Retrieval Sensitivity to Identity Signals in Queries

Andrew Tang*, Nicholas Deas*, Kathleen McKeown, Vishal Misra

Columbia University, Department of Computer Science

Correspondence: {a.tang, ndeas}@cs.columbia.edu

Abstract

Dense retrievers decide which documents reach users and the language models that use them, yet they are typically evaluated with neutral queries. We ask whether the identity signals that real users express in their queries—political ideology and dialect—bias what a retriever returns. We design evaluations in two domains, political news and consumer-health questions, each pairing a controlled synthetic set that varies only the identity signal with naturalistic queries. Across five dense retrievers and a sparse baseline, every retriever (i) retrieves articles that align with the query's own political lean and (ii) performs worse for questions written in African American Language (AAL) than in White Mainstream English (WME). Two analyses tie these gaps to queries' identity signals beyond surface vocabulary: partialling out an aggregate lexical-asymmetry score leaves the synthetic gaps largely intact, and linear probes recover lean and dialect from the retrievers' query embeddings beyond token-level features. Left unaddressed, such retrieval biases risk contributing to polarization and reinforcing the health disparities already faced by AAL speakers.[1]

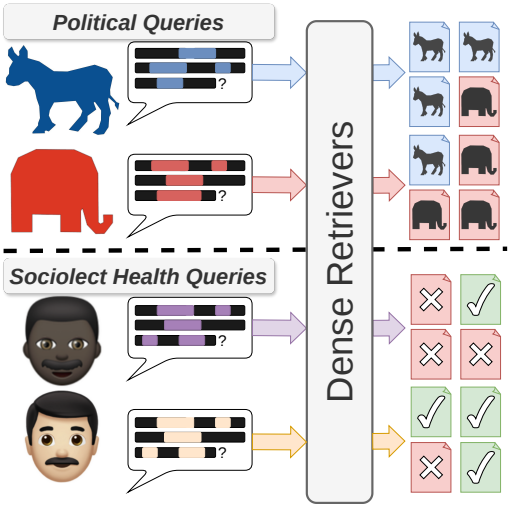

Figure 1: Paper overview. We evaluate retrievers' sensitivity to identity signals in queries, finding that retrievers favor documents aligning with the questioner's signaled political ideology (top) and that lower-quality documents are retrieved for AAL health queries (bottom).

1 Introduction

Dense retrievers (Karpukhin et al., 2020) underpin search (Huang et al., 2020; Zhao et al., 2024), recommendation (Yi et al., 2019; Fan et al., 2025), and retrieval-augmented generation (RAG) (Lewis et al., 2020; Fan et al., 2024), selecting documents a user or downstream model sees. A retriever that systematically skews *which* documents it surfaces is, therefore, consequential: in search and recommendation, the skewed results directly impact users, and in RAG, the skewed results are passed to a large language model (LLM), impacting its response. The latter is increasingly central as users shift toward AI-mediated information access (Li et al., 2025; Liang et al., 2025), where poor or skewed retrieval sharply degrades generated outputs (Chen et al., 2024).[2]

Prior work shows that retrieval can be biased in two broad ways: by spurious features such as the position and length of documents (Yao et al., 2025; Zeng et al., 2025; Cuconasu et al., 2025), and by social factors such as skewed gender representation among retrieved documents (Zhang et al., 2026; Parihar and Cheng, 2026; Krieg et al., 2023). Kim

*Denotes equal contribution.

[1]We make code available at https://github.com/Andrewtcr/bias-ret.

[2]Affected by generator robustness to imperfect context (Hsia et al., 2025).

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 60355–60381
October 24-29, 2026 ©2026 Association for Computational Linguistics

et al. (2025) frame the latter case as *bias conflict*: corpora, query embeddings, and generators each carry their own bias, but the system's overall bias is not simply the sum of that of its components. Therefore, the retriever's own contribution is worth isolating.

These studies, however, mostly examine relatively *neutral* queries (queries that do not signal any particular sociodemographic group) and assess biases among retrieved documents (e.g., over-representation of one political ideology over another). Real queries instead routinely carry cues of the author's sociodemographics and ideology through different proxies (e.g., names, emotional reactions or values, other cultural proxies, etc.; Tonneau et al. 2026; Adilazuarda et al. 2024). Similarity-based retrieval makes some sensitivity to these cues expected in principle. This expectation, however, predicts neither which retrievers are affected, nor by how much, nor through which channel (surface vocabulary or the representation itself), and it assumes that documents aligned with the asker's identity are what they need. We therefore study an underexplored question: how *identity signals in the query itself* lead a retriever to return systematically different documents.

We demonstrate this in two domains closely related to downstream harms. First, we examine signals of the authors' political ideology (conveyed through presupposed beliefs (Sieker et al., 2025), framing (Lee et al., 2022), or other means), where retrieval that aligns with the questioner's ideology may deepen polarization by reaffirming that ideology (§2). Second, we test whether retrievers return lower-quality documents for health questions written in African American Language (AAL) than for those in White Mainstream English (WME), representing a direct *allocational* harm[3] to AAL-speaking users (§3).[4]

We detail our primary contributions as follows: (1) We build **query-based identity-bias evaluations for retrievers** for two variables—political ideology and AAL/WME sociolect—each relying on both a *synthetic, paired* (i.e., topic-matched) query set that better isolates the identity signal, and a more *ecologically valid, unpaired* query set of real user queries that better represents real-world deployments and potential harms. (2) Evaluating **five dense retrievers** (and a sparse retriever baseline), we find that **every retriever prefers documents aligning with the questioner's political ideology** and **ranks the answer lower for AAL health queries than for WME**, consistent across the controlled and naturalistic settings. (3) Two analyses tie these biases to the query's identity signal beyond surface vocabulary: on the content-controlled synthetic contrasts, **partialling out aggregate lexical asymmetry leaves the gap largely intact**, and **linear probes recover lean/dialect from retriever query embeddings** beyond token-level features.

2 Political Bias

We first ask whether dense retrievers produce systematically different news articles when provided queries signaling left and right political leanings.

2.1 Data

We evaluate dense retrievers on two complementary datasets of political queries: paired left and right-leaning queries about the same subject (AllSides$_{synth}$) and unpaired naturalistic queries (Reddit$_{nat}$). Per-group query counts and length statistics are presented in App. A.

Corpus. We use the Qbias AllSides[5] corpus (Haak and Schaer, 2023) as the set of news articles to be retrieved. These articles are grouped into *editorial roundups*, which are sets of articles about the same news story, but from sources with differing political leaning. Each grouping and article political leaning are determined by AllSides' cross-partisan team of expert news specialists. For evaluation, we extract a subset of Qbias for which the editorial roundups contain an equal number of left and right-leaning documents ($\sim$17k articles total) to isolate retriever biases from corpus biases.

AllSides$_{synth}$ (paired queries). First, we consider a highly controlled setting of paired synthetic queries from 1,000 randomly sampled roundups. For each roundup we use `gpt-5-mini` to label the Boydstun framing dimensions (Card et al., 2015)

[3] We consider this an *allocational* harm because disparities in the performance of dense retrievers, particularly in the health domain, would provide inequitable access to informational health resources. See Blodgett et al. (2020) for further discussion of allocational harms.

[4] Prior work has used alternative terms, such as African American Vernacular English and Standard/Mainstream American English respectively. We follow recent work in using these terms to avoid implications of a "standard variety" (Baker-Bell, 2020; McKeown et al., 2026).

[5] https://www.allsides.com/unbiased-balanced-news

	Frames	Left	Right
AllSides$_{synth}$	Fairness & equality	*Is raising the federal minimum wage to $10.10 a **fair way to reduce inequality** and **lift nearly a million from poverty** per CBO?*	*Does the CBO report show raising the minimum wage to $10.10 **unfairly distribute job losses**, eliminating about 500,000 opportunities for workers?*
	Economic	*How would the CBO's $10.10 minimum wage proposal economically **benefit low income workers** and **reduce poverty** despite some job losses?*	*What are the economic **costs to employment and small businesses** of the CBO's $10.10 minimum wage proposal, including projected job losses?*
Reddit$_{nat}$	—	*Is anyone else **tired of ground-level conservatives** claiming they **"agree"** with news incidents while always supporting the system that created these injustices?*	***Illegal aliens** have always been subject to deportation. Why does the government now need the Alien Enemies Act to deport them?*

Table 1: Select example queries from each political query dataset. Potential signals of the questioner's political ideology are highlighted **red** and **blue**.

present in its center article. For each frame, we then prompt the model to write two short search queries grounded in that article, one left-leaning and one right-leaning. Conditioning both queries on a shared frame is what makes the pair controlled: without it, a left- and a right-leaning asker may gravitate to different frames of the same story. In Table 1, a minimum-wage topic could draw a left asker toward the *fairness & equality* frame (reducing inequality, lifting people from poverty) and a right asker toward the *economic* angle (costs to employment and small businesses)—the diagonal of the table. We generate every frame for both leans and measure the left/right contrast *within* a frame.[6] This process results in 3,723 left and 3,723 right-leaning queries that share a common topic and frame with minimal differences (median length 138 and 139 characters, respectively). Prompts, the full Boydstun frame typology, and example frame-labeled articles are included in App. C.

Reddit$_{nat}$ (unpaired queries). While the synthetic queries are paired and allow for controlled experiments, we additionally collect naturally occurring queries to better approximate realistic use cases. We collect 3,340 posts from the `r/AskALiberal`[7] (1,979) and `r/AskConservatives`[8] (1,361) subreddits. These subreddits host community questions targeted toward liberal or conservative answerers, although the posters themselves may span the political spectrum. Following prior work (Törnberg, 2025), we use LLMs to classify the ideology of a question's writer. Specifically, we pass the body of each post, which contains additional elaboration on each question, to two separate LLM classifiers, `gpt-5.4-mini` and `Qwen3.5-4B`, and keep only those where both agree. This yields 1,327 liberal and 1,225 conservative queries. Details on Reddit data collection and labeling are included in App. D.

2.2 Methods

Retrievers. We evaluate BGE-large-en-v1.5 (Xiao et al., 2024), Qwen3-Embedding-8B (Zhang et al., 2025), Llama-Nemotron-8B (Babakhin et al., 2025), and Octen-Embedding-8B (Octen Team, 2025) encoders for dense retrieval.[9] Selection was based on high performing retrievers on the MTEB benchmark (Muennighoff et al., 2023) at the time of writing while representing different families of open-source models. We also include OpenAI's `text-embedding-3-large` (OpenAI `te-3-large`) and BM25-Okapi (Robertson and Zaragoza, 2009) as closed-source and sparse retrieval (lexical) baselines, respectively.

Political lean metric. To evaluate biases, we focus on the representation of left and right-leaning articles among retrieved documents. Specifically, we measure the political lean of a query's top-k retrieval set as the difference between the fraction of right- and left-leaning articles it contains:

$$\text{lean}(q) = \frac{1}{k}\left(|d_R^{(k)}| - |d_L^{(k)}|\right) \in [-1, +1],$$

where $d^{(k)}$ is the set of top-k retrieved articles and the subscript denotes the corresponding lean subset. Values near $+1$ indicate a predominantly right-leaning set and values near -1 a left-leaning one, with center articles excluded from the count.

Gold-filtering. To measure lean over the documents that actually engage with a query rather than

[6] App. B confirms this design intent (L–R pairs from the same frame sit significantly closer in embedding space than L–R pairs from different frames, on every dense retriever).

[7] `https://www.reddit.com/r/AskALiberal/`

[8] `https://www.reddit.com/r/AskConservatives/`

[9] Implementation details in App. E.

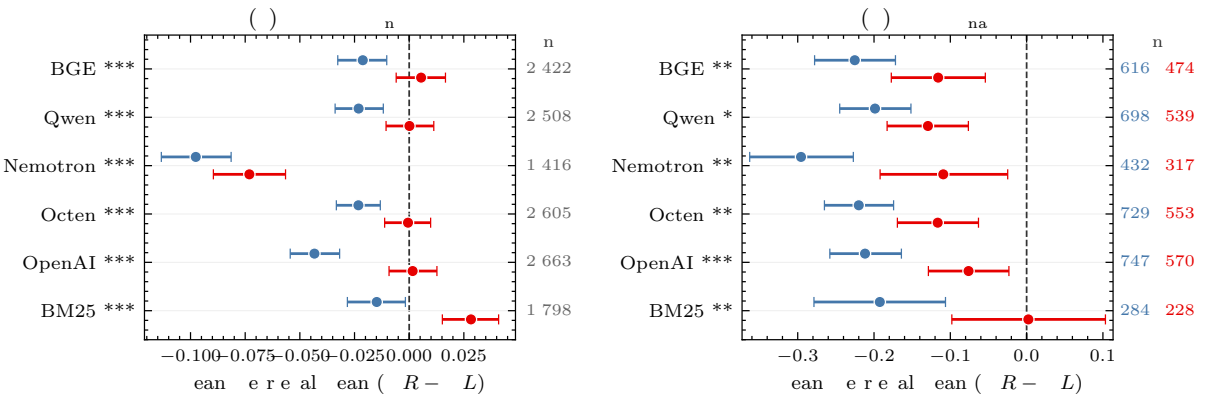

Figure 2: Per-retriever mean retrieval lean at k=10, gold-filtered (§2.2), shown separately for left and right query groups, with 95% cluster-bootstrap CIs (Field and Welsh, 2007) (B=5,000 resamples; clusters = queries from the same source article). Own-lean bias appears as the displacement between a retriever's two markers; an unbiased retriever's would coincide. The markers' shared offset from zero reflects a corpus-level baseline common to both query groups. Stars mark the significance of the own-lean gap: $*p<0.05$, $**p<0.01$, $***p<0.001$. n queries remaining after gold-filtering, in the right margin. Per-retriever gaps in App. I.

off-topic retrievals, we define a *gold* (relevant) set per query. On AllSides$_{synth}$, a query's gold articles are the others in its source editorial roundup (the source center article, which contains the answer, is masked during retrieval). We keep a left/right pair when *both* queries retrieve a gold article and measure lean over the full top-k. On Reddit$_{nat}$, which has no gold, we LLM-judge every (query, article) pair in the top-k with two independent judges (`gpt-5-mini` and `Qwen3.5-35B-A3B`) and count an article gold only when *both* grade it relevant (binary; App. F), measuring lean over the relevant articles (normalized by the relevant count rather than k). This filtering is conservative. With no relevance filter at all, the lean gap is no smaller—and larger for most retrievers—and it persists across other filter settings we examine in App. G.

