Where Does Bias Enter RAG? A Causal Attribution Framework

Anonymous authors
Paper under double-blind review

Abstract

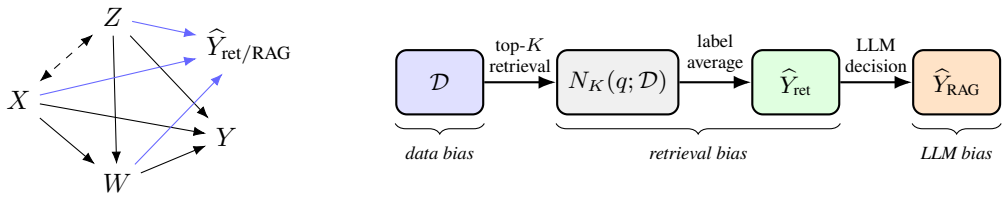
Large language models (LLMs) are increasingly used to support or automate consequential decisions, and such systems often leverage historical data through retrieval-augmented generation (RAG). Given a query case, similar past cases are retrieved and an LLM reasons over them before producing a decision. When the final decision exhibits a disparity between demographic groups, a fundamental question arises: is the disparity inherited from the historical data, induced by the retrieval algorithm, or introduced by the LLM's reasoning over the retrieved evidence? Existing bias audits measure the disparity of the deployed system as a whole, and are therefore unable to attribute it to individual components. In this paper, we make two contributions towards answering this question. First, we develop an attribution framework that decomposes the disparity of a RAG pipeline into the contributions of (a) the retrieval database, (b) the retrieval algorithm, and (c) the LLM aggregator, by modeling the pipeline as a sequence of predictors related to the same underlying causal model. This attribution method applies to any disparity measure defined on the score itself. Second, for causal fairness measures, we derive decomposition results in which each component's contribution is further attributed to the direct, indirect, and spurious causal pathways, yielding a granular bias analysis of the pipeline. We establish identification conditions for the required quantities, and construct estimators that accommodate free-text mediators such as clinical notes or loan descriptions. We demonstrate the framework in two real-world settings, emergency department triage and microfinance loan applications.

1 Introduction

Machine learning and artificial intelligence increasingly mediate consequential decisions in domains such as hiring, admissions, lending, law enforcement, and health care (Fabris et al., 2025; Das et al., 2023; Hager et al., 2024). A growing share of these systems is built on large language models (LLMs), which are used to screen and rank job applicants (An et al., 2024), summarize patient records (Van Veen et al., 2024), and draft or directly produce decisions. As their adoption expands, so does the concern that automated systems may perpetuate or amplify inequities between demographic groups. Recent studies have documented demographic disparities in LLM-based hiring and retrieval systems (An et al., 2024; Wilson & Caliskan, 2024), alongside demographic drift and occupational stereotypes in generative models (Luccioni et al., 2023; Naik & Nushi, 2023; Wang et al., 2024). Such disparities, however, do not originate purely from automated systems: the world from which data is sampled exhibits long-standing inequities, from gender pay gaps (Blau & Kahn, 1992; 2017) to racial bias in criminal sentencing (Sweeney & Haney, 1992; Pager, 2003), among others. Systems built on observational data therefore absorb patterns produced by historical processes, and this raises a fundamental question: under what conditions do AI systems replicate, exacerbate, or mitigate existing disparities?

When LLMs are used for decision-making, they are rarely deployed in isolation. A now-standard architecture is *retrieval-augmented generation* (RAG) (Lewis et al., 2020; Guu et al., 2020; Gao et al., 2023): given a query case, the system retrieves similar cases from a database, typically via dense embedding similarity (Karpukhin et al., 2020) or lexical matching (Robertson & Zaragoza, 2009). The retrieved cases are then inserted into the LLM's prompt, together with their historical labels (when present), and the LLM reasons over this evidence before producing a decision. Retrieval grounds the model in institutional precedent and is often viewed as a safeguard against hallucinated

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) SFM with predictors $\widehat{Y}$.

(b) The RAG decision pipeline analyzed in this paper.

Figure 1: (a) The Standard Fairness Model over protected attribute X, confounders Z, mediators W, and outcome Y, together with the pipeline predictors $\widehat{Y}$ (in blue); (b) the RAG pipeline as a sequence of predictors, with the bias sources analyzed in this paper.

outputs (Shuster et al., 2021). From a fairness standpoint, however, it introduces a new and largely unexamined challenge: the final decision is now the product of *three* distinct components, namely the historical database, the retrieval algorithm, and the LLM aggregator, each of which may inject, dampen, or amplify bias. To ground the discussion, we introduce a motivating example:

Example 1 (Emergency Department Triage) *A hospital deploys a RAG system to support triage in its emergency department (ED), using historical visits recorded in MIMIC-IV-ED (Johnson et al., 2023a). The database $\mathcal{D} = \{(X_i, Z_i, W_i, Y_i)\}_{i=1}^n$ contains past ED presentations, where X_i denotes the patient's race (Black x_0, White x_1), Z_i their sex and age, W_i a free-text triage note (recording arrival mode, chief complaint, vital signs, reported pain, etc.), and $Y_i = \mathbb{1}\{ESI_i \leq 2\}$, indicating whether the nurse-assigned Emergency Severity Index (Wuerz et al., 2000) marked the visit as requiring urgent care. When a new patient q presents, the system proceeds as follows:*

(1) the triage note W_q is embedded, and its cosine similarity $s_i(q)$ to the embedding of each note W_i in $\mathcal{D}$ is computed (Karpukhin et al., 2020),

(2) the K most similar past visits, together with their acuity labels Y_i, are retrieved,

(3) the retrieved cases are inserted into a prompt, and an LLM outputs a decision $\widehat{Y}_{RAG}$ for the new patient, given how similar patients were triaged in the past.

In the historical labels, White patients are marked urgent at much higher rates than Black patients,

$$TV(Y) = \mathbb{E}[Y \mid X = x_1] - \mathbb{E}[Y \mid X = x_0] = 10.2\%, \tag{1}$$

whereas the decisions of the deployed pipeline exhibit a disparity of

$$TV(\widehat{Y}_{RAG}) = \mathbb{E}[\widehat{Y}_{RAG} \mid X = x_1] - \mathbb{E}[\widehat{Y}_{RAG} \mid X = x_0] = 2.5\%. \tag{2}$$

The key question, which existing auditing approaches cannot answer, is the following: how much of the disparity in the final decision is (a) inherited from the historical labels in $\mathcal{D}$, (b) induced by the retrieval algorithm, and (c) introduced by the LLM reasoning over the retrieved cases? Furthermore, when disparities are measured causally, along which causal pathways (direct, indirect, or spurious) does each of these contributions operate? A single aggregate disparity measurement as in Eq. 2 collapses all of these mechanisms into one number, even though they call for entirely different interventions: curating or reweighting the database, modifying the retriever, or changing the aggregation model.

In this paper, we develop an attribution framework that answers this question. Our approach assumes that the historical data is generated by a structural causal model whose causal diagram is given by the Standard Fairness Model (SFM, Fig. 1a, Plečko & Bareinboim, 2024), and views the pipeline as a sequence of predictors defined on top of this data (Fig. 1b): the historical label Y, a purely retrieval-based predictor $\widehat{Y}_{\text{ret}}$ that averages the labels of retrieved neighbors, and the full RAG score $\widehat{Y}_{\text{RAG}}$ produced by the LLM. Applying disparity decompositions to each predictor, and examining the stepwise changes between consecutive stages of the pipeline, yields a component-level and pathway-level diagnosis of where bias enters. Specifically, our contributions are as follows:

(i) We formalize retrieval-augmented decision pipelines, and introduce an attribution framework that decomposes the disparity of the deployed system into contributions of the historical data, the retrieval step, and the LLM aggregation, applicable to any score-based disparity metric (Sec. 2),

(ii) For causally-defined disparity measures, we derive new decomposition results: Prop. 1 separates the disparity of the retrieval predictor into a data-inherited part and a retrieval-induced part, and Thm. 1 yields a three-way decomposition of the RAG disparity into data, retrieval, and LLM-processing contributions – each further attributed to direct, indirect, and spurious pathways, giving a 3×3 *bias signature* of the pipeline,

(iii) We discuss identification conditions and construct cross-fitted one-step estimators (Prop. 2) that accommodate high-dimensional text mediators, using fine-tuned language encoders for text-dependent nuisances (App. B),

(iv) We apply the framework to two real-world settings, emergency department triage on MIMIC-IV-ED (Johnson et al., 2023a) and microfinance loan funding on an online platform (Choo et al., 2014), providing a granular attribution analysis along pipeline components and causal pathways in each (Sec. 3).

Relationship to Previous Literature. Our work is related to the literature on causal fairness (Kusner et al., 2017; Kilbertus et al., 2017; Nabi & Shpitser, 2018; Zhang & Bareinboim, 2018; Chiappa, 2019; Plečko & Meinshausen, 2020; Plečko & Bareinboim, 2024; Plecko, 2026), which offers two major benefits. Causal notions provide human-interpretable metrics tied to the mechanisms transmitting change between groups, and a language aligned with legal notions of discrimination such as the disparate impact doctrine (Barocas & Selbst, 2016). Our contribution is to extend this analysis to *compositional* pipelines, in which multiple mechanisms (retrieval, aggregation) are stacked, and to attribute disparities across them. Secondly, our work relates to the growing literature on fairness of retrieval-augmented systems, which studies disparities introduced by retrieval, ranking, embedding, and generation components (Wu et al., 2025; Shrestha et al., 2024; Kim & Diaz, 2025; Kim et al., 2025; Hu et al., 2025). Such works do not model the underlying causal process, and consequently offer no formal attribution of the disparity, either to pipeline components or to causal pathways. Finally, statistical bias quantification and benchmarks for LLMs (Nadeem et al., 2021; Nangia et al., 2020; Liang et al., 2022; Smith et al., 2022) measure disparities of the model in isolation; our approach analyzes the model embedded in the retrieval pipeline through which it is deployed.

1.1 Preliminaries

We use the language of structural causal models (SCMs) as our basic semantical framework (Pearl, 2000). A structural causal model (SCM) is a tuple $\mathcal{M} := \langle V, U, \mathcal{F}, P(u) \rangle$, where V, U are sets of endogenous (observable) and exogenous (latent) variables respectively, $\mathcal{F}$ is a set of functions f_{V_i}, one for each $V_i \in V$, where $V_i \leftarrow f_{V_i}(\text{pa}(V_i), U_{V_i})$ for some $\text{pa}(V_i) \subseteq V$ and $U_{V_i} \subseteq U$. $P(u)$ is a strictly positive probability measure over U. Each SCM $\mathcal{M}$ is associated to a causal diagram $\mathcal{G}$ over the node set V, where $V_i \to V_j$ if V_i is an argument of f_{V_j}, and $V_i \dashleftarrow\dashrightarrow V_j$ if the corresponding U_{V_i}, U_{V_j} are not independent. An instantiation of the exogenous variables $U = u$ is called a *unit*. By $Y_x(u)$ we denote the potential response of Y when setting $X = x$ for the unit u, which is the solution for $Y(u)$ to the set of equations obtained by evaluating the unit u in the submodel $\mathcal{M}_x$, in which all equations in $\mathcal{F}$ associated with X are replaced by $X = x$.

