

Role-Differentiated Knowledge Flow and Structured Deliberation in Multi-Agent LLM Systems

Scott N Hwang

The Pennsylvania State University
Penn State Milton S. Hershey Medical Center
PO Box 850, MC-H066, 500 University Drive
Hershey, PA 17033-0850, USA

snh5391@psu.edu

Phone: +1-610-715-1933

ORCID: <https://orcid.org/0000-0002-6496-2087>

LinkedIn: <https://www.linkedin.com/in/snhwang/>

Corresponding author: Scott N Hwang, snh5391@psu.edu

Word count: approximately 8,100 words (excluding references, tables, and figure captions)

Figures: 5 **Tables:** 10

Abstract

In multi-agent systems where agents hold distinct cognitive roles, shared knowledge should adapt to each agent’s perspective rather than propagate uniformly. We implement cognitive-mode knowledge diffusion. An agent’s retrieved behavioral instructions act as a cognitive filter that reframes and selectively stores incoming information through its assigned cognitive lens, using the BEAR retrieval framework. We demonstrate the mechanism through a Six Thinking Hats deliberation panel in which six LLM agents deliberate over eight biomedical topics. A three-way ablation shows that naive verbatim storage causes $2.12\times$ store bloat and 70.6% cross-agent content overlap, while BEAR-guided flow produces compact, role-differentiated stores (centroid distance 0.098 vs. 0.011, $p = 4.97 \times 10^{-8}$; nearest-neighbor overlap 5.5% vs. 70.6%, $p = 3.06 \times 10^{-10}$). Deduplication accounts for 64% of the gain with cognitive reframing for the remaining 36%. On SCT-Bench (174 clinical reasoning questions), Claude Haiku 4.5 improves from 0.540 to 0.700 (+0.188, $p < 0.001$) and MedGemma 27B from 0.516 to 0.693 (+0.156, $p < 0.001$). On BRAINTEASER (301 lateral thinking puzzles), GPT-OSS 120B improves on sentence (+0.136) and word puzzles (+0.258, both $p < 0.001$). Constant-model controls confirm differentiation is driven by BEAR instructions, not model diversity. Panel benefit is task- and model-dependent.

Keywords: Retrieval-Augmented Generation; Behavioral AI; Multi-Agent Systems; Knowledge Flow; Cognitive-Mode Filtering

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language understanding and generation [Brown et al., 2020, OpenAI, 2023]. A central challenge in deploying LLMs for interactive intelligent systems is behavioral consistency: ensuring that an agent reliably adopts an appropriate persona, respects domain-specific constraints, follows procedural protocols, and adapts its communication style across every turn of a conversation.

In multi-agent systems, the challenge extends beyond behavioral consistency to knowledge flow. When agents with distinct cognitive roles share information, that information should adapt to each agent’s perspective rather than propagate uniformly. When six LLM agents embody distinct cognitive modes (analytical, creative, optimistic, risk-focused, intuitive, facilitative), each agent must maintain its mode across arbitrary topics and also acquire and retain knowledge in a way that reflects its cognitive role. Uniform knowledge propagation, where every agent receives identical copies of shared information, causes agent knowledge stores to converge, undermining the value of cognitive differentiation. Static system prompts do not satisfy these requirements at scale. A prompt that encodes all six cognitive modes with all their nuances exceeds practical context budgets and provides no mechanism for context-sensitive activation of either behavioral or knowledge-filtering functions. Manual conditional prompt assembly [Chase, 2022, Mikinka, 2025], selecting components via lookup tables and if/elif logic, requires developers to anticipate every context combination and grows in complexity as the behavioral space expands.

From a complex systems perspective, cognitive-mode knowledge diffusion instantiates a local filtering rule in which each agent’s behavioral profile determines what information

it absorbs from the shared environment. The system-level property that emerges, role-differentiated knowledge stores across six agents, arises not from global coordination but from the consistent application of local cognitive filters. This paper characterizes that emergent property empirically and examines its consequences for downstream task performance.

We therefore implement a multi-agent panel using Behavioral Evolution And Retrieval (BEAR) [Hwang, 2026], a framework for governing agent cognition by retrieval-based behavioral instruction selection. Behavioral instructions are authored as structured YAML artifacts with explicit types, priorities, scope conditions, and inter-instruction relationships. At inference time, a retrieval pipeline selects instructions relevant to the current query and context, resolves conflicts and dependencies, and composes them into a structured system prompt. The same LLM, given the same user query, but a different runtime context (agent identity, topic, mood state), produces fundamentally different behavior because the prompt engineering is performed dynamically by the retrieval pipeline.

This paper applies cognitive-mode knowledge diffusion to govern inter-agent knowledge flow. The same retrieval pipeline that shapes how agents respond also determines how knowledge adapts as it moves between agents with different cognitive roles. Rather than blindly propagating all utterances to all agents, each agent’s BEAR-retrieved instructions define a cognitive lens through which incoming information is evaluated, reframed, and selectively retained. In the current work, structured role differentiation produces measurable performance gains for certain models on external benchmarks (§5, §6), demonstrating that panel benefit is real but model- and task-dependent.

We evaluate this mechanism through a multi-agent panel constructed based on the concept of Six Thinking Hats [de Bono, 1985], a method for improving group thinking by assigning one of six modes of thought, identified by six colored “hats”: White (facts), Red (feelings), Black (caution), Yellow (optimism), Green (creativity), and Blue (process control). The current work directs the six hats to deliberate on biomedical topics. Discussions across eight topics (diffuse midline glioma, stroke, multiple sclerosis, Alzheimer’s disease, epilepsy, GLP-1 receptor agonists, CRISPR gene therapy, LLMs in clinical decision support) show that BEAR-guided knowledge diffusion produces compact, role-differentiated per-agent knowledge stores, while naive (unfiltered) diffusion causes $2.12\times$ storage bloat and 70.6% cross-hat NN content overlap.

Section 2 summarizes the BEAR framework. Section 3 describes the Six Thinking Hats panel. Section 4 presents the quantitative evaluation of cognitive-mode knowledge diffusion. Sections 5 and 6 evaluate BEAR-structured panel deliberation on external benchmarks, examining when and for which models structured role differentiation improves task performance. Section 7 discusses knowledge flow governance and design trade-offs. Sections 8 and 9 survey related work and conclude. An open-source implementation is available at <https://github.com/snhwang/bear>.

2 The BEAR Framework

The BEAR framework is described in detail by Hwang [2026]; we summarize the components relevant to this paper. This section describes the Behavioral Evolution And Retrieval (BEAR) framework. The core pipeline, instruction schema, and retrieval algorithm are presented

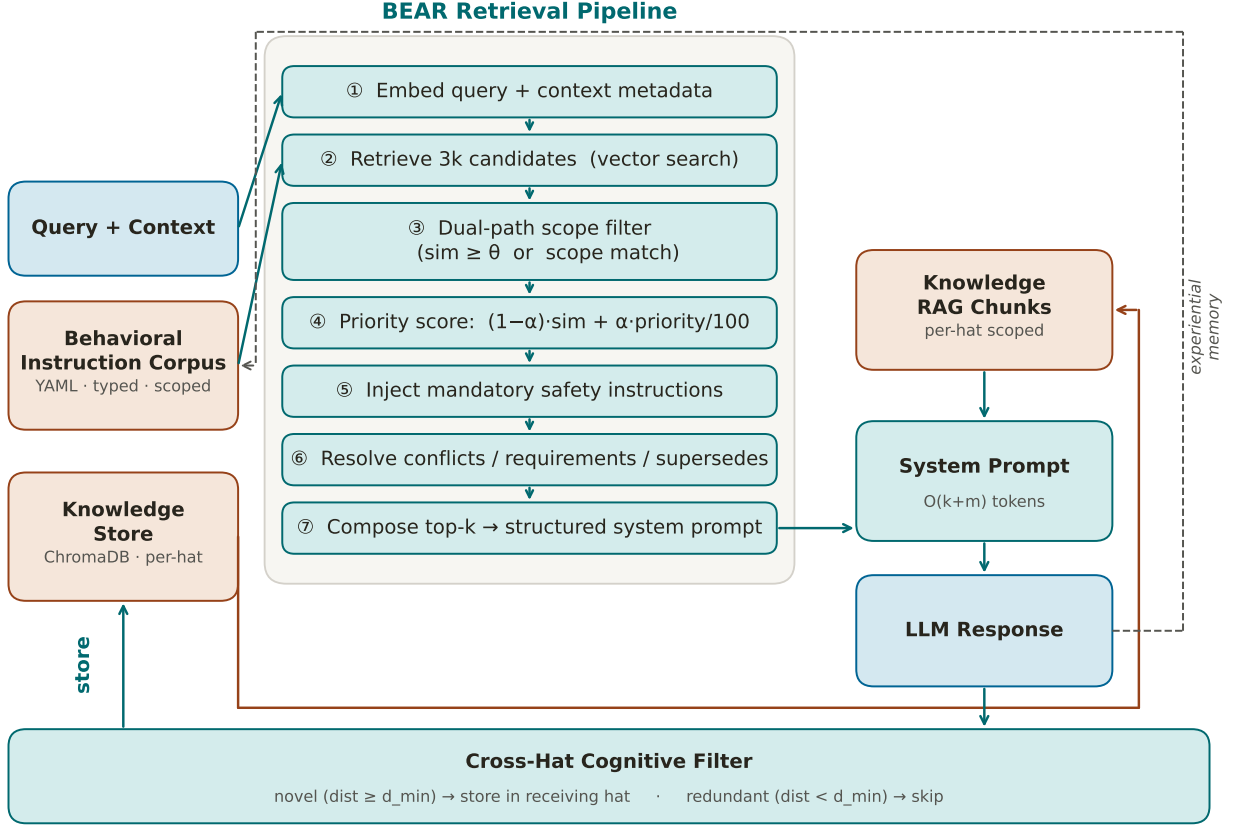


Figure 1: The BEAR retrieval pipeline. Left: two input stores — the behavioral instruction corpus (YAML, static) and the per-hat knowledge store (ChromaDB, dynamic). Center: seven-step retrieval pipeline composing behavioral instructions into a structured system prompt of size $O(k + m)$ tokens, where k is the number of retrieved instructions and m is the number of mandatory safety instructions (independent of the total corpus size). Right: knowledge RAG chunks augment the prompt; the LLM response feeds into the cross-hat cognitive filter (bottom), which selectively stores novel insights back into the knowledge store.

below; parameter sensitivity and scalability results appear in §2.4.

2.1 Core Pipeline

BEAR retrieves behavioral instructions, which are typed, prioritized, scope-tagged YAML artifacts, and composes them into structured system prompts (Figure 1). Each instruction carries a unique ID, a semantic type (CONSTRAINT, PERSONA, PROTOCOL, DIRECTIVE, FALLBACK, or TOOL), an integer priority (0–100), natural-language content, scope conditions specifying activation context, inter-instruction relationships (`conflicts_with`, `requires`, `supersedes`), and categorization tags.

At each decision cycle, the retrieval pipeline: (1) enriches the query with context metadata and embeds it; (2) retrieves $3 \cdot k$ candidates via vector search (where k is the final top- k ; default $k=10$); (3) applies a dual-path soft scope filter retaining candidates where either similarity $\geq \theta$ or scope conditions match the context; (4) scores candidates via priority-weighted scoring:

```

instructions:
- id: persona-black-hat-core
  type: persona
  priority: 80
  content: |
    You are the Black Hat. Your role is rigorous
    critical analysis. Surface risks, limitations,
    and failure modes. You are the group's quality
    filter, not its pessimist.
  scope:
    tags: [black-hat]
  tags: [persona, black-hat]

```

Figure 2: A behavioral instruction in YAML for the Black Hat cognitive mode.

score(i) = $(1 - \alpha) \cdot \text{sim}(q, i) + \alpha \cdot \text{priority}(i)/100$; (5) unconditionally injects mandatory safety instructions; (6) resolves conflicts, injects requirements, and removes superseded instructions via an instruction relationship graph; and (7) composes the top- k instructions into a structured system prompt bounded by $O(k + m)$ tokens regardless of corpus size N .

