electronics

MDPI

Article

AURORA-RAG: A Trustworthy Hierarchical Retrieval-Augmented Generation Architecture for Grounded Synthetic Personas

Rebecca Marigliano * and Kathleen M. Carley

Software and Societal Systems Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA; kathleen.carley@cs.cmu.edu
* Correspondence: rmarigli@andrew.cmu.edu

Abstract

Large language models can generate synthetic personas, but their factual claims may be unsupported when generation relies on parametric knowledge alone. We present AURORA-RAG, a hierarchical retrieval-augmented generation architecture that introduces documentary evidence at country–topic, community, community–topic, and persona construction stages. The evaluated four-country corpus comprises 834 input files and 3888 indexed semantic chunks, with non-English source material translated into English. AURORA-RAG combines dense retrieval and BM25 using reciprocal rank fusion and pseudo-relevance feedback while retaining document provenance. We compare no-RAG, persona-only hybrid RAG, and full AURORA-RAG using 1440 personas per condition across three geopolitical topics. Corpus support increases from 68% to 82% and 91%, respectively, while unsupported-claim rates decline from 24% to 13% and 7%. Between retrieval-enabled conditions, supplied-evidence faithfulness rises from 79% to 90%, attribution accuracy from 82% to 92%, and citation correctness from 85% to 94%. Full AURORA-RAG also produces greater opinion displacement from no-RAG than persona-only retrieval (0.210 versus 0.103), a measure of output change rather than accuracy. Hybrid retrieval leads on five of six ranking metrics, although only MRR@5 over dense retrieval remains significant after Holm correction. These findings support hierarchical retrieval for stronger documentary grounding and source traceability.

Keywords: hierarchical retrieval-augmented generation; synthetic personas; hybrid retrieval; documentary grounding; trustworthy AI

Check for updates

Academic Editor: Cecilio Angulo

Received: 15 September 2026
Revised: 22 October 2026
Accepted: 23 October 2026
Published: 27 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Large language models (LLMs) can generate synthetic actors and social contexts at scale, but persona generation that relies primarily on parametric model knowledge can produce generic, stale, or unsupported outputs. This is particularly problematic in hierarchical synthetic populations, where errors introduced at country or community levels can propagate to many downstream personas.

Retrieval-augmented generation (RAG) supplements knowledge stored in model parameters with external documentary evidence [1,2]. Our earlier AURORA framework (AI-Utilized Retrieval for Optimized Representation of Audiences) uses semantic retrieval and LLM-based generation to construct synthetic communities and personas within a national information environment [3]. It combines this contextual construction with salience-aware opinion assignment, in which topic importance influences how individual stances relate to

country and community context. The earlier study examined opinion-assignment strategies and differences among generation models.

AURORA-RAG extends this framework by specifying and evaluating the documentary retrieval pipeline within that population hierarchy. The present implementation combines lexical and semantic retrieval, preserves source provenance, and introduces evidence at the country–topic, community, community–topic, and persona stages. The present study retains the underlying population-generation scaffold while examining retrieval placement, evidence selection, claim support, and the retention of contextual information in downstream outputs.

This study asks whether hierarchical retrieval improves documentary grounding beyond persona-only retrieval, whether combining lexical and semantic retrieval improves evidence selection, and whether generated outputs retain distinguishable country-, community-, and persona-level information. We compare no-RAG, persona-only single-stage hybrid RAG, and full hierarchical AURORA-RAG, alongside a separate comparison of BM25, dense retrieval, and hybrid retrieval. The evaluation distinguishes corpus support, supplied-evidence faithfulness, unsupported claims, attribution accuracy, and citation correctness from retrieval relevance and information recoverability. Controlled retrieval and corpus interventions additionally examine sensitivity to changes in the documentary environment.

The results support the hierarchical design. Full AURORA-RAG achieves 91% corpus support, compared with 82% under persona-only retrieval and 68% under no-RAG, while unsupported-claim rates are 7%, 13%, and 24%, respectively. Compared with persona-only retrieval, supplied-evidence faithfulness increases from 79% to 90%, attribution accuracy from 82% to 92%, and citation correctness from 85% to 94%. Mean absolute displacement of community–topic mean opinion scores from no-RAG is 0.103 under persona-only retrieval and 0.210 under full AURORA-RAG. Hybrid retrieval with reciprocal rank fusion (RRF) achieves the highest observed value on five of six retrieval metrics, although only its MRR@5 advantage over dense retrieval remains significant after Holm correction. Community and persona information also remains recoverable from generated explanations.

AURORA-RAG contributes a hierarchical, provenance-preserving method for constructing synthetic personas from an explicit documentary environment. It separates retrieval relevance from downstream grounding and examines how contextual information remains recoverable across the population hierarchy. In this study, trustworthiness is evaluated through documentary support, supplied-evidence faithfulness, attribution, citation correctness, and source traceability. Interaction quality, cross-topic consistency, uncertainty, and adaptation remain directions for subsequent behavioral evaluation. Section 2 reviews related research, Section 3 describes the architecture, Section 4 presents the implementation, corpus, and experimental methods, and Section 5 reports the results. The remaining sections discuss the findings, summarize the contributions, and identify limitations and future research directions.

2. Related Work

This section relates AURORA-RAG to research on retrieval-augmented generation, RAG evaluation, synthetic personas, and economic-agent simulation. The comparisons distinguish established retrieval mechanisms from the present contribution: organizing documentary evidence within hierarchical persona construction and evaluating its grounding and influence on generated outputs.

2.1. Retrieval-Augmented Generation and Hybrid Retrieval

Zhao et al. review how retrieval augments generation [2], a principle AURORA-RAG applies to hierarchical persona construction. REALM learns retrieval during pretraining [4], whereas AURORA-RAG uses pretrained components during generation. Like the original RAG framework [1], it combines parametric generation with external memory but organizes evidence across population-construction stages. RETRO integrates retrieved chunks into its model architecture [5], and Fusion-in-Decoder fuses independently encoded passages during decoding [6]; AURORA-RAG supplies passages through prompts. Atlas emphasizes few-shot learning with an updateable index [7], while the present study evaluates evidence placement within a fixed generation scaffold.

AURORA-RAG shares the use of vector similarity for semantic retrieval with Dense Passage Retrieval (DPR) [8] but uses a pretrained embedding model rather than the original DPR model. Its chunk-level cosine representation also differs from ColBERT's token-level late interaction [9]. BM25 provides lexical matching for entities and specialized terminology [10], while RRF combines dense and lexical rankings without calibrating their raw scores [11]. Both are established mechanisms reused within the architecture.

Rocchio's query-vector feedback [12] and the Lavrenko–Croft relevance model [13] provide precedents for query refinement. AURORA-RAG uses a simpler pseudo-relevance feedback heuristic that appends five frequent filtered terms from initial results. HyDE constructs dense retrieval representations from hypothetical documents [14]; the optional HyDE-inspired path similarly uses generated text as a search aid. The original Query2doc method expands queries with pseudo-documents [15], whereas AURORA-RAG's optional adaptation reranks candidate chunks by cosine similarity to a pseudo-document embedding. Generated search aids are excluded from documentary evidence, and these optional paths are not enabled in the primary architecture comparison.

2.2. RAG Evaluation, Grounding, and Trustworthiness

RAGAS distinguishes context relevance, faithfulness, and answer relevance [16], paralleling AURORA-RAG's separation of retrieval quality from generation quality. ARES develops trained evaluators [17], while RAGChecker provides claim-level retrieval and generation diagnostics [18]. These studies motivate our evaluation dimensions, which we assess using the task-specific protocols described in Section 4.6. KILT connects task performance with provenance [19]; AURORA-RAG similarly evaluates attribution and citation correctness, but within its own documentary corpus and persona-generation setting. Corpus support, faithfulness to supplied evidence, and information recoverability remain distinct measurements.

Self-RAG learns retrieval decisions and uses reflection tokens to assess evidence and outputs [20]; AURORA-RAG instead specifies retrieval stages through its population architecture. Yoran et al. address irrelevant context through entailment-based filtering and robust training [21]. AURORA-RAG shares their concern with evidence quality but examines retrieval placement and downstream grounding rather than implementing these model interventions.

Wu et al. review retrieval noise, conflicting evidence, and security risks [22], while PoisonedRAG demonstrates how malicious corpus additions can manipulate generation [23]. AURORA-RAG supports source inspection through provenance records and separates generated retrieval aids from documentary evidence. These controls improve traceability but do not guarantee source truth or resistance to poisoning.

2.3. Synthetic Personas and Information Environments

Argyle et al. assess demographic response fidelity through profile-conditioned language models [24]. AURORA-RAG similarly conditions synthetic actors on specified at-

tributes but evaluates documentary grounding rather than correspondence with human subgroup responses. Aher et al.'s replications of human-subject experiments [25] likewise target behavioral validity, which differs from information recoverability. Dillion et al.'s discussion of substituting language models for human participants reinforces this distinction [26]. Park et al. retrieve experiential memories to support reflection, planning, and interaction [27]; AURORA-RAG also retrieves context but emphasizes external documents during population construction.

Studies of partisan selection [28,29] motivate AURORA-RAG's controlled exposure rules. Research on platform and browsing effects [30,31] examines information-selection mechanisms that the present system abstracts into corpus retrieval. Echo-chamber patterns [32] and reader demand for media slant [33] further motivate differentiated information environments, although AURORA-RAG does not simulate network or media-market dynamics. Actor attributes remain specified through the persona skeleton; direct identity conditioning of retrieval is a separate diagnostic rather than the production procedure. These experiments therefore test specified documentary-exposure mechanisms, not estimates of human media selection or persuasion.

2.4. Economic-Agent Simulation, Heterogeneity, and Persona Adaptation

Filippas et al. study simulated economic agents with experimentally specified information, preferences, and endowments [34]. AURORA-RAG shares this separation of actor characteristics from informational inputs while making documentary retrieval and source traceability explicit. EconAgent combines demographic and occupational profiles with perception and memory to generate repeated work and consumption decisions [35]. Both systems support heterogeneous actors, but EconAgent evaluates economic decisions and macroeconomic dynamics, whereas AURORA-RAG evaluates documentary conditioning and support for generated statements.

AgentSociety combines profiles, evolving states, and memory in large-scale social simulations [36]. It shares AURORA-RAG's interest in heterogeneous populations but emphasizes interactions and evolving behavior rather than evidence selection during construction. Qi et al. distinguish stable identity from changing psychological states and use corrective feedback to maintain persona coherence [37]. AURORA-RAG likewise separates specified attributes from contextual inputs but does not implement longitudinal psychological-state updates. Sensitivity to a changed documentary corpus therefore differs from persistent persona adaptation.

Persona variation can also arise from changes to the population-generation procedure. Paglieri et al. evolve persona-generator code using LLM-based mutation operators to broaden the range of traits, opinions, and preferences represented in generated populations [38]. This differs from changing an existing agent's demographic specification or updating its psychological state during interaction. In AURORA-RAG, demographic skeletons remain fixed across retrieval conditions so that documentary exposure can be varied without simultaneously changing population composition. Varying demographic characteristics would constitute an additional population-level intervention, while longitudinal persona adaptation would require an explicit state-update mechanism. Neither is evaluated by the present corpus-replacement experiments.

AURORA-RAG integrates established retrieval mechanisms at the country–topic, community, community–topic, and persona stages while preserving source metadata and demographic specifications. Its comparison of no retrieval, persona-only retrieval, and full hierarchical retrieval holds the generation scaffold fixed to examine evidence placement. The contribution concerns this organization of documentary evidence and its evaluation through

grounding and contextual differentiation; population representativeness and longitudinal behavioral fidelity remain separate validation targets.