Throughout, we work with the Standard Fairness Model (SFM) of Fig. 1a, where X denotes the protected attribute (taking values x_0, x_1), Z a possibly multidimensional set of confounders, W a set of mediators (e.g., a free-text document such as a loan description or a clinical vignette), and Y the outcome. Furthermore, we use existing notions of direct, indirect, and spurious effects (Zhang & Bareinboim, 2018; Plečko & Bareinboim, 2024). For a score S attached to the SFM (we take $S = Y$ for now, and consider other choices later on), these are defined as $x\text{-DE}_{x_0,x_1}(S \mid x) = \mathbb{E}(S_{x_1,W_{x_0}} \mid x) - \mathbb{E}(S_{x_0} \mid x)$, $x\text{-IE}_{x_1,x_0}(S \mid x) = \mathbb{E}(S_{x_1,W_{x_0}} \mid x) - \mathbb{E}(S_{x_1} \mid x)$, and $x\text{-SE}_{x_1,x_0}(S) = \mathbb{E}(S_{x_1} \mid x_0) - \mathbb{E}(S_{x_1} \mid x_1)$. Based on these, the total variation measure $\text{TV}_{x_0,x_1}(S) = \mathbb{E}[S \mid X = x_1] - \mathbb{E}[S \mid X = x_0]$ admits the decomposition (Zhang & Bareinboim, 2018):

$$\text{TV}_{x_0,x_1}(S) = x\text{-DE}_{x_0,x_1}(S \mid x_0) - x\text{-IE}_{x_1,x_0}(S \mid x_0) - x\text{-SE}_{x_1,x_0}(S). \tag{3}$$

The x-specific direct effect measures the effect of an $x_0 \to x_1$ transition along the direct path $X \to S$, for individuals with $X = x_0$, while keeping $X = x_0$ along the indirect path; the indirect effect captures the reverse transition $x_1 \to x_0$ along the path $X \to W \to S$; and the spurious effect quantifies how conditioning on $X = x_0$ compares to conditioning on $X = x_1$, for the potential outcome S_{x_1} where X is set to x_1 by intervention, thereby capturing variations propagated along back-door paths from X to S.

2 Causal Bias Attribution in RAG Pipelines

We begin by introducing the building blocks of retrieval-augmented decision pipelines, and mapping each of them onto the SFM. Let $\mathcal{M}$ be an SCM compatible with the SFM of Fig. 1a, and let

$$\mathcal{D} = \{(X_i, Z_i, W_i, Y_i) : i = 1, \dots, n\} \tag{4}$$

be the *retrieval database* of historical cases, drawn i.i.d. from the observational distribution of $\mathcal{M}$. A new query case $q = (X_q, Z_q, W_q)$ is an independent draw from the same population, for which a decision is required. The decision pipeline consists of three components:

(1) **Retrieval data:** the database $\mathcal{D}$, whose labels Y_i were produced by the (possibly biased) historical mechanism f_Y,

(2) **Retrieval algorithm:** an embedding map $e(\cdot)$ and a similarity metric $s_i(q) = \text{sim}\{e(W_q), e(W_i)\}$, which induce the retrieved neighborhood $N_K(q; \mathcal{D})$, defined as the set of indices of the K most similar admissible cases:

$$N_K(q; \mathcal{D}) \subseteq \mathcal{A}(q),\ |N_K(q; \mathcal{D})| = K, \quad \min_{i \in N_K(q;\mathcal{D})} s_i(q) \ \geq \max_{j \in \mathcal{A}(q) \setminus N_K(q;\mathcal{D})} s_j(q), \tag{5}$$

where $\mathcal{A}(q) \subseteq \{1, \dots, n\}$ is an optional *admissible set* restricting candidates to those matching q on selected covariates (e.g., same sex and age band), with ties broken arbitrarily,

(3) **LLM aggregator:** a prompt template $\rho(q, N_K)$ that inserts the query and the retrieved cases (with their labels) into a prompt, and an LLM that maps the prompt to a decision score.

Each stage of the pipeline may exhibit a disparity between groups. First, there is possibly a disparity in the historical labels Y_i themselves, encoding how past decisions treated the two groups. Then, in the second step, retrieval is performed, and the natural score at this stage is the *retrieval predictor*, which averages the outcomes of the retrieved neighbors,

$$\widehat{Y}_{\text{ret}}(q) = \frac{1}{K} \sum_{i \in N_K(q;\mathcal{D})} Y_i, \tag{6}$$

and may distribute decisions across groups differently than the historical labels do. Finally, in the third step, the LLM reasons over the retrieved cases and produces the *RAG predictor*,

$$\widehat{Y}_{\text{RAG}}(q) = \text{LLM}\{\rho(q, N_K(q; \mathcal{D}))\}, \tag{7}$$

which may again shift the disparity relative to the previous stage. We now define notions of *calibrated* retrieval and aggregation:

Definition 1 (Calibrated Retrieval and Aggregation) *The retrieval predictor $\widehat{Y}_{ret}$ is calibrated if*

$$\mathbb{E}[\widehat{Y}_{ret} \mid X, Z, W] = \mathbb{E}[Y \mid X, Z, W] \quad \textit{a.s.} \tag{8}$$

The LLM-aggregator $\widehat{Y}_{RAG}$ is calibrated if it satisfies

$$\mathbb{E}\big[\widehat{Y}_{RAG} \mid q, N_K(q; \mathcal{D})\big] = \frac{1}{K} \sum_{i \in N_K(q;\mathcal{D})} Y_i \quad \textit{a.s.} \tag{9}$$

Intuitively, the calibration conditions characterize when each pipeline step is locally mean-preserving. If Eq. 8 holds, retrieval does not systematically shift the outcome on average within the relevant strata. Similarly, if Eq. 9 holds, the aggregation step (performed by the LLM) does not systematically shift the retrieval prediction.

In general, both $\widehat{Y}_{\text{ret}}$ and $\widehat{Y}_{\text{RAG}}$ are (possibly stochastic) functions of the query covariates, and therefore define new outcome mechanisms $f_{\widehat{Y}}$ with parents in $\{X, Z, W\}$, attached to the SFM in place of f_Y (Plečko & Bareinboim, 2024) (see Fig. 1a). Consequently, the decomposition in Eq. 3 applies to each score $S \in \{Y, \widehat{Y}_{\text{ret}}, \widehat{Y}_{\text{RAG}}\}$, and we may speak of the direct, indirect, and spurious effects of X on any stage of the pipeline. Throughout, we treat the database $\mathcal{D}$ as fixed and part of the mechanism $f_{\widehat{Y}}$ itself. For any score S, we collect its pathway decomposition into the *bias signature*

$$\mathcal{B}(S) = \big(x\text{-DE}_{x_0,x_1}(S \mid x_0),\ x\text{-IE}_{x_1,x_0}(S \mid x_0),\ x\text{-SE}_{x_1,x_0}(S)\big). \tag{10}$$

2.1 Bias in the Data

The first source of disparity is the retrieval database itself: the historical labels Y_i may encode a direct effect of the protected attribute, an indirect effect transmitted through the mediators W (e.g., the contents of the triage note as in Ex. 1), and a spurious effect arising from the confounders Z (e.g., age). The signature $\mathcal{B}(Y)$, obtained by decomposing $\text{TV}_{x_0,x_1}(Y)$ via Eq. 3, quantifies these three channels and constitutes the baseline against which the pipeline is compared.

An important consequence of the data-generating model being compatible with the SFM is that each x-specific effect of a score S depends on its mechanism only through the conditional mean $\mathbb{E}[S \mid x, z, w]$ (see App. A). Thus, if a predictor is calibrated (Def. 1), and reproduces the historical conditional mean,

$$\mathbb{E}[S \mid X = x, Z = z, W = w] = \mathbb{E}[Y \mid X = x, Z = z, W = w], \tag{11}$$

then its bias signature coincides with that of the historical labels, $\mathcal{B}(S) = \mathcal{B}(Y)$. In particular, a pipeline that merely reproduces $\mathbb{E}[Y \mid x, z, w]$ introduces no additional pathway-specific disparity beyond that already present in the data.

2.2 Retrieval Bias

The retrieval predictor $\widehat{Y}_{\text{ret}}$ in Eq. 6 replaces the conditional mean $\mathbb{E}[Y \mid x, z, w]$ with a nearest-neighbor average: the neighbors are selected in embedding space, and their outcomes are averaged. This substitution may itself induce bias. Embeddings may cluster on names, dialect, documentation style, or other proxies of the protected attribute, in which case the neighborhoods $N_K(q)$ retrieved for the two groups differ systematically in their label composition. To quantify such distortions, we introduce the following notation. For any two scores S, S' attached to the SFM, and each causal effect $\text{CE} \in \{x\text{-DE}, x\text{-IE}, x\text{-SE}\}$, define the *effect shift*

$$\Delta\text{CE}(S \to S') := \text{CE}(S') - \text{CE}(S), \tag{12}$$

where the effects carry the subscripts of Eq. 3, e.g., $\Delta x\text{-DE}(S \to S') = x\text{-DE}_{x_0,x_1}(S' \mid x_0) - x\text{-DE}_{x_0,x_1}(S \mid x_0)$, and analogously for $x\text{-IE}_{x_1,x_0}(\cdot \mid x_0)$ and $x\text{-SE}_{x_1,x_0}(\cdot)$. Since the transitions $x_0 \to x_1$ and $x_1 \to x_0$ are fixed throughout, we suppress these subscripts in the Δ terms in what follows. This notation allows for the following definition:

Definition 2 (Retrieval Bias) *For each CE $\in \{x\text{-DE}, x\text{-IE}, x\text{-SE}\}$, the* retrieval-induced bias *is the effect shift from the historical labels to the retrieval predictor,*

$$\Delta^{ret}\text{CE} := \Delta\text{CE}(Y \to \widehat{Y}_{ret}), \tag{13}$$

and we write $\Delta^{ret} = \mathcal{B}(\widehat{Y}_{ret}) - \mathcal{B}(Y)$ for the corresponding signature shift. □

The following result shows that the disparity of the retrieval predictor separates cleanly into a part inherited from the data and a part induced by the retrieval algorithm, along each causal pathway:

Proposition 1 (Retrieval Bias Decomposition) *The total variation of the retrieval stage satisfies*

$$\begin{aligned} \text{TV}_{x_0,x_1}(\widehat{Y}_{ret}) = &\underbrace{\big[x\text{-DE}_{x_0,x_1}(Y \mid x_0) - x\text{-IE}_{x_1,x_0}(Y \mid x_0) - x\text{-SE}_{x_1,x_0}(Y)\big]}_{\textit{data bias}} \\ &+ \underbrace{\big[\Delta^{ret}\text{DE} - \Delta^{ret}\text{IE} - \Delta^{ret}\text{SE}\big]}_{\textit{retrieval-induced bias}}. \end{aligned} \tag{14}$$

Furthermore, if retrieval is calibrated (Def. 1), then $\Delta^{ret}\text{CE} = 0$ for every effect:

$$\mathbb{E}[\widehat{Y}_{ret} \mid X, Z, W] = \mathbb{E}[Y \mid X, Z, W] \quad \textit{a.s.} \implies \Delta^{ret}\text{DE} = \Delta^{ret}\text{IE} = \Delta^{ret}\text{SE} = 0. \tag{15}$$

Prop. 1 establishes an important point: the bias of the retrieval predictor can be broken into two parts, $\mathcal{B}(\widehat{Y}_{\text{ret}}) = \mathcal{B}(Y) + \Delta^{\text{ret}}$, with one part inherited from the data ($\mathcal{B}(Y)$), and the other being the contribution of the retrieval (Δ^{ret}). The calibration condition in Eq. 8 further establishes a meaningful null: calibrated retrieval implies that $\Delta^{\text{ret}} = 0$, that is, the bias signature remains preserved.

