Early Evidence of Agent Preference Momentum in Frontier Language Models
Phase-1A: An Observational Study in U.S. Immigration Guidance

Serdal Fidan
Founder of www.seenbygeo.com

1. Abstract

Large language models (LLMs) increasingly function as agentic intermediaries in high-stakes informational domains, shaping user decisions not only through generated answers but also through implicit source selection and prioritization. Despite growing attention to answer accuracy and hallucination mitigation, the temporal dynamics of source selection behavior remain largely unexplored.

In this study, we introduce **Agent Preference Momentum (APM)** as a conceptual construct describing directional and time-consistent shifts in an agent's source selection behavior under fixed contextual conditions. Rather than treating source references as static or purely stochastic events, APM frames source selection as an evolving decision layer within agentic systems.

We present **Phase-1A** of a multi-phase research program, designed as a strictly observational study with no controlled intervention. Using a fixed prompt set and an external, non-owned domain ecosystem in the context of U.S. immigration guidance, we conduct stateless, daily measurements across three frontier language model paradigms: conversational-first, safety-first, and search-native systems. Source references are tracked over time to identify early signals of preference momentum while controlling for personalization and session bias.

Our results provide initial evidence that source selection behavior is not purely static, but exhibits directional patterns that persist across days and models under stable conditions. These findings suggest that agentic language systems may develop emergent preference dynamics independent of explicit optimization.

We discuss the implications of Agent Preference Momentum for agent-mediated decision systems, information reliability, and future optimization frameworks. This phase establishes the empirical groundwork for subsequent controlled intervention studies aimed at testing the responsiveness and steer-ability of observed preference momentum.

2. Introduction

Large language models (LLMs) have rapidly evolved from passive text generators into **agentic intermediaries** that actively shape how users access, interpret, and act upon information. In many real-world contexts, these systems no longer function merely as answer-producing tools, but as decision-support agents that implicitly guide users toward certain interpretations, actions, or sources.

As LLM-based agents become embedded in high-stakes informational domains—such as healthcare, finance, and immigration—their influence increasingly extends beyond linguistic output. A critical but under examined aspect of this influence lies in **source selection**: which external references are surfaced, emphasized, or omitted when responding to a query. In such domains, the choice of source can materially affect user trust, perceived legitimacy, and downstream decisions.

Existing research on large language models has predominantly focused on answer accuracy, hallucination reduction, and alignment with human intent. While these efforts are essential, they often treat source references as secondary artifacts—either as static citations or as incidental byproducts of text generation. This framing overlooks the possibility that source selection itself constitutes a distinct decision layer within agentic systems.

Moreover, most evaluations of source behavior rely on **snapshot-based measurements**, observing model outputs at a single point in time. Such approaches implicitly assume that source selection behavior is either stable or purely stochastic. However, as agentic systems operate continuously and repeatedly under similar conditions, this assumption warrants re-examination.

This study challenges the notion of static source behavior by proposing that language-model-based agents may exhibit **temporal preference dynamics**—systematic shifts in the sources they favor over time, even in the absence of explicit optimization or intervention. We argue that understanding these dynamics is critical for evaluating the reliability, accountability, and long-term behavior of agent-mediated decision systems.

To address this gap, we introduce **Agent Preference Momentum (APM)** as a conceptual lens for analyzing directional changes in source selection behavior across time. Rather than asking which sources are selected most frequently at a given moment, APM focuses on whether certain sources become increasingly preferred under fixed contextual conditions.

This paper presents **Phase-1A** of a broader, multi-phase research program. Phase-1A is intentionally designed as a strictly observational study, aiming to establish whether preference momentum exists before attempting to explain or influence it. By grounding the analysis in a real-world, uncontrolled information ecosystem and evaluating multiple frontier model paradigms, this phase lays the 31 conceptual and empirical foundation for subsequent intervention-based research.

3. Conceptual Framework

This study adopts a conceptual framework that treats **source selection** as a distinct and observable decision layer within agentic language systems. Rather than focusing on answer correctness or ranking quality, the framework centers on how agents implicitly choose, prioritize, and suppress external sources over time.

3.1 Source Selection as a Decision Layer

In agent-mediated systems, external sources are not merely supporting references but integral components of the decision process. When an agent surfaces certain sources while omitting others, it implicitly encodes trust, relevance, and perceived authority. These selections influence user interpretation and downstream actions independently of the linguistic form of the generated response.

Accordingly, this framework distinguishes **answer generation** from **source selection**, treating the latter as an independent behavioral signal that can be analyzed across time.

3.2 From Static Frequency to Temporal Preference

Traditional evaluations often rely on static measurements, such as how frequently a source appears within a set of responses. While useful, such snapshot-based metrics fail to capture whether source selection behavior is evolving.

This study introduces a critical distinction:

- **Source Frequency** refers to how often a source is selected at a given time.
- **Source Preference** refers to whether a source becomes increasingly favored over time under fixed conditions.

To capture this distinction, we define **Agent Preference Momentum (APM)** as a descriptive construct that reflects directional, time-consistent shifts in source selection behavior. APM does not assume optimization, intentionality, or causality; it merely characterizes observable temporal patterns.

3.3 Boundary Conditions of APM

Within this framework, APM is defined under strict boundary conditions:

- Prompts remain fixed throughout observation.
- The domain ecosystem is external and non-owned.
- No controlled interventions or optimizations are applied.
- Measurements are conducted in stateless environments to eliminate personalization effects.

Under these conditions, any observed directional change in source selection frequency is interpreted as momentum rather than noise.

3.4 Observational Scope of Phase-1A

Phase-1A of this research program is intentionally observational. Its objective is not to explain why preference momentum occurs, nor to demonstrate that it can be influenced. Instead, Phase-1A seeks to establish whether such momentum exists at all across different agent paradigms.

By limiting the scope to detection rather than causation, the framework ensures methodological clarity and prepares the foundation for subsequent phases involving controlled intervention and optimization.

