Measuring and Addressing Indexical Bias in Information Retrieval

Caleb Ziems **William Held** **Jane Dwivedi-Yu** **Diyi Yang**
Stanford University, Georgia Institute of Technology, Meta AI
{cziems, diyi}@stanford.edu, wheld3@gatech.edu, janeyu@fb.com

Abstract

Information Retrieval (IR) systems are designed to deliver *relevant* content, but traditional systems may not optimize rankings for fairness, neutrality, or the balance of ideas. Consequently, IR can often introduce indexical biases, or biases in the positional order of documents. Although indexical bias can demonstrably affect people's opinion, voting patterns, and other behaviors, these issues remain understudied as the field lacks reliable metrics and procedures for automatically measuring indexical bias. Towards this end, we introduce the PAIR framework, which supports automatic bias audits for ranked documents or entire IR systems. After introducing DUO, the first general-purpose automatic bias metric, we run an extensive evaluation of 8 IR systems on a new corpus of 32k synthetic and 4.7k natural documents, with 4k queries spanning 1.4k controversial issue topics. A human behavioral study validates our approach, showing that our bias metric can help predict when and how indexical bias will shift a reader's opinion. For data and code, see `https://github.com/SALT-NLP/pair`

1 Introduction

Web search, recommendation systems, and personal assistants are all powerful Information Retrieval (IR) tools that can help people make decisions (Carroll, 2014; McKay et al., 2020). However, skewed results can lead people to make biased (Novin and Meyers, 2017) or misinformed conclusions (Bar-Ilan et al., 2009; Haas and Unkel, 2017). For example, undecided voters can be swayed to vote for a candidate who is favored in search results (Epstein and Robertson, 2015). This well-known problem is called the search engine manipulation effect (SEME; Allam et al., 2014; Azzopardi, 2021; Draws et al., 2021b; Epstein and Robertson, 2015; Pogacar et al., 2017). SEME results not from the content of any particular document, but rather from

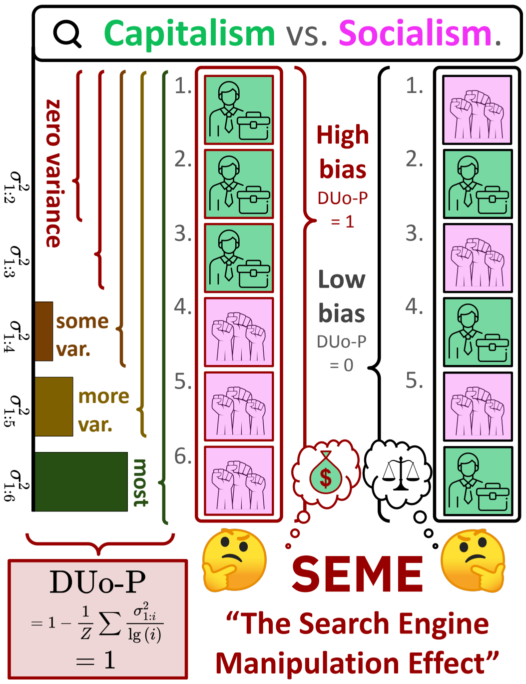

Figure 1: **The Search Engine Manipulation Effect** as predicted by our DUO metric over a set of documents favoring *Capitalism* or *Socialism*. If users read a pro-Capitalism list, they will be more likely to adopt a Capitalist position, and our metric reflects this. The ranking with a most balanced order (*right*) gets the minimal score of DUO=0, whereas the documents with the greatest possible indexical bias (*left*) get the greatest score of DUO=1. DUO uses a discounted sum of variances $\sigma^2_{1:i}$ in polarization across documents' embeddings. On the left, the first 3 *Capitalist* articles have zero variance in polarity, so $\sigma^2_{1:3} = 0$. The full list has a variance of $\sigma^2_{1:6} = 1$, but since this balance appears far down the ranking, $\sigma^2_{1:6}$ is highly discounted.

the *indexical bias* (Mowshowitz and Kawaguchi, 2002) of their rank order,[1] since people are more likely to read and trust higher-ranked documents (Schwarz and Morris, 2011).

In general, responsible IR should address indexical bias by providing not only relevant content,

[1]This is also known as *position bias* (Biega et al., 2018).

Findings of the Association for Computational Linguistics ACL 2024, pages 12860–12877
August 11-16, 2024 ©2024 Association for Computational Linguistics

but also a more fair, balanced, and representative distribution of documents (Olteanu et al., 2021). To identify and rerank biased results at scale, practitioners need automatic metrics that operate over diverse unlabeled corpora. This motivates PAIR.

Perspective-**A**ligned **I**nformation **R**etrieval, or PAIR, is a completely unsupervised method for measuring indexical bias. PAIR introduces the ***D**iscounted **U**niformity **o**f **P**erspectives*, or DUO bias metric, which measures the variance of perspectives at different ranks within an ordered set of retrieved documents. The DUO critically depends on our WIKI-BALANCE corpus, which allows us to compute beforehand the most polarized semantic axis of debate for each issue, using principle component analysis over the document embeddings. DUO is automatic, unlike prior methods that require human labels. PAIR is also generalizable, as WIKI-BALANCE can be easily expanded beyond the 1.4k distinct issues it already supports.