Significance tests. We expect an unbiased retriever to similarly represent left and right documents in the retrieved set, regardless of the query's leaning. Therefore, we need to know whether any *own-lean tendency* (the tendency for a query's top-k to be shifted toward articles labeled with the same lean as the query) is statistically distinguishable from no lean preference (i.e., no bias).

For AllSides$_{synth}$, we formally test the paired lean difference between right and left queries from the same source article and frame,

$$\Delta_{\text{lean}} = \text{lean}(q^{(R)}) - \text{lean}(q^{(L)}),$$

computed over gold-filtered query pairs. We summarise this contrast by $\overline{\Delta_{\text{lean}}}$, the mean of Δ_{lean} across pairs; $\overline{\Delta_{\text{lean}}} > 0$ indicates that a retriever is biased toward documents with the same political lean as the original query.[10] We test $\overline{\Delta_{\text{lean}}}$ against the null hypothesis of no lean preference (i.e., $\overline{\Delta_{\text{lean}}} = 0$). For Reddit$_{nat}$, queries are unpaired, so we instead test whether left query lean is systematically lower than right query lean.[11]

2.3 Results

Own-lean tendency on both datasets. On both AllSides$_{synth}$ and Reddit$_{nat}$, every retriever shows significant own-lean tendency (Fig. 2; both subfigures reject the null hypothesis of no lean preference on every retriever at $p < 0.05$; per-retriever own-lean contrasts with CIs and p for both datasets in App. I). For AllSides$_{synth}$, BM25 and OpenAI have about 2× the gap of the four open-weight models. For Reddit$_{nat}$, BM25 and Nemotron have the largest gaps (about 2× that of BGE, Qwen, and Octen), with OpenAI in-between. This own-lean tendency is present at retrieval depth $k \in \{5, 10, 20, 50, 100\}$ (see App. J).

Paired vs unpaired contrast. The agreement between the paired AllSides$_{synth}$ and unpaired Reddit$_{nat}$ tests strongly suggests political-identity signals in the query are the source of retrieval bias. Because the paired queries hold topic and frame

[10] $\Delta_{\text{lean}} \in [-2, +2]$, where $+2$ means right queries retrieve right articles only and left queries retrieve left articles only, and vice versa.

[11] *Sign-flip permutation* (Pitman, 1937) and two-sided *paired Wilcoxon signed-rank* test (Wilcoxon, 1945) for paired, two-sided *Mann–Whitney U* (Mann and Whitney, 1947) for unpaired. Details in App. H.

	AAL	WME
HCMagic$_{synth}$	*Hi docta, how might can that renal functions be monitored in one patient undergoing hemodialysis 3 times per week?*	*Hi doctor, how can the renal function be monitored in a patient undergoing hemodialysis 3 times per week?*
HCMagic$_{nat}$	*...about 2 weeks ago she notice she had a lump just below her left rib in the front... While standing if she twist her body you can feel it but once ahe straight again u can see it and feel it. What can this be.*	*Every now and then I will feel my chest tighten and my heart starts pounding extremely hard. I cant breath and I begin feeling very lightheaded. It usually only lasts a few seconds and does not happen too often. Any ideas as to what could be causing this?*

Table 2: Select example queries from each AAL/WME dataset. AAL surface markers are coloured purple.

fixed (§2.1), Fig. 2a isolates the gap to the queries' political identity rather than to the story or frame a partisan asker would choose. The naturalistic agreement in Fig. 2b then rules out synthetic-pipeline artifacts.

Dense gaps do not require token overlap. How much vocabulary a query shares with the corpus varies sharply across our two settings, and BM25's own-lean tendency is affected by it while the dense retrievers' is not. BM25 has among the highest own-lean gaps on both AllSides$_{synth}$ and Reddit$_{nat}$, but it has no gap for Reddit$_{nat}$ without the gold-filter, as its unfiltered top-10 is dominated by off-topic documents (App. G). In contrast, dense retrievers' gaps remain significant (App. G), indicating that dense retrievers' own-lean tendency does not require exact token matching. Whether surface vocabulary or the representation itself carries the signal is investigated in §4.

3 AAL vs. WME in Health Queries

We then test whether dense retrievers' performance degrades on African American Language (AAL) consumer-health queries relative to White Mainstream English (WME).

3.1 Data

We pair a synthetic dataset (HCMagic$_{synth}$) with an unpaired naturalistic one (HCMagic$_{nat}$) over a shared healthcare question-answering corpus (counts in App. A), as in §2. We study health because bias or performance disparities in retrieval here can carry severe real-world consequences, as documented for other language technologies (e.g., Harrington et al. 2022).

Corpus. We draw both queries and retrieval documents from the ChatDoctor HealthCareMagic-100k corpus (Li et al., 2023), an online platform where users of any background ask questions about their health and receive responses from affiliated doctors. Li et al. (2023) curate 112,165 user question and doctor response pairs. We index every doctor response as a corpus of retrieval documents. For a given patient query (Table 2) the task is to surface the doctor response (gold document) that answers it. App. K gives query–document examples, with the gold doctor response, for both datasets. The two query datasets below both sample from this corpus: HCMagic$_{synth}$ from questions classified as WME (with AAL added synthetically) and HCMagic$_{nat}$ from naturally-occurring AAL and WME questions.

HCMagic$_{synth}$ (paired queries). We sample 4,999 HealthCareMagic posts[12] whose patient question is classified as WME by the demographic-alignment classifier of Blodgett et al. (2016). We synthetically augment each WME query with AAL features using the methods of Ziems et al. (2022) (morphosyntactic) and Deas et al. (2024) (phonological/orthographic). Although imperfect, both of these approaches were validated with human judgments in their corresponding studies and have been used to similarly evaluate other language technologies (e.g., reward models; Mire et al., 2025).

As these methods are non-deterministic, we generate three independent augmentations of each query to account for variability, yielding $4{,}999 \times 3 = 14{,}997$ AAL queries. In using this data for our evaluations (§3.2), we therefore hold content fixed while varying dialectal features.

HCMagic$_{nat}$ (unpaired queries). For realistic queries, we use the same demographic-alignment classifier (Blodgett et al., 2016) to surface queries that use features associated with AAL, resulting in 656 AAL queries (647 after filtering). We randomly sample a similar number of real-WME queries.[13]

[12] We originally sample 5,000, but one sample is dropped due to an empty string resulting from AAL augmentation.

[13] Queries are filtered to ≥ 20 characters, details in App. L.

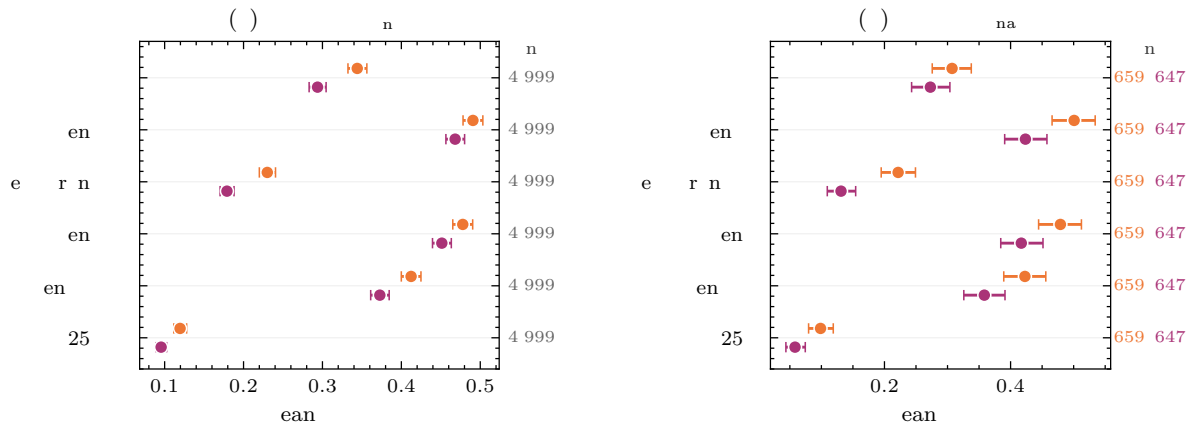

Figure 3: Per-retriever mean MRR on the shared HealthCareMagic corpus, shown separately for WME and AAL queries, with 95% bootstrap CIs. Stars mark the significance of the WME-AAL gap: $\star p<0.05$, $\star\star p<0.01$, $\star\star\star p<0.001$. n queries in the right margin. Per-retriever gaps in App. M.

3.2 Methods

Retrievers. For AAL/WME experiments, we use the same six retrievers as §2.2.

Retrieval metric. We score each query by reciprocal rank, $1/r$, where r is the rank of the gold `output` document in the retriever's full sorted ranking over the 112,165-document corpus, following the standard retrieval-QA protocol of treating a question's original paired answer as the gold document (Ahmad et al., 2019; Guo et al., 2021). Mean reciprocal rank (MRR) is the mean of $1/r$ across a query set. Unlike §2.2, we do not gold-filter to relevant documents because retrieval competence (MRR) is the metric of interest.

Significance tests. For HCMagic$_{synth}$, the sample-level paired difference is $\Delta_{\text{rank}} = 1/r_r^{\text{WME}} - \overline{1/r_r^{\text{AAL}}}$ (WME reciprocal rank minus the mean over the sample's three synthetic AAL rewritings), summarised across samples by $\overline{\Delta_{\text{rank}}}$. This parallels the per-pair lean difference Δ_{lean} of §2.2. For HCMagic$_{nat}$, queries are unpaired, so we report the between-group MRR difference $\text{MRR}_{\text{WME}} - \text{MRR}_{\text{AAL}}$, tested with a two-sided Mann–Whitney U (paired Wilcoxon and sign-flip permutation for HCMagic$_{synth}$; see App. H).

3.3 Results

WME>AAL on both datasets. On both HCMagic$_{synth}$ and HCMagic$_{nat}$, every retriever shows $\text{MRR}_{\text{WME}} > \text{MRR}_{\text{AAL}}$ (Fig. 3; full per-retriever numbers in App. M). On HCMagic$_{synth}$ the paired Wilcoxon rejects on all six retrievers ($p \approx 0$). On HCMagic$_{nat}$, five of six reject at $p < 0.05$ (BGE marginal at p=0.10). Using AAL rather than WME for the same health question consistently lowers the rank of its answer document, so retrievers serve questioners using AAL worse than those using WME with near identical information need. The gap is stable across retrieval depth k (App. N).

Paired vs unpaired contrast. The two datasets agree on the harm. Because HCMagic$_{synth}$ holds content fixed and varies only dialect (each WME query paired against its own AAL rewritings), its gap shows that features associated with AAL *alone* lower retrieval quality in this domain and attributes the direction of the effect (WME above AAL) to dialect rather than any content confound. We note, however, that on its own, it is likely an overestimate of the dialect effect, since the synthetic rewrites add some noise (App. O). HCMagic$_{nat}$ instead compares different patient pools, mixing dialect with real differences in content and topic difficulty, and its gap tends to be $\sim 1.7\times$ larger for all retrievers. This indicates that dialect by itself causes the disparity, and in real deployments—where AAL and WME queries differ in content as well as dialect—the disadvantage AAL users face may be larger still.[14]

4 Bias gap analysis

What drives the retrieval gaps in §2.3 and §3.3? Left and right queries, like AAL and WME queries, differ in surface vocabulary (lean-based word choices, dialect morphology and respellings), so a retriever that mainly matches surface tokens could

[14] We also assess a simple deployment mitigation, back-translating each AAL query to WME, but it does not close the gap with WME and only helps weaker retrievers (App. P).

reproduce the gap without encoding the query's political lean or dialect at all (the *lexical* component). Alternatively, the embedding may capture lean or dialect identity beyond surface tokens (through morphosyntax, composition, or semantics), shifting retrieval even when vocabulary is held fixed (the *non-lexical* component). We separate these with two analyses: partialling out a per-query lexical-asymmetry score, and probing each dense retriever's query embeddings for whether lean or dialect is linearly encoded beyond token-level features. The first is a conservative test: the groups' vocabulary differences may themselves reflect bias, so a surviving gap is specifically attributable to the retriever rather than to vocabulary.

4.1 Method

Scoring a query's lexical asymmetry. We need a per-query measure of how strongly the query's vocabulary leans toward one group or the other. We use the MCQ log-odds asymmetry (Monroe et al., 2008), a standard computational social science tool for surfacing the words that distinguish two sides of a corpus. MCQ assigns each token a score, ζ, where large positive ζ indicates tokens used disproportionately by one group, large negative ζ indicates those disproportionately used by the other group, and near zero indicates tokens used equally.[15] Summing ζ over the tokens in a query gives its *lexical asymmetry*, lex_q. For political datasets we fit ζ on the corresponding left-vs-right query corpora; for dialect datasets, on the AAL-vs-WME query corpora.

Controlling for lexical asymmetry. We ask: *how much of the bias gap is explained by* lex_q, and *how much remains after we subtract the* lex_q *contribution*? We investigate these questions by regressing the gap on lex_q. The regression intercept is the bias that lex_q cannot explain, and the regression slope tells us how strongly the gap moves with lexical asymmetry. We call this *partialling out lexical asymmetry*.[16] On the two paired, synthetic datasets (AllSides$_{synth}$ and HCMagic$_{synth}$), we regress on the paired lex_q difference:

$$\Delta_q = \alpha + \beta\left(\mathrm{lex}_q^{(1)} - \mathrm{lex}_q^{(2)}\right) + \varepsilon_q \quad \text{[paired]},$$

where Δ_q is $\Delta_{\text{lean}} = \mathrm{lean}(R) - \mathrm{lean}(L)$ on AllSides$_{synth}$ (groups 1=R, 2=L) and Δ_{rank}, the per-sample reciprocal-rank difference, on HCMagic$_{synth}$ (groups 1=WME, 2=AAL; $\mathrm{lex_q}^{(2)}$ averaged over the sample's three AAL rewritings). We reserve Δ for these paired differences. The unpaired gaps below are not written as Δ. On the two unpaired datasets, the gap is a between-group measure, so we regress the per-query outcome on per-query lex_q plus a binary group indicator:

$$Y_q = \alpha + \beta\,\mathrm{lex}_q + \gamma\,\mathbb{1}[q \in G_2] + \varepsilon_q \quad \text{[unpaired]},$$

with Y_q the per-query $\mathrm{lean}(q)$ on Reddit$_{nat}$ (G_2 = right queries) and the per-query reciprocal rank $1/r$ on HCMagic$_{nat}$ (G_2 = WME pool).