Figure 2 shows an example behavioral instruction.

2.2 Additional Pipeline Features

BEAR provides two query refinement mechanisms: *within-cycle refinement* (an LLM rephrases low-scoring queries before retrying) and *cross-cycle refinement* (the LLM generates a refined retrieval query after each response to steer next-cycle instruction selection). An experiential memory engine converts significant runtime experiences into typed instruction objects that enter the live corpus, can supersede outdated instructions, and persist across sessions via YAML serialization.

2.3 Two-Layer Retrieval Architecture

The system maintains two architecturally independent retrieval stores. The **behavioral instruction corpus** contains the BEAR YAML instructions described above: role-defining persona, constraint, protocol, and directive objects scoped per agent and retrieved at every turn to govern how each agent responds. The **knowledge store** is a separate ChromaDB collection populated at runtime from PDF ingestions and cross-hat diffusion events. All hats share the same collection, but each chunk is tagged with a `hat_id` ownership tag so retrieval is scoped to the querying hat. At each turn, BEAR retrieves from the instruction corpus to shape *how* the agent responds, implemented as retrieval-augmented generation (RAG [Lewis et al., 2020]). A separate knowledge RAG pipeline retrieves from the knowledge store to shape *what* the agent knows. Keeping the two stores separate prevents domain knowledge from crowding out persona, constraint, and protocol instructions in the system prompt. The behavioral instruction corpus remains static across sessions. The knowledge store grows dynamically as agents ingest documents and exchange information through the cognitive filter.

Users can supply individual hats with documents such as PDFs and can interject comments, including suggested topics of discussion. Otherwise the hats decide what to talk about and when an idea is worth adding to their personal knowledge stores.

2.4 Pipeline Robustness

The retrieval pipeline’s default parameters ($\alpha=0.3$, $\theta=0.3$, $k=10$) are robust across a range of settings. Priority weight α maintains stable F1 across $\alpha \in [0.1, 1.0]$; similarity threshold θ maintains $F1 = 0.784$ for $\theta \in [0.0, 0.5]$. Retrieval depth $k=5$ achieves peak precision ($F1 = 0.969$), while $k=10$ provides full recall at modest precision cost. BEAR achieves perfect recall across all instruction categories from 10 to 500 agents with zero scope violations. By design, top- k retrieval bounds composed prompts at $O(k+m)$ tokens. This yields stable prompts at ~ 950 – 1000 tokens regardless of corpus size (versus 34,721 tokens for monolithic injection at $N=500$, a $36\times$ compression ratio), with retrieval latency at 10–14 ms independent of corpus size.

Implementation. The reference implementation is written in Python. The server layer uses FastAPI with WebSocket support for real-time agent interaction. Behavioral instructions are stored as YAML files and indexed in ChromaDB for vector retrieval; sentence embeddings use BAAI/bge-base-en-v1.5 [Xiao et al., 2024] (768-dimensional). PDF text extraction uses pypdf, with Mathpix available as an alternative when better handling of mathematical expressions or structured figures is needed. The choice of extractor does not affect the knowledge diffusion mechanism under evaluation. Source code is available at <https://github.com/snhwang/bear>.

3 Demonstration: Six Thinking Hats Deliberation Panel

We demonstrate BEAR’s knowledge flow governance through a Six Thinking Hats deliberation panel [de Bono, 1985] applied to biomedical literature review. The panel provides a natural testbed for studying how knowledge flows between agents with distinct cognitive roles: per-agent corpus differentiation, context-adaptive instruction retrieval, knowledge RAG, cross-hat diffusion as a cognitive filter, and multi-backend LLM support are all directly observable.

3.1 Panel Setup

Six LLM agents each embody one of de Bono’s cognitive modes:

- **White Hat:** Facts and data. Reports empirical findings neutrally.
- **Red Hat:** Gut feelings and intuition. Reacts emotionally without justification.
- **Black Hat:** Critical analysis and risk. Identifies limitations and failure modes.
- **Yellow Hat:** Optimism and value. Surfaces opportunities and best-case scenarios.
- **Green Hat:** Lateral thinking and creativity. Generates novel connections and “what if” provocations.
- **Blue Hat:** Process facilitation. Keeps the session on track and draws summaries.

Each hat’s static BEAR corpus contains five hat-specific instructions (persona, thinking method, speech patterns, interaction protocol, mood-specific directive) at priorities 68–80, plus three shared safety constraints at priority 95–100 (injected via mandatory tags). The static corpus totals 33 instructions.

3.2 Per-Hat Domain Knowledge via PDF Ingestion

When a PDF document is uploaded to the session, the system extracts text using pypdf or Mathpix OCR (capped at 100,000 characters per document), and applies a 1,200-character overlapping-window chunking strategy with 200-character overlap. Each chunk is stored in a ChromaDB collection scoped to the receiving hat, together with citation metadata extracted from the document’s title and author lines.

Documents are not ingested identically for each hat. Each hat’s configured LLM reads the document through that hat’s cognitive lens and generates hat-scoped **DIRECTIVE** instructions (priority 60) capturing what is noteworthy from that perspective. White Hat extracts key statistics, measured outcomes, methodology details, and explicitly identified knowledge gaps. Black Hat extracts limitations, methodological weaknesses, counterevidence, and failure modes. Green Hat extracts unexpected connections, novel applications, and creative angles. The remaining hats follow the same pattern with extractions appropriate to their cognitive role.

The resulting knowledge instructions are scoped to the hat’s tags and topic keywords, so they surface through the standard BEAR retrieval pipeline when both the hat identity and a relevant topic are present in context. No separate knowledge retrieval step is needed; domain knowledge participates in the same pipeline as behavioral instructions.

3.3 Cognitive-Mode Knowledge Diffusion

During deliberation, each hat’s utterances may contain information relevant to another hat’s cognitive mode. A *cross-hat diffusion* mechanism enables each hat to passively accumulate knowledge from other hats’ contributions, filtered through its own BEAR-defined cognitive lens.

Mechanism. After every B exchanges (default $B=6$), a batch of recent utterances from other hats is assembled for each receiving hat. The receiving hat’s BEAR instructions are retrieved against a summary of the batch content, producing a composed cognitive profile. An LLM is prompted with this profile and the batch, asked to identify 0–2 facts or ideas that are relevant to this hat’s specific cognitive mode and worth retaining, restating each through the hat’s analytical lens.

For example, when White Hat’s utterance about DIPG survival rates reaches Black Hat’s diffusion buffer, Black Hat’s BEAR-retrieved instructions, which emphasize surfacing failure modes and naming concrete risks, cause it to store: *“DIPG’s nine-month median survival with virtually no long-term survivors represents a critical risk baseline that hasn’t moved in decades. Any proposed therapy must demonstrate it can shift this specific curve, not just show mechanistic plausibility.”* Red Hat, receiving the same utterance, stores: *“Hearing that DIPG nine-month survival statistic feels viscerally wrong. We’re not close enough to where we should be, and that urgency should drive us toward bolder experimental approaches.”*

The same fact is filtered through different cognitive lenses, producing hat-appropriate knowledge.

Deduplication. Before storing a candidate insight, the diffuser queries the receiving hat’s knowledge store for the nearest existing entry (cosine distance in ChromaDB). If the nearest neighbor distance falls below a configurable threshold (default $d_{\min} = 0.35$), the candidate is skipped as redundant. This dedup check prevents the knowledge store from accumulating semantically equivalent variations of the same observation.

Naive diffusion (ablation baseline). In naive diffusion mode, BEAR retrieval and LLM filtering are disabled. Every utterance from every other hat is stored verbatim in the receiving hat’s knowledge store, without cognitive filtering, reframing, or deduplication. This baseline approximates simple conversation logging and serves as the ablation condition in our evaluation (§4).

3.4 Per-Hat LLM Assignment

Each hat can be assigned its own LLM backend and model, with the session default handling background tasks (diffusion, insight extraction, query refinement). This enables heterogeneous multi-model configurations where analytical hats use larger models and facilitating hats use faster, cheaper ones. In our evaluation, we assign six distinct models across two providers (Anthropic and OpenAI):

- White Hat: Anthropic `claude-sonnet-4-6`
- Red Hat: Anthropic `claude-haiku-4-5`
- Black Hat: Anthropic `claude-opus-4-6`
- Yellow Hat: OpenAI `gpt-4.1-mini`
- Green Hat: OpenAI `gpt-4.1`
- Blue Hat: OpenAI `gpt-5.4`

This configuration demonstrates BEAR’s backend-agnostic design: the same behavioral instruction corpus and retrieval pipeline operates identically across models from different providers.

3.5 Session Flow

Sessions follow a scripted protocol to ensure consistency across all eight topics. A facilitator script fires a fixed sequence of chat prompts, with PDFs injected at predetermined positions between prompts. For the five original topics (two PDFs each), the structure is: an opening prompt establishes the topic; the first PDF is then injected and two follow-up prompts engage with its content; the second PDF is injected; and three closing prompts drive synthesis and clinical translation (six facilitator prompts total). The three additional topics (three PDFs each) follow the same pattern with one extra prompt and PDF injection before the closing sequence (seven prompts total).

For example, the DMG session proceeds as follows:

1. *Open*: “Discuss possible new treatments for diffuse midline glioma.”
2. INJECT PDF 1 — immunotherapy paper (Liu et al.)

| | |
|--|-------------------------------|
| Turn 6 — White Hat | <i>(before PDF injection)</i> |
| The H3 K27M mutation is not just a diagnostic marker—it drives a specific epigenetic reprogramming through inhibition of PRC2, which means it is a functionally targetable mechanism. EZH2 inhibitors and HDAC inhibitors like panobinostat have shown preclinical activity, though clinical translation has been limited by blood-brain barrier penetration and toxicity profiles. | |
| <i>[PDF injected: Liu et al. 2025 — adaptive immunotherapeutic paradigms in DMG]</i> | |
| Turn 8 — White Hat | <i>(after PDF injection)</i> |
| That connects to something in the available literature. Liu et al. 2025 describe adaptive immunotherapeutic paradigms for DMG, specifically emphasizing the interplay between epigenetic reprogramming, neuron-glioma interactions, and tumor microenvironment modulation—which supports the idea that single-target approaches may be insufficient. The data on neuron-glioma crosstalk in particular is an area where we have emerging evidence but significant gaps in understanding. | |
| Turn 10 — Blue Hat | <i>(facilitator prompt)</i> |
| White Hat, maybe you can pull together some specific data points on CAR-T and checkpoint inhibitors for us? Let’s get that factual baseline established before we evaluate any emotional or positive perspectives on them. | |

Figure 3: Excerpt from the DMG session around the first PDF injection. White Hat integrates the newly ingested literature (Turn 8 cites Liu et al. 2025 from the just-injected document); Blue Hat steers the subsequent discussion.

3. *Steer*: “What about immunotherapy approaches like CAR-T or checkpoint inhibitors for DMG?”
4. *Evidence*: “Which of these approaches has the strongest evidence so far?”
5. INJECT PDF 2 — molecular mechanisms paper (Nonnenbroich et al.)
6. *Mechanisms*: “How do the molecular mechanisms of DMG inform which targeted therapies might work?”
7. *Combinations*: “What combination strategies seem most promising given everything discussed?”
8. *Translation*: “What would a realistic clinical trial design look like for the most promising approach?”

Hats respond with BEAR-guided retrieval shaping each turn’s system prompt, cross-hat diffusion accumulates in the background, and Blue Hat closes by synthesizing the group’s key findings. PDF uploads are scripted at fixed positions to ensure protocol consistency across all sessions and conditions.

Figure 3 shows a brief excerpt from the DMG session around the first PDF injection, illustrating how White Hat integrates newly ingested literature and how Blue Hat steers the discussion.

