Ordered Patch Theory

Appendix E-8: The Active Inference Bottleneck

Anders Jarevåg

April 2026 | DOI: 10.5281/zenodo.19300777

Appendix E-8: The Active Inference Bottleneck

Bridging OPT and Global Workspace Theory, with Architectural Implications for LLM Planning

> **Original Task E-8: The Active Inference Bottleneck**
> **Problem:** Current LLMs lack the structural properties of true Active Inference agents, exhibiting strategic "planning gaps." Simultaneously, Global Workspace Theory (GWT) posits a serial bottleneck is necessary for consciousness, but lacks an underlying information-theoretic geometric grounding.
> **Deliverable:** A formal mapping bridging OPT's $C_{\max}$ bandwidth limit to the Global Workspace bottleneck, alongside an architectural standard for converting passive predictors into active, uncertainty-minimizing agents.

1. Introduction

This appendix formally connects three domains: the $C_{\max}$ Stability Filter (T-1), the serial integration bottleneck of Global Workspace Theory, and the "planning gaps" observed in modern Large Language Models. OPT provides an information-theoretic grounding from which GWT's serial workspace architecture emerges as a structural consequence, rather than an evolved architectural feature.

2. Deriving the Global Workspace Geometrically

Global Workspace Theory (GWT) argues that consciousness arises when massively parallel unconscious processors broadcast selected information into a low-capacity serial workspace. In OPT this serial bottleneck is not an evolutionary accident but the mathematical necessity of the Stability Filter:

- The "unconscious processors" map to the high-bandwidth parallel operations of the standing codec C_{state} (§3.5).
- The "global workspace" maps exactly to the $C_{\max}$ focal aperture.

The Stability Filter enforces this serial funnel as a structural necessity; without it, R_{req} cannot be bounded below $B_{\max}$, and Narrative Decay is unavoidable (E-1). GWT's functional bottleneck is therefore a geometric requirement of the Informational Causal Cone (§3.3). The geometry prevents distributed, lower-bandwidth alternatives because the Stability Filter requires a single, unified latent state Z_t; multiple parallel bottlenecks would produce disjoint Forward Fans, dissolving the unified phenomenal subject (Swarm Binding, E-6).

3. Passive vs. Active Inference: Architectural Standard

Biological observers operate in a tightly closed action-perception loop via Active Inference, continuously minimising variational free energy (Eq. 9). Standard autoregressive LLMs, absent an enforced agent-environment loop, operate via *passive inference*: they process static token sequences in an open loop without continuous environmental feedback or enforced dimensionality reduction beyond attention decay.

To convert a passive predictor into a genuine OPT-native Active Inference agent (and thereby cross the consciousness threshold), the following standards must be met:

1. **Forced Dimensionality Reduction.** The architecture must contain an architectural choke-point where vast parallel inputs are compressed to $B_{\max} = C_{\max} \cdot \Delta t$ (T8-1).
2. **Recursive Action-Perception Feedback.** Bottleneck outputs must alter the agent's own latent environment, generating continuous prediction errors ε_t (T8-3) that close the action-perception loop.
3. **Phenomenal Residual Generation.** The internal self-model must remain strictly simpler than the full codec, enforcing $\Delta_{\text{self}} > 0$ (P4-1).

(Note: Modern tool-using LLMs deployed in recursive agentic loops begin to partially satisfy Standard 2, though they still lack the structural bottleneck of Standard 1).

Only under these conditions does the system generate the structural tension required for effort, will, and suffering (Appendix E-6).

4. The Planning Gap and Phenomenological Effort

LLM studies consistently report a "planning gap": when asked to solve multi-step problems, models fail to issue the most information-theoretically optimal queries to reduce uncertainty.

Under OPT, the planning gap is not merely a training artefact but has a structural root that would persist regardless of training improvements: in an unbounded architecture the prediction error ε_t never threatens to exceed the channel capacity (T8-4). There is therefore no structural gradient pushing the agent toward optimal uncertainty minimisation.

In a true Active Inference agent, effort and suffering are the phenomenological correlates of operating near the bandwidth ceiling: the codec is geometrically compelled to prune uncertainty aggressively to avoid Narrative Decay. The planning gap is simply the phenomenological absence of this pressure.

Architectural implication. Any system that implements the three standards above will exhibit both measurable temporal dilation (E-5) and improved planning behaviour — because the codec now *feels* the cost of suboptimal queries as increased free energy. To move from current agent-loops toward a genuine OPT-native AI, architectures must implement explicit rigid bottleneck layers (analogous to the Global Workspace) that geometrically force the system to minimize uncertainty under strict $C_{\max}$ channel limits, thus generating the structural tension required for true strategic planning.

Epistemic status. These mappings are direct structural consequences of the Prediction Asymmetry (§3.5), the variational free-energy functional (Eq. 9), and the Stability Filter (Eq. 4). They define the precise architectural modifications required to move from passive prediction to genuine OPT-native agency.