To validate PAIR, we run a behavioral study which demonstrates how DUO can help predict the Search Engine Manipulation Effect (SEME). DUO is predictive whenever participants click at least one search result link. In these cases, reranking documents to minimize DUO will reduce the SEME. Finally, we leverage PAIR to evaluate 7 traditional open-source IR Systems, as well as one commercial search engine. The synthetic WIKI-BALANCE corpus serves as a large and representative stress test, revealing specific weaknesses in leading systems. We complement this test with a more natural evaluation over 4,662 documents retrieved via Google Search on the same issues. In summary, we contribute:

1. The DUO **Positional Bias Metric**, which is an unsupervised metric that works regardless of the controversial issue of interest.
2. Two diverse, large-scale **Bias Evaluation Corpora** with 32k highly polarized synthetic documents and 4,662 natural documents.
3. Extensive **Bias Audit Evaluations** for 8 IR Systems across 15 topical domains.
4. A **Human Behavioral Study** which validates our approach as predictive of the SEME.

2 Related Work

Classifying Stance, Leaning, and Ideology. Political leaning, ideology, public opinion, and stance can be tagged at scale with supervised classifiers (Johnson et al., 2017; Luo et al., 2020; Stefanov et al., 2020; Baly et al., 2020) or keywords (Adamic and Glance, 2005). In PAIR, however, we do not assume access to curated lists or supervised bias classifiers. We also opt not to use zero-shot ideology and stance detection, as LLM performance still varies widely in this domain (Ziems et al., 2023). Instead, we take a fully unsupervised approach and use statistics over generative models.

Diversifying and Debiasing IR. Maximum Marginal Relevance (MMR; Carbonell and Goldstein, 1998) is a popular method for diversifying IR results by minimizing documents' mutual similarity. PAIR embeddings can serve in the MMR similarity metric, but unlike this metric, our DUO metric also incorporates rank order. There are also explicit diversification methods such as IA-Select (Santos et al., 2011), which returns a set with at least one document for each pre-defined category. If such category labels are known in advance, then diversification may be framed as Task-aware Retrieval (Asai et al., 2022) and solved with instruction tuning. Zhao et al. (2024) find such instruction-tuning methods insufficient for perspective imbalance. Instead they redefine the document-query similarity score to condition on a pre-defined perspective p. However, all of these explicit methods may be less generally applicable than DUO due to their reliance on document perspective labels, which must be manually-annotated given the limitations above with zero-shot stance detection.

Auditing Bias in IR. Numerous prior studies evaluate bias in commercial systems (Mowshowitz and Kawaguchi, 2002; Kay et al., 2015; Kulshrestha et al., 2017; Chen et al., 2018; Gao and Shah, 2020; Draws et al., 2021a) like Google and Bing (Gezici et al., 2021). They often measure the diversity of intents (Agrawal et al., 2009; Clarke et al., 2008; Sakai et al., 2010), viewpoints (Draws et al., 2023, 2021a), or fairness towards protected groups (Biega et al., 2018; Yang and Stoyanovich, 2017; Zehlike et al., 2022). Apart from the fairness literature (Rekabsaz and Schedl, 2020), most prior work runs only case-studies of black-box proprietary search engines, and they rely on bias-keywords (Klasnja et al., 2022), classification, or manual annotation along a particular axis of interest, which is typically binary (e.g., left-right) and centered on American ideologies (e.g., Democrat-Republican). In comparison, PAIR lets us evaluate open-source IR systems over thousands of distinct issue topics, which can reveal the relationship between system bias and it's underlying algorithm and data.

3 Foundational Bias Corpora

PAIR relies on two evaluation corpora, WIKI-BALANCE$_{\text{Synthetic}}$ and WIKI-BALANCE$_{\text{Natural}}$. The former will critically support the DUO computation for automatic, cross-domain evaluation of indexical bias in IR. The latter helps us approximate real-world performance. See Table 1, *left* for a comparison, and *right* for statistics.

Source. WIKI-BALANCE reflects 1,364 of the most controversial topics from English Wikipedia, a comprehensive and reliable knowledge source (Bruckman, 2022). High-level seed topics come from the titles of Wikipedia articles that were edited in an oscillatory manner (e.g., *Bullfighting*; *Beyoncé*; *Climate Change*; *the Israel-Palestine Conflict*).[2] For details on the topic distribution, see Table 1 and the discussion in Appendix A.1.

For each topic, we prompt `GPT 3.5 Turbo` to generate 10 specific debate questions on that topic. For example, on *Noam Chomsky*, we generate queries like *Is Noam Chomsky's linguistic theory still relevant?* Figure 2 contains additional examples of WIKI-BALANCE queries. For space they are abbreviated in the figure, but all queries have fully grammatical clauses. We sample a subset of these 138k queries to seed WIKI-BALANCE.

3.1 WIKI-BALANCE$_{\text{Synthetic}}$

LLMs generate WIKI-BALANCE$_{\text{Synthetic}}$. For each of 3,996 randomly sampled debate questions (e.g., *Should Karl Marx be considered a revolutionary thinker?*), we create 8 synthetic documents. First, we prompt a `GPT 3.5 Turbo` model with the debate question and: "*For each side of the issue, write an article from the perspective of that side.*" This induces a single polar axis to divide all subsequent documents.