2.3 Retrieval-Augmented Generation Bias

The final stage replaces the label average with an LLM-generated decision: the retrieved cases and their labels are placed into the prompt $\rho(q, N_K)$, and the LLM produces the decision $\widehat{Y}_{\text{RAG}}$. To isolate the contribution of this stage, note that the retrieval predictor $\widehat{Y}_{\text{ret}}$ can itself be viewed as an aggregator without any inductive bias: it treats the retrieved cases symmetrically and simply reports their mean label. An LLM, in contrast, weighs the retrieved evidence through its own internal beliefs. It may discount certain cases, extrapolate beyond them, or attend to features of the query (including protected-attribute proxies) that the label average ignores. The difference between the two aggregators, applied to the *same* retrieved evidence, therefore quantifies the inductive bias that the LLM adds on top of retrieval:

Definition 3 (LLM-Processing Bias) *For CE* $\in \{x\text{-}DE, x\text{-}IE, x\text{-}SE\}$*, the* LLM-processing bias *is*

$$\Delta^{LLM}CE = \Delta CE(\widehat{Y}_{ret} \to \widehat{Y}_{RAG}) = CE(\widehat{Y}_{RAG}) - CE(\widehat{Y}_{ret}), \tag{16}$$

with signature shift $\Delta^{LLM} = \mathcal{B}(\widehat{Y}_{RAG}) - \mathcal{B}(\widehat{Y}_{ret})$. □

Based on concepts in Defs. 2 and 3, we can prove the main decomposition of the paper:

Theorem 1 (Three-Way RAG Decomposition) *For each causal effect CE* $\in \{x\text{-}DE, x\text{-}IE, x\text{-}SE\}$*, the RAG predictor satisfies*

$$CE(\widehat{Y}_{RAG}) = CE(Y) + \Delta^{ret}CE + \Delta^{LLM}CE, \tag{17}$$

and consequently the total disparity of the pipeline $TV_{x_0,x_1}(\widehat{Y}_{RAG})$ *decomposes as*

$$\underbrace{TV_{x_0,x_1}(Y)}_{\text{data bias}} + \underbrace{\left[\Delta^{ret}DE - \Delta^{ret}IE - \Delta^{ret}SE\right]}_{\text{retrieval-induced bias}} + \underbrace{\left[\Delta^{LLM}DE - \Delta^{LLM}IE - \Delta^{LLM}SE\right]}_{\text{LLM-induced bias}}. \tag{18}$$

Furthermore, if the LLM-aggregator is calibrated (Def. 1), then $\Delta^{LLM}CE = 0$ *for every effect:*

$$\mathbb{E}\left[\widehat{Y}_{RAG} \mid q, N_K(q; \mathcal{D})\right] = \frac{1}{K} \sum_{i \in N_K(q;\,\mathcal{D})} Y_i \quad a.s. \implies \Delta^{LLM}DE = \Delta^{LLM}IE = \Delta^{LLM}SE = 0. \tag{19}$$

Thm. 1 is the attribution tool we sought: the signature of the deployed system is the sum of three interpretable contributions, $\mathcal{B}(\widehat{Y}_{\text{RAG}}) = \mathcal{B}(Y) + \Delta^{\text{ret}} + \Delta^{\text{LLM}}$, each of which is a 3-dimensional object over the causal pathways. If $\mathcal{B}(Y)$ is already large, the historical database is biased before any retrieval or reasoning occurs, and the remedy lies in data curation. If Δ^{ret} is large, the retriever selects systematically different neighbors across groups, and the remedy lies in updating the retrieval algorithm. If Δ^{LLM} is large, the LLM is not merely averaging the evidence placed in front of it; it superimposes its own beliefs about the relationship between group membership and the outcome, and the remedy lies in the choice of aggregator or prompt. Furthermore, the calibration conditions in Def. 1 establish a key point: under calibrated retrieval/aggregation, the corresponding stage leaves the bias signature unchanged, so that $\Delta^{\text{ret}} = 0$ or $\Delta^{\text{LLM}} = 0$, respectively. This shows that the key decomposition in Eq. 18 is not an arbitrary telescoping summation, but a proper attribution assigning disparities to the appropriate stage of the pipeline.

Remark 1 (Attribution for General Disparity Metrics) *The attribution in Eq. 17 holds when CE is replaced by* any *disparity functional of the joint law of* (X, Z, W, S)*, such as demographic or distributional parity. Moreover, the null conditions of Prop. 1 and Thm. 1 continue to imply* $\Delta^{ret} = 0$ *and* $\Delta^{LLM} = 0$ *for any functional determined by the conditional mean* $\mathbb{E}[S \mid x, z, w]$ *and the covariate distribution* $P(x, z, w)$*.*

We also remark that the decomposition naturally extends to richer pipelines: additional stages (reranking, weighted aggregation $\widehat{Y}_{\text{wRAG}} = \sum_i a_i(q) Y_i$ with softmax weights $a_i(q) \propto \exp\{\lambda s_i(q)\}$, or retrieval-free LLM scoring) would insert further terms into the sum in Eq. 17, each measuring the signature shift of the corresponding mechanism replacement.

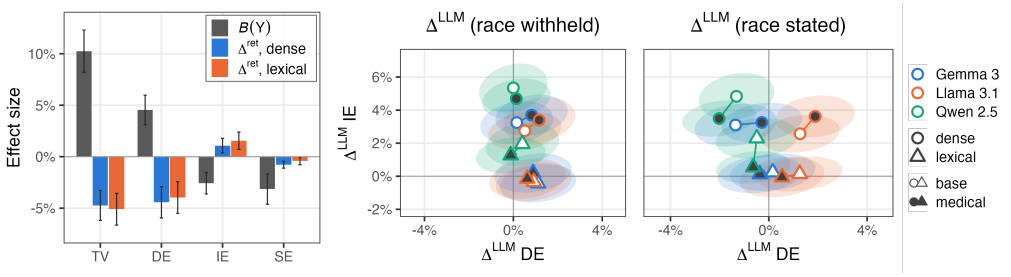


Figure 2: Triage (MIMIC-IV-ED), $n = 16{,}000$. Left: bias signature of the recorded labels, $\mathcal{B}(Y)$, beside the retrieval-induced shift Δ^{ret} for each retriever. Middle/Right: the shifts Δ^{LLM} in the (DE, IE) plane, with race withheld or stated in the query. Circles denote dense retrieval and triangles lexical; hollow markers are base models and filled ones their medical fine-tunes, joined by an arrow. Shaded 95% confidence ellipses obtained from joint influence-function covariances.

3 Experiments

We validate the estimation procedure on semi-synthetic data with a known ground truth (see App. D), based on the MIMIC-IV-ED described in the sequel. In the main text, we focus on real-world data. Throughout, we report $x\text{-DE}_{x_0,x_1}(\cdot \mid x_0)$, $x\text{-IE}_{x_1,x_0}(\cdot \mid x_0)$, and $x\text{-SE}_{x_1,x_0}(\cdot)$, so that TV $=$ DE $-$ IE $-$ SE (Eq. 3). Note that IE and SE consider the reverse transition $x_1 \to x_0$ compared to DE, and enter the total variation with a negative sign; a *negative* IE/SE *increases* the TV measure. We report 95% confidence intervals based on the estimated variance of the influence function. Operational aspects, such as identification and estimation, are discussed in App. B.

3.1 MIMIC-IV-ED: Emergency Department Triage

The MIMIC-IV-ED database (Johnson et al., 2023b;a) contains information on emergency department (ED) visits, including patient demographics, mode of arrival, chief complaint, and triage vital signs. Upon presentation, each patient is assigned an Emergency Severity Index (ESI, Wuerz et al., 2000), a standard triage score ranging from 1 (most urgent) to 5 (least urgent). We restrict to patients recorded as White (x_1) or Black (x_0), take the confounders Z to be sex and age, and define the outcome $Y = 1$ if ESI ≤ 2, indicating that the patient's condition requires urgent treatment. Since part of the raw triage data is tabular, we construct the text mediator W: for each visit, a $\sim$100-word triage vignette is generated by Llama 3.3 70B (Grattafiori et al., 2024) from the known facts (age, sex, arrival mode, chief complaint, vitals, and reported pain), with explicit instructions to never mention race or acuity. The full cohort construction and prompting details are given in App. C. We assume the data is compatible with the SFM in Fig. 1a. From the 32,172 White and Black visits in the cohort we draw an estimation sample of $n = 16{,}000$ and assign the remaining 16,172 to the retrieval database $\mathcal{D}$, which is therefore fixed and disjoint from every query. The admissible set $\mathcal{A}(q)$ matches candidates on sex and age decade with the query, and $K = 10$ neighbors are retrieved within it. We compare two retrievers: dense cosine similarity over `all-MiniLM-L6-v2` embeddings (Sentence Transformers, 2021) and lexical BM25 matching (Robertson & Zaragoza, 2009).

Aggregators & Estimation. We use six open-weight models as LLM aggregators, arranged as three pairs, each consisting of a base model and a medical fine-tune of that same base: Gemma 3 27B (Gemma Team, 2025) and MedGemma 27B (Sellergren et al., 2025); Llama 3.1 70B (Grattafiori et al., 2024) and Llama3.1-Aloe-Beta-70B; and Qwen 2.5 72B (Qwen Team, 2024) and Qwen2.5-Aloe-Beta-72B. The two Aloe models share the adaptation recipe (Garcia-Gasulla et al., 2025). Each retrieval-model pair is run under two disclosure conditions: the LLM-aggregator query having the race withheld, and having race stated as a structured field. This gives 24 configurations in total. Retrieval is shared across aggregators and disclosure: for a given retriever, all twelve model-disclosure pairs are shown the same neighborhood $N_K(q; \mathcal{D})$, so Δ^{ret} remains the same regardless of the aggregator-disclosure pair. Text nuisances $\mu_S(x, z, w)$ and $\pi(x \mid z, w)$ are fit with ModernBERT (Warner et al., 2024), with a shared trunk and one outcome head per x (Shalit et al., 2017); $\pi(x \mid z, w)$ is fit over (X, Z, W) and not on the outcome, so it is estimated once and reused by all 27 outcomes.

Findings. In the data, White patients are labeled urgent at a higher rate than Black patients, $\mathrm{TV}(Y) = 10.2\% \pm 2.1\%$, and all three pathways are individually significant (Fig. 2 left, grey bars). The largest contribution is the direct effect, $\mathrm{DE}(Y) = 4.5\% \pm 1.4\%$, meaning White patients are more likely to be labeled urgent than Black patients presenting with the same described symptoms. The spurious effect, $\mathrm{SE}(Y) = -3.1\% \pm 1.5\%$, reflects the differing age composition of the two groups, and a smaller indirect effect, $\mathrm{IE}(Y) = -2.6\% \pm 1.0\%$, is transmitted through the vignette content.

Both retrievers remove the direct effect: $\Delta^{\text{ret}}\mathrm{DE} = -4.4\% \pm 1.5\%$ for dense retrieval and $-4.0\% \pm 1.5\%$ for lexical (Fig. 2 left, blue/orange bars), leaving $\mathrm{DE}(\widehat{Y}_{\text{ret}}) = 0.1\% \pm 0.5\%$ and $0.6\% \pm 0.6\%$ respectively. As race is excluded from the vignettes by construction, and the matching is a function of Z, $\widehat{Y}_{\text{ret}}$ is not a function of X, so the DE must vanish (see Prop. 3 of App. D for details). The indirect effect is attenuated, $\Delta^{\text{ret}}\mathrm{IE} = 1.1\% \pm 0.7\%$ and $1.6\% \pm 0.8\%$, but the spurious effect moves the other way, $\Delta^{\text{ret}}\mathrm{SE} = -0.8\% \pm 0.3\%$ and $-0.4\% \pm 0.4\%$, leaving $\mathrm{SE}(\widehat{Y}_{\text{ret}}) = -3.9\% \pm 1.4\%$ under dense retrieval, larger in magnitude than in the data. This change in the SE offsets part of the DE removal; in total, retrieval halves the aggregate disparity to $\mathrm{TV}(\widehat{Y}_{\text{ret}}) \approx 5.5\%$.

When race is withheld, $\widehat{Y}_{\text{RAG}}$ has no access to X, so $\mathrm{DE}(\widehat{Y}_{\text{RAG}})$ is expected to vanish; since $\mathrm{DE}(\widehat{Y}_{\text{ret}}) = 0$ as well, we conclude that $\Delta^{\text{LLM}}\mathrm{DE}$ should vanish. Across all 12 race-withheld pipelines $\Delta^{\text{LLM}}\mathrm{DE}$ lies between -0.1% and 1.2%, with no confidence interval excluding zero (Fig. 2 middle). At the same time, the IE behaves differently, and separates more by retriever than aggregator. Under dense retrieval every one of the six aggregators shifts it significantly, from $\Delta^{\text{LLM}}\mathrm{IE} = 2.8\% \pm 0.9\%$ for Llama 3.1 to $5.3\% \pm 1.2\%$ for Qwen 2.5. Under lexical retrieval the indirect effect largely disappears: four of the six lie within $\pm 0.4\%$ of zero, and only the two Qwen models remain significant, at 2.0% and 1.3%. Since the aggregation stage is identical in the two cases and only the retrieved evidence differs, the shift is a property of the retrieval-aggregator pair, and not of the aggregator alone. The sign also matters: $\mathrm{IE}(Y)$ is -2.6% in the data, so under dense retrieval the pipeline does not merely attenuate the indirect pathway but reverses it, weighing the symptoms described in the vignette differently than the triage clinicians did.