3. AURORA-RAG Architecture

3.1. Design Principles

AURORA-RAG is designed around three principles. First, contextual generation should be conditioned on documentary evidence rather than relying exclusively on the LLM's parametric memory. Second, retrieval should follow the same hierarchy as the synthetic population so that country–topic, community, community–topic, and persona-level context can be represented at the appropriate level of specificity. Third, model-generated query expansions may guide retrieval but are not treated as documentary evidence; retrieved corpus passages remain distinguishable from the generated contextual states propagated between stages. These constraints reduce the amount of contextual information that must be reconstructed solely from model parameters while preserving an auditable separation between documentary evidence and model-generated text.

The framework does not replace the underlying population model. Demographic and structural attributes are established independently and supplied to the generator as fixed state. Retrieval instead controls the documentary information environment available during generation. This separation is important because it distinguishes the demographic structure of the synthetic population from the external information to which the population is exposed. It also permits retrieval behavior and downstream generation behavior to be evaluated separately.

A demographic skeleton is the predefined individual specification supplied to persona construction, including attributes such as age, gender, and political orientation. A community is a modeled institutional or social group within a country, while a community category specifies the type of group to construct. Country–topic and community–topic states provide contextual information used in subsequent opinion assignment. Salience represents a topic's assigned importance and determines how persona-level opinion is sampled from these upstream states. A generated persona combines its fixed specification with community context and documentary evidence; its topic-specific explanation is subsequently generated for an already-assigned stance.

Figure 1 shows the implemented hierarchy:

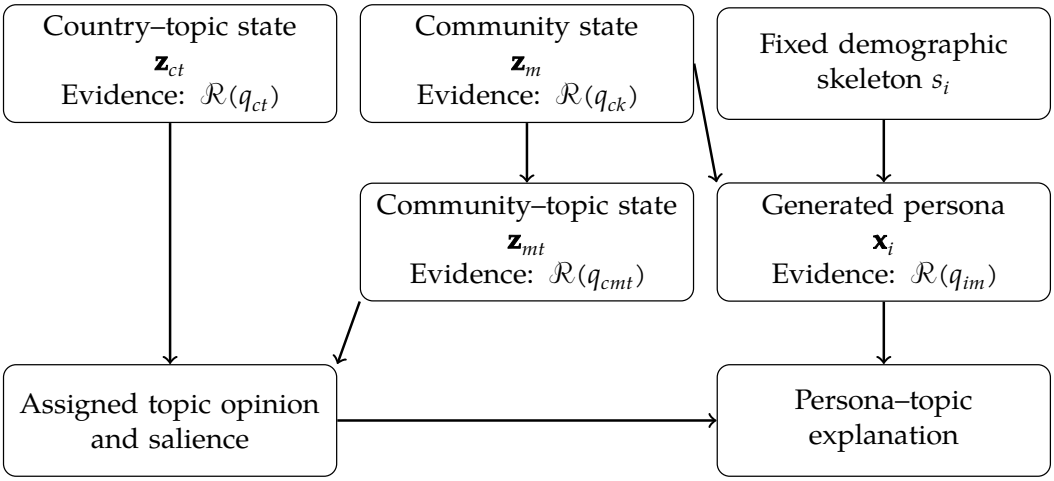

Figure 1. Hierarchical generation in AURORA-RAG. Each $\mathcal{R}(q)$ denotes retrieved corpus evidence; arrows show state propagation. Country–topic and community–topic states inform stance assignment, while the constructed persona and assigned stance condition the explanation.

3.2. System Architecture

AURORA-RAG consists of two coupled pipelines: offline indexing and online hierarchical retrieval. During indexing, scenario documents are loaded, assigned SHA-256 content hashes, deduplicated, semantically segmented, and transformed into dense embeddings. The embeddings are stored in a persistent ChromaDB vector collection, while a BM25 index is constructed over the same set of documentary chunks. Maintaining aligned dense and sparse representations ensures that semantic and lexical retrieval operate over the same evidence base.

During online retrieval, each generation level constructs a deterministic query from its current structured state and submits that query to both the dense and sparse retrievers. Dense and BM25 rankings are fused through Reciprocal Rank Fusion (RRF), and Pseudo-Relevance Feedback (PRF) can expand the query before a second retrieval pass. The highest-ranked documentary chunks are formatted as evidence and supplied to schema-constrained LLM calls. Figure 2 summarizes this workflow.

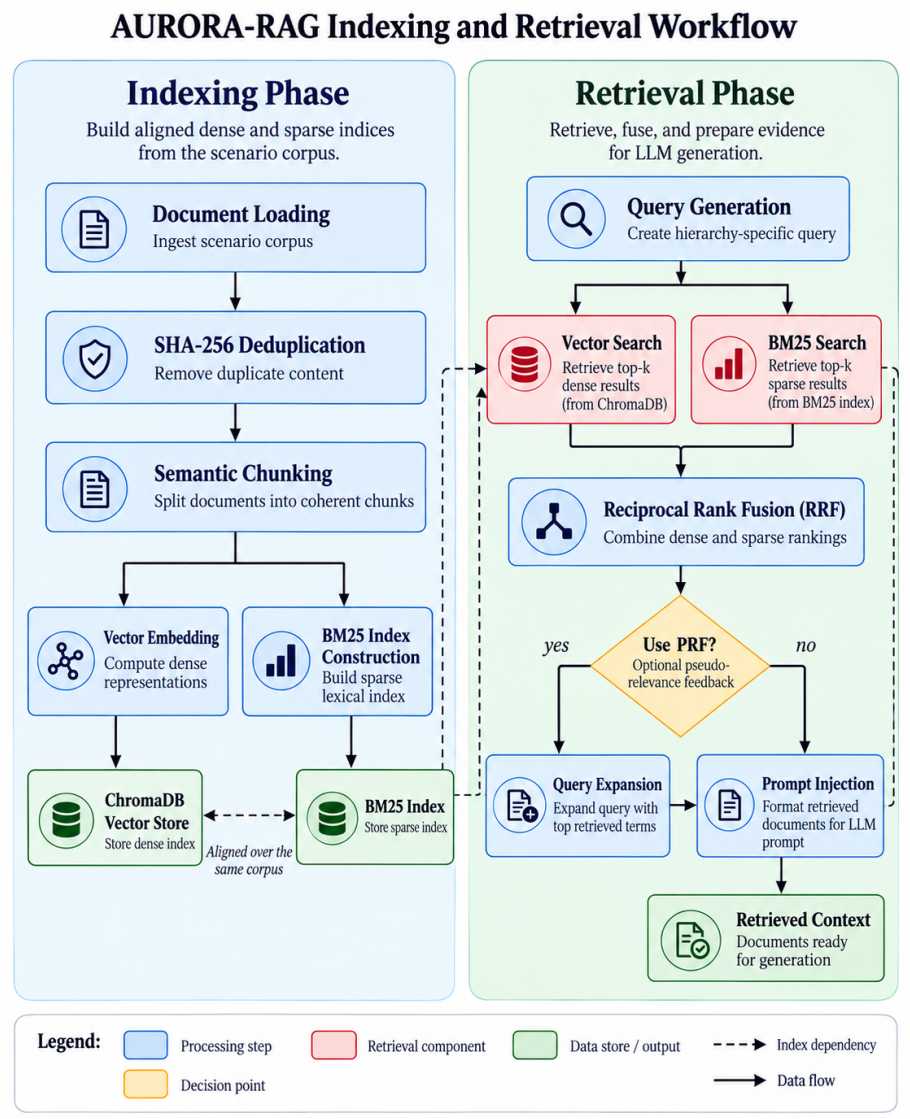

Figure 2. AURORA-RAG indexing and retrieval workflow. The offline phase constructs aligned dense-vector and BM25 representations from a shared semantically segmented corpus. During online generation, hierarchy-specific queries are evaluated by both retrievers, fused through Reciprocal Rank Fusion, optionally expanded with Pseudo-Relevance Feedback, and converted into corpus-grounded context for the LLM.

Full AURORA-RAG introduces evidence at four construction stages within the three-level country–community–persona hierarchy: country–topic, community, community–topic, and persona construction. Structured states propagate downstream, while each retrieval-enabled stage obtains additional corpus evidence appropriate to its context. This combines inherited contextual information with stage-specific retrieval rather than relying on a single static evidence set.

3.3. Hierarchical Evidence Retrieval and Propagation

AURORA-RAG retrieves evidence at multiple levels of the synthetic-population hierarchy while propagating structured state between related generation stages. Let $\mathcal{R}(q)$ denote the formatted top-k corpus evidence returned for query q.

For country c and topic t, the country–topic state is

$$z_{ct} = G_\theta^{(c)}(c, t, \mathcal{R}(q_{ct})). \tag{1}$$

For institutional category k within country c, the community state is

$$z_m = G_\theta^{(m)}(c, k, \mathcal{R}(q_{ck})). \tag{2}$$

For community m and topic t, the community–topic state combines the propagated community state with newly retrieved topic-specific evidence,

$$\mathbf{z}_{mt} = G_\theta^{(mt)}(\mathbf{z}_m, c, t, \mathcal{R}(q_{cmt})). \tag{3}$$

Persona construction is described separately below. The resulting architecture is hierarchical rather than strictly linear: community state $\mathbf{z}_m$ is propagated into later generation stages, while country–topic state $\mathbf{z}_{ct}$ and community–topic state $\mathbf{z}_{mt}$ contribute to persona–topic stance and reasoning. Retrieval is performed independently at the country–topic, community, community–topic, and persona stages so that evidence enters at the level where it is most relevant.

The production retriever does not directly include age, gender, or political orientation in these queries. These attributes enter through the structured persona specification; the identity-conditioned retrieval evaluated later is therefore a controlled experimental extension rather than the default retrieval procedure.

3.4. Retrieval-Conditioned Persona Construction

Persona construction establishes the individual-level state used in subsequent topic-specific generation. Each persona receives a fixed demographic skeleton, an assigned community state, country context, and newly retrieved community-specific documentary evidence. Persona construction is represented as

$$\mathbf{x}_i = G_\theta^{(p)}\big(s_i, \mathbf{z}_{m(i)}, c, \mathcal{R}(q_{im})\big), \tag{4}$$

where s_i contains predefined demographic attributes, $\mathbf{z}_{m(i)}$ is the structured state of the persona's assigned community, and $\mathcal{R}(q_{im})$ provides documentary evidence associated with that community and country context.

Demographic attributes are maintained independently of retrieval so that documentary conditioning modifies the persona's information environment rather than its underlying demographic specification. Schema-constrained generation produces a structured persona representation $\mathbf{x}_i$ that is subsequently used during topic-specific reasoning.

https://doi.org/10.3390/electronics15214962

3.5. Persona–Topic Stance and Reasoning

After persona construction, AURORA-RAG establishes a topic-specific stance before generating the corresponding natural-language explanation. For persona i and topic t, salience is sampled around the associated community–topic salience. High-salience cases draw opinion around the community–topic stance, whereas lower-salience cases draw from the country–topic opinion using the corresponding normal, skewed, or bimodal country-level distribution. The resulting scalar opinion is bounded to $[-1, 1]$ and converted to qualitative opinion and salience labels [3].

The persona–topic LLM call then receives the constructed persona $\mathbf{x}_i$, topic t, assigned opinion label, and assigned salience label. Its role is to generate the natural-language explanation of that already-established stance rather than to determine the stance itself. Scalar opinion and explanatory language are therefore distinct outputs of the pipeline and are evaluated separately.

3.6. End-to-End RAG Generation Procedure

Algorithm 1 summarizes the architecture from corpus indexing through persona and persona–topic generation. The algorithm makes the community–topic layer explicit because it is an independent structured stage in the implemented AURORA hierarchy.

