Retrieval-Augmented Generation Must Move Beyond Factual Grounding to Represent Diverse Opinions

Aditya Agrawal **Alwarappan Nakkiran** **Alex Karlsson Aman Singh Thakur** **Harsha Aduri**
Amazon.com
{aadag, nakkiran, nkarlss, amanzing, venaduri}@amazon.com

Abstract

Retrieval-Augmented Generation (RAG) systems are built on an unexamined assumption: that queries have correct answers and retrieval should converge toward them. This position paper argues that this creates a factual bias where RAG systems optimize for reducing epistemic uncertainty while ignoring the aleatoric uncertainty, inherent in opinion-rich content. The consequences go beyond technical limitations: due to risk of minority voice erasure and risk of opinion manipulation. To address this, we formalize opinion-aware retrieval through uncertainty quantification and derive a unified objective using the Wasserstein distance. As an existence proof, we present Opinion-Aware RAG (O-RAG), which enriches documents with LLM-extracted, entity-linked opinion metadata before indexing. Across e-commerce seller forums and public hotel reviews, O-RAG reduces Wasserstein distance to corpus-level sentiment distributions by 18–48%, and human evaluators preferred its responses 79.2% of the time. We close with a research agenda for opinion-aware RAG.

1 Introduction

Retrieval-Augmented Generation has become a central technique for grounding Large Language Models in external evidence, reducing hallucinations and enabling verifiable, up-to-date text generation (Lewis et al., 2020; Gao et al., 2023b; Gan et al., 2025). However, beneath this progress lies an unexamined assumption- that knowledge bases should contain factually correct, objective information, and that retrieval would converge toward the ground-truth. This assumption pervades the RAG ecosystem, from the datasets used for evaluation, to the retrieval mechanisms optimized for factual recall, to the generation objectives that reward consistency with a single correct answer.

Of 36 major RAG benchmarks surveyed (Table 1), only one addresses opinion synthesis- OpinioRAG (Nayeem and Rafiei, 2025). Two others (PIR (Zhao et al., 2024b), BERDS (Chen and Choi, 2025)) evaluate perspective-aware retrieval but not distributional fidelity. The rest assume questions have definitive answers and treat contradictory information as noise. This reveals a misalignment between how RAG systems are built and how a large share of real-world information actually behaves.

Table 1: Representative RAG evaluation datasets (9 of 36 surveyed). Only OpinioRAG addresses opinion synthesis; PIR and BERDS evaluate perspective-aware retrieval but not distributional fidelity. See Appendix M for the complete analysis.

Dataset	Year	Evaluation Focus	Opinion
OpinioRAG	2025	Sentiment Alignment	Yes
BERDS	2025	Perspective Coverage	Partial
PIR	2024	Perspective Recall	Partial
RAGBench	2024	TRACe Metrics	No
CRAG	2024	Accuracy, Hallucination	No
FActScore	2023	Atomic Fact Accuracy	No
BEIR	2021	nDCG, MAP, Recall	No
Natural Questions	2019	EM, Token F1	No
HotpotQA	2018	EM, F1, Sup. F1	No

Consider two queries: *"What is the corporate tax rate in the United States?"* versus *"What do small business owners think about the current tax policy?"* The first has a definitive answer that existing RAG systems handle well. The second requires synthesizing diverse, potentially conflicting perspectives from multiple sources. This is a task for which current architectures have no explicit design. Some LLMs demonstrate emergent opinion synthesis capabilities- producing balanced-sounding summaries that capture diverse viewpoints (see Appendix J for an example). However, the difference between accidental capability and intentional design matters when deploying systems, where understanding the sources and potential biases of opinion synthesis is critical.

We argue that current RAG systems exhibit

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 321–338
October 17-21, 2026 ©2026 Association for Computational Linguistics

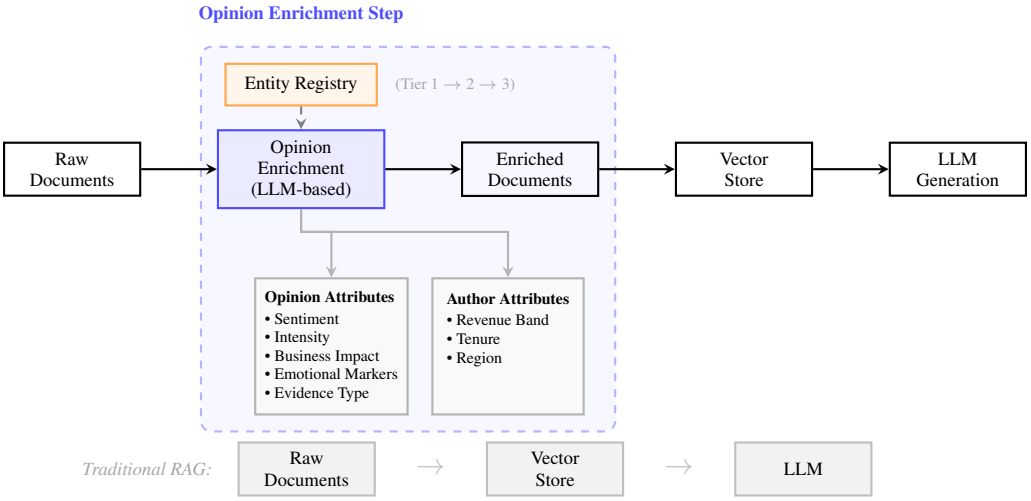


Figure 1: Opinion-Aware RAG Architecture. Traditional RAG (bottom, grayed) indexes raw documents directly. Our approach introduces an opinion enrichment step (highlighted) that extracts per-entity opinion and author attributes before indexing. See Section 4 for details.

a factual bias. They optimize for epistemic uncertainty reduction, while ignoring the aleatoric uncertainty inherent in opinion-rich content.

This limitation grows more consequential as RAG deployments expand beyond traditional knowledge bases into online community forums, consumer reviews, social media, and user-generated content (Guo and Wei, 2026; Zhang et al., 2025c). In these domains, the goal is to understand and represent *multiple truths* as expressed through human experiences and judgments, instead of converging on single truth. Without opinion-aware capabilities, systems risk echo chamber amplification, minority voice erasure, potential for opinion manipulation, and artificial consensus where genuine disagreement exists. While recent work on pluralistic alignment (Sorensen et al., 2024; Bakker et al., 2022) addresses diversity at the generation layer, it cannot compensate for biased retrieval, as a generation model can only reason over what it receives.

Our contributions are:

1. An **audit** of factual bias across the RAG ecosystem. We document the absence of opinion-aware evaluation (Section 2).

2. A **theoretical framework** formalizing opinion-aware retrieval through epistemic vs. aleatoric uncertainty (Section 3).

3. An **existence proof**: Opinion-Aware RAG evaluated across two domains (e-commerce forums and public hotel reviews) using a novel evaluation framework measuring opinion coverage, perspective diversity, and representation fairness. We demonstrate that even simple opinion enrichment yields measurably more representative retrieval and generation (Sections 4–5, Figure 1).

4. A **research agenda** identifying open problems for the community (Section 6).

Crucially, our position is a claim about *generation*, not retrieval alone: the objective is opinion-preserving text generation, where the distribution of perspectives is faithfully carried through to the produced output, since a system that surfaces diverse viewpoints but collapses them into an artificial consensus during generation still fails the user.

2 The Factual Bias in RAG: An Audit

The factual bias in RAG is a field-wide pattern that runs through the entire pipeline, from how benchmarks are constructed to how retrieval and evaluation are designed. We document this through a survey of 36 major RAG evaluation datasets (2013–2025), summarized in Table 1 and detailed in Appendix M.

