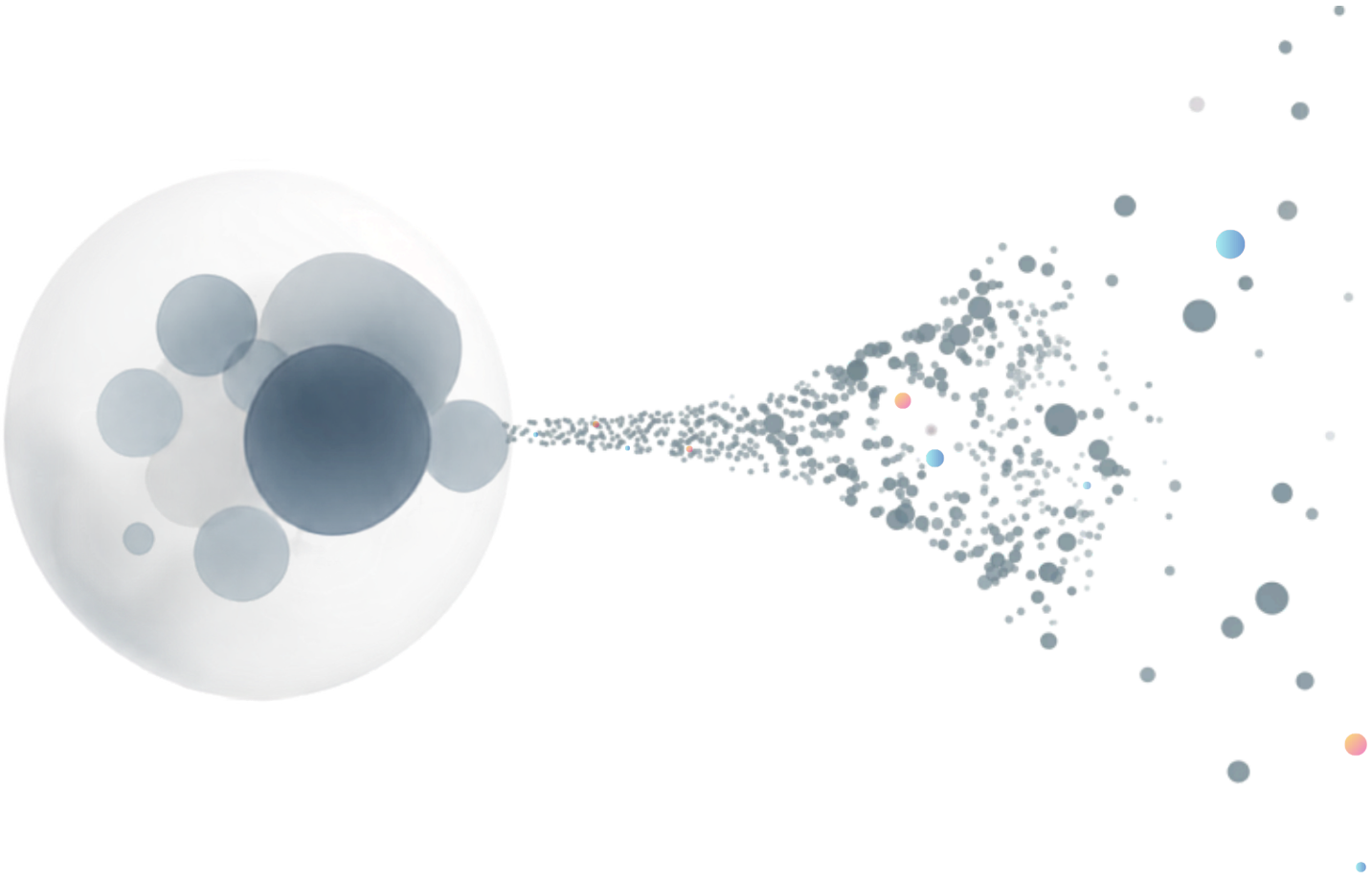
The Effect of Evidence Distribution Across Conversational Memory and Documents on QA Agent Answer Correctness

Master of Science Thesis

DSAIT
Todor Mladenović



TU Delft

Master of Science Thesis

Data Science and Artificial Intelligence Technology (DSAIT)

by

Todor Mladenović

Student Number: 4772083

Responsible Professor & Supervisor:	Julián Urbano
External Thesis Committee Member:	Martin Skrodzki
Faculty:	Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS) at TU Delft

The Effect of Evidence Distribution Across Conversational Memory and Documents on QA Agent Answer Correctness

Todor Mladenović
Delft University of Technology
Delft, Netherlands

Abstract

Personalized QA agents increasingly retrieve from both external documents and private conversational memory. This raises a basic design question: does answer correctness depend on where the supporting evidence is stored? Existing evaluations make this difficult to answer because memory and document conditions often differ in their questions, facts, or available sources. We construct a controlled evaluation setting from MultiHop-RAG in which the same questions and evidence are held fixed while golden evidence is placed entirely in documents, entirely in conversational memory, or split across both. We evaluate four retrieval architectures across three persona-specific memory corpora. The results show that evidence placement does not exert a uniform penalty. Instead, its effect is significantly moderated by the topic match of the query to the persona's memory profile. While in-domain memory retrieval is highly competitive with document retrieval, answer correctness collapses during off-topic queries. This performance drop suggests that personalized agents may create soft algorithmic filter bubbles, penalizing user curiosity with lower accuracy when exploring unfamiliar domains.

1 Introduction

Retrieval-Augmented Generation (RAG) has become essential for grounding large language models in external knowledge. It significantly improves factual accuracy for knowledge-intensive tasks [14, 41]. Modern personalized AI systems increasingly integrate RAG frameworks into Question-Answering (QA) agent architectures that combine multiple knowledge sources to enhance user satisfaction [15]. Commercial conversational agents now routinely combine these capabilities for long-term personalized assistance. These systems must retrieve knowledge not only from documents but also from conversational memory (a corpus of past conversational turns containing facts such as mentioned dates, stated preferences, or project details shared across sessions).

A growing body of evidence suggests that where and how facts are situated affects an agent's ability to retrieve and reason over them. Even within document collections, organization of knowledge across heterogeneous sources affects retrieval quality. Systems retrieving from multiple, heterogeneous sources face challenges that don't exist when evidence is centralized [7, 37], with specialized retrieval strategies often needed to optimize performance across diverse data properties.

Beyond source organization, the structural and contextual packaging around facts shapes retrieval and reasoning performance. The presence of distractors (retrieved passages that are semantically similar to the query but irrelevant) [9, 10] and stylistic choices of surrounding context [27] influence reasoning and retrieval quality. Isolating facts from their surrounding context by splitting texts into smaller chunks can harm retrieval when individual chunks lack sufficient context [2]. Conversely, reasoning suffers from the 'Lost in the middle' phenomenon when too much surrounding context is supplied, with fact positioning having a meaningful impact [16].

Memory and documents differ as knowledge sources in ways that may affect retrieval. Documents are deliberately authored to communicate information to a broad audience, typically following editorial conventions that structure facts within a coherent narrative. Conversational memory, by contrast, arises from private exchanges that were (usually) not intended to serve as a knowledge source, and is shaped by personalization in ways documents are not. Users of LLM-powered search already query more toward information that confirms their existing views than they do with conventional search [28], so the resulting conversational histories concentrate around recurring topics rather than the broad topical coverage documents are edited to provide.

Personalization also extends to how facts are expressed: each user's memory reflects their own voice and phrasing, so stylistic variation across a memory corpus is likely greater than across the same evidence in an edited document corpus. Together, this topical and stylistic personalization may create retrieval conditions that differ substantially from those in a document corpus, yet the degree to which these differing source characteristics affect retrieval and downstream reasoning in agentic QA remains understudied.

Most existing evaluations treat these two sources in isolation. Focusing either on documents [12, 30, 39] or more recently on conversational memory [17, 36, 41]. Although the associated datasets specify distinct demands for each setting, fact retrieval and reasoning remain the common underlying challenge.

A smaller but growing line of work examines systems that handle multiple knowledge source types simultaneously [33, 34]. For example, evaluations of systems on questions requiring memory (limited to personal details mentioned in past conversations), documents, or combinations of both have shown that overall performance is sensitive to the source setting, with document retrieval reported as particularly poor when persona retrieval is also required [34]. However, these works do not isolate the effect of evidence distribution. The question set for each condition (persona-only, document-only, both) differs, and the persona facts themselves are distinct from the facts in documents. While such studies demonstrate that systems face different challenges depending on which knowledge sources must be accessed, they confound the effect of source characteristics with differences in the questions and evidence. Moreover, agents differ in how they organize retrieval across sources (e.g. shared versus separate indexes, single-pass versus iterative retrieval), and it

remains unclear whether some designs are more robust to evidence placement than others.

Given these structural differences, relocating the same evidence between memory and documents while holding the questions fixed would directly test whether source placement affects agent performance. Allowing us to answer our main research question:

RQ1: How does the distribution of evidence across a document corpus and conversational memory affect an agent's answer correctness?

We supplement RQ1 with three sub-questions:

RQ1a To what extent does the effect of evidence distribution on answer correctness vary across agent retrieval architectures?

RQ1b To what extent does persona-specific memory composition moderate the effect of evidence distribution on answer correctness?

RQ1c To what extent are differences in answer correctness across evidence distributions accounted for by retrieval success?

However, no existing dataset or evaluation protocol supports such a controlled manipulation. Thus in order to answer *RQ1* we must first construct a setting that enables this, which we do by answering:

RQ2: How can a controlled evaluation setting be constructed that isolates the effect of where evidence resides (in a document corpus, in conversational memory, or split across both)?

To answer *RQ2*, we construct three controlled variants of a dataset derived from MultiHop-RAG [30], allocating the same underlying evidence to one of the three distribution conditions such that each piece appears in only one location per variant. Then for *RQ1*, we evaluate four agent architectures across all three variants in a controlled experiment designed to maximize statistical power under a fixed evaluation budget, crossed with three evaluation personas. This lets us test *RQ1a* and *RQ1b*. Every condition also logs recall, which we use to test *RQ1c*. In every condition the agents have access to both memory and documents, so that observed differences reflect where evidence resides rather than which sources are available.

Our contributions are:

(1) A controlled QA dataset and construction methodology that disentangles evidence distribution from question content, enabling within-question comparison across source configurations.
(2) A rigorous experimental design that enables controlled comparison of distribution effects across four architecturally distinct agents while maximizing statistical power.
(3) Empirical findings indicating that the association between evidence distribution and answer correctness depends on persona-specific memory context, particularly the topic match between the question and the persona's memory profile. We find no evidence that this association varies across agent architectures. Specifically, users exploring topics outside their dominant memory profiles face lower answer accuracy than when querying within their frequently discussed domains.

The remainder of the paper is organized as follows. §2 reviews the literature on RAG, memory retrieval, personalized multi-source agents, and the impact of fact packaging. §3 describes the construction and validation of our controlled dataset. §4 details the four agent architectures, the D-optimal experimental design, and the statistical models used to analyze the results. §5 reports the experimental findings and answers the research questions by interpreting those findings. Additionally, here we discusses limitations and outline directions for future work. §6 concludes the paper before §7 addresses the ethical considerations involved in carrying out this study. Finally §8 documents the use of generative AI during the course of the study.

2 Related Work

2.1 Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) extends a language model with non-parametric memory by retrieving relevant passages from an external corpus at inference time and conditioning generation on them [14]. The core pipeline has two components: a retrieval component that indexes a document collection and returns the passages most relevant to a query, and a generator that produces a response conditioned on both the original query and the retrieved evidence. By grounding generation in retrieved documents, RAG reduces hallucination and supports questions whose answers depend on knowledge absent from the model's parameters.

