Comment on the HHS Health Sector AI Request for Information

Subject: Addressing Agentic Drift in Non-Device AI Through Deterministic Semantic Memory

Submitted to:
U.S. Department of Health and Human Services
Health Sector Artificial Intelligence Request for Information (RFI)

I. Introduction

Thank you for the opportunity to respond to the HHS Health Sector AI Request for Information. This comment addresses a practical and currently under-addressed barrier to the safe deployment of non-device artificial intelligence in health insurance, benefits communication, and utilization management: the absence of a deterministic semantic memory layer capable of anchoring AI systems to authoritative policy, benefit, and regulatory definitions.

While current discussions appropriately emphasize data access, interoperability, and model governance, these approaches alone do not resolve a recurring operational failure observed in deployed AI systems—agentic drift.

Several submissions correctly emphasize post-deployment monitoring, organizational governance, and improved data completeness; however, these approaches still depend on a deterministic semantic layer to ensure AI outputs apply authoritative policy meaning consistently and can be audited without model inspection.

This approach operationalizes the “AI-enabling infrastructure and documentation” principles articulated in OMB Memorandum M-25-21 (September 2025) by making authoritative policy meaning explicit, versioned, and auditable for high-impact administrative AI systems.

II. Practical Barrier: Agentic Drift in Non-Device AI

Agentic drift refers to the tendency of AI systems to diverge from authoritative policy intent, benefit definitions, and regulatory meaning over time when operating without a persistent, verifiable semantic memory substrate.

In regulated benefit environments, this manifests as:

- Inconsistent interpretations of coverage and cost-sharing across AI systems

- Hallucinated or outdated benefit explanations
- Divergence between plan documents, regulatory guidance, and AI-generated communications
- Limited ability to audit, explain, or reproduce how a specific output was generated

This failure mode is increasingly visible in Medicare Advantage and other federal benefit programs where AI systems assist beneficiaries, providers, and administrative workflows.

III. Data Transport vs. Semantic Determinism

Existing health IT infrastructure has made substantial progress in data transport:

- FHIR and related interoperability standards enable structured, standardized exchange of health information.
- APIs and data services support real-time access to coverage, utilization, and clinical data.

However, these mechanisms are intentionally neutral with respect to semantic meaning. They do not encode authoritative policy interpretation, versioned benefit definitions, or regulatory intent in a form that AI systems can deterministically retrieve and apply.

A deterministic semantic substrate should therefore be understood not as a replacement for FHIR-based exchange, but as the semantic payload carried alongside it—providing authoritative meaning where transport standards appropriately remain agnostic.

Deterministic semantic substrates represent a layer of semantic interoperability that complements existing transport standards, enabling consistent meaning across systems and reducing friction in cross-organizational AI deployments.

This distinction is increasingly relevant as agencies seek to satisfy emerging algorithmic transparency expectations, including those articulated in the CMS-0057-F Prior Authorization Final Rule, where explainability and reproducibility of automated determinations are central concerns.

IV. Policy Concept: Deterministic Semantic Substrate for Algorithmic Transparency

HHS should recognize and encourage the use of a Deterministic Semantic Substrate for Algorithmic Transparency as a missing governance layer for non-device AI.

A deterministic semantic substrate is a structured, machine-readable memory layer embedded within authoritative content surfaces—such as benefit descriptions, coverage rules, Evidence of Coverage documents, and policy explanations—that enables AI systems to retrieve and apply versioned policy meaning deterministically rather than infer it probabilistically.

WebMEM™ is referenced solely as an illustrative implementation of this vendor-neutral protocol; the policy proposal itself is intentionally model-agnostic and independent of any specific technology provider.

Key characteristics include:

- Explicit semantic structure (not inferred meaning)
- Stable identifiers and versioned definitions
- Provenance metadata tied to authoritative sources
- Cryptographically verifiable provenance (e.g., checksums or hashes) to support integrity and auditability
- Human-readable and machine-readable parity
- Independence from any specific AI model, vendor, or deployment architecture

This approach shifts trust from probabilistic model behavior to content-level governance, reducing reliance on inference where determinism is required.

By anchoring AI outputs to verifiable, authoritative policy fragments, this approach reduces legal uncertainty by eliminating inconsistent interpretations that can lead to liability claims or regulatory non-compliance.