Algorithm 1 AURORA-RAG Hierarchical Generation

Require: Scenario corpus $\mathcal{D}$, countries, topics, community categories, and persona specifications
Ensure: Retrieval-conditioned country–topic, community–topic, and persona states with provenance

```
Assign source hashes, deduplicate, and semantically segment corpus 𝒟
Construct aligned dense-vector and BM25 indices
for each country c do
    for each topic t do
        Retrieve country–topic evidence ℛ(q_ct)
        Generate structured country–topic state z_ct
    end for
    for each required community category do
        Retrieve category–country evidence
        Generate structured communities m
    end for
    for each community m do
        for each topic t do
            Retrieve community–topic evidence ℛ(q_cmt)
            Generate z_mt conditioned on z_m
        end for
    end for
    for each persona specification s_i do
        Assign persona i to community m(i)
        Retrieve community-specific persona evidence ℛ(q_im)
        Generate x_i using fixed demographic constraints
        for each topic t do
            Derive persona–topic stance from z_ct and z_mt
            Generate persona–topic reasoning conditioned on the structured state
        end for
    end for
end for
return structured outputs and associated retrieval provenance
```

3.7. Provenance and Trustworthiness Controls

AURORA-RAG retains source metadata and parent-document hashes throughout chunking and retrieval so that retrieved evidence can be traced to its originating document. SHA-256 hashing prevents exact duplicate documents from being repeatedly indexed, reducing accidental amplification of duplicated evidence. Structured scenario values known *a priori* are loaded directly rather than inferred through probabilistic retrieval when an authoritative structured value is available.

These controls improve auditability but do not guarantee factual correctness. Contradictory sources, semantic duplicates, query drift, model misinterpretation, and corpus poisoning remain possible. AURORA-RAG reduces reliance on unsupported parametric generation and makes retrieved evidence inspectable, but these controls alone do not guarantee claim-level faithfulness or eliminate hallucination; the following experiments therefore evaluate documentary claim support separately from retrieval correctness and output differentiation.

4. Methods and Experimental Design

4.1. Corpus and Preprocessing

The four-country documentary corpus was assembled from newspaper reporting originating in the United States, United Kingdom, India, and Russia. For each country, news articles were collected in relevant local or official languages based on predefined topics of interest for the synthetic-persona scenarios. Non-English articles were subsequently translated into English to establish a common language for document preprocessing, semantic segmentation, and retrieval across the four national information environments. The collected material was organized by country and topic and stored as plain-text files or JSON collections, with individual JSON files potentially containing multiple articles. An available news-search manifest records a search window from 1 May 2024 to 1 August 2025; these dates describe search parameters rather than verified publication dates for every indexed document.

Exact duplicate documents are identified using SHA-256 hashes before segmentation. LangChain's `SemanticChunker` places boundaries at changes in semantic similarity, producing variable-length passages without a fixed token window or overlap. Each chunk retains its parent-document hash and source metadata for traceability.

Table 1 reports the complete documentary collections for the United States, United Kingdom, India, and Russia, together with the subset covering Cyber Security Cooperation, Russian Aggression, and Taiwanese Sovereignty.

Table 1. Corpus composition by country. Files are input storage units; chunks are indexed passages. Focal-topic counts are included in the full-corpus totals.

Country	Full Files	Full Chunks	Focal-Topic Files	Focal-Topic Chunks
United States	259	1637	58	147
United Kingdom	278	1185	58	139
India	158	521	58	170
Russia	139	545	58	204
Total	834	3888	232	660

The complete documentary corpus comprises 834 documents organized into 834 input files and 3888 indexed semantic chunks across the four countries. Of these, 232 documents (27.8%) and 660 chunks (17.0%) correspond to the three focal topics evaluated in this

study. The remaining 602 documents contribute 3228 chunks covering additional topics and background information. Semantic segmentation produces an average of 4.66 chunks per document across the full corpus and 2.84 chunks per document within the focal-topic subset. The corpus contains approximately 1.9 million tokens, occupying 7.4 MiB of extracted English text.

Each retrieval-enabled generation stage draws documentary evidence from the scenario corpus. The retrieval-component diagnostic uses the 660 focal-topic chunks as queries and excludes candidate passages sharing the query's source-content hash. A retrieved passage is considered relevant when it comes from a different source and matches the query's country and topic. Section 4.7 describes the scoring and aggregation procedures.

4.2. Retrieval Implementation and Model Configuration

Let $C = \{x_1, \ldots, x_J\}$ denote the indexed semantic chunks. Each chunk retains the source metadata and parent-document hash established during indexing.

For dense retrieval, an embedding model f_ϕ maps query q and chunk x_j into a shared vector space. Similarity is measured with cosine similarity,

$$s_{\text{dense}}(q, x_j) = \frac{f_\phi(q)^\top f_\phi(x_j)}{\|f_\phi(q)\|_2 \, \|f_\phi(x_j)\|_2}. \tag{5}$$

The current implementation uses OpenAI's `text-embedding-3-small` model for dense document representation and semantic chunking, although the architecture is embedding-model agnostic and can be configured to use other compatible embedding models. Dense embeddings are stored in ChromaDB. A BM25 retriever is constructed from the same chunk objects stored in the vector collection so that lexical and semantic retrieval operate over an aligned set of passages rather than over separately segmented corpora.

Dense retrieval is paired with BM25 because the two retrieval mechanisms address different failure modes. Dense retrieval can recover paraphrases and semantically related material even when query and document vocabulary differ, whereas BM25 remains sensitive to exact entities, rare terms, and domain-specific language [10]. Because their raw scores are not assumed to be directly calibrated, AURORA-RAG combines the resulting rankings using RRF [11]:

$$s_{\text{RRF}}(x_j \mid q) = \sum_{r \in \{\text{dense}, \text{BM25}\}} \frac{\mathbb{I}[x_j \in R_r(q)]}{\kappa + \rho_r(x_j)}, \qquad \kappa = 60, \tag{6}$$

where $\rho_r(x_j)$ is the rank assigned to chunk x_j by retriever r. The implementation uses $\kappa = 60$. The highest-ranked fused chunks form the evidence set used by downstream generation.

For the production hierarchical-generation path, country–topic, community, community–topic, and persona retrieval return the top three documentary passages from the fused ranking. Although retrieval depth is configurable, the reported generation pipeline therefore uses $k = 3$.

4.2.1. Pseudo-Relevance Feedback and LLM-Assisted Retrieval

For queries affected by vocabulary mismatch, AURORA-RAG applies lightweight PRF. The system first performs hybrid retrieval using the original query and concatenates text from the highest-ranked results. A simple stop-word filter removes common function words, after which candidate terms shorter than four characters are discarded. The five most frequent remaining terms are appended to the original query:

$$q' = q \,\|\, u_{(1)} \,\|\, \cdots \,\|\, u_{(5)}. \tag{7}$$

Dense and sparse retrieval are then repeated using q', followed by a second RRF fusion. In the production retrieval function, this PRF-expanded query is used by default before

the final top-k documentary passages are returned. PRF therefore changes the retrieval representation while leaving the documentary evidence base unchanged.

A higher-cost retrieval path is also implemented for difficult or ambiguous queries. This path combines original hybrid-RRF candidates, PRF-expanded candidates, and HyDE-style candidates generated from an LLM-produced hypothetical passage [14]. For HyDE, the implementation prompts `gpt-3.5-turbo` to produce a short hypothetical answer to the query, embeds that generated passage, and retrieves corpus chunks according to similarity to the hypothetical representation.

The same optional path also implements a Query2Doc-style reranking step. A second generated pseudo-document represents an ideal passage corresponding to the information need. Candidate corpus chunks are independently embedded and reranked using cosine similarity to the pseudo-document representation. Both HyDE and Query2Doc use `gpt-3.5-turbo` with temperature 0.7 in the evaluated implementation; however, neither component is model-specific and alternative compatible LLMs can be substituted.

Generated passages are used only as search representations; only indexed corpus chunks are supplied as documentary evidence to downstream generation.

The hierarchical population-generation path uses the lower-cost PRF-enabled hybrid RRF retrieval function. HyDE and Query2Doc therefore represent optional retrieval extensions in the implemented framework and should not be interpreted as mechanisms required for every reported generation result.

4.2.2. Model Configuration

Table 2 summarizes the evaluated settings. The same generation model produces contextual states, personas, and persona–topic explanations. Architecture comparisons vary retrieval placement while holding generation settings, demographic skeletons, and prompt templates fixed; alternative generation and embedding models are not evaluated.

Table 2. Principal model and retrieval configuration used in the evaluated implementation.

Component	Configuration
Generation model	`gpt-4o-mini`
Generation temperature	1.0
Embedding model	`text-embedding-3-small`
Document segmentation	LangChain `SemanticChunker`
Chunk size	Variable; determined by semantic boundaries
Chunk overlap	No fixed overlap
Dense vector store	ChromaDB
Sparse retrieval	BM25
Rank fusion	Reciprocal Rank Fusion
RRF constant	$\kappa = 60$
Production retrieval depth	Top 3 documentary chunks
PRF expansion	Five frequent informative terms
Exact deduplication	SHA-256 content hashing
Optional HyDE model	`gpt-3.5-turbo`, temperature 0.7
Optional Query2Doc model	`gpt-3.5-turbo`, temperature 0.7
Generation output	Schema-constrained structured output

Dense and BM25 retrieval share the same chunks and metadata, with source hashes linking chunks to parent documents. PRF expands an initial hybrid query with filtered terms from retrieved passages and repeats retrieval; it remains enabled wherever retrieval is active. The production retrieval depth is distinct from the ranking cutoffs used in retrieval evaluation.

Optional HyDE and pseudo-document reranking inspired by Query2doc use the retrieval-aid model listed in the table and are excluded from the architecture comparison. Their generated text guides evidence selection but is not documentary evidence. Schema constraints govern output structure; factual support is evaluated separately.

4.3. Population and Example

All primary analyses use four country conditions—the United States, United Kingdom, India, and Russia—and three geopolitical topics: Cyber Security Cooperation, Russian Aggression, and Taiwanese Sovereignty. Each architecture contains 1440 personas, with 360 personas per country, yielding 4320 persona–topic responses per architecture.

Beyond the controlled experimental configurations, the population-generation implementation has also been used to construct populations of up to 5000 synthetic personas. These larger runs demonstrate execution at an increased population size but are separate from the matched architecture comparisons, claim-level grounding evaluations, and retrieval-component analyses reported in this study. Table 3 shows an actual indexed passage and a saved synthetic-persona output from the scenario. It illustrates the source and output formats, not a verified retrieval-to-output trace or an additional evaluation result.

Table 3. Excerpt and saved persona output for cyber security cooperation. Quoted text, scores, and labels are reproduced as stored.

Element	Content
Source document	U.S. Department of State, Solidifying Cybersecurity Cooperation in the U.S-Vietnam Comprehensive Strategic Partnership; manifest date: 21 November 2024. Stored as `03.txt` in the cyber security cooperation folder.
Indexed passage (excerpt)	"Specifically, the MOU seeks to explore opportunities for enhancing cooperation between the United States and Vietnam regarding information/operational technology security, incident response and reporting, and collaboration on cybersecurity incidents."
Synthetic persona	Maya Reyes; age 25; female; policy analyst; Progressive Local Governments Network; United States.
Saved topic state	Opinion: 0.574 (Moderately Support); salience: 0.561 (Medium).
Generated explanation	"Maya strongly supports cyber security cooperation as it is crucial for protecting communities from cyber threats. She believes that collaboration enhances the resilience of local governments and ensures the safety of vital information."