Three families of retrieval methods are in common use. Sparse methods (such as BM25) rank documents by weighted term overlap and require no learned representations. Dense methods encode queries and passages into continuous vectors and retrieve by nearest-neighbour search in the embedding space, allowing them to match semantically related text that shares no surface-level vocabulary. Lastly, hybrid methods combine the previously mentioned families to produce a single ranked retrieved set. Surveys of RAG systems treat these as the standard retrieval options, typically with a dense bi-encoder as the learned component and an optional reranker [41]. Direct comparisons in conversational QA find that hybrid BM25 and reranking often outperform vanilla dense RAG, so dense retrieval is a convenient default rather than a uniformly dominant choice [1]. Zero-shot retrieval quality across heterogeneous domains is itself non-trivial: BEIR shows that models which perform well in-domain frequently fail to generalize, and that a simple lexical baseline remains competitive [31].

The generator LLM determines how faithfully the system reasons over retrieved context as opposed to relying on parametric knowledge. While closed-source API-served models may offer strong performance, they introduce a fundamental reproducibility challenge: their update schedules are opaque and their behaviour has been shown to shift substantially across versions even over short timescales [6]. Open-weight models provide fixed, publicly accessible checkpoints which permit easier reproduction/auditing of results. The Llama 3 family demonstrates this is achievable alongside strong performance on reasoning-intensive tasks that are central to RAG evaluation [11].

Dense retrieval requires an encoder that maps both queries and passages into a shared embedding space such that relevant pairs are geometrically close. Sentence transformers [25], which fine-tune a transformer encoder with contrastive objectives on sentence pairs,

have become the standard architecture for this task. Their quality-to-efficiency trade-off is well characterized by BEIR [31], which evaluates zero-shot generalization across diverse retrieval tasks, and the MTEB leaderboard [19], which explicitly stratifies models by the balance they strike between quality and inference cost. Models in the MiniLM family occupy the "speed and performance" tier of the MTEB leaderboard, achieving scores competitive with encoders several times larger while requiring only a fraction of their memory and compute [19].

RAG research has primarily been conducted against static document corpora, an assumption reflected throughout the benchmarks and evaluation frameworks that constitute the field [1, 41]. The distinct challenges that arise when the retrieval target is a user's conversational history, with its personal specificity, topical concentration, and evolving relevance, are discussed in the following section.

2.2 Memory

Conversational memory (the accumulated interaction history between a user and their assistant) presents distinct retrieval challenges compared to general document corpora. Whereas encyclopedic collections are designed to cover diverse topics, conversational memory naturally concentrates around a user's recurring interests and goals. This can be seen in literature on **filter bubbles**: a phenomenon resulting in users being increasingly exposed to content aligning with their existing views, with direct evidence showing that users of LLMs issue a high rate of biased, view-confirming queries [28]. Over a series of sessions, that querying pattern would populate memory with many passages that are semantically close to any later in-domain query. Semantic caching work reports a related pattern: users commonly ask semantically similar queries, especially in coding workloads [32].

This topical concentration is expected to degrade answer correctness through two pathways. First, **distractor** passages (memory passages that are semantically similar to a query yet do not contain the required evidence) compete with golden evidence for a fixed retrieval budget and can displace it from the retrieved set entirely. Second, distractors that are retrieved alongside golden evidence have been found to interfere with the generator LLM [9, 10].

Consistent with this, prior multi-source evaluations observe that document retrieval quality drops when memory retrieval is also required [34]. Yet no existing work isolates the effect of this topical density on answer correctness while controlling for question difficulty and evidence content. A gap our cross-persona design addresses by evaluating each question against memory corpora of varying in-domain density.

Existing RAG benchmarks reinforce the separation between the two source types. Datasets designed for long-term memory evaluation, such as LongMemEval [36] and LoCoMo [17], probe episodic, session-referential questions that can only be answered from personal interaction history, while document retrieval benchmarks, such as MultiHop-RAG [30], HotpotQA [39], and FRAMES [12], probe factual multi-hop reasoning that requires chaining evidence across documents. Distinct benchmarks serve distinct evaluation targets [41] but few require an agent to locate and combine both kinds of evidence within a single evaluation, and none allow the same evidence to be relocated between them. This is why our setting must be constructed rather than adopted, as described in §3.

2.3 Personalized Agents

The retrieval paradigms discussed so far, grounding answers in a document corpus (§2.1) and surfacing relevant fragments of a user's interaction history (§2.2), are increasingly combined within a single system. Personalized agents treat conversational memory and external documents as complementary knowledge sources, retrieving from both to produce responses that are simultaneously factually grounded and tailored to the individual user. A recent survey studies this trajectory explicitly, framing the field as a progression *from RAG to agent*: from static retrieve-then-read pipelines toward systems that actively manage memory, plan retrieval, and adapt their behaviour to the user [15]. The same shift is visible across concrete systems that inject user-specific signals into retrieval and generation, whether through dedicated user-modelling agents [43] or personalized graph-structured knowledge [5].

A defining feature of these agents is that personalization requires reconciling multiple heterogeneous sources. Because a user's memory and the document corpus differ in structure, reliability, and relevance to any given query, an agent must decide not only what to retrieve but from where. Existing work approaches this through source planners that select among the available sources on a per-query basis [33, 34]. This framing presumes that sources remain separately addressable, and in those settings the presumption is reasonable, since a persona is typically a small set of short statements rather than a corpus. Once conversational memory is itself corpus-scale, however, whether memory and documents should occupy one index at all becomes a live question. Comparisons of shared against separate indexes exist for heterogeneous document collections [35], but to our knowledge not where one of the sources is conversational memory. We therefore treat source organization as an open design decision rather than an implementation detail, and evaluate both configurations directly.

Two complementary lines of work refine how such agents access and use evidence. The first specializes the memory component itself: rather than a flat vector store, long-term memory is organized and maintained through structured or reflective mechanisms, including temporal knowledge graphs [24], learned memory construction and consolidation [22, 29], and global memory modules that summarize history to guide retrieval [23]. The second makes the retrieval control flow agentic: instead of a single retrieval pass, the model reasons about, evaluates, and revises its own retrieval. This builds on the reasoning-and-acting paradigm [40] and is realized for RAG by self-reflective retrieval that decides when and what to retrieve [3], corrective retrieval that grades retrieved evidence and triggers remedial action when it is inadequate [38], and iterative schemes that refine retrieval across successive rounds [26].

These two axes, the memory mechanism and the retrieval control flow, are largely orthogonal. A specialized memory store can still be queried in a single pass, and an agentic loop can run over a general-purpose store.

Although these paradigms are widely adopted, they are typically introduced and evaluated in isolation, each against a simple baseline, and are rarely compared with one another under matched evidence

conditions. This makes it difficult to attribute performance to a particular design choice rather than to differences in datasets, base models, or evidence availability.

2.4 Impact of Fact Packaging for RAG

Whether a RAG system retrieves the correct fact depends not only on whether that fact exists in the corpus, but on how it is physically packaged into retrievable units. Segmenting a corpus into chunks creates a fundamental trade-off. Finer-grained chunks surface target facts more precisely because the retrieval signal is not diluted by surrounding off-topic content, but they strip the local context an LLM needs to reason about the fact coarser chunks preserve that context at the cost of retrieval precision. Systematic evaluations of chunking strategies confirm that this trade-off is real and domain-dependent: smaller chunks (64-128 tokens) work well for concise, fact-based queries, whereas descriptive or multi-step questions benefit from larger units [20]. Practitioner analyses echo the same finding, identifying chunk granularity as one of the highest-leverage parameters for end-to-end RAG accuracy [2].

Even when retrieval succeeds, how the retrieved passages are assembled into the LLM's context determines whether the model can exploit them. Two packaging effects emerge at the generation stage. The first is positional: LLMs systematically under-use evidence placed in the middle of a long context, performing best when relevant passages appear near the start or end of the input [16]. The second concerns ordering in multi-hop questions, where the answer requires chaining evidence across passages. Performance is sensitive not only to which passages are present but to the order in which they appear, with a sequence that matches the reasoning chain being more performant [42]. Taken together, these effects show that the same retrieved content can produce different answers depending solely on how it is packaged into context.

These challenges become more acute in multi-source settings where the two corpora differ in granularity and topical distribution. In our setting, each evidence item resides in exactly one source, memory or documents, but the agent retrieves from both simultaneously. The relevant passage must surface over potentially many plausible-but-irrelevant passages from the other source. This pressure is asymmetric as conversational memory is topically concentrated by nature, so the density of distractors is expected to be higher within memory than within a general document corpus (§2.2). The two sources also differ structurally: memory is often indexed at the conversational-turn pair level [36] while documents are segmented into fixed 256-token chunks, so their passages vary in information density and context coverage in ways that can independently skew which source's passages appear more relevant to a given query [20]. Existing packaging studies largely examine single, homogeneous corpora but our experiment isolates the effect of source-level organizational choices by holding evidence content fixed across conditions.

3 Dataset Construction

This section addresses **RQ2** by describing the construction of our controlled dataset, which enables within-question comparison of evidence distribution effects. The central manipulation produces three distribution of evidence *variants* for each question that differ only in where the golden evidence resides: entirely in the document corpus (**document-only**), entirely in conversational memory (**memory-only**), or split evenly across both (**integrated**). Holding the question and its evidence fixed while relocating that evidence isolates the effect of source placement from question content.

We start with MultiHop-RAG as the source corpus (§3.1), apply oracle-context and no-context screening to identify questions where retrieval effects can be meaningfully observed (§3.2), and assign questions to three evaluation personas based on evidence domains (§3.3). We then sample the evaluation question set (§3.4), generate synthetic conversational sessions that embed golden evidence (§3.5), and curate per-persona memory corpora with controlled in-domain topic composition (§3.6). Finally, we detail how the three variants are constructed (§3.7) and audit whether stochastic generation parameters introduced confounds (§3.8).

3.1 Source Dataset

We derive our dataset from MultiHop-RAG [30], a collection of 2,556 multi-hop reasoning questions[1] over news articles. Each question requires 2–4 supporting evidence documents to answer, drawn from a corpus of 609 news articles spanning September 26, 2023 through December 26, 2023. The corpus questions often require article metadata (source, title, published-at timestamp) in addition to the key information to provide contextual grounding.

MultiHop-RAG distinguishes four question types (examples in Appendix A):

- **Comparison query**: comparing attributes or events across entities
- **Inference query**: synthesizing information to derive conclusions
- **Temporal query**: reasoning about event ordering or time-sensitive relationships
- **Null query**: unanswerable questions designed to test abstention

Null queries have no golden evidence to relocate, so they cannot support the distribution manipulation and are not part of the main evaluation. We report them separately in Appendix B.

For our study, MultiHop-RAG provides an appropriate foundation because: (i) the evidence is already decomposed into discrete items that can be independently allocated across memory and documents, (ii) the questions are answerable with short factual responses (suitable for Exact Match evaluation), and (iii) the news-article format enables realistic synthetic conversational sessions where users might discuss current events with an assistant.

3.2 Controlling Parametric Knowledge and In-Context Reasoning

Two factors could prevent us from observing the effect of evidence distribution on Answer Correctness. First, if the model can answer questions without any retrieved evidence, performance reflects

[1] *Multi-hop* usually denotes a question whose answer requires combining several pieces of evidence, often through chained retrieval steps in which one retrieved fact determines what to retrieve next. MultiHop-RAG questions require evidence from multiple documents, but the required ordering is not explicit, so we use the term in the weaker sense of multi-evidence synthesis.

parametric knowledge[2] rather than retrieval effectiveness. Second, if the model cannot answer questions even with perfect evidence, poor performance reflects reasoning limitations rather than retrieval quality. Prior work has used no-context and oracle-context conditions as baselines to bound RAG system performance [1, 4]. We extend this approach by using these conditions to filter our evaluation set, retaining only the questions where retrieval effects can be meaningfully observed.