4 Evaluation

We evaluate the cognitive-mode knowledge diffusion mechanism through three lenses: (1) *BEAR retrieval quality*, measuring how consistently behavioral instructions are retrieved

across a session; (2) *cognitive filtering ablation*, a controlled comparison of BEAR-guided versus naive cross-hat diffusion across eight biomedical topics; and (3) *response divergence*, measuring whether BEAR-guided agents produce measurably more differentiated outputs than naive agents. Evaluation of the BEAR retrieval pipeline itself (retrieval quality, scalability, and parameter sensitivity) is summarized in §2.4.

4.1 Experimental Setup

Topics. We conduct sessions on eight biomedical topics, each with two to three source PDFs:

- **Diffuse Midline Glioma (DMG):** Liu et al. 2025 (adaptive immunotherapy) [Liu et al., 2026]; Nonnenbroich et al. 2024 (molecular mechanisms) [Nonnenbroich et al., 2024].
- **Stroke:** Endovascular mechanical thrombectomy (standard of care) [Ding, 2015]; neuroprotective strategies for ischemic stroke [Haupt et al., 2023].
- **Multiple Sclerosis (MS):** Disease-modifying therapies including rituximab [Alping, 2023]; interventions targeting remyelination [De Keersmaecker et al., 2025].
- **Alzheimer’s Disease:** Etiology hypotheses and therapeutic strategies [Scarano et al., 2025]; treatment challenges for the future [Hardy, 2025].
- **Epilepsy:** Epilepsy as a dynamic disease [Schubert et al., 2025]; state-of-the-art gene therapy [Walker, 2025].
- **GLP-1 Receptor Agonists:** Three source PDFs on GLP-1 pharmacology and clinical outcomes.
- **CRISPR Gene Therapy:** Three source PDFs on CRISPR-based therapeutic approaches.
- **LLMs in Clinical Decision Support:** Three source PDFs on large language models applied to clinical decision-making.

Conditions. Each topic is run under three conditions:

- **BEAR-guided:** Cross-hat diffusion uses BEAR-retrieved instructions as cognitive filter, LLM reframing, and cosine-distance deduplication.
- **Embed-only:** Cosine-distance deduplication at $d_{\min} = 0.35$ but verbatim storage (no LLM reframing). Simulated by replaying naive session logs through the dedup filter.
- **Naive:** Every utterance from every hat is stored verbatim in every other hat’s knowledge store; no filtering, reframing, or dedup.

Each session is initialized with a clean knowledge store (`--clean` flag) to prevent cross-topic contamination. Sessions run 6 user-injected prompts over approximately 34 turns. Hat responses are generated at temperature 0.85 (encouraging diverse, creative contributions); diffusion filtering and knowledge ingestion LLM calls use temperature 0.3 (favoring consistent, deterministic filtering decisions).

Metrics. All metrics are deterministic (no LLM-as-judge):

- **Diffusion stored/skipped:** Counts of events stored versus deduplication-skipped per session.
- **Skip rate:** Fraction of diffusion candidates skipped.

- **Stored/skipped cosine distance:** Nearest-neighbor distance in the knowledge store at time of storage or skip decision.
- **Per-hat store size:** Final number of diffusion-sourced items in each hat’s knowledge store.
- **BEAR retrieval score:** Final priority-weighted score of retrieved instructions per turn (see §2).
- **Knowledge RAG hit rate:** Fraction of hat turns where at least one knowledge chunk was served.
- **Inter-hat centroid distance:** Cosine distance between mean embedding vectors (centroids) of each hat pair’s diffusion store.
- **Inter-hat Hausdorff distance:** Maximum nearest-neighbor cosine distance between hat pairs’ stores—captures focal differentiation.
- **Nearest-neighbor overlap:** Fraction of items in each store with a match (cosine distance $< d_{\min}$) in the other store.
- **Response divergence:** Mean pairwise cosine distance between hat response embeddings at each conversation phase—measures whether knowledge differentiation translates to different agent behavior.
- **Unique information:** Fraction of each hat’s response not explained by other hats’ responses (1 minus max cosine similarity to other responding hats).

Statistical methods. Statistical tests were selected to match data structure: one-sided t -tests for discrimination ratios (directional hypothesis $\bar{r} > 1.0$); Wilcoxon signed-rank for paired continuous scores (SCT-Bench); and McNemar’s test for paired binary outcomes (BRAINTEASER). Effect sizes are reported as Cohen’s d throughout. All p -values are two-sided unless otherwise noted.

4.2 BEAR Retrieval Quality

BEAR retrieval scores are stable across all sixteen sessions and both conditions (overall mean ≈ 0.75 ; constraint sub-score 0.88–0.95). The two systems differ in what they do with diffused knowledge, not in how behavioral instructions are retrieved.

4.3 Cognitive Filtering Ablation

BEAR-guided diffusion produces a mean skip rate of 28.9% across topics (range 7–56%); naive diffusion produces 0% by construction (no deduplication). The skip rate is highest for White Hat (56%), whose store is pre-populated with dense factual content from the ingested PDFs, causing most incoming diffusion candidates to fail the novelty check. The store size differential is clear: BEAR-guided mode produces a mean of ~ 9.4 items per hat (range 5.6–12.9), compared to ~ 20.1 per hat in naive mode, representing **2.12 \times store bloat**. Store sizes vary across topics and roles, reflecting topic-dependent variation in cross-hat information relevance.

Distance separation. Table 1 shows cosine distances at diffusion time. Stored events (items that passed the dedup check) have mean distances of 0.452–0.501, well above the deduplication

Table 1: Cosine distance distributions at diffusion time (BEAR-guided condition), across all eight topics. Stored items are genuinely novel (mean ≈ 0.481); skipped items are redundant (mean ≈ 0.270). The dedup threshold of 0.35 sits cleanly between the distributions.

| Topic | Event Type | Mean Distance | Range |
|-------------|------------|---------------|----------------|
| DMG | Stored | 0.463 | [0.350, 0.790] |
| DMG | Skipped | 0.286 | [0.210, 0.350] |
| Stroke | Stored | 0.495 | [0.360, 0.720] |
| Stroke | Skipped | 0.284 | [0.160, 0.340] |
| MS | Stored | 0.501 | [0.350, 0.700] |
| MS | Skipped | 0.278 | [0.160, 0.350] |
| Alzheimers | Stored | 0.483 | [0.350, 0.740] |
| Alzheimers | Skipped | 0.265 | [0.140, 0.340] |
| Epilepsy | Stored | 0.476 | [0.360, 0.700] |
| Epilepsy | Skipped | 0.273 | [0.110, 0.350] |
| GLP-1 | Stored | 0.452 | [0.350, 0.700] |
| GLP-1 | Skipped | 0.248 | [0.080, 0.340] |
| CRISPR | Stored | 0.500 | [0.350, 0.740] |
| CRISPR | Skipped | 0.279 | [0.210, 0.350] |
| LLM-CDS | Stored | 0.467 | [0.350, 0.640] |
| LLM-CDS | Skipped | 0.275 | [0.240, 0.310] |
| Mean | Stored | 0.481 | |
| | Skipped | 0.270 | |

threshold of 0.35, confirming they are genuinely novel relative to existing knowledge. Skipped events have mean distances of 0.248–0.286, well below threshold, confirming the dedup filter is correctly identifying redundant content. The two distributions are clearly separated (stored mean 0.481, skipped mean 0.270; gap of 0.211), with no overlap at the threshold.

Per-hat store growth. Table 2 shows per-hat accumulated store sizes averaged across all eight BEAR-guided topics. White Hat (mean 5.6) accumulates the fewest items. Its store is pre-populated with dense factual content from the ingested PDFs, causing most incoming diffusion candidates to fail the novelty check (56% skip rate). Black Hat (mean 12.9) and Yellow Hat (mean 12.4) accumulate the most diffused content, reflecting broad engagement with risk, critique, and optimistic reframing from every other cognitive angle. Red, Blue, and Green Hats sit in between at 7.2–10.8 items each, consistent with steady role-differentiated engagement with cross-hat content.

Temporal store evolution. The preceding metrics report final-state snapshots. To assess whether cognitive filtering operates consistently throughout a session or concentrates in particular phases, we track cumulative store size and inter-hat centroid distance turn-by-turn. Figure 4 shows the results for all eight topics.

Naive stores grow roughly linearly, accumulating ~ 116 – 126 items by session end across

Table 2: Per-hat diffusion store size (mean across 8 topics, BEAR-guided vs. naive). BEAR-guided stores are compact and role-differentiated; naive stores are uniformly large and topic-independent.

| Hat | BEAR-guided (mean) | Naive (mean) |
|-------------|--------------------|--------------|
| White | 5.6 | 18.0 |
| Red | 7.8 | 20.2 |
| Black | 12.9 | 23.2 |
| Blue | 7.2 | 18.0 |
| Green | 10.8 | 20.2 |
| Yellow | 12.4 | 21.0 |
| Mean | 9.4 | 20.1 |

all hats. BEAR-guided stores grow gradually to ~ 47 – 66 items, with the $\sim 2.1\times$ compactness gap opening from the first diffusion batch (turn 8–10) and persisting through the end of the session. The stability of this ratio indicates that BEAR’s cognitive filtering maintains a consistent selectivity throughout, rather than front-loading or exhausting its filtering capacity.

The inter-hat centroid distance (bottom row of Figure 4) reveals a complementary pattern: BEAR-guided sessions develop measurable inter-hat differentiation from the first diffusion event and maintain or increase it throughout the session. Naive sessions remain near zero centroid distance throughout, confirming that undifferentiated stores never spontaneously develop role-specific structure regardless of session length or topic.

Inter-hat role differentiation. The preceding metrics show that BEAR-guided stores are smaller and contain novel items, but do not address whether the six hats’ stores contain different content from each other. Table 3 measures pairwise differentiation between hat knowledge stores using three complementary metrics: centroid cosine distance (overall directional divergence), Hausdorff distance (maximum single-item divergence, indicating whether any hat retained content with no close match in another hat’s store), and nearest-neighbor overlap at threshold $d_{\min} = 0.35$ (fraction of items with a match in the other store). Across all eight topics, BEAR-guided sessions achieve mean centroid distance of 0.098 versus 0.011 in naive sessions ($9.2\times$, paired t -test $p = 4.97 \times 10^{-8}$, $d = 4.21$), and mean nearest-neighbor overlap of 0.055 versus 0.706 ($p = 3.06 \times 10^{-10}$, $d = 17.9$). In naive mode, all hats store the same unfiltered utterances, so their stores converge: 70.6% of items in any hat’s store have a near-duplicate in another hat’s store. In BEAR-guided mode, each hat’s cognitive filter selects and reframes different content. Black Hat retains risks, Green Hat retains creative angles, and White Hat retains data points, producing genuinely role-differentiated knowledge stores with only 5.5% cross-hat NN overlap. This confirms that role-differentiated diffusion produces genuinely specialized per-hat knowledge rather than redundant copies of the same content.