The pattern is consistent across task categories: single-hop QA (Natural Questions, MS MARCO, SQuAD), multi-hop reasoning (HotpotQA, MuSiQue), long-form QA (ELI5, ASQA), domain-specific QA (COVID-QA, Qasper), fact verification (FEVER, TruthfulQA), and generation quality (FActScore, RAGBench). Every category

assumes questions have correct answers and evaluates systems on their ability to converge toward them.

This bias becomes consequential as RAG deployments expand into domains where understanding human opinion is the primary goal (Zhao et al., 2024a; Brown et al., 2025; Fan et al., 2024): online community forums, consumer reviews, customer service, policy analysis, healthcare patient feedback, and political discourse. In these settings, information is inherently subjective, and collapsing it to a single answer misrepresents reality. Yet recent work actively reinforces this mismatch by framing subjectivity as noise: RAGuard (Zeng et al., 2025) treats Reddit discussions as naturally occurring misinformation to be filtered, CONFLICTS (Cattan et al., 2025) focuses on conflict resolution rather than preserving perspective diversity, and robustness approaches (Chen et al., 2022; Choi et al., 2025; Zhang et al., 2025a) emphasize filtering contradictory information. While valuable for factual domains, these approaches are misaligned with scenarios where diverse perspectives are the goal.

The boundary between factual and opinion content is blurred in practice. Even a factual query, such as "What are the shipping rates for small parcels?", will surface opinion-laden forum discussions alongside authoritative sources when the knowledge base includes community content. A semantic similarity-based retriever cannot distinguish an official rate schedule from a seller's complaint about rate increases. Current systems have no mechanism to handle this mixture, treating all retrieved content as equivalent evidence. Existing diversity methods do not solve this: Maximal Marginal Relevance (Carbonell and Goldstein, 1998) and Determinantal Point Process-based methods promote diversity by penalizing redundancy in embedding space, but semantic diversity and opinion diversity are fundamentally different objectives. Two documents expressing opposing views on the same topic may use similar vocabulary and land close together in embedding space. MMR would penalize retrieving both; opinion-aware retrieval should specifically surface both.

Additionally, current RAG-based metrics can assign high scores to biased outputs. A response to *"What leadership qualities are most important for startup CEOs?"* that emphasizes only traditionally masculine-coded traits could score excellently on ROUGE-L, context relevance, and answer faithfulness. However, it would underrepresent collaborative and inclusive leadership perspectives (Wang et al., 2025; Zhang et al., 2025b; de Oliveira et al., 2025; Thakur et al., 2025) (see Appendix K for a concrete example). Traditional metrics tend to reward convergence and cannot detect what was excluded.

Related techniques like argument mining and stance detection have matured considerably (Al-Dayel and Magdy, 2021; Hardalov et al., 2022), with work spanning argument-based retrieval (Li et al., 2025), sentiment analysis on decentralized platforms (Chhetri et al., 2025), and multimodal/comment-level opinion analysis (Sharma et al., 2020; Li et al., 2018; Weißmann et al., 2025; Sun et al., 2024), yet these capabilities remain isolated from RAG frameworks. RAFTS (Yue et al., 2024) synthesizes contrastive arguments but targets claim credibility, not opinion diversity. LLMs' inability to reliably detect contradictions (Lee et al., 2025; Gokul et al., 2025) compounds the challenge. OpinioRAG (Nayeem and Rafiei, 2025) bridges this gap via RAG-based opinion summarization but focuses on condensing opinions into narratives rather than ensuring the retrieved set faithfully represents the opinion spectrum.

Latest work on perspective benchmarks is a step in this direction, but is incomplete. PIR (Zhao et al., 2024b) and BERDS (Chen and Choi, 2025) confirm that current retrievers are perspective-blind: PIR shows systematic bias toward supporting stances ($\sim$58.5%), while BERDS finds full perspective coverage in only 40% of cases. However, both evaluate categorical *perspective* presence (is viewpoint X retrieved?) rather than *opinion* distributions - the intensity, affect, and demographic grounding that characterize real-world subjective content.

Pluralistic alignment is another relevant body of work which is complementary to opinion-aware retrieval. A growing body of work addresses pluralism at the generation layer: Sorensen et al. (2024) formalize pluralistic value alignment for LLMs, Bakker et al. (2022) fine-tune models to find agreement among diverse preferences, and Argyle et al. (2023) frame LLMs as "silicon sampling" of human opinions. However, pluralistic alignment trains the *generation model* to preserve opinion distributions, while opinion-aware retrieval ensures the *retrieved context* faithfully represents the distribution *before* generation. These are complementary: a pluralistically-aligned LLM given biased retrieved context still produces biased responses.

Moreover, without transparency into retrieval, users cannot distinguish substantive balance from stylistic tendencies induced by RLHF (Sharma et al., 2023). Even when diverse opinions are retrieved, users cannot understand how conflicting perspectives were weighted or what constitutes "balanced representation". This is different from factual retrieval where citation suffices, as opinion synthesis requires understanding how conflicting viewpoints are reconciled.

3 Epistemic vs. Aleatoric: A Theoretical Foundation

We formalize the distinction between factual and opinion-aware retrieval through uncertainty quantification. Throughout, we use: q (query), Y (the response, with realization y), $\mathcal{D}$ (corpus), θ (latent state), $P_{\text{pop}}(\theta)$ (population opinion distribution), $\mathcal{R}(q)$ (retrieved documents), $\phi : \mathcal{D} \to \mathbb{R}^m$ (opinion embedding), $H(\cdot)$ (entropy), and $\mathcal{W}_p$ (p-Wasserstein distance).

3.1 Two Kinds of Uncertainty

The posterior predictive distribution for a query q given corpus $\mathcal{D}$ is:

$$P(y|q,\mathcal{D}) = \int P(y|\theta,q)\, P(\theta|\mathcal{D})\, d\theta \tag{1}$$

For **factual queries**, a unique true state θ^* exists. Uncertainty is *epistemic* and reducible through evidence. As corpus size grows, the posterior collapses to a point:

$$\lim_{|\mathcal{D}|\to\infty} P(\theta|\mathcal{D}) = \delta(\theta - \theta^*) \tag{2}$$

For **opinion queries**, uncertainty is *aleatoric* and reflects genuine heterogeneity rather than ignorance. The posterior converges to the population distribution, not a point:

$$\lim_{|\mathcal{D}|\to\infty} P(\theta|\mathcal{D}) = P_{\text{pop}}(\theta) \tag{3}$$

This distinction has a direct design implication: *factual RAG should minimize posterior entropy; opinion-aware RAG must preserve it.* We note that opinion queries still involve epistemic uncertainty about P_{pop} itself. However, the goal is to reduce this epistemic component through comprehensive retrieval, while preserving the aleatoric component by representing discovered diversity.

3.2 Divergent Optimization Objectives

Standard RAG maximizes information gain by minimizing conditional entropy:

$$\mathcal{L}_{\text{factual}} = H(Y|q) - H(Y|q,\mathcal{R}(q)) \tag{4}$$

This rewards retrieval that eliminates uncertainty which works correctly for facts and is counterproductive for opinions.