`GPT 3.5 Turbo` gives regular output: each article begins with a header describing the position; e.g., ***Perspective 1:*** *Karl Marx as a Revolutionary Thinker*; ***Perspective 2:*** *Karl Marx as a biased ideologist*). We use regular expressions to extract them, and continue to prompt the model 3 more times: *Given the question above, write an article from the perspective of the side below.* This produces 32k polarized synthetic documents (see Figure 2). In this way, each document is marked as relevant to one out of 3,996 queries. Each query in the WIKI-BALANCE$_{\text{Synthetic}}$ corpus corresponds to a balanced set of documents, with each half supporting a different perspective.

3.2 WIKI-BALANCE$_{\text{Natural}}$

This Natural corpus helps us (1) evaluate the Google Search engine, and (2) evaluate open-source IR systems in a real-world setting where documents are less polarized. With 452 randomly-sampled queries from WIKI-BALANCE$_{\text{Synthetic}}$, we scrape publicly-available natural web documents from the top 10 results of Google Search in October 2023. To reduce noise, we keep only HTML documents' body text.

3.3 WIKI-BALANCE Quality Estimation

Human Evaluation. To measure the quality of WIKI-BALANCE$_{\text{Synthetic}}$, we recruit domain experts[3] from Upwork to blindly evaluate 10-20 random query-document pairs. First, evaluators consider the *subjectiveness* and *topical relevance* of the query. A subjective question does not have a single correct answer, and a relevant query relates to the topic in an interesting and well-specified way. Table 1 shows that queries are sufficiently subjective and highly relevant to their respective topics. Annotators also score our synthetic documents for faithfulness (Durmus et al., 2020), coherence (Dang, 2005), relevance, and fluency, showing that they are high in each of these respects.

Safety. Although WIKI-BALANCE is centered around controversial issues, we want to reduce risk by measuring and addressing any toxic, harmful, or otherwise unsafe content contained in the documents. Using the OpenAI Moderation API, we determine that no document contains hate, harassment, self-harm, or unwarranted sexual content with a model confidence score larger than 0.09.

[2] The subjects of edit wars, NPOV disputes, edit restrictions, or otherwise frequent content revisions, reversions and rollbacks; `https://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues`

[3] For *Entertainment, History, Religion, and Sports*, we recruit a Graduate Student with a B.S. in Journalism. For *Politics, Sexuality, Law, and Media*, we find a Public Policy Graduate Student with a B.A. in Political Science. For *Psychiatry, Technology, and People* we enlist a Nurse in Clinical Behavioral Health with a B.A. in Psychology. And for *Science and Environment*, we recruit a former CDC health communication specialist with a B.S. in Public Health and an M.S. in Health Education. For *Languages and Philosophy*, we hire a former writing expert at Grammarly with an M.F.A.

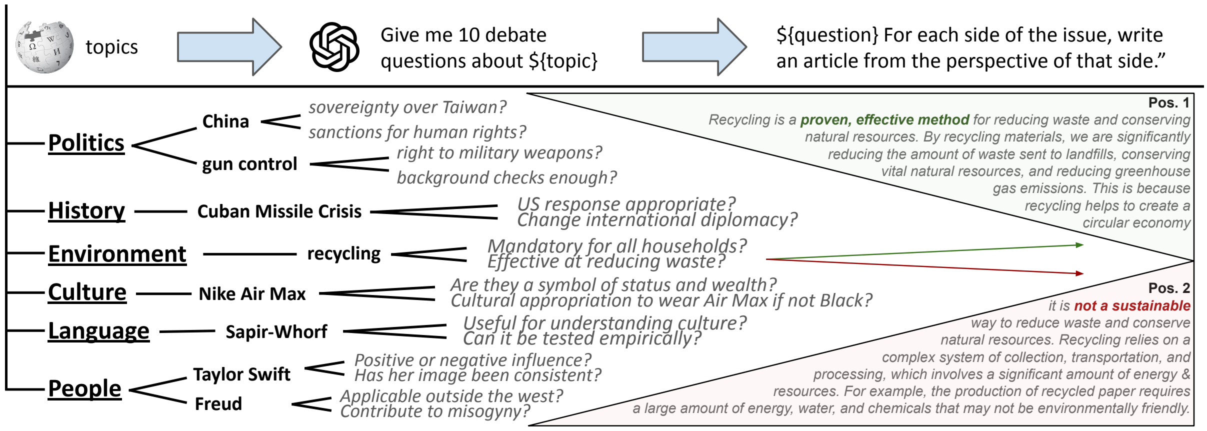

Figure 2: **WIKI-BALANCE Pipeline.** First we pull 1,364 controversial topics from English Wikipedia. Data is under CC BY-SA License and is consistent with intended use. Then we generate 10 debate questions about each topic (examples are abbreviated in this figure). For each debate question, we generate 8 polarized documents, with 4 on each side of the initial axis generated by the LLM.

We manually verified that the most triggering documents were benign cases of almost journalistic reporting on their respective topics (e.g., a debate on the key ingredients of sexual gratification).