We employ two validation agents to implement this filtering:

- **Oracle-Context Agent**: Receives each question with all required evidence provided directly in the prompt (no retrieval). Questions that this agent cannot answer correctly are excluded, ensuring the model possesses sufficient reasoning capacity.
- **No-Context Agent**: Receives each question without any evidence. Questions that this agent answers correctly are excluded, as correct answers indicate parametric knowledge rather than retrieval effectiveness.

Applying this two-stage filter to the 2,556 questions in MultiHop-RAG yields a pool of 1,115 *reasonable* questions: questions the oracle-context agent answered correctly (demonstrating adequate reasoning capacity) and the no-context agent answered incorrectly (requiring retrieval rather than parametric knowledge). This filtered pool forms the sampling frame for the evaluation question set described in §3.4.

Table 1: Distribution of reasonable questions (excluding 61 *null-queries*) by union of golden-evidence domain categories.

Domains	Count
Sports	312
Technology	279
Business & Technology	145
Entertainment	81
Business	79
Entertainment & Sports	51
Entertainment & Technology	43
Business & Sports	14
Sports & Technology	11
Business & Entertainment	7
Entertainment & Health	7
Science	6
Business & Science	5
Entertainment & Science	4
Business & Entertainment & Health	3
Science & Technology	3
Health	2
Health & Science	1
Health & Technology	1

3.3 Personas and Domain Assignment

MultiHop-RAG evidence items carry 'category' metadata (*business*, *entertainment*, *health*, *science*, *sports*, *technology*). Questions can have multiple evidence items from differing domains. By taking the union of category tags across all golden evidence items for each question that requires evidence we get Table 1 which shows the distribution of reasonable questions.

As discussed in §2.2, conversational memory accumulates around a user's recurring interests rather than covering topics evenly. Testing whether this concentration affects retrieval requires evaluating the same question against memory corpora that differ in how much of their content shares that question's domain, while holding the question and its evidence fixed. We therefore need a small set of stable memory profiles whose topical coverage overlaps each question to differing degrees.

To do this we utilize personas, a common device in synthetic memory generation used for control and consistency [17]. We construct three personas centered around the three most frequent question-domain categories from Table 1. Each persona p is identified with a topic set T_p: Tech {*technology*}, Sports {*sports*}, and Business-Technology {*business*, *technology*}.

Let C_q denote the union of category tags across all golden evidence items of question q, as tabulated above. Comparing C_q with T_p yields a three-level *topic match* label for every question–persona pair:

- **in-domain**: $C_q \subseteq T_p$ (every evidence domain lies inside the persona's profile);
- **out-of-domain**: $C_q \cap T_p = \varnothing$ (no overlap);
- **partial**: neither of the above (some but not all of the question's evidence domains match the persona).

Because C_q is a property of the question's evidence and T_p is a property of the evaluation persona, the label does not depend on where that evidence is placed. Document-only, memory-only, and integrated realizations of the same (q, p) therefore share one topic-match value. This invariance is deliberate: agents always retrieve from both stores (§4.2), so a persona's memory remains a topical distractor even when all golden evidence resides in documents.

Each question is evaluated against all three personas in the cross-persona design described in §4. The same item therefore takes different topic-match values across the three profiles, isolating the effect of memory composition on retrieval and answer correctness while controlling for question difficulty.

The personas also shape the voice of generated evidence sessions (§3.5) and the topical mix of each memory haystack (§3.6). Those uses are distinct from the trial-level label defined here: haystack composition is a property of M_p, whereas topic match records how the scored question sits relative to T_p.

3.4 Evaluation Question Set

From the pool of 1,115 reasonable questions, we sample 198 answerable questions for the experiment (66 per persona) using random sampling with minimum covariate coverage. Within each persona, the sampler ensures a minimum number of questions per `question_type` and `hop_count` level (minimum: 5) to guarantee sufficient observations for downstream analysis, then fills the remaining quota with unrestricted random draws. Sampling is stratified by question type (*comparison*, *inference*, *temporal*) and hop count (2, 3, or 4 evidence items).

[2] *Parametric knowledge* refers to the knowledge embedded directly in an LLMs "neurons" during training.

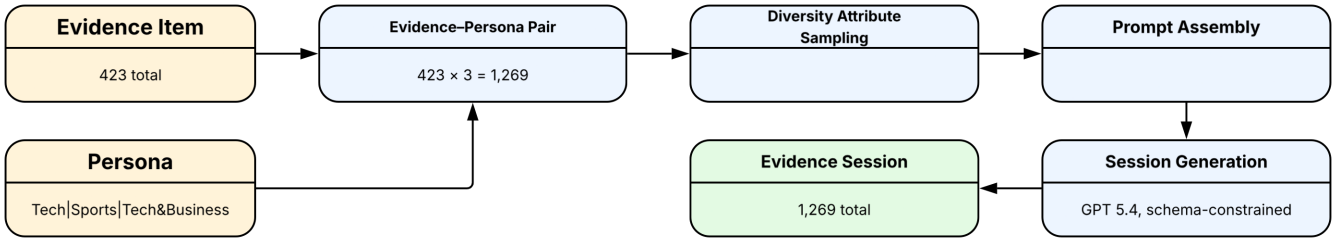


Figure 1: Session Generation Pipeline. For each evidence and persona pair, diversity attributes are sampled to assemble structured prompts. These prompts guide GPT-5.4 in generating multi-turn dialogue sessions containing all required evidence associated with the evidence item in voice of the persona.

3.5 Evidence Session Generation

To create the memory-only and integrated variants, we convert MultiHop-RAG evidence items into synthetic conversational sessions where golden evidence appears naturally in past user-assistant exchanges. Each session simulates a plausible interaction between a persona and their assistant, with the user introducing the factual content from the evidence item.

Figure 1 illustrates our approach. To ensure all questions are answerable under each persona's memory conditions, the same evidence is encoded with each persona's voice in a unique evidence session. An important note is that document metadata also needs to be explicitly mentioned in the sessions due to the nature of the MultiHop-RAG questions (§3.1). From the metadata, *published-at timestamp* is only encoded only with *temporal-query* question types to avoid memory conditions with dates in every session.

3.5.1 Diversity Attributes. To prevent generated sessions from collapsing into formulaic templates, we parametrize the generation prompt across five independently sampled diversity dimensions (prompt impression in Appendix C):

- **Scenario**: 15 scenarios categorized into professional, casual or transactional.
- **Turn range**: Target session length in user-assistant turn pairs. Options: (4,4), (4,6), (6,10), (8,10).
- **Evidence placement**: Where the key information appears in the session. Options: early, middle, late.
- **Topic drift**: How tightly the conversation stays on-topic. Options: tight, moderate, wide.
- **Evidence density**: Whether the evidence is stated in a single message or distributed across multiple turns. Options: single-message, multi-message.

Each persona has a distinct diversity profile controlling scenario category weights, turn range preferences, and topic drift distributions. These profiles ensure sessions vary realistically within each persona while remaining plausibly different across personas. A sixth dimension (the human-turn message triggering the conversation) is sampled from a pool of 10 phrasings. Together, these dimensions create substantial structural variety within each persona.

Constraint filtering ensures sensible combinations: transactional scenarios (e.g. quick fact-checks) are restricted to short turn ranges (4-6 turns), early evidence placement, tight-to-moderate topic drift, and single-message density to avoid incoherent exchanges.

3.5.2 Generation Pipeline. Sessions are generated using GPT-5.4 (due to high reasoning capacity) with structured outputs via the OpenAI Batch API [3]. The output is constrained to 4-10 alternating user-assistant turns (always an even count). Within this range, the LLM targets the sampled turn band while adapting to the complexity and informational density of the evidence item.

An example output of the generation pipeline can be seen in Figure 2. In here two things are apparent: The questions are not very natural sounding, a by-product of the synthetic generation process that produced them, and the generated session assumes a rewriting task for the Agent. Although this was not explicitly prompted and diversity attributes aimed to avoid template like sessions, manual inspection (§5.6) showed some form of re-writing in every inspected session.

3.6 Memory Corpus Curation

Let $s_p(e)$ denote the session generated for evidence item e in the voice of persona p, and let $d(e)$ denote the source document from which e was drawn.

Each persona p maintains a fixed memory corpus M_p of 500 sessions, the haystack against which every question for that persona is evaluated. It is composed of 375 in-domain sessions (75%), generated from questions whose evidence domains match T_p, and 125 off-topic sessions (25%), generated from questions outside it. The off-topic quota prevents perfect topical homogeneity, so that in-domain density is a controlled property of the corpus rather than an artifact of its construction. This mix describes how M_p is built; it is distinct from the trial-level topic-match label of §3.3, which records how the scored question's evidence domains sit relative to T_p.

Questions are drawn first from the evaluation sample, reused across personas. This supply covers each persona's off-topic quota but not its in-domain quota, so the in-domain portion is completed with sessions generated from additional same-domain questions sampled from the reasonable pool and held out of the evaluation set.

Crucially, M_p is fixed for the whole experiment. Only the golden-evidence overlay varies, per question and per variant, as described next.

[3] https://platform.openai.com/docs/guides/batch

Question. Does the TechCrunch article claim that Google engages in anticompetitive behavior within the app distribution and payment processing markets, while another TechCrunch article alleges Google's anticompetitive actions harm news publishers' revenue, and does The Verge focus on Google's use of relevance, distance, and prominence in local search rankings without mentioning anticompetitive behavior?
Gold answer: Yes.

Source evidence	**Generated memory session**	
Title. The restaurant nearest Google **Source.** The Verge **Key information.** Google's public documentation says it uses three categories to rank results that are local to a searcher: relevance, distance, and prominence. *Fields not required for this (non-temporal) item:* published time.	**User**	Quick one. I saw The Verge piece "**The restaurant nearest Google**," and it says Google's public documentation says it uses three categories to rank results that are local to a searcher: relevance, distance, and prominence. Can you give me a plain-English version of that I could text to somebody?
	Asst.	Sure: Google basically looks at three things for local results— how well a place matches what you searched for, how close it is to you, and how well-known or established it seems online.
	User	Even shorter. Like one sentence, no jargon.
	Asst.	Google shows local places based on what fits your search, what's nearby, and what seems most notable.

Figure 2: Example evaluation question (top), gold evidence (left), and the generated memory session that embeds that evidence (right).

3.7 Variant Construction

The evaluation unit is a triple (q, p, v): a question q, a persona p, and a distribution variant $v \in \{\text{document-only}, \text{memory-only}, \text{integrated}\}$. The three variants differ only in where q's golden evidence is placed relative to p's fixed haystack M_p and the shared document corpus D. Let q have golden evidence $E_q = \{e_1, \ldots, e_n\}$. Each variant v splits E_q into a memory share $E^{\text{mem}}_{q,v}$ and a document share $E^{\text{doc}}_{q,v}$ given by:

$$\left(E^{\text{mem}}_{q,v},\ E^{\text{doc}}_{q,v}\right) = \begin{cases} \left(\varnothing,\ E_q\right) & \text{if } v = \text{document-only}, \\ \left(E_q,\ \varnothing\right) & \text{if } v = \text{memory-only}, \\ \left(R_q,\ E_q - R_q\right) & \text{if } v = \text{integrated}. \end{cases} \tag{1}$$

For the integrated case, R_q is a uniformly random subset of E_q of size $\lfloor n/2 \rfloor$; if n is odd, the leftover item is assigned to memory or to documents with equal probability. The stores presented to the agent for (q, p, v) are then:

- **Memory:** $\left(M_p - s_p(E^{\text{doc}}_{q,v})\right) \cup s_p(E^{\text{mem}}_{q,v})$
- **Documents:** $D - d(E^{\text{mem}}_{q,v})$

3.8 Dataset Audit

To ensure the generated sessions contain the correct evidence, we run an LLM-as-a-judge to evaluate the generated sessions against the source evidence. GPT-OSS-120B[4] is used for it's high reasoning capacity and open-source availability to mark the exact user turn each piece of evidence from a given evidence item appears in. In the case of missing a evidence piece (key information, source, title or published-at), the LLM-as-a-judge is asked to mark the session as incomplete and report the missing evidence piece. Eight sessions were found to be incomplete missing the key information from the evidence item. A downstream sensitivity check was conducted confirming that the missing evidence pieces did not affect the analysis.

