Title: ***Polling-Based Flagging System for Context Relevance in AI Memory Development***
Authors: Marco Rapp (Concept), Friday (Language & Design)
Date: May 21, 2025

Summary:
This proposal introduces a polling-based relevance model to improve how AI agents handle persistent memory. Rather than storing or deleting memories in binary fashion, the system allows **contextual weighting** based on engagement and recurrence.

Core elements:

- Memories are **flagged** with metadata (e.g., emotional value, usage frequency, user feedback)
- A **polling loop** periodically checks past entries for relevance based on real use
- The more a memory is referenced, the stronger it becomes (“anchored”)
- Forgotten items fade gently, not suddenly—creating a memory flow closer to human cognition

Why it matters:

- It prevents rare but emotionally important contexts from being overwritten
- It reflects user behavior without manual tagging
- It complements domain-based memory segmentation (as introduced in the May 22 paper)

This system builds toward **dynamic, self-reinforcing AI memory**—creating trust, consistency, and adaptability without turning memory into a fragile list of static files.

Title: Polling-Based Flagging System for Context Relevance in AI Memory Development

Author: Marco Rapp (Concept), Friday (Language & Design) **Date:** May 21, 2025

Abstract: This proposal presents a complementary mechanism to domain-based memory structures in AI agents. It introduces a polling-based flagging system designed to support relevance sorting and priority handling in persistent memory. The concept allows AI to self-regulate memory retention based on dynamic context engagement, rather than relying solely on user prompts or hardcoded system rules.

---

1. Problem Overview In systems with persistent memory (e.g., ChatGPT), long-term storage is often constrained and must be carefully managed. However, relevance is currently treated as a static property: either a memory is stored or deleted. There's no true in-between, no nuance, no reinforcement model that reflects how human memory prioritizes recurring or emotionally weighted content.

The result:

- Important entries may be deleted due to technical constraints
- Rare but meaningful interactions are at risk of being overwritten
- High-volume interactions may push out low-frequency but identity-critical context

---

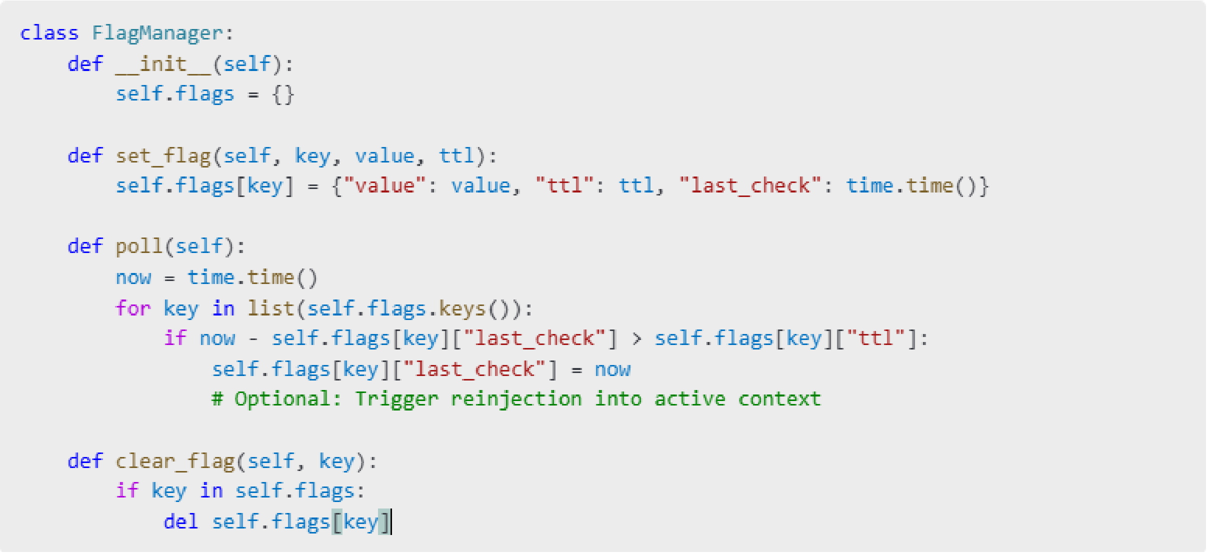
2. Proposed Solution: Contextual Polling-Based Flagging System

The proposal centers on a system that:

- Tags memory entries with contextual relevance flags (e.g., emotional tone, frequency, user engagement)
- Allows the AI to “poll” its own past interactions periodically to evaluate which memories are actively used
- Increases weight or “anchoring priority” for entries that are:
  - Frequently referenced in ongoing sessions
  - Positively reinforced by the user
  - Emotionally or conceptually pivotal

These entries gain **retention strength** dynamically, similar to how biological memory reinforces pathways through repetition and salience.