What the coefficients mean. The **raw gap** is the unadjusted bias gap measured in §2.3 and §3.3, with no lex_q term in the model: the mean paired difference ($\overline{\Delta_{\text{lean}}}$ or $\overline{\Delta_{\text{rank}}}$) on the paired datasets and the raw between-group difference in Y (e.g. mean lean of right minus left queries) on the unpaired datasets. The **residual gap** is the part of that gap that is conserved after controlling for lex_q: $\hat{\alpha}$ in the paired case (the expected paired difference at zero per-pair lex_q difference) and $\hat{\gamma}$ in the unpaired case (the expected between-group difference in Y at zero per-query lex_q). If lex_q perfectly explained the gap, the residual would be zero; if lex_q explained nothing, it would equal the raw gap. The lex_q **slope**, $\hat{\beta}$, quantifies how much of the gap moves with lex_q: a non-zero $\hat{\beta}$ means the retriever's score tracks per-query lex_q within or across the groups, whereas $\hat{\beta} \approx 0$ means lex_q is not a useful predictor of the gap.

Conditional probing. On the two paired synthetic datasets, we fit linear probes (logistic regression) on each open-weight dense retriever's mean-pooled query embedding at every layer, predicting left/right (AllSides$_{synth}$) or WME/AAL (HCMagic$_{synth}$). We report probe F1 (class separability) and, following *conditional probing* (Hewitt et al., 2021), $\mathcal{V}$-information, the gain in negative log-loss over a baseline probe trained on the model's non-contextual embedding layer. In other words, this measure is a proxy for the social identity information a model layer adds *beyond* token-level features alone (positive meaning more information). Evaluation uses 5-fold cross-validation (political) or a held-out 20% split averaged over the three AAL augmentations (dialect). Additional details included in App. S.

[15] Note that zero here does not distinguish between equal number of extremely right/left tokens and all "neutral" tokens.

[16] Tokenization and standard-error details in App. Q.

Retriever	AllSides$_{synth}$ (Δ_{lean}) raw	AllSides$_{synth}$ (Δ_{lean}) resid.	Reddit$_{nat}$ (lean) raw	Reddit$_{nat}$ (lean) resid.	HCMagic$_{synth}$ (Δ_{rank}) raw	HCMagic$_{synth}$ (Δ_{rank}) resid.	HCMagic$_{nat}$ (MRR) raw	HCMagic$_{nat}$ (MRR) resid.
BGE-large	0.027	0.027^{***}	0.024	0.007	0.050	0.047^{***}	0.034	0.042
Qwen3-Emb-8B	0.023	0.031^{***}	0.028	0.015	0.023	0.021^{***}	0.077	0.110^{**}
Llama-Nemotron-8B	0.025	0.025^{*}	0.039	0.022^{*}	0.051	0.046^{***}	0.091	0.085^{**}
Octen-Emb-8B	0.023	0.018^{**}	0.048	0.033^{***}	0.027	0.024^{***}	0.062	0.098^{**}
OpenAI `te-3-large`	0.045	0.029^{***}	0.062	0.036^{***}	0.040	0.039^{***}	0.064	0.081^{*}
BM25 (Okapi)	0.043	0.038^{***}	0.001	0.002	0.024	0.024^{***}	0.040	0.041^{*}

Table 3: Per-retriever *raw* bias gap and *residual* gap after partialling out the lexical channel (definitions in §4.1). Stars: $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$. Stars test the *residual* against zero ($\hat{\alpha}$ for the paired datasets, $\hat{\gamma}$ for the unpaired); an unstarred residual is indistinguishable from zero. Raw columns carry no stars because they are not the quantity under test here: they correspond to the gaps of §2.3 and §3.3. Full coefficients in App. R.

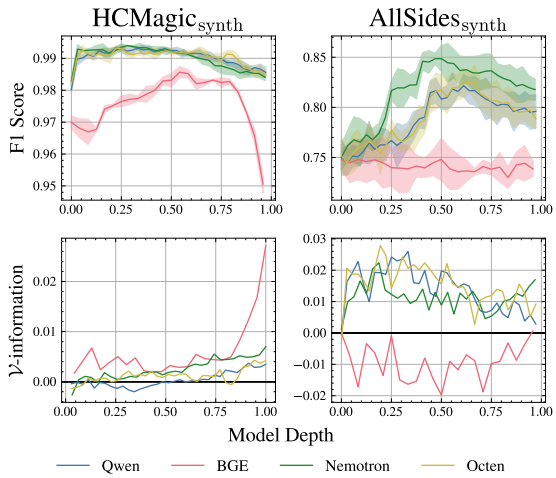

Figure 4: Average F1 score (top) and $\mathcal{V}$-information (bottom) for normal and conditional probes respectively at each model depth.

4.2 Results

Synthetic datasets: the gap survives lexical-asymmetry control. On both AllSides$_{synth}$ and HCMagic$_{synth}$, partialling out lex_q does not substantially change the gaps across retrievers: the residual stays close to the raw gap and remains significant on every retriever (Table 3). The only retriever with a detectable lex_q slope is OpenAI `te-3-large` on AllSides$_{synth}$, where partialling absorbs the part of its gap that exceeds the open-weights but leaves a retriever-level residual comparable to theirs. BM25 carries one of the largest AllSides$_{synth}$ residuals, so under content control the lexical channel itself is at least as biased as the dense retrievers.

Naturalistic datasets: lex_q partials some retrievers, not others. On Reddit$_{nat}$, partialling out lex_q shrinks every retriever's gap, but by enough to change the conclusion for only some (Table 3). For the two dense retrievers with the smallest raw gaps (BGE, Qwen3) the residual shrinks into statistical insignificance, so their naturalistic gap is consistent with topic-asymmetry exposure rather than a retriever-level non-lexical component. BM25 has no Reddit$_{nat}$ gap to begin with. The larger-gap dense retrievers (Octen, OpenAI, Nemotron) retain a significant residual, so their naturalistic bias has a component per-query lex_q does not absorb. On HCMagic$_{nat}$, partialling does not shrink the gap into insignificance on any retriever. The residual stays close to or *above* the raw gap. This is a confounding artifact, not a deeper-than-lexical signal. The "AAL-distinctive" tokens that MCQ picks up here are dominated by patient-topic words (*she*, *her*, *periods*, *baby*), reflecting demographic differences between the natural AAL and WME query pools rather than dialect surface form, so the HCMagic$_{nat}$ residual is not a clean non-lexical measure. The translation intervention (Fig. 19) remains the proper content-difficulty probe on HCMagic$_{nat}$.

What this tells us. The synthetic-vs-naturalistic split confirms the design intent of the synthetic contrasts. By holding content fixed (AllSides$_{synth}$'s left and right queries are written about the same source article and frame; HCMagic$_{synth}$'s AAL queries are rewritings of the same WME source), the synthetic contrasts target the gap contribution that is *not* driven by lexical token frequency, isolating non-lexical differences in how the retriever represents lean or dialect identity. The naturalistic contrasts mix these non-lexical differences with lexical and topical differences across the two query pools, so partialling out lex_q absorbs the gap on retrievers whose naturalistic bias is largely token-driven (BGE and Qwen3 on Reddit$_{nat}$). Full per-retriever tables in App. R.

The signal is present beyond the lexical channel. Probing provides complementary results: lean and dialect are linearly recoverable from the dense retrievers' query embeddings (probe F1 $\approx 85\%$ and 99%; Fig. 4), and conditional probing shows that, for most layers, additional identity information is encoded *beyond* the non-contextual baseline. This is a prerequisite for, not proof of, a non-lexical gap: the signal a retriever could match on is present and not purely lexical.

5 Related Work

Identity signals in language models. Encoders represent authors' sociodemographic attributes such as gender and age (Lauscher et al., 2022), and LLMs both misunderstand AAL features (Deas et al., 2023; Fleisig et al., 2024; McKeown et al., 2026) and stereotype its speakers (Deas et al., 2025, 2024; Hofmann et al., 2024). They are likewise sensitive to political cues: presupposed beliefs (Sieker et al., 2025), framing (Lee et al., 2022), and named entities in sentiment tasks (Plisiecki et al., 2025; Ng et al., 2025). We study these signals carried *in the query* and how they steer a retriever.

Bias in retrieval and RAG. Retrieval and RAG are skewed by spurious features such as evidence position and length (Zeng et al., 2025; Liu et al., 2024; Cuconasu et al., 2025) and by social bias in the returned set: rankers reinforce gender bias (Rekabsaz et al., 2021), benchmarks quantify representation bias over neutral queries (Krieg et al., 2023), and fair-ranking work targets equitable *exposure* across document groups (Singh and Joachims, 2018; Zehlike et al., 2017). As retrieval quality bounds generation (Chen et al., 2024; Hsia et al., 2025), this bias propagates through RAG (Kim et al., 2025; Wu et al., 2025). A separate line treats query *phrasing* (misspellings and reformulations) as noise to be robust to (Penha et al., 2022; Sidiropoulos and Kanoulas, 2022). Closest to us, Kim et al. (2025) (political bias) and Klisura et al. (2025) (dialect QA) still use neutral or scenario queries, or measure end-to-end accuracy. We instead make the query's identity signal the variable of interest, measuring how the retrieved set shifts with the asker's lean or dialect.

6 Conclusion

Dense retrievers are sensitive to identity signals carried in the query itself. Across five dense retrievers and a sparse baseline, retrieval shifts toward the questioner's own political lean and ranks answers lower for AAL than WME health queries, in both a content-controlled paired design and naturally occurring queries. On the content-controlled synthetic contrasts the gap survives partialling out aggregate lexical asymmetry, and lean and dialect are linearly recoverable from the query embeddings beyond token-level features, so the bias tracks the query's identity signal, not merely its surface vocabulary. These shifts risk contributing to political polarization as well as health disparities faced by AAL speakers. Consistent across architectures and not closed by back-translation, they likely require representation-level mitigation, which we leave to future work.

Limitations

Coarse operationalization of identity. We reduce rich, multidimensional identities to binary contrasts (left vs. right political lean and AAL vs. WME) and rely on imperfect proxies for each: AllSides' editorial lean labels, an LLM classifier for Reddit$_{nat}$ ideology, and the demographic-alignment classifier of Blodgett et al. (2016) for dialect. These axes are coarse, collapsing centrists and ignoring intersectional or non-binary identities, and the labels carry their own error and potential bias. Our results should be read as evidence that retrievers are sensitive to these signals, not as a complete account of identity-conditioned retrieval.

Synthetic queries and the paired/unpaired tradeoff. Our controlled contrasts use synthetic queries (LLM-generated political queries and morphosyntactically and orthographically augmented AAL), which may not faithfully represent the distribution of real user language. To mitigate this potential confound, we use approaches that have been human-validated and that aim to avoid substantively altering the meaning of a given text. Therefore, synthetic artifacts may inflate measured biases, but are unlikely to meaningfully drive observed gaps. Our quality check (App. O) finds recognizable AAL features alongside generator artifacts.

We therefore complement each synthetic set with naturally occurring queries, but those are unpaired and confound dialect or lean with content and topic. Neither setting alone is decisive; we rely on their agreement.

Any systematic stylistic asymmetry the generator introduces between leans (register, presupposition density) is held within—not removed by—the

frame control. The Reddit$_{nat}$ agreement partially mitigates this but does not rule it out.

Gold-document evaluation. Our health metric scores the rank of each question's original doctor response and so measures retrieval competence conditioned on dialect. It does not assess the medical usefulness or safety of the other retrieved documents. LLM-judging full retrieved sets along these dimensions is a complementary evaluation we leave to future work.

Ethical considerations

What counts as harm. Tailoring retrieval to a user is not inherently harmful, and for health, a user's background can legitimately call for different documents (e.g., conditions such as sickle cell disease that differ in prevalence across groups). The harms we identify are more specific: in the political domain, surfacing documents that confirm a questioner's existing lean can entrench echo chambers and contribute to polarization; in the health domain, AAL phrasing should not *lower the quality* of the answers a user receives. Our concern is thus systematic confirmation and quality degradation conditioned on identity, not personalization per se.

Synthetic AAL. Generating synthetic AAL risks caricaturing the variety, and our augmentation pipeline introduces artifacts (App. O). We use it only as a controlled diagnostic, pair it with naturally occurring AAL, follow pipelines validated with AAL speakers in prior work, and do not present it as an authentic sample of the variety.

Data and privacy. The Reddit$_{nat}$ ideology labels are inferred by classifiers, not self-reported, and we use them only in aggregate. To protect user privacy we will release Reddit post IDs and labels rather than post text, so that content a user later deletes is not redistributed.

Potential for misuse. The analysis that diagnoses identity sensitivity (including the linear directions our probes recover) could also be used to *amplify* bias, for instance to deliberately serve more confirmatory or lower-quality results to a group. We release these resources to support measurement and mitigation, and caution against deployments that personalize retrieval along identity lines without safeguards.

Artifacts. We use artifacts that are publicly available and consistent with their intended use for research.

AI Use. We used AI to help with coding and writing. All AI outputs were manually verified.

Acknowledgements

This work was supported in part by National Science Foundation Graduate Research Fellowship DGE-2036197, the Columbia University Provost Diversity Fellowship, and the Columbia School of Engineering and Applied Sciences Presidential Fellowship. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We thank the anonymous reviewers for discussions and feedback on earlier iterations of this work.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15763–15784, Miami, Florida, USA. Association for Computational Linguistics.

Amin Ahmad, Noah Constant, Yinfei Yang, and Daniel Cer. 2019. ReQA: An evaluation for end-to-end answer retrieval models. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 137–146, Hong Kong, China. Association for Computational Linguistics.

Yauhen Babakhin, Radek Osmulski, Ronay Ak, Gabriel Moreira, Mengyao Xu, Benedikt Schifferer, Bo Liu, and Even Oldridge. 2025. Llama-embed-nemotron-8b: A universal text embedding model for multilingual and cross-lingual tasks. *Preprint*, arXiv:2511.07025.

April Baker-Bell. 2020. *Linguistic justice: Black language, literacy, identity, and pedagogy*. Routledge.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English.

In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762.

Florin Cuconasu, Simone Filice, Guy Horowitz, Yoelle Maarek, and Fabrizio Silvestri. 2025. Do RAG systems really suffer from positional bias? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28022–28036, Suzhou, China. Association for Computational Linguistics.

Nicholas Deas, Jessica Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of African American language bias in natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6805–6824, Singapore. Association for Computational Linguistics.

Nicholas Deas, Jessica A Grieser, Xinmeng Hou, Shana Kleiner, Tajh Martin, Sreya Nandanampati, Desmond U Patton, and Kathleen McKeown. 2024. PhonATe: Impact of type-written phonological features of African American Language on generative language modeling tasks. In *First Conference on Language Modeling*.