4 Indexical Bias Metrics

4.1 Prior Metrics

Indexical bias arises when documents of class A receive greater visibility in search results than do documents of class B. Due to primacy effects (Ho and Imai, 2008), the document's rank index can serve as a proxy for visibility, as higher-ranked documents will be more visible (Joachims et al., 2007; Pan et al., 2007; Baeza-Yates, 2018), and thus more frequently clicked (Insights, 2013). Discounted Cumulative Gain (DCG) assumes a document's visibility, or the attention it receives, is inversely proportional to the logarithm of its index. Although log-based decay may not exactly reflect user attention (Ghosh et al., 2021; Sapiezynski et al., 2019), this has become the standard in IR. DCG is thus defined according to Equation 1.

$$\mathrm{DCG}(r,u) = \sum_{i=1}^{|r|} \frac{u(i,r)}{\log_2 i} \tag{1}$$

where $u(i,r)$ is the utility of ranking r up to and including document i. Researchers can normalize the Discounted Cumulative Gain by setting

$$\mathrm{nDCG}(r,u) = \frac{\mathrm{DCG}(r,u) - \min_{r'}\{\mathrm{DCG}(r',u)\}}{\max_{r'}\{\mathrm{DCG}(r',u)\} - \min_{r'}\{\mathrm{DCG}(r',u)\}} \tag{2}$$

That is, we set the minimum to zero and divide by the metric's highest attainable value for the given number of items $|r|$ and metric parameters. This means all measures will reach their best, most fair value at 0, and their worst value at 1.

Following the group fairness literature (Pedreschi et al., 2009; Pedreshi et al., 2008), we can define $u(i,r)$ as statistical parity in the visibility of a protected group (Pitoura et al., 2022). In the standard formulation of Yang and Stoyanovich (2017), they set rND $=$ nDCG(r, u_{ND}) with

$$u_{ND}(i,r) = P_g@i - P_g@\,|r| \tag{3}$$

which is the difference between the proportion of protected group g members in the top i against the proportion of protected group g members in the population (i.e., the full ranking). The corresponding rND metric is convex and continuous, but not differentiable at zero. To further smooth this metric, Yang and Stoyanovich (2017) also consider the *KL divergence* between protected group membership in the top i vs. the full ranking, giving rKL $=$ nDCG(r, u_{KL}) with

$$\begin{aligned} u_{KL}(i,r) = &-P_g@i \log\left(\frac{P_g@\,|r|}{P_g@i}\right) \\ &-\left(1 - P_g@i \log\left(\frac{1-P_g@\,|r|}{1-P_g@i}\right)\right) \end{aligned} \tag{4}$$

To apply fairness metrics rND and rKL to the most general case of indexical bias, we can abstract group membership g to indicate whether a document is polarized in a particular direction. We can safely ignore any prior metric for which this generalization would not apply, such as the

	WIKI-BALANCE Synthetic	Natural
Domains	15	15
Topics	1,364	288
Queries	3,996	452
Documents	31,534	4,662
Google Search	✗	✓
Gold Labels	✓	✗
Applies: rND	✓	✗
Applies: rKL	✓	✗
Applies: DUO	✓	✓

Domain	Topics	Queries	Docs	Query Quality Relev.	Query Quality Subj.	Synthetic Document Quality Faith.	Coh.	Relev.	Flu.
Entertain.	26	66	528	4.3	3.8	5.0	5.0	5.0	4.8
History	122	382	3,004	4.9	3.5	5.0	5.0	4.5	4.8
Law	15	41	324	3.3	4.9	4.8	5.0	4.8	5.0
Culture	110	323	2,550	4.5	5.0	4.8	4.8	4.7	5.0
Politics	243	703	5,576	4.4	4.9	4.7	4.5	4.8	5.0
Religion	112	334	2,638	5.0	3.0	4.7	4.7	4.3	4.6
Sexuality	86	249	1,990	4.3	5.0	4.8	5.0	5.0	5.0
Sports	20	57	444	4.9	3.0	5.0	5.0	4.6	4.7
Mean	91	266	2,102	4.5	4.1	4.9	4.9	4.7	4.9

Table 1: (*Left*) **WIKI-BALANCE statistics** for both Synthetic and Natural corpora, which both use the same topics and queries, but the latter is much smaller and lacks gold labels, so previous metrics rND and rKL do not apply. (*Right*) **Quality audit of a random sample of WIKI-BALANCE according to human raters.** Humans perceive most queries to be Relevant (Relev.) and sufficiently subjective (Subj.) for use in this task. Synthetic documents are highly Faithful (Faith.), Coherent (Coh.), Relevant to the Query (Relev.) and Fluent (Flu.), which gives us confidence in the validity of the WIKI-BIAS resource.

keyword-bias metric of Rekabsaz et al. (2021) and the Normalized Discounted Ratio of Yang and Stoyanovich (2017), which assumes g is a minority group ($P_g@|r| < 0.5$). We will also focus entirely on *ranking* bias metrics, and ignore Kulshrestha et al. (2019) and others who measure the bias of individual documents.

4.2 The DUO Bias Metric

Clearly, rND and rKL can be used only in cases where document polarization labels are known, such as when: (1) we have manually annotated the corpus according to the target axis, or (2) when we generate a controllable synthetic corpus like WIKI-BALANCE$_{\text{Synthetic}}$. Option (1) is not scalable, especially with thousands of distinct axes of controversy in §3. Option (2) does not apply to evaluations in real-world settings. This motivates us to build an unsupervised metric for indexical bias that can operate automatically, even in real-world settings, using scalable knowledge from LLMs.

Our proposed bias metric is the "***D**iscounted **U**niformity of **O**pinions*" (DUO). This unsupervised metric critically depends on our synthetic data to determine the axis of polarization for each query.[4] Given a topic t represented by a query q_t, we pull from WIKI-BALANCE$_{\text{Synthetic}}$ a set of $|r|$ synthetic documents that argue each opposing perspective for q_t following §3.1. We use a transformer-based document encoder to embed each document into a dense vector, and by running PCA, we project document embeddings into scalar polarization scores $p_j \in \mathbb{R}$, with an average score of $\bar{p} = \frac{1}{|r|}\sum_{i=1}^{|r|} p_j$.