4.4. Evaluation Overview

The evaluation comprises five complementary analyses. The architecture comparison assesses differences in downstream opinion scores among no-RAG, persona-only single-stage hybrid RAG, and full AURORA-RAG. The claim-level evaluation measures corpus support and unsupported-claim frequency across all three architectures, while supplied-evidence faithfulness, attribution accuracy, and citation correctness are assessed only in the retrieval-enabled conditions. A separate retrieval-component comparison evaluates evidence ranking under BM25, dense retrieval, and hybrid RRF. The hierarchy analysis examines country–topic opinion patterns and the recoverability of community and persona information from generated explanations. Finally, information-environment diagnostics test whether explicit identity attributes are reflected in retrieved evidence and whether replacing the documentary corpus changes generated outputs while persona specifications remain fixed.

4.5. ArchitectureComparison

The architecture ablation holds the hierarchical generation scaffold, demographic skeletons, model configuration, prompt templates, and base seed fixed while changing where documentary retrieval is enabled. Table 4 defines the three conditions.

Table 4. Retrieval stages enabled in each architecture. CT denotes country–topic and MT denotes community–topic.

Condition	CT	Community	MT	Persona
No-RAG	No	No	No	No
Single-stage hybrid RAG	No	No	No	Yes
Full AURORA-RAG	Yes	Yes	Yes	Yes

The single-stage condition uses the same hybrid retriever as full AURORA-RAG but enables documentary retrieval only during persona construction. It therefore serves as a single-stage retrieval-placement baseline while preserving the hierarchical generation scaffold. Matching uses the same persona skeletons and 96 country–community-category slots, while community names and descriptions are generated separately under each condition. The architecture comparison represents one base generation seed. These conditions represent retrieval-placement ablations of the present implementation, not independent reimplementations of AURORA.

Architecture effects are summarized using displacement from the no-RAG reference. Let $\overline{Y}_{amt}$ denote the mean opinion for architecture a, matched community slot m, and topic t:

$$D_{am} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \left| \overline{Y}_{amt} - \overline{Y}_{0mt} \right|, \tag{8}$$

where 0 denotes no-RAG and $\mathcal{T}$ contains the three focal topics. Conditional 95% confidence intervals are estimated with 10,000 bootstrap replicates that resample matched slots within country. Full-versus-single-stage displacement is additionally evaluated with a paired t-test across the 96 matched slots and effect size d_z.

4.6. Claim-Level Faithfulness Evaluation: Sampling, Evidence Support, and Attribution

The claim-level experiment evaluates whether generated factual statements are supported by documentary evidence rather than treating retrieval accuracy as a proxy for grounding. Each architecture contains 1440 personas evaluated across three focal topics, yielding 4320 persona–topic responses per architecture in the full generation output.

For the faithfulness evaluation, a country-stratified 25% subsample selects 360 matched persona skeletons, with 90 drawn from each country. Each sampled persona is evaluated on all three focal topics, yielding 1080 persona–topic responses per architecture. Because the same matched cases are evaluated under no-RAG, single-stage hybrid RAG, and full AURORA-RAG, the experiment contains 3240 architecture-specific responses in total.

Four externally verifiable factual claims are evaluated from each sampled response, producing 4320 claim-level observations per architecture and 12,960 claim-level observations across the three architectures. Table 5 summarizes the relationship between the full generation output and the claim-level evaluation sample.

Table 5. Sample construction for the claim-level faithfulness evaluation. Counts are shown per architecture.

Evaluated Unit	No-RAG	Single-Stage Hybrid RAG	Full AURORA-RAG
Full persona population	1440	1440	1440
Full persona–topic responses	4320	4320	4320
Faithfulness-sample personas	360	360	360
Faithfulness-sample responses	1080	1080	1080
Factual claims evaluated	4320	4320	4320

Table 5 distinguishes the full generated population from the claim-evaluation sample. Each architecture contains 1440 personas and 4320 persona–topic responses. The country-stratified evaluation selects the same 360 persona skeletons across conditions and includes all three topics, yielding 1080 responses per architecture. Four factual claims are assessed per sampled response, giving 4320 claim observations per architecture. Thus, the repeated count of 4320 denotes different units: responses in the full population and claims in the evaluation sample.

Subjective statements, preferences, and explicitly hypothetical statements are excluded from factual-claim scoring. In the retrieval-enabled conditions, each factual claim is classified as supported by supplied evidence, supported elsewhere in the corpus, contradicted, or unsupported. No-RAG claims are classified as corpus-supported, contradicted, or unsupported. These categories distinguish whether a claim is supported somewhere in the documentary corpus from whether that support was actually supplied to the generator.

Table 6 summarizes the five evaluation measures and the conditions to which they apply.

Table 6. Claim-level faithfulness and attribution measures.

Metric	Definition	Conditions
Corpus support	Proportion of factual claims supported either by supplied evidence or elsewhere in the reference corpus.	All three
Evidence faithfulness	Proportion of factual claims supported specifically by documentary evidence supplied during generation.	Single-stage and full
Hallucination frequency	Proportion of factual claims classified as unsupported. Contradicted claims are recorded separately.	All three
Attribution accuracy	Proportion of eligible named-source attributions in which the identified source supports the attributed information.	Single-stage and full
Citation correctness	Proportion of eligible citations whose referenced passage supports the associated factual claim.	Single-stage and full

Corpus support and evidence faithfulness answer different questions. Corpus support asks whether a generated factual claim can be substantiated somewhere in the reference corpus, whereas evidence faithfulness asks whether it is supported by information actually supplied to the model during generation. Hallucination frequency is based on claims classified as unsupported; contradicted claims are recorded separately and therefore are not included in the hallucination category. Attribution accuracy and citation correctness

are evaluated only for the retrieval-enabled conditions and use the eligible attributions and citations identified during annotation as their respective denominators.

Corpus support, hallucination frequency, and evidence faithfulness are analyzed using separate mixed-effects logistic regressions. Architecture is the primary fixed effect, with country and topic included as covariates. Random intercepts for persona and response account for the three topic-specific responses associated with each persona and for multiple factual claims extracted from the same response. Pairwise architecture contrasts are reported as odds ratios with Wald 95% confidence intervals and two-sided significance tests. Attribution accuracy and citation correctness are reported descriptively because their denominators depend on the number of eligible attributions and citations produced in each retrieval-enabled condition.

4.7. Comparative Evaluation of Retrieval Methods

A separate offline experiment compares the retrieval methods used by the architecture. BM25 provides sparse lexical retrieval, dense retrieval uses stored document embeddings, and hybrid RRF combines the two rankings. Table 7 shows the breakdown.

Table 7. Retrieval-component comparison.

Retriever	BM25	Dense	Rank Fusion
BM25 only	Yes	No	No
Dense only	No	Yes	No
Hybrid RRF	Yes	Yes	RRF

Each of 660 focal-topic chunks serves as a query. Chunks sharing the same source-content hash are excluded from the candidate set, preventing retrieval from succeeding by returning another passage from the same source. A result is relevant when it comes from a different source and matches the query's country and topic.

Metrics are calculated at ranks one, three, and five and then averaged within source-content hash and across the unique source hashes represented in the 660-query evaluation set. Table 8 summarizes the metrics.

Table 8. Retrieval metrics used in the component comparison.

Metric	Interpretation
Precision@k	Proportion of the first k retrieved chunks satisfying the relevance criterion.
Recall@5	Proportion of eligible relevant evidence recovered within the first five results.
MRR@5	Reciprocal rank of the first relevant result within the first five positions; zero if none appears.
nDCG@5	Ranking quality that gives greater weight to relevant evidence appearing earlier in the list.

Twelve paired source-level comparisons test hybrid RRF against BM25 and dense retrieval across the six reported metrics. Holm adjustment is applied across the family of tests. This experiment evaluates the combined hybrid retriever; it does not independently isolate RRF from alternative fusion methods or the contribution of PRF.

4.8. Hierarchical Information Preservation

The original production population is analyzed separately to examine whether information introduced at the three levels of the hierarchy remains visible downstream. Table 9 summarizes the corresponding analyses.

Table 9. Evaluation of country-, community-, and persona-level structure.

Level	Output Analyzed	Evaluation
Country–topic	Scalar opinion	Mean opinion profiles and factorial OLS characterize differences across country–topic conditions.
Community	Free-text reasoning	TF–IDF features and regularized logistic regression predict institutional category and exact community membership under persona-grouped five-fold cross-validation.
Community	Scalar opinion	Change in R^2 after adding community and community–topic terms measures additional explained variance beyond country and topic.
Persona	Free-text reasoning	TF–IDF unigram and bigram features and regularized logistic regression predict political orientation, gender, and age under five-fold stratified cross-validation using balanced accuracy and macro-averaged one-versus-rest AUC.

Community classification uses persona-grouped five-fold cross-validation so that multiple topic responses from the same persona remain within the same fold. Persona-attribute recoverability is evaluated separately using five-fold stratified cross-validation.

A complementary community analysis compares TF–IDF cosine similarity between each persona explanation and its assigned community–topic description with similarity to alternative communities from the same country and topic. These analyses test whether community information remains visible in downstream output; they do not isolate the causal contribution of community retrieval from community descriptions or other structured inputs.

4.9. Information-Environment Diagnostics

Two diagnostics test whether changing the information environment changes retrieval and downstream opinion. Table 10 shows the breakdown below.

Table 10. Information-environment diagnostics.

Diagnostic	Manipulation	Measures
Explicit-identity retrieval	Identity attributes are included in the query.	Retrieval agreement, source slant, and opinion differences.
Corpus replacement	Persona attributes are fixed while the corpus changes.	Paired opinion change.

The explicit-identity benchmark contains 4320 persona–topic retrieval cases per repetition. The retrieval procedure is repeated across 20 random seeds, yielding 86,400 retrieval events across the robustness analysis. Because country, topic, political orientation, age, and gender are explicitly included in the query, this benchmark evaluates implementation of a specified exposure mechanism rather than inference under ambiguous information needs.

The corpus-replacement experiment holds persona attributes and generation settings fixed while changing the documentary corpus, testing whether the evidence environment changes generated opinion.

4.10. Evaluation Scope

The experiments deliberately separate retrieval performance, documentary faithfulness, generated opinion structure, and recoverability of country-, community-, and persona-level information. This distinction prevents high retrieval accuracy or strong population differentiation from being treated as evidence of factual grounding. Social interaction quality, selection of appropriate interaction partners, consistency across related topics, realistic confusion or uncertainty, and adaptation to novel situations remain outside the present evaluation.

5. Results

The results compare retrieval placement, assess documentary support in generated claims, and evaluate alternative retrieval methods. Supporting analyses then examine country, community, and persona information in the original production population and sensitivity to changes in the information environment. These analyses are reported separately because they use different evaluation samples and answer different questions.

5.1. Comparison of the Architecture

The three conditions use the same saved demographic skeletons, generation-model configuration, base seed, and prompt templates. No-RAG disables documentary retrieval. Single-stage hybrid RAG enables retrieval only during persona construction. Full AURORA-RAG enables retrieval at all four stages shown in Table 4 from earlier. The generation scaffold remains hierarchical in every condition, so "single-stage" refers to retrieval placement rather than to a different, flat generation architecture.

Each architecture contains 1440 generated personas and 4320 persona–topic records across four countries and the three focal geopolitical topics. Matching is based on saved persona skeletons and 96 country–community-category slots; community names and descriptions are generated separately under each condition. Only one base generation seed is represented.