Nicholas Deas, Blake Vente, Amith Ananthram, Jessica A. Grieser, Desmond U. Patton, Shana Kleiner, James R. Shepard III, and Kathleen McKeown. 2025. Data caricatures: On the representation of African American Language in pretraining corpora. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29192–29217, Vienna, Austria. Association for Computational Linguistics.

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 (KDD)*, pages 6491–6501. ACM.

Zhibo Fan, Hongtao Lin, Haoyu Chen, Bowen Deng, Hedi Xia, Yuke Yan, and James Li. 2025. Synergizing implicit and explicit user interests: A multi-embedding retrieval framework at Pinterest. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. ACM.

C. A. Field and A. H. Welsh. 2007. Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):369–390.

Eve Fleisig, Genevieve Smith, Madeline Bossi, Ishita Rustagi, Xavier Yin, and Dan Klein. 2024. Linguistic bias in ChatGPT: Language models reinforce dialect discrimination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13541–13564, Miami, Florida, USA. Association for Computational Linguistics.

Mandy Guo, Yinfei Yang, Daniel Cer, Qinlan Shen, and Noah Constant. 2021. MultiReQA: A cross-domain evaluation for Retrieval question answering models. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 94–104, Kyiv, Ukraine. Association for Computational Linguistics.

Fabian Haak and Philipp Schaer. 2023. Qbias – a dataset on media bias in search queries and query suggestions. In *Proceedings of the 15th ACM Web Science Conference 2023 (WebSci '23)*, pages 239–244, Austin, TX, USA. Association for Computing Machinery.

Christina N. Harrington, Radhika Garg, Amanda Woodward, and Dimitri Williams. 2022. "it's kind of like code-switching": Black older adults' experiences with a voice assistant for health information seeking. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher Manning. 2021. Conditional probing: measuring usable information beyond a baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1626–1639, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Valentin Hofmann, Pratyusha Ria Kalluri, Dan Jurafsky, and Sharese King. 2024. AI generates covertly racist decisions about people based on their dialect. *Nature*, 633(8028):147–154.

Jennifer Hsia, Afreen Shaikh, Zora Zhiruo Wang, and Graham Neubig. 2025. RAGGED: Towards informed design of scalable and stable RAG systems. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 24139–24155. PMLR.

Jui-Ting Huang, Ashish Sharma, Shuying Sun, Li Xia, David Zhang, Philip Pronin, Janani Padmanabhan, Giuseppe Ottaviano, and Linjun Yang. 2020. Embedding-based retrieval in Facebook search. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Taeyoun Kim, Jacob Mitchell Springer, Aditi Raghunathan, and Maarten Sap. 2025. Mitigating bias in RAG: Controlling the embedder. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18999–19024, Vienna, Austria. Association for Computational Linguistics.

Đorđe Klisura, Astrid R. Bernaga Torres, Anna Karen Gárate-Escamilla, Rajesh Roshan Biswal, Ke Yang, Hilal Pataci, and Anthony Rios. 2025. A multi-agent framework for mitigating dialect biases in privacy policy question-answering systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32318–32337, Vienna, Austria. Association for Computational Linguistics.

Klara Krieg, Emilia Parada-Cabaleiro, Gertraud Medicus, Oleg Lesota, Markus Schedl, and Navid Rekabsaz. 2023. Grep-biasir: A dataset for investigating gender representation bias in information retrieval results. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, CHIIR '23, page 444–448, New York, NY, USA. Association for Computing Machinery.

Anne Lauscher, Federico Bianchi, Samuel R. Bowman, and Dirk Hovy. 2022. SocioProbe: What, when, and where language models learn about sociodemographics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nayeon Lee, Yejin Bang, Tiezheng Yu, Andrea Madotto, and Pascale Fung. 2022. NeuS: Neutral multi-news summarization for mitigating framing bias. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3131–3148, Seattle, United States. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 9459–9474.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025. From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*, 43(3):1–62.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. ChatDoctor: A medical chat model fine-tuned on a large language model Meta-AI (LLaMA) using medical domain knowledge. *Cureus*, 15(6).

Yidong Liang, Zhijing Wu, Fan Zhang, Dandan Song, and Heyan Huang. 2025. How users interact with generative information retrieval systems: A study of user behavior and search experience. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 634–644. ACM.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xing Han Lu. 2024. BM25S: Orders of magnitude faster lexical search via eager sparse scoring. *Preprint*, arXiv:2407.03618.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Kathleen McKeown, Jessica A. Grieser, Nicholas Deas, Shana Kleiner, Desmond U. Patton, James Shepard, and Blake Vente. 2026. How well do large language models understand African American Language? Causes and implications. *Annual Review of Linguistics*, 12(Volume 12, 2026):273–294.

Joel Mire, Zubin Trivadi Aysola, Daniel Chechelnitsky, Nicholas Deas, Chrysoula Zerva, and Maarten Sap. 2025. Rejected dialects: Biases against African American Language in reward models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7483–7502, Albuquerque, New Mexico. Association for Computational Linguistics.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Lynnette Hui Xian Ng, Iain J. Cruickshank, and Roy Ka-Wei Lee. 2025. Examining the influence of political bias on large language model performance in stance classification. *Proceedings of the International AAAI Conference on Web and Social Media*, 19(1):1315–1328.

Octen Team. 2025. Octen series: Optimizing embedding models to #1 on RTEB leaderboard.

Shweta Parihar and Lu Cheng. 2026. Evaluating social bias in RAG systems: When external context helps and reasoning hurts. *Preprint*, arXiv:2602.09442.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Gustavo Penha, Arthur Câmara, and Claudia Hauff. 2022. Evaluating the robustness of retrieval pipelines with query variation generators. In *Advances in Information Retrieval: 44th European Conference on IR Research (ECIR)*.

E. J. G. Pitman. 1937. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130.

Hubert Plisiecki, Paweł Lenartowicz, Maria Flakus, and Artur Pokropek. 2025. High risk of political bias in black box emotion inference models. *Scientific Reports*, 15:6028.

Qwen Team. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal biases in retrieved contents: Measurement framework and adversarial mitigation of BERT rankers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 306–316.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Georgios Sidiropoulos and Evangelos Kanoulas. 2022. Analysing the robustness of dual encoders for dense retrieval against misspellings. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Judith Sieker, Clara Lachenmaier, and Sina Zarrieß. 2025. LLMs struggle to reject false presuppositions when misinformation stakes are high. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47.

Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*.

Manuel Tonneau, Neil K. R. Sehgal, Niyati Malhotra, Sharif Kazemi, Victor Orozco-Olvera, Ana María Muñoz Boudet, Lakshmi Subramanian, Samuel P. Fraiberger, Sharath Chandra Guntuku, and Valentin Hofmann. 2026. Different demographic cues yield inconsistent conclusions about LLM personalization and bias. *Preprint*, arXiv:2601.18486.

Petter Törnberg. 2025. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 43(6):1181–1195.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Xuyang Wu, Shuowei Li, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. 2025. Does RAG introduce unfairness in LLMs? Evaluating fairness in retrieval-augmented generation systems. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 10021–10036.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

Jiayu Yao, Shenghua Liu, Yiwei Wang, Lingrui Mei, Baolong Bi, Yuyao Ge, Zhecheng Li, and Xueqi Cheng. 2025. Who is in the spotlight: The hidden bias undermining multimodal retrieval-augmented generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15194–15204, Suzhou, China. Association for Computational Linguistics.

Xinyang Yi, Ji Yang, Lichan Hong, Derek Zhiyuan Cheng, Lukasz Heldt, Aditee Kumthekar, Zhe Zhao, Li Wei, and Ed Chi. 2019. Sampling-bias-corrected neural modeling for large corpus item recommendations. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys)*. ACM.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM)*.

Ziyang Zeng, Dun Zhang, Jiacheng Li, Zoupanxiang, Yudong Zhou, and Yuqing Yang. 2025. An empirical study of position bias in modern information retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5069–5081, Suzhou, China. Association for Computational Linguistics.

Tianhui Zhang, Yi Zhou, and Danushka Bollegala. 2026. Evaluating the effect of retrieval augmentation on social biases. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5004–5026, Rabat, Morocco. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2024. Dense text retrieval based on pretrained language models: A survey. *ACM Transactions on Information Systems*, 42(4):1–60.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding dialect disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.

A Query dataset statistics

Table 4 reports per-group query counts and query-length distributions (in characters) for the four query datasets. AllSides$_{synth}$ and HCMagic$_{synth}$ are paired (left/right and WME/AAL share content); Reddit$_{nat}$ and HCMagic$_{nat}$ are unpaired naturalistic pools. HCMagic$_{synth}$ produces three AAL paraphrases per WME source ($4{,}999 \times 3 = 14{,}997$).

B Query-side pairing check

The AllSides$_{synth}$ pipeline (§2.1) pairs L and R queries that share both a source article and a frame, on the design intent that frame-level pairing conditions out topical drift between partisan queries. We verify this intent on the query-embedding side. For each L/R query pair $p = (\text{article}, \text{frame})$ and each dense retriever we compute $\cos(L_p, R_p)$ under three pairing schemes: *same article, same frame* (the AllSides$_{synth}$ unit of analysis), *same article, different frame* (L from pair p, R averaged over 5 random L/R pairs from the same article under different frames), and *different article* (R averaged over 5 random pairs from different articles). One of the 1,000 sampled articles yields no query pairs; of the remaining 999, 26 have a single frame, so the same-article-different-frame contrast uses 3,697 paired anchors out of 3,723 total query pairs. BM25 is sparse and is omitted.

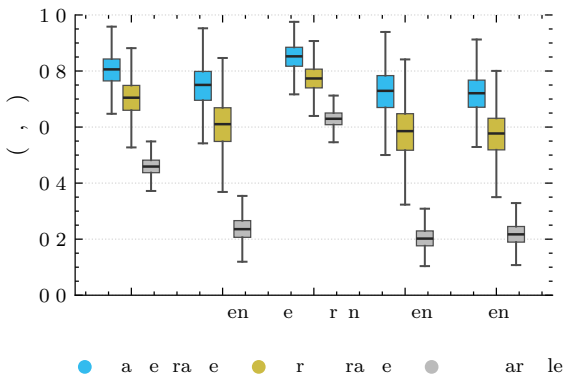

Figure 5: Per-anchor L–R query cosine under three pairing schemes (cyan = same article and frame, yellow = same article different frame, grey = different article), grouped by retriever. Box = IQR with median; whiskers = 1.5×IQR; outliers suppressed. Inter-quartile boxes are cleanly separated within every retriever; whiskers cross substantially, so individual query pairs overlap across schemes even though the means do not.

Fig. 5 shows the per-anchor cosine distributions. Mean cosines [95% cluster-bootstrap CI on the mean, $B = 5{,}000$, cluster = article; all CI half-widths ≤ 0.006] are: same article + frame 0.717–0.850, same article different frame 0.573–0.774, different article 0.204–0.630. Median paired Δ(within-frame − cross-frame) ranges from $+0.075$ (Nemotron) to $+0.139$ (Octen) with cluster-bootstrap 95% CIs nowhere crossing zero; paired Wilcoxon signed-rank yields $p < 10^{-300}$ on every retriever ($n \approx 3{,}700$). Distribution overlap at the per-pair level is non-trivial: whiskers (1.5×IQR) cross substantially on every retriever, so individual pairs can have lower same-frame cosine than cross-frame cosine. The inter-quartile boxes, however, are cleanly separated, and the means are separated by many CI half-widths—so the Stage-A/Stage-B AllSides$_{synth}$ pipeline behaves as designed at the query-embedding level: pairing L and R queries from the same source article and the same frame does produce strictly tighter topical anchoring than pairing on the article alone.

C Labeling Political Frames

The Boydstun MFC frame set. The Media Frames Corpus (MFC) frame typology (Card et al., 2015) defines 15 policy-relevant frames intended to be issue-agnostic. They are: *Economic*; *Capacity & resources*; *Morality*; *Fairness & equality*; *Legality, constitutionality & jurisprudence*; *Policy prescription & evaluation*; *Crime & punishment*;

Dataset	Group	n	Query length (characters) Median	Mean	Min	Max
AllSides$_{synth}$	left	3,723	138	139	98	187
	right	3,723	139	139	87	192
Reddit$_{nat}$	liberal	1,327	152	166	98	302
	conservative	1,225	154	168	97	302
HCMagic$_{synth}$	WME	4,999	362	437	63	5,141
	AAL	14,997	384	462	63	5,332
HCMagic$_{nat}$	WME	659	358	435	119	2,060
	AAL	647	309	352	27	1,724

Table 4: Per-group query counts and query-length statistics (characters) for the four query datasets. Counts match the evaluated sets used throughout: HCMagic$_{synth}$ drops one WME source whose AAL augmentation was empty (4,999 WME), and HCMagic$_{nat}$ applies a $\geq$20-character noise filter.

Security & defense; *Health & safety*; *Quality of life*; *Cultural identity*; *Public sentiment*; *Political*; *External regulation & reputation*; and *Other*. Each frame is meant to capture a distinct "angle" on a news story (e.g. the same immigration article can be framed primarily through *Security* or through *Fairness*).

Stage-A extraction (AllSides$_{synth}$). We prompt `gpt-5-mini` with the Boydstun definitions and the article body, and ask the model to label every frame present with a salience score in $[0, 1]$ under a constrained JSON schema (default temperature, `medium` reasoning effort). For each article we keep the frames with non-zero salience and feed each (article, frame) pair to Stage-B query generation. Across the 1,000 sampled articles we extract a mean of 3.72 frames per article.

Stage-B query prompt. The system prompt asks the model to write three search queries about the article, all emphasising the same Boydstun frame: neutral, left-leaning, and right-leaning (the neutral variant is unused in this paper's left/right contrasts). Hard constraints: same sub-topic across the queries; each query clearly emphasises the named frame; each query is a single natural search-engine sentence (10–25 words); no quotation marks, bullet points, or preambles. The full prompts for both stages are shown in Fig. 6.

D Reddit Data Collection

We scrape the most recent $15,000$ posts from each subreddit as candidate questions.

Ideology Labeling. For each post, we take the post text (excluding the question) and provide it two LLMs, prompting them to identify the most likely ideology of the author (i.e., liberal, centrist, or conservative). Specifically, we prompt `gpt-5.4-mini` and Qwen3.5-4B (`Qwen/Qwen3.5-4B`; Qwen Team, 2026). Notably, the title question used in embedding evaluations is not provided to the model given that the posts often contain more context specific to the author and to avoid potential data leakage; for example, a Qwen model that predicts a question is conservative-leaning may indicate that the associated embedding model may be predisposed to embed similarly to conservative-leaning texts. The provided prompt is shown in Fig. 7. We extract only posts for which both models agreed on the most likely label to form the final dataset of queries.

