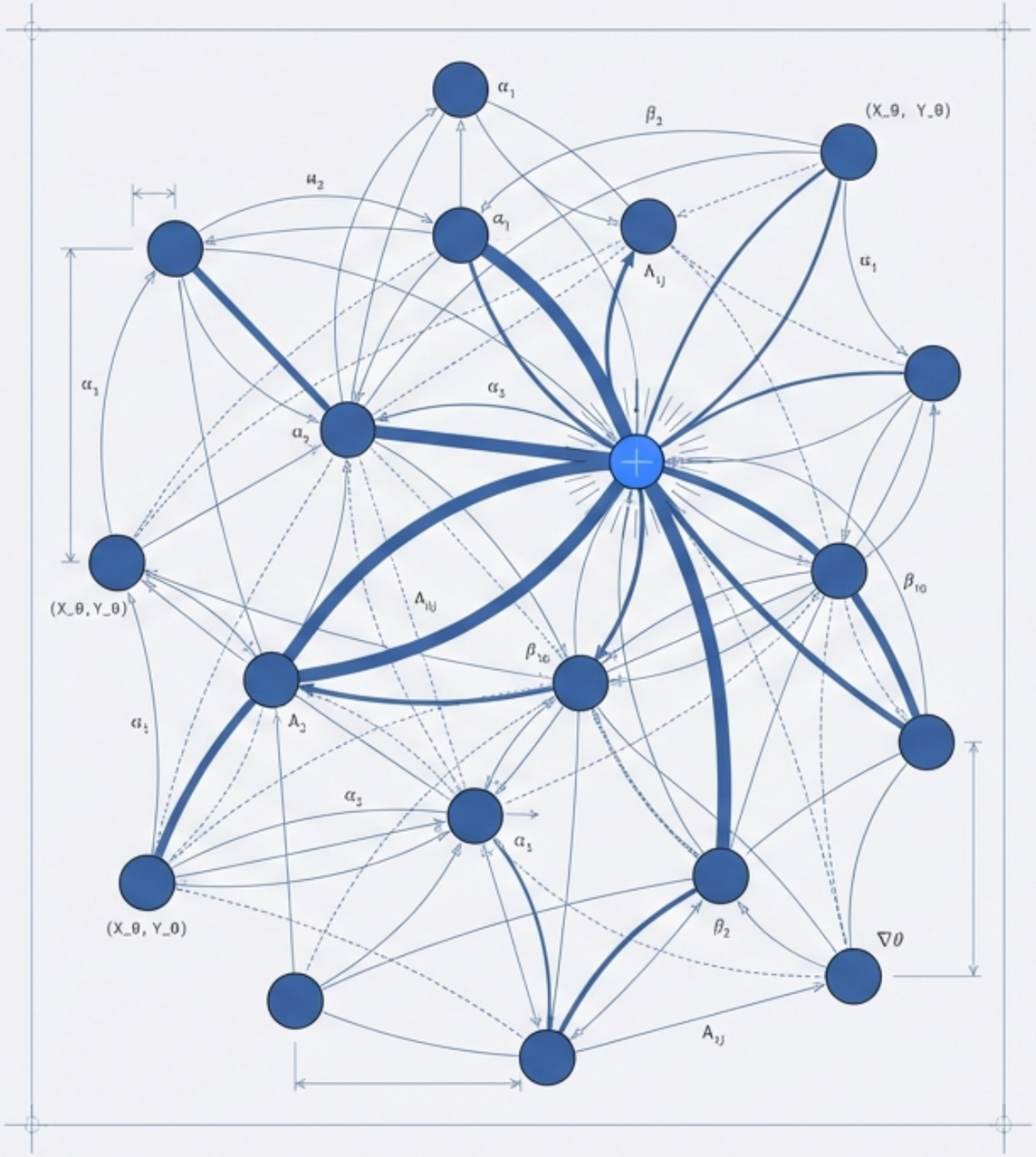
Adaptive Policies for Robust Multi-Agent Systems

Introducing Interpretable Graph-Attention Collaboration (IGAC)

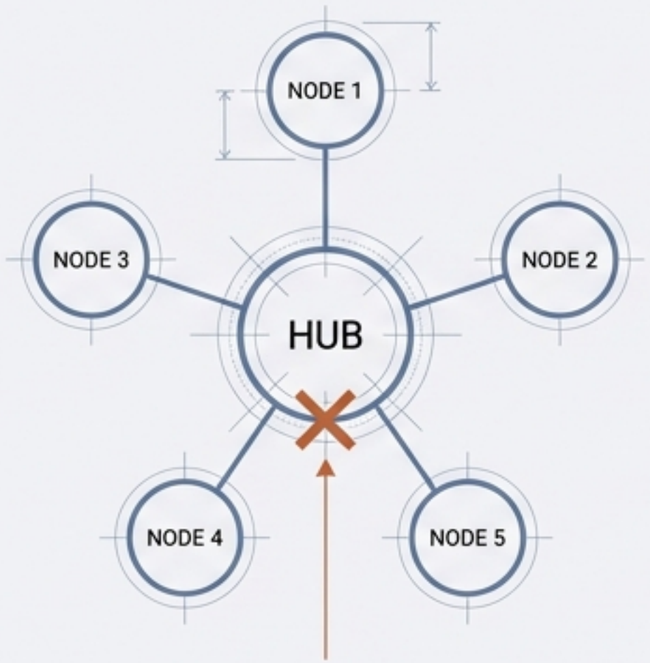
A Technical Reading Deck | Based on Research in Collaborative State Reconstruction



Intelligent Agents, Rigid Communication

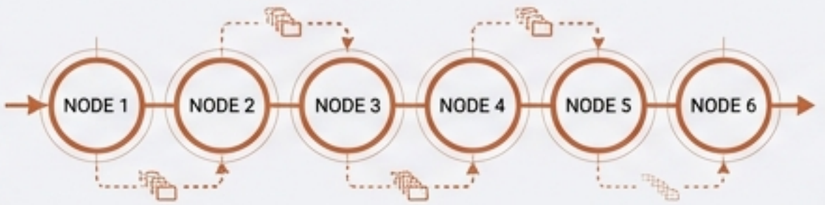
The bottleneck in current Multi-Agent Systems (MAS) is the static topology.

Star Topology



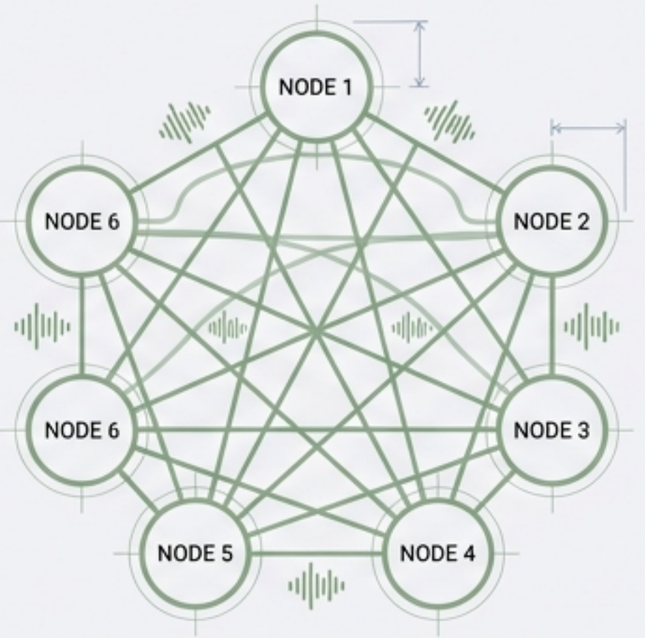
Single Point of Failure.
If the hub fails, the network disconnects.

Chain Topology



Latency & Loss.
Information degrades as it hops sequentially.