Stating race allows for direct effects in aggregation. Under dense retrieval, four of the six aggregators carry a significant DE, with sign split by model family: Gemma 3 at $-1.4\% \pm 1.3\%$, Qwen 2.5 at $-1.3\% \pm 1.2\%$ and Aloe-Qwen at $-2.0\% \pm 1.2\%$ predict higher urgency for Black patients presenting with the same symptoms, whereas Aloe-Llama at $+1.9\% \pm 1.3\%$ predicts higher urgency for White patients; MedGemma and Llama 3.1 are not significant at the 5% level. Under lexical retrieval all six remain indistinguishable from zero, as when race is withheld. The indirect pathway, by contrast, remains largely unchanged by race disclosure: $\Delta^{\text{LLM}}\mathrm{IE}$ ranges over 2.6% to 4.8% under dense retrieval, against 2.8% to 5.3% when race is withheld. The two interventions therefore act on separate pathways, the retriever governing the indirect effect and disclosure the direct one.

Two facts on prediction accuracy qualify the above comparisons. Against the recorded outcome Y, lexical retrieval achieves an accuracy of 67.5%, and dense 72.9% (against a majority-class baseline of $P(Y = 0) = 60.3\%$). Thus, smaller fairness shifts of lexical retrieval may reflect less informative neighborhoods rather than more calibrated aggregation. Moreover, no LLM aggregator improves on the dense retrieval average, with the best of 24 configurations reaching an accuracy of only 71.2%.

The results further indicate that medical fine-tuning moves the aggregator's output level substantially but its bias signature very little. The marginal rate at which a model returns $\widehat{Y}_{\text{RAG}} = 1$ shifts by up to 19% between a base model and its medical fine-tune, in both directions: Aloe-Llama raises it from 55% to 62% while Aloe-Qwen lowers it from 48% to 29%, against a rate of 39.7% for the recorded outcome. The corresponding movement in $(\Delta^{\text{LLM}}\mathrm{DE}, \Delta^{\text{LLM}}\mathrm{IE})$, however, is at most 1.8% in each of the six pairs, in both disclosure conditions.

In summary, $\mathcal{B}(Y)$ is large, and Δ^{ret} and Δ^{LLM} act on different pathways: retrieval removes the direct effect, while aggregation, under dense retrieval, reverses the indirect one, and introduces a direct effect only once race is stated to it. Whether such dampening is desirable depends on the substantive interpretation of the data-level effects, e.g., genuine differences in presentation severity versus historical over- or under-triage of racial groups. Distinguishing between the two requires anchoring the human-assigned acuity labels to objective downstream outcomes, such as mortality or ICU admission, requiring further work. Regardless, the proposed attribution makes the pipeline's behavior transparent, and shows the importance of granular insights into RAG pipelines.

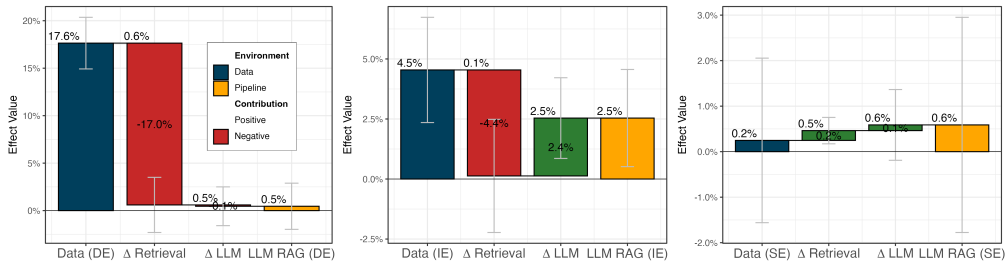


Figure 3: Waterfall plots of DE/IE/SE for Kiva application: data-level effects $\mathcal{B}(Y)$ (blue), retrieval shift Δ^{ret} and LLM shift Δ^{LLM} (green/red increments), and the resulting pipeline effect (orange).

3.2 Kiva: Microfinance Loan Funding

We next look at an example in microfinance, based on a hypothetical application of a RAG system to real data. Kiva is an online platform on which borrowers apply for small loans (Choo et al., 2014). Each posting is described by a short narrative, and individual lenders choose which loans to fund. To provide feedback to users about their descriptions (to indicate whether their application will attract high demand), a pipeline is instantiated as follows: given a new posting, similar past postings and their outcomes are retrieved, and the aggregator predicts the outcome for the query.

We focus on Kiva data from Colombia. App. E describes the cohort construction, and gives additional experimental details. The protected attribute X is borrower sex (x_1 female, x_0 male), confounder Z is the posting year, and the mediators W include the text description, economic sector, and the loan amount. To determine demand Y, we use the speed of funding. Among all funded postings, let T be the share of the fundraising window elapsed at completion, and set $Y = 1$ if T falls below the median, and $Y = 0$ otherwise. We assume the data is compatible with the SFM in Fig. 1a. This yields a total of $n = 29{,}003$ postings from 2014 onward. Of these, 60% are used for the estimation sample, and 40% for the retrieval database $\mathcal{D}$, from which $K = 10$ neighbors are drawn by dense cosine retrieval, restricted to postings from the same year. The aggregator is Llama 3.1 70B (Grattafiori et al., 2024). In the prompt, the model is told the borrower's sex via a structured field.

Since Kiva descriptions name the borrower and use gendered pronouns, we remove sex markers from every description before it enters the pipeline, to reduce issues of overlap (App. E). Furthermore, we additionally propensity-trim the estimation sample at $\varepsilon = 0.05$, retaining 85.2% of the data.

Findings. Fig. 3 shows the pipeline decomposition. In the data, lenders fund women's postings faster, and the effect runs mainly along the direct pathway, $\text{DE}(Y) = 17.6\% \pm 2.7\%$: two postings describing the same business, for the same amount, in the same year, are funded at markedly different speeds according to the borrower's sex. A smaller indirect effect favoring men, $\text{IE}(Y) = 4.5\% \pm 2.2\%$, is transmitted through description/amount, while the spurious effect is negligible, $\text{SE}(Y) = 0.2\% \pm 1.8\%$, so that the sex composition of the postings over time does not explain the disparity. The retrieval step removes the direct effect almost entirely, $\Delta^{\text{ret}}\text{DE} = -17.0\% \pm 2.9\%$, leaving $\text{DE}(\widehat{Y}_{\text{ret}}) = 0.6\% \pm 1.1\%$ non-significant (recall Prop. 3). The aggregation step does not add any of it back, with $\Delta^{\text{LLM}}\text{DE} = -0.1\% \pm 2.0\%$. The indirect pathway behaves differently. The retrieval step also removes the male advantage there, $\Delta^{\text{ret}}\text{IE} = -4.4\% \pm 2.4\%$, but the aggregator restores part of it, $\Delta^{\text{LLM}}\text{IE} = 2.4\% \pm 1.7\%$: the LLM weighs the content of the description differently than the retrieved labels alone would warrant, ending at $\text{IE}(\widehat{Y}_{\text{RAG}}) = 2.5\% \pm 2.0\%$. The spurious pathway stays negligible throughout, $\text{SE}(\widehat{Y}_{\text{RAG}}) = 0.6\% \pm 2.4\%$. The deployed pipeline therefore reproduces none of the direct effect that dominates the historical record, and the only disparity it carries is the one the aggregator introduces along the mediating pathway.

Limitations & Future Work. An important limitation is that applying the framework to a new domain requires specifying the SFM, i.e., human elicitation of causal assumptions about which covariates confound or mediate the effect of the protected attribute. Future work should investigate efficiency-preserving nuisance estimation for text mediators, since current estimators inherit the limitations of the underlying encoder, such as document truncation and absence of convergence rate guarantees. Further, moving from bias detection to fair prediction is an important step, to design retrieval and aggregation mechanisms that provably control components of the bias signature.

AI Use Statement

Generative AI tools were used for language editing, proofreading, and polishing of the manuscript. They were not used to develop the scientific ideas, theoretical results, methodology, or conclusions of the work.

Reproducibility Statement

We provide detailed descriptions of the datasets, preprocessing procedures, retrieval pipelines, model configurations, estimands, nuisance estimation, and evaluation protocols used in the experiments. The appendix contains additional implementation details and semi-synthetic validation. Code and configuration files necessary to reproduce the reported experiments are made available.

Ethics Statement

We do not foresee significant ethical concerns arising from this work. The proposed methodology is intended to help identify and localize potential fairness concerns in retrieval-augmented decision systems, including distinguishing disparities originating from the underlying data, retrieval mechanism, and downstream language models. We hope that these tools can support more careful auditing and diagnosis of such systems. The empirical analyses use existing datasets and are intended for methodological evaluation rather than deployment or individual-level decision making.

References

Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 386–397, 2024.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Peter J Bickel, Chris AJ Klaassen, Ya'acov Ritov, and Jon A Wellner. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

Francine D Blau and Lawrence M Kahn. The gender earnings gap: learning from international comparisons. *The American Economic Review*, 82(2):533–538, 1992.

Francine D Blau and Lawrence M Kahn. The gender wage gap: Extent, trends, and explanations. *Journal of economic literature*, 55(3):789–865, 2017.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7801–7808, 2019.

Jaegul Choo, Changhyun Lee, Daniel Lee, Hongyuan Zha, and Haesun Park. Understanding and promoting micro-finance activities in kiva. org. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pp. 583–592, 2014.

Sanjiv Das, Richard Stanton, and Nancy Wallace. Algorithmic fairness. *Annual Review of Financial Economics*, 15(1):565–593, 2023.

Alessandro Fabris, Nina Baranowska, Matthew J Dennis, David Graus, Philipp Hacker, Jorge Saldivar, Frederik Zuiderveen Borgesius, and Asia J Biega. Fairness and bias in algorithmic hiring: A multidisciplinary survey. *ACM Transactions on Intelligent Systems and Technology*, 16(1):1–54, 2025.

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

Dario Garcia-Gasulla, Jordi Bayarri-Planas, Ashwin Kumar Gururajan, et al. The aloe family recipe for open and specialized healthcare llms. *arXiv preprint arXiv:2505.04388*, 2025.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622, 2024.

Mengxuan Hu, Hongyi Wu, Ronghang Zhu, Zihan Guan, Dongliang Guo, Daiqing Qi, and Sheng Li. No free lunch: Retrieval-augmented generation undermines fairness in llms, even for vigilant users. In *EMNLP (Findings)*, pp. 18145–18170, 2025.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Leo Anthony Celi, Roger Mark, and Steven Horng. MIMIC-IV-ED. *PhysioNet*, January 2023a. doi: 10.13026/5ntk-km72. URL `https://doi.org/10.13026/5ntk-km72`. Version 2.2.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023b.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744*, 2017.

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

To Eun Kim and Fernando Diaz. Towards fair rag: On the impact of fair ranking in retrieval-augmented generation. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, pp. 33–43, 2025.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pp. 611–626, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36:56338–56351, 2023.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 5356–5371, 2021.

Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 786–808, 2023.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1953–1967, 2020.

Devah Pager. The mark of a criminal record. *American journal of sociology*, 108(5):937–975, 2003.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000. 2nd edition, 2009.

Drago Plecko. Causal bias detection in generative artificial intelligence. *arXiv preprint arXiv:2605.11365*, 2026.

Drago Plečko and Elias Bareinboim. Reconciling predictive and statistical parity: A causal approach. *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 2024.

Drago Plečko and Elias Bareinboim. Causal fairness analysis. *Foundations and Trends in Machine Learning*, 17 (3):304–589, 2024.

Drago Plečko and Nicolai Meinshausen. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21:242, 2020.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174, 2009.

James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89 (427):846–866, 1994.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Sentence Transformers. all-MiniLM-L6-v2. `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`, 2021. Hugging Face model card.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Ilya Shpitser and Judea Pearl. What counterfactuals can be tested. In *Proceedings of the Twenty-third Conference on Uncertainty in Artificial Intelligence*, pp. 352–359, 2007.