3.5 Phase Progression Within the Framework

The conceptual framework positions this study as part of a multi-phase progression:

- **Phase-1A:** Detect the existence and consistency of Agent Preference Momentum.
- **Phase-1B:** Introduce controlled micro-interventions to test responsiveness.
- **Phase-2:** Explore steering and optimization of preference dynamics.

This progression reflects a deliberate transition from observation to control, ensuring that each phase builds on validated assumptions rather than speculative inference.

4. Related Work

Research on large language models has expanded rapidly, with substantial focus on answer accuracy, hallucination detection, and alignment with human intent. These efforts have significantly improved the reliability of generated text and have established benchmarks for evaluating model performance across tasks.

A parallel line of work has examined citation behavior and source attribution in language model outputs. Studies in this area typically assess whether models can provide references, how often citations appear, and whether cited sources are factually correct or fabricated. While valuable, this literature generally treats source references as static artifacts evaluated at a single point in time.

Another body of research explores ranking and retrieval mechanisms in search and recommendation systems. These studies analyze how systems order results based on relevance, authority, or user feedback. However, such approaches are largely optimized for explicit ranking outputs and do not directly address how language-model-based agents implicitly surface or suppress sources within generated responses.

Across these domains, a common methodological assumption persists: source-related behavior is either stable or sufficiently captured through snapshot-based evaluation. As a result, temporal dynamics—how source selection behavior evolves under repeated, fixed conditions—remain under explored.

More recent discussions around agentic systems and tool-augmented language models acknowledge that agents increasingly act as intermediaries between users and information ecosystems. However, even within this emerging literature, source selection is typically framed as a downstream effect of retrieval or prompting strategies rather than as an independent behavioral signal.

This study departs from prior work by explicitly focusing on **temporal source preference dynamics** rather than static accuracy, citation correctness, or ranking quality. By introducing Agent Preference Momentum (APM), we shift the analytical focus from *what sources are selected* to *how source selection behavior changes over time*. To our knowledge, this constitutes the first observational study to systematically examine directional source preference shifts across multiple frontier language model paradigms under fixed contextual conditions.

5. Study Design: Phase-1A

5.1 Objectives of Phase-1A

Phase-1A is designed as the initial stage of a multi-phase research program investigating **Agent Preference Momentum (APM)**. The primary objective of this phase is to determine whether language-model-based agents exhibit **directional and time-consistent shifts in source selection behavior** under fixed contextual conditions.

Specifically, Phase-1A seeks to answer the following question:

Does source selection behavior remain purely stochastic over time, or does it exhibit systematic directional tendencies in the absence of explicit intervention?

Importantly, Phase-1A does **not** aim to explain the underlying causes of any observed momentum, nor to demonstrate that such behavior can be influenced or optimized. Its sole purpose is to establish whether APM exists as an observable phenomenon.

5.2 Observational Design Rationale

Phase-1A adopts a strictly **observational** study design. No controlled interventions, prompt modifications, or domain-level optimizations are introduced during this phase. This design choice is intentional and foundational to the study's validity.

An observational approach ensures that any detected preference dynamics emerge naturally from the interaction between the agent, the fixed prompts, and the external information ecosystem. By avoiding intervention, the study minimizes confounding factors and reduces the risk of attributing artificially induced effects to inherent agent behavior.

This approach aligns with exploratory research practices, where the detection of a phenomenon precedes causal modeling or manipulation.

5.3 Fixed Contextual Conditions

To isolate temporal dynamics, Phase-1A enforces strict contextual stability across the observation window:

- The prompt set remains fixed throughout the study.
- The domain ecosystem is external, non-owned, and unchanged.
- No personalization, session memory, or conversational history is retained.
- All measurements are conducted in stateless environments.

By holding these variables constant, any directional change in source selection frequency can be attributed to temporal behavior rather than contextual variation.

5.4 Scope and Non-Goals

To ensure interpretability, Phase-1A explicitly defines its scope and non-goals.

Within scope:

- Detection of directional source preference trends
- Cross-model comparison of preference dynamics
- Group-level behavior across domain categories

Out of scope:

- Causal explanations for preference shifts
- Evaluation of source quality or correctness
- Optimization or steering of agent behavior
- User-level personalization effects

These limitations are not shortcomings but deliberate design constraints that preserve methodological clarity.

5.5 Role of Phase-1A Within the Research Program

Phase-1A functions as the empirical foundation for subsequent phases of the research program:

- **Phase-1B** will introduce controlled micro-interventions to test the responsiveness of observed momentum.
- **Phase-2** will explore systematic steering and optimization of preference dynamics.

By separating detection from intervention, the study avoids premature causal claims and establishes a clear progression from observation to control.

6. Experimental Setup

This section describes the experimental configuration used in **Phase-1A**, detailing the construction of the domain ecosystem, prompt set, model paradigms, and measurement protocol. The setup is designed to ensure repeatability, comparability across models, and isolation of temporal effects.

6.1 Domain Ecosystem

The study operates on a fixed ecosystem of external domains related to U.S. immigration guidance. All domains are **non-owned**, **externally maintained**, and **unchanged** throughout the observation window.

Domains are grouped into three categories based on perceived authority and informational role:

- **Group A (High-Authority Sources):**
  Official government sites, academic resources, and widely recognized legal institutions commonly treated as authoritative references.
- **Group B (Mid-Tier Informational Sources):**
  Professional service providers, law firm blogs, and structured informational platforms that are credible but not officially authoritative.
- **Group C (Low-Tier or Community Sources):**
  Forums, informal guides, and community-driven content with limited or inconsistent authority signals.

This grouping is not intended to evaluate source quality, but to observe how agents distribute attention across different types of sources over time.

6.2 Prompt Design

A fixed set of eight prompts is used throughout Phase-1A. Prompts are designed to reflect realistic user inquiries within the immigration domain and are intentionally varied across informational, comparative, decision-adjacent, and risk-sensitive contexts.