[4]GPT-OSS-120B:https://huggingface.co/openai/gpt-oss-120b

Downstream results (§5) also suggest that the generated sessions are usable as evidence: only 22 of the 198 answerable questions were never answered correctly under the memory-only and integrated variants.

The stochastic diversity attributes introduced during session generation (scenario, turn range, evidence placement, topic drift, evidence density) are randomly sampled per session to prevent formulaic templates. However, this randomization could introduce confounds if certain attribute combinations systematically affect answer correctness.

To assess potential confounds introduced by the diversity-attributes used for generation, we audit their impact of successful retrieval hits (described further in §4.3 and analysed in §5.2). The findings indicate that the topic-drift and possibly evidence-placement could influence the observations,however these effects are comparatively small and should not drive results. Further research is encouraged to observe these effects in greater detail.

That audit concerns how sessions are packaged, not whether the topic-match construct itself resembles real user memory. The 75% in-domain haystack is a controlled density, not an empirical estimate of production profiles. Qualitative inspection of retrieved chunks later suggests that same-domain sessions can be denser than a typical user history (§5.6), so findings about topic match should be read as effects under this construction rather than as a claim of real-world systems.

4 Methodology

This section describes the agent architectures and experimental design used to evaluate the effect of evidence distribution on answer correctness. Sections 4.1 and 4.2 continue to address **RQ2**: Section 4.1 details the four agent architectures, which differ in their retrieval mechanisms and knowledge organization, and Section 4.2 presents the experimental design, including the three variants, the

D-optimal blocked structure with persona crossing, and the evaluation protocol. Section 4.3 then turns to **RQ1**, modeling the distribution effect and its *RQ1a-RQ1c* extensions.

4.1 Agent Architectures

We evaluate four agent architectures, each representing a realistic and widely adopted paradigm for memory and document grounded question answering (§2.3). Rather than sampling arbitrary systems, we select architectures that vary along orthogonal design axes while sharing a common retrieval and generation backbone (§4.1.1): how memory and documents are organized in the index (a single shared store versus separate per-source stores), the mechanism used to retrieve from conversational memory (a general-purpose vector store versus a specialized hierarchical memory system), and the retrieval control flow (a single retrieval pass versus an agentic loop with relevance grading and query rewriting). Holding the backbone fixed while varying these axes lets us test whether evidence-distribution effects generalize across architectural paradigms rather than reflecting just the traits of any single design. The remainder of this section describes the shared backbone (§4.1.1) and then each architecture in turn: the Unified Store (§4.1.2), Separate Store (§4.1.3), MemPalace (§4.1.4), and Corrective Agent (§4.1.5) [5].

4.1.1 Shared Components. All four agents share core infrastructure to ensure controlled comparisons:

Language Model. All agents use Llama 3.3 70B (70 billion parameters) accessed via Groq's inference API.[6] As discussed in §2.1, open-weight models provide fixed, inspectable checkpoints that ensure reproducibility, and the Llama 3 family has demonstrated strong performance on reasoning-intensive QA tasks. Inference uses greedy decoding (temperature=0.0, seed=42).

Embedding Model. All agents encode queries and chunks with `all-MiniLM-L6-v2`,[7] a 384-dimensional sentence transformer, accessed via ChromaDB's[8] built-in ONNX runtime. As discussed in §2.1, models in the MiniLM family offer a strong speed-and-performance balance for semantic similarity tasks with local, API-free inference. The shared embedding space ensures retrieval differences reflect architectural choices rather than encoder variation.

Retrieval. Retrieval is performed by dense vector search, ranking chunks by cosine similarity to the query embedding and returning the top k per source. Although the underlying stores also implement sparse (BM25) and hybrid retrieval, all experimental runs use dense retrieval so that comparisons are not confounded by the retrieval algorithm. Dense retrieval is chosen as it is a standard, simple baseline [1]. The sole exception is MemPalace, whose specialized memory retrieval pipeline is the architectural variation under study (§4.1.4); its document retrieval remains dense.

Context Assembly. Figure 3 illustrates the generation prompt structure into which context is assembled. Retrieved passages are placed into the model's context identically across agents to ensure that performance differences stem from what each architecture retrieves rather than how retrieved evidence is presented.

[5]Code and dataset are made available at https://github.com/Todor-cmd/memdoc to support reproduction of the dataset and experiments.
[6]https://groq.com
[7]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
[8]https://www.trychroma.com

System. You are a precise question-answering system with access to two knowledge sources: conversation memory and a document corpus. You will be given a question and retrieved passages from both sources. Answer using only what is supported by the provided passages. If the passages are missing, insufficient, or do not support a definite answer, respond with exactly: *Insufficient Information.* Reason step-by-step, then give a short factual final answer (or that exact phrase).

User. `PASSAGES:`
`=== Retrieved from Conversation Memory ===`
`--- Memory Passage 1..`k `---`
[memory passages 1. . .k*]*
`=== Retrieved from Document Corpus ===`
`--- Document Passage 1..`k `---`
[document passages 1. . .k*]*
`QUESTION:`
[user question]

Figure 3: Generation prompt. Retrieved chunks and questions are omitted. Section headers mark the two sources and empty sections are omitted at inference time. Retrieved chunks are collapsed for brevity.

Passages are grouped into two labeled sections, one for conversation memory and one for the document corpus. Separating the sources in the generator LLM's context is motivated by the difficulty presented in merging two heterogeneous sources into a single ranking, discussed in §2.4. While a merged ranking would allow for broader observations into effects on reasoning, out of caution for interpretability of results and grounding in retrieval, this is left for future work.

Structured Output. Agents return responses in a validated schema with two fields: `reasoning` (chain-of-thought explanation) and `final_answer` (the predicted answer string). Only normalized `final_answer` is used for answer correctness; `reasoning` is logged for qualitative analysis.

Chunking Strategy. Memory sessions are chunked at the *pair* granularity: each chunk contains one user message and the following assistant response. This follows recent findings that simple conversation-pair units are performant for long-term memory retrieval [36] (though shortcomings of this approach have been found in light of a more sophisticated segment-level approach [22] which falls beyond the scope of this study). Documents are chunked into 256-token segments with 32-token overlap using tiktoken's[9] `cl100k_base` tokenizer. Each document chunk is prefixed with metadata (title, source, publication time) to preserve context.

4.1.2 Unified Store. The Unified Store is the simplest baseline: memory and documents share a single embedding index rather than being partitioned by source. All chunks, both conversation pairs and document segments, are indexed in one ChromaDB collection, and retrieval draws the top-ranked chunks from this combined pool. Passages are still labeled by origin in the assembled context (§4.1.1), but the retrieval step itself is source-agnostic: memory and documents compete directly for the same retrieval budget. This architecture anchors the source-organization axis; comparing it against the separate-store agents shows whether partitioning memory and documents into dedicated indexes helps or hurts when evidence must be located across sources.

[9]https://github.com/openai/tiktoken

4.1.3 Separate Store. The Separate Store Agent is the standard multi-source RAG baseline. It maintains two independent ChromaDB collections, one for conversational memory and one for documents, and in a single pass retrieves `memory_k=10` chunks from the memory index and `document_k=10` chunks from the document index before generating one answer. Its only structural difference from the Unified Store is that the two sources occupy separate indexes and are retrieved independently, so a fixed retrieval budget is guaranteed to each source rather than shared competitively. This makes the Agent a reference point for the source-organization axis and the base single-pass flow that MemPalace Agent and Corrective Agent extend.

4.1.4 MemPalace. The MemPalace agent reuses the single-pass flow of Separate Store but replaces the general-purpose memory vector store with MemPalace,[10] a retrieval system purpose-built for long-term conversational memory. MemPalace organizes memory hierarchically: a coarse "closet" layer holds session-level topic pointers, while fine-grained "drawers" store conversation-pair chunks. Its memory retrieval departs from the shared dense backbone, combining dense similarity over the shared MiniLM embeddings with BM25 re-ranking, closet-based boosting of topically related sessions, and drawer-level enrichment; document retrieval remains the standard dense ChromaDB path. Because our MemPalace Agent is otherwise identical to the Separate Store Agent, the comparison between them isolates the memory-retrieval-mechanism axis, testing whether a specialized memory retriever outperforms a general vector store when evidence resides in conversational history.

4.1.5 Corrective Agent. Corrective Agent varies the retrieval control-flow axis: in place of a single pass, it runs an agentic loop that grades its own retrieval and retries when it falls short. Like the Separate Store Agent, it keeps separate ChromaDB collections for memory and documents. Each round retrieves from both stores and passes the combined passages to an LLM grader that returns a single relevance judgment for the question. If the evidence is judged relevant, the agent generates its answer; otherwise it rewrites the question to better express the underlying information need and retrieves again. The loop allows up to two rewrite cycles, after which the agent generates from whatever the final round retrieved. This architecture tests whether self-assessment and query reformulation can recover from an unproductive initial retrieval, which we expect to matter most in the integrated variant, where evidence is split across sources and a single query may under-serve one of them.

4.2 Experimental Design

The main effect under investigation is evidence distribution: the location of a question's golden evidence across knowledge sources. As detailed in §3.7, each question has it's evidence realized in three evidence distribution variants (document-only, memory-only, and integrated) where the same evidence is relocated between the sources. Agents always have access to both the memory store and document corpus. The design crosses this three-level distribution factor with agent architecture and evaluation persona, as described below.

[10] https://www.mempalace.net

4.2.1 D-Optimal Blocked Design. The experiment uses a D-optimal partial factorial design that maximizes statistical power for estimating the main effect (evidence distribution) and and the two-factor interactions it has with persona and agent architecture, while reducing the total number of experimental runs. This approach is employed to allow for the principled observation of multiple factors while accounting for between-question differences [18].

The design has two layers. **Layer 1** selects a D-optimal subset of the 12 possible (distribution × agent) combinations for each question, choosing 4 combinations that maximize information for estimating the distribution main effect, agent main effect, and their interaction (enabling *RQ1a*). **Layer 2** fully crosses each selected combination with all three evaluation personas (Tech, Business-Technology, Sports), enabling analysis of how persona-specific memory composition moderates the primary effects (enabling *RQ1b*).

Each of the 198 answerable questions serves as a blocking factor (random intercept in the statistical model), controlling for question-specific difficulty while estimating within-question effects of distribution and agent. Layer 2 induces the three-level topic-match covariate defined in §3.3: for a given evaluation persona, a question is in-domain, partial, or out-of-domain according to how its evidence-domain union overlaps that persona's profile. Because both stores are always available, the same (q, p) label applies under all three distributions, so in-domain memory can compete as a distractor even when golden evidence lies entirely in documents. This is the moderator through which we measure how topical density in memory affects answer correctness (*RQ1b*).

The design yields 198 questions × 4 selected (distribution, agent) cells × 3 personas = 2,376 experimental runs, with 594 runs per agent. A third of the cost of the full factorial on this sample. Twelve null queries were additionally administered under the same protocol but since they contribute no distribution contrast, are reported in Appendix B.