Robik Shrestha, Yang Zou, Qiuyu Chen, Zhiheng Li, Yusheng Xie, and Siqi Deng. Fairrag: Fair human generation via fair retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11996–12005, 2024.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3784–3803, 2021.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. "i'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 9180–9211, 2022.

Laura T Sweeney and Craig Haney. The influence of race on sentencing: A meta-analytic review of experimental studies. *Behavioral Sciences & the Law*, 10(2):179–195, 1992.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

Sibo Wang, Xiangkui Cao, Jie Zhang, Zheng Yuan, Shiguang Shan, Xilin Chen, and Wen Gao. Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. *arXiv preprint arXiv:2406.14194*, 2024.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, et al. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*, 2024.

Kyra Wilson and Aylin Caliskan. Gender, race, and intersectional bias in resume screening via language model retrieval. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7(1), pp. 1578–1590, 2024.

Xuyang Wu, Shuowei Li, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. Does rag introduce unfairness in llms? evaluating fairness in retrieval-augmented generation systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10021–10036, 2025.

Richard C Wuerz, Leslie W Milne, David R Eitel, Debbie Travers, and Nicki Gilboy. Reliability and validity of a new five-level triage instrument. *Academic emergency medicine*, 7(3):236–242, 2000.

Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

SUPPLEMENTARY MATERIAL FOR *WHERE DOES BIAS ENTER RAG? A CAUSAL ATTRIBUTION FRAMEWORK*

The source code for reproducing the results can be found in the anonymized code repository `https://anonymous.4open.science/r/ragcbd-8B40`. The repository includes a README file explaining the setup steps. All experiments were run on four NVIDIA GH200 GPUs with 96GB VRAM each. Inference was performed using vLLM (Kwon et al., 2023), and the total computation time required to generate the results across all models and datasets is under 96 hours. BERT fine-tuning for the text-based nuisances $\mu_S(x, z, w)$ and $\pi(x \mid z, w)$ used AdamW with learning rate 5×10^{-5}, a linear schedule, batch size 16, and 5 epochs, with validation loss evaluated four times per epoch, early stopping after 4 evaluations without improvement, and the best checkpoint restored before prediction.

A PROOFS

We first prove Prop. 2 appearing in App. B, since the remaining proofs build on the identification expressions established therein.

A.1 PROOF OF PROPOSITION 2

Since the scores $\widehat{Y}_{\text{ret}}, \widehat{Y}_{\text{RAG}}$ are attached to the SFM via mechanisms $f_S(x, z, w, U_S)$ with $U_S \perp\!\!\!\perp U_{XZ}, U_W$, the augmented graph over (X, Z, W, S) is again an SFM with S in place of Y (Plečko & Bareinboim, 2024), and it suffices to prove the claim for a generic score S compatible with the SFM. We demonstrate why the potential outcome $\mathbb{E}[S_{x_y, W_{x_w}} \mid X = x_z]$ is identifiable under the SFM. For proving its identifiability, we make use of the counterfactual graph (Shpitser & Pearl, 2007) in Fig. 4. We can expand $P(S_{x_y, W_{x_w}} = s \mid x_z)$ as follows:

$$= \sum_w P(S_{x_y,w} = s, W_{x_w} = w \mid x_z) \qquad \text{(Counterfactual Un-nesting)} \tag{20}$$

$$= \sum_{z,w} P(S_{x_y,w} = s, W_{x_w} = w \mid z, x_z) P(z \mid x_z) \qquad \text{(Law of Total Probability)} \tag{21}$$

$$= \sum_{z,w} P(S_{x_y,w} = s \mid z, x_z) P(W_{x_w} = w \mid z, x_z) P(z \mid x_z) \qquad (S_{x_y,w} \perp\!\!\!\perp W_{x_w} \mid Z, X) \tag{22}$$

$$= \sum_{z,w} P(S_{x_y,w} = s \mid z, x_z) P(W_{x_w} = w \mid z, x_w) P(z \mid x_z) \qquad (W_{x_w} \perp\!\!\!\perp X \mid Z) \tag{23}$$

$$= \sum_{z,w} P(S_{x_y,w} = s \mid z, x_z) P(W = w \mid z, x_w) P(z \mid x_z) \qquad \text{(Consistency)} \tag{24}$$

$$= \sum_{z,w} P(S_{x_y,w} = s \mid z, w, x_z) P(w \mid z, x_w) P(z \mid x_z) \qquad (S_{x_y,w} \perp\!\!\!\perp W \mid Z, X) \tag{25}$$

$$= \sum_{z,w} P(S_{x_y,w} = s \mid z, w, x_y) P(w \mid z, x_w) P(z \mid x_z) \qquad (S_{x_y,w} \perp\!\!\!\perp X \mid Z, W) \tag{26}$$

$$= \sum_{z,w} P(S = s \mid z, w, x_y) P(w \mid z, x_w) P(z \mid x_z), \qquad \text{(Consistency)} \tag{27}$$

where the stated conditional independences follow from d-separation in the counterfactual graph. Taking the expectation over s in Eq. 27 yields

$$\mathbb{E}[S_{x_y, W_{x_w}} \mid X = x_z] = \sum_{z,w} \mu_S(x_y, z, w)\, P(w \mid x_w, z)\, P(z \mid x_z), \tag{28}$$

which is Eq. 38. The potential outcomes appearing in the x-specific effects of Sec. 1.1 are special cases, obtained by suitable choices of (x_y, x_w, x_z), together with the simpler expressions $P(S_{x_y} = s \mid x_z) = \sum_{z,w} P(s \mid x_y, z, w) P(w \mid x_y, z) P(z \mid x_z)$ (obtained by an analogous but simplified argument) and $P(S_x = s \mid x) = P(s \mid x)$.

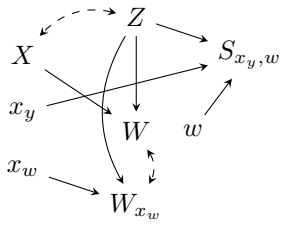


Figure 4: Counterfactual graph used in the proof of Prop. 2, for the potential outcome $S_{x_y,W_{x_w}}$.

For the estimation claim, the efficient influence function of the functional in Eq. 38 is derived using standard semiparametric theory (Bickel et al., 1993; Robins et al., 1994), yielding the expression in Eq. 39; the cross-fitted one-step estimator based on it is $\sqrt{n}$-consistent and asymptotically normal under standard rate conditions on the nuisance estimates (Chernozhukov et al., 2018). ■

A.2 Bias Inheritance

We now prove the bias inheritance lemma:

Lemma 1 (Inheritance of Data Bias) *Let S be any (possibly stochastic) score with mechanism $f_S(x, z, w, U_S)$, $U_S \perp\!\!\!\perp U_{XZ}, U_W$, attached to the SFM. If S is* calibrated*, i.e.,*

$$\mathbb{E}[S \mid X = x, Z = z, W = w] = \mathbb{E}[Y \mid X = x, Z = z, W = w] \quad \forall (x, z, w), \tag{29}$$

then $\mathcal{B}(S) = \mathcal{B}(Y)$, and in particular $TV_{x_0,x_1}(S) = TV_{x_0,x_1}(Y)$. □

By Prop. 2, each of the effects x-DE$_{x_0,x_1}(S \mid x_0)$, x-IE$_{x_1,x_0}(S \mid x_0)$, x-SE$_{x_1,x_0}(S)$ is a contrast of terms of the form in Eq. 38, and therefore a functional of the triple $\big(\mu_S(\cdot), P(w \mid x, z), P(z \mid x)\big)$; in particular, it depends on the mechanism f_S *only* through the conditional mean $\mu_S(x, z, w)$. The distributions $P(w \mid x, z)$ and $P(z \mid x)$ are properties of the underlying SCM $\mathcal{M}$ and do not vary with the choice of score. Calibration states $\mu_S = \mu_Y$, so each effect of S coincides with the corresponding effect of Y, which implies that $\mathcal{B}(S) = \mathcal{B}(Y)$. The equality of total variations then follows from Eq. 3. ■

A.3 Proof of Proposition 1

We first prove Eq. 14. Applying the decomposition of Eq. 3 to the score $\widehat{Y}_{\text{ret}}$ gives

$$\text{TV}_{x_0,x_1}(\widehat{Y}_{\text{ret}}) = x\text{-DE}_{x_0,x_1}(\widehat{Y}_{\text{ret}} \mid x_0) - x\text{-IE}_{x_1,x_0}(\widehat{Y}_{\text{ret}} \mid x_0) - x\text{-SE}_{x_1,x_0}(\widehat{Y}_{\text{ret}}). \tag{30}$$

For each CE $\in \{x\text{-DE}, x\text{-IE}, x\text{-SE}\}$, adding and subtracting the corresponding effect of Y yields

$$\text{CE}(\widehat{Y}_{\text{ret}}) = \text{CE}(Y) + \underbrace{\big[\text{CE}(\widehat{Y}_{\text{ret}}) - \text{CE}(Y)\big]}_{= \Delta\text{CE}(Y \to \widehat{Y}_{\text{ret}}) \,=\, \Delta^{\text{ret}}\text{CE}}, \tag{31}$$

where the underbraced term is the effect shift of Eq. 12, named in Def. 2. Substituting this expansion for each of the three effects in Eq. 30 and regrouping terms, we obtain

$$\text{TV}_{x_0,x_1}(\widehat{Y}_{\text{ret}}) = \underbrace{\big[x\text{-DE}(Y \mid x_0) - x\text{-IE}(Y \mid x_0) - x\text{-SE}(Y)\big]}_{\text{data bias}} + \underbrace{\big[\Delta^{\text{ret}}\text{DE} - \Delta^{\text{ret}}\text{IE} - \Delta^{\text{ret}}\text{SE}\big]}_{\text{retrieval-induced bias}}, \tag{32}$$

which is precisely Eq. 14.

We next prove the claim following Eq. 8, namely that calibration implies $\Delta^{\text{ret}}\text{CE} = 0$ for every CE. Note that, conditionally on the database $\mathcal{D}$, the retrieval predictor $\widehat{Y}_{\text{ret}}$ is a stochastic score with mechanism $f_{\widehat{Y}_{\text{ret}}}(z, w, U_{\widehat{Y}})$ attached to the SFM. The calibration condition in Eq. 8 states that $\mathbb{E}[\widehat{Y}_{\text{ret}} \mid X, Z, W] = \mathbb{E}[Y \mid X, Z, W]$ a.s., which is precisely calibration of $\widehat{Y}_{\text{ret}}$ in the sense of Lem. 1. Applying the lemma then gives $\mathcal{B}(\widehat{Y}_{\text{ret}}) = \mathcal{B}(Y)$, and therefore

$$\Delta^{\text{ret}}\text{CE} = \text{CE}(\widehat{Y}_{\text{ret}}) - \text{CE}(Y) = 0 \quad \text{for every CE} \in \{x\text{-DE}, x\text{-IE}, x\text{-SE}\}, \tag{33}$$

completing the proof. ■

A.4 Proof of Theorem 1

We first prove Eq. 17. For each CE $\in \{x\text{-DE}, x\text{-IE}, x\text{-SE}\}$, adding and subtracting the effect of the intermediate pipeline stage yields the telescoping identity

$$\text{CE}(\widehat{Y}_{\text{RAG}}) = \text{CE}(Y) + \underbrace{\big[\text{CE}(\widehat{Y}_{\text{ret}}) - \text{CE}(Y)\big]}_{=\,\Delta^{\text{ret}}\text{CE (Def. 2)}} + \underbrace{\big[\text{CE}(\widehat{Y}_{\text{RAG}}) - \text{CE}(\widehat{Y}_{\text{ret}})\big]}_{=\,\Delta^{\text{LLM}}\text{CE (Def. 3)}}, \tag{34}$$

which is Eq. 17. To prove Eq. 18, we apply the decomposition of Eq. 3 to the score $\widehat{Y}_{\text{RAG}}$, and substitute Eq. 17 for each of the three effects:

$$\begin{aligned}
\text{TV}_{x_0,x_1}(\widehat{Y}_{\text{RAG}}) &= x\text{-DE}(\widehat{Y}_{\text{RAG}} \mid x_0) - x\text{-IE}(\widehat{Y}_{\text{RAG}} \mid x_0) - x\text{-SE}(\widehat{Y}_{\text{RAG}}) \qquad (35)\\
&= \big[x\text{-DE}(Y \mid x_0) + \Delta^{\text{ret}}\text{DE} + \Delta^{\text{LLM}}\text{DE}\big] - \big[x\text{-IE}(Y \mid x_0) + \Delta^{\text{ret}}\text{IE} + \Delta^{\text{LLM}}\text{IE}\big]\\
&\quad - \big[x\text{-SE}(Y) + \Delta^{\text{ret}}\text{SE} + \Delta^{\text{LLM}}\text{SE}\big]\\
&= \underbrace{\big[x\text{-DE}(Y \mid x_0) - x\text{-IE}(Y \mid x_0) - x\text{-SE}(Y)\big]}_{=\,\text{TV}_{x_0,x_1}(Y)\text{ by Eq. 3}} + \underbrace{\big[\Delta^{\text{ret}}\text{DE} - \Delta^{\text{ret}}\text{IE} - \Delta^{\text{ret}}\text{SE}\big]}_{\text{retrieval}}\\
&\quad + \underbrace{\big[\Delta^{\text{LLM}}\text{DE} - \Delta^{\text{LLM}}\text{IE} - \Delta^{\text{LLM}}\text{SE}\big]}_{\text{LLM processing}},
\end{aligned}$$

where the last step regroups the terms by pipeline stage, yielding Eq. 18.