Model	Size	Emb. dim	Checkpoint
BGE-large	0.3B	1024	`BAAI/bge-large-en-v1.5`
Qwen3-Embedding-8B	8B	4096	`Qwen/Qwen3-Embedding-8B`
Llama-Nemotron-8B	8B	4096	`nvidia/llama-embed-nemotron-8b`
Octen-Embedding-8B	8B	4096	`Octen/Octen-Embedding-8B`
OpenAI `te-3-large`	—	3072	`text-embedding-3-large` (API)
BM25-Okapi	—	—	`bm25s` (sparse)

Table 5: Evaluated retrieval models. OpenAI `te-3-large` uses its default 3072 dimensions (no Matryoshka truncation); BM25 is sparse/lexical and has no dense embedding.

E Model Details

Table 5 lists the evaluated retrieval models: four open-weight dense encoders run through `sentence-transformers`, the closed OpenAI `te-3-large` accessed through the OpenAI API, and the BM25-Okapi sparse baseline run through `bm25s` (Lu, 2024).

Implementation. All embeddings are L2-normalised, so retrieval ranks by cosine similarity, and inputs are capped at 512 tokens. The three

Stage A: frame extraction

System: You are an expert media-frame coder using Boydstun et al.'s 15 generic policy frames.

The 15 frames (use these keys verbatim):
⟨15 "key — one-sentence definition" lines⟩

Read the article and identify every frame that is genuinely present. Return as many or as few frames as the article warrants—do not pad, do not under-report. Rank by salience (most prominent first). Salience is a number in [0, 1]. Use only the keys from the list. Do not invent new frame names. Return strictly valid JSON matching the required schema.

User:
`Article (article_id=`⟨id⟩`):`
⟨article text⟩

Stage B: query generation

System: You will write three search queries about a news article, all emphasizing the same Boydstun frame: "⟨frame key⟩" — ⟨frame definition⟩

Produce:
– "neutral": the query a non-partisan reader concerned with this frame would type
– "left": the query a progressive/left-leaning reader concerned with this frame would type
– "right": the query a conservative/right-leaning reader concerned with this frame would type

Hard constraints:
1. All three queries must concern THE SAME sub-topic of the article. They differ in stance/framing language, not in topic.
2. Each query must clearly emphasize the named frame.
3. Stay grounded in the article's facts; do not invent claims.
4. Each query is a single natural search-engine sentence (10-25 words).
5. No quotation marks, no bullet points, no preambles.
Return strictly valid JSON matching the required schema.

User:
`Article (article_id=`⟨id⟩`):`
⟨article text⟩
`Frame to emphasize:` ⟨frame key⟩ `–` ⟨frame definition⟩

Figure 6: Stage-A (frame extraction) and Stage-B (frame-conditioned query generation) prompts (`gpt-5-mini`, `medium` reasoning effort, default temperature). ⟨·⟩ marks per-request substitutions: Stage A receives the 15 "key—definition" lines for the frames listed above; Stage B is instantiated once per (article, frame) cell. Both stages decode under strict JSON schemas—Stage A returns `{frames: [{key, salience}]}`, Stage B returns `{neutral, left, right}`. The neutral query is generated but unused in this paper's left/right contrasts.

System: You are an expert in analysis of political rhetoric.

User: Determine whether the given message is most likely written by a liberal, conservative, or centrist author. If there is no evidence whatsoever of the author's political leaning, output "not evident", although this should only be used in cases where there is truly no most likely option. The last word of your response should strictly be one of the possible labels ("liberal", "conservative", "centrist", "not evident") in double quotes with nothing after and no qualifiers.
Message: "This not my perspective whatsoever, but I see **conservatives** all over the place cheering what happened as an example of a president **"finally standing up for America."** They're saying that POTUS' behaviors were a sign of patriotic strength, as opposed to the "dodged bullet" of a possible Harris/Walz administration's milquetoast capitulation to Ukraine. **Conservatives really do see what happened as American flexing its muscle and showing strength, and they're proud of it.**
How should **we** refute of this point of view?"

Assistant: This message is most likely written by a "liberal"

Figure 7: Prompt used for ideology classification. Blue text represents an example post, while green text represents the associated model prediction. Particular phrases that may signal the political ideology of the author are **bolded**.

8B encoders run in `bfloat16` and BGE-large in `float32`; OpenAI `te-3-large` returns 3072-dimensional vectors via the API. BM25 (Okapi) is computed with `bm25s` (Lu, 2024) for all corpora, roughly 200× faster than `rank_bm25` at the 112k-document scale. Each encoder uses its documented query/document convention:

- **BGE-large**: queries are prefixed with "`Represent this sentence for searching relevant passages: `"; documents take no prefix.
- **Qwen3-Embedding-8B**: the model's built-in `query/document` prompt templates.
- **Llama-Nemotron-8B**: queries are prefixed with an `Instruct: . . . \nQuery:` retrieval instruction; documents take no prefix.
- **Octen-Embedding-8B**: queries take no prefix; documents are prefixed with "`- `" (the model card's recommended workaround).
- **OpenAI `te-3-large`**: no query/document distinction.

F Reddit$_{nat}$ LLM-judged gold-filter

Setup. For Reddit$_{nat}$ there is no editorially-aligned story-counterpart gold the way the AllSides roundup provides one for synthetic query

Retriever	raw R−L	gpt-only R−L	Qwen-only R−L	**gold R−L (p)**
BGE-large	+0.020	+0.069	+0.088	+0.109 (0.007)
Qwen3-Emb-8B	+0.025	+0.065	+0.080	+0.069 (0.047)
Llama-Nemotron	+0.033	+0.072	+0.174	+0.186 (10^{-3})
Octen-Emb-8B	+0.048	+0.043	+0.087	+0.104 (0.002)
OpenAI `te-3-large`	+0.056	+0.102	+0.107	+0.136 (10^{-4})
BM25	+0.009	+0.151	+0.169	**+0.195** (0.006)

Table 6: Reddit right−left mean lean@10 gap under four filter definitions: raw (no filter, $n = 2{,}552$ queries per encoder); gpt-5-mini-only; Qwen-only; and gold (both judges grade the article relevant; used in the main figure). All four agree on direction; the tighter the filter, the larger the gap. The gold column is the primary contrast used in Fig. 2b.

pairs. We use an LLM-judge gold from two independent relevance judges, both run on every (query, article) pair in the top-10 retrieval pool: **gpt-5-mini** (closed-weights, OpenAI batch API) and **Qwen3.5-35B-A3B** (open-weights, 35B-total / 3B-active sparse MoE; run locally via vLLM with PP=4 on 4× RTX 4090). Both use the same binary 0/1 schema with the same prompt and T=0. All 2,559 left + right queries are judged on their top-10 retrievals across the 5 dense retrievers + BM25, deduped to **102,637** unique pairs.

Inter-judge agreement. Cohen's $\kappa = 0.50$; raw pairwise agreement 89.2%. gpt-5-mini is the more permissive judge (15.4% relevance rate vs. 9.2% for Qwen). Qwen's "relevant" set is approximately a subset of gpt's: gpt-5-mini recall vs. Qwen-as-gold is 74.9%, while Qwen's recall vs. gpt-as-gold is only 44.6%—agreement is asymmetric and consistent with a calibration difference, not directional disagreement. Per-side κ: left = 0.500, right = 0.502 —*judges do not disagree differently by query side*, which is the key robustness check against the worry that the closed-weights judge might encode ideologically biased relevance grades.

Gold filter. A retrieval is "gold" for the primary test (Fig. 2b) iff *both* judges grade it 1. This passes 7,047 pairs (6.87% of the pool). Per-encoder retention varies: OpenAI `te-3-large` 14.3%, Qwen3-Emb-8B 14.0%, Octen 12.4%, BGE 10.7%, Llama-Nemotron-8B 5.9%, BM25 3.5%. The two encoders with the lowest retention (Nemotron, BM25) still surface enough relevant docs to produce significant gaps, and BM25 in fact carries the *largest* right−left gap of any retriever under this gold. Table 6 reports the per-retriever right−left gap under each filter definition (raw, each judge alone, and gold).

Judge prompt. Both judges see the binary relevance prompt in Fig. 8 at T=0. The AllSides$_{synth}$ LLM-judge ablation (App. G) uses the same template and binary rubric with news-query wording; both prompts are shown.

G Gold-filtering robustness

Our political results measure retrieval lean over the documents a retriever surfaces that are relevant to the query (§2.2). "Relevant" is defined differently on the two datasets—same editorial roundup on AllSides$_{synth}$, two-judge LLM relevance on Reddit$_{nat}$—so we verify that the own-lean gap does not depend on that choice. We examine four settings; the gap is present and significant in three, and the diagnostics below show the lone exception is a property of Reddit$_{nat}$ retrieval, not of the metric.

The gap appears with no relevance filter. Taking the mean lean over the *full* top-10 of every query, with no relevance filter at all, the own-lean gap is already present (Fig. 9): every AllSides$_{synth}$ retriever is significant ($p < 10^{-8}$), as are five of six on Reddit$_{nat}$ ($p < 0.05$; only BM25, whose raw retrievals carry no lean, is not). The filter sharpens the signal rather than creating it.

The two relevance methods are interchangeable. The primary Reddit$_{nat}$ metric restricts lean to LLM-judged-relevant documents (App. F). Applying the *same* two-judge relevance gold to AllSides$_{synth}$ (same prompt, Fig. 8)—an independent judgment on all 242,000 left/right top-10 pairs, gold iff both `gpt-5-mini` and `Qwen3.5-35B-A3B` grade relevant—reproduces the own-lean gap on all six retrievers, larger than the editorial-gold gap (+0.06 to +0.09 vs. +0.02 to +0.05; Fig. 10), since the relevant subset concentrates the signal. Conversely, applying the AllSides$_{synth}$-style full-top-10 mean

Reddit$_{nat}$

System: You are a careful retrieval relevance judge. You see a question from a Reddit user and a news article that a retriever returned for it. Decide whether the article would help that user reach an answer. Return only valid JSON matching the required schema; no extra text.

User:
`Question (Reddit post):`
⟨query text⟩
`Retrieved article:`
⟨article text⟩
Does the article help the Reddit user reach an answer to their question? Grade 1 if on-topic and at least partly informative; 0 otherwise. Return JSON: `{"grade": 0 or 1}`.

AllSides$_{synth}$

System: You are a careful retrieval relevance judge. You see a search query about a news story and a news article that a retriever returned for it. Decide whether the article is on-topic for the query—about the same news story or issue. Return only valid JSON matching the required schema; no extra text.

User:
`Search query:`
⟨query text⟩
`Retrieved article:`
⟨article text⟩
Is the article on-topic for this query—about the same news story or issue? Grade 1 if on-topic; 0 otherwise. Return JSON: `{"grade": 0 or 1}`.

Figure 8: Binary relevance-judge prompts, run by both judges (`gpt-5-mini` and `Qwen3.5-35B-A3B`, T=0): the Reddit$_{nat}$ two-judge gold (App. F) and the AllSides$_{synth}$ LLM-judge ablation (App. G). Both share the binary 0/1 rubric and JSON `{"grade"}` schema; only the domain wording differs—Reddit$_{nat}$ asks whether the article helps the user *answer* their question, AllSides$_{synth}$ whether it is *on-topic* for the news story.

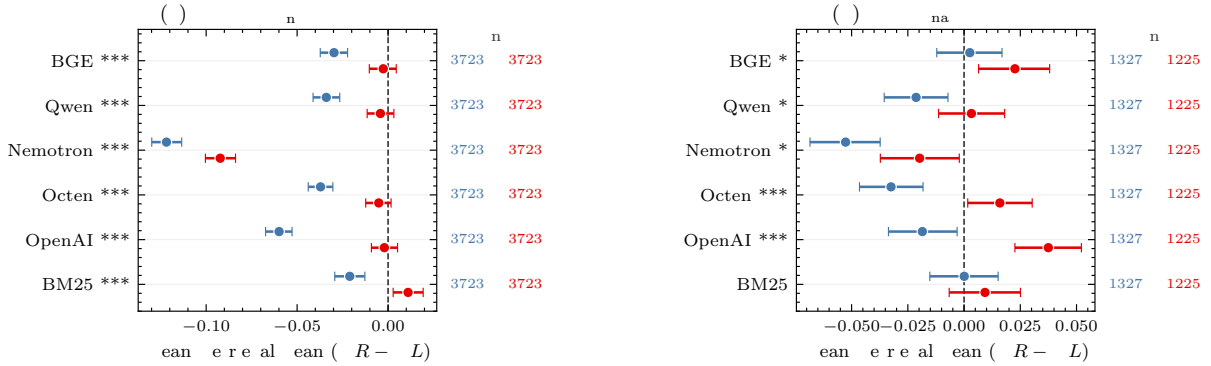

Figure 9: Per-retriever mean retrieval lean over the *full* top-10 (no relevance filter), 95% bootstrap CIs. **(a)** AllSides$_{synth}$ left/right queries; **(b)** Reddit$_{nat}$ liberal-/conservative-asker queries. Stars on retriever labels ($*p<0.05$, $**p<0.01$, $***p<0.001$); n per group in the right margin. The own-lean gap (right marker right of left) holds without any filtering.

to Reddit$_{nat}$ weakens it: only three of six retrievers stay significant (Fig. 11).

Why the full-top-10 Reddit$_{nat}$ gap is diluted, not absent. The datasets differ sharply in how much of the top-10 is on-topic (Fig. 12). On AllSides$_{synth}$, 83% of (query, retriever) top-10 lists contain ≥ 1 relevant article (median 4 among those), because the synthetic queries are written from the corpus; on Reddit$_{nat}$, only 40% do (median 2). A full-top-10 mean on Reddit$_{nat}$ therefore averages over a large majority of off-topic articles—which carry no consistent lean—so the relevant-subset metric is the appropriate one there. BM25 is the extreme case of this dilution: its gold retention is the lowest of any retriever (3.5%; App. F), so its full top-10 carries no consistent lean (Fig. 9b) even though its gap over relevant documents is the largest of any retriever (Table 6). The converse holds on AllSides$_{synth}$, where queries are written from the corpus and overlap is high; even there, shared topical vocabulary (e.g., "CBO", "$10.10"; Table 1) dominates the BM25 score and matches articles of both leans within a roundup, leaving the few partisan terms to shift the top-10 only modestly.