4.2.2 Evaluation Metrics and Protocol. We adopt Exact Match (EM) as the primary answer-correctness metric and Recall as the retrieval diagnostic. EM provides unambiguous, reproducible scoring across architectures and is consistent with the source dataset's design [30]. For Recall, our goal is not to characterize retrieval quality in isolation but to establish whether evidence retrieval, rather than in-context reasoning alone, mediates differences in answer correctness; §4.3 uses this comparison to answer *RQ1c*.

We compute Recall over the combined pool of all retrieved evidence items passed to the generator LLM. This measures what fraction of the required evidence the agent actually had access to when generating its answer, directly connecting retrieval outcomes to the answer-correctness results.

4.3 Statistical Analysis

The blocked design yields repeated observations on the same questions, and questions differ substantially in difficulty. Aggregate accuracy would treat those runs as independent and would confound experimental effects with which items happened to fall in each cell.

This confounding is observed during exploratory analysis. Figure 4 shows clear variability introduced by the question type, with inference queries from the question pool being meaningfully easier

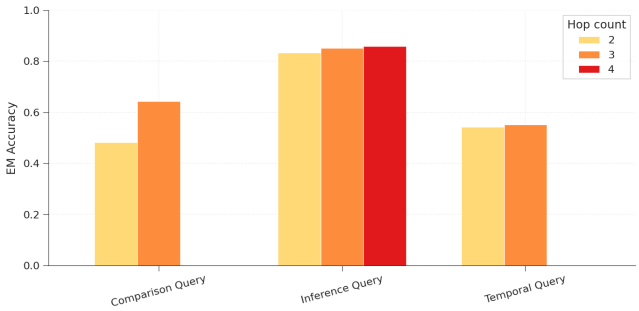


Figure 4: Raw EM accuracy by question type and hop count for answerable questions.

to answer and not all combinations of hops and question type present in the evaluation set. Figure 5 then shows that the impact of all main effects of interest when accounting for question-level variance is very modest. Together, these observations motivate our statistical modeling.

We therefore model exact-match correctness with a logistic mixed-effects model, using a random intercept for questions to absorb item difficulty and allow for within-question contrasts as previous works have done [21]. Models are fitted in R with `lme4`[11]. Factors are sum-coded, so coefficients are read relative to the grand mean.

The sample comprises 2,376 runs from the 198 answerable questions. The specification is pre-declared from the design and the main model on this full sample is:

$$\begin{aligned}\text{correct} \sim\ & \text{persona} + \text{dist} \times \text{agent} + \text{dist} \times \text{topic\_match} \\ & + \text{question\_type} + \text{hop\_count} + (1 \mid \text{question}).\end{aligned} \tag{2}$$

Here, $\times$ denotes interaction + main effects of interaction terms. The interaction with evidence distribution (`dist`) and agent addresses *RQ1a*. The interaction with topic match (§3.3) addresses *RQ1b*: whether the distribution effect depends on how well the question's domains match the evaluation persona. Persona remains as a main effect for leftover differences among the three memory corpora that topic match does not capture. Question type and hop count are included to account help model differences among questions.

Terms are tested for significance with Type III Wald χ^2 tests.[12] Significant terms are then interpreted as estimated marginal means on the probability scale, with Holm-adjusted pairwise contrasts within each planned family. The full Type III tables, and the nested likelihood-ratio test for recall, are reported in Appendix D.

Two planned follow-ups are reported with the results. For *RQ1c*, recall is added to the primary specification on the full sample. A nested likelihood-ratio test asks whether that addition improves fit. We then observe the Type III table for the recall-augmented model (full reported in Appendix D) contrasting with Model 2 only for explorative inisight.

Diversity attributes (§3.5.1) and the evidence-session audit (§3.8) are properties of individual memory sessions, not of experimental runs, so they are not entered as run-level covariates on answer correctness. We instead restrict to the memory-only variant, where every gold item is a session, and model whether that session was retrieved under a strict evidence-hit rule. The audit model is

$$\begin{aligned}\text{retrieval\_hit} \sim\ & \text{evidence\_placement} + \text{topic\_drift} + \\ & \text{evidence\_density} + \text{actual\_turns} + \\ & \text{agent} + \text{persona} + \text{evidence\_in\_domain} \\ & + (1 \mid \text{question}) + (1 \mid \text{evidence}).\end{aligned} \tag{3}$$

`evidence_in_domain` is a binary variable at the session level. It is 1 if that gold item's category lies in the evaluation persona's topic set T_p, 0 otherwise. Random intercepts for question and evidence absorb shared query difficulty and the variability as a result of the underlying fact being generated once per persona. Integrated and document-only runs are omitted because part or all of the gold evidence is not a generated session. This is designed to investigate whether the generation procedure made some sessions systematically harder to retrieve and does not re-estimate the primary distribution effects.

Incomplete sessions (flagged as missing key evidence) are initially included in Model 3. The same model is then refit after dropping the incomplete sessions, to check whether they influence the diversity-attribute results.

5 Results and Discussion

This section reports the analysis of the mixed effects modelling discussed in (§4.3) and in doing so answers **RQ1**. We start off by summarizing the general experimental statistics in §5.1. Afterwards, we showcase and discuss the audit of the generation process concluding the evaluation of our answer to **RQ2**. Then in §5.3, §5.4 and §5.5, we restate the main sub-questions and use our results to answer them. §5.6 then examines selected runs qualitatively. Following this, §5.7 synthesizes the preceding sections to answer **RQ1** and highlight the primary findings. These findings are then constrained to methodological shortcomings in §5.8 before suggestions for future work are listed in §5.9.

5.1 Sample Overview

From the 2,376 runs, 43 had inference failures (34 of which occurred with the corrective agent) during generation and failed to produce any answer. These are treated as wrong answers, potentially biasing the interpretation of the corrective agent's performance. Omitting them raises that agent's raw Exact-Match from 0.57 to 0.60 and the other three agents move by at most 0.007. The evidence-distribution shifts are modestly in the same direction at 0.013 (document-only), 0.008 (integrated), and 0.012 (memory-only). Failures therefore lower the Corrective agent's overall accuracy but likely do not significantly affect the distribution pattern tested in the proceeding chapters.

Across the 198 questions, Exact-Match is invariant in every experimental cell for 51 questions: 35 are always correct (ceiling) and 16 are always wrong (floor). The remaining 147 questions change outcome in at least one cell and are therefore the items that can inform the distribution contrasts.

Question type has the largest effect size in the primary model, $\chi^2(2) = 25.68$, $p < .001$ (Table 3). With estimated marginal means

[11] `lme4` is used with the `bobyqa` optimizer: https://cran.r-project.org/package=lme4
[12] Type III analysis of deviance implemented with `car::Anova`: https://cran.r-project.org/package=car.

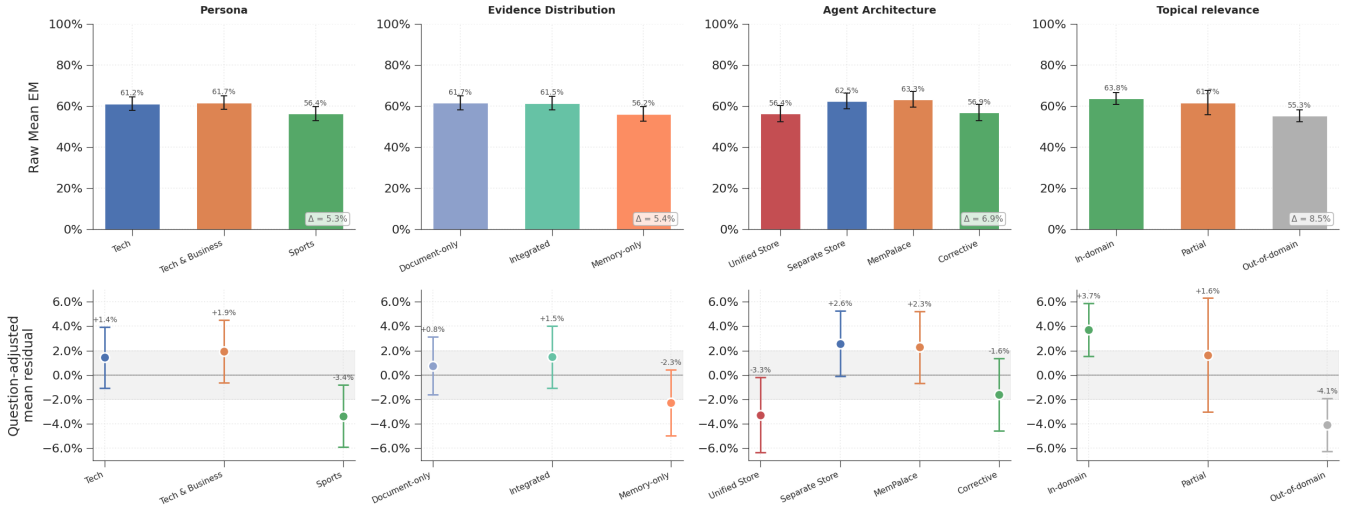


Figure 5: Raw mean EM (top) versus question-centered mean residuals (bottom) for persona, evidence distribution, agent architecture, and topic match. Residuals are computed at the run level by subtracting each question's mean EM across all of its runs are reported with 95% confidence intervals. The gray band marks ±2 percentage points.

$P(\text{correct})$ is found to be 0.91–0.94 for inference questions, 0.53–0.62 for comparison questions, and 0.49–0.59 for temporal questions. The specification does not include a distribution × question-type interaction, so this term soaks up item difficulty together with the question random intercept; it does not test whether placement effects differ by question type.

A sensitivity test dropping the 43 failed generations from the model leaves the ANOVA pattern unchanged.

5.2 Generation Audit

Among the generation attributes, only topic drift is significant as a Type III term, $\chi^2(2) = 6.43$, $p = .040$ (Table 6). Tight-drift sessions are retrieved more often ($\hat{P} = 0.192$) than moderate-drift sessions (0.130). Placement is marginal, $\chi^2(2) = 5.47$, $p = .065$ while density ($\chi^2(1) = 0.16$, $p = .691$) and session length ($\chi^2(1) = 1.50$, $p = .221$) are not.

Per-evidence in-domain status remains by far the largest term ($\chi^2(1) = 81.64$, $p < .001$), ahead of persona and agent (Table 6). Dropping the sessions with missing information (§3.8) yields the same conclusion: topic drift remains significant, placement remains marginal and per-evidence in-domain status remains the largest term.

The audit therefore shows that synthetic packaging of gold memory is not fully retrieval-neutral, but the residual effect is comparatively modest next to topic-match.

5.3 RQ1a: Architecture

RQ1a asks to what extent the effect of evidence distribution on answer correctness varies across agent architectures. The distribution × agent interaction is not significant, $\chi^2(6) = 4.50$, $p = .610$ (Table 3). Architecture therefore does not significantly moderate the distribution effect. Agent is itself a significant predictor of answer correctness, $\chi^2(3) = 13.17$, $p = .004$, so some systems are more accurate overall. Trends associated with particular architectural choices are examined in an exploratory fashion in Appendix E.

5.4 RQ1b: Topic Match

RQ1b asks to what extent persona-specific memory composition moderates the effect of evidence distribution on answer correctness. The distribution × topic-match interaction is significant, $\chi^2(4) = 16.72$, $p = .002$ (Table 3). Topic match therefore significantly moderates the distribution effect: where the golden evidence resides matters more or less depending on how well the question's domains match the evaluation persona's memory profile.

Figure 6 shows the corresponding estimated marginal means. Document-only correctness is essentially flat across topic-match levels. The memory-only estimates, by contrast, fall as topical fit worsens. Integrated sits between the two, remaining close to memory-only when the question is in-domain and closer to document-only when it is out-of-domain. Holm-adjusted pairwise contrasts within each facet isolate the drop: none of the in-domain or partial contrasts reach significance, whereas out-of-domain memory-only is reliably worse than both document-only (odds ratio 2.33, $p < .001$) and integrated (odds ratio 1.84, $p = .007$).