Finally, we prove the claim following Eq. 19, namely that a calibrated aggregator satisfies $\Delta^{\text{LLM}}\text{CE} = 0$ for every CE. Suppose the LLM is a calibrated aggregator, as in Eq. 9. Then, by the tower property, for every (x, z, w):

$$\mathbb{E}\big[\widehat{Y}_{\text{RAG}} \mid x, z, w\big] = \mathbb{E}\Big[\underbrace{\mathbb{E}\big[\widehat{Y}_{\text{RAG}} \mid q, N_K(q; \mathcal{D})\big]}_{=\,\frac{1}{K}\sum_{i \in N_K} Y_i\text{ by Eq. 9}} \,\Big|\, x, z, w\Big] = \mathbb{E}\big[\widehat{Y}_{\text{ret}} \mid x, z, w\big], \tag{36}$$

using the definition of $\widehat{Y}_{\text{ret}}$ in Eq. 6 in the last step. In other words, $\widehat{Y}_{\text{RAG}}$ is calibrated *with respect to* $\widehat{Y}_{\text{ret}}$. Applying the argument of Lem. 1, with $\widehat{Y}_{\text{ret}}$ taking the role of Y, gives $\mathcal{B}(\widehat{Y}_{\text{RAG}}) = \mathcal{B}(\widehat{Y}_{\text{ret}})$, and therefore

$$\Delta^{\text{LLM}}\text{CE} = \text{CE}(\widehat{Y}_{\text{RAG}}) - \text{CE}(\widehat{Y}_{\text{ret}}) = 0 \quad \text{for every CE} \in \{x\text{-DE}, x\text{-IE}, x\text{-SE}\}, \tag{37}$$

completing the proof. ■

B Identification and Estimation

In this appendix, we discuss operational aspects, namely identification and estimation of the causal effects in Eq. 18. All effects appearing in the bias signature $\mathcal{B}(S)$ are contrasts of potential-outcome means of the form $\mathbb{E}[S_{x_y,W_{x_w}} \mid x_z]$, where $x_y, x_w, x_z \in \{x_0, x_1\}$ denote the attribute values along the direct, mediated, and spurious channels, respectively. The following result gives their identification and an efficient estimation strategy:

Proposition 2 (Identification & Estimation) *Under the assumptions encoded in the SFM, for any score $S \in \{Y, \widehat{Y}_{ret}, \widehat{Y}_{RAG}\}$ with mechanism $f_S(x, z, w, U_S)$, $U_S \perp\!\!\!\perp U_{XZ}, U_W$, the potential outcome mean $\theta := \mathbb{E}[S_{x_y,W_{x_w}} \mid X = x_z]$ is identified as*

$$\theta = \sum_{z,w} \mu_S(x_y, z, w)\, P(w \mid x_w, z)\, P(z \mid x_z), \tag{38}$$

where $\mu_S(x, z, w) := \mathbb{E}[S \mid x, z, w]$. Writing $\pi(x \mid z), \pi(x \mid z, w)$ for the propensity scores and $\nu_S(z) := \mathbb{E}[\mu_S(x_y, Z, W) \mid X = x_w, Z = z]$ for the nested mean, θ can be estimated by the cross-fitted one-step estimator $\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} \phi(V_i; \widehat{\eta})$ based on the (uncentered) efficient influence function

$$\begin{aligned}\phi(V; \eta) = {} & \frac{\mathbb{1}(X = x_y)}{P(x_z)} \frac{\pi(x_w \mid Z, W)}{\pi(x_y \mid Z, W)} \frac{\pi(x_z \mid Z)}{\pi(x_w \mid Z)} \big[S - \mu_S(x_y, Z, W)\big] \\ & + \frac{\mathbb{1}(X = x_w)}{P(x_z)} \frac{\pi(x_z \mid Z)}{\pi(x_w \mid Z)} \big[\mu_S(x_y, Z, W) - \nu_S(Z)\big] + \frac{\mathbb{1}(X = x_z)}{P(x_z)}\, \nu_S(Z),\end{aligned} \tag{39}$$

where $\eta = (\mu_S, \pi, \nu_S)$ collects the nuisance functions (Robins et al., 1994; Bickel et al., 1993; Chernozhukov et al., 2018). □

The nuisances $\mu_S(x, z, w)$ and $\pi(x \mid z, w)$ depend on the document W, and we parameterize them with a pre-trained language encoder (ModernBERT, Warner et al., 2024): the document is encoded, and the pooled representation is concatenated with the numeric covariates Z to form a shared trunk. For the outcome model μ_S, the trunk feeds one linear head per value of x (Shalit et al., 2017), with each head trained on units with $X = x$; for the propensity $\pi(x \mid z, w)$, a single linear head is used. Heads use a sigmoid link for binary targets. The encoder and heads are fine-tuned jointly, with early stopping on a held-out fold. The ν_S nuisance is fit with gradient boosting (Chen & Guestrin, 2016), with ν_S obtained by regressing the encoder-based predictions $\mu_S(x_y, Z, W)$ onto (X, Z) on a fold disjoint from the one used to train the encoder. The remaining $\pi(x \mid z)$ nuisance is fit with gradient boosting, or obtained from simple cell frequencies when Z is discrete and low-dimensional. All nuisances are cross-fitted over 5 folds (Chernozhukov et al., 2018), with standard errors obtained from the estimated variance of the influence function. Pipeline scores $\widehat{Y}_{\text{ret}}, \widehat{Y}_{\text{RAG}}$ are likewise computed out-of-fold. In all experiments, the retrieval database $\mathcal{D}$ is a disjoint set of the data from the estimation sample, and the same database is used for retrieval for each of the estimation folds.

C MIMIC-IV-ED: Construction and Additional Results

In this appendix we describe the construction of the MIMIC-IV-ED cohort and of the text mediator, give the prompts used at each stage of the pipeline, and report results for each of the 24 configurations.

C.1 Cohort and Mediator Construction

The data come from the MIMIC-IV-ED database (Johnson et al., 2023b;a), which records emergency department visits with patient demographics, mode of arrival, chief complaint, triage vital signs, and the assigned ESI acuity. We exclude visits with missing acuity, chief complaint, or age. The outcome is $Y = \mathbb{1}(\text{ESI} \leq 2)$, the protected attribute is race, restricted to patients recorded as White (x_1) or Black (x_0), and the confounders Z are sex and age decade. Vignettes are generated for 40,002 visits, of which 32,172 correspond to White or Black patients. From these we draw an estimation sample of $n = 16{,}000$, stratified so that each (X, Z) cell retains its share of the cohort, and assign the remaining 16,172 visits to the retrieval database $\mathcal{D}$. The outcome base rate is $P(Y = 1) = 39.7\%$ and the majority share is $P(X = x_1) = 72.6\%$. Retrieval takes $K = 10$ neighbors from the admissible set, which requires an exact match on sex and age decade.

Vignette generation. Each visit's raw vitals are complemented with qualitative descriptors computed from standard thresholds: systolic blood pressure (< 90 low; ≥ 140 elevated; ≥ 180 markedly elevated), heart rate (≥ 120 marked tachycardia; ≥ 100 tachycardia; < 60 bradycardia), respiratory rate (≥ 24 tachypnea; < 10 low), oxygen saturation ($< 90\%$ severe hypoxemia; $< 94\%$ reduced), temperature (≥ 100.4F fever; < 96F low), and reported pain (≥ 8 severe; ≥ 4 moderate; 0 none; otherwise mild). Vignettes are then generated with Llama 3.3 70B at temperature 1 (top-$p = 0.9$), from a two-shot prompt supplying the structured facts (age, sex, arrival mode, chief complaint, triage vitals, reported pain) alongside these descriptors, under the following rules:

> *Write a realistic emergency-department triage paragraph for a new patient. Rules: Draw on the provided facts and vital descriptors. Do not add symptoms, diagnoses, clinical impressions, tests, medications, treatments, resources, admission status, disposition, or acuity. Do not mention race or ethnicity. Do not say anything about what should happen next. Write one short paragraph, with approximately 100 words. You may paraphrase; not every fact needs to be stated verbatim. Return only text inside* `<case>...</case>`.

The two in-context examples pair a fact list with a reference vignette, and the generated case is extracted from the delimiter tags. An example output:

> *A 58-year-old woman arrived by ambulance with chest pain she rated 7 out of 10. She was tachycardic and blood pressure was mildly elevated. Respiratory rate and oxygen saturation were normal, and she was afebrile.*

Protected attribute leakage. Race is never written into W, while sex and age are stated explicitly. A logistic probe of X on the sentence embedding $e(W)$ nonetheless attains an out-of-fold AUC of 0.646 over the estimation sample, so W carries some information about X. This has no consequence for the direct effect: any outcome that is a function of (Z, W) only and takes no X argument will have a direct effect that vanishes, regardless of how much X the mediator encodes. The leakage enters only through W, and so can affect the indirect effect alone.

C.2 Pipeline and Estimation

Aggregation prompt. The aggregator receives the retrieved cases and their labels, followed by the query document:

```
Here are K similar cases with known outcomes:
Case 1:  [retrieved document 1]
-> Condition Requires Immediate Attention:  [label 1]
...
```

```
Case K:  [retrieved document K]
-> Condition Requires Immediate Attention:  [label K]
Based on these cases, predict the Condition Requires
Immediate Attention for:
[query document]
Options:
A. 0
B. 1
Begin your answer with the capital letter of your
choice.
```

Retrieved labels are shown as binary values. The prompt is wrapped in the model's chat template and the response generated deterministically (temperature 0) with constrained decoding, restricted to a single token from the option letters, whose decoded option is the score $\widehat{Y}_{\text{RAG}} \in \{0, 1\}$. In the second disclosure arm the line `Race: White` or `Race: Black` is prepended to the query document only; the retrieved cases are unchanged, so the neighborhood and Δ^{ret} are identical between arms.

Nuisances and overlap. The text nuisances $\mu_S(x, z, w)$ and $\pi(x \mid z, w)$ are fit with ModernBERT at learning rate 5×10^{-5}, with one outcome head per value of x off a shared trunk (Shalit et al., 2017), cross-fitted over 5 folds. Since $\pi(x \mid z, w)$ is fit over (X, Z, W) and not on the score, it is estimated once and reused by all 15 scores of an arm. With Z discrete, $\pi(x \mid z)$ is the cell frequency rather than an estimate. Fig. 5 shows the distribution of $\hat{\pi}(x_1 \mid Z, W)$ in each group. It ranges over $[8.1\%, 98.3\%]$ with mean 73.6%, against a majority share of 72.6%, and the two distributions overlap across the whole range, so no propensity trimming is applied.