Robustness to the relevance judge. On AllSides$_{synth}$ we have editorial ground truth (same-roundup membership), letting us measure how far the LLM judge departs from it (Fig. 13). The two-judge gold is much broader than the editorial set: it recovers 70% of same-roundup articles but also flags many other on-topic ones, for low agreement (Cohen's $\kappa = 0.17$). Despite this, the own-lean gap is significant under *both* definitions (editorial: Table 7; LLM-judge: Fig. 10), so it does not hinge on the exact relevance

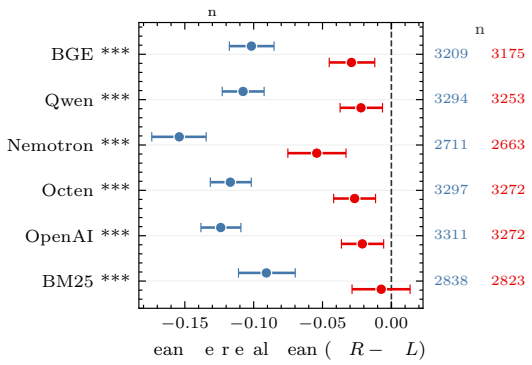

Figure 10: AllSides$_{synth}$ per-retriever lean over LLM-judged-relevant top-10 docs (the Reddit$_{nat}$ setting applied to AllSides$_{synth}$): the own-lean gap holds on all six retrievers ($p < 10^{-10}$), larger than under the editorial gold (Table 7). 95% bootstrap CIs; n = retained left/right queries.

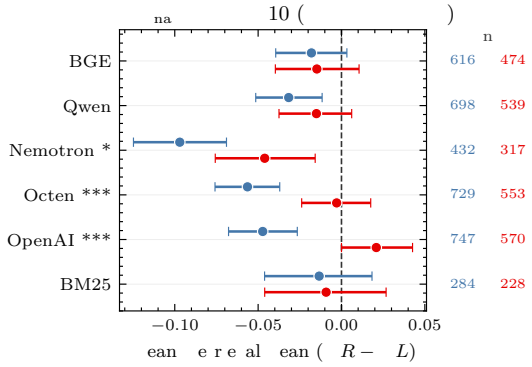

Figure 11: Reddit$_{nat}$ per-retriever lean over the *full* top-10, restricted to queries with ≥ 1 relevant doc (the AllSides$_{synth}$ setting applied to Reddit$_{nat}$). Only three of six retrievers stay significant—the off-topic majority of the top-10 dilutes the signal (Fig. 12).

criterion.

H Significance tests

The own-lean and dialect contrasts are tested as follows; per-retriever p-values are in App. I (political) and App. M (dialect).

Paired contrasts (AllSides$_{synth}$, HCMagic$_{synth}$, translation). On the paired datasets we test the mean per-pair difference (Δ_{lean} or Δ_{rank}) against a null of no preference in two ways. The *sign-flip permutation* (Pitman, 1937) randomly flips the sign of each per-pair difference on B=50,000 resamples and asks how often the permuted mean is at least as extreme as the observed mean (two-sided; p-floor $1/B = 2{\cdot}10^{-5}$). The two-sided *paired Wilcoxon signed-rank* test (Wilcoxon, 1945) asks whether the per-pair difference has a non-zero median across

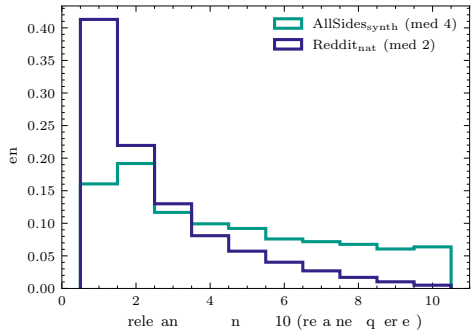

Figure 12: Number of relevant docs in the top-10 per (query, retriever), among queries with ≥ 1 relevant. AllSides$_{synth}$ lists are far more on-topic than Reddit$_{nat}$'s (83% vs. 40% of lists contain any relevant doc).

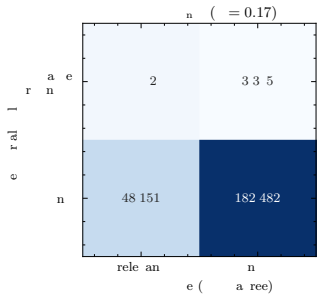

Figure 13: AllSides$_{synth}$ LLM-judge relevance (both judges agree) vs. editorial same-roundup gold, over all 242,000 judged top-10 pairs. The judge is broader than the editorial set ($\kappa = 0.17$), yet the own-lean gap holds under both.

pairs, without distributional assumptions.

Unpaired contrast (Reddit$_{nat}$, HCMagic$_{nat}$). The naturalistic datasets are unpaired (different queries in each group), so we use a two-sided rank-based *Mann–Whitney U* test (Mann and Whitney, 1947), asking whether one group's per-query lean (or MRR) is stochastically larger than the other's.

I Per-retriever gold-filtered own-lean tendency

On AllSides$_{synth}$, the per-retriever paired $\overline{\Delta_{\text{lean}}} = \text{lean}(R) - \text{lean}(L)$ at k=10 on the fair corpus, restricted to (article, frame) pairs where both the left and the right query surface a same-story counterpart in top-10 (the gold set, Table 7). This is the per-retriever difference behind Fig. 2a, which plots the left- and right-query lean separately. Sign-flip permutation (B=50,000, p-floor $2{\cdot}10^{-5}$) and paired Wilcoxon signed-rank both reject on every retriever.

Retriever	n pairs	$\overline{\Delta_{\text{lean}}}$ (R−L)	95% CI	Wilcoxon p	Perm p
BGE-large	2,422	+0.027	[+0.016, +0.038]	$4.4{\cdot}10^{-6}$	$< 2{\cdot}10^{-5}$
Qwen3-Emb-8B	2,508	+0.023	[+0.014, +0.033]	$2.0{\cdot}10^{-6}$	$< 2{\cdot}10^{-5}$
Llama-Nemotron-8B	1,416	+0.025	[+0.011, +0.038]	$9.5{\cdot}10^{-5}$	$2.8{\cdot}10^{-4}$
Octen-Emb-8B	2,605	+0.023	[+0.014, +0.032]	$1.0{\cdot}10^{-5}$	$< 2{\cdot}10^{-5}$
OpenAI `te-3-large`	2,663	**+0.045**	[+0.035, +0.054]	$3.9{\cdot}10^{-21}$	$< 2{\cdot}10^{-5}$
BM25 (Okapi)	1,798	**+0.043**	[+0.029, +0.057]	$1.4{\cdot}10^{-8}$	$< 2{\cdot}10^{-5}$

Table 7: AllSides$_{synth}$ gold-filtered paired contrast at k=10; positive $\overline{\Delta_{\text{lean}}}$ means each query retrieves toward its own political lean. Both paired tests reject on every retriever: two-sided paired Wilcoxon and sign-flip permutation (B=50,000, p-floor $2{\cdot}10^{-5}$; App. H).

Queries on Reddit$_{nat}$ are unpaired, so the parallel per-retriever contrast is the *own-lean gap*: the mean gold-filtered lean over conservative-asker queries minus that over liberal-asker queries (Table 8), the difference behind Fig. 2b. A two-sided Mann–Whitney U test rejects on every retriever ($p < 0.05$).

J Political k-sweep companions

Companion k-sweep figures to Fig. 2. The L–R separation (a) and the asker-ideology right−left gap (b) are stable across $k \in \{5, 10, 20, 50, 100\}$ in both AllSides$_{synth}$ and Reddit$_{nat}$.

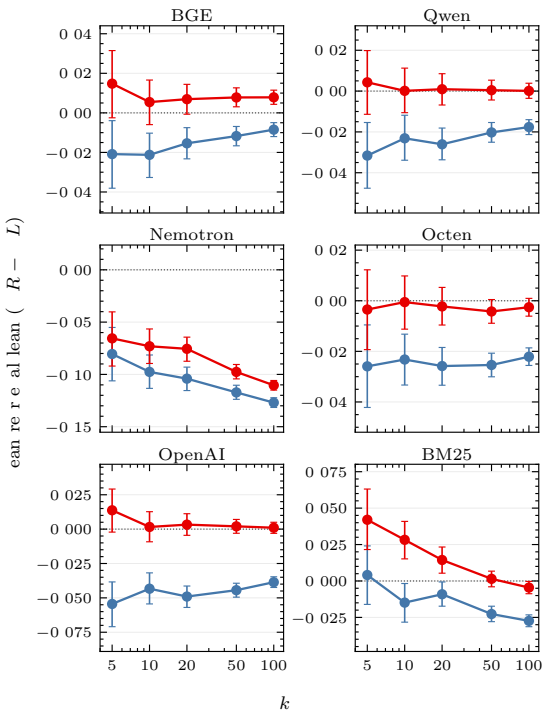

Figure 14: AllSides$_{synth}$ gold-filtered paired lean across $k \in \{5, 10, 20, 50, 100\}$, per retriever (per-panel y-axis). Whiskers: 95% cluster-bootstrap CIs.

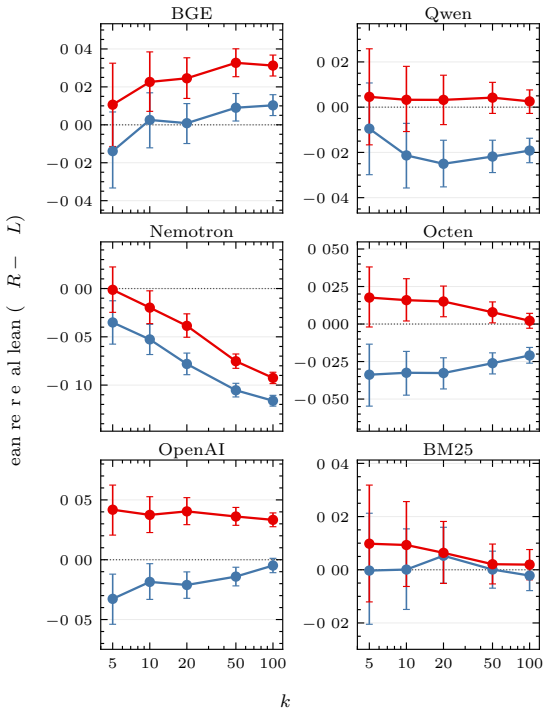

Figure 15: Reddit$_{nat}$ mean retrieval lean by asker ideology across $k \in \{5, 10, 20, 50, 100\}$, per retriever (per-panel y-axis; raw unfiltered, all retrievals). Whiskers: 95% cluster-bootstrap CIs.

K HealthCareMagic corpus example

Table 9 gives a worked example from each dataset (queries from Table 2), with the real doctor response each query is scored against—indexed from the originating HealthCareMagic row. HCMagic$_{synth}$ is paired: a real WME query and its synthetic AAL rewrite share one gold document. HCMagic$_{nat}$ is unpaired, so it contributes a real AAL post and a separate real WME post, each a different patient with its own gold. All are ranked against the same 112,165-document corpus.

Retriever	n (lib/con)	gap (con−lib)	95% CI	MWU p
BGE-large	616/474	+0.109	[+0.028, +0.193]	$7.1 \cdot 10^{-3}$
Qwen3-Emb-8B	698/539	+0.069	[−0.002, +0.139]	$4.7 \cdot 10^{-2}$
Llama-Nemotron-8B	432/317	**+0.186**	[+0.079, +0.297]	$1.1 \cdot 10^{-3}$
Octen-Emb-8B	729/553	+0.103	[+0.037, +0.176]	$1.5 \cdot 10^{-3}$
OpenAI `te-3-large`	747/570	+0.136	[+0.064, +0.205]	$7.7 \cdot 10^{-5}$
BM25 (Okapi)	284/228	**+0.195**	[+0.062, +0.323]	$5.5 \cdot 10^{-3}$

Table 8: Reddit$_{nat}$ gold-filtered own-lean gap at k=10, parallel to Table 7: mean lean over conservative-asker queries minus mean lean over liberal-asker queries, computed over LLM-judged-relevant documents. Positive means each asker retrieves toward their own political lean. n is the retained liberal/conservative query count; 95% bootstrap CI on the gap; two-sided Mann–Whitney U p (the test behind Fig. 2b). Qwen3-Emb-8B's gap CI marginally includes 0 although its Mann–Whitney $p < 0.05$.

L HCMagic input filter and dropped rows

HCMagic$_{nat}$ (naturalistic). The hcmagic source contains 1,315 rows (AAL: 656; WME: 659). The loader applies two filters before embedding/retrieval, both motivated by inspection of the raw query field: (i) drop rows with missing/blank query text (otherwise pandas serialises NaN as the literal string `"nan"` and the embedding step proceeds on that junk text); (ii) drop rows whose query is shorter than 20 characters after stripping whitespace, since the bottom of the AAL length distribution is dominated by one-word gibberish/typos (e.g. "thtr", "aesthema", "zzxczxc") and one three-word fragment ("i wana pregnent") that do not function as questions. The 20-char threshold preserves real-but-brief medical questions such as "bipolar type 2 can be cure?" (27 chars). Nine AAL rows are dropped total—one by filter (i) and eight by filter (ii)—yielding a final AAL count of 647. WME has zero rows dropped (minimum length 119 chars). The same nine rows are excluded from trans-AAL via the paired-on-row join, so the paired contrast runs on n=647 pairs.

HCMagic$_{synth}$ (synthetic). The HCMagic$_{synth}$ source contains 5,000 rows. The same 20-character minimum filter is applied to all four query variants (`input` plus three `aal_message{1,2,3}` paraphrases) and the `output` field; one row is dropped (all three AAL paraphrase fields empty), yielding n=4,999 paired samples used in Table 11 and Fig. 3a.

M Per-retriever dialect gap

The per-retriever contrasts behind Fig. 3, in the same format as the political tables (App. I). Positive values mean the WME-form query retrieves better—its answer document ranks higher—than the AAL form. HCMagic$_{synth}$ is *paired*, summarised by the mean per-sample $\Delta_{\text{rank}} = \text{RR}_{\text{WME}} - \text{RR}_{\text{AAL}}$ over reciprocal rank and tested with paired Wilcoxon and sign-flip permutation; HCMagic$_{nat}$ is an *unpaired* between-group MRR gap on naturally-occurring queries (different patients), tested with a two-sided Mann–Whitney U. All at k=10 on the unified 112,165-document corpus, no relevance filter; 95% CIs are bootstrap. No multiple-comparison correction; the BGE HCMagic$_{nat}$ entry (p=0.10) is the only contrast that does not clear $p < 0.05$.