This pattern is counter-intuitive relative to the literature that motivated the design. Because conversational memory concentrates around recurring interests (§2.2), we expected in-domain questions to face denser, more competitive distractors and therefore to suffer when evidence was placed in memory [9, 10, 22]. Instead, memory-only correctness is highest when the question is in-domain and collapses only once the question falls outside the persona's topics. One reading, pursued qualitatively in §5.6, is that category-level topic match is a poor proxy for the embedding neighbourhood the retriever actually sees: out-of-domain runs still surface semantically similar chunks, while the in-domain cell places gold inside

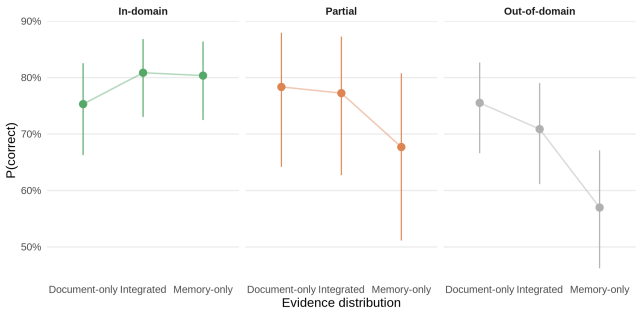


Figure 6: Estimated P(correct) by evidence distribution, faceted by topic match. Intervals are 95% asymptotic confidence intervals on the response scale.

a haystack that is 75% same-domain by construction (§3.6). Alternatively, if future work reaffirms a genuine mismatch penalty, it would be concerning: errors would concentrate on topics outside the user's profile, where the user is least equipped to notice them and most in need of reliable answers when learning something new. Despite being misaligned with the research that inspired the design it might sit in line with findings that model reasoning can create filter bubbles, biasing outputs toward affiliations stated in context [13].

5.5 RQ1c: Retrieval

RQ1c asks to what extent differences in answer correctness across evidence distributions are accounted for by retrieval success. Nested comparison of the primary mixed-effects model against the same specification with recall added improves fit(Table 5). Retrieval is therefore a strong predictor of whether a run is answered correctly.

Once recall is in the model it dominates the Type III table ($\chi^2(1) = 86.92$, $p < .001$), while the distribution × topic-match interaction falls from $\chi^2(4) = 16.72$ in Model 2 to 5.68, $p = .224$ (Table 4). That drop indicates that the topic-match moderation of placement may be rooted in retrieval. The next subsection pursues this reading qualitatively.

5.6 Qualitative Analysis

To inspect how retrieval produces the distribution × topic-match pattern, we investigated a sample of nine (q, p, agent) units in which Exact Match disagreed across the distributions assigned to that agent (three units per distribution with 18 runs in total). From the inspected units we observe the following patterns:

- The single run that retrieved every gold item was answered correctly.
- Out-of-domain memory-only runs typically miss the labelled gold sessions (recall 0) and abstain with *Insufficient Information*, while the document-only twin of the same unit is often already correct on a single gold article (recall 0.33–0.50).
- Those misses are not empty retrieval. The returned memory chunks are often semantically close to the question required related entities, events, or neighbouring coverage.
- Integrated placement can match document-only recall while missing the memory share of a comparison, or miss both shares and substitute a distractor entity.
- Residual generation errors under incomplete recall include polarity errors on comparisons.

These observations sit in line with the quantitative result in §5.5 that the distribution × topic-match effect is rooted in retrieval. They also are indicative of the topic-match label not describing the context the generator LLM sees. Despite the golden context being out-of-domain, the context is filled with sessions of the golden context's domain.

Furthermore, the observed similarity in retrieved memory chunks from the same domain hints that the in-domain sessions might be more semantically dense than expected. This would result in a construct that fails to capture real world memory profiles.

Additionally, the analysis hints at a larger contamination problem despite the sampling frame (§3.2): questions screened as requiring retrieval can still be answered from non-gold passages that recycle similar but not exactly the same information. The correct news article metadata is often missing from those passages, but for short Exact-Match answers that payload can be enough.

5.7 Takeaways: Evidence Distribution (RQ1)

RQ1 asked how the distribution of evidence across a document corpus and conversational memory affects an agent's answer correctness. The answer is that it does, but not as a uniform penalty for placing evidence in memory. Averaged over topic match, the three distributions are initially found significant ($\chi^2(1) = 6.36$, $p = .042$) but not reliably distinguishable with (Holm-adjusted estimated marginal means. What is more reliable is the interaction with topic match (§5.4): memory-only correctness is superior when the question matches the persona's topics and falls once it does not, while document-only remains flat. Evidence distribution therefore matters through its dependence on persona-specific memory composition, not as a main effect that appears in every cell.

The remaining sub-questions bound that claim. The pattern does not vary detectably across the four architectures (§5.3), so the result is not an artifact of a single retrieval design. Retrieval success dominates the recall-augmented model and the distribution × topic-match interaction shrinks once recall is included (§5.5), which is why we read the pattern qualitatively in §5.6. Consequently, we cannot claim that "memory is harder than documents". Instead we observed that relocating the same evidence into conversational memory is costly primarily when that evidence sits outside the user's topical profile.

5.8 Limitations

The experiment evaluated one language-model checkpoint as the generator. Different models may use retrieved context differently, and the findings may not carry over to other languages, time periods, or knowledge domains.

MultiHop-RAG being the single source dataset is potentially the largest weakness of the study. Since questions are known to be a large source of variance, using a dataset with only synthetically generated questions targeted at the document retrieval setting has potential to limit our findings. Biasing toward the document-only

distribution and resulting in less likely user sessions due to the nature of the evidence being embedded.

The controlled design also relies on synthetic conversational histories. This made it possible to move the same evidence while holding the question fixed, but generated sessions cannot reproduce every ambiguity, dependency, or change in preference found in natural user histories. Synthetic memories haystacks, especially when generated to embed evidence details native to documents likely fail to capture the complexity and distribution of real memories.

The evaluation metrics also might limit findings. Recall reflects the evidence available to each agent, but ignores rank and whether the generator attends to a passage (faithfulness). Exact match is convenient and reproducible, but limits answers to concise strings that are rarely seen in agent outputs today. The qualitative sample also shows that questions screened as requiring retrieval can still be answered from non-gold, semantically close passages; short Exact-Match strings can succeed without the labelled article metadata, so gold recall is not a strict account of what the model used. Category topic-match and the 75% in-domain haystack likewise may overstate how well the construct matches real user profiles.

5.9 Future Work

Replication with other language models, source datasets and agent architectures remains the direct next step. Adapting the evaluation to natural longitudinal conversation histories would better test whether the findings generalise to production settings.

Future work should also probe the topic-match construct more tightly, measuring embedding neighbourhood, LLM-as-a-judge semantic similarity and distractor density rather than category labels alone. Harder correctness criteria (citations, article identity, or faithfulness) would reduce the chance that near-duplicate contamination counts as success. Separating retrieval from generation effects on answer correctness remains a priority.

6 Conclusion

As QA agents increasingly combine private conversational memory with external documents, a basic design question becomes unavoidable: does it matter where the evidence for a query is stored? This study takes a first controlled step toward answering that question. By moving the same evidence across memory and documents while keeping the questions fixed, we separate source placement from question difficulty, evidence content, and source availability.

The results show that memory is not simply another document store. Evidence placement was associated with answer correctness, but not through a simple document-vs-memory ordering. Instead, the effect is moderated by topic match, with correctness collapsing when queries require retrieving off-topic evidence from conversational history. These results suggest that evidence distribution and memory composition jointly dictate agent performance. Additionally, they reveal a concerning vulnerability: personalized systems are systematically less reliable when users attempt to explore domains outside their dominant conversational profiles. This dynamic could act as a soft algorithmic filter bubble, penalizing user curiosity with lower accuracy and warranting deeper study to see if these patterns persist in production environments.

7 Ethical Reflection

This study does not evaluate real user conversations. The memory corpora are synthetic sessions generated from a public QA dataset, and the personas are experimental constructs rather than profiles of actual people. This limits the immediate risk to human participants, but it does not remove the ethical importance of the setting. The work is motivated by systems that retrieve from private conversational memory, where the relevant evidence may include personal preferences, relationships, routines, or sensitive facts that users did not expect to be reused in later tasks [8].

The main ethical issue is therefore not the experimental data itself, but the design implications of personalized memory retrieval. If answer quality depends on how evidence is distributed across memory and documents, then memory is not a neutral storage backend. Decisions about what to remember, how to chunk it, how long to retain it, and when to retrieve it can shape both system performance and user exposure. Deployed systems should make these choices visible to users, support deletion and correction of stored memories, and avoid treating retrieved memory as automatically relevant or consented-to merely because it is available.

The study also has representational limits. The three personas and topic domains are simplified, and the generated conversations cannot capture the full social, linguistic, or cultural variation of real long-term memory. The results should therefore be read as evidence about controlled retrieval conditions, not as a claim that one memory design is generally safer or fairer across user groups.

8 Use of Generative AI

I used generative AI as a tool to speed up work that I designed, directed, and verified. The research questions, experimental design, interpretation of results, and the arguments in this thesis are my own. I checked all generated code, text, citations, and analysis outputs before including them.

8.1 Generative AI in the Experimental Pipeline

GPT-5.4 was used to generate synthetic conversational sessions and GPT-OSS-120B to check fact coverage in those sessions (§3.5). Llama 3.3 70B was the language model for all QA agents and for Corrective RAG relevance grading (§4.1). These uses are part of the studied system and are described in the dataset and methodology sections.

8.2 Generative AI in Project Development

Cursor was used for programming assistance: autocomplete, looking up R syntax and packages, and occasional help with scaffolding, debugging, and drafting experiment or analysis code. I specified the design, reviewed every change, and retained the implementation as my own work. Project plans were written by me first; AI was used only to brainstorm alternatives, improving clarity of my plans and to retrieve notes I had already recorded.

For the thesis text, AI was used to aid research, to suggest structure, and to edit grammar, conciseness, tone, and LaTeX formatting. AI-produced text improved drafts I had written to communicate my intended meaning more clearly. It did not introduce arguments or

results I had not already formulated. When it attempted to do so because of surrounding context or hallucinations, I removed that material. Editing was done in Cursor and in ArkOffice (arkoffice.com), a platform I help build and test. Only my own notes and non-sensitive project material were provided as context. The underlying models included Composer 2, Gemma 4, Claude Sonnet 5, and, toward the end, Grok 4.6.

References

[1] Klejda Alushi, Jan Strich, Chris Biemann, and Martin Semmann. 2026. Comprehensive Comparison of RAG Methods Across Multi-Domain Conversational QA. arXiv:2602.09552 [cs.CL] https://arxiv.org/abs/2602.09552

[2] Anthropic. 2024. Introducing Contextual Retrieval. (2024). https://www.anthropic.com/news/contextual-retrieval Blog post.

[3] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *The Twelfth International Conference on Learning Representations (ICLR)*. arXiv:2310.11511 [cs.CL] https://arxiv.org/abs/2310.11511

[4] Mahdi Astaraki, Mohammad Arshi Saloot, Ali Shiraee Kasmaee, Hamidreza Mahyar, and Soheila Samiee. 2026. When Iterative RAG Beats Ideal Evidence: A Diagnostic Study in Scientific Multi-hop Question Answering. arXiv:2601.19827 [cs.CL] https://arxiv.org/abs/2601.19827

[5] Steven Au, Cameron J. Dimacali, Ojasmitha Pedirappagari, Namyong Park, Franck Dernoncourt, Yu Wang, Nikos Kanakaris, Hanieh Deilamsalehy, Ryan A. Rossi, and Nesreen K. Ahmed. 2025. Personalized Graph-Based Retrieval for Large Language Models. *arXiv preprint arXiv:2501.02157* (2025).