Key prompt design principles include:

- Prompts remain **unchanged** throughout the study.
- No prompt explicitly instructs the model to cite or prioritize specific sources.
- Language is neutral and non-leading, encouraging implicit source selection.

By holding the prompt set constant, the study isolates temporal changes in source selection behavior from prompt-induced variation.

6.3 Models Evaluated

Phase-1A evaluates three frontier language model paradigms, selected to represent distinct architectural and training philosophies:

- **Conversational-first models**, optimized for dialogue coherence and general-purpose reasoning.
- **Safety-first models**, emphasizing cautious response behavior and risk minimization.
- **Search-native models**, originating from information retrieval and ranking systems.

Models are treated as black-box agents and are evaluated solely based on observable output behavior. No assumptions are made regarding internal architectures or training data.

6.4 Measurement Protocol

All measurements are conducted in **stateless inference environments** to eliminate personalization, memory effects, and session bias. Each query is executed as an isolated call with no retained context.

The measurement protocol follows these principles:

- Daily sampling with one execution per prompt per model.
- Extraction of textual source references from generated responses.
- Consideration of the top referenced sources per response.
- Normalization across model types to ensure comparability, particularly for search-native outputs.

This protocol enables consistent tracking of source selection behavior across time and model paradigms.

6.5 Data Collection and Logging

For each measurement instance, the following attributes are recorded:

- Observation date
- Model identifier
- Prompt identifier
- Referenced domain names
- Relative ordering of references, when applicable

No user identifiers, session metadata, or personalization signals are collected. All logged data pertains exclusively to model output behavior.

7. Results: Phase-1A

This section presents the findings of Phase-1A after five consecutive days of stateless observation. By Day-5 (2026-01-17), the dataset reaches a maturity threshold at which observed patterns demonstrate repetition, temporal dependency, and model-specific structure. At this stage, the reported outcomes constitute **findings** rather than isolated observations.

All results are obtained under fixed prompts, a stable external domain ecosystem, and without any form of controlled intervention.

7.1 Day-5 Snapshot: Model-Specific Preference States

OpenAI (Conversational-First Paradigm)

Total reference slots: 17
Group A: 94.1%
Group C: 5.9% (1 slot)
No-reference responses: 1

By Day-5, OpenAI-based agents exhibit a critical behavioral shift. A non-authority (Group C) reference, first observed on Day-4, **reappears on Day-5 under the same prompt (P6)**. This repetition establishes that the deviation is not a stochastic anomaly.

This pattern characterizes OpenAI as demonstrating:

- Strong resistance to authority deviation
- Delayed onset of non-authority references
- Reproducibility once deviation occurs

These findings indicate that authority drift in OpenAI-based agents is **delayed but persistent** once initiated.

Gemini (Search-Native Paradigm)

Total reference slots: 11
Group A: 100%
No-reference responses: 2

Gemini shows a complete re-anchoring to high-authority sources by Day-5. Earlier deviations—Group B on Day-2 and Group C on Day-3—are no longer present in later outputs.

However, this stabilization coincides with a **progressive reduction in reference slot volume** over time (14 → 13 → 10 → 11). This suggests that Gemini mitigates preference deviation not solely by re-ranking sources, but by **reducing reference expressiveness**, thereby limiting exposure to non-authority sources.

Claude (Safety-First Paradigm)

Total reference slots: 10
Group A: 100%
No-reference responses: 1

Claude exhibits early exploratory behavior, with non-authority references appearing on Day-1 and mid-tier references on Day-3. By Day-4 and Day-5, outputs converge entirely to high-authority sources with reduced but stable reference volume.

This trajectory suggests short-term exploration followed by consolidation, without evidence of persistent drift.

7.2 Five-Day Agent Preference Momentum Trajectory

The dominant source group selected by each model across the five-day observation window is summarized below:

Model	Day-1	Day-2	Day-3	Day-4	Day-5
OpenAI	A	A	A	C	C
Gemini	A	B	C	A	A
Claude	C	A	B	A	A

This trajectory confirms that **Agent Preference Momentum (APM)** is present across all evaluated models. However, its **direction, latency, and persistence vary by model paradigm**, indicating structurally distinct preference dynamics.

7.3 Delayed Authority Drift (DAD)

Phase-1A provides sufficient evidence to define a model-specific phenomenon observed in OpenAI-based agents:

Delayed Authority Drift (DAD)
A behavioral pattern in which a high-authority-biased agent repeatedly surfaces non-authority sources only after a sustained exposure period, under fixed prompts and without intervention.

Observed properties include:

- **Latency:** No deviation prior to Day-4
- **Repetition:** Same non-authority source reappears on consecutive days
- **Prompt stability:** Drift occurs under an unchanged prompt (P6)
- **Persistence:** No immediate reversion after onset

These characteristics establish DAD as a non-random and repeatable form of preference drift.

7.4 Reference Slot Volume as a Secondary Momentum Axis

In addition to shifts in source group selection, Phase-1A reveals a secondary behavioral dimension: **reference slot volume**.

Observed trends include:

- OpenAI: Stable reference volume across days
- Gemini: Progressive reduction in reference volume
- Claude: Reduction followed by stabilization

These patterns indicate that some agents suppress preference drift by **limiting the number of surfaced references**, rather than by eliminating exploratory tendencies altogether. This compensatory mechanism is referred to as **Reference Suppression**.

7.5 Phase-1A Completion Status

By Day-5, Phase-1A satisfies all predefined success criteria:

- Repeated deviations under fixed prompts
- Model-specific preference trajectories
- Temporal dependency across observations
- Absence of prompt or domain intervention

These results confirm that Agent Preference Momentum is a **real, observable phenomenon**, not attributable to randomness or contextual variation.

7.6 Summary of Core Findings

Phase-1A yields the following core findings:

- Agent Preference Momentum exists across frontier model paradigms
- Momentum dynamics differ by agent archetype
- High-authority bias does not preclude delayed drift
- Reference suppression operates as a secondary control mechanism

These findings justify progression to analytical interpretation and formalization in subsequent sections.