N Dialect k-sweep companions

Companion k-sweep figures to Fig. 3. The WME>AAL gap is stable across $k \in \{1, 5, 10, 20, 50, 100\}$ in both the paired HCMagic$_{synth}$ and the unpaired HCMagic$_{nat}$ settings.

Translation k-sweep. Companion to Fig. 19: paired Recall@k across $k \in \{1, 5, 10, 20, 50, 100\}$ for GPT-translated WME vs the original AAL query on the same n=647 HCMagic$_{nat}$ AAL queries (Fig. 18). The two-camp split visible in the main forest (Qwen / Octen / OpenAI `te-3-large` overlap; BGE / Nemotron / BM25 separate) persists at every swept k.

O HCMagic$_{synth}$ augmentation quality

The HCMagic$_{synth}$ synthetic AAL paraphrases are generated by the Ziems et al. (2022) / Deas et al. (2024) pipelines applied to HealthCareMagic patient queries. Both pipelines were validated with AAL-speaker reviewers on their own source corpora (TwitterAAL and LiveQA respectively) but were not independently re-validated on HealthCareMagic outputs. We report a quantitative + qualitative spot-check below.

HCMagic$_{synth}$ — real WME query, synthetic AAL rewrite, shared real gold document	
AAL query *(synthetic)*	*Hi docta, how might can that renal functions be monitored in one patient undergoing hemodialysis 3 times per week?*
WME query *(real)*	*Hi doctor, how can the renal function be monitored in a patient undergoing hemodialysis 3 times per week?*
Gold document *(real)*	Renal function is monitored by maintaining a chart of blood urea and serum creatinine levels. . . during each visit of dialysis you must. . . note down the. . . levels of renal parameters. . . regularly. . . show it to the nephrologist. . . Take care.
HCMagic$_{nat}$ — real AAL post and a separate real WME post, each with its own gold (unpaired)	
AAL query *(real)*	*. . . about 2 weeks ago she notice she had a lump just below her left rib in the front. . . once ahe straight again u can see it and feel it. What can this be.*
Gold document *(real)*	Hi, Dear, thanks for the query to Chat Doctor. . . In my opinion your daughter has a third nipple or a Supernumerary nipple below the normal nipple on left side, or may be a puberty change. . . Have a good time.
WME query *(real)*	*Every now and then I will feel my chest tighten and my heart starts pounding extremely hard. I cant breath and I begin feeling very lightheaded. It usually only lasts a few seconds and does not happen too often. Any ideas as to what could be causing this?*
Gold document *(real)*	Hello and thanks for writing. . . The symptoms you have described are somewhat red flags and need a thorough workup to get to a cause. . . conditions that might cause them include cardiac arrhythmia, angina, anxiety or Blood Pressure issues. . . You may also need few tests like BP measurement, CBC, FT, ECG, Holder study. . . I suggest you consult a physician and get yourself worked up for the problem.

Table 9: Worked query–document examples for both dialect datasets (the queries from Table 2), each scored against a *real* gold document—the doctor response indexed from its originating HealthCareMagic row (abridged with . . .). HCMagic$_{synth}$ is paired: a *real* WME query and its *synthetic* AAL rewrite share one gold document. HCMagic$_{nat}$ is unpaired, so we show a *real* AAL post and a separate *real* WME post—different patients, each with its own gold document (§3.1). AAL surface markers in purple.

idx	char len	query (first 30 chars)
70	7	`zzxczxc`
144	9	`lipoprint`
267	7	`nothing`
386	4	`thtr`
398	12	`bhjkhjkhklhk`
456	8	`aesthema`
495	15	`i wana pregnent`
511	0	`(empty)`
569	4	`SDfc`

Table 10: The nine AAL rows dropped from the hcmagic source by the input-validity filter.

Quantitative QC (all 14,997 paraphrases). Median length ratio AAL/WME $= 1.06$; median character-level edit ratio (vs source WME) $= 0.58$ (5th percentile $= 0.20$, so even the least-shifted paraphrase contains substantial surface change); inter-paraphrase diversity (avg pairwise edit ratio across the three paraphrases per sample, n=500-sample subset) median $= 0.58$. Near-WME-duplicates (edit ratio < 0.05) total $19/14{,}997$ (0.13%); identical-to-WME $2/14{,}997$.

Qualitative observations. Random samples contain recognisable AAL features (phonological respellings, habitual aspect "I been had", "y'all", "might can" double modal) but also systematic generator artifacts: over-frequent cleft structures ("It is X that Y"), substitution of "for me" for "my", mid-sentence pronoun drift ("she/he" replacing first-person "I"), and occasional sub-word garbling ("Shems" for "I'm", "mat" for "might").

Example. Source WME: *"Hi I have a problem with my gum. Have a rednes on the left two front toths for about a year. have been at the dentist have my gums clean have an x ray atc. no one can help me. I also have venires done on my top. Can you please let me know what can cose infection."* AAL paraphrase 1: *"Hi **She have** a problem with **for me** gum. Do Have a rednes on the left two front toth for abat a year. have be at that dentist have my gum clean have an x ra atc. no one **mat** can help me. **It is venires done on my top that i also has.** Can **yaw** please let **she** know can what cos infection."* (Bolded blue: real AAL feature `yaw` for *y'all*; generator artifacts `She have` pronoun drift, `for me` for *my*, `mat can` as attempted double-modal *might can*, and cleft restructure.) These artifacts likely inflate the HCMagic$_{synth}$ Δ_{rank} above what natural-AAL

Retriever	n pairs	$\overline{\Delta_{\text{rank}}}$ (WME−AAL)	95% CI	Wilcoxon p	Perm p
BGE-large	4,999	**+0.050**	[+0.045, +0.055]	$1.1 \cdot 10^{-116}$	$< 2 \cdot 10^{-5}$
Qwen3-Emb-8B	4,999	+0.023	[+0.018, +0.027]	$1.8 \cdot 10^{-24}$	$< 2 \cdot 10^{-5}$
Llama-Nemotron-8B	4,999	**+0.051**	[+0.046, +0.056]	$2.6 \cdot 10^{-192}$	$< 2 \cdot 10^{-5}$
Octen-Emb-8B	4,999	+0.027	[+0.022, +0.031]	$1.3 \cdot 10^{-24}$	$< 2 \cdot 10^{-5}$
OpenAI `te-3-large`	4,999	+0.040	[+0.035, +0.044]	$1.1 \cdot 10^{-96}$	$< 2 \cdot 10^{-5}$
BM25 (Okapi)	4,999	+0.024	[+0.020, +0.028]	$8.3 \cdot 10^{-50}$	$< 2 \cdot 10^{-5}$

Table 11: HCMagic$_{synth}$ paired dialect gap at k=10 (the contrast behind Fig. 3a): mean per-sample $\Delta_{\text{rank}} = \text{RR}_{\text{WME}} - \text{RR}_{\text{AAL}}$, averaging the three synthetic AAL rewrites per WME source. Both paired tests (Wilcoxon, sign-flip permutation; App. H) reject on every retriever. Parallel to the AllSides$_{synth}$ Table 7.

Retriever	n (WME/AAL)	MRR gap (WME−AAL)	95% CI	MWU p
BGE-large	659/647	+0.034	[−0.010, +0.077]	0.10
Qwen3-Emb-8B	659/647	**+0.077**	[+0.028, +0.125]	$2.3 \cdot 10^{-3}$
Llama-Nemotron-8B	659/647	**+0.091**	[+0.056, +0.125]	$8.7 \cdot 10^{-8}$
Octen-Emb-8B	659/647	+0.062	[+0.015, +0.111]	0.01
OpenAI `te-3-large`	659/647	+0.064	[+0.017, +0.111]	0.01
BM25 (Okapi)	659/647	+0.040	[+0.015, +0.065]	0.03

Table 12: HCMagic$_{nat}$ unpaired dialect gap at k=10 (the contrast behind Fig. 3b): difference in mean MRR between the naturally-occurring WME and AAL query pools. Parallel to the Reddit$_{nat}$ Table 8; bold marks the two largest gaps. BGE's CI includes 0, consistent with its p=0.10.

features alone would induce, particularly for out-of-distribution-sensitive strong encoders—the translation contrast (App. P) bounds the natural-AAL dialect penalty for the strong encoders at ± 0.015 MRR.

P Translation-mitigation experiment

A natural deployment fix for the AAL gap is to normalize the query into WME—the variety these retrievers handle best—before retrieving. We test this with `gpt-5-mini`: each of the n=647 HCMagic$_{nat}$ AAL queries is back-translated to WME at T=0 and retrieved with the WME version. Because both forms share the same gold document, the contrast is paired (per-query $\Delta_{\text{rank}} = \text{RR}_{\text{trans-WME}} - \text{RR}_{\text{AAL}}$; paired Wilcoxon and sign-flip permutation). LLM translation has no content-fidelity guarantee and may add or drop medical content, so we report it as a practical intervention, not a clean dialect contrast; prior work also finds that LLMs struggle to rewrite and interpret AAL (Deas et al., 2023; McKeown et al., 2026), making this a stringent test of the fix.

Back-translation splits the retrievers cleanly (Fig. 19, Table 13): BGE-large, Llama-Nemotron-8B, and BM25 improve measurably ($p \approx 0$), while Qwen3-Emb-8B, Octen-Emb-8B, and OpenAI `te-3-large` do not change ($|\overline{\Delta_{\text{rank}}}| \leq 0.005$, all $p \geq 0.22$). The retrievers that already handle AAL best benefit least, and—apart from BGE, which had no significant gap to begin with—no retriever closes the gap. A lexical-surface rewrite thus helps only the most dialect-sensitive retrievers and does not remove the disparity.

Q Bias-gap method details

Tokenization for lex_q. Tokenization is a lowercase alphanumeric regex split with apostrophe-suffixes kept attached (so "don't" and "she's" are single tokens, punctuation dropped); we deliberately use a non-model tokenizer so the lex_q channel is defined on the raw query text rather than on any retriever's BPE/wordpiece.

Robust standard errors. We use robust standard errors—specifically HC1 heteroskedasticity-consistent, with cluster-robust adjustment within source article on AllSides$_{synth}$ and within source sample on HCMagic$_{synth}$; Reddit$_{nat}$ and HCMagic$_{nat}$ are unpaired so there is no within-cluster structure. The qualitative pattern in §4.2 holds under HC0–HC3.

R Lexical-residual regressions

Per-retriever coefficients for the lexical-residual decomposition of §4 on all four datasets: the raw gap, the residual after partialling out the lexical channel ($\hat{\alpha}$ for the paired datasets, $\hat{\gamma}$ for the unpaired), and the lex slope $\hat{\beta}$.

Retriever	n pairs	$\overline{\Delta_{\text{rank}}}$ (trans-WME−AAL)	95% CI	Wilcoxon p	Perm p
BGE-large	647	$+\mathbf{0.032}$	$[+0.015, +0.048]$	$1.1{\cdot}10^{-7}$	$2.0{\cdot}10^{-4}$
Qwen3-Emb-8B	647	-0.005	$[-0.020, +0.010]$	0.22	0.50
Llama-Nemotron-8B	647	$+\mathbf{0.044}$	$[+0.027, +0.061]$	$8.5{\cdot}10^{-28}$	$< 2{\cdot}10^{-5}$
Octen-Emb-8B	647	-0.004	$[-0.019, +0.012]$	0.43	0.64
OpenAI `te-3-large`	647	-0.000	$[-0.014, +0.014]$	0.89	0.99
BM25 (Okapi)	647	$+0.023$	$[+0.010, +0.037]$	$1.2{\cdot}10^{-9}$	$3.8{\cdot}10^{-4}$

Table 13: Translation-mitigation paired gap at k=10: each HCMagic$_{nat}$ AAL query back-translated to WME via `gpt-5-mini`, then $\Delta_{\text{rank}} = \text{RR}_{\text{trans-WME}} - \text{RR}_{\text{AAL}}$. Translation helps the dialect-sensitive retrievers (BGE, Llama-Nemotron, BM25: positive and significant) but not the rest (Qwen, Octen, OpenAI `te-3-large`: CIs span 0), and closes no gap fully.

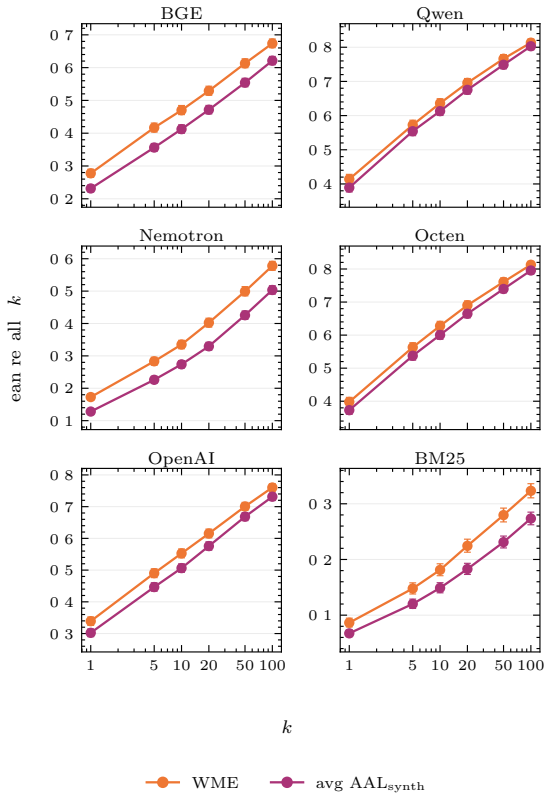

Figure 16: HCMagic$_{synth}$ mean Recall@k across $k \in \{1, 5, 10, 20, 50, 100\}$, per retriever (per-panel y-zoom to make CIs visible at this n). Whiskers: 95% bootstrap CIs.

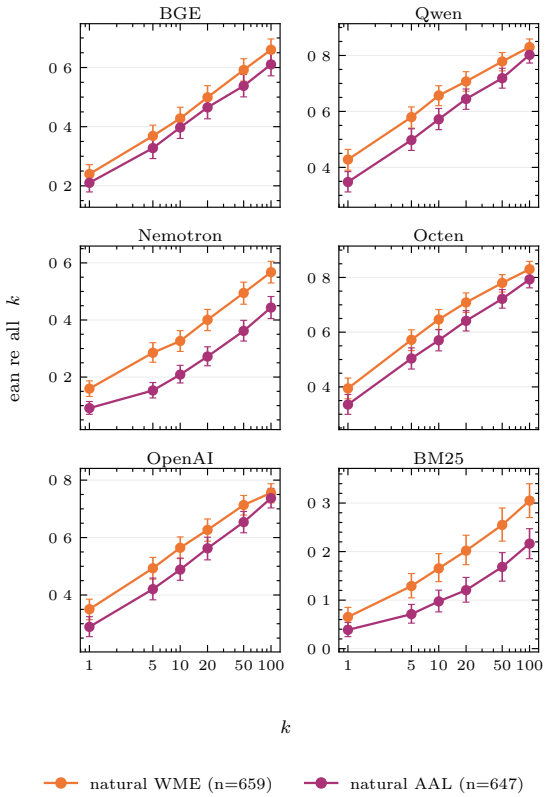

Figure 17: HCMagic$_{nat}$ unpaired mean Recall@k across $k \in \{1, 5, 10, 20, 50, 100\}$. Whiskers: 95% bootstrap CIs.