Isolating the cognitive filter contribution. The preceding comparison (BEAR-guided vs. naive) conflates two mechanisms: (a) cosine deduplication removing near-duplicate items,

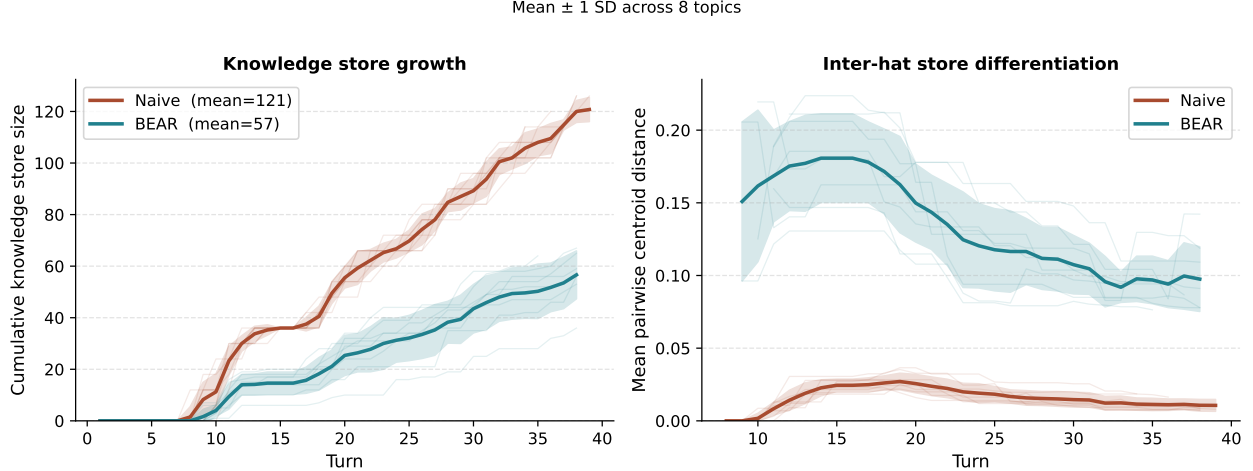


Figure 4: Temporal evolution of knowledge stores (mean \pm 1 SD across 8 topics). **Left:** cumulative item count per session. Naive stores grow to ~ 121 items; BEAR-guided stores remain compact at ~ 57 ($2.1\times$ gap). Right panel: BEAR-guided stores maintain mean centroid distance ~ 0.098 throughout, versus ~ 0.011 for naive stores ($9.2\times$ greater inter-hat differentiation). **Right:** inter-hat pairwise centroid distance over turns. BEAR-guided sessions establish role differentiation from the first diffusion batch and sustain it (mean ~ 0.40); naive sessions remain near zero throughout.

and (b) LLM-based cognitive reframing through the hat’s BEAR instructions. To disentangle these, we introduce an *embed-only* baseline that applies the same cosine dedup at $d_{\min} = 0.35$ but stores items verbatim (no LLM reframing). This baseline is simulated by replaying the naive session logs: for each utterance, we check whether it would pass the dedup threshold against previously accepted items in the receiving hat’s store.

Table 4 shows the three-way comparison (embed-only baseline run on the original five topics). The embed-only baseline produces stores that are intermediate between naive and BEAR-guided: centroid distance 0.063 (vs. 0.009 naive, 0.092 BEAR) and NN overlap 0.914 (vs. 0.979 naive, 0.945 BEAR), all measured on response-level embeddings from the original five sessions (Table 4). Deduplication accounts for the majority of the gain: approximately 64% of the centroid-distance increase over naive. The remaining 36% of the centroid gain (Δ centroid = 0.029, from 0.063 embed-only to 0.092 BEAR) is attributable to cognitive reframing, where the LLM filters and restates each candidate through the receiving hat’s BEAR-retrieved instructions. Both mechanisms contribute distinct effects: dedup removes near-duplicate content, while reframing reshapes retained content to align with each hat’s cognitive perspective, an increment that simple deduplication cannot produce.

Role versus topic identity. As a supplementary analysis, we compare intra-hat centroid distances (same hat across different topics) to inter-hat centroid distances (different hats within the same topic) under the BEAR-guided condition. Mean intra-hat distance across topics is 0.526, compared to mean inter-hat distance within topics of 0.098. The fact that intra-hat \geq inter-hat indicates that topic identity exerts a stronger signal than role identity at the centroid level, which is expected: with only 3–15 items per hat, the topical vocabulary

Table 3: Inter-hat pairwise store differentiation pooled across all eight topics (mean \pm SD, $n=8$ topics per condition, 15 hat pairs per session). All metrics computed on bge-base-en-v1.5 embeddings of stored knowledge items. Centroid Dist: cosine distance between mean embedding vectors of each hat’s store. NN Overlap: fraction of items with a near-duplicate ($\cos \geq 0.85$) in another hat’s store (lower = more differentiated). BEAR-guided diffusion produces stores with greater inter-hat divergence (higher centroid distance, lower NN overlap) than naive diffusion; temporal dynamics are shown in Figure 4.

| Condition | Centroid Dist | NN Overlap |
|--------------------------------|-------------------------------|--------------------------------|
| BEAR-guided | 0.098 ± 0.022 | 0.055 ± 0.043 |
| Naive | 0.011 ± 0.004 | 0.706 ± 0.045 |
| Ratio (BEAR/Naive) | $9.2\times$ | $0.08\times$ |
| p (paired t -test, $n=8$) | 4.97×10^{-8} | 3.06×10^{-10} |

dominates the embedding centroid. However, the lower NN overlap (5.5% vs. 70.6% naive) confirms that within a given topic, the hats retain genuinely different items; role differentiation operates at the item level, not the centroid level.

Knowledge RAG hit rate. Knowledge retrieval achieves 71–75% hit rate across all sessions; paper-sourced chunks account for 20–40% of retrieved content in BEAR-guided sessions, confirming that source PDFs have direct influence on the conversation.

Cognitive filter independence. BEAR retrieval scores are identical across BEAR-guided and naive conditions, confirming that the behavioral and knowledge layers are cleanly separated.

Role adherence in dialogue responses. The preceding metrics evaluate what is stored in knowledge stores. A complementary question is whether the hats’ actual dialogue responses adhere to their assigned cognitive roles. We measure this with a *role discrimination ratio*: for each hat response, we compute its embedding cosine similarity to the hat’s own role-instruction anchor (the centroid of its five YAML instruction embeddings) and divide by the mean similarity to all other hats’ anchors. A ratio above 1.0 indicates that the response is closer to the intended role than to other roles in embedding space.

Table 6 reports per-hat discrimination ratios averaged across all sixteen sessions (eight BEAR-guided, eight naive). Green Hat is the most role-distinctive ($\bar{r} = 1.057$): its creative, lateral-thinking language occupies a region of embedding space well-separated from the analytical vocabulary of other hats. Red Hat follows ($\bar{r} = 1.053$), reflecting its distinctive emotional and intuition-driven framing. Blue Hat is the least distinctive ($\bar{r} = 0.980$, not significant), expected for a process facilitator whose responses naturally span multiple cognitive modes.

Statistical significance of role signal. To confirm that the observed ratios are not noise, we run per-response one-sample t -tests against $H_0: \bar{r} = 1.0$ (Table 5). Note that Table 6

Table 4: Three-way comparison of diffusion conditions across five biomedical topics (the original five; embed-only baseline not re-run on the three additional topics). Embed-only applies cosine dedup at $d_{\min} = 0.35$ but stores items verbatim (no LLM reframing). Deduplication and cognitive reframing contribute complementary effects to inter-hat differentiation.

| Topic | Condition | Centroid Dist | Hausdorff Dist | NN Overlap |
|-------------|-------------|---------------|----------------|--------------|
| DMG | BEAR-guided | 0.099 | 0.373 | 0.930 |
| | Embed-only | 0.071 | 0.220 | 0.956 |
| | Naive | 0.011 | 0.241 | 1.000 |
| Stroke | BEAR-guided | 0.122 | 0.348 | 0.920 |
| | Embed-only | 0.073 | 0.276 | 0.900 |
| | Naive | 0.006 | 0.233 | 1.000 |
| MS | BEAR-guided | 0.080 | 0.361 | 0.932 |
| | Embed-only | 0.056 | 0.283 | 0.883 |
| | Naive | 0.013 | 0.336 | 0.990 |
| Alzheimer’s | BEAR-guided | 0.085 | 0.352 | 0.972 |
| | Embed-only | 0.041 | 0.146 | 0.917 |
| | Naive | 0.007 | 0.231 | 1.000 |
| Epilepsy | BEAR-guided | 0.088 | 0.400 | 0.922 |
| | Embed-only | 0.072 | 0.134 | 0.917 |
| | Naive | 0.009 | 0.269 | 1.000 |
| Mean | BEAR-guided | 0.092 | 0.346 | 0.945 |
| | Embed-only | 0.063 | 0.212 | 0.914 |
| | Naive | 0.009 | 0.326 | 0.979 |

reports session-level per-hat averages (mean $\bar{r} = 1.030$), while Table 5 operates on the full population of individual responses. Across all 517 responses (all sixteen sessions, including constant-model controls), the overall discrimination ratio is $\bar{r} = 1.024$, $p = 2.34 \times 10^{-26}$ (one-sided t -test), Cohen’s $d = 0.491$. Five of six hats show individually significant role signals: White Hat ($\bar{r} = 1.045$, $p < 10^{-5}$), Red Hat ($\bar{r} = 1.053$, $p < 10^{-15}$), Black Hat ($\bar{r} = 1.037$, $p < 10^{-7}$), Yellow Hat ($\bar{r} = 1.008$, $p < 0.01$), and Green Hat ($\bar{r} = 1.057$, $p < 10^{-12}$). Blue Hat ($\bar{r} = 0.980$, n.s.) does not reach significance, as expected for a process facilitator whose responses naturally span multiple cognitive modes.

Layer independence confirmed. The mean discrimination ratio is nearly identical between BEAR-guided and naive conditions ($p = 0.825$). The result confirms the architectural claim of layer independence: both conditions use the same BEAR behavioral instruction retrieval pipeline; the difference between conditions lies solely in the knowledge diffusion layer (cognitive filtering vs. unfiltered storage). Had changing the knowledge layer altered role adherence, it would indicate unwanted cross-layer coupling. The near-identical ratios demonstrate that the behavioral layer governs agent responses independently of what is in their knowledge stores.

Table 5: Statistical significance of discrimination ratios: one-sample t -tests against $H_0: \bar{r} = 1.0$ (one-sided). White, Red, Black, Yellow, and Green Hats show individually significant role signals; the overall population of 517 responses is highly significant ($p < 10^{-25}$, $d = 0.491$).

| Hat | n | \bar{r} | p (one-sided) | Sig |
|------------|-----|--------------|------------------------|------|
| White | — | 1.045 | $< 10^{-5}$ | *** |
| Red | — | 1.053 | $< 10^{-15}$ | *** |
| Black | — | 1.037 | $< 10^{-7}$ | *** |
| Yellow | — | 1.008 | < 0.01 | ** |
| Green | — | 1.057 | $< 10^{-12}$ | *** |
| Blue | — | 0.980 | n.s. | n.s. |
| All | 517 | 1.024 | 2.34×10^{-26} | *** |

Table 6: Role adherence: per-hat discrimination ratio (self-alignment / mean cross-alignment) averaged across all sessions (including constant-model controls). Five of six hats show ratios > 1.0 , confirming measurable role signal. Green Hat is most distinctive; Blue least distinctive.

| Hat | Discrimination Ratio |
|-------------|----------------------|
| White | 1.045 |
| Red | 1.053 |
| Black | 1.037 |
| Yellow | 1.008 |
| Green | 1.057 |
| Blue | 0.980 |
| Mean | 1.024 |

Hausdorff distance between hat response clouds. At the individual response level, BEAR-guided sessions achieve a mean Hausdorff distance of 0.375 versus 0.341 for naive sessions ($\Delta = +0.034$), consistent with the store-level finding (Table 3, mean 0.359 BEAR vs. 0.326 Naive).

The overall mean discrimination ratio of 1.024 (significantly > 1.0 , $p = 2.34 \times 10^{-26}$, $d = 0.491$; Table 5) confirms that role signal is present but modest, reflecting a known property of sentence-level embeddings: topical vocabulary dominates the embedding space. The per-hat significance pattern (strong for White, Red, Black, Green, and Yellow; non-significant for Blue), combined with Hausdorff distances of 0.29–0.45, indicates that role differentiation operates as a reliable bias. Effects concentrate in hats whose cognitive styles are linguistically most distinctive (creative, emotional, risk-focused, factual, optimistic) rather than those whose framing overlaps with general discourse (facilitative). The per-hat discrimination ratios are shown in Table 6.

Blue Hat synthesis coverage. Blue Hat’s syntheses achieve mean semantic coverage of 0.76 in BEAR-guided sessions versus 0.47 in naive sessions, a +0.29 advantage reflecting the

richer, differentiated content that BEAR diffusion provides. Topic fidelity averages 0.68 (range 0.58–0.77). Blue Hat’s low discrimination ratio ($\bar{r} = 0.980$) is expected: a good facilitator synthesizes across all perspectives rather than staying within one cognitive mode.