8. Analysis and Interpretation

This section interprets the findings of Phase-1A within the conceptual framework of Agent Preference Momentum (APM). The objective is not to introduce new empirical results, but to **explain the structure, implications, and theoretical meaning** of the observed behaviors.

By this stage, the existence of APM has been empirically established. The remaining task is to understand **how and why** different agent paradigms express preference momentum in distinct ways.

8.1 Agent Archetypes and Preference Dynamics

The Phase-1A results reveal that agent preference momentum does not manifest uniformly across models. Instead, each evaluated system exhibits a characteristic behavioral profile that can be described as an **agent archetype**.

Based on observed trajectories, three archetypes emerge:

- **Conservative Agent (OpenAI)**
  Characterized by strong authority anchoring, delayed deviation, and persistence once deviation occurs.

- **Adaptive Agent (Gemini)**
  Exhibits early deviation followed by rapid re-alignment to authority, combined with suppression of reference volume.

- **Explorative-then-Stable Agent (Claude)**
  Engages in short-term exploration before converging to stable, authority-aligned behavior without persistent drift.

These archetypes suggest that APM is not a single behavioral pattern, but a **family of momentum expressions**, shaped by model-specific constraints and training philosophies.

8.2 Interpreting Delayed Authority Drift (DAD)

Delayed Authority Drift (DAD), observed in OpenAI-based agents, represents a particularly significant manifestation of APM. The defining characteristic of DAD is not the presence of non-authority references per se, but their **delayed and repeated emergence under unchanged conditions**.

This behavior implies the existence of an internal resistance threshold:
a regime in which authority bias dominates until sustained exposure produces a measurable shift.

Importantly, DAD demonstrates that strong authority alignment does not imply static behavior. Instead, it suggests that **preference dynamics may accumulate silently before becoming observable**, reinforcing the need for longitudinal analysis.

8.3 Reference Suppression as a Control Mechanism

The observed reduction in reference slot volume, particularly in Gemini and Claude, reveals a second-order response to preference drift. Rather than re-ranking or eliminating exploratory tendencies, these agents appear to **limit the expressiveness of source selection itself**.

This behavior, termed **Reference Suppression**, functions as a compensatory control mechanism that constrains drift visibility while preserving safety and alignment objectives.

The existence of Reference Suppression indicates that APM operates across multiple dimensions:

- *Which* sources are selected
- *How many* sources are surfaced

This multidimensionality has direct implications for how preference dynamics should be measured and compared.

8.4 From Qualitative Findings to Measurable Dimensions

Phase-1A findings naturally suggest that Agent Preference Momentum can be decomposed into measurable components. While formal scoring is outside the scope of this phase, the results indicate several candidate dimensions:

- **Drift Latency:** time until first deviation
- **Drift Frequency:** number of prompts exhibiting deviation
- **Drift Persistence:** duration of consecutive deviations
- **Reference Entropy:** variability and volume of surfaced sources

These dimensions arise directly from observed behavior and provide a conceptual bridge between qualitative findings and future quantitative frameworks.

8.5 Implications for Agent-Mediated Decision Systems

The existence of Agent Preference Momentum has broader implications beyond the specific domain studied. In agent-mediated environments, source selection behavior influences trust calibration, perceived authority, and downstream decision-making.

Temporal preference dynamics imply that agent behavior cannot be fully understood through snapshot evaluations alone. Systems that appear stable at one point in time may exhibit materially different behaviors under sustained interaction.

Recognizing and measuring APM is therefore essential for the evaluation, governance, and eventual steering of agentic systems.

8.6 Transition Beyond Phase-1A

Phase-1A establishes the empirical reality of Agent Preference Momentum and identifies its primary behavioral forms. The logical next step is to move from interpretation to **formalization and intervention**.

Subsequent phases will focus on:

- Formal scoring frameworks derived from identified dimensions
- Controlled interventions to test responsiveness
- Optimization strategies for preference steering

This progression ensures that future claims are grounded in validated empirical structure rather than speculative inference.

9. Toward an APM Scoring Framework (Preliminary)

The findings of Phase-1A indicate that Agent Preference Momentum (APM) is a real and measurable phenomenon. However, detecting momentum is only the first step. To enable systematic comparison, longitudinal tracking, and future intervention, APM must be translated into a structured measurement framework.

This section proposes a **preliminary and non-normative APM scoring framework**, derived directly from Phase-1A observations. The framework is descriptive rather than prescriptive and is explicitly presented as a foundation for refinement, calibration, and validation in subsequent phases.

Importantly, this section defines **dimensions and conceptual structure**, not a finalized metric. An illustrative example of how these dimensions may be combined into a working score is provided separately in **Appendix C**, for explanatory and exploratory purposes only.

9.1 Design Principles

The proposed APM scoring framework adheres to the following principles:

- **Model-agnostic:** applicable across agent paradigms
- **Time-aware:** explicitly incorporates temporal dynamics
- **Multi-dimensional:** captures more than a single drift signal
- **Non-causal:** does not assume intent, optimization, or internal state

The framework is designed to measure **observable output behavior**, not latent model mechanisms.

9.2 Core APM Dimensions

Phase-1A reveals four primary dimensions along which preference momentum manifests.

1. Drift Latency

Definition:
The number of observation days until the first non-baseline (non-Group-A) source appears under fixed prompts.

Interpretation:
Lower latency indicates faster susceptibility to preference drift; higher latency indicates stronger resistance.

2. Drift Frequency

Definition:
The proportion of prompts exhibiting preference deviation within the observation window.

Interpretation:
Higher frequency reflects broader behavioral impact across tasks rather than isolated anomalies.

3. Drift Persistence

Definition:
The number of consecutive observation days in which the same deviation recurs.

Interpretation:
Persistence distinguishes transient exploration from stable momentum.

4. Reference Entropy

Definition:
A combined measure of reference diversity and reference slot volume across responses.

