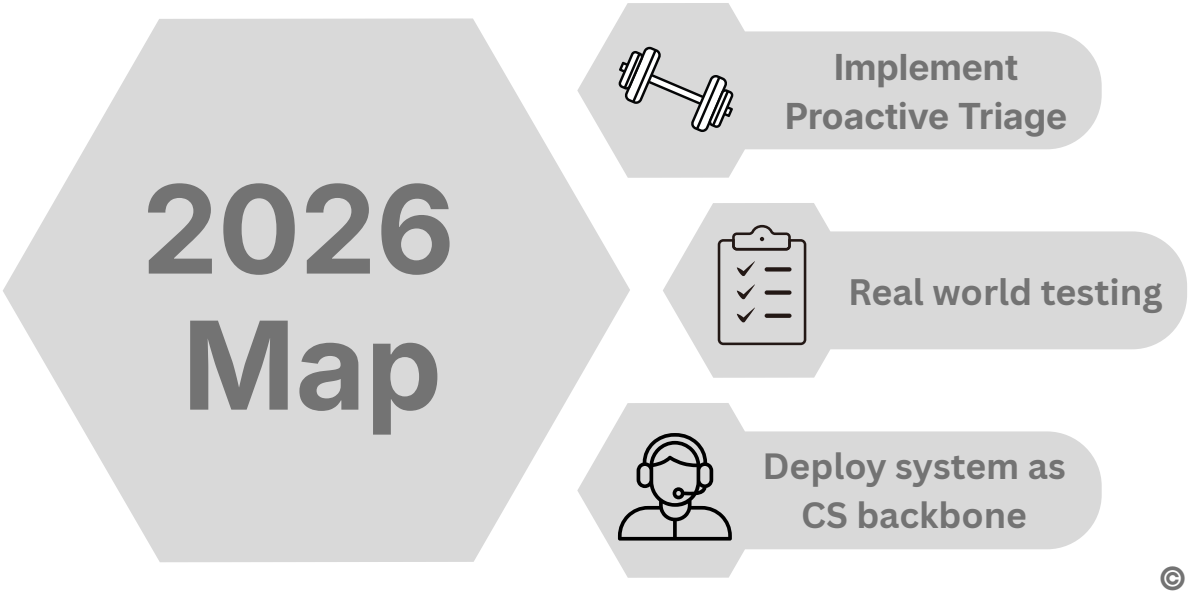
- **Target:** Scalability and precision.
- **Logic:** Transitioning from a singular MongoDB collection to Domain-Specific Clusters (e.g., Customer Support, Provider Logistics, and Marketing). By isolating these datasets into separate clusters within n8n, the system narrows the search vector before querying. This eliminates Cross-Domain Interference, ensuring only relevant data is considered, significantly boosting accuracy and speed as the Knowledge Base expands.

Pillar 2: Dynamic Texting

- **Target:** User resonance and mild dialect adaptation.
- **Logic:** Introduces a post-governance synthesis layer. After the logic engine retrieves a factual match, a secondary pass adapts the dialect and tone to the user's specific query. Factuality remains anchored while the delivery becomes empathetic and fluid.

Pillar 3: Dynamic Intent Aggregation

- **Target:** Multi-burst input resolution (Context Stability).
- **Logic**: Implements a Soft Debounce System that resets its execution window with every new input event. This prevents fragmented processing by waiting for a natural pause in user cadence to aggregate message streams. By capturing "Complete Thought Units" before triggering the governance gate, the system ensures holistic analysis and eliminates race conditions.