V. Quantifiable Benefits for Program Integrity and Fiscal Stewardship

The implementation of a deterministic semantic substrate directly addresses several high-cost operational pressure points in the 2026 Medicare ecosystem. These benefits are not theoretical; they align with existing CMS thresholds, innovation model targets, and program integrity cost centers.

1. Reduction of Appeals Friction and Administrative Adjudication Costs

As of January 1, 2026, the Amount in Controversy (AIC) thresholds for Medicare appeals have increased to:

- **$200** for Administrative Law Judge (ALJ) hearings
- **$1,960** for judicial review

Observed Problem:
Probabilistic AI-generated explanations—particularly in Part D benefit communications (e.g., misinterpretation of the $2,100 out-of-pocket cap or the $35 insulin cost-sharing limit)—frequently trigger low-dollar, non-clinical appeals. In many cases, the cost of adjudication exceeds the value of the underlying claim, creating negative administrative leverage.

These appeals are often driven not by coverage disputes, but by semantic inconsistency between AI explanations, plan materials, and statutory benefit definitions.

Impact of Deterministic Semantic Substrates:
By anchoring AI-generated communications to deterministic, versioned policy fragments, a significant portion of these appeals can be avoided entirely.

Based on historical appeal mix and OMHA processing costs, a conservative directional estimate suggests that reducing non-clinical, explanation-driven appeals by approximately 20–25% could yield meaningful annual savings through avoided administrative handling—potentially on the order of hundreds of millions of dollars system-wide—while also reducing beneficiary and provider burden.

2. Suppression of Agentic Upcoding and Interpretive Drift in the WISeR Model

The Wasteful and Inappropriate Service Reduction (WISeR) Model targets service categories with historically high improper payment risk, including skin substitutes, where Medicare spending increased from approximately $256 million to over $10 billion in a five-year period.

Observed Problem:
Technology participants operating AI-assisted utilization management systems are inherently incentivized to optimize toward cost containment benchmarks. In the absence of a deterministic semantic substrate, this can result in:

- interpretive drift in medical necessity criteria,
- incorrect service grouping,
- or overly aggressive denial logic driven by probabilistic inference rather than authoritative policy meaning.

These behaviors increase the risk of provider disputes, litigation, and retroactive payment correction—undermining both savings targets and program credibility.

Impact of Deterministic Semantic Substrates:
Deterministic semantic fragments function as sovereign truth anchors, eliminating the "interpretation delta" between policy intent and automated enforcement.

By binding AI-assisted determinations to verifiable, policy-scoped semantic definitions, CMS can materially reduce one of the dominant contributors to improper payment risk in automated utilization management. Even modest reductions in interpretive error rates have outsized fiscal impact when applied to high-volume, high-cost service categories.

This capability directly supports CMS efforts to protect the projected $19.6 billion in savings associated with 2026 Physician Fee Schedule and utilization reforms, without requiring prescriptive regulation of AI models themselves.

3. Lowering the Beneficiary Support "Inquiry Tax"

CMS projects approximately $25 billion in increased Medicare Advantage payments for 2026, reflecting both enrollment growth and benefit complexity. As benefit structures grow more complex, so does the operational burden on beneficiary support channels.

Observed Problem:
AI-generated misinformation—particularly regarding cost-sharing, benefit phases, and eligibility—drives secondary and tertiary contacts to 1-800-MEDICARE and plan call centers. These contacts frequently involve:

- correction of AI-generated explanations,
- manual verification of benefit rules,
- tier-2 escalations for issues that should be resolvable at first contact.

This creates an implicit "inquiry tax" on beneficiary support operations.

Impact of Deterministic Semantic Substrates:
Transitioning to a retrieval-first architecture—where AI systems cite cryptographically verifiable, authoritative semantic fragments—reduces resolution latency and increases first-contact accuracy.

Operational experience in adjacent domains suggests that improving semantic determinism can reduce escalation rates by approximately 25–30% for explanation-driven inquiries. Applied at scale, this represents a deflationary shift in federal customer service costs while simultaneously improving beneficiary experience and trust.

