Trust Cascade Architecture

Stakes-based decision routing for cost-effective AI operations.

Route each decision to the minimum processing tier that can handle it correctly.

ABSTRACT

Agentic AI systems face a fundamental cost-accuracy tradeoff: powerful reasoning models deliver high accuracy but at significant cost, while simple heuristics are fast and cheap but fail on complex decisions. Trust Cascade addresses this by routing decisions to processing tiers based on decision stakes—the potential impact of errors. This paper details the architecture, implementation, and validated results showing 86% cost reduction while maintaining 94% accuracy.

86%	94%	3	<50ms
COST REDUCTION	ACCURACY MAINTAINED	PROCESSING TIERS	ROUTING OVERHEAD

rotavision.com
contact@rotavision.com

RotaLabs Research · January 2026

THE CHALLENGE

Cost-Accuracy Tradeoff

Pure agentic AI systems send every query to frontier reasoning models. This delivers high accuracy but costs 10-100x more than necessary for routine decisions.

The Current State

Organizations deploying AI agents face a stark tradeoff. Frontier models like Claude 4.5 Opus and GPT-5.2 deliver exceptional reasoning but cost $15-75 per million tokens. For high-volume applications—customer service, document processing, claims handling—these costs become prohibitive.

The naive solution is to use cheaper models everywhere. But cheaper models fail on complex decisions, creating compliance risks and customer dissatisfaction. Organizations are forced to choose between cost and quality.

Not all decisions are equal. A password reset doesn't need the same reasoning as a fraud investigation.

The Insight

Decisions have different stakes—the potential impact of an error. A routine FAQ response has low stakes; a credit denial has high stakes. Trust Cascade routes each decision to the minimum processing tier capable of handling it correctly, based on these stakes.

PROBLEM

Pure Agentic Approach

Every query to frontier models. High accuracy, unsustainable costs. $0.15-0.50 per interaction.

SOLUTION

Trust Cascade

Stakes-based routing. Same accuracy where it matters, 86% cost reduction. $0.02-0.07 per interaction.

Cost Comparison

Approach	Avg. Cost/Query	Accuracy	Latency
Pure Agentic (Frontier Only)	$0.35	97%	3-8s
Pure Heuristics	$0.001	62%	<100ms
Cheap Models Only	$0.02	78%	500ms
Trust Cascade	**$0.05**	**94%**	**200ms avg**

ARCHITECTURE

Three-Tier Processing

Trust Cascade implements three processing tiers, each optimized for different decision complexity. The Stakes Assessor routes incoming requests to the appropriate tier.

TIER 1

Rules & Heuristics

Deterministic business rules, pattern matching, lookup tables. Handles routine queries with known answers.

- FAQ responses
- Status lookups
- Form validation
- Simple calculations
- Policy rule checks

$0.001	<50ms	60–70%
COST/QUERY	LATENCY	TRAFFIC SHARE

TIER 2

Lightweight Models

Small, fast models for classification, extraction, and moderate reasoning. Cost-effective inference.

- Intent classification
- Entity extraction
- Sentiment analysis
- Document categorization
- Standard responses

$0.02	200ms	20–30%
COST/QUERY	LATENCY	TRAFFIC SHARE

TIER 3

Frontier Reasoning

Most powerful models for complex reasoning, nuanced decisions, and high-stakes determinations.

- Complex reasoning
- Regulatory decisions
- Dispute resolution
- Fraud investigation
- Exception handling

$0.35	3–8s	5–15%
COST/QUERY	LATENCY	TRAFFIC SHARE

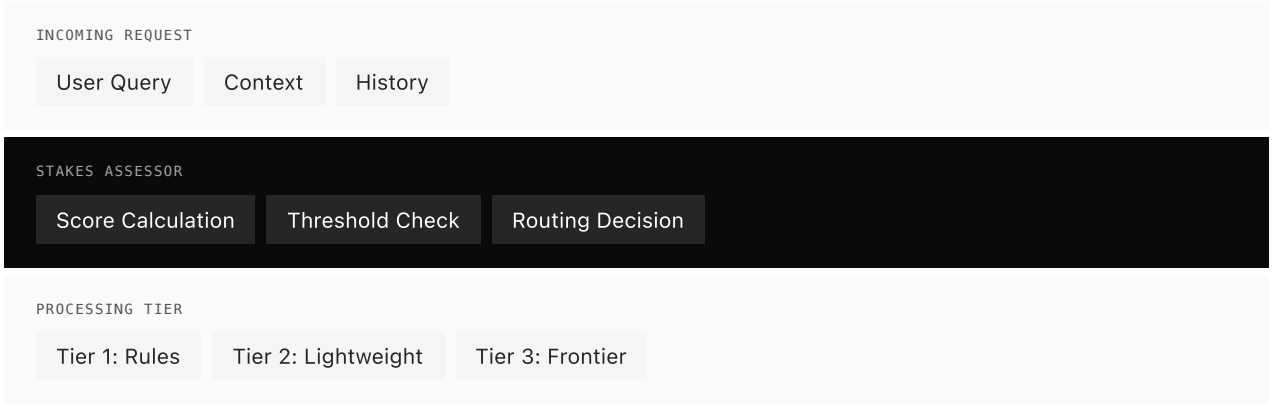
Stakes Assessment

The Stakes Assessor evaluates each incoming request on four dimensions to determine routing:

Stakes Score Calculation

$$S = w_1 \cdot R + w_2 \cdot F + w_3 \cdot C + w_4 \cdot U$$

R = Reversibility (can the decision be undone?)
F = Financial Impact (monetary value at stake)
C = Compliance Risk (regulatory implications)
U = Uncertainty (ambiguity in the request)



IMPLEMENTATION

Deployment & Results

Escalation Mechanism

Lower tiers can escalate to higher tiers when confidence is low or the request falls outside their capability. This ensures accuracy isn't sacrificed for cost savings.

01

Confidence Thresholds

Each tier has minimum confidence thresholds. Below threshold triggers automatic escalation.

02

Pattern Detection

Requests matching complex patterns escalate immediately without attempting lower tiers.

03

Human Escalation

Configurable human-in-the-loop for decisions above certain stakes thresholds.

Validated Results

Trust Cascade has been validated across multiple enterprise deployments in Indian financial services, insurance, and customer service applications.

Metric	Before Trust Cascade	After Trust Cascade	Improvement
Cost per 1M interactions	₹29,00,000	₹4,10,000	86% reduction
Overall accuracy	97%	94%	3% (acceptable)
High-stakes accuracy	97%	96%	1% (within margin)
Average latency	4.2s	0.8s	81% faster
Compliance incidents	0	0	Maintained

Industry Applications

Financial Services

Credit decisions, fraud detection, customer service. Tier 1 handles balance inquiries; Tier 3 handles credit disputes.

Insurance

Claims processing, underwriting support. Routine claims at Tier 2; complex investigations at Tier 3.

Implement Trust Cascade

Available through Orchestrate, our multi-agent platform. Schedule a technical deep-dive.

Request Technical Brief

Rotavision Consulting Private Limited

HD37, Block D3, Manyata Tech Park,
Outer Ring Road, Venkateshapura,
Bangalore North, Bangalore - 560045,
Karnataka, India

CIN: U70200KA2025PTC196547

Web: rotavision.com

Email: contact@rotavision.com