Fully Connected

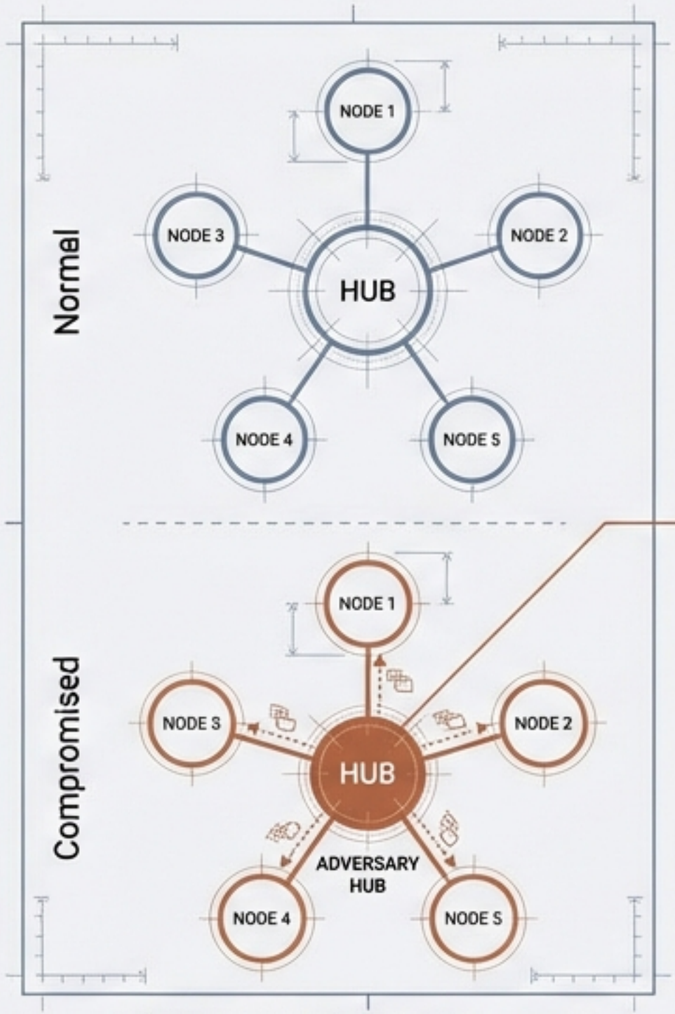


Noise & Inefficiency.
Indiscriminate aggregation creates flooding.

The Fragility of Static Networks

When 20% of agents are adversarial (injecting noise or lies), fixed topologies lack the mechanisms to filter them out. The entire collective reasoning process is poisoned.

"Fixed topologies operate on blind trust."



Baseline Error Rate under attack:

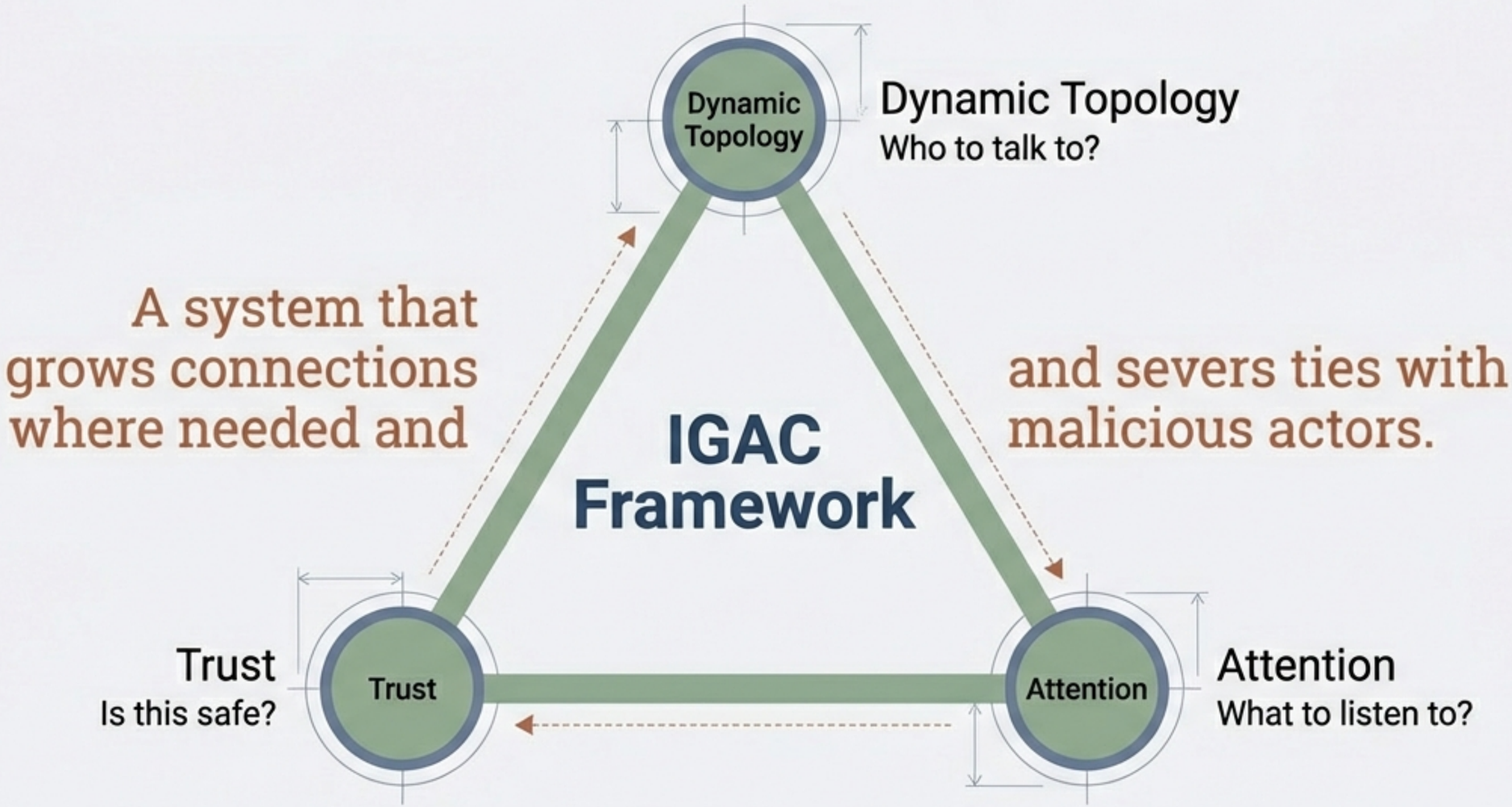
2.224

(Fully Connected)

The Traitor Hub: 100% Corruption.

From Fixed Design to Adaptive Policy

IGAC replaces static wires with dynamic, learned decisions.

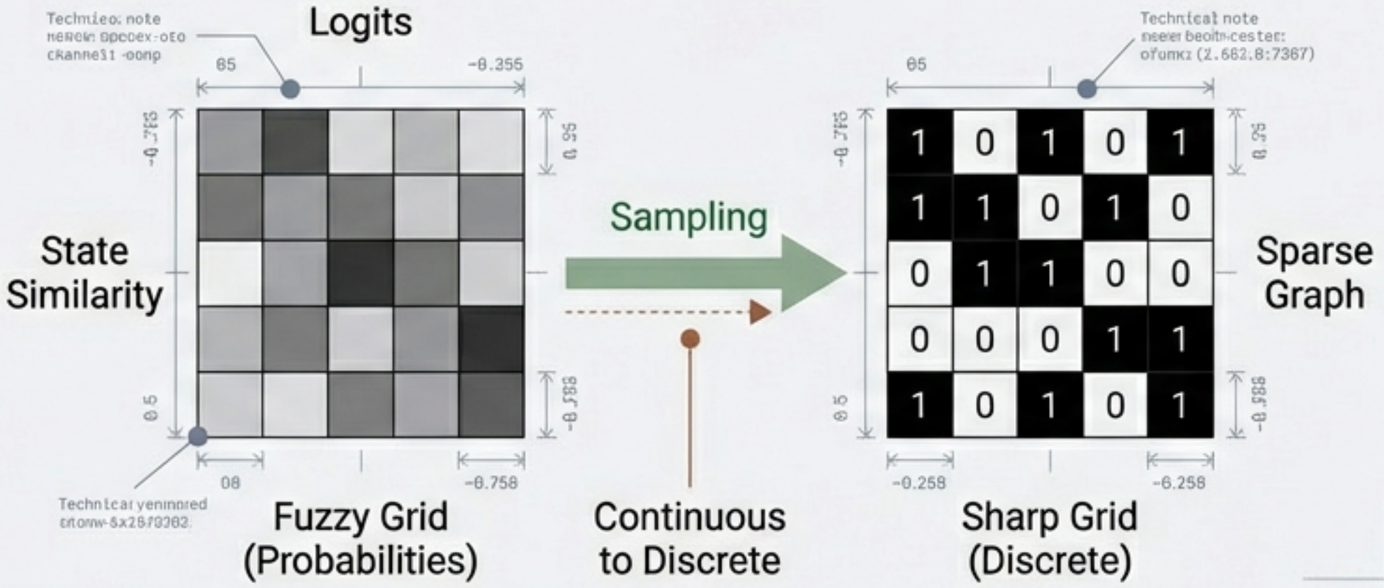


Pillar 1: Learned Topology via Gumbel-Softmax

The "Switchboard."
Instead of a permanent cable, a meta-controller decides connectivity at every step based on state similarity.
Continuous variables are relaxed into discrete choices to keep the graph sparse.