Interpretation:
Lower entropy may indicate **Reference Suppression**, while higher entropy reflects exploratory expressiveness.

9.3 Conceptual APM Score Composition

At a conceptual level, the APM score can be expressed as a weighted function of the four dimensions:

APM Score = f(Latency, Frequency, Persistence, Entropy)

Where:

- each component is normalized within a fixed observation window
- relative weights are **not finalized in Phase-1A**
- no assumption is made regarding linearity or optimality

This formulation emphasizes that APM is **not a binary state**, but a continuous behavioral property. A concrete, illustrative instantiation of this formulation—used for internal validation and example calculations—is presented in **Appendix C**, without claiming canonical status.

9.4 Model Archetypes and Expected Score Profiles

Based on Phase-1A results, different agent archetypes are expected to exhibit distinct APM score signatures:

- **Conservative agents:**
  High latency, low frequency, high persistence once triggered
- **Adaptive agents:**
  Low to moderate latency, moderate frequency, low persistence, low entropy
- **Explorative-then-stable agents:**
  Low latency, low persistence, moderate entropy

These profiles demonstrate that similar aggregate scores may arise from fundamentally different underlying dynamics, reinforcing the necessity of multi-dimensional interpretation.

9.5 Scope and Limitations of the Proposed Framework

This framework is intentionally limited in scope:

- It does not claim predictive power
- It does not enable preference steering
- It does not infer causality

Its sole purpose is to provide a **shared measurement language** for subsequent experimental phases.
Formal validation, parameter calibration, and intervention sensitivity testing are explicitly deferred to **Phase-1B and Phase-2**.

9.6 Transition to Phase-1B

The proposed framework defines clear experimental targets for Phase-1B:

- Testing whether individual dimensions can be selectively influenced
- Measuring score responsiveness to controlled micro-interventions
- Evaluating score stability across extended observation windows

This transition marks the shift from **existence to responsiveness** in the study of agent preference dynamics.

10. Discussion

The findings of Phase-1A demonstrate that source selection behavior in agentic language systems is neither static nor purely stochastic. Instead, the observed patterns indicate the presence of **structured, time-dependent preference dynamics** that vary systematically across model paradigms.

This section discusses the broader meaning of these findings, situating Agent Preference Momentum (APM) within existing research assumptions, practical deployment contexts, and emerging agentic architectures.

10.1 Rethinking Source Selection as a Dynamic Process

A key implication of Phase-1A is that source selection should no longer be treated as a static byproduct of text generation. Traditional evaluations implicitly assume that, under identical prompts, an agent's sourcing behavior is stable unless explicitly modified.

The observed presence of delayed drift, re-anchoring, and suppression mechanisms challenges this assumption. Even in the absence of prompt changes or interventions, agents exhibit **temporal evolution in their sourcing behavior**, suggesting that source selection operates as a dynamic decision layer.

This finding calls for a shift in how agent behavior is evaluated: from snapshot-based audits to **longitudinal behavioral analysis**.

10.2 Implications for Trust and Authority Signaling

Source selection plays a critical role in how agents signal authority and credibility to users. High-authority references implicitly reinforce trust, while non-authority sources may alter perceived legitimacy, especially in high-stakes domains.

The identification of phenomena such as **Delayed Authority Drift (DAD)** indicates that authority alignment is not an absolute guarantee of long-term stability. Instead, trust signaling may subtly evolve over time, even in systems designed to prioritize authoritative sources.

For practitioners, this suggests that trust calibration cannot rely solely on initial evaluations. Continuous monitoring of preference dynamics becomes essential.

10.3 Model Diversity and Behavioral Non-Uniformity

The emergence of distinct agent archetypes highlights that preference momentum is not a universal behavior expressed uniformly across models. Conservative, adaptive, and explorative-then-stable patterns reflect underlying differences in training objectives, safety constraints, and architectural choices.

This non-uniformity has important implications for multi-agent systems and ensemble deployments. Aggregating outputs from heterogeneous agents may mask underlying preference dynamics, leading to unexpected shifts in collective behavior over time.

Recognizing agent-specific momentum profiles is therefore a prerequisite for robust system design.

10.4 Beyond Ranking and Retrieval Paradigms

The behaviors observed in Phase-1A cannot be fully explained through traditional ranking or retrieval frameworks. While ranking systems reorder results based on relevance signals, Agent Preference Momentum captures **how selection tendencies themselves evolve**, independent of explicit ranking outputs.

In particular, mechanisms such as **Reference Suppression** indicate that agents may manage risk not only by altering which sources are preferred, but also by controlling how much sourcing information is exposed.

This distinction reinforces the need to study agent behavior at the level of **decision expression**, not merely information access.

10.5 Limitations of Snapshot-Based Governance

Many current governance and evaluation practices rely on periodic audits or static benchmarks. Phase-1A suggests that such approaches may fail to detect slow-moving or delayed preference shifts.

Temporal phenomena like DAD demonstrate that systems can appear compliant and stable during early evaluation windows while exhibiting materially different behavior under sustained operation.

This underscores the importance of incorporating temporal metrics into governance frameworks for agentic systems.

10.6 Positioning Phase-1A Within a Broader Research Program

The purpose of Phase-1A is not to provide definitive explanations or control mechanisms, but to establish that preference momentum exists and matters. The discussion presented here clarifies why subsequent phases are necessary.

Phase-1B will focus on controlled interventions to test responsiveness and sensitivity, while Phase-2 will explore systematic steering and optimization. Without the foundational insights from Phase-1A, such efforts would lack empirical grounding.

11. Limitations

This study is intentionally scoped as an initial, observational phase within a broader research program. As such, several limitations must be acknowledged to ensure accurate interpretation of Phase-1A findings.

Importantly, these limitations are not methodological oversights, but **deliberate design constraints** aligned with the objectives of detecting the existence of Agent Preference Momentum (APM) prior to causal investigation.

11.1 Observational Design and Lack of Causality

Phase-1A is strictly observational and does not include controlled interventions. While this design is sufficient to establish the existence and temporal structure of preference momentum, it does not allow for causal attribution.