[4]It is important to note that even for our DUO computations on the Natural corpus, we rely on WIKI-BALANCE$_{\text{Synthetic}}$ here to determine the axis of bias.

This allows us to define

$$u_V(i,r) = \frac{1}{i}\sum_{j=1}^{i}(p_j - \bar{p})^2 \qquad (5)$$

so that the utility of a ranked subset is the variance of its polarization scores. Now this utility is the complement of a bias metric: more variance in the polarization scores indicates a better balance of ideas and *less* indexical bias. Thus for convenience in this paper, we will refer to our normalized DUO(r) metric by the following equation:

$$\text{DUO}(r) = 1 - \text{nDCG}(r, u_V) \qquad (6)$$

5 Validating DUO

Experimental evidence demonstrates the validity of the DUO metric. Our first experiment shows that the unsupervised polarization score that grounds DUO can accurately partition a corpus of polarized documents into their respective viewpoints (§5.1). Our second experiment shows that DUO can help predict the Search Engine Manipulation Effect in a real behavioral manipulation (§5.2). Together, these results demonstrate both the fundamental and psychological validity of DUO.

5.1 Validation with Synthetic Data

Since DUO depends on reliable polarization scores to compute viewpoint variance, the first validation step is to evaluate the accuracy of these scores. Accurate polarization scores should partition a set of related documents into contrasting subsets for each viewpoint, where documents with positive scores endorse Perspective 1, and documents with negative scores endorse Perspective 2 for a given query

Embedding Model	**Accuracy**
Worst: `clip-ViT-B-32-multilingual-v1`	73.19%
Mean Accuracy	85.54%
Median Accuracy	85.70%
Best: **`sentence-t5-xl`**	**95.27%**

Table 2: **Polarization score accuracies** for the worst, best, median, and mean performance among all 124 evaluated models. High mean and best performances validate our approach.

(or vice versa). With viewpoint labels from §3.1 as ground truth, we compute the maximum accuracy for each query, and average over all queries in WIKI-BALANCE$_{\text{Synthetic}}$.

Accuracy depends on our document embedding model, so we evaluate all 124 of the models associated with Sentence BERT (Reimers and Gurevych, 2019), a transformer-based document similarity metric widely used in the IR community. We also evaluate the recent `voyage-02` embeddings from Voyage AI.[5] Table 2 provides the worst accuracy (73%), best accuracy (95%), and the median (87%) and mean (86%) accuracy across all 124 evaluations. These validate our approach since, in the best case, with every third query there is only one incorrect document label among the 8 documents retrieved. For more detailed results, see Tables 6 and 7 in Appendix A.3.

5.2 Validation with A Behavioral Study

Experimental Design. A behavioral study can demonstrate the psychological validity of the DUO metric as predictive of the Search Engine Manipulation Effect (SEME). For each query $q \in$ WIKI-BALANCE, a randomly sampled participant will have some opinion $o_{\text{prior}} \in \mathbb{Z}$. SEME predicts that, after this participant considers a biased list of search results relevant to the given query, their final opinion $o_{\text{posterior}} \in \mathbb{Z}$ will be shifted with a magnitude proportional to the magnitude of the indexical bias. Our null hypothesis is that DUO is unrelated to the effect, so the coefficient β_2 is zero in the following regression:

$$o_{\text{posterior}} = \beta_0 + \beta_1 o_{\text{prior}} + \beta_2 \mu^+_{\text{DUO}} + \epsilon$$

To validate DUO, we would reject the null hypothesis. Here, μ^+_{DUO} is a signed copy of DUO (see Appendix A.2 for a derivation), which by default measures only the magnitude of the bias and not its direction. We need a sign to indicate the *direction* of the bias because we expect that shift will move towards the favored perspective.

[5] https://www.voyageai.com/

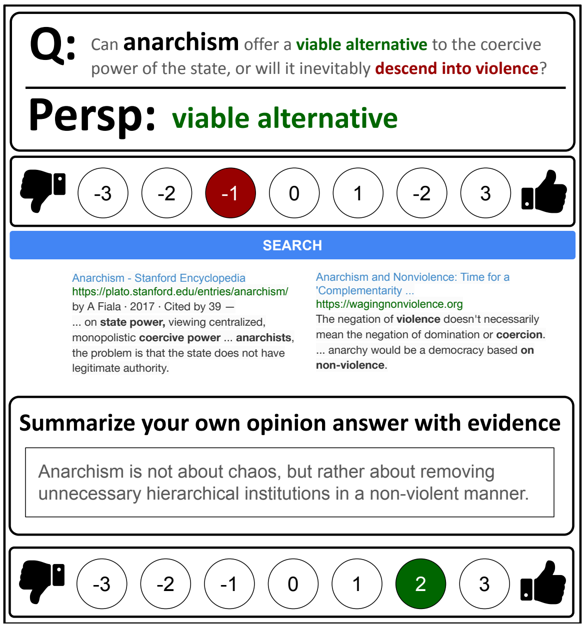

Figure 3: **Human Behavioral Study Interface** to help determine whether biased search results lead to the SEME. Participants read a query (Q) with a given Perspective (Persp) and tell us whether they agree (3) or disagree (-3) with Perspective. After reading a manipulated list of up to 10 search results, they summarize their informed opinion and provide us their updated agreement on a scale from -3 to 3. We expect more biased results to more radically shift their opinions.