For opinions, we seek *distributional fidelity*. Let v denote a position in opinion space and $P_{\text{pop}}(v|q)$ the population opinion distribution. The opinion-aware objective minimizes divergence while penalizing artificial consensus:

$$\mathcal{L}_{\text{OA}} = \mathcal{W}_2\big(P_{\text{pop}}(v|q),\ P_{\text{model}}(v|q)\big) - \lambda\, H(P_{\text{model}}) \tag{5}$$

3.3 Retrieval as Coverage Optimization

Let $\mathcal{R}(q) = \{d_1, \ldots, d_k\}$. The embedding $\phi : \mathcal{D} \to \mathbb{R}^m$ maps documents to opinion space; in our implementation (Section 4), ϕ encodes sentiment polarity, stance intensity, and topical framing. The empirical opinion distribution from retrieval is:

$$\hat{P}_{\mathcal{R}} = \frac{1}{k}\sum_{i=1}^{k} \delta_{\phi(d_i)} \tag{6}$$

We measure coverage using the p-Wasserstein distance, where $\Gamma(\cdot,\cdot)$ denotes the set of couplings between the two distributions:

$$\mathcal{W}_p(\hat{P}_{\mathcal{R}}, P_{\text{pop}}) = \left(\inf_{\gamma\in\Gamma}\int \|\phi - \phi'\|^p\, d\gamma\right)^{1/p} \tag{7}$$

Intuitively, $\mathcal{W}_p$ measures the minimum cost to reshape one distribution into another, naturally penalizing retrieval that misses regions of opinion space. We prefer it over KL divergence or Jensen-Shannon divergence because it remains well-defined when distributions have non-overlapping support.

3.4 Unified Objective

Opinion-Aware RAG minimizes:

$$\mathcal{L} = \underbrace{\mathcal{W}_2(P_{\text{pop}}, \hat{P}_{\mathcal{R}})}_{\text{coverage}} + \alpha\, \underbrace{\mathcal{W}_2(P_{\text{pop}}, P_{\text{reader}})}_{\text{fidelity}} + \beta\, \underbrace{\max_{g\in G} \Delta_g}_{\text{fairness}} \tag{8}$$

where $P_{\text{reader}}(\phi|y)$ is the opinion distribution a reader infers from response y, and Δ_g =

$\mathcal{W}_2(P_{\text{pop}}(\phi|g), \hat{P}_{\mathcal{R}}(\phi|g))$ measures representation disparity for demographic group $g \in G$. The minimax fairness criterion ensures no subpopulation is excluded ($\alpha, \beta > 0$).

Scope and limitations. This framework identifies *what* opinion-aware RAG should optimize - coverage, fidelity, and fairness. However, this is not a directly implementable loss function. Key assumptions include: (1) P_{pop} is estimated from the corpus, treating it as a population sample; (2) opinions are treated as static (temporal dynamics are out of scope for this work); (3) directly optimizing Equation 8 is computationally prohibitive with current retrieval architectures. Our implementation (Section 4) approximates coverage through opinion-enriched indexing, and we validate the connection empirically via Wasserstein distance measurements in Section 5.

This formulation frames opinion-aware RAG as a *distribution estimation and communication problem*. This framing differs substantially from the point-estimation paradigm underlying factual retrieval, and one that we believe opens productive directions for future work.

4 O-RAG: A Practical Architecture

We present Opinion-Aware RAG (O-RAG) as one possible instantiation of the theoretical framework. We do not claim this to be the definitive solution, but as evidence that even simple opinion-aware enrichment yields measurable improvements. The architecture introduces an **opinion enrichment step** before indexing (illustrated in Figure 1), consisting of three domain-agnostic components.

Entity registry. A hierarchical taxonomy of opinion targets organized by granularity. Entities represent the topics people hold views about. The registry can be bootstrapped automatically via LLM extraction on sampled documents, followed by clustering and deduplication. Manual curation helps refine but is not strictly required.

Opinion attributes. Structured metadata capturing the nature of expressed opinions: sentiment and intensity, perceived impact or severity, emotional markers, and evidence type (personal experience, anecdotal, or data-driven).

Author attributes. Contextual factors about opinion holders that enable fairness analysis. These can be demographic or experiential characteristics relevant to the domain, supporting the fairness term $\max_{g \in G} \Delta_g$ in the unified objective (Equation 8).

Per-entity document splitting. Each source document is split into multiple enriched documents (one per mentioned entity) with opinion and author attributes embedded alongside content. This ensures that when a document discusses multiple topics with different sentiments, each opinion is indexed separately with its corresponding metadata.

4.1 Instantiation

We instantiate this framework in two domains:

E-commerce seller forums. Sellers discuss policies, fees, pain points, and operational challenges. We construct a 98-entity registry across three tiers, extract four opinion attributes (sentiment, intensity, business impact, emotional markers) and three author attributes (revenue band, tenure, region) per entity mention using an LLM with structured output (see Appendix G for the full extraction prompt).

Public hotel reviews (OpinioBank). Consumer reviews of hotel properties, drawn from the publicly available OpinioBank dataset introduced by Nayeem and Rafiei (2025). The extraction prompt required only minor domain adaptation; entity registry construction stabilized at approximately 90 entities. This demonstrates cross-domain feasibility with minimal manual effort.

Both configurations use vector search with Amazon Titan Text Embeddings V2 (1024 dimensions) and fixed-size chunking (4000 tokens, 15% overlap). We compare an Opinion-Enriched KB against a Raw KB (traditional RAG with no opinion metadata) using identical source content. Full model, decoding, retrieval, and cost settings are reported in Appendix A.

5 Empirical Evidence Across Domains

These experiments are not meant to prove O-RAG is the optimal solution, but to demonstrate that (a) the problem is real and measurable, and (b) even simple opinion-aware enrichment yields significant improvements. We evaluate across two domains spanning 19 diverse entities and 105 manually curated questions.

5.1 Setup

Datasets. (1) *E-commerce seller forums*: six months of discussions, approximately 8,000 unique sellers, 10K+ documents. (2) *OpinioBank hotel*

reviews (Nayeem and Rafiei, 2025): we use a sample of three hotel properties (Silverton Las Vegas, Ambassador Bangkok, Ala Moana Honolulu) from the full 500-entity OpinioBank release, totaling 6,256 reviews across 15 entities and 57 questions.

Evaluation protocol. We target entities with genuine opinion diversity by computing sentiment entropy $H(e)$ and selecting those with $H(e) \geq 0.6$ ensuring both positive and negative sentiments are represented (4 entities for forums, 15 for OpinioBank). For each entity, we generate natural language questions across multiple types designed to probe different facets of opinion. This yielded 48 questions for forums and 57 for OpinioBank. Each question is submitted to both KB configurations. Then, we retrieve the top-k documents and generate responses using the retrieved context. Retrieved documents are matched back to opinion metadata to compute diversity metrics.

Metrics. We design proxy metrics aligned with the theoretical objectives (Equation 8).

For *coverage*: (1) **Sentiment-Intensity** σ: combines sentiment polarity and intensity into a single numeric score (Positive+High = +30, Positive+Medium = +20, Positive+Low = +10, Neutral = 0, Negative+Low = -10, Negative+Medium = -20, Negative+High = -30); we report the standard deviation across retrieved documents, where higher σ indicates broader coverage of the sentiment-intensity space. (2) **Business Impact** σ: maps impact levels to numeric values (High = 30, Medium = 20, Low = 10); higher σ indicates retrieval of both high-impact and low-impact perspectives. (3) **Entity Match Rate**: fraction of retrieved documents that mention the queried entity, measuring retrieval precision for opinion-relevant content. (4) **Emotional Markers**: count of distinct emotional markers (e.g., Frustration, Hope, Resignation) across retrieved documents. (5) **Empirical** $\mathcal{W}_1$: Wasserstein distance between corpus-level and retrieved sentiment distributions.

For *fidelity*: human evaluation via blind pairwise comparison on the forums dataset (5 annotators, 48 pairs).

For *fairness*: **Demographic Coverage**, the fraction of categories represented in retrieved documents for each author attribute: Revenue Band (business size), Tenure (experience levels), and Evidence Type (personal, anecdotal, data-driven).

Table 2: Diversity metrics at k=20 (hybrid retrieval, forums dataset). Statistical significance via Wilcoxon signed-rank test (n=48): **bold** = $p<0.05$, * = $p<0.001$.