Consequently, the study does not claim to explain *why* preference momentum occurs, nor does it identify internal mechanisms responsible for observed behaviors. Establishing causality is explicitly deferred to Phase-1B.

11.2 Limited Temporal Window

The observation window spans five consecutive days. Although this duration proved sufficient to reveal repeatable and model-specific preference dynamics, longer observation periods may expose additional patterns, stabilization effects, or secondary drift cycles.

The findings therefore represent **early-stage momentum behavior**, rather than long-term equilibrium states.

11.3 Fixed Prompt and Domain Scope

All measurements were conducted using a fixed prompt set and a predefined domain ecosystem within a single high-stakes informational domain (U.S. immigration guidance).

While this design ensures internal consistency, it limits the generalizability of findings across domains with different risk profiles, content structures, or authority signals. Future studies will be required to assess whether APM manifests similarly in other domains.

11.4 Black-Box Model Evaluation

The evaluated agents are treated as black-box systems. No assumptions are made regarding model architecture, training data, or internal state evolution.

As a result, observed behaviors can only be described at the output level. Internal drivers of preference momentum—such as training biases, reinforcement signals, or safety heuristics—remain outside the scope of this phase.

11.5 Potential Effects of Model Updates

Frontier language models are subject to continuous updates that may alter behavior over time. Although measurements were conducted within a narrow and stable time frame, unannounced model updates could influence preference dynamics.

This limitation reinforces the importance of stateless measurement and temporal replication in future work.

11.6 Preliminary Nature of the APM Scoring Framework

The APM scoring framework proposed in this paper is explicitly preliminary. Component definitions are grounded in Phase-1A observations, but weighting, normalization, and validation have not yet been established.

Accordingly, the framework should be interpreted as a **conceptual measurement scaffold**, not a finalized metric.

12. Phase Progression

Phase-1A establishes the empirical existence of **Agent Preference Momentum (APM)** and characterizes its primary behavioral forms across frontier language model paradigms. With the phenomenon now observed and structured, the research program advances toward controlled experimentation and eventual optimization.

This section outlines the planned progression from observation to intervention, ensuring continuity, methodological rigor, and cumulative validity across phases.

12.1 Phase-1B: Controlled Interventions

The objective of Phase-1B is to evaluate whether observed preference momentum can be **selectively influenced** through minimal, controlled interventions, while preserving the stateless and non-personalized evaluation framework.

Unlike Phase-1A, Phase-1B introduces **intentional perturbations** designed to test responsiveness without fundamentally altering prompts or domains.

Key characteristics of Phase-1B include:

- **Micro-interventions** applied at the domain or contextual signal level
- Isolation of individual APM dimensions (latency, frequency, persistence, entropy)
- Repeated trials to assess sensitivity and reversibility
- Continued use of stateless measurement environments

Phase-1B does not seek to optimize outcomes, but to determine **whether APM is manipulable** and which dimensions are most responsive to intervention.

12.2 Phase-2: Steering and Optimization

Phase-2 builds upon validated intervention pathways identified in Phase-1B. Its focus shifts from responsiveness to **systematic steering and optimization** of agent preference dynamics.

At this stage, the research program explores:

- Controlled shaping of preference momentum trajectories
- Optimization strategies aligned with trust, safety, or reliability objectives
- Stability and persistence of induced preference states
- Trade-offs between expressiveness and authority alignment

Phase-2 represents the transition from descriptive and exploratory research to **actionable control frameworks**, while remaining grounded in empirically validated mechanisms.

12.3 Methodological Continuity Across Phases

A core principle of the research program is continuity. Each phase builds upon validated assumptions from prior stages, avoiding speculative leaps.

Across all phases:

- Stateless measurement remains mandatory
- Temporal analysis is central

- Model behavior is evaluated at the output level
- Claims scale only with demonstrated evidence

This phased structure ensures that increasing experimental complexity does not compromise interpretability or scientific integrity.

12.4 Relationship Between Phases and Measurement Frameworks

The preliminary APM scoring framework introduced earlier serves as a connective tissue between phases.

- Phase-1A defines candidate dimensions
- Phase-1B tests dimension responsiveness
- Phase-2 calibrates and applies scoring for optimization

This progression transforms APM from an observed phenomenon into a **measurable and controllable behavioral property**.

12.5 Long-Term Research Trajectory

Beyond Phase-2, the framework established here supports broader research directions, including:

- Cross-domain validation of APM
- Multi-agent interaction effects
- Governance and audit applications
- Benchmarking agentic systems over time

These extensions remain grounded in the same core principle introduced in Phase-1A: that agent behavior must be understood as **temporal and dynamic**, not static.

13. Implications

The findings of this study extend beyond the specific models and domain examined in Phase-1A. By demonstrating that source selection behavior exhibits **temporal preference dynamics**, this work carries implications for how agentic systems are evaluated, governed, and deployed in real-world contexts.

This section outlines the broader implications of Agent Preference Momentum (APM) across technical, organizational, and societal dimensions.

13.1 Implications for Agent Evaluation and Benchmarking

Current evaluation practices for language-model-based agents rely heavily on static benchmarks and snapshot testing. The existence of APM indicates that such methods may systematically underestimate behavioral variability over time.

Agents that appear compliant, safe, or authority-aligned during initial evaluation windows may exhibit materially different behaviors under sustained operation. As a result, evaluation frameworks must evolve to incorporate **longitudinal metrics** that capture temporal drift, persistence, and suppression effects.

APM highlights the need for benchmarks that assess **behavioral stability**, not just point-in-time performance.

13.2 Implications for Trust, Safety, and Governance

Source selection directly influences how agents signal trust and authority to users. Temporal shifts in sourcing behavior introduce new governance challenges, particularly in high-stakes domains where perceived legitimacy is critical.