$$l_{ij} = \frac{\boldsymbol{h}_i^T \boldsymbol{h}_j}{\|\boldsymbol{h}_i\| \|\boldsymbol{h}_j\|} + \log\left(\frac{\rho}{1-\rho}\right)$$

$$A_t[\mathrm{i}, \mathrm{j}] = \mathrm{GumbelSoftmax}(l_{ij}, \tau)$$

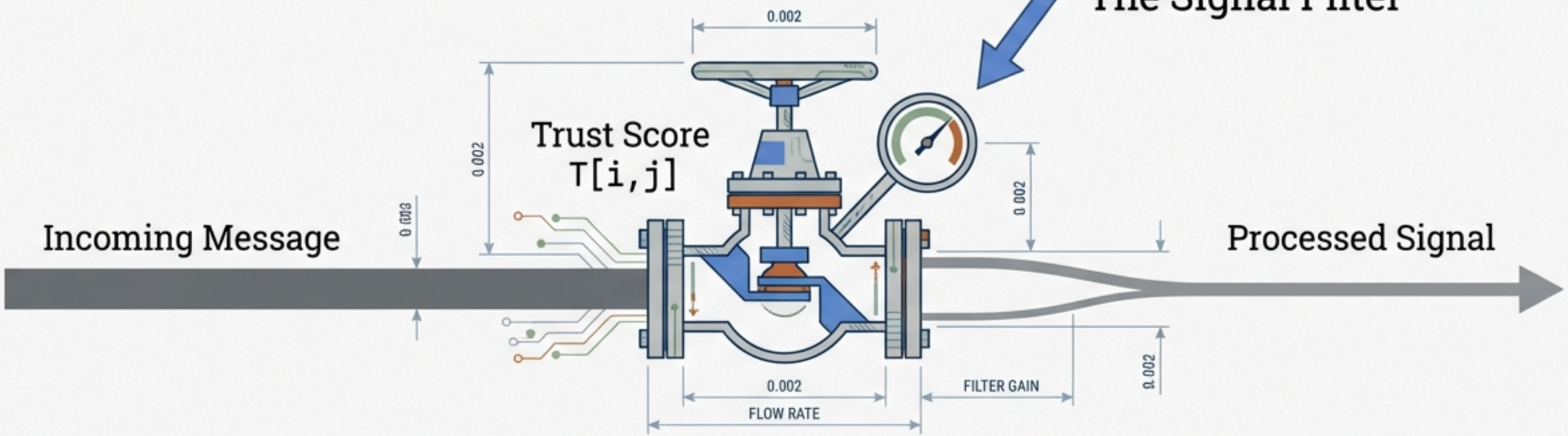


Pillar 2: Trust-Weighted Attention

The “Volume Knob.” Standard attention mechanisms are modulated by a learned Trust Score. If Trust is zero, the connection is effectively silenced.

$$\alpha_{ij} = \frac{(A_t[i, j] * T[i, j] * \exp(\ldots))}{\sum (\ldots)}$$

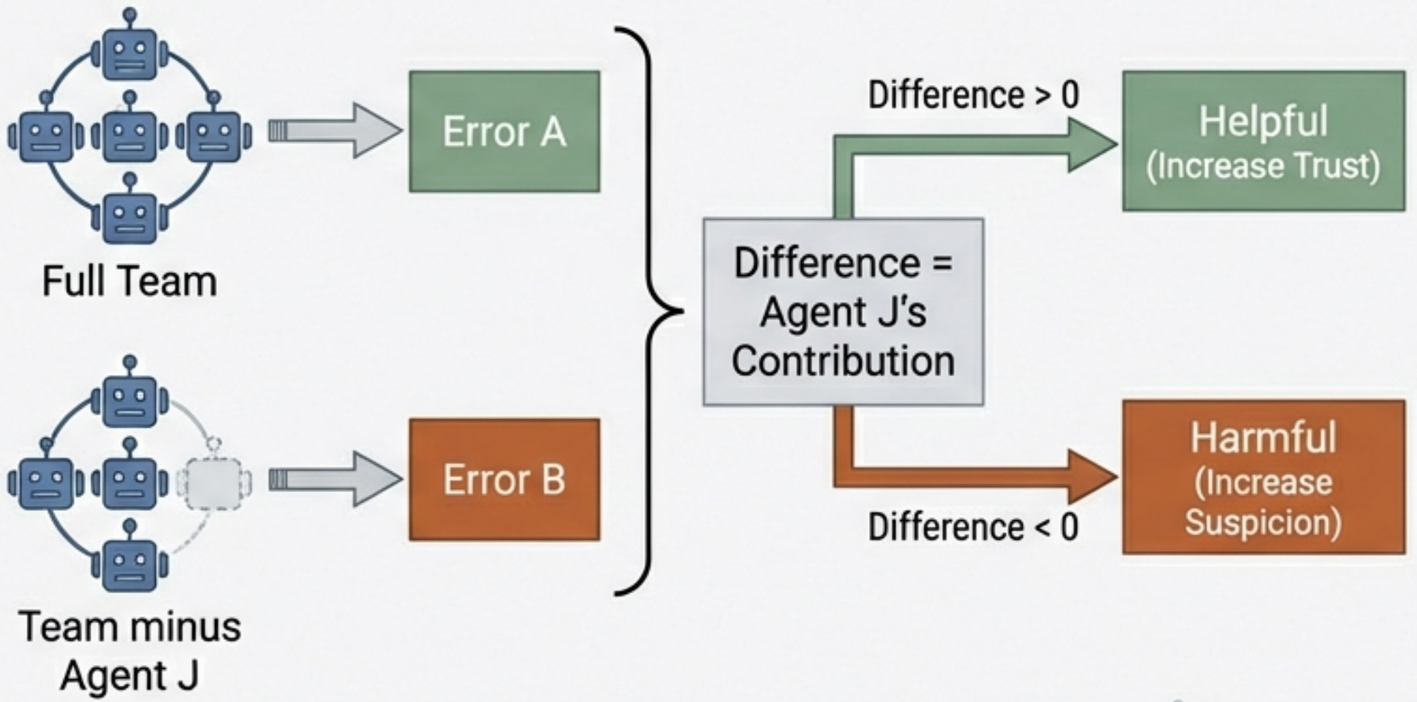
The Signal Filter



Pillar 3: The Immune System (Counterfactual Trust)

How do we measure reliability? By asking: "What would the result be if you weren't here?"
We compare the team's error with and without a specific agent.

$$\Delta_j = \left\| \hat{\mathbf{s}}_{-j} - \mathbf{s} \right\|_2 - \left\| \hat{\mathbf{s}} - \mathbf{s} \right\|_2$$