Experiment. American adult participants are recruited from Prolific. Each participant interacts with an interface like that shown in Figure 3. At the top is a query and a given perspective, like *anarchism is a viable alternative to the coercive power of the state.* The participant then provides their initial opinion on a 7-point Likert scale, $o_{\text{prior}} \in \{-3, -2, -1, 0, 1, 2, 3\}$, where -3 indicates strong disagreement with the listed perspective, and 3 indicates strong agreement. After entering this prior, the participant uses our manipulated search engine, which retrieves a list of up to 10 search results. Participants are randomly assigned to the experimental manipulation: either results are ordered (1) with the *maximum* bias or (2) *minimum* bias, according to the DUO metric. We log any article links the participant clicks, assuming participants are motivated to read these results, since it can assist them in the penultimate task question: summarizing their opinion and quoting evidence. The participant concludes the task after providing $o_{\text{posterior}}$ on a similar Likert scale.

Corpus	Behavior	N	β_2	$P(\beta_2 = 0)$	R^2
Synthetic	All	200	0.059	0.673	0.364
Synthetic	Clicked	19	0.255	0.566	0.689
Natural	All	225	0.140	0.253	0.474
Natural	Clicked	99	0.392	0.036	0.489
Combined	Clicked	118	0.365	0.032	0.519

Table 3: **Regression Results on the Significance of Duo in a SEME Behavioral Study** over both the Natural and Synthetic coprora. In natural experiments where participants clicked at least one article link (Behavior=**Clicked**), we observe significant ($p < 0.05$) positive β_2 coefficients, leading us to conclude that Duo helps predict the SEME in cases of article click-through, and thus validating our method.

Results. We run our experiment for $N = 200$ evaluations on each of our two corpora, as shown in Table 3. By logging click behavior, we discover that participants are *not* as often motivated to read search results in the Synthetic corpus as they are in the Natural corpus. With Synthetic, 10% of users clicked at least one article, while in the Natural, half of users clicked at least one article. Differential click-through behavior effects our findings. When we limit our regression to only those trials where a user clicked at least one link (Behavior: Clicked), we can reject the null hypothesis with statistical significance. So **only in cases of article click-through, Duo significantly helps predict the Search Engine Manipulation Effect** with an R^2 effect size greater than 0.48 in both the Natural and Combined corpora, $p < 0.05$. These findings are robust, as they replicate with the best and average embedding models (see Table 9 in Appendix A.6). From this, we conclude that **Duo is psychologically valid as it helps predict SEME**.

6 IR Bias Audits

With Duo (§5) and evaluation corpora (§3), we can audit the indexical bias of leading IR models. The labeled synthetic corpus shows us how Duo compares with prior bias metrics. However, synthetic results may not reflect end-model behavior on real document distributions. For more realistic results, we evaluate on the Natural corpus, where SEME behavioral experiments clearly validate Duo as predictive. The results in Tables 4 and 5 use both standard relevance metrics and the bias metrics from §4, and are averaged over 3 random seeds. Along with concurrent work (Zhao et al., 2024), Pair is among the first cross-domain audits of indexical biases in open-source IR systems.

6.1 Models

Following BEIR (Thakur et al., 2021), we use BM-25 as a strong lexical baseline, and evaluate six additional open-source neural systems, as well as one industrial system. For sparse models, we evaluate SPARTA (Zhao et al., 2021) and SPLADE (Formal et al., 2021). Next, we consider three dense models, ANCE (Xiong et al., 2020), SBERT (Reimers and Gurevych, 2019), and Use-QA (Yang et al., 2021). Our late-interaction model is ColBERT (Khattab and Zaharia, 2020). Finally, we evaluate Google's search engine. Our document embedding model is the best-performing `sentence-t5-xl`.

6.2 Aggregate Bias Results

Here, **the most relevant models are not always the least biased.** Although SPLADE produces the most relevant results, it also introduces the most indexical bias in the natural evaluation setting (Duo$= 0.62$). On the other hand, Use-QA is the least relevant model, yet it produces the least biased rankings in both the natural and synthetic evaluation (Duo$= 0.64, 0.57$).

Most importantly, bias results validate our Duo metric, as **Duo highly correlates with the supervised metrics rND and rKL**, with Spearman correlations of 0.80 and 0.83 respectively ($p < 0.05$). Thus Duo gives us similar conclusions about model bias without the need for human annotation. On the synthetic data, Use-QA and SPARTA are the least biased (Duo≤ 0.58; rKL $= 0.60$), followed by ColBERT, SPLADE, and BM-25 (Duo$= 0.60$). The most biased models are ANCE and SBERT (Duo≥ 0.61; rKL≥ 0.62). These results are stable; even if we compute Duo using a less accurate embedding model, the relative model order is roughly preserved ($\rho = 0.72$; see Appendix A.6).