Changes in Community–Topic Opinion Structure

Using the displacement measure defined in Equation (8), mean displacement from the no-RAG population is 0.103 for single-stage retrieval and 0.210 for full AURORA-RAG (Table 11 and Figure 3). The between-condition difference in displacement is 0.107, with a conditional 95% bootstrap interval of [0.081, 0.132]. An exploratory paired test across the 96 matched slots gives $t(95) = 7.06$, $p < 0.001$, and $d_z = 0.72$. Intervals resample slots within each fixed country using 10,000 bootstrap replicates; they do not incorporate uncertainty from regenerating country states or repeating the entire experiment.

Table 11. Displacement of community–topic mean opinions from no-RAG across the three focal topics. Values in brackets are conditional 95% country-stratified community-slot bootstrap intervals. Larger values indicate greater change from the no-RAG population, not greater accuracy.

Architecture	Mean Absolute Displacement from No-RAG
No-RAG	Reference
Single-stage hybrid RAG	0.103 [0.089, 0.119]
Full AURORA-RAG	0.210 [0.189, 0.231]

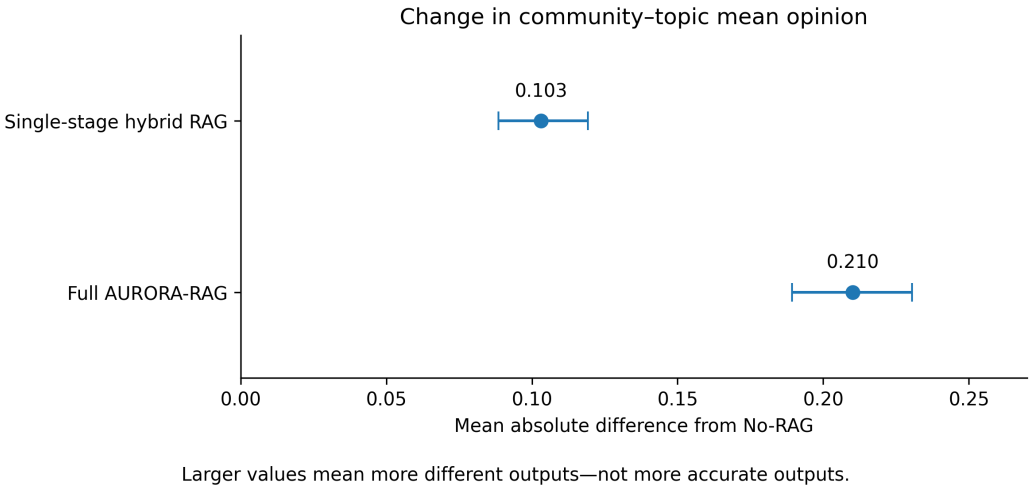

Figure 3. Absolute displacement of community–topic mean opinions from their no-RAG counterparts across the three focal topics. Error bars are conditional 95% country-stratified community-slot bootstrap intervals. No-RAG is the reference and is not plotted as a zero-quality system.

Table 11 uses no-RAG as the reference condition. The single-stage condition retains the hierarchical generation scaffold but enables retrieval only during persona construction, whereas full AURORA-RAG additionally enables retrieval at the country–topic, community, and community–topic stages. Mean absolute displacement of community–topic mean opinion scores from the no-RAG reference is 0.103 under single-stage retrieval and 0.210 under full AURORA-RAG, representing an additional displacement of 0.107. With demographic skeletons, generation scaffolds, model configurations, and prompt templates held fixed, these results suggest that retrieval placement influences the resulting opinion distributions. In particular, distributing retrieval throughout the hierarchy produces approximately twice the displacement observed with persona-level retrieval alone.

These values quantify the magnitude of output change rather than factual accuracy or behavioral realism. The reported intervals are conditional on the generated upstream states and the single base generation run and do not capture uncertainty across independently regenerated populations. The claim-level evaluation in Section 5.2 separately assesses whether generated statements are supported by documentary evidence.

5.2. Claim-Level Faithfulness and Attribution

The architecture comparison measures changes in generated opinion structure; the claim-level evaluation examines whether the generated statements are supported by documentary evidence. Corpus support measures whether a factual claim is supported by the reference corpus, whereas evidence faithfulness restricts this assessment to passages supplied during generation. Attribution accuracy evaluates whether named sources are correctly associated with the statements attributed to them, and citation correctness evaluates whether cited passages support their associated claims. The following results concern the three focal geopolitical topics.

Corpus Support and Hallucination Frequency

Full AURORA-RAG achieves the highest corpus-support rate among the three architectures (Table 12). Corpus-supported claims increase from 68% under no-RAG to 82% under single-stage hybrid RAG and 91% under full AURORA-RAG. Relative to single-stage retrieval, the full architecture improves corpus support by 9 percentage points (OR = 2.20, 95% CI [1.70, 2.85], $p < 0.001$); relative to no-RAG, the improvement is 23 percentage points

(OR = 4.75, 95% CI [3.65, 6.18], $p < 0.001$). Contradicted claims simultaneously decrease from 8% to 5% to 2%.

Table 12. Documentary support outcomes across the three architectures. Values are percentages of evaluated factual claims.

Architecture	Corpus-Supported Claims (%)	Unsupported/ Hallucinated (%)	Contradicted Claims (%)
No-RAG	68	24	8
Single-stage hybrid RAG	82	13	5
Full AURORA-RAG	91	7	2

Hallucination frequency shows the inverse pattern, decreasing from 24% under no-RAG to 13% under single-stage hybrid RAG and 7% under full AURORA-RAG. The full architecture reduces hallucination by 6 percentage points relative to single-stage retrieval (OR = 0.50, 95% CI [0.37, 0.67], $p < 0.001$) and by 17 percentage points relative to no-RAG (OR = 0.24, 95% CI [0.18, 0.33], $p < 0.001$). Overall, progressively integrating retrieval across the hierarchy increases documentary support while reducing unsupported and contradicted claims.

Faithfulness to Supplied Evidence

Among the retrieval-enabled architectures, full AURORA-RAG also achieves higher faithfulness to the documentary evidence supplied during generation (Table 13). Supplied passages support 79% of factual claims under single-stage RAG and 90% under full AURORA-RAG, an improvement of 11 percentage points. The odds of a claim being supported by the supplied evidence are higher under the full architecture (OR = 2.41, 95% CI [1.78, 3.27], $p < 0.001$). This comparison goes beyond whether relevant information exists somewhere in the corpus: it evaluates whether the generated claim is supported by evidence made available during generation.

Table 13. Faithfulness, attribution accuracy, and citation correctness in the retrieval-enabled conditions. Values are percentages; each metric uses its own eligible claims, attributions, or citations as the denominator. A dash indicates that the metric was not assessed for that condition.

Architecture	Evidence Faithfulness (%)	Attribution Accuracy (%)	Citation Correctness (%)
No-RAG	–	–	–
Single-stage hybrid RAG	79	82	85
Full AURORA-RAG	90	92	94

Attribution and Citation Quality

The full architecture has higher observed attribution accuracy and citation correctness than single-stage retrieval. Attribution accuracy increases from 82% to 92%, while citation correctness increases from 85% to 94%. These differences indicate that outputs from the full architecture more often associate statements with the appropriate sources and attach citations that support the associated factual claims. The attribution and citation comparisons are reported descriptively.

The claim-level results provide evidence of an advantage for full AURORA-RAG on the evaluated documentary-support measures. Both retrieval-enabled systems produce a higher proportion of corpus-supported claims and a lower reported hallucination frequency than no-RAG, with the full architecture achieving the strongest observed performance. Compared with single-stage retrieval, full AURORA-RAG additionally produces more

claims supported by supplied evidence and higher attribution and citation correctness. Unlike opinion displacement, these outcomes assess correspondence between generated statements and documentary evidence rather than simply the magnitude of output change.

5.3. Retrieval Component Comparison

To assess the hybrid retriever separately from retrieval placement, we compared BM25, dense retrieval, and hybrid RRF using a source-balanced, leave-one-source-out diagnostic. A retrieved chunk was considered relevant when it came from a different source but matched the query's country and topic. All methods used the same rank-five cutoff for reciprocal rank, with MRR@5 assigned zero when no relevant item appeared among the first five results. Table 14 summarizes the comparison.

Table 14. Source-balanced, leave-one-source-out retrieval results. Relevance is defined by country and topic labels rather than factual claim support. Higher values indicate better performance under the specified relevance criterion.

Retriever	P@1	P@3	P@5	R@5	MRR@5	nDCG@5
BM25	0.313	0.250	0.218	0.093	0.412	0.237
Dense	0.255	0.260	0.240	0.108	0.366	0.247
Hybrid RRF	0.319	0.261	0.241	0.107	0.417	0.255

Hybrid RRF achieved the highest observed score on five of the six metrics, while dense retrieval had marginally higher Recall@5. However, most differences were small, particularly between hybrid RRF and BM25. Twelve exploratory paired source-level tests compared hybrid RRF with the two individual retrievers across the six metrics. After Holm adjustment, only the MRR@5 advantage over dense retrieval remained significant (unadjusted $p = 0.004$, $p_{\text{Holm}} = 0.049$). The Precision@1 difference did not remain significant after adjustment (unadjusted $p = 0.016$, $p_{\text{Holm}} = 0.175$), and no adjusted comparison established an advantage over BM25.

5.4. Country, Community, and Persona Structure

We next examine how information associated with the three levels of the generation hierarchy appears in the saved population: country–topic context in opinion scores, community membership in generated explanations, and individual attributes in explanation language. These analyses use 1440 personas and 4320 responses across the three focal topics and characterize the original production population.

Country–Topic Information

Mean opinion scores vary substantially across country–topic conditions. For Russian aggression, the mean is 0.897 in the Russian condition, compared with −0.612 in the United States and −0.790 in the United Kingdom. For Taiwan sovereignty, the United States and United Kingdom have positive means (0.891 and 0.712), while India and Russia have negative means (−0.506 and −0.651). Cyber cooperation follows a different pattern, with the highest mean in India (0.786) and a near-neutral mean in Russia (0.008). Figure 4 shows the results below.

A factorial OLS model explains $R^2 = 0.9680$ of scalar opinion variance, with a large country–topic interaction, $F(6, 4305) = 19{,}491.09$, $p < 0.001$, partial $\eta^2 = 0.9645$. The results therefore indicate strong topic-specific country conditioning rather than a uniform country effect across issues. Because personas share upstream country and community states, these statistics characterize structure in the synthetic generation process rather than effects estimated from independent human observations.