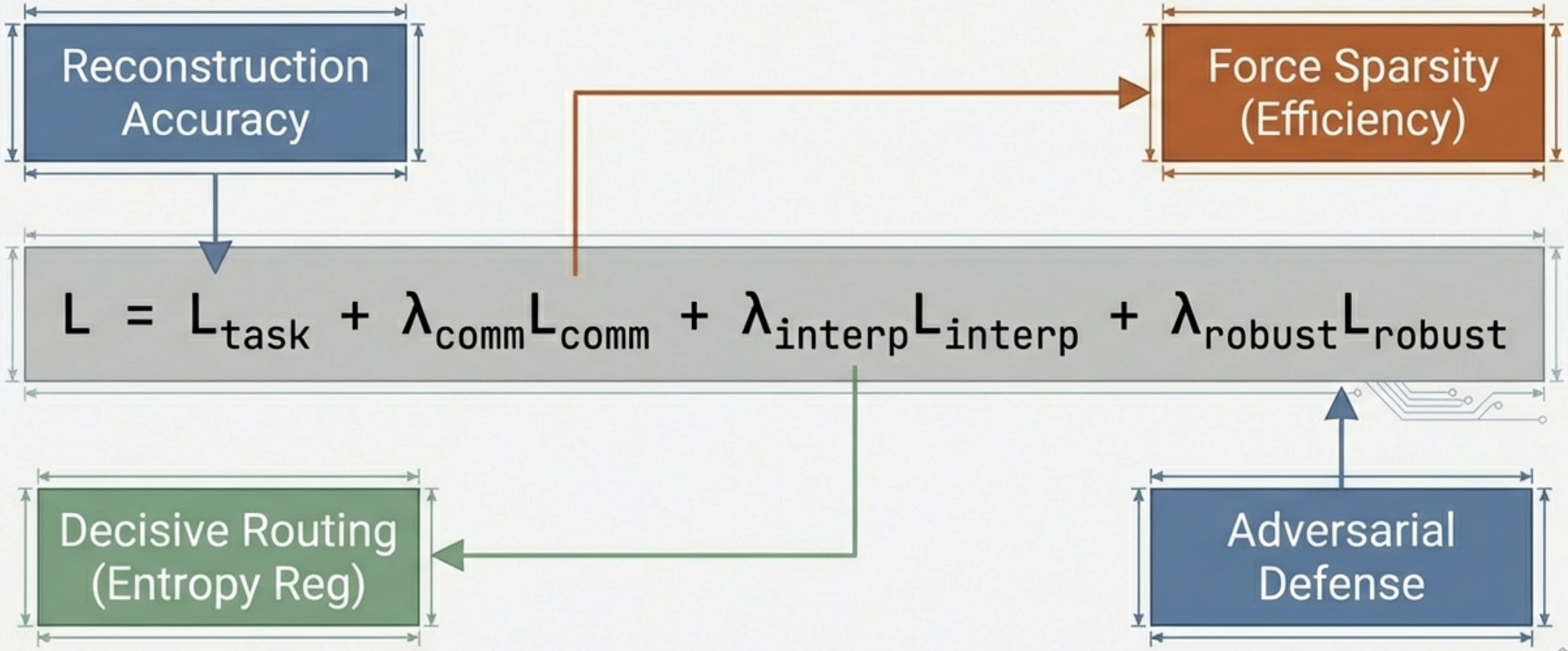
Modeling Uncertainty with Beta Distributions

Trust is not a single number, but a probability distribution. This allows the system to distinguish between “lucky guessers” and “proven experts”.


```
Positive interaction: Increment Alpha.
Negative interaction: Increment Beta.
Expected Trust = Alpha / (Alpha + Beta).
```