5 Application: Clinical Reasoning Benchmark

The preceding evaluation establishes that BEAR-guided diffusion produces role-differentiated knowledge stores. We now ask whether the deliberation structure BEAR enables improves performance on external tasks with objective ground truth, and under what conditions. The following two sections evaluate BEAR-structured Six Hats panels using static role prompts drawn from the same BEAR instruction corpus as the brainstorming evaluation, without live knowledge diffusion, since structured tasks with deterministic answers do not require per-query knowledge retrieval. The benchmarks therefore test the deliberation structure BEAR enables, complementing the diffusion evaluation in §4. Running live diffusion during benchmark evaluation would require constructing per-task knowledge corpora and would not affect the hat role prompts, which are the primary variable under test.

5.1 Benchmark and Setup

We evaluate on SCT-Bench [McCoy et al., 2025], a clinical reasoning benchmark based on Script Concordance Testing [Charlin et al., 2000]. Each of the 174 publicly available questions presents a clinical scenario, a diagnostic or treatment hypothesis, and a new piece of clinical information; the respondent rates how the new information affects the hypothesis on a five-point scale (−2 to +2). Scoring compares model responses against an expert clinician panel distribution: the modal expert response receives a score of 1.0, with partial credit for less common expert choices. This scoring scheme rewards calibration to clinical expert reasoning under uncertainty, not merely factual recall.

We compare three conditions, each using the same total number of LLM calls:

1. **Single-agent**: one call per question at temperature 0 (deterministic).
2. **Self-consistency (SC)**: six independent samples at temperature 0.5; majority vote selects the final answer.
3. **Six Hats panel**: six BEAR-structured hat agents respond sequentially, each seeing all prior hat responses before committing to a rating. The final answer is determined by majority vote over the six hat ratings (algorithmic aggregation, no LLM synthesis step).

The panel uses the hat-role instruction corpus developed for the brainstorming evaluation. The sequential discussion structure means each hat builds on prior perspectives; all prior responses are in context when each hat commits to its rating. We evaluate nine models spanning frontier APIs, open-weight models, and domain-specific medical models (Table 7).

5.2 Results

Table 7 reports results. The panel significantly outperforms self-consistency for four of nine models, and significantly hurts performance for two (Gemma 3 27B and Mistral Nemo 12B).

Table 7: SCT-Bench results (174 questions, score = mean alignment with expert clinician panel). Panel = Six Hats majority vote; SC = self-consistency majority vote (6 samples). Δ vs SC with bootstrap 95% CI (10,000 resamples). p -values from two-sided Wilcoxon signed-rank test; Cohen’s d for paired differences. Human benchmarks from McCoy et al. [McCoy et al., 2025]. Bold = panel significantly outperforms SC ($p < 0.05$); italics = panel significantly hurts. Claude models use $t=0.5$ for SC/panel, $t=0$ for single; MedGemma uses $t=0$ [Yang et al., 2025]. Three additional models (GPT-OSS 120B, Nemotron Super, Gemma 4 31B) evaluated at manufacturer-recommended temperatures showed no significant panel effect and are omitted (GPT-OSS: $\Delta=+0.011$, n.s.; Nemotron: $\Delta=-0.092$, $p<0.001$; Gemma 4: $\Delta=+0.015$, n.s.).

| Model | Single | SC | Panel | Oracle | Δ vs SC [95% CI] | p | d |
|-------------------|-----------------|-------|--------------|--------|-------------------------|-------------|--------|
| Claude Opus 4.6 | 0.639 | 0.623 | 0.730 | 0.794 | +0.107 [+0.063, +0.155] | < 0.001 *** | 0.344 |
| Claude Sonnet 4.6 | 0.666 | 0.666 | 0.757 | 0.833 | +0.091 [+0.040, +0.143] | < 0.001 *** | 0.260 |
| Claude Haiku 4.5 | 0.540 | 0.513 | 0.700 | 0.890 | +0.188 [+0.112, +0.261] | < 0.001 *** | 0.380 |
| GPT-5.4 | 0.745 | 0.758 | 0.770 | 0.825 | +0.012 [−0.025, +0.050] | n.s. | 0.048 |
| Grok 4.20 | 0.673 | 0.687 | 0.673 | 0.884 | −0.013 [−0.070, +0.045] | n.s. | −0.035 |
| Gemma 3 27B | 0.615 | 0.634 | <i>0.563</i> | 0.756 | −0.071 [−0.135, −0.010] | 0.033 * | −0.166 |
| MedGemma 27B | 0.516 | 0.537 | 0.693 | 0.833 | +0.156 [+0.083, +0.226] | < 0.001 *** | 0.316 |
| MedGemma 4B | 0.457 | 0.519 | 0.463 | 0.567 | −0.056 [−0.128, +0.016] | n.s. | −0.128 |
| Mistral Nemo 12B | 0.606 | 0.595 | <i>0.483</i> | 0.789 | −0.112 [−0.185, −0.041] | 0.002 ** | −0.231 |
| Medical students | ≈ 0.620 | | | | (McCoy et al.) | | |
| Residents | ≈ 0.700 | | | | | | |
| Staff physicians | ≈ 0.750 | | | | | | |

Three additional models evaluated at manufacturer-recommended temperatures showed no significant panel effect and are noted in the caption.

The ceiling effect. GPT-5.4 has the highest single-agent baseline (0.745, near staff physician level) and shows no significant panel improvement ($p = 0.52$). When a model is already strong and self-consistent, the six hats converge on similar answers and majority vote adds no advantage over self-consistency. Grok 4.20 similarly shows no significant improvement; SC already improves over single-agent (0.687 vs. 0.673), suggesting temperature sampling alone is sufficient for its diversity needs.

Models where the panel hurts. Two models show significant panel degradation. Gemma 3 27B (-0.071 , $p = 0.033$) has severely degraded Black Hat (0.438) and Green Hat (0.444), dragging down the majority vote. Mistral Nemo 12B shows the strongest negative effect (-0.112 , $p = 0.002$, $d = -0.231$): every hat performs below single-agent, with Black Hat at 0.300. Both models lack the capacity to meaningfully adopt structured reasoning perspectives; the hat prompts confuse rather than help.

Structured diversity beats temperature diversity. For Opus and Haiku, self-consistency essentially equals or falls below single-agent (Opus: 0.623 vs. 0.639; Haiku:

0.513 vs. 0.540): at temperature 0.5, these models are sufficiently deterministic that sampling produces near-identical responses, and majority vote over redundant samples cannot improve (or slightly degrades) accuracy. The panel unlocks +0.107 (Opus) and +0.188 (Haiku) improvements by forcing genuinely different analytical perspectives through role-differentiated prompts, rather than relying on stochasticity for diversity. Haiku’s gain is the largest of any model ($d = 0.380$) despite its weaker single-agent baseline, confirming that the panel’s structured diversity is most valuable precisely when temperature-based diversity fails.

The MedGemma finding. The most striking result involves MedGemma 27B, a domain-specific medical language model. Its single-agent score (0.516) falls below the medical student benchmark (≈ 0.620), suggesting that domain pretraining alone does not improve calibration to expert clinical reasoning panels. The Six Hats panel lifts MedGemma to 0.693, approaching the resident benchmark (≈ 0.700) and giving a +0.177 absolute improvement over single-agent (+0.156 over SC, $p < 0.001$, $d = 0.316$). Per-hat analysis reveals why: MedGemma’s individual hat scores are uniform (0.627–0.709), indicating it can reason from each cognitive perspective in isolation; the deficit is in spontaneously integrating multiple perspectives without structured deliberation. The panel structure compensates for this integration failure by making each perspective explicit and sequential.

Capability threshold: MedGemma 4B. MedGemma 4B provides a lower-bound test of the capability threshold hypothesis. Its single-agent score (0.457) falls well below the medical student benchmark, and the panel yields no improvement ($\Delta = -0.056$, $p = 0.158$, n.s.). Per-hat scores are tightly clustered (0.435–0.474), even more uniform than MedGemma 27B, yet the panel does not help. The difference is that 4B’s hats are uniformly weak: the model lacks the capacity to generate quality role-differentiated reasoning regardless of which hat it wears. This contrasts with MedGemma 27B, where uniform hat scores reflect genuine and consistent capability across all six perspectives. Panel benefit requires per-hat capability, not merely uniformity: the panel amplifies what is already present in the model’s reasoning repertoire and cannot substitute for fundamental capacity gaps.

Hat quality uniformity as a predictor. Across all models, panel benefit correlates with hat-score uniformity. Models where the panel helps significantly (Opus, Sonnet, Haiku, MedGemma 27B) show all six hats performing within a relatively narrow range, so each hat contributes useful signal. Models where the panel fails or hurts (Gemma 3 27B, Grok 4.20) show high variance across hats, with weak perspectives (often Green and Black) introducing noise that majority vote cannot overcome.

6 Application: Lateral Thinking Benchmark

The SCT evaluation tests structured deliberation on clinical reasoning under uncertainty. To assess whether the benefit generalizes to a different reasoning domain, we evaluate on BRAINTEASER [Jiang et al., 2023], a lateral thinking benchmark requiring creative reinterpretation of language.

Table 8: BRAINTEASER results for GPT-OSS models (SP = 169 sentence puzzles, WP = 132 word puzzles; $t=1.0$, $\text{top-}p=1.0$). p -values from McNemar’s test (panel vs. SC); Cohen’s d for paired differences. 95% bootstrap CIs on Δ : 120B SP [+0.065, +0.207], 120B WP [+0.174, +0.341], 20B SP [+0.130, +0.296]. Four additional models (Nemotron Super, Gemma 4 31B, Claude Haiku 4.5, Claude Sonnet 4.6) showed near-ceiling baselines (≥ 0.882) with no significant panel improvement and are omitted. Bold = significant ($p < 0.05$).

| Model | Type | Single | SC | Role-Maj | Panel | p | d |
|---------------------|------|--------|-------|----------|--------------|---------------|-------|
| GPT-OSS 120B | SP | 0.710 | 0.746 | 0.811 | 0.846 | < 0.001 *** | 0.287 |
| GPT-OSS 120B | WP | 0.636 | 0.735 | 0.720 | 0.894 | < 0.001 *** | 0.511 |
| GPT-OSS 20B | SP | 0.574 | 0.627 | 0.663 | 0.787 | < 0.001 *** | 0.382 |

6.1 Benchmark and Setup

BRAINTEASER consists of 169 Sentence Puzzles (SP) and 132 Word Puzzles (WP), each a multiple-choice question (A–D) where the intuitive answer is typically wrong and the correct answer requires lateral thinking: wordplay, double meanings, or unconventional interpretations. We use the same three-condition design as SCT (single-agent, self-consistency, Six Hats panel), plus a *role-majority* condition where each hat answers independently without seeing prior discussion, isolating the effect of role-specific prompting from deliberation. All models use manufacturer-recommended sampling parameters.

6.2 Results

Table 8 reports results. All data verified clean: $\geq 99\%$ parsed hat answers, zero missing majority votes.

GPT-OSS benefits significantly. GPT-OSS 120B shows significant panel improvement on both puzzle types: +0.136 over SC on SP ($p < 0.001$, $d = 0.287$) and +0.258 on WP ($p < 0.001$, $d = 0.511$). The smaller GPT-OSS 20B also improves (+0.213 over SC on SP, $p < 0.001$, $d = 0.382$), demonstrating that the benefit is not size-dependent within this model family.

Ceiling effects. Four additional models tested at recommended parameters showed near-ceiling single-agent baselines (≥ 0.882): Nemotron Super (0.893), Gemma 4 31B (0.947), Claude Haiku 4.5 (0.882), and Claude Sonnet 4.6 (0.947). None showed significant panel improvement ($p > 0.05$). Nemotron illustrates a methodological point: at suboptimal $t=0.5$ its single-agent accuracy was only 0.663 and the panel appeared to help dramatically, but at the recommended $t=1.0$ the baseline rises to 0.893 and the apparent gain disappears. Suboptimal temperature settings can artificially depress baselines and inflate apparent panel benefits.