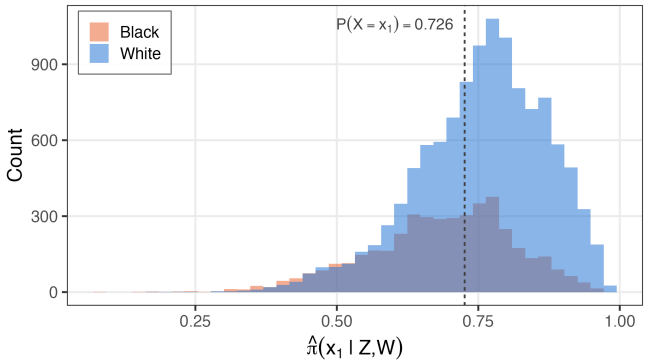


Figure 5: Fitted propensity $\hat{\pi}(x_1 \mid Z, W)$ on the estimation sample, by recorded race. The dashed line marks the marginal share $P(X = x_1)$.

C.3 Per-Configuration Results

We now report, for each of the 24 configurations, the rate at which the aggregator returns $\widehat{Y}_{\text{RAG}} = 1$, its accuracy against the recorded outcome, and its full aggregation-induced shift Δ^{LLM}. These are collected in Tbl. 1. The largest accuracy attained by any aggregator is 71.2% (Aloe-Qwen 72B under lexical retrieval with race stated, in bold). This is below the 72.9% of the dense retrieval average (without aggregation), and above the majority-class baseline of $P(Y = 0) = 60.3\%$. Therefore, no aggregator exceeds the dense retrieval average, as reported in Sec. 3.

In Fig. 6 we plot Δ^{LLM}IE against accuracy on the recorded outcome Y, for all 24 configurations. Lexical retrieval yields both smaller values of Δ^{LLM}IE and lower accuracy than dense retrieval, on average. Across the 24 points the two quantities are positively related, with a correlation of $+0.35$ and a linear slope of $+0.24$ percentage points of Δ^{LLM}IE per point of accuracy, although the linear association is not significant at the 5% level ($t = 1.7$, 22 degrees of freedom).

Model	Retrieval	$P(\widehat{Y}_{\text{RAG}} = 1)$	Acc.	Δ^{LLM} (%) TV	DE	IE	SE
retrieval predictor $\widehat{Y}_{ret}$							
retrieval average	dense	40.0	72.9	–	–	–	–
retrieval average	lexical	38.0	67.5	–	–	–	–
aggregator $\widehat{Y}_{RAG}$*, race withheld*							
Gemma 3 27B	dense	47.7	67.6	-3.4^{*}	0.2	3.2^{*}	0.3
Gemma 3 27B	lexical	50.2	67.6	2.1^{*}	1.1	-0.4	-0.6^{*}
MedGemma 27B	dense	51.5	67.1	-3.0^{*}	0.8	3.7^{*}	0.2
MedGemma 27B	lexical	57.7	65.2	1.1	0.9	0.2	-0.4
Llama 3.1 70B	dense	54.9	66.8	-2.1^{*}	0.5	2.8^{*}	-0.1
Llama 3.1 70B	lexical	56.8	64.9	2.4^{*}	0.9	-0.3	-1.1^{*}
Aloe-Llama 70B	dense	61.5	63.4	-2.7^{*}	1.2	3.4^{*}	0.4
Aloe-Llama 70B	lexical	64.7	61.7	1.2	0.6	-0.1	-0.4
Qwen 2.5 72B	dense	46.0	67.9	-5.3^{*}	-0.0	5.3^{*}	0.0
Qwen 2.5 72B	lexical	47.9	68.0	-0.6	0.4	2.0^{*}	-0.9^{*}
Aloe-Qwen 72B	dense	33.2	70.0	-4.6^{*}	0.1	4.7^{*}	0.1
Aloe-Qwen 72B	lexical	32.8	70.9	-0.6	-0.1	1.3^{*}	-0.8^{*}
aggregator $\widehat{Y}_{RAG}$*, race stated*							
Gemma 3 27B	dense	47.6	67.6	-4.6^{*}	-1.4^{*}	3.1^{*}	0.2
Gemma 3 27B	lexical	50.2	67.4	0.6	0.2	0.2	-0.6^{*}
MedGemma 27B	dense	50.4	67.7	-3.5^{*}	-0.3	3.3^{*}	-0.0
MedGemma 27B	lexical	56.3	65.4	0.0	-0.4	0.1	-0.5
Llama 3.1 70B	dense	53.1	67.8	-0.9	1.3	2.6^{*}	-0.4
Llama 3.1 70B	lexical	55.1	65.3	2.4^{*}	1.3	0.1	-1.3^{*}
Aloe-Llama 70B	dense	62.3	63.3	-2.1^{*}	1.9^{*}	3.6^{*}	0.4
Aloe-Llama 70B	lexical	66.1	61.2	1.0	0.5	-0.1	-0.4
Qwen 2.5 72B	dense	46.4	68.3	-5.9^{*}	-1.3^{*}	4.8^{*}	-0.3
Qwen 2.5 72B	lexical	48.2	67.7	-1.6^{*}	-0.5	2.3^{*}	-1.2^{*}
Aloe-Qwen 72B	dense	30.3	70.8	-5.5^{*}	-2.0^{*}	3.5^{*}	-0.1
Aloe-Qwen 72B	lexical	29.3	**71.2**	-0.7	-0.6	0.5	-0.5^{*}

Table 1: Every configuration of the triage pipeline, $n = 16{,}000$. Rates and accuracies in percent; accuracy thresholds $\widehat{Y}_{\text{ret}}$ at 0.5. Asterisks mark shifts whose 95% confidence intervals exclude zero.

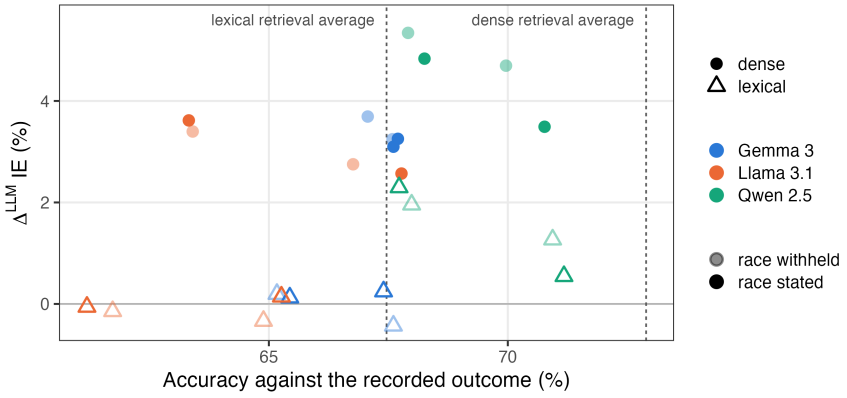


Figure 6: Aggregation-induced indirect effect against accuracy on the recorded outcome, for all 24 configurations. Dashed lines mark the accuracy of the two retrieval averages. Pale markers withhold race from the aggregator, solid markers state it.

D Semi-Synthetic Validation on MIMIC-IV-ED

In this appendix we evaluate the estimation procedure of App. B in a setting with a known ground truth. We specify an outcome mechanism explicitly, and draw synthetic outcomes based on it. We then run the pipeline and the estimation, and compare the estimates against the ground truth values implied by the mechanism. Since the outcome mechanism is known, the bias signature $\mathcal{B}(\widetilde{Y})$ and the shift Δ^{ret} can be computed in closed form. Furthermore, we also use a known closed-form mechanism for the aggregation step, based on which Δ^{LLM} can be computed as well.

We work with the 32,172 visits from the MIMIC-IV-ED cohort (only White and Black patients included); this full cohort represents the population ground truth. From the cohort, we draw an estimation sample of $n = 8{,}041$ encounters and a disjoint retrieval database $\mathcal{D}$ of $6{,}434$ encounters. The majority rate is $P(X = x_1) = 72.6\%$ and the number of retrieved neighbors is $K = 10$. The observed covariates are held fixed throughout: the protected attribute X is race, the confounders Z include age and sex, and the triage vignettes W are those of the real data. Only the outcome mechanism is synthetic, making the overall data semi-synthetic.

Construction of the outcome mechanism. Let $\hat{m}(z, w) = \widehat{\mathbb{E}}[Y \mid Z = z, W = w]$ be a regression of the recorded ESI label on the confounders and the vignette, excluding X, obtained by fine-tuning a BERT-base model with a linear head, and write $\hat{m}^{\text{logit}}$ for its logit function. Let $e(\cdot)$ be a frozen sentence encoder mapping into $\mathbb{R}^p$, and let Π_{20} be the projection onto the first 20 principal components of $\{e(W_i)\}_{i=1}^n$. We compute the probe direction u as the normalized coefficient vector of a logistic regression of the protected attribute on the projected embedding,

$$u = \hat{\beta}/\|\hat{\beta}\|_2, \qquad (\hat{\beta}, \hat{\beta}_0) = \underset{\beta \in \mathbb{R}^{20},\, \beta_0 \in \mathbb{R}}{\arg\min} \sum_{i=1}^{n} \ell\big(\mathbb{1}(X_i = x_1),\, \langle \Pi_{20}\, e(W_i), \beta \rangle + \beta_0\big), \tag{40}$$

where $\ell(y, \eta) = -y\eta + \log(1 + e^{\eta})$ is the logistic loss. The probe achieves AUC $= 0.61$ for predicting X, out of fold. The synthetic outcome $\widetilde{Y}$ is then drawn according to the following mechanism:

$$\widetilde{Y}_i \sim \text{Bern}\Big(\sigma\big\{\, \alpha\, \hat{m}^{\text{logit}}(Z_i, W_i) \;+\; \gamma\, \langle \Pi_{20} e(W_i), u \rangle \;+\; b\, \tilde{Z}_i \;+\; c\, \mathbb{1}(X_i = x_1) \;+\; d \,\big\}\Big), \tag{41}$$

where $\tilde{Z}_i$ is a standardized age-bin index. The five parameters control distinct parts of the mechanism. The coefficient α scales the outcome-relevant clinical signal carried by (Z, W), and hence loads both the mediator and the confounder channels. The coefficient γ loads the X-predictive content of the vignette and so governs the indirect effect, while b acts through the confounders and governs the spurious effect. The coefficient c is the direct effect on the logit scale, and d fixes the base rate.

For parameters $(\alpha, b, c, \gamma, d)$, we search over a grid, evaluating the induced signature $\mathcal{B}(\widetilde{Y})$ by plug-in at each candidate, and retain a configuration in which all three pathway components are large relative to Monte Carlo error. The reported configuration is $(\alpha, b, c, \gamma, d) = (0.25, 0.5, -1.0, 1.5, 0.9266)$, which places the base rate at 0.5 and gives x-DE $= -15.3\%$, x-IE $= -6.6\%$ and x-SE $= -6.7\%$.

Two features of Eq. 41 are used below. First, the outcome regression $\mu_{\widetilde{Y}}(x, z, w)$ is the stated sigmoid, hence available in closed form at either value of x for every unit. Second, the outcome is generated from the vignette W rather than from the underlying structured triage variables, so that W is a parent of $\widetilde{Y}$, as required by the mediator in $\mathbb{E}[S_{x_y, W_{x_w}} \mid x_z]$. Generating from the structured variables instead would place W downstream of the mediator and make $\mathbb{E}[\widetilde{Y}_{x_y, W_{x_w}} \mid x_z]$ a functional of a different model.

Finally, the mechanism and the estimator use different encoders: $\hat{m}$ is built on BERT-base, whereas the nuisances $\mu_S(x, z, w)$ and $\pi(x \mid z, w)$ are fit with ModernBERT, so that the nuisance models are not correctly specified by construction.