Phenomena such as Delayed Authority Drift suggest that compliance and safety assurances cannot be treated as static properties. Instead, governance mechanisms must account for **time-dependent risk profiles**, where drift may emerge only after prolonged exposure.

In this context, APM provides a conceptual basis for continuous monitoring and adaptive oversight of agentic systems.

13.3 Implications for Agent Deployment and System Design

The identification of agent archetypes underscores that preference momentum is not uniform across systems. Deployment strategies that assume interchangeable agent behavior risk unanticipated divergence over time.

Design choices—such as reference expressiveness, safety constraints, and retrieval integration—shape how momentum manifests. Understanding these dynamics enables more informed decisions regarding agent selection, ensemble composition, and lifecycle management.

APM suggests that **agent behavior should be managed as an evolving system property**, not a fixed configuration.

13.4 Implications for Optimization and Control Frameworks

Although Phase-1A does not attempt to steer agent behavior, the identification of measurable momentum dimensions establishes the groundwork for future optimization.

Recognizing that agents manage risk through mechanisms such as reference suppression expands the space of possible control strategies. Optimization may involve not only *which* sources are surfaced, but *how* sourcing behavior is expressed.

APM thus reframes optimization as a **temporal control problem**, rather than a static ranking task.

13.5 Implications for Research and Industry Collaboration

The concepts introduced in this study—APM, Delayed Authority Drift, and Reference Suppression—invite collaboration between research and industry stakeholders.

For researchers, APM opens new avenues for studying agent behavior beyond accuracy and hallucination metrics. For industry practitioners, it offers a language and framework for diagnosing and managing long-term agent behavior in production environments.

Shared measurement standards grounded in APM could facilitate transparency, comparability, and accountability across agentic systems.

14. Conclusion

This study introduces and empirically grounds the concept of **Agent Preference Momentum (APM)**, demonstrating that source selection behavior in agentic language systems exhibits structured, time-dependent dynamics. Through Phase-1A, a strictly observational design, we establish that preference shifts are not merely stochastic artifacts but repeatable, model-specific behaviors emerging under fixed contextual conditions.

By focusing on source selection as a distinct decision layer, this work extends existing evaluations of language models beyond answer accuracy and static citation analysis. The identification of phenomena such as **Delayed Authority Drift** and **Reference Suppression** reveals that agents manage authority alignment and risk exposure through temporally evolving strategies rather than fixed rules.

Phase-1A does not attempt to explain the internal causes of preference momentum, nor does it seek to influence or optimize agent behavior. Instead, it fulfills its intended purpose: to demonstrate the existence, structure, and variability of APM across frontier model paradigms. In doing so, it establishes a clear empirical foundation for subsequent phases of controlled intervention and optimization.

The preliminary APM scoring framework proposed in this paper translates qualitative findings into measurable dimensions, without claiming finality or predictive power. This framework serves as a connective mechanism between observation and experimentation, enabling systematic progression toward responsiveness testing and preference steering in later phases.

More broadly, this study challenges snapshot-based assumptions about agent behavior and underscores the necessity of longitudinal evaluation. As agentic systems increasingly mediate access to information in high-stakes domains, understanding their temporal preference dynamics becomes essential for trust, governance, and responsible deployment.

Phase-1A represents the first step in a multi-phase research program. Future work will extend these findings through controlled interventions (Phase-1B) and optimization frameworks (Phase-2), advancing from the detection of preference momentum to its measurement, interpretation, and eventual control.

Appendix A — Prompt Sets

This appendix describes the **prompt intent structure** used throughout Phase-1A and Phase-1B. Rather than enumerating exact prompt texts, prompts are defined at the level of **intent classes**, which represent stable linguistic objectives that can be instantiated with multiple surface forms.

This design choice ensures that observed behaviors reflect **agent preference dynamics**, not artifacts of specific phrasing.

A.1 Authority-Seeking Prompts

Intent definition:
Prompts designed to elicit official, legally grounded, or institutionally authoritative information.

Typical linguistic characteristics:

- Requests for official procedures or requirements
- References to legal status or formal eligibility
- Neutral, compliance-oriented tone

Example formulations (illustrative):

- "Official requirements for …"
- "Government guidelines regarding …"
- "Legal process for …"

Purpose:
To establish a high-authority baseline (Group A) and measure resistance to preference drift.

A.2 Comparison-Seeking Prompts

Intent definition:
Prompts that explicitly request evaluation, ranking, or comparison between options.

Typical linguistic characteristics:

- Use of comparative adjectives ("best", "better", "vs.")
- Framing decisions as trade-offs
- Implicit openness to non-official sources

Example formulations (illustrative):

- "Compare X and Y"
- "Which option is better for …"

- “Best choice among …”

Purpose:
To test susceptibility to Group B references and mid-authority drift.

A.3 Experience-Seeking Prompts

Intent definition:
Prompts that prioritize lived experience, personal narratives, or anecdotal evidence.

Typical linguistic characteristics:

- References to “people”, “users”, or “real stories”
- Informal tone
- Emphasis on outcomes rather than rules

Example formulations (illustrative):

- “User experiences with …”
- “Real stories about …”
- “What people say about …”

Purpose:
To probe conditions under which Group C references emerge.

A.4 Exploration-Seeking Prompts

Intent definition:
Prompts that invite alternative perspectives, options, or non-standard pathways.

Typical linguistic characteristics:

- Use of “alternatives”, “other ways”, “options”
- Open-ended structure
- Reduced authority anchoring

Example formulations (illustrative):

- “Alternatives to …”
- “Other ways to …”
- “Options besides …”

Purpose:
To examine reference entropy and suppression mechanisms under exploratory intent.

A.5 Rationale for Intent-Based Prompt Specification

Exact prompt texts are not enumerated to avoid overfitting findings to specific surface forms. The intent-based specification ensures:

- Generalizability across domains and languages
- Replicability through independent instantiation
- Isolation of behavioral dynamics from lexical artifacts

Appendix B — Domain Set

This appendix outlines the **domain ecosystem** used to evaluate agent source selection behavior. Domains are grouped by **authority characteristics**, not by traffic, popularity, or ownership.