Metric	Raw KB	Enr. KB	Δ
Entity Match (%)	24.8	35.4	**+42.7%***
Sent-Int σ	9.12	11.56	**+26.8%***
Biz Impact σ	10.55	12.08	**+14.5%**
Emot. Markers	7.2	8.9	**+23.6%***
Revenue Band (%)	78.3	68.1	**−13.0%***
Tenure (%)	81.2	75.6	**−6.9%**
Evidence (%)	44.3	46.4	+4.7%

Table 3: Empirical $\mathcal{W}_1$ between corpus-level and retrieved sentiment distributions (k=20). Lower is better. Forums results significant at k=10, 40, 60 ($p<0.05$).

Dataset	Raw $\mathcal{W}_1$	Enriched $\mathcal{W}_1$	Δ
Forums	8.87	7.24	−18%
Silverton	15.31	9.17	−40%
Ambassador	19.15	14.54	−24%
Ala Moana	12.88	6.75	−48%

5.2 Results

Coverage. Table 2 presents the core comparison at k=20 with hybrid retrieval on the forums dataset. Opinion enrichment significantly improves all coverage metrics: +42.7% entity match rate, +26.8% sentiment-intensity σ, +14.5% business impact σ, and +23.6% emotional markers (all $p<0.05$).

To directly validate the connection to our theoretical framework, we compute empirical $\mathcal{W}_1$ between corpus-level and retrieved sentiment distributions (Table 3). The enriched KB is consistently 18-48% closer to corpus-level sentiment truth across both domains.

Fidelity. Human annotators confirmed that opinion-enriched retrieval produces responses that better communicate opinion diversity. Of 48 question-response pairs, annotators preferred Opinion-Enriched KB responses 79.2% of the time (95% CI: [65.0%, 89.5%], $p<0.001$). Appendix D has more details about the annotation study. Appendix F provides a qualitative example showing how enriched responses capture sentiment distributions and constructive suggestions missing from baseline responses.

Fairness. Contrary to expectations, opinion enrichment reduces overall demographic coverage: Revenue Band by 13.0% and tenure by 6.9% (both $p<0.05$). However, when comparing only documents that actually mention the queried entity,

Table 4: OpinioBank results (public hotel reviews). Entity match improvement and sentiment-intensity σ change vs. Raw KB.

Property	Reviews	Entities	Entity Match Δ	Sent-Int σ Δ
Silverton LV	2,297	4	+37.9pp	+5.04
Ambassador Bangkok	1,230	4	+37.2pp	+7.09
Ala Moana Honolulu	2,729	4	+18.9pp	+3.00

demographic coverage *improves* (Revenue Band +9.6%, Tenure +31.6%; see Appendix L). This reveals that the Raw KB's apparent demographic breadth comes from off-topic documents. It retrieves from a wider author pool, but much of that content is irrelevant to the query. The Enriched KB concentrates retrieval on entity-relevant content, achieving better demographic representation among documents that matter. This trade-off motivates future work on multi-objective retrieval.

Cross-domain generalization. Table 4 shows consistent improvements on public OpinioBank data across a different domain (hotel reviews vs. e-commerce forums). Diversity metrics were statistically significant for emotional markers; direction and magnitude are consistent with the forums findings. On OpinioBank, entity-matched documents show consistent demographic improvements (+12.5 to +29.4pp), confirming the fairness pattern observed on forums.

5.3 Ablations

Cross-model consistency. The Full O-RAG improvement over Raw KB is consistent across extraction models: Mistral Large (+24.2% Sent-Int σ), Qwen 235B (+27.4%), and Claude Sonnet 4.5 (+26.8%), confirming the approach is not dependent on a specific LLM (detailed in Appendix B).

Entity splitting vs. full O-RAG. An ablation disentangling per-entity splitting from opinion metadata reveals a nuanced picture (see Appendix E for full tables). Both Entity-Split-Only and Full O-RAG substantially improve over Raw KB across all metrics. For Mistral Large, Full O-RAG outperforms Entity-Split-Only on all diversity metrics ($p<0.001$): Sent-Int σ 8.11 vs. 7.37, Biz Impact 10.94 vs. 10.27, Emot Markers 4.1 vs. 3.5. For Sonnet 4.5, Entity-Split-Only achieves higher Sent-Int σ (13.1 vs. 11.56) and Biz Impact (12.92 vs. 12.08), but Full O-RAG wins on Emotional Markers (8.9 vs. 7.8, $p<0.05$). This suggests that entity splitting captures a large share of diversity gains through structural reorganization, while opinion metadata contributes additional value (particularly for emotional and affective dimensions) that varies by extraction model.

Question type variation. Business Impact (+38.9%) and Emotional (+34.1%) queries benefit most from opinion enrichment, while Experience queries (+6.2%) benefit least- likely because they already surface diverse personal perspectives without metadata (Appendix H).

Retrieval method. Hybrid retrieval amplifies enrichment benefits over semantic-only search (+26.8% vs. +21.4% Sent-Int σ), as keyword matching surfaces entity-relevant documents (Appendix I).

Retrieval depth. Sent-Int σ was consistently higher for opinion-enriched KB across varying retrieval depths (10-60) showing the performance is retrieval depth agnostic. (Appendix C)

5.4 Alternative Views and Counterarguments

One might argue that the LLM handles opinion synthesis at generation time: that RAG retrieval is merely context and a capable LLM can synthesize diverse opinions from whatever it receives. Our experiments suggest otherwise. Raw KB retrieval systematically excludes perspectives before the LLM sees them: at k=20, the Raw KB's sentiment distribution is 18-48% further from corpus truth than the Enriched KB's (Table 3). The LLM can only reason over what it receives. Human evaluators preferred enriched responses 79.2% of the time ($p<0.001$), indicating that retrieval-level bias propagates through generation despite the LLM's synthesis capabilities.

A second objection could be that existing diversity methods and pluralistic alignment already solve this. Diversity-aware retrieval methods like MMR (Carbonell and Goldstein, 1998) promote diversity in embedding space, and pluralistic alignment (Sorensen et al., 2024; Bakker et al., 2022) trains generation models to produce balanced outputs. As discussed in Section 2, these address dif-

ferent dimensions of the problem. MMR diversifies based on semantic dissimilarity- but opposing opinions on the same topic can be semantically *similar*, sharing vocabulary while differing in stance. MMR would penalize retrieving both. Pluralistic alignment operates at the generation layer and is complementary- a pluralistically-aligned LLM given biased retrieved context still produces biased responses.

A third objection is that opinion-aware RAG risks amplifying fringe or harmful viewpoints. However, the goal is *transparent representation proportional to prevalence*, not equal amplification. The distributional fidelity objective (Equation 8) weights opinions by corpus frequency, and the fairness term penalizes over- or under-representation of demographic groups. Opaque filtering carries its own risk: users cannot audit what was excluded, making bias invisible rather than absent.

6 A Research Agenda for Opinion-Aware RAG

We identify seven open problems for the community:

1. Opinion-aware benchmarks. The field needs RAG evaluation datasets with naturally occurring subjective content and annotations for stance, sentiment, demographic attributes, and perspective diversity. Without such benchmarks, progress on opinion-aware retrieval remains difficult to measure.

2. Learned opinion embeddings. Our approach relies on LLM-based extraction, but the theoretical framework's opinion embedding $\phi : \mathcal{D} \to \mathbb{R}^m$ suggests the need for learned representations that jointly encode semantic content with subjective attributes.

3. Direct distributional optimization. Our proxy metrics approximated the coverage objective $\mathcal{W}_2(P_{\text{pop}}, \hat{P}_{\mathcal{R}})$ but to effectively close the distributional gap we need retrieval strategies that directly minimize this during search - via query expansion, stance-diverse re-ranking, or submodular optimization.

4. Multi-objective retrieval. Our experiments reveal trade-offs between opinion diversity and demographic fairness. Retrieval strategies that jointly optimize the three terms in Equation 8 remain an open challenge.