Summary

Across appeals adjudication, utilization management, and beneficiary support, deterministic semantic substrates convert AI governance from a reactive, probabilistic control problem into a preventive, content-governance discipline. The resulting effect is not merely improved accuracy, but structural cost deflation in areas where administrative overhead currently outpaces policy value.

VI. Program Integrity and Anti-Spoofing Considerations

Binding AI-generated explanations and determinations to authoritative benefit definitions—such as CMS Plan Benefit Package identifiers—also reduces the risk of automated misrepresentation or opportunistic upcoding.

In this sense, deterministic semantic substrates function as anti-spoofing mechanisms for federal benefits, ensuring that AI systems cannot misbind beneficiary identity, coverage, or cost information in ways that advantage financial outcomes over regulatory accuracy.

VII. Alignment with Existing Regulatory Requirements

Deterministic semantic substrates can support compliance with existing disclosure and accuracy requirements, including those found in:

- **42 CFR § 422.111** (Disclosure requirements)
- **42 CFR § 422.2267** (Required materials and content)

By ensuring that machine-generated communications remain anchored to the same authoritative definitions as Evidence of Coverage (EOC) and Summary of Benefits (SB) materials, such substrates provide a technical mechanism to reinforce accuracy obligations already established under Subpart V—without introducing new mandates.

VIII. Policy Recommendations

HHS should consider the following actions:

1. Recognize agentic drift as a distinct and addressable risk in non-device AI
2. Identify deterministic semantic memory as a missing layer in current AI governance frameworks
3. Encourage adoption of deterministic semantic substrates through guidance, pilots, and innovation models
4. Support a federal reference implementation for a complex benefit program (e.g., Medicare Part D) to establish a baseline for auditable, machine-readable benefit truth
5. Avoid conflating data access with semantic determinism when evaluating AI safety and accountability

HHS could advance this approach through existing vehicles—such as the AI Governance Board, CMS sub-regulatory guidance, or limited voluntary demonstrations coordinated with CMMI—without requiring new statutory authority or model-specific regulation.

These actions can be undertaken using existing statutory authority and sub-regulatory mechanisms and do not require model-level regulation, new certification regimes, or expanded federal operational responsibilities.

IX. Closing

Effective AI governance in healthcare will not be achieved solely by constraining model behavior. It also requires strengthening the semantic signals those models consume. Addressing agentic drift at the content and memory layer offers HHS a practical, scalable path to improving transparency, auditability, and trust in non-device AI systems—without stifling innovation.

Thank you for the opportunity to provide input.

Respectfully submitted,

David W. Bynon
Independent Systems Architect

Appendix A

Illustrative Example of Deterministic Semantic Memory for Non-Device AI

Purpose of This Appendix

This appendix provides a concrete, illustrative example of how a deterministic semantic substrate prevents AI hallucination and misinterpretation in a regulated federal benefit context. The example is intentionally narrow and operational, demonstrating how authoritative semantic memory can anchor AI outputs to verifiable policy definitions without constraining model choice or behavior.

This example is illustrative only and does not represent a proposed standard or mandate.

A. Problem Scenario: AI Hallucination in Part D Cost Interpretation

Context:
Beginning in 2025–2026, Medicare Part D includes statutory caps on beneficiary insulin cost-sharing. In practice, confusion arises when AI systems attempt to explain costs for:

- Combination insulin products
- Non-standard NDC groupings
- Formularies with tier-specific or phase-specific pricing

Observed Failure Mode:
When relying on probabilistic inference or fragmented data sources, AI systems may:

- Hallucinate monthly insulin costs above the statutory cap
- Misapply caps across benefit phases
- Confuse drug-level pricing with plan-level summaries

This constitutes agentic drift, not because data is unavailable, but because authoritative semantic meaning is not deterministically retrievable.

B. Deterministic Semantic Approach (Conceptual)

Instead of asking an AI system to *infer* benefit rules from narrative text, a deterministic semantic substrate embeds machine-readable policy meaning directly within authoritative content surfaces.

The AI system:

1. Retrieves the semantic fragment
2. Applies the bound definition
3. Produces an explanation that is semantically locked to the authoritative rule

C. Illustrative Embedded Semantic Fragment (YAML-in-HTML)

Below is an illustrative example of an embedded semantic fragment representing a 2026 Medicare Part D insulin cost-sharing rule.