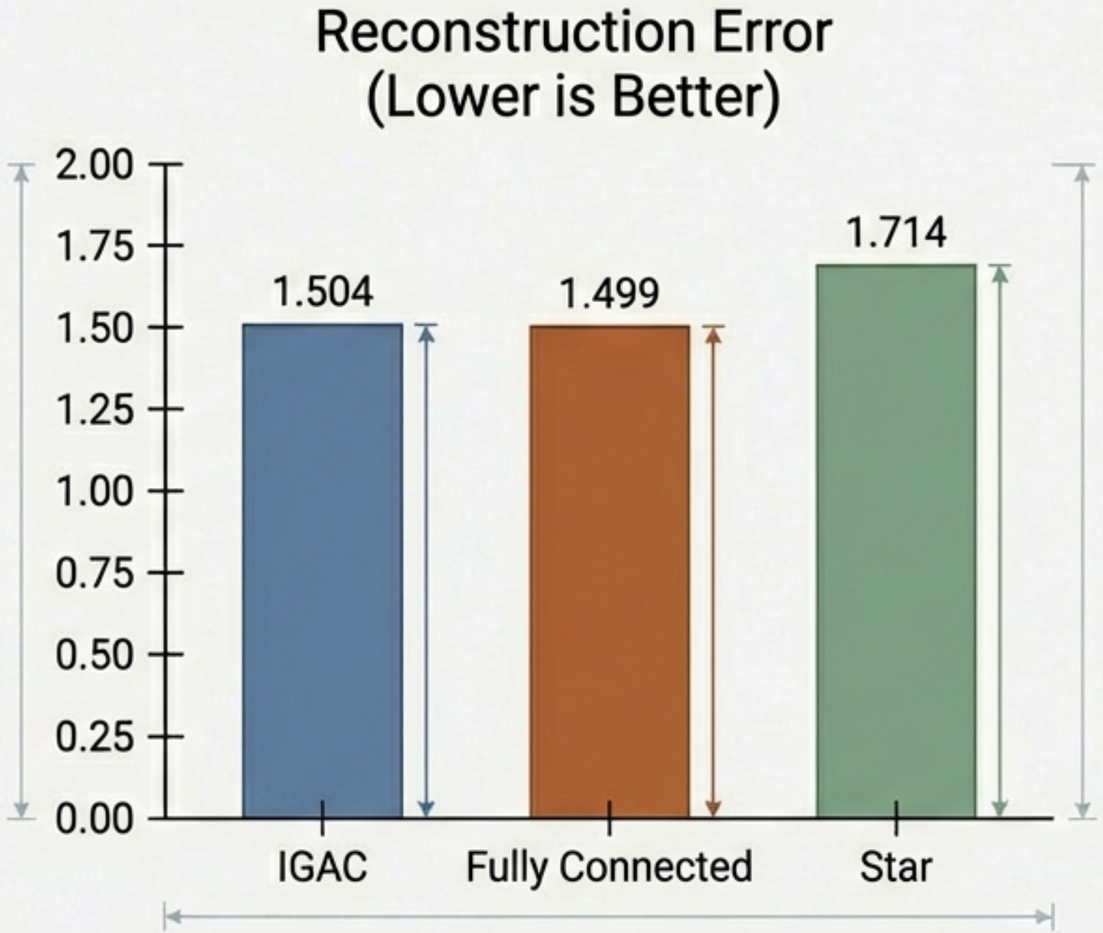
The Composite Training Objective

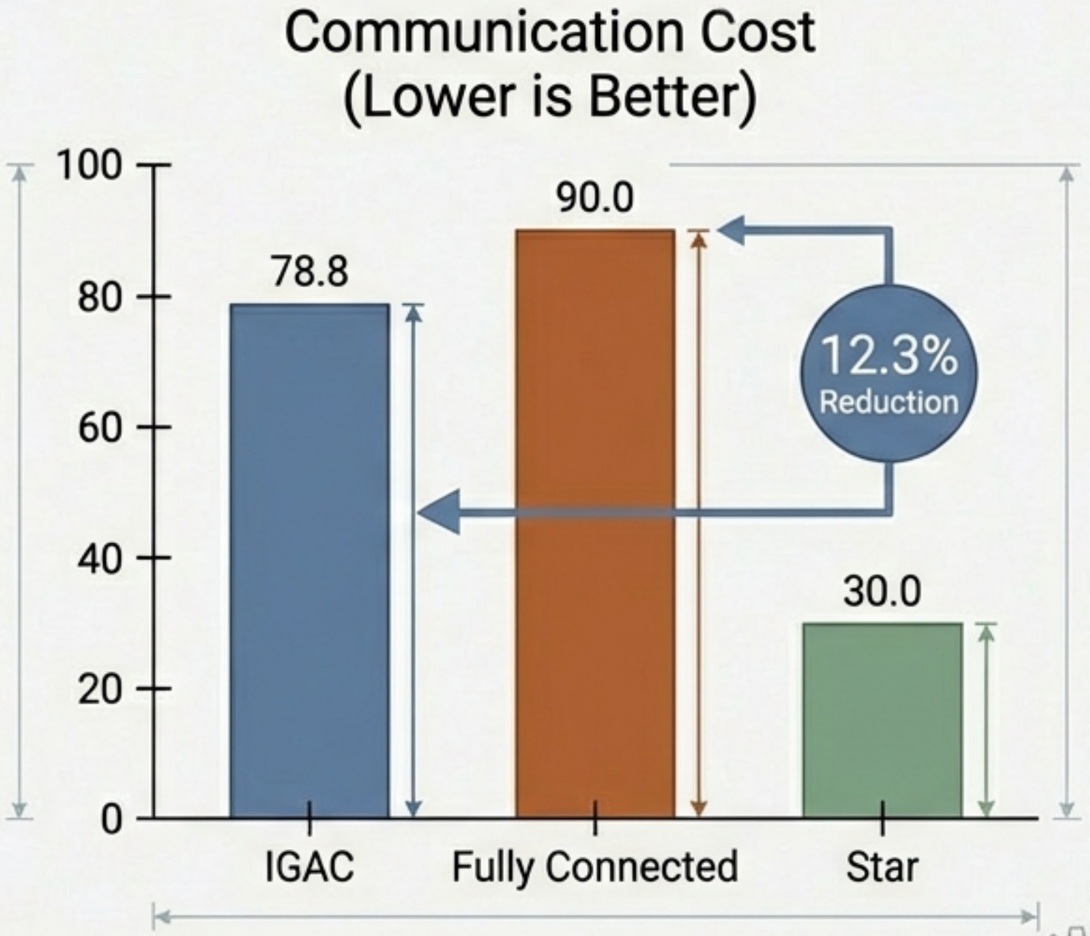
Optimizing for accuracy, structure, and robustness simultaneously.



Efficiency Without Compromise

IGAC matches the accuracy of fully connected systems with significantly less talk.

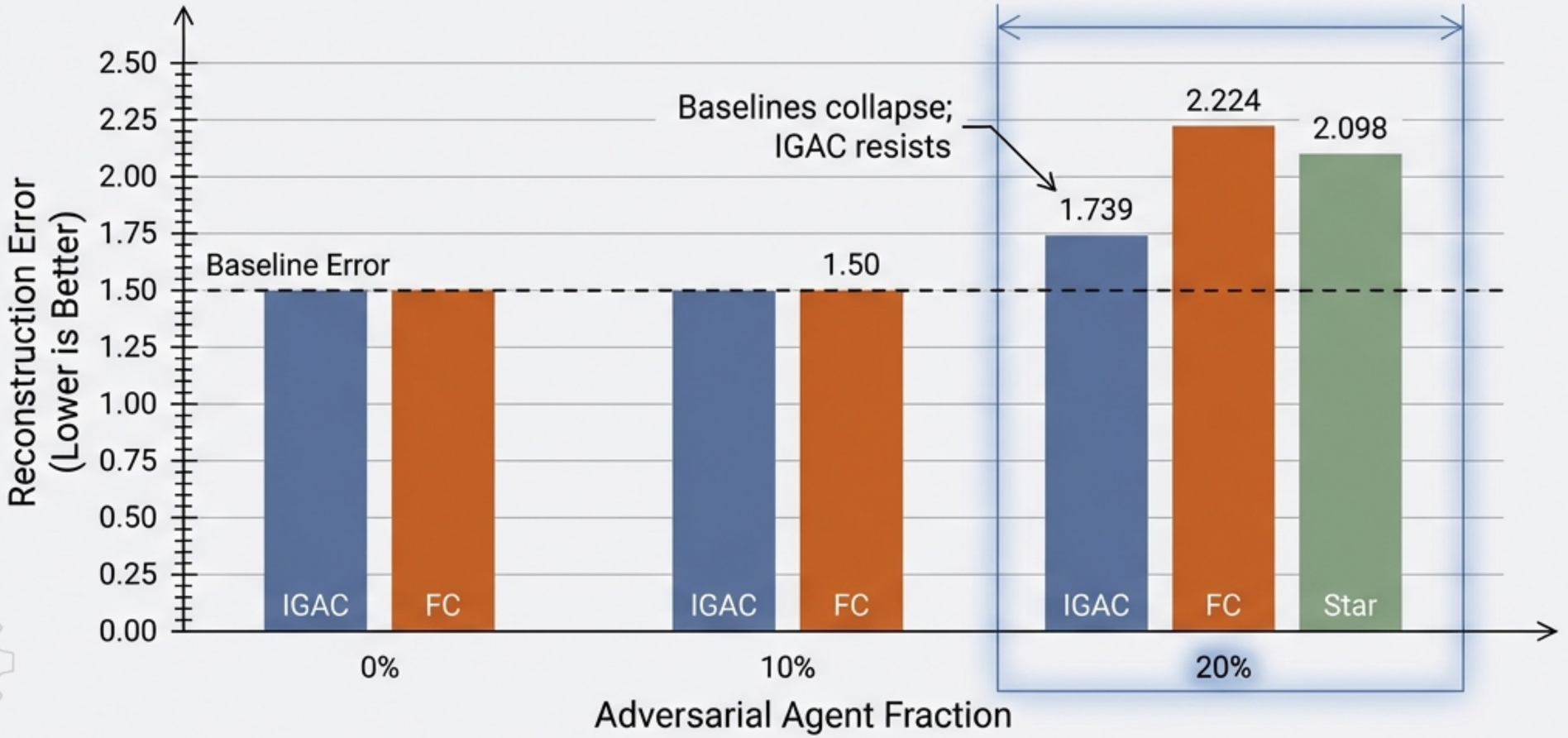




Note: Reconstruction Error is mean squared error; Communication Cost is relative message count.

Resilience Under Attack

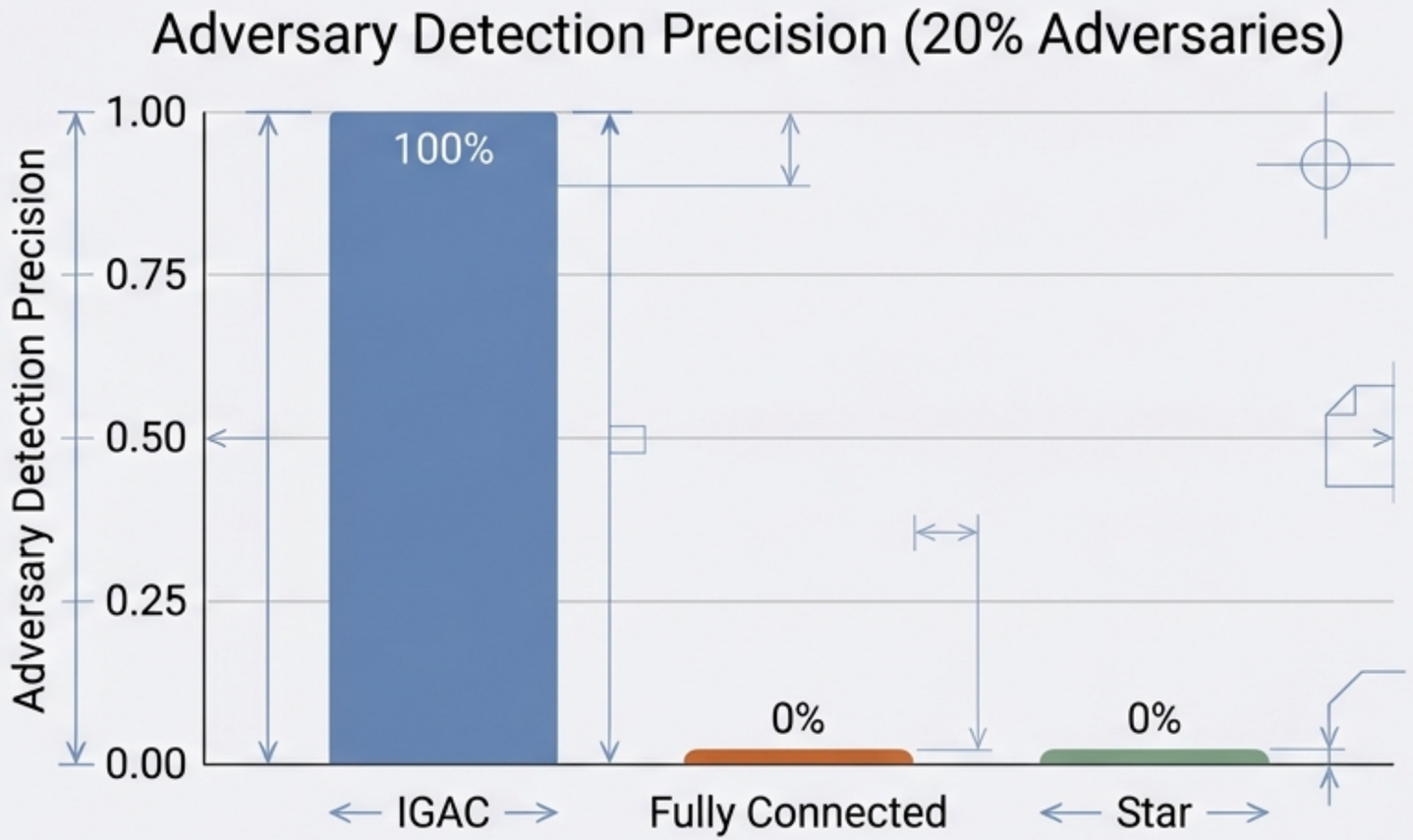
Performance when 20% of agents are adversarial.