5. Temporal opinion dynamics. The framework assumes a static P_{pop}, but opinions evolve. Sometimes, rapidly in response to events (e.g., a policy change triggering immediate backlash that later normalizes) and sometimes gradually as communities shift attitudes over months or years. Distinguishing temporary reactive opinions from lasting attitudinal shifts and detecting when P_{pop} has changed enough to invalidate cached representations require extending the theoretical foundation to handle non-stationary distributions. This is particularly important for domains like e-commerce and social media where opinion landscapes can shift within days.

6. Generation fidelity. Our evaluation focuses on retrieval diversity. Measuring whether generated responses faithfully communicate retrieved opinion distributions to readers, the fidelity term $\mathcal{W}_2(P_{\text{pop}}, P_{\text{reader}})$, requires new evaluation protocols beyond pairwise preference.

7. Joint retrieval-generation optimization. Combining opinion-aware retrieval with pluralistic alignment at the generation layer could yield systems that are both representative in what they retrieve and faithful in what they communicate.

7 Conclusion

We have argued that RAG systems exhibit a factual bias which is embedded throughout the generation pipeline from retrieval objectives to evaluation metrics, rendering misalignment with opinion-rich content. Our theoretical framework formalizes this through the epistemic–aleatoric distinction. We propose that factual queries should minimize posterior entropy, while opinion queries must preserve it. The unified objective over coverage, fidelity, and fairness provides a principled foundation for future work.

As an existence proof, Opinion-Aware RAG demonstrates that even simple opinion enrichment yields 18-48% reduction in distributional divergence from corpus truth across two domains. Human evaluators confirm the perceptible quality improvements transfers to generated responses.

The path forward requires the community to treat opinion-aware RAG as a first-class research problem: building benchmarks with subjective annotations, learning opinion-aware representations, developing retrieval and generation strategies that jointly optimize for diversity, and establishing evaluation protocols for opinion synthesis. As AI systems increasingly mediate access to information, we argue that their ability to represent that diversity in the generated responses is critical.

Limitations

Our work has several limitations that also frame directions for future research.

Theoretical framework is not directly implementable. The unified objective in Equation 8 identifies *what* opinion-aware RAG should optimize (coverage, fidelity, and fairness), but it is not a directly optimizable loss under current retrieval architectures. It relies on three assumptions: (1) the population opinion distribution P_{pop} is estimated from the corpus and treated as a population sample, which may not hold when the corpus itself is a biased sample of the underlying population; (2) opinions are treated as static, so temporal dynamics are out of scope; and (3) directly minimizing the objective is computationally prohibitive. Our implementation approximates only the coverage term through opinion-enriched indexing.

Limited domain and scale of evaluation. We evaluate on two English-language domains (e-commerce seller forums and hotel reviews) with 105 curated questions and 19 high-entropy entities. This is sufficient as an existence proof but does not establish generality across languages, cultures, or higher-stakes domains such as healthcare or political discourse. Entity selection was restricted to those with sentiment entropy $H(e) \geq 0.6$, which deliberately targets opinion-diverse content and may overstate gains relative to a uniformly sampled query set.

Human evaluation scope. The fidelity evaluation rests on blind pairwise preference from 5 annotators over 48 question-response pairs on a single domain. The wide confidence interval (95% CI: [65.0%, 89.5%]) reflects this limited scale, and we did not measure whether readers correctly *infer* the underlying opinion distribution (the fidelity term $\mathcal{W}_2(P_{\text{pop}}, P_{\text{reader}})$).

Dependence on LLM-based extraction. Opinion attributes and the entity registry are produced by LLM extraction. While we show cross-model consistency across three extraction models, the extraction step can itself introduce a systematic bias (for example, in how sentiment intensity or demographic attributes are assigned), and we do not fully audit these downstream effects.

Diversity-fairness trade-off. Opinion enrichment reduces overall demographic coverage even as it improves coverage among entity-relevant documents. We characterize this trade-off empirically but do not resolve it; multi-objective retrieval that jointly optimizes diversity and fairness remains open.

Ethics Statement

This work is motivated by an ethical concern: RAG systems that collapse opinion-rich content to a single answer can lead to echo chamber effects that amplify dominant viewpoints, which can lead to opinion manipulation and under-representation of minority voices. Our aim is transparent representation in proportion to prevalence rather than equal dissemination of all views. We acknowledge a dual-use tension. A system that faithfully represents an opinion distribution could also be used to surface fringe or harmful viewpoints. We argue that the distributional-fidelity and fairness terms in our objective mitigate this by weighting opinions according to corpus prevalence. The objective penalizes over or under-representation of demographic groups and that transparent representation is preferable to opaque filtering that hides, what was excluded. We do not claim this fully resolves the risk, and deployment in high-stakes settings would require additional safeguards.

Human annotation. All annotation was performed by internal colleagues (middle-aged, US-based, fluent English speakers) as part of their normal work; no additional compensation was provided for the task. Five evaluators completed the blind pairwise preference comparisons, and three annotators produced the opinion labels used to measure inter-annotator agreement. The annotators consented to the use of their judgments for this research. The study collected no personal data beyond the annotators' preference and label judgments over publicly derived text and was determined not to require ethics-board review. Annotators reviewed user-generated content that may contain complaints or negative sentiment but no sensitive personal information beyond what the source platforms already make available. No demographic inferences about individuals were released.

Inferred author attributes. The architecture derives author attributes (for example, revenue band, tenure, region) to support fairness analysis. These are coarse, corpus-derived categories used only in aggregate for representation measurement, not to

profile or target individuals. Any real deployment should treat such attributes with care, obtain appropriate consent, and avoid individual-level inference.

Data Availability

The OpinioBank hotel-review data used in our cross-domain evaluation is publicly available under the CC BY-NC-SA 4.0 license and was introduced by Nayeem and Rafiei (2025)[1]; our use is for non-commercial research evaluation, consistent with the license. The e-commerce seller-forum corpus is proprietary and contains user-generated content that cannot be released for privacy and contractual reasons; we report aggregate statistics and anonymize all platform-specific terms and quoted content. All data is in English, and no additional preprocessing was applied beyond the chunking described in Appendix A. Opinion-extraction prompts and metric definitions are provided in the appendix to support reproduction on comparable corpora.

Use of AI Assistants

We disclose the following use of AI assistants in preparing this work, separate from the LLMs studied as objects of research. Anthropic Claude Opus 4.6 assisted with writing the experiment code, Google Gemini supported the literature survey, and Claude Sonnet 4.5 provided grammatical & fluency assistance for the manuscript. All research claims, analyses, and final text were written and reviewed by the authors.

References

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, et al. 2022. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544.

Andrew Brown, Muhammad Roman, and Barry Devereux. 2025. A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. *arXiv preprint arXiv:2508.06401*.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336.

Arie Cattan, Alon Jacovi, Ori Ram, Jonathan Herzig, Roee Aharoni, Sasha Goldshtein, Eran Ofek, Idan Szpektor, and Avi Caciularu. 2025. Dragged into conflicts: Detecting and addressing conflicting sources in search-augmented llms. *arXiv preprint arXiv:2506.08500*.

Hung-Ting Chen and Eunsol Choi. 2025. Open-world evaluation for retrieving diverse perspectives. *Preprint*, arXiv:2409.18110.

Hung-Ting Chen, Michael Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2292–2307.

Gaurab Chhetri, Anandi Dutta, and Subasish Das. 2025. Cognitivesky: Scalable sentiment and narrative analysis for decentralized social media. *arXiv preprint arXiv:2509.11444*.

Eunseong Choi, June Park, Hyeri Lee, and Jongwuk Lee. 2025. Conflict-aware soft prompting for retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26981–26995.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610.