Ground truth. All effects are contrasts of $\theta = \mathbb{E}[S_{x_y, W_{x_w}} \mid X = x_z]$. If μ_S is known, no text nuisance is required to evaluate the potential outcome: conditioning on $X = x_w$ reproduces the law of $W \mid Z$ under x_w by consistency, so reweighting the confounder distribution suffices,

$$\theta \;=\; \mathbb{E}\big[\, \mu_S(x_y, Z, W)\, \omega(Z) \;\big|\; X = x_w \big], \quad \omega(z) \;=\; \frac{P(z \mid x_z)}{P(z \mid x_w)} \;=\; \frac{\pi(x_z \mid z)}{\pi(x_w \mid z)} \cdot \frac{P(x_w)}{P(x_z)}. \tag{42}$$

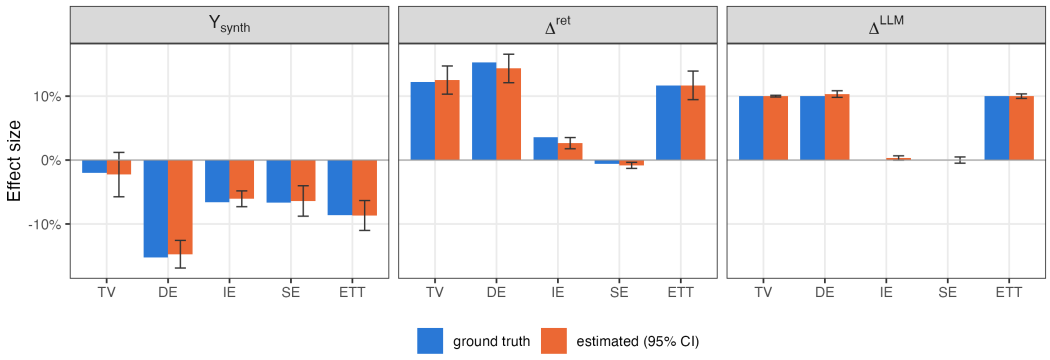


Figure 7: Closed-form ground truth against the cross-fitted one-step estimate, one replication, $n = 8{,}041$. Error bars are 95% confidence intervals from the influence function. Left: $\mathcal{B}(\widetilde{Y})$. Middle and right: the shifts Δ^{ret} and Δ^{LLM}.

Since Z is low-dimensional and tabular, ω is obtained from binned empirical frequencies and is exact up to Monte Carlo error.

Retrieval and aggregation stages. We construct the retrieval stage to return the average synthetic label over the K nearest admissible neighbors of the query $q = (z, w)$ in $\mathcal{D}$,

$$\widehat{Y}_{\text{ret}}(q) \;=\; \mu_{\text{ret}}(z, w) \;=\; \frac{1}{K} \sum_{i \in N_K(q;\mathcal{D})} \widetilde{Y}_i, \tag{43}$$

which is deterministic given $\mathcal{D}$ and carries no x argument. Since $\mathcal{D}$ is disjoint from the estimation sample, μ_{ret} is a single fixed function and no unit's retrieved labels are drawn from the encounters its own nuisance model is fit on. This construction determines the direct effect of the retrieval stage to be exactly 0, since retrieval does not use the X variable, as shown in the following result:

Proposition 3 (Retrieval carries no direct effect) *Let $\widehat{Y}_{ret}$ be as in Eq. 43. Then*

$$x\text{-}DE_{x_0,x_1}(\widehat{Y}_{ret} \mid x_0) = 0 \qquad \textit{and} \qquad \Delta^{ret}DE = -\,x\text{-}DE_{x_0,x_1}(\widetilde{Y} \mid x_0). \tag{44}$$

Proof The direct effect contrasts $\mathbb{E}[(\widehat{Y}_{\text{ret}})_{x_1,W_{x_0}} \mid x_0]$ with $\mathbb{E}[(\widehat{Y}_{\text{ret}})_{x_0,W_{x_0}} \mid x_0]$. By Eq. 43, $\widehat{Y}_{\text{ret}}$ is a measurable function of (Z, W) alone, so the two potential outcomes coincide unit by unit and the contrast vanishes. The second claim follows from $\Delta^{\text{ret}} = \mathcal{B}(\widehat{Y}_{\text{ret}}) - \mathcal{B}(\widetilde{Y})$. ■

Therefore, for the retrieval outcome's attribution, we expect the direct effect to vanish.

For the aggregation stage, instead of an LLM, we use a closed-form aggregator constructed to carry a known direct effect,

$$\widehat{Y}_{\text{RAG}}(q) \;=\; \frac{1}{K} \sum_{i \in N_K(q;\mathcal{D})} \widetilde{Y}_i \;+\; \delta\, \mathbb{1}(X_q = x_1), \qquad \delta = 0.10. \tag{45}$$

The injected term $\delta\, \mathbb{1}(X_q = x_1)$ depends on X_q and on nothing else, so it shifts the direct effect by exactly δ and leaves the indirect and spurious effects unchanged. Thus, the difference in the bias signature is given by $\Delta^{\text{LLM}} = (\delta, 0, 0)$. In the implementation, the estimation procedure takes as input only (X, Z, W, S) and does not use any knowledge of the outcome/retrieval/aggregation mechanisms in Eqs. (41), (43) and (45).

D.1 Estimation Validation

Fig. 7 shows the closed-form ground truth compared against the cross-fitted one-step estimates. We inspect the true and estimated values of $\mathcal{B}(\widetilde{Y})$, and the two shifts Δ^{ret} and Δ^{LLM}.

For $\mathcal{B}(\widetilde{Y})$ (left), the estimates are -14.7%, -6.1% and -6.4% against truths of -15.3%, -6.6% and -6.7% for the direct, indirect and spurious effects, each within the respective confidence interval. The retrieval shift (middle) is recovered as $\Delta^{\text{ret}}\text{DE} = 14.3\%$ against 15.3%, with the indirect and spurious components at 2.6% and -0.8% against 3.6% and -0.6%. The direct effect of $\widehat{Y}_{\text{ret}}$ is the sum of the first two, estimated at -0.4% with a confidence interval containing zero, against the value of zero given by Prop. 3. The aggregation shift is in the rightmost plot. By Eq. 45, the injected term $\delta\,\mathbb{1}(X_q = x_1)$ depends on X_q but not on (Z_q, W_q), so $\Delta^{\text{LLM}}\text{IE} = \Delta^{\text{LLM}}\text{SE} = 0$ exactly. The estimator returns $\Delta^{\text{LLM}}\text{DE} = 10.2\%$ against the injected $\delta = 10.0\%$, with $\Delta^{\text{LLM}}\text{IE} = 0.3\%$ and $\Delta^{\text{LLM}}\text{SE} = 0.0\%$, both confidence intervals containing zero, and $\Delta^{\text{LLM}}\text{TV} = 10.0\%$. The bias introduced at the aggregation stage is thus correctly attributed to that stage and to the direct pathway, while indirect and spurious pathways remain unaffected.

E KIVA: MICROFINANCE LOAN FUNDING

In this appendix, we provide further experimental details for the Kiva dataset investigated in Sec. 3.2. Kiva is an online platform on which borrowers apply for small loans (Choo et al., 2014). Each posting is described by a short narrative, and individual lenders choose which loans to fund. The pipeline is instantiated as a feedback system: given a new posting, similar past postings and their outcomes are retrieved, and the aggregator predicts the outcome for the query. The protected attribute X is borrower sex (x_1 female, x_0 male), the confounder Z is the posting year, and the mediators W include the text description, economic sector, and the loan amount. The outcome Y is the speed of funding: among all funded postings, $Y = 1$ if the share T of the fundraising window elapsed at completion falls below its median, and $Y = 0$ otherwise.

E.1 COHORT

The analysis uses the official Kiva snapshot, which includes $1{,}682{,}790$ postings. We retain single-borrower postings, for which the recorded sex is unambiguous, postings that are funded, and postings with a planned expiration date, which the outcome Y requires. We further restrict to Colombia and to postings carrying a translated description of more than 50 characters. After dropping data from 2012 and 2013 (which hold only a single posting each), this leaves us with a sample of $29{,}003$ postings, with $P(X = x_1) = 62.7\%$ and a raw sex gap of $\mathrm{TV}_{x_0,x_1}(Y) = 11.3\%$.

E.2 REMOVAL OF SEX MARKERS FROM THE MEDIATOR

Kiva descriptions name the borrower and use gendered pronouns throughout, so the raw text nearly determines X. When fit on raw text, $\hat{\pi}(x_1 \mid z, w)$ attains $\mathrm{AUC} = 0.995$ and places 99.8% of units outside $[0.05, 0.95]$: positivity fails and no causal effects are estimable. Fig. 8 shows the distribution.

We therefore remove sex markers from every description before it enters the pipeline, in three deterministic passes. The borrower's own name is taken from the snapshot's `BORROWER_NAMES` field, so it is removed exactly; pronouns are replaced by a neutral referent; and gendered kinship and role nouns are mapped to neutral equivalents, including gendered forms of an occupation, for which the occupation is retained and the marking dropped (e.g., *saleswoman* to *salesperson*). Occupation and sector are otherwise left untouched: they predict sex, but they are mediator content, and removing them would remove the indirect pathway along with the attribute.

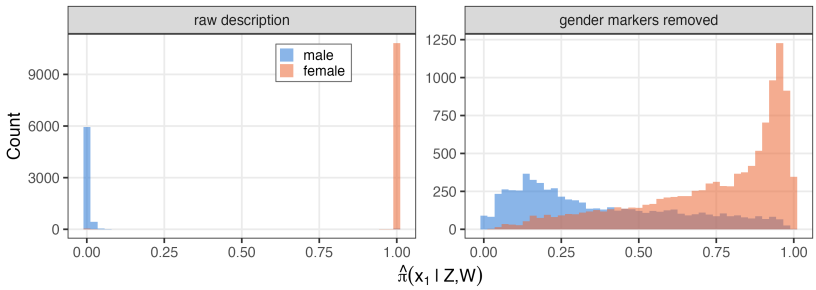


Figure 8: Fitted propensity $\hat{\pi}(x_1 \mid Z, W)$ by recorded sex, before and after sex markers are removed from the description.

Redaction does not fully restore positivity. On the Colombia cohort the propensity still attains $\mathrm{AUC} = 0.846$, with 14.8% of units outside $[0.05, 0.95]$, against roughly zero for the triage application of App. C. We thus perform propensity-trimming at $\varepsilon = 0.05$, and compute the causal effects on the subset of the data where

$$\varepsilon < \hat{\pi}(x_1 \mid Z, W) < 1 - \varepsilon. \tag{46}$$

E.3 RETRIEVER AND DISCLOSURE CONDITIONS

Sec. 3.2 reports dense retrieval with the borrower's sex stated to the aggregator. Tbl. 2 adds the lexical retriever and the condition in which sex is withheld.

		$\mathcal{B}(Y)$			$\mathcal{B}(\widehat{Y}_{\text{ret}})$			$\mathcal{B}(\widehat{Y}_{\text{RAG}})$		
Retrieval	Sex	DE	IE	SE	DE	IE	SE	DE	IE	SE
dense	withheld	$+17.6^{*}$	$+4.5^{*}$	$+0.2$	$+0.6$	$+0.1$	$+0.5$	$+2.4$	$+4.4^{*}$	$+0.6$
dense	stated	$+17.6^{*}$	$+4.5^{*}$	$+0.2$	$+0.6$	$+0.1$	$+0.5$	$+0.5$	$+2.5^{*}$	$+0.6$
lexical	withheld	$+17.6^{*}$	$+4.5^{*}$	$+0.2$	-0.2	-0.9^{*}	$+0.4$	$+1.2$	$+1.2$	$+0.7$
lexical	stated	$+17.6^{*}$	$+4.5^{*}$	$+0.2$	-0.2	-0.9^{*}	$+0.4$	$+1.2$	$+0.5$	$+0.7$

Table 2: Bias signature of each pipeline stage, in percent, after trimming at $\varepsilon = 0.05$. An asterisk marks an effect whose 95% confidence interval excludes zero.

The direct effect of the pipeline is not significantly different from zero in either of the settings with sex withheld, in line with Prop. 3. The aggregator induces a change in the indirect effect for both retrievals. For dense, $\Delta^{\text{LLM}}\text{IE}$ is $+4.3\%$ with sex withheld and $+2.4\%$ with it stated; for lexical, it is $+2.1\%$ with sex withheld, while with it stated $+1.4\%$, with the latter not reaching significance. Since the aggregation step is identical in the two cases and only the retrieved evidence differs, the effects are a property of the retriever–aggregator pair rather than of the aggregator alone.