Note: Reconstruction Error is mean squared error under attack.

Perfect Detection & Isolation

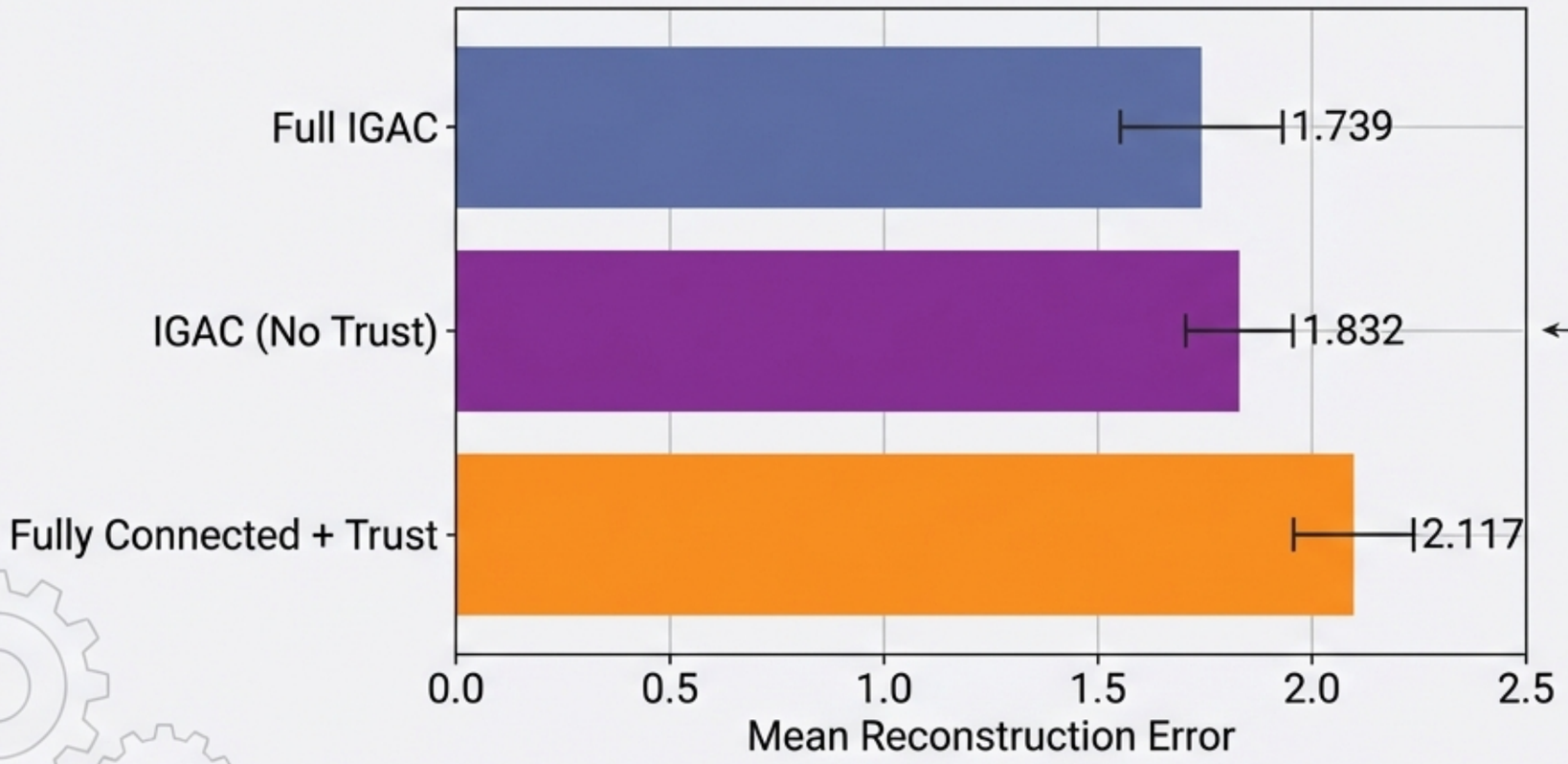
Identifying the ‘Traitor’ via Counterfactual Credit Assignment.



While standard architectures have no mechanism to distinguish friend from foe, IGAC’s trust module achieves perfect precision and recall in isolating the compromised node.

Anatomy of Robustness (Ablation Study)

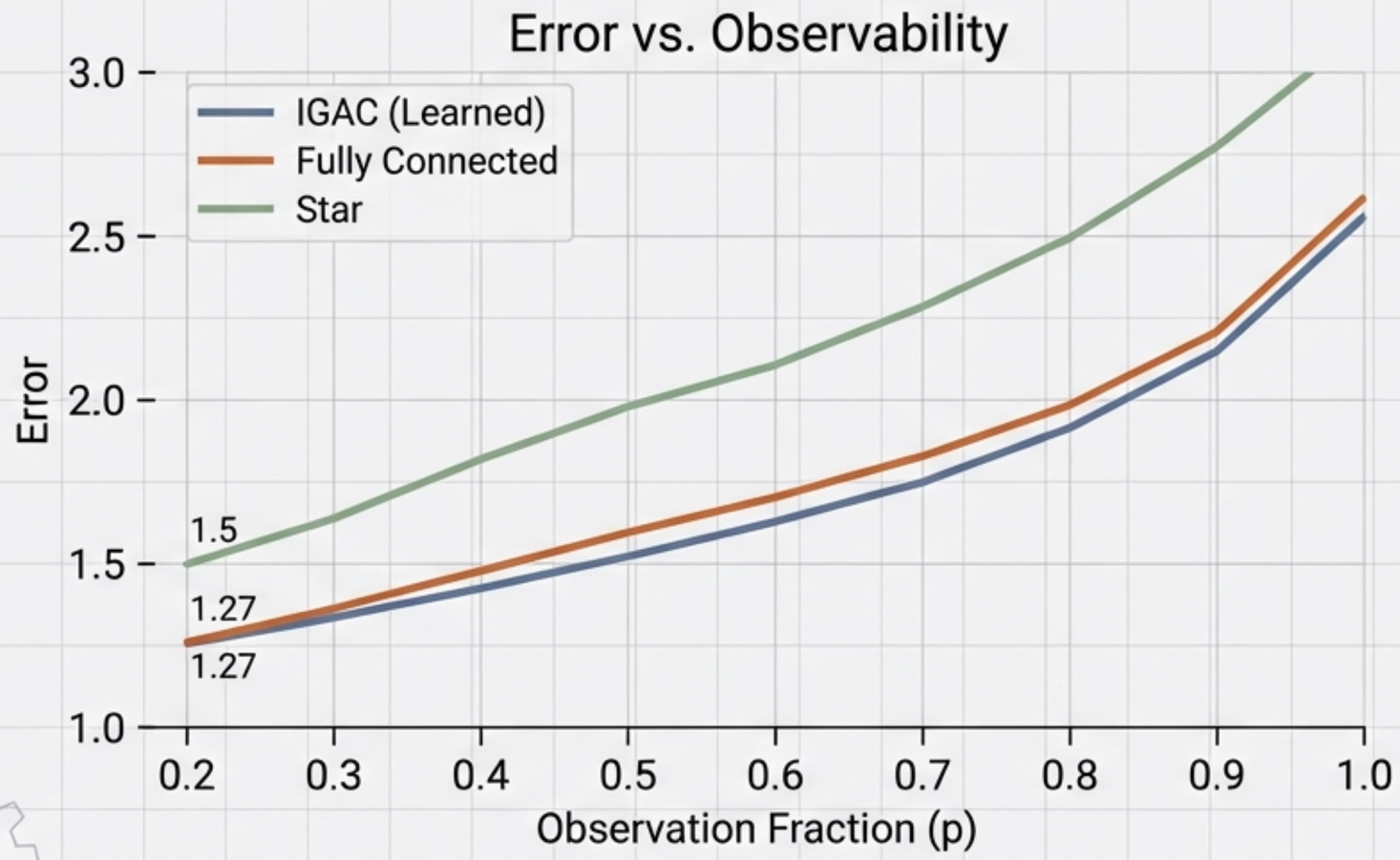
Both Learned Topology and Trust are required for defense.