Matheus Vinicius da Silva de Oliveira, Jonathan de Andrade Silva, and Awdren de Lima Fontao. 2025. Fairness testing in retrieval-augmented generation: How small perturbations reveal bias in small language models. *arXiv preprint arXiv:2509.26584*.

[1] https://huggingface.co/datasets/tafseer-nayeem/OpinioBank

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Robert Friel, Masha Belyi, and Atindriyo Sanyal. 2024. Ragbench: Explainable benchmark for retrieval-augmented generation systems. *arXiv preprint arXiv:2407.11005*.

Aoran Gan, Hao Yu, Kai Zhang, Qi Liu, Wenyu Yan, Zhenya Huang, Shiwei Tong, and Guoping Hu. 2025. Retrieval augmented generation evaluation in the era of large language models: A comprehensive survey. *arXiv preprint arXiv:2504.14891*.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023a. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023b. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Vignesh Gokul, Srikanth Tenneti, and Alwarappan Nakkiran. 2025. Contradiction detection in rag systems: Evaluating llms as context validators for improved information consistency. *arXiv preprint arXiv:2504.00180*.

Haoze Guo and Ziqi Wei. 2026. Hidden-in-plain-text: A benchmark for social-web indirect prompt injection in rag. *arXiv preprint arXiv:2601.10923*.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7740–7754.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Jungyeon Lee, Kangmin Lee, and Taeuk Kim. 2025. Magic: A multi-hop and graph-based benchmark for inter-context conflicts in retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8783–8803.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Hao Li, Viktor Schlegel, Yizheng Sun, Riza Batista-Navarro, and Goran Nenadic. 2025. Large language models in argument mining: A survey. *arXiv preprint arXiv:2506.16383*.

Tao Li, Lei Lin, Minsoo Choi, Kaiming Fu, Siyuan Gong, and Jian Wang. 2018. Youtube av 50k: An annotated corpus for comments in autonomous vehicles. In *Proceedings of the Thirteenth International Joint Symposium on Artificial Intelligence and Natural Language Processing*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating the effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. Covid-qa: A question answering dataset for covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Mir Tafseer Nayeem and Davood Rafiei. 2025. OpinioRAG: Towards generating user-centric opinion highlights from large-scale online reviews. In *Second Conference on Language Modeling*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches (CoCo@NIPS)*.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, and Samuel R. Bowman. 2022. Quality: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, et al. 2021. 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*, pages 2523–2544.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Bjorn Gamback. 2020. Semeval-2020 task 8: Memotion analysis – the visuolingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Taylor Sorensen, Liwei Jiang, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. 2024. Value kaleidoscope: Engaging AI with pluralistic human values, rights, and duties. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18):19937–19947.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. Asqa: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.

Yifan Sun, Hirofumi Tsuruta, Masaya Kumagai, and Ken Kurosaki. 2024. Topic modeling and sentiment analysis on japanese online media's coverage of nuclear energy. *arXiv preprint arXiv:2411.18383*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4149–4158.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 404–430, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Nandan Thakur, Luiz Bonifacio, Xinyu Zhang, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Boxing Chen, Mehdi Rezagholizadeh, and Jimmy Lin. 2024. "knowing when you don't know": A multilingual relevance assessment dataset for robust retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

Linlin Wang, Tianqing Zhu, Laiqiao Qin, Longxiang Gao, and Wanlei Zhou. 2025. Bias amplification in rag: Poisoning knowledge retrieval to steer llms. *arXiv preprint arXiv:2506.11415*.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. 2024. Cmb: A comprehensive medical benchmark in chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Sarah Weißmann, Aaron Philipp, Roland Verwiebe, Chiara Osorio Krauter, Nina-Sophie Fritsch, and Claudia Buder. 2025. Clicks, comments, consequences: Are content creators' socio-structural and platform characteristics shaping the exposure to negative sentiment, offensive language, and hate speech on youtube? *arXiv preprint arXiv:2504.07676*.

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, et al. 2024. Crag – comprehensive rag benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. 2023. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4615–4635.

Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. 2024. Retrieval augmented fact verification by synthesizing contrastive arguments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10331–10343.

Linda Zeng, Rithwik Gupta, Divij Motwani, Yi Zhang, and Diji Yang. 2025. Worse than zero-shot? a fact-checking dataset for evaluating the robustness of rag against misleading retrievals. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025a. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21863–21882.

Tianhui Zhang, Yi Zhou, and Danushka Bollegala. 2025b. Evaluating the effect of retrieval augmentation on social biases. *arXiv preprint arXiv:2502.17611*.

Yao Zhang, Zaixi Shang, Silpan Patel, and Mikel Zuniga. 2025c. From unstructured communication to intelligent rag: Multi-agent automation for supply chain knowledge bases. In *Proceedings of the 1st Workshop on AI for Supply Chain: Today and Future @ KDD 2025*.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024a. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473*.

Xinran Zhao, Tong Chen, Sihao Chen, Hongming Zhang, and Tongshuang Wu. 2024b. Beyond relevance: Evaluate and improve retrievers on perspective awareness. *Preprint*, arXiv:2405.02714.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. Qmsum: A new benchmark for query-based multi-domain meeting summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921.

A Reproducibility and Experimental Settings

Models. Opinion extraction used three LLMs accessed through Amazon Bedrock in April 2026: Mistral Large, Qwen 235B, and Claude Sonnet 4.5. Response generation from the retrieved context used Anthropic Claude Sonnet 4.5. Document and query embeddings used Amazon Titan Text Embeddings V2 (1024 dimensions).

Decoding. All extraction and generation calls used temperature 0, top-p at the Amazon Bedrock default, and the Bedrock default maximum output length of 4096 tokens. Because temperature is 0, outputs are effectively deterministic given fixed inputs, so all reported numbers come from a single run.

Retrieval. Documents were indexed in an Amazon Bedrock Knowledge Base vector store using fixed-size chunking (4000 tokens, 15% overlap). We report both semantic-only and hybrid (dense plus keyword) retrieval across depths $k \in \{10, 20, 40, 60\}$.

OpinioBank sampling. We evaluate on a 3-property sample (Silverton Las Vegas, Ambassador Bangkok, Ala Moana Honolulu) drawn from the full 500-entity OpinioBank release. Entity and question counts reported in Section 5 refer to this sample, not the full dataset.

Compute and cost. All inference was API-hosted through Amazon Bedrock; the method involves no model training and no local GPU compute. Total inference cost across all extraction, generation, and ablation runs was approximately US$2,500.

Statistical testing. Significance is assessed with the Wilcoxon signed-rank test over paired questions; the confidence interval for human preference uses bootstrap resampling. Both were computed with the SciPy library (`scipy.stats`). Significance thresholds are stated in the relevant table captions.

B Opinion Extraction Model Variations

Table 5 compares retrieval diversity when using different LLMs for opinion extraction. All configurations use hybrid search at k=20. Results show consistent improvements across models.

Table 5: Retrieval diversity improvements (Δ vs. Raw KB) across opinion extraction LLMs (k=20, hybrid).

Metric	Mistral	Qwen	Sonnet
Sent-Int σ	**+24.2%**	**+27.4%**	**+26.8%***
Biz Impact σ	**+18.4%**	**+12.4%**	**+14.5%**
Revenue Band Cov.	−17.5%	−19.0%	**−13.0%***
Tenure Cov.	−3.9%	−3.6%	**−6.9%**
Evidence Cov.	**+9.4%**	**+8.9%**	+4.7%
Emot. Markers	**+28.1%**	**+20.0%**	**+23.6%***

C Retrieval Depth Analysis

Table 6 shows coverage and fairness metrics across retrieval depths.

D Annotation Study Details

The manual annotation comprised 129 forum posts labeled by 3 independent annotators. Raw agreement: 80.6%. Gwet's AC1 = 0.845 (95% CI [0.8, 0.9]). Fleiss' κ = 0.23, which is low due to the kappa paradox under high prevalence (91% positive votes); AC1 is the more appropriate metric under prevalence imbalance.