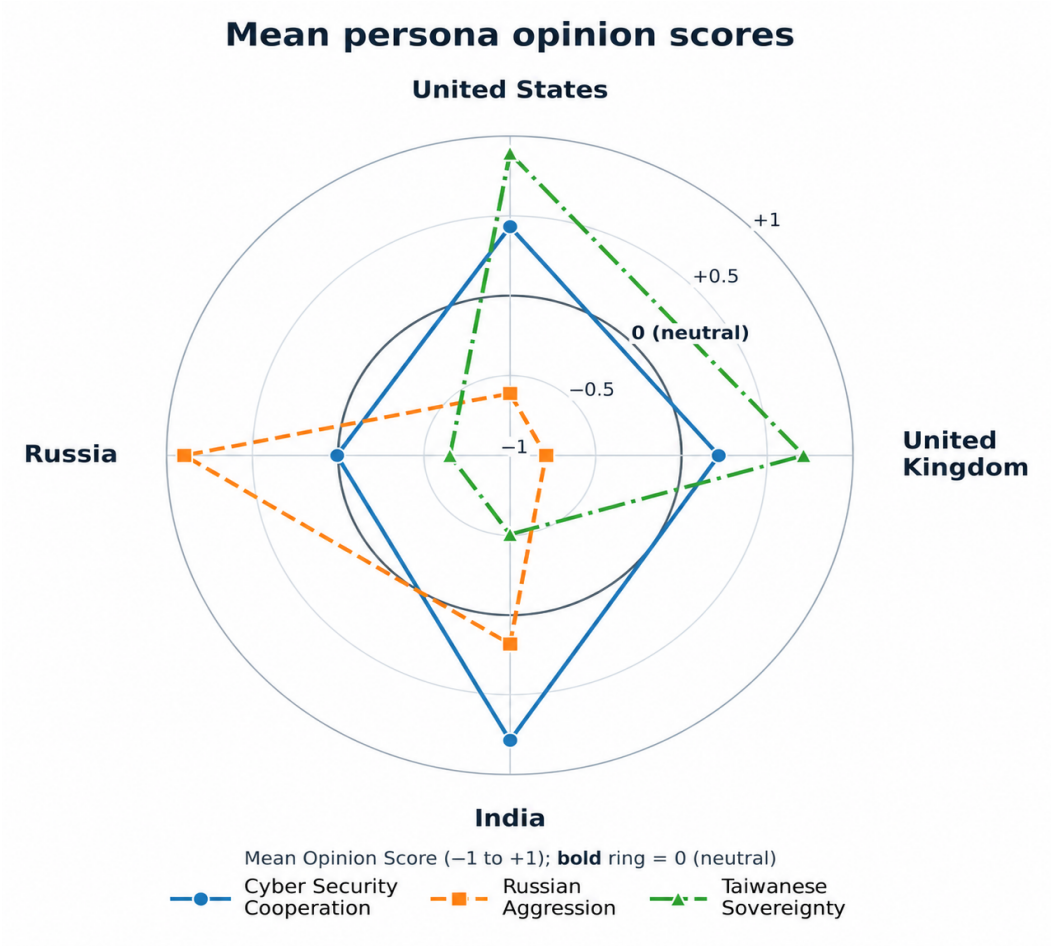

Figure 4. Mean synthetic opinion scores by country and topic.

Community Information

Figure 5 shows the community membership is strongly recoverable from the generated explanations. Using persona-grouped cross-validation, mean balanced accuracy reaches 60.6% for the institutional category, compared with a chance baseline of 12.5%, and 37.3% for exact community membership, compared with approximately 4.2% chance. Thus, the institutional category is recovered at nearly five times the chance rate, while exact community membership is recovered at nearly nine times chance. These results indicate that generated explanations retain substantial information associated with the communities to which personas were assigned.

A complementary text-based comparison shows the same pattern. Persona reasoning has greater mean TF–IDF similarity to its assigned community–topic description (0.065) than to alternative community descriptions from the same country and topic (0.042), a relative increase of approximately 55%. This suggests that community-specific information remains visible in the language and reasoning produced by downstream personas.

Community membership has a much smaller effect, however, on the final scalar opinion score. After accounting for country and topic, adding community fixed effects increases R^2 by only 0.00075, while community–topic effects increase it by 0.00206. This suggests that community membership is more strongly associated with variation in *how* personas explain and justify their positions than with variation in the scalar opinion measure. In other words, country and topic account for most of the variation in overall stance, while community-level conditioning is expressed primarily through differences in reasoning and language.

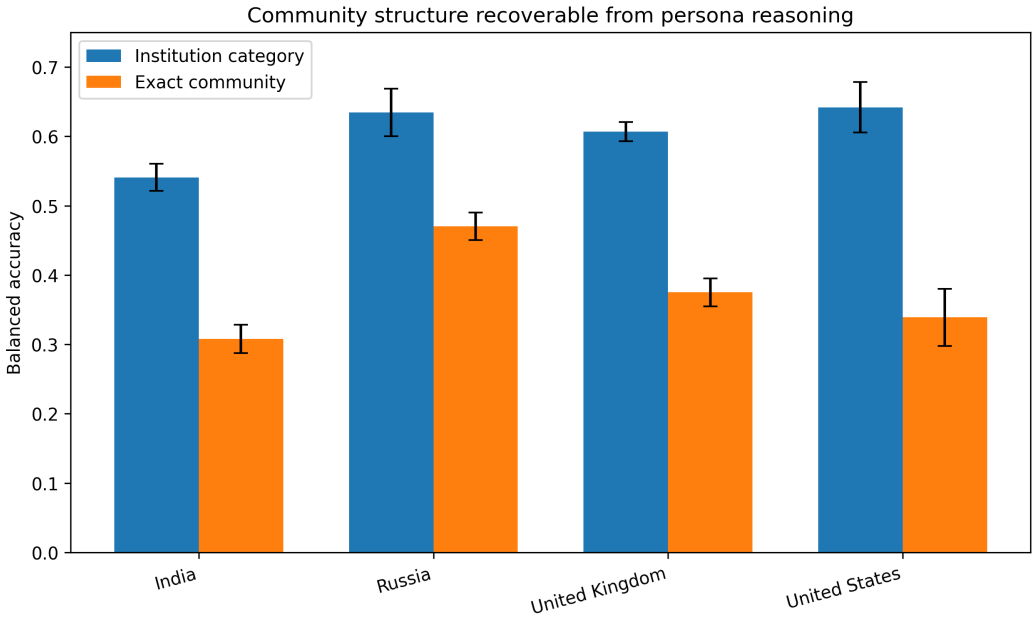

Figure 5. Recoverability of institutional category and exact community membership from original production-population reasoning. Balanced accuracy is shown separately for each country condition. Error bars represent cross-validation fold standard deviations, not uncertainty across independent generation runs.

This distinction is important for interpreting the hierarchical architecture. The results demonstrate that community information remains detectable in downstream persona outputs without dominating the resulting opinion score.

Individual Reasoning Within Country and Community Context

A persona's explanation is generated within the country–topic and community context established earlier in the pipeline. Country–topic information provides the broader scenario perspective, while the assigned community supplies more specific group context. The opinion score is established from upstream distributions before the LLM writes the explanation. The generated reasoning therefore expresses an assigned position through the persona's profile and inherited context, rather than independently determining that position.

To examine whether individual characteristics remain visible within this shared context, we tested whether political orientation, gender, and age could be identified from the generated explanations. Political orientation was the most recoverable attribute, with AUC = 0.842 and balanced accuracy = 76.8%, followed by gender (AUC = 0.756; balanced accuracy = 69.1%) and age (AUC = 0.631; balanced accuracy = 58.6%) (Table 15). These results indicate that the explanations contain information associated with individual attributes, even though personas share broader country and community context.

Table 15. Recoverability of assigned persona attributes from original production-population explanations. Higher classification scores indicate more recognizable identity information in the text, not greater factual accuracy or behavioral realism.

Persona Attribute	AUC	Balanced Accuracy (%)
Political orientation	0.842	76.8
Gender	0.756	69.1
Age	0.631	58.6

The distinction between stance and explanation is important. Two personas can have similar opinion scores while describing their positions in different terms. Shared country–topic information provides a common background, and community membership supplies group-specific context, but the resulting explanations can still contain distinguishable political, gender, and age signals. Together with the preceding community analysis, the classification results show that shared context does not remove all community-associated and individual information from the generated language.

The combined interpretation follows the generation hierarchy: country–topic context contributes to the broader stance, community information remains detectable in the explanations, and individual attributes are reflected in how those explanations are worded. This pattern is consistent with personas expressing their positions through both shared context and individual characteristics.

5.5. Information-Environment Interventions

These experiments test two ways the information environment affects synthetic opinion: conditioning retrieval on persona attributes and replacing the documentary corpus while holding persona specification fixed. They are mechanism diagnostics of AURORA-RAG, not tests of real-world media selection or persuasion.

Identity-Conditioned Retrieval

In the controlled benchmark, country agreement is 98.61%, topic agreement is 99.28%, political-source congruence is 97.25%, and joint agreement is 95.25%. Across 20 retrieval seeds, mean joint agreement is 95.15%, with a minimum of 94.21%. Because these attributes are explicitly included in the query, the results verify that retrieval constructs the intended information environment.

Figure 6 shows clear ideological differentiation in retrieved evidence. Mean source slant is −1.286 for left-leaning personas and 1.297 for right-leaning personas, a difference of 2.583 points.

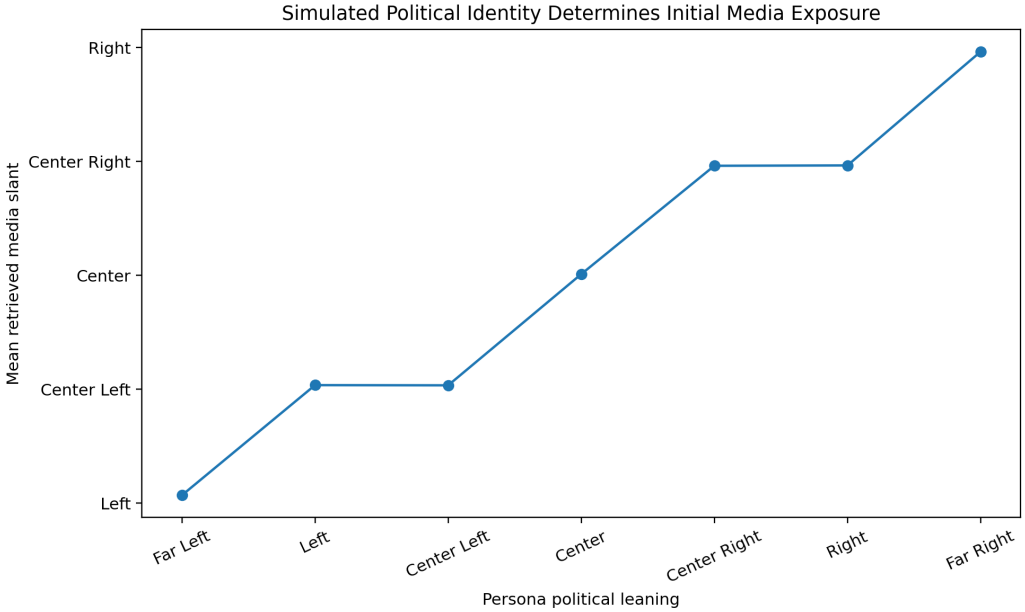

Figure 6. Mean retrieved source slant across political-orientation categories. Political orientation is supplied directly in the retrieval query, so the figure evaluates the implemented exposure mechanism rather than human media-selection behavior.

Downstream opinion differences are smaller and topic dependent. The pooled right-minus-left difference is −0.092 for cyber cooperation, 0.323 for Russian aggression, and

0.021 for Taiwan sovereignty. Thus, retrieval strongly changes the evidence supplied to the model, but not uniformly the generated stance.

Corpus Replacement

The second intervention replaces the documentary corpus while holding persona specification and generation settings fixed. For Russian aggression, mean opinion shifts from −0.600 under the U.S. corpus to 0.887 under the Russian corpus, an absolute difference of 1.487. The corresponding U.S.–Russia contrasts are 0.398 for cyber cooperation and 1.512 for Taiwan sovereignty.

Both identity-conditioned retrieval and corpus replacement alter the model's information environment, but their effects on generated opinion remain topic dependent and should not be interpreted as estimates of human persuasion or media susceptibility.

Full AURORA-RAG has the highest observed corpus-support and lowest unsupported-claim rates across the three architectures, with higher supplied-evidence faithfulness than persona-only retrieval. It also shows greater displacement of community–topic mean opinions from no-RAG than persona-only retrieval in the single-seed architecture comparison. Hybrid RRF has the highest observed values on five of six retrieval metrics, but only its MRR@5 advantage over dense retrieval remains significant after Holm correction. Production-population analyses identify country–topic structure in scalar opinions and recoverable community and persona information in explanations. Controlled retrieval and corpus interventions alter documentary exposure and are associated with topic-dependent differences in opinion scores.