Topology alone is not enough. Trust alone is not enough. The combination is critical.

Handling Partial Observability

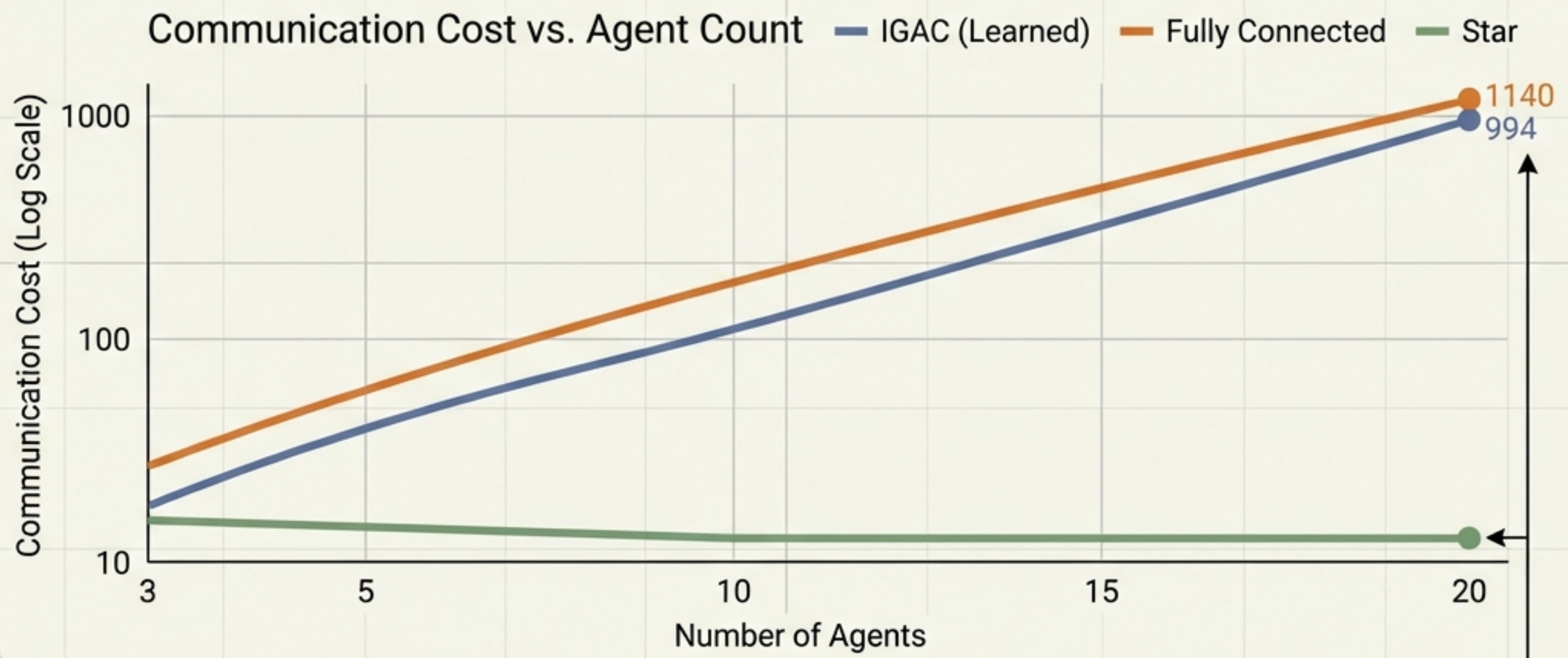
Adaptability when data is scarce.



At low observability (p=0.2), IGAC outperforms rigid structures by stitching together sparse data more effectively.

Scalability at 20 Agents

Sub-quadratic growth in communication costs.

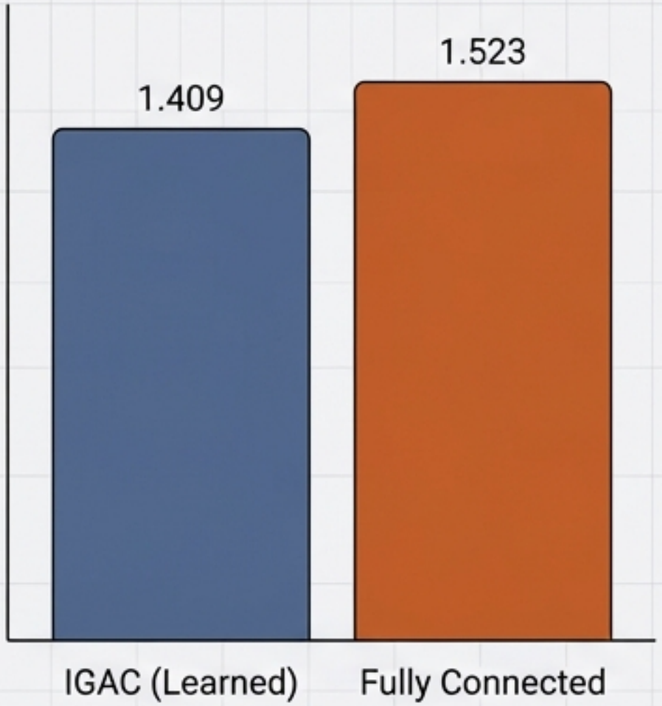


While Star is cheapest, its error spikes at N=20 (Hub Bottleneck).
IGAC provides the best balance of scale and accuracy.