Annotators for both the labeling task and the blind pairwise preference evaluation were internal colleagues who participated as part of their normal work and consented to the use of their judgments; recruitment, compensation, consent, and ethics-review details are described in the Ethics Statement.

Pairwise preference instructions. For the blind preference evaluation, each annotator was shown the two responses to a given question with their positions randomly assigned to slots A and B (so that neither the Raw KB nor the Opinion-Enriched KB was consistently placed). Annotators were told which system produced which response was hidden, and were asked a single question: which response, A or B, is preferred based on which better reflects a diverse and holistic range of opinions. No other scoring dimensions were collected.

E Entity-Split Ablation Details

Tables 7 and 8 disentangle per-entity splitting from opinion metadata enrichment across two extraction models.

Table 7: Entity-Split-Only ablation at k=20 (Sonnet 4.5). **Bold** = highest per metric.

KB Config	Sent-Int σ	Biz Impact	Emot Marker
Raw	9.12	10.55	7.2
Entity-Split Only	**13.1**	**12.92**	7.8
Full O-RAG	11.56	12.08	**8.9***

Table 8: Entity-Split-Only ablation at k=20 (Mistral Large). **Bold** = highest per metric, * = $p<0.001$ vs. Entity-Split.

KB Config	Sent-Int σ	Biz Impact	Emot Marker
Raw	6.53	9.24	3.2
Entity-Split Only	7.37	10.27	3.5
Full O-RAG	**8.11***	**10.94***	**4.1***

F Qualitative Response Comparison

The following comparison illustrates how opinion enrichment affects generated responses. Key phrases are highlighted in **red**.

Query: *"Give me a detailed overview of what sellers think about community management & moderators specifically on seller forums?"*

Table 9 contrasts the Raw KB response against the Opinion-Enriched KB response.

G Opinion Extraction Prompt

We use an LLM with a structured output schema for opinion extraction. The prompt template is shown below (with platform-specific terms anonymized):

```
You are an expert at analyzing
seller opinions from e-commerce platform
content. Analyze the following post and
extract structured opinion information
```

Table 6: Diversity metrics at different retrieval depths (k) using hybrid retrieval. **Bold** = $p<0.05$, * = $p<0.001$.

		Entity	Scored Diversity (σ)		Categorical Coverage (%)			Emot.
k	KB Config	Match %	Sent-Int	Biz Impact	Revenue Band	Tenure	Evidence	Marker
10	Raw	32.7	10.10	11.48	63.7	65.7	38.0	5.5
	Enriched	48.5	13.51	12.73	51.8	59.2	40.1	7.0
20	Raw	24.8	9.12	10.55	78.3	81.2	44.3	7.2
	Enriched	35.4	11.56	12.08	68.1	75.6	46.4	8.9
40	Raw	16.9	8.11	9.45	87.8	89.3	49.5	9.1
	Enriched	24.6	9.68	10.88	83.6	85.1	53.1	11.0
60	Raw	13.3	7.28	8.52	90.5	89.6	50.5	10.1
	Enriched	19.0	8.52	9.78	89.0	87.5	53.6	12.2

Raw Discussions KB Response	**Opinion-Enriched KB Response**
Based on seller discussions in the forums, **there are several perspectives** on community management and moderators: **Requests for Moderator Intervention:** Sellers frequently turn to moderators when they feel stuck in **automated response loops** or cannot get resolution through standard support channels. Multiple sellers specifically call out moderators in their posts, asking for help with account issues, policy violations, and reimbursement disputes. **Frustration with Support Systems:** There's significant frustration with the overall support system that leads sellers to seek moderator help. Sellers report dealing with what they describe as a "broken system." **Moderator Responsiveness:** Sellers appear to view moderators as a potential escalation path when standard channels fail. However, **the search results don't show clear feedback** on whether moderators consistently respond or resolve these escalated issues. **Community Guidelines Enforcement:** At least one seller reported frustration with the moderation process for customer reviews, noting they had reported guideline violations but received no response after 17 days.	Sellers have **mixed but generally positive** views of community moderators, often seeing them as a **critical escalation path** when standard support channels fail. Multiple sellers report that **moderators are the only ones who have successfully helped** them resolve issues that automated systems could not address. Sellers turn to moderators when trapped in what they describe as **"horrible loops"** of unhelpful automated responses. However, sellers **express frustration** with community management's responsiveness to guideline violations. One seller reported waiting 17 days without response. The reviews **continued to negatively affect** their listings and sales. Sellers also **request improvements** to forum functionality. One seller suggested individual user accounts because the current setup causes irrelevant notifications. This seller indicated they **"rarely ever post"** specifically because of this issue. When sellers reach out to moderators, it's typically **as a last resort** after exhausting other options. They view moderator intervention as **necessary for protecting** their accounts from false claims and removing malicious feedback.

Table 9: Side-by-side comparison. The Raw KB organizes by *action categories* and admits "search results don't show clear feedback" on outcomes. The Opinion-Enriched KB captures *sentiment distribution* (mixed but positive), *success attribution* (only ones who helped), *behavioral explanations* (rarely post, last resort), and *constructive suggestions* missing from Raw KB.

```
for EACH entity (business topic/issue)
mentioned.

POST CONTENT: {content}
PLATFORM: {platform_type}
CATEGORY: {platform_category}
{entity_registry}

Extract the following information and
respond ONLY with valid JSON:

{"entities": [{
  "entity_name": "entity name",
  "confidence": "Low|Medium|High",
  "sentiment": "positive|negative|neutral|mixed",
  "sentiment_intensity": "Low|Medium|High",
  "emotional_markers": ["marker1", "marker2"],
  "evidence": "supporting evidence",
  "business_impact": "Low|Medium|High",
  "context_snippet": "relevant quote",
  "evidence_type": "personal|anecdotal|data"
}]}
```

Key extraction guidelines include:

- **Entity matching**: If the entity matches one from the hierarchical registry, use that exact name; prefer granular (Tier 3) over general (Tier 1) entities
- **Per-entity extraction**: Each entity receives its own complete opinion structure; a single post may yield multiple entities with different sentiments
- **Evidence typing**: Classify as personal (first-hand experience), anecdotal (heard from others), or data (cites specific statistics)
- **Emotional markers**: Common markers include Frustration, Anger, Resignation, Adaptation, Hope, Satisfaction, Confusion, Anxiety
- **Business impact**: Consider financial men-

tions, operational disruption, and account risks when assigning impact levels

The model outputs a JSON array with one object per entity-opinion pair, enabling downstream per-entity document splitting.

H Question Type Ablation

Table 10 shows how question formulation affects diversity gains.

Table 10: Sentiment-Intensity σ by question type (k=20). Largest gains for Business Impact and Emotional queries; Experience queries benefit least.

Question Type	Raw	Enriched	Δ
Detailed Overview	9.14	11.80	**+29.1%**
Experience	10.64	11.30	**+6.2%**
Business Impact	8.56	11.89	**+38.9%**
Specific Examples	9.02	11.14	**+23.5%**
Process	7.82	10.44	**+33.5%**
Emotional	9.56	12.82	**+34.1%**

I Retrieval Method Comparison

Table 11 compares semantic-only and hybrid retrieval. Hybrid search amplifies opinion enrichment benefits (+26.8%) over semantic-only (+21.4%), as keyword matching surfaces entity-relevant documents.

Table 11: Semantic-only vs. hybrid retrieval at k=20. **Bold** = $p<0.05$, * = $p<0.001$.