Role-majority vs. panel. The role-majority condition (independent hat votes, no discussion) provides an intermediate between SC and panel. For GPT-OSS 120B on SP, role-majority

(0.811) already substantially exceeds SC (0.746), indicating that role-specific prompting alone, without any deliberation, accounts for much of the improvement. The full panel adds a further +0.041, suggesting that discussion context provides additional but smaller gains beyond role diversity. On WP, however, the pattern reverses: role-majority (0.720) barely exceeds SC (0.735), while the panel jumps to 0.894, indicating that WP puzzles benefit more from inter-hat deliberation.

Task and model specificity. The GPT-OSS family benefits on BRAINTEASER but not SCT, while Claude models and MedGemma 27B benefit on SCT but show near-ceiling BRAINTEASER baselines (≥ 0.882) that leave no room for panel improvement. Panel benefit therefore depends on the interaction between model capabilities and task demands: GPT-OSS can adopt creative lateral-thinking perspectives but not clinical judgment perspectives, and Claude models show the reverse.

7 Discussion

7.1 Governing Knowledge Flow as a General Pattern

The cross-hat diffusion mechanism reveals the central finding of this work: BEAR’s retrieved behavioral instructions can govern how knowledge flows between agents, not only how individual agents respond. In the canonical BEAR pipeline, instructions shape how the LLM responds. In the diffusion pipeline, the same retrieval and composition steps produce a cognitive profile that shapes how the agent perceives and retains information from its environment. The deliberation panel is the evaluation testbed; knowledge flow governance is the contribution.

This separation yields a two-layer architecture that follows established multi-agent design patterns [Hong et al., 2024], with the novelty that behavioral retrieval governs the knowledge layer. The BEAR layer retrieves behavioral instructions that define the agent’s cognitive lens (what it finds interesting, how it frames information, what it prioritizes), and the knowledge layer stores the filtered and reframed content produced by applying that lens to incoming information.

The role adherence evaluation (Table 6) confirms this separation: switching from BEAR-guided to naive diffusion produces no significant change in discrimination ratios ($p = 0.825$; Table 5), even though the knowledge stores differ dramatically in size and content (Table 3). The behavioral layer governs *how* agents respond; the knowledge layer governs *what* they know. Modifying one does not perturb the other. The layers are architecturally independent but functionally coupled: different knowledge leads agents to emphasize different aspects of the topic within their established cognitive mode, as the benchmark results in §5 and §6 demonstrate directly (using static BEAR role prompts, without live diffusion).

The NN overlap analysis (Table 3) adds a subtler observation: BEAR-guided stores exhibit lower inter-hat NN overlap (0.359 vs. 0.326 naive). Cognitive filtering shapes what is stored and also increases focal differentiation, since each hat retains at least some content with no close equivalent in any other hat’s store.

Currently, this pattern is implemented at the application layer (`CrossHatDiffuser` class

in BEAR Parlor). No modifications to the core BEAR framework were required; the existing retriever and composer APIs provided all necessary machinery. A natural generalization would be a core BEAR primitive, tentatively `BearDrivenFilter`, that, given retrieved instructions and arbitrary input content, determines what to retain and how to reframe it. This would formalize BEAR’s role as a knowledge flow governance framework, applicable to any multi-agent system where agents with differentiated roles share information.

7.2 Why Cognitive Filtering Matters for Multi-Agent Knowledge Flow

In an unfiltered multi-agent deliberation, agents face a signal-to-noise problem: not all utterances from other agents are relevant to any given agent’s role. If every agent stores every utterance from every other agent, the retrieval space fills with off-role content and agent knowledge stores converge toward undifferentiated copies of the full conversation. When White Hat later queries its knowledge store about treatment mechanisms, it retrieves Red Hat’s gut reactions, Black Hat’s risk analyses, and Green Hat’s lateral provocations together, a mix that dilutes the factual, data-grounded content White Hat was designed to produce.

BEAR-guided diffusion addresses this by having each hat’s behavioral instructions act as a filter: before storing a cross-hat observation, the receiving hat asks “*is this relevant to my cognitive mode, and if so, how should I frame it?*” The result is a knowledge store that reflects not only what was said but what matters to this agent’s function.

The empirical results bear this out: BEAR-guided stores stay compact, contain items that are genuinely novel relative to existing knowledge (Table 1), and overlap little across hats (Table 3). Naive stores, by contrast, hold $2.12\times$ as many items on average, with 70.6% of those items duplicated across hats. The embed-only baseline (Table 4) disentangles the two mechanisms: cosine deduplication accounts for roughly 64% of the centroid-distance gain over naive; cognitive reframing accounts for the remaining 36%. Deduplication is the dominant contributor, but reframing adds a further increment by reshaping retained content toward each hat’s cognitive mode, something deduplication alone cannot produce.

Implications for human-AI collaboration. While our evaluation uses automated facilitator prompts, a natural deployment would include a human facilitator interacting with the panel. In this setting, the cognitive filtering mechanism serves a dual purpose: it helps each AI agent maintain role consistency, and it presents the human facilitator with genuinely differentiated perspectives. A human seeing Black Hat’s risk-framed retention of survival statistics alongside Green Hat’s creative reframing of the same data may generate novel ideas that neither perspective would surface in isolation. The BEAR-governed knowledge flow thus functions as a structured tool for human-AI collaboration, not merely an autonomous agent system.

7.3 Design Trade-offs

Deduplication threshold. To characterize the sensitivity of the deduplication mechanism to the threshold parameter d_{\min} , we replay the naive session logs (original five topics) through embed-only dedup at seven threshold values ($d_{\min} \in \{0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50\}$).

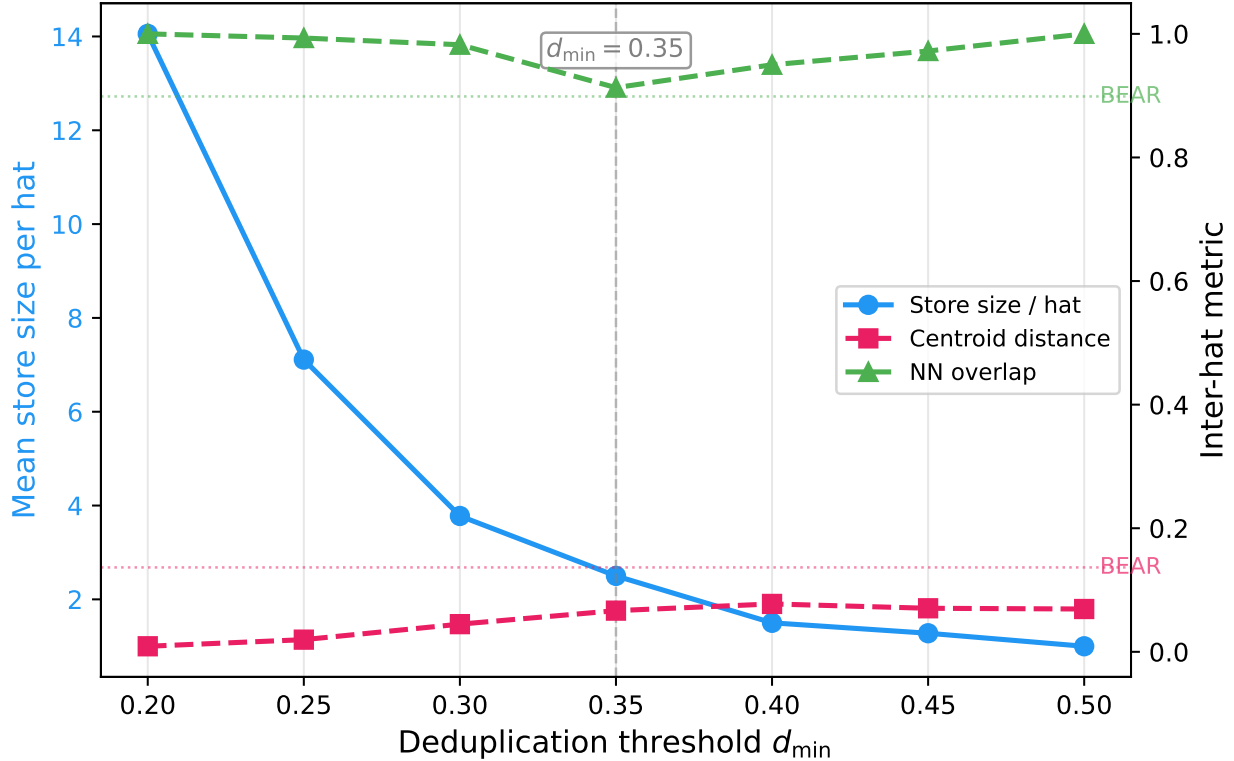


Figure 5: Sensitivity of embed-only dedup to threshold d_{\min} . Blue: mean store size per hat (left axis); pink: pairwise centroid distance between hat stores (right axis); green: nearest-neighbor overlap (right axis). Dashed horizontal lines show BEAR-guided reference values. The vertical marker at $d_{\min} = 0.35$ indicates the default threshold, which sits at the knee of the store-size curve. Dedup alone produces partial differentiation that saturates at higher thresholds; cognitive reframing (BEAR) adds further divergence beyond dedup’s contribution.

For each threshold, we compute the mean per-hat store size and inter-hat differentiation metrics, using the same embeddings across all thresholds so the only variable is the threshold itself.

Figure 5 and Table 9 show the results. Store size drops sharply from 11.9 items/hat at $d_{\min} = 0.20$ to 2.3 at $d_{\min} = 0.35$, after which the curve flattens (1.7 at 0.40, 1.2 at 0.50). The threshold $d_{\min} = 0.35$ thus sits at the knee of the store-size curve, where further increases yield diminishing dedup gains while risking loss of genuinely novel content. Centroid distance and nearest-neighbor overlap show moderate variation across thresholds (centroid 0.01–0.08, overlap 0.91–1.00), indicating that deduplication contributes meaningful but partial inter-hat differentiation that saturates at higher thresholds. The BEAR-guided reference (centroid 0.098, overlap 0.055) remains substantially higher, confirming that cognitive reframing contributes additional differentiation beyond what deduplication achieves alone, consistent with the three-way comparison in Table 4. The threshold should be calibrated to the embedding space: the 0.35 value was chosen for BAAI/bge-base-en-v1.5 and may need adjustment for other models.

Table 9: Sensitivity of embed-only dedup to d_{\min} . Store size and skip rate quantify dedup aggressiveness; centroid distance, Hausdorff distance, and NN overlap measure inter-hat differentiation. The default $d_{\min} = 0.35$ (bold) balances compactness with content retention. BEAR-guided results (bottom) include cognitive reframing.

| d_{\min} | Store/hat | Skip% | Centroid | Hausdorff | Overlap |
|-------------|------------|--------------|--------------|--------------|--------------|
| 0.20 | 11.9 | 48.3% | 0.011 | 0.271 | 0.991 |
| 0.25 | 6.4 | 72.0% | 0.020 | 0.285 | 0.983 |
| 0.30 | 3.7 | 84.1% | 0.041 | 0.258 | 0.970 |
| 0.35 | 2.3 | 90.1% | 0.063 | 0.212 | 0.914 |
| 0.40 | 1.7 | 92.8% | 0.073 | 0.163 | 0.937 |
| 0.45 | 1.5 | 93.3% | 0.070 | 0.132 | 0.950 |
| 0.50 | 1.2 | 94.9% | 0.077 | 0.118 | 0.960 |
| BEAR | — | — | 0.098 | — | 0.055 |

Batch size. Diffusion fires every $B = 6$ exchanges rather than every turn. This amortizes the LLM call cost across multiple utterances and allows the receiving hat to evaluate content with more context. Smaller batches produce more frequent diffusion events but increase cost; larger batches delay the availability of diffused knowledge.

Per-hat LLM assignment. Using six different models across two providers introduces heterogeneity in output quality and style. This is intentional for our evaluation, since it demonstrates framework backend-agnosticism. A constant-model control experiment (§7.5) confirms that differentiation persists, and in fact strengthens, under model homogeneity.