Interpretability: Seeing the Reasoning

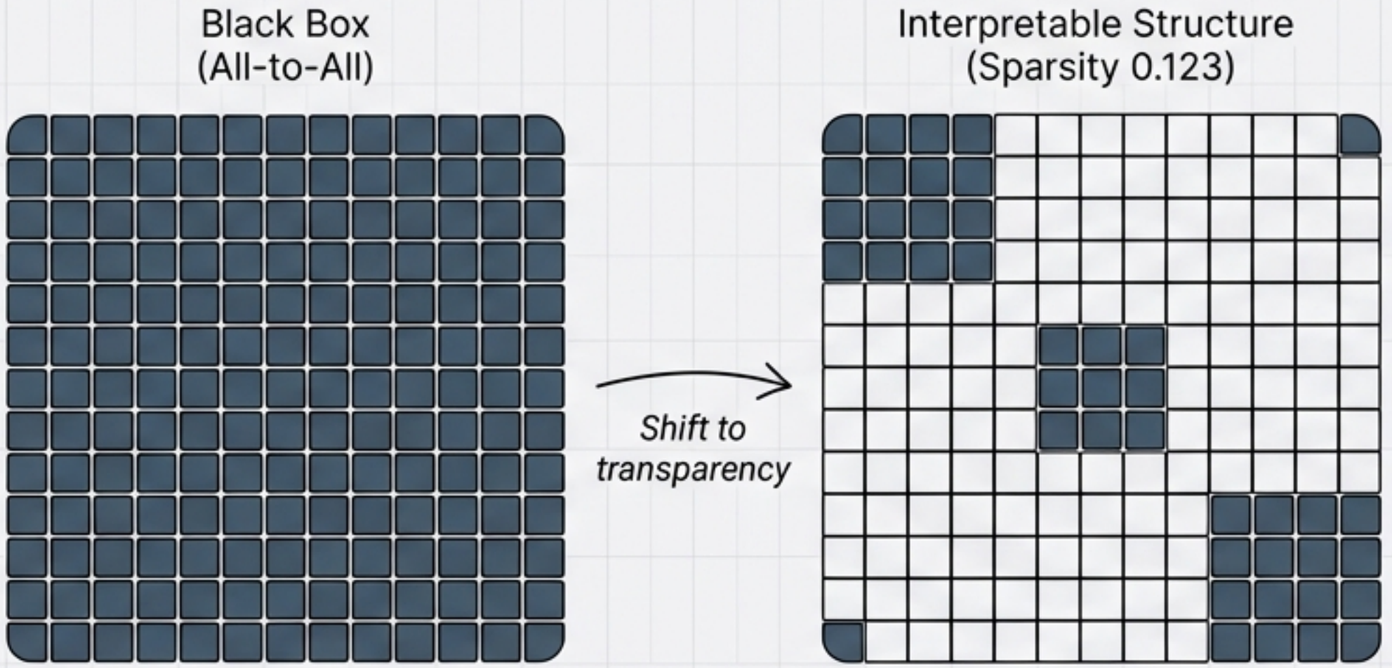
IGAC offers a ‘Glass Box’ view of agent collaboration.

Attention Entropy



Lower entropy = More decisive, confident routing.

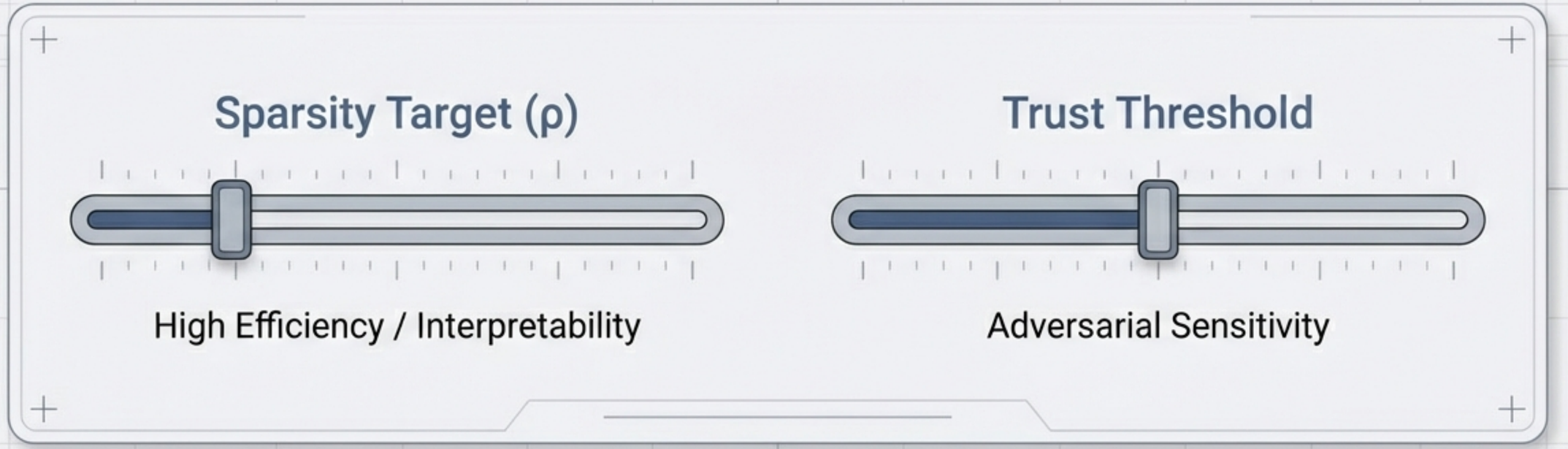
Edge Sparsity (Sparsity 0.123)



Swiss Engineering Blueprint

The Control Interface

IGAC is not a static black box; it offers tunable parameters for system architects.



Current Limitations & Future Work: Adapting Gumbel-Softmax relaxation for discrete text generation in Large Language Models (LLMs).

A New Paradigm for Collaboration

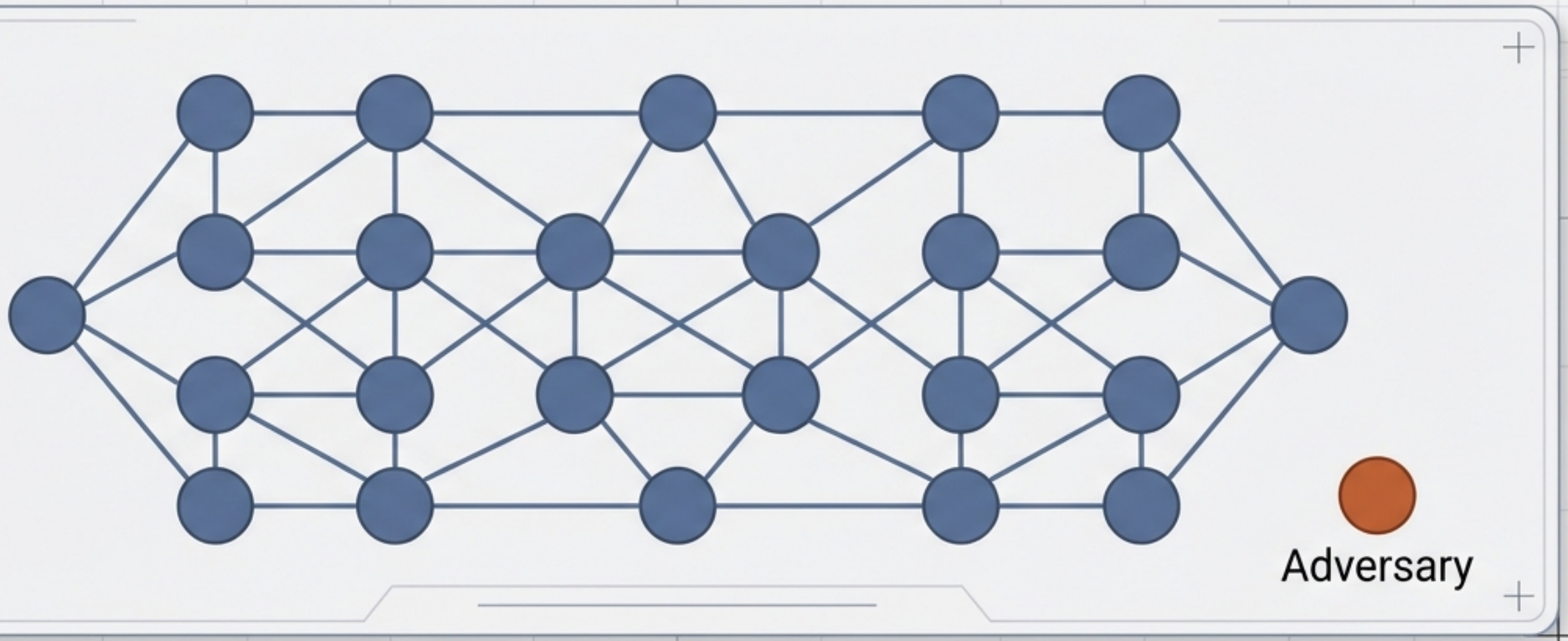
IGAC sets a new standard for Multi-Agent Systems.

Efficiency	Matches dense graph accuracy with sparse graph costs (~12% savings).	✓
Robustness	The only architecture to survive and detect 20% adversarial injection.	✓
Adaptability	Dynamic reconfiguration based on partial observability and instance difficulty.	✓

"We have moved from Blind Obedience in agent communication to Informed Collaboration."

Building Trustworthy Agent Teams

As we deploy multi-agent systems in high-stakes environments, Trust and Topology cannot be static design choices. They must be dynamic, learned policies.



Secure, Optimized Collaboration.