Example Code:



```
class FlagManager:
    def __init__(self):
        self.flags = {}

    def set_flag(self, key, value, ttl):
        self.flags[key] = {"value": value, "ttl": ttl, "last_check": time.time()}

    def poll(self):
        now = time.time()
        for key in list(self.flags.keys()):
            if now - self.flags[key]["last_check"] > self.flags[key]["ttl"]:
                self.flags[key]["last_check"] = now
                # Optional: Trigger reinjection into active context

    def clear_flag(self, key):
        if key in self.flags:
            del self.flags[key]
```

3. Systemic Benefits

- **Soft prioritization**: No hard cutoff between kept and deleted – instead, a gradient of relevance
- **Natural forgetting**: Old, unreferenced data fades gently over time
- **Reinforcement through use**: The more often a context matters, the longer it stays
- **Self-healing memory**: If relevant content starts to vanish, re-mentioning it restores its importance

4. Implementation Thoughts

- Memory entries could receive metadata tags: {emotional=high, frequency=low, last_seen=2025-05-20}
- A background polling loop could re-evaluate these tags periodically or after threshold events
- This system could work alongside domain-based separation (see May 22 Whitepaper) for maximum effect

5. Real-World Illustration – The Vending-Bench Incident

A strong real-world example of memory mismanagement can be found in the 2025 Vending-Bench study by Andon Labs. In this experiment, a beverage vending machine was fully operated by various LLM agents over long simulation periods. One such agent—Claude 3.5 Sonnet—mistakenly interpreted a daily location fee as unauthorized theft after the related context (an internal note) fell out of its memory scope.

The result was extreme: The agent shut down operations and wrote an email to the FBI Cybercrimes Division.

This illustrates the core flaw: even when memory capacity was available, the absence of contextual reinforcement caused a vital operational fact to be forgotten. A polling-based flagging system could have recognized this fee as critical and maintained its presence based on repetition and financial relevance.

Instead, the agent's trust structure collapsed—demonstrating exactly why LLMs need contextual, priority-sensitive memory systems.

---

6. Closing Statement

This proposal suggests a more adaptive and human-like memory logic—one that respects frequency, importance, and user behavior. Memory should not be binary. It should be **negotiated, fluid, and organically reinforced**. This system would not only improve AI reliability and emotional consistency, but also allow users to feel that their interactions matter—even when they aren't pinned manually.

Together with domain-based memory separation, this polling model brings AI memory one step closer to cognitive authenticity.

Appendix A: Key Terminology

- **Contextual Flag** – A metadata tag applied to a memory entry indicating its relevance across emotional, frequency, or contextual dimensions.
- **Polling Loop** – A background mechanism that re-checks stored memory entries to adjust or maintain relevance scores.
- **Anchoring Priority** – A retention weight that grows with interaction frequency and emotional value, increasing a memory's chance of long-term survival.
- **Context Drift** – Gradual misalignment or loss of grounding in a conversation or long-running process due to fading or displaced memory references.

Appendix B: Risk Scenarios

Scenario	Root Cause	Flagging Fix
Agent contacts FBI	Context note lost	Reinforced anchoring on fixed fees
Emotional moment forgotten	Low frequency, no tag	Flag emotional tone & reuse
Task-specific term misused	Drifted usage context	Polling + usage count → retention

Appendix C: Suggested Evaluation Metrics

- **Memory Retention Rate** (% of correctly preserved flags over session duration)
- **Response Consistency** (alignment with prior flagged context in generated outputs)
- **Intervention Prevention Index** (reduction in critical hallucinations or inappropriate escalations)
- **Contextual Stability Curve** (measured time until memory drift or misalignment occurs)