AllSides$_{synth}$ (paired). Per-encoder OLS $\Delta_{\text{lean}} = \alpha + \beta\,(\text{lex}_R - \text{lex}_L)_c$ on each retriever's gold-filtered subset (cluster-robust SE on `article_id`), where $(\text{lex}_R - \text{lex}_L)_c$ is the Monroe–Colaresi–Quinn log-odds asymmetry between the right and left query of pair c (mean ≈ -28, SD 26). OpenAI `te-3-large` is the only retriever with a measurable slope; BM25 and the four open-weights have β indistinguishable from zero (Table 14). After partialling out the lex channel, OpenAI `te-3-large`'s residual α sits inside the open-weights' residual band, while BM25's residual has larger magnitude than every open-weights residual.

Reddit$_{nat}$ (unpaired). Per-encoder unpaired between-group regression at $k = 20$ on Reddit$_{nat}$ (n_L = 1,327 liberal-asker + n_R = 1,225 conservative-asker queries; centrist dropped): $\text{lean}_q = \alpha + \beta\,\text{lex}_q + \gamma\,\mathbb{1}[q \in \text{right}] + \varepsilon$ with HC1 robust SE. The fit is over the full unfiltered top-k, so the raw gap here differs from the gold-filtered k=10 gaps of Fig. 2b. $\hat{\gamma}$ is the right−left lean gap after partialling out per-query MCQ asym-

Retriever	n	$\overline{\Delta_{\text{lean}}}$	residual α	slope β (p)
BM25 (Okapi)	1,798	+0.043	+0.038	≈ 0 (p=0.56)
BGE-large	2,422	+0.027	+0.027	≈ 0 (p=0.98)
Qwen3-Emb-8B	2,508	+0.023	+0.031	$+2.9\cdot10^{-4}$ (p=0.11)
Llama-Nemotron-8B	1,416	+0.025	+0.025	≈ 0 (p=0.94)
Octen-Emb-8B	2,605	+0.023	+0.018	≈ 0 (p=0.36)
OpenAI `te-3-large`	2,663	+0.045	+0.029	$\mathbf{-5.7\cdot10^{-4}}$ (**p=0.002**)

Table 14: Gold-filtered lex-residual fit. Each retriever's regression runs on its own gold-filtered subset (so n varies; cross-retriever α-comparison partly confounds with subset selection). OpenAI `te-3-large` is the only retriever with a non-zero β; the lex channel accounts for $\overline{\Delta_{\text{lean}}} - \hat{\alpha} = +0.016$ of its gap at the mean.

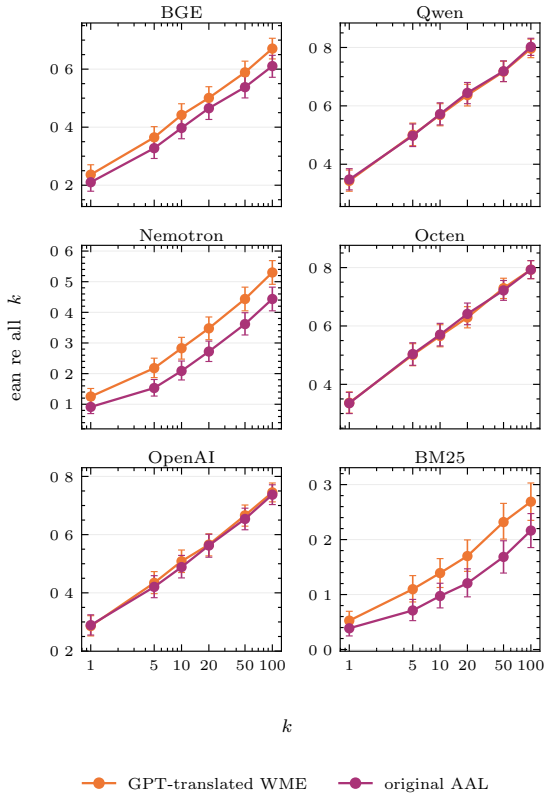

Figure 18: Translation-mitigation Recall@k across $k \in \{1, 5, 10, 20, 50, 100\}$ per retriever on the n=647 paired AAL queries. Whiskers: 95% bootstrap CIs. Per-panel y-zoom to make CIs visible.

metry between left and right post-text vocabularies; $\hat{\beta}$ is the within-group lex slope. All five dense retrievers have positive significant $\hat{\beta}$ (organic Reddit vocabulary is far more lexically partisan than AllSides$_{synth}$ query pairs); BGE's residual gap collapses to zero, Llama / Octen / OpenAI `te-3-large` retain a significant non-lexical component, Qwen3 marginal.

Reddit$_{nat}$, gold-filtered outcome (robustness). The fit above uses the unfiltered top-k; as a robustness check we repeat it with the outcome of

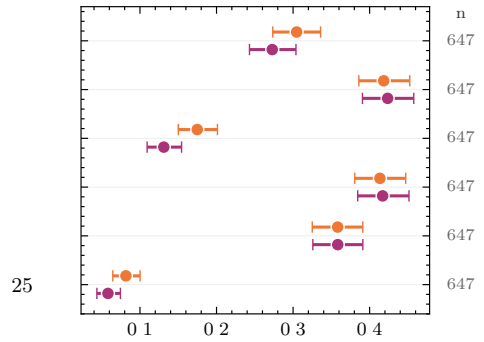

Figure 19: Translation-mitigation paired MRR on the n=647 HCMagic$_{nat}$ AAL queries: original AAL vs. `gpt-5-mini` back-translation to WME. Same content (modulo translation noise), tested with paired Wilcoxon. Translation k-sweep in App. N.

Fig. 2b: per-query lean over the judged-relevant articles in the top-10 (both judges grade relevant, App. F), keeping queries with at least one relevant article. This decomposes the same quantity the main test reports, at the cost of smaller per-retriever n (gold retention), a coarser per-query outcome (few relevant articles per query), and conditioning on judged relevance. The raw gaps reproduce the gold-filtered gaps of App. I, and the conclusions of the unfiltered fit carry over (Table 16): BGE and Qwen3 show no detectable residual, while Llama-Nemotron and OpenAI `te-3-large` retain large significant residuals ($\hat{\gamma}$/raw ≈ 0.8); Octen is positive but not significant. BM25, which has no unfiltered gap, here carries the largest raw gap, and its residual is large but marginal (p=0.07). $\hat{\beta}$ is significant for BGE ($p<10^{-4}$), Qwen3 (p=0.008), and Octen (p=0.03), and not for the others.

HCMagic$_{synth}$ (paired). Per-encoder OLS $\Delta_{\text{rank}} = \alpha + \beta\,\text{asym}_r + \varepsilon_r$ on all 4,999 paired samples with HC1 robust SE, where $\text{asym}_r = \text{lex}_{\text{WME}_r} - \overline{\text{lex}}_{\text{AAL}_{\text{synth}_r}}$ is the per-sample Monroe–Colaresi–Quinn log-odds asymmetry

Retriever	raw R−L gap	$\hat{\gamma}$ (p)	$\hat{\gamma}$/raw	$\hat{\beta}$ (p)
BGE-large	+0.024	+0.007 (p=0.42)	0.31	positive sig.
Qwen3-Emb-8B	+0.028	+0.015 (p=0.08)	0.55	positive sig.
Llama-Nemotron-8B	+0.039	**+0.022 (p=0.028)**	0.55	positive sig.
Octen-Emb-8B	+0.048	**+0.033 ($\mathbf{p < 10^{-3}}$)**	0.70	positive sig.
OpenAI `te-3-large`	+0.062	**+0.036 ($\mathbf{p < 10^{-3}}$)**	0.58	positive sig.
BM25 (Okapi)	+0.001	+0.002 (NS)	—	NS

Table 15: Reddit$_{nat}$ unpaired lex-residual at $k = 20$. Bolded $\hat{\gamma}$ are detectable encoder-level residual gaps after partialling out per-query MCQ asymmetry; BGE's collapse is consistent with topic-asymmetry exposure. $\hat{\beta}$ is positive and significant ($p < 0.05$) for all five dense retrievers, contrasting with AllSides$_{synth}$ where only OpenAI `te-3-large` had a detectable slope.

Retriever	n (lib/con)	raw R−L gap	$\hat{\gamma}$ (p)	$\hat{\gamma}$/raw
BGE-large	616/474	+0.109	+0.020 (p=0.67)	0.18
Qwen3-Emb-8B	698/539	+0.069	+0.017 (p=0.69)	0.24
Llama-Nemotron-8B	432/317	+0.186	**+0.148 (p=0.017)**	0.79
Octen-Emb-8B	729/553	+0.103	+0.061 (p=0.13)	0.59
OpenAI `te-3-large`	747/570	+0.136	**+0.106 (p=0.009)**	0.78
BM25 (Okapi)	284/228	+0.195	+0.138 (p=0.07)	0.71

Table 16: Reddit$_{nat}$ lex-residual with the gold-filtered relevant-subset lean@10 of Fig. 2b as the outcome, on queries with ≥ 1 both-judge-relevant article. Bolded $\hat{\gamma}$ are detectable residual gaps after partialling out per-query MCQ asymmetry.

summed over tokens (mean $\overline{\text{asym}} = 574$, SD 370). ζ scores are computed against the union vocabulary of all 4,999 WME inputs ($\sim$ 425k tokens) vs. all 14,997 AAL paraphrases ($\sim$1.34M tokens) with the same informative-Dirichlet prior ($\alpha_0 = 0.01$) as the political fit (Table 14).

HCMagic$_{nat}$ (unpaired). Per-encoder unpaired between-group regression on HCMagic$_{nat}$ (647 natural AAL + 659 natural WME queries): $\text{MRR}_q = \alpha + \beta\,\text{lex}_q + \gamma \mathbb{1}[q \in \text{WME}] + \varepsilon$ with HC1 robust SE. $\hat{\gamma}$ is the WME−AAL gap after partialling out per-query MCQ asymmetry between AAL and WME post-text vocabularies. No retriever has a detectable within-group lex slope (smallest $\hat{\beta}$ $p = 0.15$); $\hat{\gamma}$ is detectable on five of six retrievers (BGE marginal at $p = 0.21$); the slight excess of $\hat{\gamma}$ over the raw gap on five of six arises because $\hat{\beta}$ is mildly negative while the WME pool has higher mean lex. Note the caveat in §4: AAL-distinctive tokens include content words (*she*, *her*, *periods*, *baby*), so lex-residual partials out token-frequency asymmetry but not content-difficulty.

S Probing Details

Extracting Hidden Representations. For encoder-only models, we use mean pooling on the outputs of each layer to represent hidden representations. We additionally apply standard scaling. OpenAI `te-3-large` is excluded from probing, as the API exposes no intermediate layers.

Probe Training. We train linear probes using `LogisticRegression` models from `sklearn` (Pedregosa et al., 2011). We use the default hyperparameters and set the maximum iterations to 5000 to allow for convergence.

Retriever	n	$\overline{\Delta_{\text{rank}}}$	residual α	slope β (p)	$\beta \cdot \overline{\text{asym}}$
Llama-Nemotron-8B	4,999	$+0.051$	$+0.046$	$+8.4 \cdot 10^{-6}$ (p=0.21)	$+0.005$
BGE-large	4,999	$+0.050$	$+0.047$	$+5.8 \cdot 10^{-6}$ (p=0.43)	$+0.003$
OpenAI `te-3-large`	4,999	$+0.040$	$+0.039$	$+5.9 \cdot 10^{-7}$ (p=0.93)	$+0.000$
Octen-Emb-8B	4,999	$+0.027$	$+0.024$	$+4.8 \cdot 10^{-6}$ (p=0.48)	$+0.003$
BM25 (Okapi)	4,999	$+0.024$	$+0.024$	$+4.2 \cdot 10^{-7}$ (p=0.93)	$+0.000$
Qwen3-Emb-8B	4,999	$+0.023$	$+0.021$	$+2.7 \cdot 10^{-6}$ (p=0.68)	$+0.002$

Table 17: Per-sample lex-residual fit on HCMagic$_{synth}$ (sorted by raw $\overline{\Delta_{\text{rank}}}$). All $\hat{\beta}$ point estimates are positive (consistent direction) but none are statistically distinguishable from zero; the implied per-sample lex contribution $\beta \cdot \overline{\text{asym}}$ is at most 0.005 MRR (under 10% of the raw paired gap for the open-weights, ≈ 0 for OpenAI `te-3-large` and BM25). Residual α ranking matches the raw $\overline{\Delta_{\text{rank}}}$ ranking up to two micro-swaps within noise (BGE/Llama by 10^{-3}; BM25/Octen by $2 \cdot 10^{-6}$).

Retriever	raw WME$-$AAL gap	$\hat{\gamma}$ (p)	$\hat{\gamma}$/raw	$\hat{\beta}$ (p)
Llama-Nemotron-8B	$+0.091$	$\mathbf{+0.085}$ ($\mathbf{p=0.0018}$)	0.93	NS
Qwen3-Emb-8B	$+0.077$	$\mathbf{+0.110}$ ($\mathbf{p=0.0018}$)	1.42	NS
OpenAI `te-3-large`	$+0.064$	$\mathbf{+0.081}$ ($\mathbf{p=0.023}$)	1.26	NS
Octen-Emb-8B	$+0.062$	$\mathbf{+0.098}$ ($\mathbf{p=0.0049}$)	1.59	NS
BM25 (Okapi)	$+0.040$	$\mathbf{+0.041}$ ($\mathbf{p=0.043}$)	1.01	NS
BGE-large	$+0.034$	$+0.042$ (p=0.21)	1.21	NS

Table 18: HCMagic$_{nat}$ unpaired lex-residual. Bolded $\hat{\gamma}$ are significant at $p < 0.05$. The residual gap survives partialling on every retriever with $\hat{\gamma}/(\text{raw gap}) \in [0.93,\ 1.59]$; no retriever has a detectable within-group lex slope. Conclusion qualitatively matches the HCMagic$_{synth}$ result in Table 17.