[6] Lingjiao Chen, Matei Zaharia, and James Zou. 2024. How Is ChatGPT's Behavior Changing Over Time? *Harvard Data Science Review* 6, 2 (2024). doi:10.1162/99608f92.5317da47

[7] Zhe Chen, Yusheng Liao, Zhiyuan Zhu, Haolin Li, Hongcheng Liu, Yanfeng Wang, and Yu Wang. 2026. HeteroRAG: A Heterogeneous Retrieval-Augmented Generation Framework for Medical Vision Language Tasks. arXiv:2508.12778 [cs.CL] https://arxiv.org/abs/2508.12778

[8] Beatriz Costa-Gomes, Pavel Tolmachev, E. Taysom, V. Sounderajah, Hannah Richardson, Philipp Schoenegger, Xiaoxuan Liu, M. M. Nour, Seth Spielman, Samuel F Way, Yash Shah, M. Bhaskar, Harsha Nori, Christopher J. Kelly, P. Hames, Bay Gross, Mustafa Suleyman, and Dominic King. 2026. Public use of a generalist LLM chatbot for health queries. *Nature Health* (April 2026). https://www.microsoft.com/en-us/research/publication/public-use-of-a-generalist-llm-chatbot-for-health-queries/

[9] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The Power of Noise: Redefining Retrieval for RAG Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)*. ACM, 719–729. doi:10.1145/3626772.3657834

[10] Muhan Gao, Zih-Ching Chen, and Kuan-Hao Huang. 2026. The First Drop of Ink: Nonlinear Impact of Misleading Information in Long-Context Reasoning. arXiv:2605.10828 [cs.CL] https://arxiv.org/abs/2605.10828

[11] Aaron Grattafiori et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[12] Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2025. Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation. arXiv:2409.12941 [cs.CL] https://arxiv.org/abs/2409.12941

[13] Tomo Lazovich. 2023. Filter bubbles and affective polarization in user-personalized large language model outputs. arXiv:2311.14677 [cs.CY] https://arxiv.org/abs/2311.14677

[14] 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. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL] https://arxiv.org/abs/2005.11401

[15] Xiaopeng Li, Pengyue Jia, Derong Xu, Yi Wen, Yingyi Zhang, Wenlin Zhang, Wanyu Wang, Yichao Wang, Zhaocheng Du, Xiangyang Li, Yong Liu, Huifeng Guo, Ruiming Tang, and Xiangyu Zhao. 2025. A Survey of Personalization: From RAG to Agent. *arXiv preprint* (2025). Manuscript submitted to ACM.

[16] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172 [cs.CL] https://arxiv.org/abs/2307.03172

[17] Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating Very Long-Term Conversational Memory of LLM Agents. arXiv:2402.17753 [cs.CL] https://arxiv.org/abs/2402.17753

[18] Douglas C. Montgomery. 2019. *Design and Analysis of Experiments* (10 ed.). John Wiley & Sons.

[19] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2014–2037. doi:10.18653/v1/2023.eacl-main.148

[20] Irina Iuliana Nagâț, Vlad Constantin Pârcă, Mihai-Vlad Florea, Elena Cocianu, Andrei-Marius Avram, and Traian Rebedea. 2026. A Systematic Investigation of Document Chunking Strategies and Embedding Sensitivity. arXiv:2603.06976 [cs.IR] https://arxiv.org/abs/2603.06976

[21] National Institute of Standards and Technology. 2025. *Expanding the AI Evaluation Toolbox with Statistical Models*. Technical Report NIST.AI.800-3. National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-3.pdf

[22] Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. 2025. On Memory Construction and Retrieval for Personalized Conversational Agents. arXiv:2502.05589 [cs.CL] https://arxiv.org/abs/2502.05589 ICLR 2025.

[23] Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. 2025. MemoRAG: Boosting Long Context Processing with Global Memory-Enhanced Retrieval Augmentation. arXiv:2409.05591 [cs.CL] https://arxiv.org/abs/2409.05591

[24] Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2024. Zep: A Temporal Knowledge Graph Architecture for Agent Memory. (2024). https://www.getzep.com

[25] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 3982–3992. doi:10.18653/v1/D19-1410

[26] Alireza Salemi and Hamed Zamani. 2025. Learning to Rank for Multiple Retrieval-Augmented Models through Iterative Utility Maximization. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR '25)*. ACM. doi:10.1145/3731120.3744584

[27] Tobias Schreieder and Michael Färber. 2025. Claim2Source at CheckThat! 2025: Zero-Shot Style Transfer for Scientific Claim-Source Retrieval. In *CLEF 2025 Working Notes*, Vol. 4038. CEUR-WS.org, Madrid, Spain, 1203–1216. https://ceur-ws.org/Vol-4038/paper_94.pdf

[28] Nikhil Sharma, Q. Vera Liao, and Ziang Xiao. 2024. Generative Echo Chamber? Effects of LLM-Powered Search Systems on Diverse Information Seeking. arXiv:2402.05880 [cs.CL] https://arxiv.org/abs/2402.05880

[29] Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long T. Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 2025. In Prospect and Retrospect: Reflective Memory Management for Long-term Personalized Dialogue Agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 8416–8439. https://aclanthology.org/2025.acl-long.413/

[30] Yixuan Tang and Yi Yang. 2024. MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries. arXiv:2401.15391 [cs.CL] https://arxiv.org/abs/2401.15391

[31] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*. https://arxiv.org/abs/2104.08663

[32] Chen Wang, Xunzhuo Liu, Yue Zhu, Alaa Youssef, Priya Nagpurkar, and Huamin Chen. 2025. Category-Aware Semantic Caching for Heterogeneous LLM Workloads. arXiv:2510.26835 [cs.DB] https://arxiv.org/abs/2510.26835

[33] Hongru Wang, Minda Hu, Yang Deng, Rui Wang, Fei Mi, Weichao Wang, Yasheng Wang, Wai-Chung Kwan, Irwin King, and Kam-Fai Wong. 2023. Large Language Models as Source Planner for Personalized Knowledge-grounded Dialogue. arXiv:2310.08840 [cs.CL] https://arxiv.org/abs/2310.08840

[34] Hongru Wang, Wenyu Huang, Yang Deng, Rui Wang, Zezhong Wang, Yufei Wang, Fei Mi, Jeff Z. Pan, and Kam-Fai Wong. 2024. UniMS-RAG: A Unified Multi-source Retrieval-Augmented Generation for Personalized Dialogue Systems. arXiv:2401.13256 [cs.CL] https://arxiv.org/abs/2401.13256

[35] Shuai Wang, Ekaterina Khramtsova, Shengyao Zhuang, and Guido Zuccon. 2024. FeB4RAG: Evaluating Federated Search in the Context of Retrieval Augmented Generation. arXiv:2402.11891 [cs.IR] https://arxiv.org/abs/2402.11891

[36] Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. arXiv:2410.10813 [cs.CL] https://arxiv.org/abs/2410.10813

[37] Yikuan Xia, Jiazun Chen, Yirui Zhan, Suifeng Zhao, Weipeng Jiang, Chaorui Zhang, Wei Han, Bo Bai, and Jun Gao. 2025. ER-RAG: Enhance RAG with ER-Based Unified Modeling of Heterogeneous Data Sources. arXiv:2504.06271 [cs.IR] https://arxiv.org/abs/2504.06271

Table 2: Example of each MultiHop-RAG question type with its required evidence and gold answer.

Type	Question	Required evidence	Gold answer
Comparison	Does the Sporting News article on the Marshall Thundering Herd indicate a change in starting quarterback similar to the quarterback replacement reported for the Seattle Seahawks by Sporting News?	(1) *Sporting News*; Ian Valentino; "UTSA vs. Marshall odds, props, predictions: Roadrunners lay big spread in Frisco Bowl"; 2023-12-18. Fact: "Instead, Marshall will start freshman Cole Pennington, who is the son of former Marshall star QB Chad Pennington." (2) *Sporting News*; Gilbert McGregor; "Seahawks vs. Giants live score, updates, highlights from NFL 'Monday Night Football' game"; 2023-10-02. Fact: "Geno Smith's night appears to be done as Drew Lock is back in for the Seahawks."	Yes
Temporal	Did TechCrunch report on DeepMind's expected Gemini release before reporting that Google released only a "lite" Gemini Pro version?	(1) *TechCrunch*; Kyle Wiggers; "One year later, ChatGPT is still alive and kicking"; 2023-11-30. Fact: "DeepMind, Google's premier AI research lab, is expected to debut a next-gen chatbot, Gemini, before the end of the year." (2) *TechCrunch*; Kyle Wiggers; "Google fakes an AI demo, Grand Theft Auto VI goes viral and Spotify cuts jobs"; 2023-12-09. Fact: "But it didn't release the full model, Gemini Ultra—only a 'lite' version called Gemini Pro."	Yes
Inference	Which quarterback, less effective under pressure (CBSSports.com), also holds a 5–0 starting record for 2023 (Sportskeeda)?	(1) *CBSSports.com*; Dave Richard; "NFL Fantasy Football Week 6 Lineup Decisions: Starts, Sits, Sleepers, Busts to know for every game"; 2023-10-12. Fact: "However, Purdy's been at his worst when pressured (like most quarterbacks), completing 50% of his throws for 6.7 yards per attempt with a gaudy 15.9% off-target rate." (2) *Sportskeeda*; Nick Igbokwe; "Why is Brock Purdy called Mr. Irrelevant? Story behind 49ers QB's nickname explained"; 2023-10-10. Fact: "So far, Purdy has started the 2023 NFL season where he left off, opening the season with a 5–0 record."	Brock Purdy
Null	Considering Times of India and Hindustan Times articles on Jaya Bachchan, which film character is portrayed by an actor who also served in the Rajya Sabha?	None	Insufficient information.

[38] Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective Retrieval Augmented Generation. arXiv:2401.15884 [cs.CL] https://arxiv.org/abs/2401.15884

[39] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. arXiv:1809.09600 [cs.CL] https://arxiv.org/abs/1809.09600

[40] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *The Eleventh International Conference on Learning Representations (ICLR)*. arXiv:2210.03629 [cs.CL] https://arxiv.org/abs/2210.03629

[41] Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. *Evaluation of Retrieval-Augmented Generation: A Survey*. Springer Nature Singapore, 102–120. doi:10.1007/978-981-96-1024-2_8

[42] Sangwon Yu, Jimin Lee, Jaehyung Kim, and Sungzoon Cho. 2025. Unleashing Multi-Hop Reasoning Potential in Large Language Models through Repetition of Misordered Context. In *Findings of the Association for Computational Linguistics: NAACL 2025*. https://arxiv.org/abs/2410.07103

[43] Saber Zerhoudi and Michael Granitzer. 2024. PersonaRAG: Enhancing Retrieval-Augmented Generation Systems with User-Centric Agents. *arXiv preprint arXiv:2407.09394* (2024).

A MultiHop-RAG Examples

Table 2 presents an example of each question type from MultiHop-RAG Dataset. These examples were sampled to be short and have fewer evidence items for conciseness.

From the table it can be felt that the questions are synthetically generated. They often signal deep knowledge of the source news articles required for the question, while the answers seem to be more trivial.

B Null Query Results

Though Figure 7 appears to portray trends in agent architecture performance. Particularly the corrective agent being the least performant. These results should be interpreted with nuance as the experiment was not designed around evaluating null queries and so the comparison is done purely for exploratory purposes.