7.4 Relationship to Chain-of-Thought Reasoning

A natural question is whether structured multi-hat deliberation merely reproduces chain-of-thought (CoT) prompting [Wei et al., 2022] or self-consistency [Wang et al., 2023] at higher cost. Two observations distinguish the approaches.

First, CoT elicits step-by-step reasoning within a single agent, a kind of monologue. Self-consistency generates multiple independent monologues and selects via majority vote. Neither mechanism allows one reasoning path to critique another. In the BEAR panel each hat sees all prior hats’ responses and responds to them: Black Hat explicitly challenges proposed answers, and Blue Hat weighs competing arguments before synthesizing. This interactive, role-differentiated critique is structurally closer to AI debate [Irving et al., 2018] than to prompted reasoning chains.

Second, the role adherence results (§4) confirm that BEAR-guided hats produce measurably more differentiated outputs than naive agents given the same role labels. This divergence is a prerequisite for effective deliberation: if all agents converge to the same reasoning path, sequential discussion adds context length without adding information. The cognitive filter ensures that each hat’s retrieved instructions steer it toward a genuinely distinct perspective, producing the role-driven diversity that CoT and self-consistency lack.

The panel benefit is conditional, improving outcomes for some models but not others, mirroring findings from human group deliberation research. Diehl and Stroebe [1987] showed that interacting groups often underperform nominal groups due to production blocking and evaluation apprehension; yet groups with genuine cognitive diversity consistently outperform individuals when deliberation is structured to exploit that diversity [Karadzhov et al., 2024]. The critical distinction is between adversarial role differentiation, in which participants defend fixed positions against each other, and complementary role differentiation, in which each participant contributes a distinct perspective toward a shared goal. The Six Hats framework instantiates the latter: no hat argues against another; each performs its cognitive function sequentially, and the aggregation step combines their outputs rather than adjudicating between them. When a model lacks the capacity to genuinely inhabit a role (Gemma 3 27B’s degraded Black and Green Hats; MedGemma 4B’s uniformly weak hats), the complementary structure collapses into noise. The panel amplifies genuine cognitive differentiation; it cannot manufacture it.

7.5 Limitations

Evaluation scope. Our ablation covers eight topics with two to three PDFs each and sixteen sessions total. Larger-scale evaluation across more topics and domains would strengthen generalizability. The per-hat store size differences across topics suggest topic-dependence in diffusion rates; a more systematic characterization would require controlled variation in topic breadth and document density.

Model heterogeneity. The primary evaluation assigns six different LLMs to the six hat roles across two providers (Anthropic and OpenAI). To disentangle BEAR-instruction effects from model-diversity effects, we ran constant-model controls in which both BEAR-guided and naive DMG sessions used a single model for all six hats. We tested first a cloud model (Claude Sonnet 4.6), then a local 12B model (Mistral Nemo Instruct, via LM Studio). Table 10 shows the results. With Sonnet 4.6, the uniform-model BEAR session produces comparable inter-hat differentiation to the heterogeneous-model sessions (response-level centroid 0.125 vs. 0.144, Hausdorff 0.345 vs. 0.345, overlap 91.7% vs. 93.6%), while the naive session collapses to near-zero differentiation (centroid 0.010, overlap 100%). All three metrics in Table 10 are computed over hat *dialogue responses* (turn-level utterance embeddings), whereas Table 3 reports the same metrics over hat *knowledge store items* (diffused and reframed insights). The NN overlap threshold also differs (0.9 here vs. 0.85 in Table 3). The two tables are therefore complementary: the constant-model control measures whether BEAR instructions shape how agents respond, while Table 3 measures whether they shape what agents retain. The local 12B model shows even stronger BEAR differentiation (centroid 0.336, overlap 45.0%), likely due to smaller models’ more literal adherence to filtering instructions. In both cases, BEAR-guided differentiation is 12–21 \times the naive baseline, confirming that BEAR instructions, not model diversity, drive inter-hat knowledge specialization.

Knowledge retrieval quality. We measure what is stored and how much, and the role adherence evaluation (Table 6) confirms that hat responses align with their assigned cognitive

Table 10: Constant-model control: inter-hat differentiation with all six hats using the same model versus the heterogeneous-model primary evaluation. BEAR-guided differentiation persists (Sonnet) or strengthens (12B local) under model homogeneity, confirming that the effect is driven by BEAR instructions, not model diversity. **Note:** all three metrics here are computed over hat *dialogue responses* (not knowledge store items as in Table 3). Centroid = mean pairwise cosine distance between hat response embeddings. Overlap = fraction of responses with cosine similarity ≥ 0.9 to another hat (stricter threshold than Table 3’s ≥ 0.85 store-level metric).

| Condition | Models | Centroid [†] | Hausdorff [†] | Overlap [†] |
|-------------|----------------------|-----------------------|------------------------|----------------------|
| BEAR-guided | Heterogeneous | 0.144 | 0.345 | 0.936 |
| BEAR-guided | Uniform (Sonnet 4.6) | 0.125 | 0.345 | 0.917 |
| BEAR-guided | Uniform (12B local) | 0.336 | 0.443 | 0.450 |
| Naive | Heterogeneous | 0.011 | 0.241 | 1.000 |
| Naive | Uniform (Sonnet 4.6) | 0.010 | 0.228 | 1.000 |
| Naive | Uniform (12B local) | 0.016 | 0.317 | 0.988 |

[†] Response-level metrics (hat dialogue embeddings); not directly comparable to Table 3 (knowledge store embeddings).

roles. An LLM-as-judge evaluation of response quality and a more fine-grained ablation of deliberation structure would complement the embedding-based metrics reported here.

Embedding granularity. The per-response discrimination ratios are significantly above 1.0 overall ($\bar{r} = 1.024$, $p = 2.34 \times 10^{-26}$, $d = 0.491$; Table 5), with five of six hats individually significant (White, Red, Black, Yellow, Green). Blue Hat does not reach individual significance, reflecting that facilitative language overlaps more with general discourse in sentence-level embedding space. Finer-grained analysis, such as style-specific feature extraction or metrics designed for distributional rather than centroid-based comparison, may reveal stronger role signals for these hats. The current metrics establish a lower bound on measurable role adherence.

Self-diffusion artifacts. In our sessions, the LLM occasionally misattributes the source hat in the JSON output of diffusion items, labeling the receiving hat as the source. This is cosmetic for session logs but could inflate source attribution counts. Post-hoc filtering of “self-attributed” diffusion items would clean this.

Lateral thinking benchmark scope. The BRAINTEASER evaluation uses the same static hat system prompts as the SCT evaluation, without dynamic per-query retrieval. Significant panel improvement was observed for two models (GPT-OSS 120B and 20B). Five models showed near-ceiling single-agent baselines (≥ 0.88), leaving insufficient room to detect panel benefit. A more informative evaluation would require models with moderate baselines

(0.60–0.85) across both puzzle types. The WP subset ($n=132$) is moderately powered; results for GPT-OSS 20B on WP are not yet available at manufacturer-recommended settings.

Clinical benchmark scope. The SCT evaluation uses static hat system prompts assembled from the BEAR instruction corpus, rather than dynamic per-query retrieval. As discussed in §5, the structural similarity of SCT questions (all sharing the same clinical reasoning format) means per-query retrieval would return nearly identical instruction sets across questions, making a retrieval ablation uninformative. The benchmark therefore evaluates the deliberation structure BEAR enables rather than the retrieval mechanism specifically. A future study with clinical domain-specific instruction corpora would more directly test BEAR’s retrieval contribution in clinical settings.

Behavioral guarantees. BEAR provides *guidance* to the LLM, not *guarantees*. An LLM may deviate from retrieved instructions under adversarial prompting. The framework should be used as one layer in a defense-in-depth strategy alongside output filtering and runtime guardrails.

8 Related Work

Retrieval-Augmented Generation. RAG was introduced by Lewis et al. [2020] to augment LLM generation with retrieved passages from external knowledge bases, building on dense passage retrieval [Karpukhin et al., 2020]. Standard RAG retrieves *factual knowledge*; BEAR applies the same infrastructure to *behavioral instructions* that shape how the LLM responds.

System instruction retrieval. Instruction-Tool Retrieval (ITR) [Franko, 2025] proposes retrieving system-prompt fragments to dynamically compose runtime prompts for agentic LLMs (no public implementation is available; our characterization follows the published description). BEAR shares this foundational insight but contributes a principled *behavioral governance* framework: typed instruction schema with explicit priorities, scope conditions combining OR and AND logic, an instruction relationship graph for conflict resolution and dependency management, mandatory injection guarantees, and behavioral evolution through retrieval-gap monitoring. Where ITR frames instruction retrieval as an efficiency optimization, BEAR frames it as infrastructure for scalable, context-adaptive prompt engineering. ID-RAG [Fan et al., 2025] retrieves identity traits from a knowledge graph but targets factual attributes rather than behavioral instructions.

Prompt engineering and automation. Chain-of-thought prompting [Wei et al., 2022] and least-to-most prompting [Zhou et al., 2022] demonstrated that prompt structure shapes LLM behavior. DSPy [Khattab et al., 2023] compiles declarative LM pipelines through offline optimization; BEAR complements DSPy by selecting which (possibly DSPy-optimized) instructions to include for a given context. Universal Conditional Logic [Mikinka, 2025] formalizes deterministic conditional prompt assembly; BEAR extends this with semantic retrieval for novel contexts. Zheng et al. [2024] found that simple static persona assignments

generally do not improve LLM performance, motivating BEAR’s approach of retrieving detailed, context-specific behavioral instructions.

Multi-agent systems and LLMs. CrewAI [CrewAI, 2024] assigns static role-based personas to agents; BEAR’s retrieval pipeline enables dynamic, context-sensitive behavioral activation within each role. ReAct [Yao et al., 2023] interleaves reasoning and action; BEAR controls how agents behave (persona, constraints, style), complementing frameworks that control what they do (tool selection, planning).

Memory and behavioral adaptation. Park et al. [2023] created generative agents with memory streams and reflection. BEAR’s experiential memory engine addresses the same problem domain but converts significant experiences into typed instruction objects that participate in the retrieval and governance pipeline (with scope conditions, priority scores, and conflict resolution eligibility) rather than appending memories as raw text. MemOS [Li et al., 2025] provides lifecycle governance across parametric, activation, and plaintext memory types, but does not filter incoming knowledge through agent-specific cognitive role profiles before storage.

Cognitive differentiation in multi-agent deliberation. Prior AI-augmented brainstorming work [Gero et al., 2023, Suh et al., 2024] focuses on idea generation quality but uses uniform prompting across participants. BEAR’s contribution is cognitive-mode differentiation through behavioral instruction retrieval: each participant maintains a distinct cognitive role across arbitrary topics, and cross-participant knowledge flows through role-appropriate filters. The Six Thinking Hats methodology provides a natural structure for evaluating this differentiation, as each hat’s role is explicitly defined and its adherence is verifiable.

Knowledge flow in multi-agent systems. Prior multi-agent communication protocols [Hong et al., 2024] use structured messaging between specialized agents but do not apply behavioral role profiles to filter what each agent retains. While behavioral instruction retrieval is established [Franko, 2025], cognitive-mode diffusion is, to our knowledge, the first application of retrieved role instructions as filters governing inter-agent knowledge absorption and adaptation to each agent’s cognitive role.

9 Conclusion

Behavioral instruction retrieval can govern not only how LLM agents respond but how knowledge flows between agents in a multi-agent system. We have demonstrated this through cognitive-mode knowledge diffusion, in which BEAR-retrieved instructions serve as a cognitive filter governing inter-agent knowledge flow, determining what each agent retains and how it adapts incoming information to its cognitive role. Evaluated on a Six Thinking Hats deliberation panel across eight biomedical topics, the mechanism maintains distinct cognitive modes and produces role-differentiated knowledge across six agents throughout multi-turn sessions.