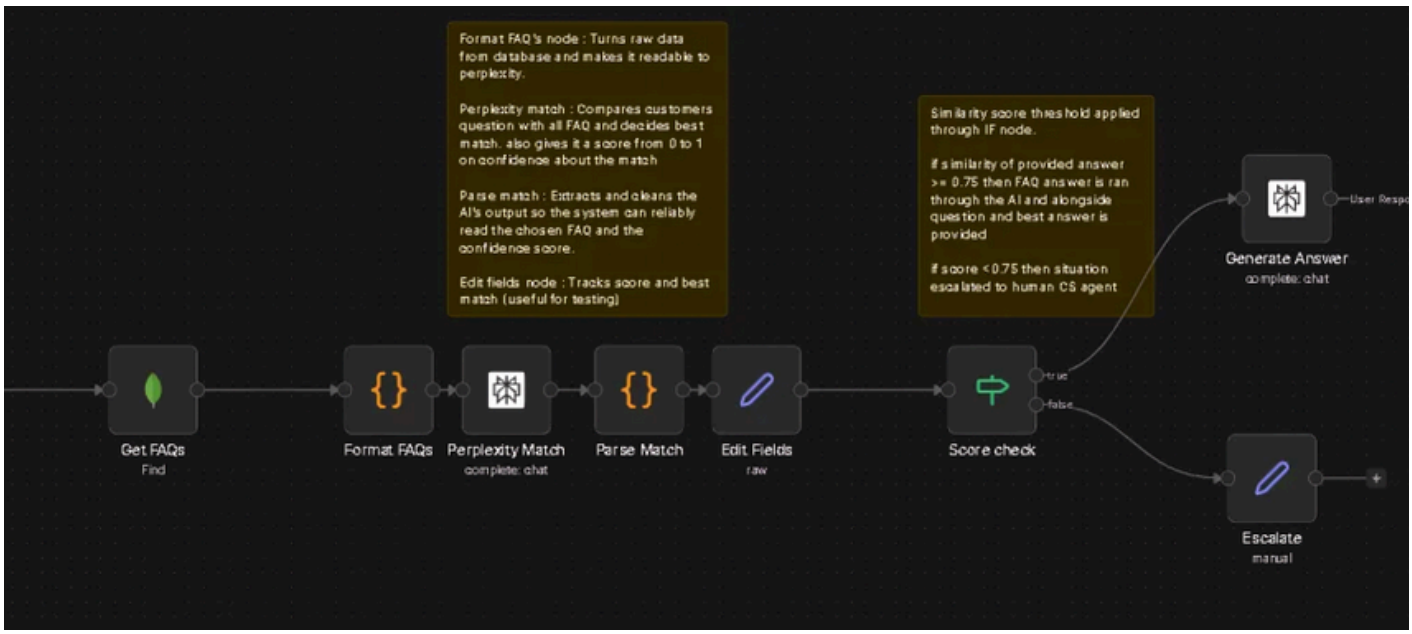
© 2026 SVMP

07- Appendix

Technology is often deployed for efficiency at the cost of accuracy. Our founding trio chose a different path. By focusing on Deterministic AI Governance (the 0.75 gate) and Social Impact (NGO deployment), we have built a system that prioritizes human trust over automated noise. Our goal is not just to build a company, but to establish a new standard for how AI interacts with local communities.

As we move toward May 2026, our primary focus remains the democratization of this technology. By providing our v3.0 framework to non-profit organizations for free, we are stress-testing our logic in the most high-stakes environments possible: where information isn't just data, but a lifeline.

The v3 Intent-Classifier Node Cluster.



The Database Schema

JSON

```
{
  "intent_id": "SVMP_REQ_102",
  "domain": "customer_support",
  "question": "How do I track my active service request?",
  "answer": "You can track your request in real-time via the 'My Bookings' ta
}
```

© 2026 SVMP

The Live Knowledge Base

Scan to view the SVMP Knowledge Base on Notion. This repository contains our live Intent Taxonomy, Schema Definitions, and Version Logs

The SVMP v3.0 framework is a living architecture. We are currently recruiting pilot partners for our upcoming v4.0 stress-testing phase.

Team Credentials

Lead Architect: Pranav H - LinkedIn
Product Lead: Samarth Magi - LinkedIn
Operations Lead: Shravan Kumar - LinkedIn | Portfolio

© 2026 SVMP