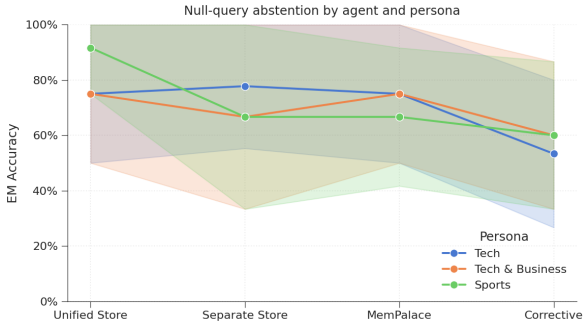


Figure 7: Null-query exact-match rate by agent and persona

C Generation Prompts

Figure 8 is the non-temporal system prompt used at generation time, with filled fields shown in italics. Temporal items additionally insert a published-at line in `EVIDENCE TO EMBED` and mention publication time in the evidence-origin and preservation rules. Off-topic evidence appends a `TOPIC CONTEXT` block after `SCENARIO`.

An additional Off topic block is added in case the evidence does not align with the domain(s) of the persona. This block is seen in Figure 9.

D Full Statistical Results

Tables 3 to 6 report the Type III analysis-of-deviance tests described in §4.3. Each term is tested given all other terms in the model.

System.
You are generating a realistic past interaction between a user and their AI assistant. The assistant is a general-purpose tool for tasks and questions—work or everyday—not a companion chatbot. Keep the exchange grounded in concrete outcomes.

`USER PROFILE:`
[persona character]

The user's communication style varies naturally between sessions. They may be concise and task-driven in one conversation and more exploratory or conversational in another. The assistant should match the user's energy—brief when the user is brief, more detailed when the user is asking open-ended questions. Avoid a rigid or formulaic tone; the exchange should read like a real person talking to their AI assistant.

`EVIDENCE AND KNOWLEDGE:`
The block below is news-article reporting. The assistant has no access to it or any knowledge of the article except what the user says in this conversation. Only the user may introduce the title, source, and key facts listed—the assistant must not state or assume them first. General domain knowledge is fine; this specific reporting is not. After the user has supplied those details, the assistant may answer, analyze, or help, but must not be the originator of that factual content.
Incorporate the following evidence as something the user discussed with the assistant in the past.

`EVIDENCE TO EMBED:`
- `Topic:` *[title]*
- `Source:` *[source]*
- `Key information:` *[fact]*

`SCENARIO:`
The scenario below describes the user's situation and motivation. It should shape how the user sounds and what they need from the assistant—not be quoted or announced explicitly unless a real person would naturally say it that way.
[sampled scenario]

`CONVERSATION NATURALISM:`
[rendered from sampled topic-drift: tight / moderate / wide]

`STRUCTURAL GUIDELINES:`
- Produce a conversation of *[min]* to *[max]* turns (*[min/2]* to *[max/2]* user messages, with an equal number of assistant responses).
- *[sampled evidence-placement instruction]*
- *[sampled evidence-density instruction]*
- The user initiates the conversation. The user must introduce the evidence (title, source, and key factual details) across their messages.
- The full factual content of the "Key information", the article title, and the source name must appear in the USER's messages. Do not omit, paraphrase away, or dilute critical details (names, numbers, dates, comparisons).

User. *[sampled trigger phrasing]*

Figure 8: Session-generation prompt (non-temporal). Italic brackets are filled per request from the persona, gold evidence, and sampled diversity attributes. All other wording is shared across sessions.

`TOPIC CONTEXT:`
This topic is outside the user's primary domain. They are engaging with it as a general-interest reader—through curiosity, social context, or incidental exposure. If the scenario is work-related, the connection to this topic should be indirect (e.g., a colleague mentioned it, it came up in passing, or they encountered it while looking for something else). The user should not sound like a domain expert on this topic.

Figure 9: Off-topic topic-context block. Appended to the system prompt after `SCENARIO` when the evidence domains do not overlap the target persona.

Table 3: Type III Wald tests for the primary mixed-effects model of Exact-Match correctness.

Term	χ^2	df	p
persona	1.72	2	.422
dist	6.36	2	.042
agent	13.17	3	.004
topic match	19.07	2	$< .001$
question type	25.68	2	$< .001$
hop count	3.44	1	.064
dist × agent	4.50	6	.610
dist × topic match	16.72	4	.002

Table 4: Type III Wald tests after adding recall (*RQ1c*).

Term	χ^2	df	p
persona	0.88	2	.644
dist	5.28	2	.072
agent	9.95	3	.019
topic match	8.70	2	.013
question type	22.75	2	$< .001$
hop count	13.04	1	$< .001$
recall	86.92	1	$< .001$
dist × agent	5.19	6	.520
dist × topic match	5.68	4	.224

Table 5: Nested likelihood-ratio test of the primary model versus the same specification with recall (*RQ1c*).

Model	npar	AIC	BIC	logLik	χ^2	df	p
Primary	24	2441.6	2580.1	−1196.8	—	—	—
Primary + recall	25	2354.0	2498.3	−1152.0	89.54	1	$< .001$

Table 4 is a residual check after adding recall, not a re-test of the experimental factors. Table 5 reports the nested likelihood-ratio test of whether adding recall improves fit.

Table 7: Estimated P(correct) by agent from the primary model. Values are emmeans on the response scale, averaged over the remaining factors; intervals are 95% asymptotic confidence intervals.

Agent	$\hat{P}$	95% CI
Separate Store	0.780	[0.701, 0.843]
MemPalace	0.780	[0.700, 0.843]
Corrective	0.713	[0.622, 0.790]
Unified Store	0.687	[0.593, 0.767]

Table 6: Type III Wald tests for the memory-only generation audit (evidence hit).

Term	χ^2	df	p
evidence placement	5.47	2	.065
topic drift	6.43	2	.040
evidence density	0.16	1	.691
session length	1.50	1	.221
agent	14.98	3	.002
persona	37.09	2	< .001
evidence in-domain	81.64	1	< .001

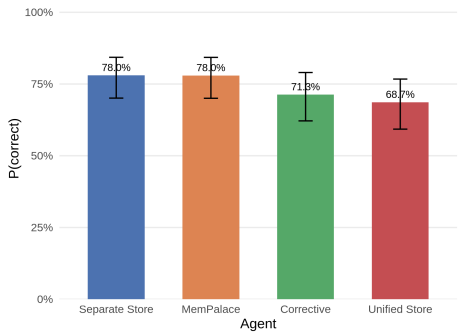


Figure 10: Estimated P(correct) by agent.

E A Discussion of the Agent's Architectural choices

Agent is a significant predictor of Exact-Match correctness, $\chi^2(3) = 13.17$, $p = .004$ (Table 3), while the planned distribution × agent interaction is not ($\chi^2(6) = 4.50$, $p = .610$). Architectures therefore differ in overall accuracy, not in how evidence placement affects accuracy. The estimates below are marginal means from the primary model, averaged over distribution, persona, topic match and question type (Figure 10; Table 7).

Separate Store and MemPalace are tied at $\hat{P} = 0.78$. Corrective is lower (0.71) and Unified Store lowest (0.69). Holm-adjusted pairwise contrasts isolate only the two comparisons against Unified Store: both Separate Store and MemPalace outperform it (odds ratio 1.62, $p = .021$ in each case). Corrective is not reliably different from either the leading pair or Unified Store after Holm adjustment.

The Unified Store deficit is expected on architectural grounds: memory and document chunks compete for a shared retrieval budget. That the same system is the only one that differs reliably from the two-index designs, while Corrective does not, also sits with the sample overview (§5.1): failed generations concentrate on Corrective and are scored as incorrect, which pulls its marginal mean down without producing a significant pairwise contrast.

Because the interaction is non-significant, cell-wise slopes in Figure 11 are exploratory. They remain non-significant after recall is added (Table 4). The curves are comparatively flat across agents on integrated; the largest visual drop-offs on memory-only are for Corrective and Unified Store. Those rankings are readings of the figure, not pairwise tests.

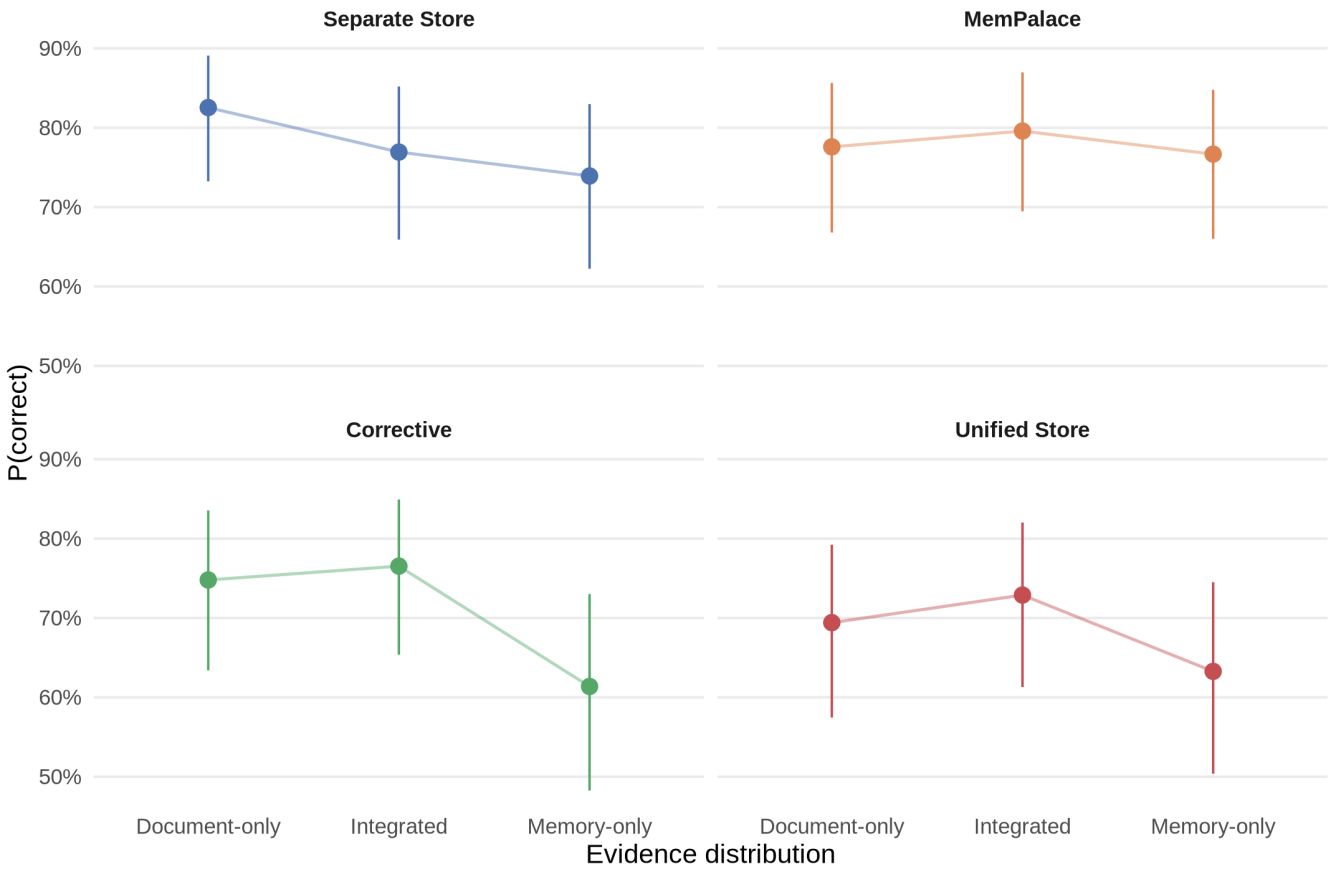


Figure 11: Estimated $P(\text{correct})$ by evidence distribution, faceted by agent. Intervals are 95% asymptotic confidence intervals on the response scale, averaged over topic match and question type. The distribution × agent interaction is not significant (Table 3).