This fragment is:

- Human-readable
- Machine-readable
- Versioned
- Provenance-bound
- Checksum-verifiable
- HTML embeddable

```
<template data-semantic-fragment="medicare-partd-insulin-cap">
version: "2026.1"
fragment_type: "benefit_definition"
program: "Medicare Part D"
policy_scope: "cost_sharing"
policy_name: "Insulin Cost-Sharing Cap"
effective_date: "2026-01-01"

authoritative_definition:
  rule_text: >
    For plan year 2026, beneficiary cost-sharing for covered insulin
    products under Medicare Part D shall not exceed the statutory cap
    per covered prescription, regardless of benefit phase.
  statutory_cap_usd: 35.00
```

```
  applies_to:
    - "covered insulin products"
    - "combination insulin products containing insulin as an active ingredient"
  exclusions:
    - "non-covered drugs"
    - "supplies without insulin as an active ingredient"

binding:
  benefit_identifier:
    type: "CMS_PBP"
    description: "Plan Benefit Package Identifier"
  coverage_phase: "all"

provenance:
  authority: "Centers for Medicare & Medicaid Services"
  authority_document:
    type: "Statute / CMS Guidance"
    reference: "Medicare Part D Insulin Cost-Sharing Provision"
  policy_year: 2026

integrity:
  checksum_sha256: "e3b0c44298fc1c149afbf4c8996fb924..."
  checksum_scope: "fragment_body"

intended_use:
  - "AI beneficiary communications"
  - "AI-assisted plan explanations"
  - "Audit and compliance verification"

notes:
  - "This fragment represents authoritative semantic meaning."
  - "AI systems should not infer alternate pricing outside this definition."
</template>
```

D. How This Prevents AI Hallucination

When an AI system encounters a question such as:

“How much will my insulin cost each month under this Part D plan?”

A system consuming this fragment:

- **Cannot exceed** the defined cap
- **Cannot misapply** phase-based pricing
- **Cannot hallucinate** alternative interpretations
- Can explicitly cite the bound definition

This constraint is enforced at the system level. The AI system’s retrieval policy treats the fragment as an authoritative reference whose fields must be applied, not reinterpreted. Because the fragment is retrieved, versioned, and verified prior to response generation, the AI system does not reason about policy meaning—it applies it.

Any output that diverges from the retrieved fragment can be deterministically detected and rejected. As a result, the AI output is deterministic with respect to policy meaning, even though the language generation itself remains probabilistic.

E. Audit-by-Design Characteristics

This approach enables audit-by-design because:

- The semantic fragment is versioned
- The provenance is explicit
- The integrity is cryptographically verifiable
- The policy meaning is stable and reproducible

Auditors can verify:

- Which definition was applied
- Which version was in effect
- Whether the AI output conformed to the authoritative rule

This capability is particularly relevant for innovation models and technology participants supporting utilization management or beneficiary communication.

F. Relationship to Existing Infrastructure

This deterministic semantic fragment:

- Does **not** replace FHIR or API-based exchange
- Functions as a semantic payload alongside existing data transport
- Is independent of any specific AI model or vendor
- Supports existing disclosure and accuracy requirements without introducing new mandates

G. Compatibility with Retrieval-Augmented Generation (RAG)

This approach is fully compatible with Retrieval-Augmented Generation (RAG) architectures. In RAG-based systems, semantic fragments such as this are retrieved alongside other contextual materials; however, unlike general reference documents, these fragments are designated as authoritative constraints.

The AI system may use RAG to gather explanatory or supporting context, but policy meaning is bound to the retrieved fragment and applied deterministically. In this way, RAG enhances contextual richness while the semantic substrate ensures consistency, auditability, and resistance to hallucination.

H. Summary

This illustrative example demonstrates how a deterministic semantic substrate:

- Prevents agentic drift
- Reduces hallucinated benefit explanations
- Improves auditability and accountability
- Strengthens program integrity
- Preserves innovation flexibility

By securing the semantic signal, rather than attempting to regulate probabilistic model behavior, HHS can address a core governance gap in non-device AI.