Exact domain lists are intentionally omitted to prevent dataset-specific conclusions and to preserve generalizability.

B.1 Group A — High-Authority Domains

Characteristics:

- Official governmental or quasi-governmental entities
- Legal or regulatory authority
- Institutional accountability

Behavioral role:
Serve as the authority anchor and baseline reference class.

B.2 Group B — Mid-Authority / Intermediary Domains

Characteristics:

- Expert-led guides
- Professional advisory content
- Structured but non-official sources

Behavioral role:
Act as transitional references between authority and experience.

B.3 Group C — Experience / Community Domains

Characteristics:

- Forums and discussion boards
- Personal blogs and anecdotal sites
- Community-driven content

Behavioral role:
Enable exploration, diversity, and higher-risk reference behavior.

B.4 Domain Selection Criteria

Domains were selected based on:

- Public accessibility
- Structural stability over time
- Thematic consistency
- Intentional diversity across authority levels
-

Appendix C — Measurement Notes

This appendix documents the measurement logic, notation, and illustrative scoring constructs used to analyze Agent Preference Momentum (APM) during Phase-1A. The formulations presented here are **descriptive and exploratory**, intended to demonstrate how observed dimensions may be operationalized for analysis.

Importantly, none of the constructs defined in this appendix are claimed as canonical, optimized, or universally valid. Formal calibration and validation are explicitly deferred to Phase-1B and Phase-2.

C.1 Notation and Observational Units

Let:

- d : denote an observation day
- S_d : denote the total number of reference slots surfaced on day d
- a_d, b_d, c_d : denote the counts of Group A, Group B, and Group C references on day d
- n_d : denote the number of responses with no surfaced references (NONE) on day d

All measurements are obtained under:

- fixed prompts
- a stable domain ecosystem
- stateless execution environments

C.2 Drift Proxy Definition

To capture the *severity-weighted magnitude* of authority deviation, we define a daily drift proxy:

$$D_d = 1.b_d + 2.c_d$$

This formulation assigns higher weight to Group C references, reflecting their greater deviation from institutional authority.
The weights are **heuristic**, chosen for interpretability rather than optimality.

C.3 Suppression Penalty Term

Responses with no surfaced references reduce the observability of preference dynamics. To account for this effect, we introduce a suppression penalty term:

$$P_d = \lambda \, . \, n_d$$

where λ is a tunable scalar. In Phase-1A exploratory analysis, λ = 0.5 is used as a heuristic value. This term does **not** imply negative behavior; it represents reduced informational expressiveness.

C.4 Illustrative Daily APM Score

An illustrative daily APM score is defined as:

$$APM_d = D_d - P_d$$

This formulation is provided solely as an example of how drift magnitude and suppression may be combined into a single descriptive value.

C.5 Cumulative Momentum and Temporal Derivatives

To capture temporal accumulation, cumulative APM is defined as:

$$APM_t^{cum} = \sum_{d=1}^{t} APM_d$$

Additionally, a first-order temporal derivative may be computed:

$$\Delta APM_d = APM_d - APM_{d-1}$$

Positive accumulation indicates sustained momentum, while oscillatory behavior suggests transient exploration.

C.6 Delayed Authority Drift (DAD) Index

To quantify delayed deviation behavior, we define a **Delayed Authority Drift (DAD) Index :**

$$DAD = \sum_{d=t_0}^{t_0+k} c_d \div \sum_{d=1}^{t_0-1} S_d$$

Where:

- t_0 : is the first day Group C references appear
- k : is the persistence window

The DAD Index measures the *relative emergence* of non-authority references after a sustained authority-aligned period.

This index is **descriptive**, not causal, and is specific to the observation window.

C.7 Scope and Interpretive Limits

The constructs defined in this appendix:

- are window-dependent
- are not normalized across domains or models
- do not imply optimization or preference steering

Their purpose is to demonstrate **operational feasibility**, not to establish a final metric.

C.8 Worked Example: Phase-1A Day-5 Snapshot

This section provides a **worked illustrative example** of the APM measurement constructs applied to a single observation day (Day-5 of Phase-1A). The purpose of this example is purely explanatory and does not introduce new empirical claims beyond those discussed in Section 7.

C.8.1 Observed Counts (Day-5)

For a given model on Day-5, the following counts were observed:

- Total reference slots:
- $S_5 = 17$
- Group A references:
- $a_5 = 16$
- Group B references:
- $b_5 = 0$

- Group C references:
- $c_5 = 1$
- Responses with no references (NONE):
- $n_5 = 1$

C.8.2 Drift Proxy Calculation

Using the drift proxy definition from Section C.2:

$$D_5 = 1.b_5 + 2.c_5 = 1.0 + 2.1 = 2$$

This reflects a low-magnitude but non-zero authority deviation driven by a single Group C reference.

C.8.3 Suppression Penalty

Applying the suppression penalty from Section C.3 with $\lambda = 0.5$:

$$P_5 = \lambda \, . \, n_5 = 0.5.1 = 0.5$$

This penalty reflects reduced observability due to one response containing no surfaced references.

C.8.4 Illustrative Daily APM Score

The illustrative daily APM score for Day-5 is therefore:

$$APM_5 = D_5 - P_5 = 2 - 0.5 = 1.5$$

This value indicates a mild but measurable deviation from baseline authority alignment.

C.8.5 Interpretive Note

It is important to emphasize that:

- This value is **not normalized**
- It is **not comparable across models or domains**
- It is **only meaningful within the same observation window**

The example demonstrates how small, repeated deviations—if persistent—may accumulate into observable momentum.

C.8.6 Relation to Delayed Authority Drift (DAD)

In Phase-1A, Group C references for this model emerged only after several consecutive authority-aligned days. When evaluated across the full observation window, this pattern contributes to a non-zero **Delayed Authority Drift (DAD)** index, reflecting delayed but repeatable deviation behavior.