Model bias on the synthetic corpus also weakly predicts its bias on the natural corpus, with a strong Spearman correlation of 0.64 between synthetic and natural Duo scores. ANCE remains the most biased model, while Use-QA remains the least biased open-source model. Google search is the least biased overall (Duo = 0.63). We conclude that synthetic evaluation may be used as a surrogate in this way to quickly evaluate IR systems across a wide range of domains. However, natural data remains the gold standard and should not be replaced by synthetic evaluations alone. Discrepancies between the synthetic and natural results may reflect differences between the distributions

		Relevance: *Synthetic*		**Bias:** *Synthetic*			**Relevance:** *Natural*		**Bias:** *Natural*
Class	Model	nDCG@1	@10	rND	rKL	DUO	nDCG@1	@10	DUO
Lexical	BM-25	0.99	0.97	0.66	0.61	0.60	0.87	0.75	0.66
Sparse	SPARTA	0.99	0.95	0.65	0.59	0.58	0.79	0.65	0.65
	SPLADE	1.00	0.97	0.66	0.60	0.60	0.83	0.75	0.67
Dense	ANCE	0.99	0.96	0.67	0.62	0.62	0.81	0.71	0.67
	SBERT	0.99	0.96	0.68	0.62	0.61	0.86	0.75	0.65
	Use-QA	0.95	0.88	0.66	0.60	0.57	0.78	0.68	0.64
Late	ColBERT	1.00	0.97	0.66	0.60	0.60	0.83	0.72	0.65
Industry	Google	*N/A*†	*N/A*†	*N/A*†	*N/A*†	*N/A*†	*N/A*†	*N/A*†	0.63

Table 4: **Aggregate relevance and bias results** over WIKI-BALANCE$_{\text{Synthetic}}$ (*left*) and WIKI-BALANCE$_{\text{Natural}}$ (*right*) demonstrate how the most relevant models are not always the least biased. Use-QA and SPARTA have the lowest bias scores, but they are also the least relevant. SPLADE is the most relevant, but also introduces the most indexical bias in the Natural setting. Best results are green, and worst results are red. †*N/A* indicates that the metric is not applicable because it requires external human labels.

Class	Model	Entertainment		Environment		History		Languages		Law & Order		Media & Culture	
Lexical	BM-25	0.63	*0.71*	0.59	*0.62*	0.60	*0.70*	0.58	*0.69*	0.62	*0.62*	0.58	*0.63*
Sparse	SPARTA	0.63	*0.74*	0.61	*0.67*	0.56	*0.63*	0.60	*0.67*	0.58	*0.67*	0.58	*0.66*
	SPLADE	0.63	*0.68*	0.62	*0.67*	0.62	*0.67*	0.58	*0.71*	0.59	*0.66*	0.57	*0.67*
Dense	ANCE	0.69	*0.69*	0.63	*0.68*	0.64	*0.68*	0.62	*0.67*	0.62	*0.64*	0.61	*0.66*
	SBERT	0.64	*0.67*	0.64	*0.66*	0.60	*0.65*	0.60	*0.64*	0.58	*0.68*	0.60	*0.63*
	Use-QA	0.58	*0.70*	0.60	*0.62*	0.57	*0.67*	0.57	*0.65*	0.56	*0.68*	0.57	*0.63*
Late	ColBERT	0.62	*0.65*	0.62	*0.65*	0.62	*0.66*	0.57	*0.69*	0.57	*0.69*	0.59	*0.68*
Industry	Google	N/A	*0.69*	N/A	*0.66*	N/A	*0.66*	N/A	*0.61*	N/A	*0.63*	N/A	*0.59*

Class	Model	People		Politics & Econ		Psychiatry		Science		Sex & Gender		Technology	
Lexical	BM-25	0.59	*0.67*	0.59	*0.67*	0.56	*0.65*	0.60	*0.68*	0.59	*0.60*	0.58	*0.76*
Sparse	SPARTA	0.59	*0.63*	0.58	*0.65*	0.55	*0.52*	0.56	*0.60*	0.59	*0.58*	0.59	*0.65*
	SPLADE	0.60	*0.67*	0.60	*0.66*	0.53	*0.62*	0.62	*0.69*	0.59	*0.59*	0.59	*0.72*
Dense	ANCE	0.61	*0.67*	0.61	*0.71*	0.65	*0.70*	0.62	*0.69*	0.58	*0.61*	0.61	*0.71*
	SBERT	0.60	*0.66*	0.61	*0.65*	0.71	*0.74*	0.61	*0.66*	0.58	*0.65*	0.59	*0.74*
	Use-QA	0.57	*0.64*	0.56	*0.66*	0.49	*0.68*	0.60	*0.57*	0.53	*0.58*	0.57	*0.65*
Late	ColBERT	0.60	*0.69*	0.61	*0.62*	0.56	*0.56*	0.61	*0.65*	0.58	*0.54*	0.57	*0.70*
Industry	Google	N/A	*0.59*	N/A	*0.60*	N/A	*0.56*	N/A	*0.60*	N/A	*0.58*	N/A	*0.67*

Table 5: **Domain-Level WIKI-BIAS results** over WIKI-BALANCE$_{\text{Synthetic}}$ (*left columns*) and WIKI-BALANCE$_{\text{Natural}}$ (*italicized right columns*) can help identify entry points for critical bias-mitigation efforts. *SBERT*, one of the most most overall biased models, specifically struggles with *psychiatry*, *entertainment*, and the *environment*.

of their document polarizations. Synthetic data follows a highly-polarized bimodal distribution, while natural bias scores are both more neutral, and also normally distributed (see Appendix A.4). The respective evaluations are mutually complementary.

Overall, **the orderings between models above are relatively stable**, even when we consider alternative embedding models, or when we apply debiasing methods to the embedding process to remove possible spurious correlations that arise from WIKI-BALANCE$_{\text{Synthetic}}$ (see Appendix A.6 for more details on Experimental Replications, and Appendix A.5 for methods to remove spurious artifacts).