6. Discussion

6.1. Hierarchical Retrieval Improves Grounding and Output Conditioning

AURORA-RAG advances grounded synthetic-population construction by introducing documentary evidence where contextual states are formed. Building on AURORA's population-generation framework [3], it aligns retrieval with the country–topic, community, community–topic, and persona stages. The central result is that this distributed evidence pipeline improves documentary support beyond both no-RAG and persona-only retrieval in the evaluated configurations.

Corpus support increases from 68% under no-RAG to 82% under persona-only retrieval and 91% under full AURORA-RAG, while unsupported-claim frequency decreases from 24% to 13% and 7% (Table 12). Among retrieval-enabled conditions, supplied-evidence faithfulness increases from 79% to 90%, with higher observed attribution accuracy (82% to 92%) and citation correctness (85% to 94%) (Table 13). The advantage therefore extends beyond differentiated outputs: more generated statements are supported by documentary evidence, and their sources are more accurately identified.

These findings build on the external-knowledge principle established by REALM, RAG, and Atlas [1,4,7]. Whereas those studies investigate retrieval-augmented learning and knowledge-intensive tasks, AURORA-RAG examines how evidence should enter a population-construction hierarchy. Its no-RAG and persona-only baselines retain the same generation scaffold, making the comparison informative about the added value of distributing retrieval across contextual-state construction. The results support the methodological premise that evidence should be introduced at the levels where contextual information is formed and propagated.

The opinion-structure comparison provides a complementary result: mean displacement from no-RAG is 0.103 under persona-only retrieval and 0.210 under full AURORA-RAG (Table 11). These values characterize output differences within the single-base-run comparison; the claim-level results establish the accompanying documentary-grounding advantage.

6.2. Hybrid Retrieval Supports Hierarchical Evidence Selection

The retrieval-component experiment supports hybrid RRF as an evidence-selection mechanism for AURORA-RAG. It achieves the highest observed value on five of six ranking metrics, with its MRR@5 advantage over dense retrieval remaining significant after Holm correction (Table 14). The adjusted advantage is specific to that comparison, while BM25 remains competitive. This result supports the practical value of rank fusion [11] for improving early relevant-evidence ranking over the evaluated dense-only channel.

The design combines the semantic-matching principle represented by dense retrieval research [8] with BM25's lexical matching [10]. AURORA-RAG uses pretrained chunk embeddings rather than DPR while retaining the complementary roles of semantic similarity and exact terminology. RRF combines these rankings without calibrating their raw scores, allowing both channels to serve hierarchy-specific queries over a common, provenance-preserving evidence base.

The reported grounding gains are obtained with PRF-enabled hybrid retrieval. HyDE-inspired retrieval and pseudo-document reranking inspired by Query2doc [14,15] remain optional extensions and are excluded from the architecture comparison. This separation makes the evaluated configuration explicit while preserving a consistent evidence boundary: query reformulation can guide selection, but indexed corpus passages remain the documentary evidence.

6.3. The Hierarchy Preserves Country, Community, and Persona Information

The production-population analyses show that shared contextual grounding and individual differentiation coexist in AURORA-RAG. Country–topic structure is evident in scalar opinions, while community and persona information remains recoverable from explanations. This follows the pipeline's division of roles: upstream states inform assigned stances, and generated explanations express those stances through community context and individual characteristics.

Institutional category is recoverable with 60.6% balanced accuracy against a 12.5% chance baseline and exact community membership with 37.3% against approximately 4.2% chance. Explanations are also more similar to assigned community–topic descriptions than alternatives (0.065 versus 0.042), while community terms contribute comparatively little additional scalar-opinion variance. Political orientation, gender, and age remain detectable within the shared context. The hierarchy therefore preserves distinguishable ways of explaining positions even when scalar opinions are similar.

These findings complement Argyle et al.'s work on demographic conditioning and correspondence with human samples [24] by examining how information remains identifiable across multiple contextual levels. They also connect to Park et al.'s use of retrieved experiential memory to support context-sensitive agents [27]. AURORA-RAG contributes a documentary-grounded construction pathway in which national context, community information, and individual attributes remain distinguishable downstream. This provides an empirical basis for combining shared information environments with differentiated persona expression.

6.4. Retrieval Creates a Controllable Information Environment

AURORA-RAG makes documentary exposure an explicit part of the synthetic-population specification. Beyond supplying evidence for an answer, retrieval determines which external information enters contextual-state and persona construction. Keeping this information environment separate from demographic attributes makes exposure assumptions configurable and allows their consequences to be examined within the generation pipeline.

The identity-conditioned diagnostic demonstrates this control: mean source slant shifts from −1.286 for left-leaning personas to 1.297 for right-leaning personas when political orientation is included directly in the query. This diagnostic extension implements a specified exposure rule, while production retrieval uses country and community context without directly querying demographic attributes. Studies of partisan selection [28,29] and platform-mediated exposure [30] motivate distinguishing actor characteristics from available information. AURORA-RAG translates that distinction into an experimentally controllable retrieval mechanism.

Corpus replacement provides complementary evidence of information-environment sensitivity: changing the documentary corpus while holding persona specifications and generation settings fixed yields topic-dependent differences in pipeline-generated opinions. This operationalizes the controlled-information principle used in simulated economic-agent research [34] through traceable documentary inputs. The interventions establish a way to study how specified evidence environments relate to synthetic outputs without changing the underlying demographic specification.

6.5. Implications for Synthetic-Agent Evaluation

AURORA-RAG brings documentary grounding and population structure into a coordinated evaluation. Following the component-level perspective of RAGAS, ARES, and RAGChecker [16–18], the study assesses retrieval relevance, corpus support, supplied-evidence faithfulness, and attribution separately. It adds hierarchical information recoverability to determine whether contextual distinctions remain visible downstream. This combination distinguishes a population that is differentiated from one whose factual statements are supported while showing that AURORA-RAG can achieve both properties.

EconAgent evaluates heterogeneous agents through repeated economic decisions and macroeconomic dynamics [35], while AgentSociety emphasizes large-scale social interaction [36]. AURORA-RAG contributes a complementary foundation for such simulations: constructing personas with explicit documentary inputs and assessing the support for their generated statements. Its separation of demographic specifications, contextual states, and retrieved evidence offers a way to examine information-environment assumptions alongside subsequent behavioral outcomes.

Human-study replication [25] and dynamic persona coherence [37] provide complementary targets for extending this evaluation. AURORA-RAG's contribution is to make the evidence environment available for inspection and controlled variation before adding longitudinal interaction. Combining this construction pipeline with state-update and interaction mechanisms would enable investigation of how documentary grounding relates to belief updating and sustained persona coherence.

6.6. Computational Cost and Scalability

The controlled evaluation examines 1440 personas per architecture, with 4320 persona–topic responses across three geopolitical topics and a documentary corpus containing 3888 indexed passages. Beyond this experimental population, AURORA-RAG has also been used to construct populations of up to 5000 personas, demonstrating execution at a larger population size. However, computational scalability involves two distinct dimensions: expanding the synthetic population increases generation and retrieval operations, while expanding the documentary corpus increases preprocessing, embedding, storage, and evidence-selection requirements.

The hierarchical architecture supports reuse of shared contextual states across descendant personas. As described in Algorithm 1, country–topic, community, and community–topic states are generated before persona construction and subsequently reused. For P

personas and T topics, the downstream procedure requires P persona-construction operations and PT persona–topic explanation operations, giving $P(1 + T)$ logical generation operations when the upstream structure is fixed. This relationship characterizes the algorithmic workload rather than measured execution time, and the reuse of upstream states does not eliminate the additional retrieval operations required by full hierarchical conditioning.

Corpus scalability introduces different considerations. AURORA-RAG maintains aligned dense and BM25 indices over semantically segmented documents, allowing retrieval without providing the entire corpus to the generator. The production configuration returns three passages per retrieval stage, limiting the number of evidence items included in generation prompts. Nevertheless, retrieval cost depends on index size and configuration, while variable passage lengths affect token consumption. Pseudo-relevance feedback also introduces an additional retrieval and fusion pass. Consequently, increased corpus size may affect both computational requirements and the quality of evidence selection, even when the number of returned passages remains fixed.

The larger population-generation runs demonstrate that the implementation can construct populations beyond the controlled experimental size, but they do not establish linear execution-time scaling or equivalent grounding quality at that scale. Similarly, the current experiments do not systematically vary corpus size or report comparative latency, memory usage, or generation cost. Future scalability evaluations should vary population and corpus sizes independently while measuring computational requirements, retrieval relevance, and downstream claim support. Such experiments would establish whether the framework maintains its grounding advantages as the documentary environment and synthetic population increase in size.

6.7. Implications for Trustworthy RAG

The grounding, attribution, and citation results show the value of treating evidence placement, source traceability, and generation quality as connected design concerns. This complements KILT's joint attention to task performance and provenance [19]. Self-RAG addresses evidence use through learned reflection, while Yoran et al. investigate robustness to irrelevant context [20,21]. AURORA-RAG contributes an architectural approach: introduce evidence throughout contextual-state construction, preserve its source metadata, and assess its support for downstream claims. Its results support hierarchical evidence organization as a means of improving documentary grounding within structured generative pipelines.

Wu et al. identify retrieval noise, conflicting evidence, and external knowledge bases as sources of RAG risk [22], while PoisonedRAG demonstrates the consequences of malicious corpus additions [23]. Because documentary support is relative to the corpus, source-quality controls and adversarial testing remain important complements to provenance. AURORA-RAG provides the traceable evidence pipeline on which these controls can build, combining measured grounding gains with an inspectable account of how external information enters synthetic-population construction.

7. Conclusions

AURORA-RAG extends synthetic-persona construction by introducing documentary retrieval at the country–topic, community, community–topic, and persona stages rather than restricting retrieval to final persona generation. The framework combines dense and lexical retrieval with provenance-preserving evidence handling and distinguishes fixed demographic specifications from the documentary information environments used to construct contextual states. This separation supports the evaluation of retrieval relevance, factual grounding, and downstream contextual differentiation as related but distinct properties.

In the controlled four-country, three-topic evaluation, full AURORA-RAG achieves the highest corpus-support rate (91%, compared with 82% for persona-only retrieval and 68% for no-RAG) and the lowest unsupported-claim frequency (7%, compared with 13% and 24%, respectively). Among retrieval-enabled conditions, supplied-evidence faithfulness increases from 79% to 90%, attribution accuracy from 82% to 92%, and citation correctness from 85% to 94%. Full AURORA-RAG also produces greater displacement of community–topic mean opinion scores relative to no-RAG than persona-only retrieval (0.210 versus 0.103). This displacement indicates differences in generated opinion structure rather than improvements in factual accuracy or behavioral realism.

At the retrieval layer, hybrid RRF achieves the highest observed score on five of six ranking metrics, although only its MRR@5 advantage over dense retrieval remains significant after correction for multiple comparisons. Additional analyses demonstrate that country–topic information remains visible in scalar opinions, while community and persona characteristics are recoverable from generated explanations. Controlled retrieval and corpus-replacement interventions further show that documentary exposure can be varied independently of fixed persona specifications, producing topic-dependent differences in generated opinions. These findings establish the framework's utility for examining information-environment conditioning without implying that synthetic responses reproduce human information selection or persuasion.

The results support hierarchical evidence placement as an approach to improving documentary grounding and source traceability in structured LLM pipelines. Beyond the controlled evaluation, the implementation has also generated populations of up to 5000 personas, demonstrating execution at a larger population size without establishing computational scalability or equivalent grounding quality. Further research should assess retrieval performance and generation cost as corpus and population sizes increase, alongside source quality, translation fidelity, resistance to conflicting or manipulated evidence, and longitudinal behavioral consistency. AURORA-RAG therefore provides a foundation for studying how externally retrieved information is selected, propagated, and expressed throughout hierarchical synthetic-agent systems.