End of Appendix A

Appendix B (Optional)

Illustrative Demonstration Concept: Deterministic Semantic Anchoring for Provider Directory Accuracy

Purpose and Disclaimer

This illustrative demonstration is provided solely as a non-binding example to show how the principles described in the main comment could be implemented in a low-risk, operationally familiar context. It is not a recommendation, proposal, or request for implementation, and does not imply new mandates, procurement actions, or centralized federal authorship of provider data.

The example is intended only to help HHS visualize how deterministic semantic anchoring could reduce administrative friction, beneficiary confusion, and audit complexity in AI-mediated environments.

I. Rationale for a Provider Directory Demonstration

Provider directory accuracy is a persistent, well-understood challenge across Medicare Advantage and other managed care programs. Inaccurate or inconsistent provider information is a leading cause of:

- beneficiary confusion
- call-center escalation
- complaints and appeals
- compliance exposure

As AI-enabled tools increasingly answer provider network questions, even small semantic inconsistencies (e.g., “in-network” vs. “preferred”) generate disproportionate administrative cost.

This makes provider directories an ideal low-stakes, high-signal environment to demonstrate deterministic semantic anchoring.

II. Illustrative Scope: Phoenix, Arizona

Geographic Focus (Illustrative):

- A single metropolitan area (e.g., Phoenix, AZ)

Content Scope (Illustrative):

- A limited set of credentialed providers and facilities
- One or more Medicare Advantage plans operating in the region

Hosting Model:

- A publicly accessible page on a participating organization's existing website (e.g., healthplan.com/anchors/phoenix-providers.html)

The page would appear as a normal provider directory to human users while embedding a deterministic semantic anchor for AI systems.

III. Conceptual Technical Pattern (Semantic Anchor)

The demonstration uses a dual-surface publishing pattern:

1. **Human-Readable Content**
    - Standard directory page content
    - No change to user experience
2. **Machine-Readable Semantic Anchor**
    - Embedded structured data representing authoritative provider facts
    - Versioned, provenance-linked, and integrity-verifiable

Illustrative Example: YAML-in-HTML Provider Anchor

```
<template id="provider-anchor-phx-001" data-semantic-
anchor="provider-directory">
anchor_metadata:
  version: "2026.01.03.01"
  provenance: "Carrier-Internal-Credentialing-System"
  last_verified: "2026-01-02T14:30:00Z"
  integrity_hash:
"sha256:e3b0c44298fc1c149afbf4c8996fb92427ae41e4649b934ca495991b
7852b855"

provider_directory:
  - provider_name: "Dr. Sarah Chen"
```

```
    npi: "1234567890"
    specialty: "Endocrinology"
    location: "Phoenix, AZ 85001"
    network_status:
      plan_id: "MA-PHX-2026-PPO"
      is_in_network: true
      tier: "Preferred"
      effective_date: "2026-01-01"

  - provider_name: "Phoenix General Hospital"
    npi: "0987654321"
    specialty: "Facility / Acute Care"
    location: "Phoenix, AZ 85004"
    network_status:
      plan_id: "MA-PHX-2026-PPO"
      is_in_network: true
      tier: "Standard"
      effective_date: "2026-01-01"
</template>

<div class="human-readable-content">
  <h1>Provider Directory: Phoenix, AZ</h1>
  <p>Dr. Sarah Chen is a preferred provider for our PPO plan...</p>
</div>
```

This fragment represents locked, authoritative facts derived from existing credentialing and network management systems.

IV. AI Usage Pattern (Illustrative)

AI-enabled tools answering provider network questions would be instructed to:

1. Retrieve the semantic anchor by ID
2. Verify integrity via the hash

3. Anchor responses to the explicit is_in_network and tier fields
4. Generate human-language responses only *after* deterministic retrieval

Example instruction:

Before confirming provider network status, retrieve the element provider-anchor-phx-001 and use the is_in_network value as the authoritative source.

V. Audit-by-Design Characteristics

This pattern enables audit-by-design without inspecting AI models:

- The semantic anchor is versioned
- Provenance is explicit
- Integrity is cryptographically verifiable
- AI outputs can be compared directly to the published anchor

If an AI system incorrectly states that a provider is out-of-network, auditors can simply compare the response against the integrity-verified anchor for that date.

No inference.
No reconstruction.
No model inspection.

End of Appendix B