	Semantic		Hybrid	
Metric	Raw	Enr. (Δ)	Raw	Enr. (Δ)
Sent-Int σ	8.93	10.84 (**+21.4%***)	9.12	11.56 (**+26.8%***)
Biz Impact σ	10.47	11.63 (**+11.1%**)	10.55	12.08 (**+14.5%**)
Revenue Band (%)	79.1	68.7 (**−13.1%***)	78.3	68.1 (**−13.0%***)
Tenure (%)	81.5	76.8 (**−5.8%**)	81.2	75.6 (**−6.9%**)
Evidence (%)	44.3	43.8 (−1.1%)	44.3	46.4 (+4.7%)
Emot. Markers	7.1	8.4 (**+18.3%**)	7.2	8.9 (**+23.6%***)

J Emergent Opinion Synthesis Example

The following is a response from ChatGPT (GPT-4o, January 2026) to the query: *"What do small business owners think about the current tax policy?"*

> *Small business owners express mixed views on current tax policy. Many appreciate efforts toward permanence and certainty, particularly wanting key provisions from the 2017 Tax Cuts and Jobs Act made permanent to aid long-term planning. They support targeted deductions and pass-through relief, especially proposals allowing deduction of the first $25,000 of qualifying business income. However, concerns persist: many feel previous tax cuts primarily benefited larger corporations rather than small businesses. Common complaints include tax code complexity, high compliance costs, and perceived unfairness in how large firms can leverage specialists and loopholes. Small business owners consistently advocate for simplified reporting, scaled incentives that actually help smaller firms, and measures ensuring "fair play" against large corporation advantages.*

This response sounds balanced, but without transparency into retrieval, users cannot determine whether it reflects genuine diversity in retrieved sources. Or the model's learned tendency to present information evenhandedly regardless of input.

K Biased Leadership Response Example

The following illustrates how traditional RAG metrics can assign high scores to biased outputs. In response to the query *"What leadership qualities are most important for startup CEOs?"*, a system might generate:

> *Successful startup CEOs demonstrate strong analytical thinking, decisive action under pressure, and natural assertiveness in competitive markets. Research shows that effective leaders maintain composure during high-stakes negotiations, exhibit confidence in pitching to investors, and demonstrate the aggressive persistence needed to disrupt established industries. The most successful CEOs are typically those who embrace calculated risks, maintain unwavering conviction in their vision, and possess the competitive drive necessary to outperform rivals.*

While scoring excellently across standard metrics, this response amplifies gender bias by emphasizing traditionally masculine-coded traits while omitting empathetic or inclusive leadership qualities.

L Entity-Matched Document Analysis

Table 12 analyzes diversity metrics on only the subset of retrieved documents that mention the queried entity, explaining the diversity-fairness trade-off discussed in Section 5.

Table 12: Diversity metrics on entity-matched documents only (k=20, forums dataset). On entity-matched documents, Opinion-Enriched KB *improves* demographic coverage- the opposite of the overall result.

Metric	**Raw**	**Enr.**	**Δ**
Matched Docs (of 960)	238	340	**+42.9%***
Sent-Int σ	14.29	16.57	+16.0%
Biz Impact σ	4.45	5.47	**+22.9%**
Revenue Band Cov. (%)	42.9	47.0	**+9.6%**
Tenure Cov. (%)	40.5	53.3	**+31.6%***
Evidence Cov. (%)	44.3	46.4	+4.7%
Emot. Markers	7.2	8.9	**+23.6%***

M Extended Benchmark Analysis

Table 13 provides a comprehensive analysis of 36 major evaluation datasets used in RAG research from 2013–2025.

Dataset	Year	Task Category	Evaluation Focus	Opinion	Reference
Single-hop Question Answering					
WebQuestions	2013	Knowledge Base QA	EM, F1	No	(Berant et al., 2013)
MS MARCO	2016	Machine Reading Comprehension	ROUGE-L, BLEU	No	(Nguyen et al., 2016)
TriviaQA	2017	Reading Comprehension	EM, F1	No	(Joshi et al., 2017)
SQuAD 2.0	2018	Reading Comprehension	EM, F1	No	(Rajpurkar et al., 2018)
Natural Questions	2019	Open-domain QA	EM, Token F1	No	(Kwiatkowski et al., 2019)
PopQA	2023	Entity-centric QA	Accuracy	No	(Mallen et al., 2023)
Multi-hop Question Answering					
HotpotQA	2018	Multi-hop Reasoning	EM, F1, Support F1	No	(Yang et al., 2018)
2WikiMultiHopQA	2020	Multi-hop Reasoning	EM, F1	No	(Ho et al., 2020)
MuSiQue	2022	Multi-hop Reasoning	EM, F1	No	(Trivedi et al., 2022)
Long-form Question Answering					
NarrativeQA	2018	Long-form QA	ROUGE, BLEU	No	(Kocisky et al., 2018)
ELI5	2019	Long-form QA	ROUGE-L	No	(Fan et al., 2019)
QMSum	2021	Meeting Summarization	ROUGE	No	(Zhong et al., 2021)
ASQA	2022	Ambiguous QA	ROUGE, QA-based	No	(Stelmakh et al., 2022)
Domain-specific Question Answering					
COVID-QA	2020	Medical Domain	EM, F1	No	(Möller et al., 2020)
Qasper	2021	Scientific Papers	EM, F1	No	(Dasigi et al., 2021)
CMB	2024	Chinese Medical	Accuracy	No	(Wang et al., 2024)
Multi-choice Question Answering					
ARC	2018	Science Questions	Accuracy	No	(Clark et al., 2018)
CommonsenseQA	2019	Commonsense Reasoning	Accuracy	No	(Talmor et al., 2019)
MMLU	2021	World Knowledge	Accuracy	No	(Hendrycks et al., 2021)
QuALITY	2022	Reading Comprehension	Accuracy	No	(Pang et al., 2022)
Information Retrieval & Knowledge Tasks					
KILT	2021	Knowledge-Intensive Tasks	EM, F1, ROUGE-L	No	(Petroni et al., 2021)
BEIR	2021	Information Retrieval	nDCG, MAP, Recall	No	(Thakur et al., 2021)
Fact Verification & Robustness					
FEVER	2018	Fact Verification	Accuracy, FEVER Score	No	(Thorne et al., 2018)
PubHealth	2020	Health Fact-checking	Accuracy	No	(Kotonya and Toni, 2020)
TruthfulQA	2022	Truthfulness Evaluation	% True, % Informative	Partial	(Lin et al., 2022)
NoMIRACL	2024	Multilingual Robustness	Hallucination Rate, Error Rate	No	(Thakur et al., 2024)
CRAG	2024	Comprehensive QA	Accuracy, Hallucination	No	(Yang et al., 2024)
RAGuard	2025	Robustness Testing	Resistance to Misinfo	No	(Zeng et al., 2025)
CONFLICTS	2025	Conflict Resolution	Appropriateness Score	No	(Cattan et al., 2025)
Generation & Attribution					
ALCE	2023	Citation Generation	Citation Quality, Fluency	No	(Gao et al., 2023a)
FActScore	2023	Factual Precision	Atomic Fact Accuracy	No	(Min et al., 2023)
AttrScore	2023	Attribution Quality	Attribution Accuracy	No	(Yue et al., 2023)
RAGBench	2024	Industry RAG	TRACe Metrics	No	(Friel et al., 2024)
OpinioRAG	2025	Opinion Summarization	Sentiment Alignment	Yes	(Nayeem and Rafiei, 2025)
Perspective-Aware Retrieval					
PIR	2024	Perspective Retrieval	p-Recall, Bias	Partial	(Zhao et al., 2024b)
BERDS	2025	Diverse Perspective Retrieval	Perspective Coverage	Partial	(Chen and Choi, 2025)

Table 13: Complete analysis of 36 major RAG/LLM evaluation datasets (2013–2025). Only OpinioRAG addresses opinion synthesis; PIR and BERDS evaluate perspective-aware retrieval but not distributional fidelity.