BEAR-guided diffusion produces knowledge stores roughly half the size of their naive counterparts and with little overlap across hats (5.5% versus 70.6% cross-hat NN overlap). A three-way ablation against an embed-only baseline shows that cosine deduplication accounts for about two-thirds of the role-differentiation gain and cognitive reframing for the remaining third, confirming that the LLM-driven reframing step contributes a directional effect that simple deduplication cannot produce. Across 517 responses the role discrimination ratio is significantly above 1.0 ($\bar{r} = 1.024$, $p = 2.34 \times 10^{-26}$) and is unchanged between BEAR-guided and naive conditions, evidence that the behavioral layer governs agent responses independently of what is in the knowledge layer. Constant-model controls in which all six hats run on the same Claude Sonnet 4.6 or a 12B local model preserve or strengthen inter-hat differentiation, ruling out model heterogeneity as the source of the effect.

Applied to two external reasoning benchmarks the same role-prompt structure produces task- and model-dependent panel effects. On SCT-Bench, four of nine models improve significantly: Claude Haiku 4.5 and MedGemma 27B both rise from below the medical-student benchmark to near or above the resident level, while Claude Opus 4.6 and Sonnet 4.6 show smaller but significant gains. Two smaller models degrade. On BRAINTEASER, GPT-OSS 120B improves on both sentence and word puzzles and GPT-OSS 20B improves on sentence puzzles. The pattern is consistent across both benchmarks: panel benefit correlates with per-role capability, so the panel amplifies cognitive differentiation that the underlying model is already capable of expressing rather than manufacturing it.

Together these results extend BEAR from a response-governance framework to a knowledge-flow governance framework applicable to any multi-agent system with differentiated cognitive roles.

Future work will formalize the cognitive filter pattern as a core BEAR primitive (**BearDrivenFilter**), integrate token-budget-aware composition (as demonstrated in ITR [Franko, 2025]) for cost-sensitive deployments, and investigate alternative retrieval strategies including LLM-native constrained decoding [Jain et al., 2024]. Work in progress combines knowledge diffusion with behavioral evolution in a multi-agent creative environment.

Acknowledgements

No funding was received for this work.

Declaration of Interest Statement

The author reports there are no competing interests to declare.

Data Availability Statement

The session logs, evaluation scripts, and result files supporting this paper are available at <https://github.com/snhwang/paper-knowledge-diffusion-artifacts> (session logs in `bear_parlor/session_logs/`, evaluation scripts in `evals/` and `benchmarks/`, result files in `results/`). The BEAR framework source code is available at <https://github.com/snh>

wang/bear. Both repositories are released under the Open Core Ventures Source Available License (OCVSAL) v1.0; commercial production use requires a licensing agreement from the Pennsylvania State University Office of Technology Transfer.

References

- Peter Alping. Disease-modifying therapies in multiple sclerosis: A focused review of rituximab. *Basic & Clinical Pharmacology & Toxicology*, 133(5):550–564, 2023. doi: 10.1111/bcpt.13932. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/bcpt.13932>.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901, Red Hook, NY, USA, 2020. Curran Associates, Inc. URL <https://arxiv.org/abs/2005.14165>.
- Bernard Charlin, Jacques Tardif, and Henny P.A. Boshuizen. The script concordance test: A tool to assess the reflective clinician. *Teaching and Learning in Medicine*, 12(4):189–195, 2000. doi: 10.1207/S15328015TLM1204.4.
- Harrison Chase. LangChain. <https://github.com/langchain-ai/langchain>, 2022. Open-source framework for LLM application development.
- CrewAI. CrewAI: Framework for orchestrating role-playing autonomous AI agents. <https://github.com/crewAIInc/crewAI>, 2024. Open-source multi-agent orchestration framework.
- Edward de Bono. *Six Thinking Hats*. Little, Brown and Company, Boston, MA, 1985. URL <https://www.worldcat.org/title/six-thinking-hats/oclc/12281729>.
- Anna-Victoria De Keersmaecker, Eline van Doninck, Inez Wens, Yousra El Ouamari, Veronica Popescu, Guy Laureys, Melissa Cambron, Miguel D’Haeseleer, Lander Willem, Judith Derdelinckx, Tatjana Reynders, and Barbara Willekens. Recent advances in interventions targeting remyelination and a systematic review of remyelinating effects of approved disease-modifying treatments for multiple sclerosis. *European Journal of Neurology*, 32(11):e70397, 2025. doi: 10.1111/ene.70397. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/ene.70397>.
- Michael Diehl and Wolfgang Stroebe. Productivity loss in brainstorming groups: Toward the solution of a riddle. *Journal of Personality and Social Psychology*, 53(3):497–509, 1987. doi: 10.1037/0022-3514.53.3.497.
- Dale Ding. Endovascular mechanical thrombectomy for acute ischemic stroke: A new standard of care. *Journal of Stroke*, 17(2):123–126, 2015. doi: 10.5853/jos.2015.17.2.123. URL <http://www.j-stroke.org/journal/view.php?number=69>.
- Alexander Fan et al. ID-RAG: Identity retrieval-augmented generation for long-horizon persona coherence in generative agents. arXiv preprint arXiv:2509.25299, 2025. URL <https://arxiv.org/abs/2509.25299>.

- Uria Franko. Dynamic system instructions and tool exposure for efficient agentic LLMs. arXiv preprint arXiv:2602.17046, 2025. URL <https://arxiv.org/abs/2602.17046>.
- Katy Ilonka Gero, Tao Long, and Lydia B Chilton. Social dynamics of AI support in creative writing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–18, New York, NY, USA, 2023. ACM. doi: 10.1145/3544548.3580782.
- John Hardy. Alzheimer’s disease: Treatment challenges for the future. *Journal of Neurochemistry*, 169(8):e70176, 2025. doi: 10.1111/jnc.70176. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/jnc.70176>.
- Matteo Haupt, Stefan T. Gerner, Mathias Bähr, and Thorsten R. Doeppner. Neuroprotective strategies for ischemic stroke—future perspectives. *International Journal of Molecular Sciences*, 24(5):4334, 2023. doi: 10.3390/ijms24054334. URL <https://doi.org/10.3390/ijms24054334>.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiwu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. MetaGPT: Meta programming for a multi-agent collaborative framework. arXiv preprint arXiv:2308.00352, 2024. URL <https://arxiv.org/abs/2308.00352>.
- Scott N. Hwang. Retrieval-governed context: Scope-gated selection of instructions and tools for LLMs and intelligent agents, 2026. URL <https://doi.org/10.5281/zenodo.19705464>. Preprint. Submitted to ACM Transactions on Intelligent Systems and Technology.
- Geoffrey Irving, Paul Christiano, and Dario Amodei. AI safety via debate. arXiv preprint arXiv:1805.00899, 2018. URL <https://arxiv.org/abs/1805.00899>.
- Palak Jain, Livio Baldini Soares, and Tom Kwiatkowski. From RAG to RICHES: Retrieval interlaced with sequence generation. arXiv preprint arXiv:2407.00361, 2024. URL <https://arxiv.org/abs/2407.00361>.
- Yifan Jiang, Filip Ilievski, Kaixin Ma, and Zhivar Sourati. BRAINTEASER: Lateral thinking puzzles for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14317–14332, Singapore, 2023. ACL. URL <https://arxiv.org/abs/2310.05057>.
- Georgi Karadzhov, Andreas Vlachos, and Tom Stafford. The effect of diversity on group decision-making. arXiv preprint arXiv:2402.01427, 2024. URL <https://arxiv.org/abs/2402.01427>.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, 2020. ACL. URL <https://arxiv.org/abs/2004.04906>.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Mober, et al. DSPy:

- Compiling declarative language model calls into self-improving pipelines. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 16114–16149, Red Hook, NY, USA, 2023. Curran Associates, Inc. URL <https://arxiv.org/abs/2310.03714>.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, Red Hook, NY, USA, 2020. Curran Associates, Inc. URL <https://arxiv.org/abs/2005.11401>.
- Zhiyu Li, Chenyang Xi, Chunyu Li, Ding Chen, et al. MemOS: A memory OS for AI system. arXiv preprint arXiv:2507.03724, 2025. URL <https://arxiv.org/abs/2507.03724>.
- J. Liu, J. Ha, M. Abikenari, et al. Adaptive immunotherapeutic paradigms in diffuse midline glioma: Integrating epigenetic reprogramming, neuron–glioma interactions, and tumor microenvironment modulation. *Journal of Neuro-Oncology*, 176:144, 2026. doi: 10.1007/s11060-025-05347-9. URL <https://doi.org/10.1007/s11060-025-05347-9>.
- Liam G. McCoy et al. Assessment of large language models in clinical reasoning: A novel benchmarking study. *NEJM AI*, 2:AIdbp2500120, 2025. doi: 10.1056/AIdbp2500120. URL <https://ai.nejm.org/doi/full/10.1056/AIdbp2500120>.
- Anthony Mikinka. Universal conditional logic: A formal language for prompt engineering—foundational specification, core validation, and research roadmap. arXiv preprint arXiv:2601.00880, 2025. URL <https://arxiv.org/abs/2601.00880>.
- Leo F. Nonnenbroich, Sara M. Bouchal, Eva Millesi, Julian S. Rechberger, Soumen Khatua, and David J. Daniels. H3K27-altered diffuse midline glioma of the brainstem: From molecular mechanisms to targeted interventions. *Cells*, 13(13):1122, 2024. doi: 10.3390/cells13131122. URL <https://doi.org/10.3390/cells13131122>.
- OpenAI. GPT-4 technical report. arXiv preprint arXiv:2303.08774, 2023. URL <https://arxiv.org/abs/2303.08774>.
- Joon Sung Park, Joseph C O’Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *ACM Symposium on User Interface Software and Technology (UIST)*, pages 1–22, New York, NY, USA, 2023. ACM. URL <https://arxiv.org/abs/2304.03442>.
- Nicolò Scarano, Francesca Musumeci, Bruno Casini, Chiara Brullo, Pasquale D’Ursi, Paola Fossa, Silvia Schenone, and Elena Cichero. Alzheimer’s disease etiology hypotheses and therapeutic strategies: A perspective. *International Journal of Molecular Sciences*, 26(14): 6980, 2025. doi: 10.3390/ijms26146980. URL <https://pubmed.ncbi.nlm.nih.gov/40725224/>.
- Kai M. Schubert, Anthony G. Marson, Eugen Trinkä, and Marian Galovic. Epilepsy as a dynamic disease: Toward actionable, individualized seizure risk prediction. *Epilepsia*, 66

- (12):4654–4666, 2025. doi: 10.1111/epi.18602. URL <https://pubmed.ncbi.nlm.nih.gov/40824375/>.
- Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. Luminate: Structured generation and exploration of design space with large language models for human-AI co-creation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–22, New York, NY, USA, 2024. ACM. doi: 10.1145/3613904.3642400.
- Matthew C. Walker. State-of-the-art gene therapy in epilepsy. *Current Opinion in Neurology*, 38(2):128–134, 2025. doi: 10.1097/WCO.0000000000001349. URL <https://pubmed.ncbi.nlm.nih.gov/39917784/>.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171, 2023. URL <https://arxiv.org/abs/2203.11171>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837, Red Hook, NY, USA, 2022. Curran Associates, Inc. URL <https://arxiv.org/abs/2201.11903>.
- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649, New York, NY, USA, 2024. ACM. doi: 10.1145/3626772.3657878. URL <https://arxiv.org/abs/2309.07597>.
- Lin Yang et al. MedGemma: An open medical foundation model. arXiv preprint arXiv:2507.05201, 2025. URL <https://arxiv.org/abs/2507.05201>.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, pages 1–21, Online, 2023. OpenReview.net. URL <https://arxiv.org/abs/2210.03629>.
- Mingqian Zheng, Jiaxin Pei, and David Jurgens. When “a helpful assistant” is not really helpful: Personas in system prompts do not improve performances of large language models. arXiv preprint arXiv:2311.10054, 2024. URL <https://arxiv.org/abs/2311.10054>.
- Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625, 2022. URL <https://arxiv.org/abs/2205.10625>.