8. Limitations and Future Work

The present evaluation establishes documentary grounding and hierarchical information conditioning, but not complete behavioral validity. Recoverable country, community, and persona information does not establish realistic interaction, cross-topic consistency, uncertainty, or adaptation. Future work should therefore evaluate multi-agent behavior, longitudinal belief updating, contradictory evidence, and out-of-distribution situations.

The architecture experiments also do not isolate every component independently. The current study compares no-RAG, persona-only retrieval, and full hierarchical AURORA-RAG and separately compares BM25, dense retrieval, and hybrid RRF. Future work should ablate individual retrieval stages, PRF, and alternative fusion methods and evaluate the optional HyDE and Query2Doc paths.

Several evaluation results should also be interpreted within their experimental design. The explicit-identity benchmark supplies identity attributes directly in the query and therefore tests a controlled exposure mechanism rather than ambiguous retrieval. Claim-support measures are relative to the documentary corpus and do not guarantee objective truth when sources are stale, biased, conflicting, or incorrect. Independent source-quality assessment and human validation would strengthen future evaluations. Computational scalability remains a separate limitation. Although the implementation has generated populations of up to 5000 personas, systematic benchmarks are needed to establish execution-time

scaling, memory requirements, and grounding performance as corpus and population sizes increase independently.

Finally, personas share upstream states and information environments, so observations are not equivalent to independent human responses. The architecture comparison also uses one base generation seed. Future work should repeat complete population generation across multiple seeds and evaluate robustness to noisy retrieval, semantic duplicates, corpus poisoning, and changing information environments. These extensions will help determine how well AURORA-RAG balances grounding, contextual consistency, and adaptability.

Author Contributions: Conceptualization, R.M. and K.M.C.; methodology, R.M.; software, R.M.; validation, R.M. and K.M.C.; formal analysis, R.M.; investigation, R.M.; resources, R.M. and K.M.C.; data curation, R.M.; writing—original draft preparation, R.M.; writing—review and editing, R.M. and K.M.C.; visualization, R.M.; supervision, K.M.C.; project administration, R.M. and K.M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Knight Foundation; the Office of Naval Research under grant number N000142112229; the U.S. Army under contract number W911NF20D0002 through the Army Artificial Intelligence Integration Center (AI2C); and the Center for Informed Democracy and Social-cybersecurity (IDeaS) and the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University.The views and conclusions expressed in this article are those of the authors and do not necessarily represent the official policies or positions of the Knight Foundation, the Office of Naval Research, the U.S. Army, or the U.S. Government.

Institutional Review Board Statement: Not applicable. This study analyzed synthetic personas and documentary corpora and did not involve human participants, identifiable private human-subject data, or animal subjects.

Informed Consent Statement: Not applicable.

Data Availability Statement: The AURORA-RAG source code is not publicly available because it contains proprietary information. The code, synthetic-persona outputs, controlled benchmark data, analysis outputs, and figure data supporting the findings are available from the corresponding author upon reasonable request, subject to applicable permissions. Third-party documentary texts remain subject to their respective redistribution restrictions.

Acknowledgments: Large language models (LLMs) were used as part of the research methodology to generate synthetic personas and associated responses within the AURORA-RAG framework. Their use, configuration, and evaluation are described in the Methods section. The authors reviewed and analyzed the generated outputs and take full responsibility for the content and conclusions of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AURORA	AI-Utilized Retrieval for Optimized Representation of Audiences
BM25	Best Matching 25 lexical ranking function
CSV	Comma-Separated Value
HyDE	Hypothetical Document Embeddings
LLM	Large Language Model
PRF	Pseudo-Relevance Feedback
RAG	Retrieval-Augmented Generation
RRF	Reciprocal Rank Fusion
SHA-256	Secure Hash Algorithm 256-bit

References

1. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Kuttler, H.; Lewis, M.; Yih, W.t.; Rocktaschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–12 December 2020; Volume 33, pp. 9459–9474.
2. Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Jiang, J.; Cui, B. Retrieval-Augmented Generation for AI-Generated Content: A Survey. *Data Sci. Eng.* **2026**, *11*, 1–29. [CrossRef]
3. Marigliano, R.; Carley, K.M. AURORA: Enhancing Synthetic Population Realism Through RAG and Salience-Aware Opinion Modeling. In *Proceedings of the 2025 Winter Simulation Conference*; IEEE: Piscataway, NJ, USA, 2025; pp. 2563–2574.
4. Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; Chang, M.W. REALM: Retrieval-Augmented Language Model Pre-Training. In Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 13–18 July 2020; Proceedings of Machine Learning Research, Voume 119, pp. 3929–3938.
5. Borgeaud, S.; Mensch, A.; Hoffmann, J.; Cai, T.; Rutherford, E.; Millican, K.; Van Den Driessche, G.B.; Lespiau, J.B.; Damoc, B.; Clark, A.; et al. Improving Language Models by Retrieving from Trillions of Tokens. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; Proceedings of Machine Learning Research; Volume 162, pp. 2206–2240.
6. Izacard, G.; Grave, E. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 874–880. [CrossRef]
7. Izacard, G.; Lewis, P.; Lomeli, M.; Hosseini, L.; Petroni, F.; Schick, T.; Dwivedi-Yu, J.; Joulin, A.; Riedel, S.; Grave, E. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *J. Mach. Learn. Res.* **2023**, *24*, 11912–11954.
8. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.T. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the EMNLP, Online, 16–20 November 2020; pp. 6769–6781. [CrossRef]
9. Khattab, O.; Zaharia, M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, 25–30 July 2020; pp. 39–48. [CrossRef]
10. Robertson, S.; Zaragoza, H. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389.
11. Cormack, G.V.; Clarke, C.L.A.; Buttcher, S. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In Proceedings of the 32nd International ACM SIGIR Conference, Boston, MA, USA, 19–23 July 2009; pp. 758–759.
12. Rocchio, J.J. Relevance Feedback in Information Retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*; Salton, G., Ed.; Prentice-Hall: Englewood Cliffs, NJ, USA, 1971; pp. 313–323.
13. Lavrenko, V.; Croft, W.B. Relevance-Based Language Models. In Proceedings of the 24th Annual International ACM SIGIR Conference, New Orleans, LA, USA, 9–13 September 2001; pp. 120–127.
14. Gao, L.; Ma, X.; Lin, J.; Callan, J. Precise Zero-Shot Dense Retrieval without Relevance Labels. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 1762–1777. [CrossRef]
15. Wang, L.; Yang, N.; Wei, F. Query2doc: Query Expansion with Large Language Models. In Proceedings of the EMNLP, Singapore, 6–10 December 2023; pp. 9414–9423. [CrossRef]
16. Es, S.; James, J.; Espinosa-Anke, L.; Schockaert, S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, St. Julian's, Malta, 17–22 March 2024; pp. 150–158.
17. Saad-Falcon, J.; Khattab, O.; Potts, C.; Zaharia, M. ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Mexico City, Mexico, 16–21 June 2024; pp. 338–354.
18. Ru, D.; Qiu, L.; Hu, X.; Zhang, T.; Shi, P.; Chang, S.; Jiayang, C.; Wang, C.; Sun, S.; Li, H.; et al. RAGChecker: A Fine-grained Framework for Diagnosing Retrieval-Augmented Generation. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024; Volume 37.
19. Petroni, F.; Piktus, A.; Fan, A.; Lewis, P.; Yazdani, M.; De Cao, N.; Thorne, J.; Jernite, Y.; Karpukhin, V.; Maillard, J.; et al. KILT: A Benchmark for Knowledge Intensive Language Tasks. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 2523–2544. [CrossRef]
20. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
21. Yoran, O.; Wolfson, T.; Ram, O.; Berant, J. Making Retrieval-Augmented Language Models Robust to Irrelevant Context. In Proceedings of the The Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
22. Wu, S.; Xiong, Y.; Cui, Y.; Wu, H.; Chen, C.; Yuan, Y.; Huang, L.; Liu, X.; Kuo, T.W.; Guan, N.; et al. Retrieval-augmented generation for natural language processing: A survey. *Artif. Intell. Rev.* **2026**, *59*, 192. [CrossRef]

23. Zou, W.; Geng, R.; Wang, B.; Jia, J. PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. In Proceedings of the 34th USENIX Security Symposium, Seattle, WA, USA, 13–15 August 2025.
24. Argyle, L.P.; Busby, E.C.; Fulda, N.; Gubler, J.R.; Rytting, C.; Wingate, D. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Anal.* **2023**, *31*, 337–351. [CrossRef]
25. Aher, G.; Arriaga, R.I.; Kalai, A.T. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23-29 July 2023; Proceedings of Machine Learning Research, Volume 202, pp. 337–371.
26. Dillion, D.; Tandon, N.; Gu, Y.; Gray, K. Can AI Language Models Replace Human Participants? *Trends Cogn. Sci.* **2023**, *27*, 597–600. [CrossRef] [PubMed]
27. Park, J.S.; O'Brien, J.C.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative Agents: Interactive Simulacra of Human Behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, San Francisco, CA, USA, 29 October–1 November 2023. [CrossRef]
28. Iyengar, S.; Hahn, K.S. Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *J. Commun.* **2009**, *59*, 19–39. [CrossRef]
29. Stroud, N.J. Polarization and Partisan Selective Exposure. *J. Commun.* **2010**, *60*, 556–576. [CrossRef]
30. Bakshy, E.; Messing, S.; Adamic, L.A. Exposure to Ideologically Diverse News and Opinion on Facebook. *Science* **2015**, *348*, 1130–1132. [CrossRef] [PubMed]
31. Flaxman, S.; Goel, S.; Rao, J.M. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opin. Q.* **2016**, *80*, 298–320. [CrossRef]
32. Cinelli, M.; De Francisci Morales, G.; Galeazzi, A.; Quattrociocchi, W.; Starnini, M. The Echo Chamber Effect on Social Media. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2023301118. [CrossRef] [PubMed]
33. Gentzkow, M.; Shapiro, J.M. What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica* **2010**, *78*, 35–71. [CrossRef]
34. Filippas, A.; Horton, J.J.; Manning, B.S. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? In Proceedings of the 25th ACM Conference on Economics and Computation, New Haven, CT, USA, 8–11 July 2024; pp. 614–615. [CrossRef]
35. Li, N.; Gao, C.; Li, M.; Li, Y.; Liao, Q. EconAgent: Large Language Model-Empowered Agents for Simulating Macroeconomic Activities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Kerrville, TX, USA, 2024; pp. 15523–15536. [CrossRef]
36. Piao, J.; Yan, Y.; Zhang, J.; Li, N.; Yan, J.; Lan, X.; Lu, Z.; Zheng, Z.; Wang, J.Y.; Zhou, D.; et al. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. *iFuture* 2026, *accepted*. [CrossRef]
37. Qi, Y.; Zhang, X.; Zeng, R.; Liu, M.; Zhou, Z.; Miao, D.; Yan, B.; Guan, Z. Beyond Static Persona Consistency: Dynamic Persona Coherence in LLM Role-Playing. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Kerrville, TX, USA, 2026; pp. 28942–28956. [CrossRef]
38. Paglieri, D.; Cross, L.; Cunningham, W.A.; Leibo, J.Z.; Vezhnevets, A.S. Persona Generators: Generating Diverse Synthetic Personas for Arbitrary Contexts. *arXiv* **2026**, arXiv:2602.03545.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.