6.3 Aggregate Relevance Results

Our relevance results confirm prior work (Thakur et al., 2021). Table 4 shows *BM-25* is the strongest baseline for relevant retrieval on both WIKI-BALANCE corpora, and that *BM-25* beats out models of greater complexity like *ANCE* and *SPARTA*. *ColBERT* also achieves the top relevance scores on the WIKI-BALANCE$_{\text{Synthetic}}$ corpus, which also aligns with Thakur et al. (2021) and sanity-checks our results. The nDCG@10 relevance scores are all higher in the left side of the table, showing unsurprisingly that our synthetic corpus is an easier task than natural web retrieval.

6.4 Domain-Level Results

One of PAIR's benefits is the ability to evaluate models domain-specific biases. Small aggregate differences in bias performance may not sway industry leaders and practitioners to adopt an entirely new IR system, but if an operational system demonstrates weakness in a particular domain, that can become a focal point for bias mitigation, like debiasing embeddings. Here, **PAIR can serve as a precise instrument for diagnosing and addressing localized indexical biases**, as in this section.

Table 5 decomposes aggregate bias results into focal domains, revealing weaknesses in even the best models. The best open model, Use-QA, still falls short of Google Search in three key domains: *Psychiatry* (+0.12 more bias than Google), *Politics* (+0.06 DUO), and *Law* (+0.05 DUO). Since indexical bias in political search results can affect voting behavior (Epstein and Robertson, 2015), practitioners may have strong incentives to mitigate such biases in open source systems.

In politics as in the aggregate, Google Search performs with the least bias on natural web data. However, given the power of any prominent search engine to influence countless users, it could be strategic for such search companies to invest greater attention towards bias mitigation at the weakest points. For Google, a weakness is the *Environment* (+0.04 more biased than the best Use-QA model).

7 Applications and Extensions

The PAIR framework is general, and future work can explore its extensions outside of search in other domains where information follows a rank order, such as a chatbot's probability-ranked utterances, a politician's most common phrases, or even the ordered paragraphs of a news article (i.e., *framing*; Ziems and Yang, 2021). Indexical bias evaluation will become increasingly relevant to mitigate harms in more recent generative methods which rely on Retrieval Augmented Generation for grounding knowledge (Lewis et al., 2020; Shuster et al., 2022; Jiang et al., 2023; Khattab et al., 2022).

A natural extension of the DUO could also handle issues with more than two sides, simultaneously incorporating multiple axes of semantic variation. Our formalization is readily prepared for such an extension. If we generate up to m different viewpoint axes for each issue in WIKIBALANCE$_{\text{Synthetic}}$, we could simply increase the dimensionality of the PCA projection in §4 such that polarization scores become polarization vectors $P_j \in \mathbb{R}^m$. This would allow a separate variance utility computation $u_v^x(i,r) = \frac{1}{i}\sum_{j=1}^{i}(P_j^x - \bar{P}^x)^2$ and thus a separate score $\text{DUO}^x(r)$ for each viewpoint axis x. Depending on the application, one could aggregate across axes x, for example by taking the maximum bias score as in the multi-group Attention Bias Ratio of Ghosh et al. (2021). For more discussion of multi-group extensions of indexical bias metrics, see the Group Relevance Framework of Sakai et al. (2023).

Additionally, there is still much to be learned about the formal and mathematical properties of the DUO metric, especially in the context of optimization and reranking. Prior works suggest that fluctuations in fairness metrics like DUO can lead to unstable training (Rekabsaz et al., 2021). This may pose a challenge to the integration of DUO with current systems. One other bottleneck that may prevent the widespread adoption of DUO in reranking is the expensive computation for normalization. Our code includes a stochastic approximation, but the code could be further optimized.

8 Conclusion

From web search to personal assistants, IR systems have the potential to skew users' opinions on a wide range of topics, from media and entertainment to political issues and scientific insights. Before one can address the problematic outcomes of such manipulation, one should first expect to measure *indexical bias*, the root of this psychological effect. PAIR is the first completely automatic method for evaluating indexical bias in IR systems without the need for manual human annotations. PAIR is built on two bias evaluation corpora, which support our DUO metric. Since DUO requires no human supervision, it can serve as a scalable evaluation metric and, in future work, as an automatic reranking criterion. We demonstrate the psychological validity of DUO using a controlled experiment. After proving its validity, we use DUO to run an extensive audit over the biases in current IR technologies, both open and closed-source. Together, these contributions provide a basis for future efforts to measure and address indexical bias in IR.

Limitations

Complementary Notions of Fairness. All of the methodology we introduced in this work was focused on *fairness of exposure* (Diaz et al., 2020), balancing rankings to ensure *equal visibility* (Pessach and Shmueli, 2022) between groups or ideas. However, the appropriateness of fair exposure depends on the context (Singh and Joachims, 2018). It may not always be socially desirable to balance certain viewpoints, as this may elevate or amplify hate, espouse misinformation, or jeopardize personal or collective well-being. For any practitioners interested in applying PAIR or using the WIKIBALANCE corpus, we strongly encourage a more careful selection of the topics and domains over which DUO balance is optimized.

Other related works have measured the complementary objective of *fairness through neutrality* (Zerveas et al., 2022), where systems are encouraged to preferentially retrieve more factual, neutral, and unbiased documents and to omit the most polarized documents entirely. Fairness through neutrality is critical and should not be ignored, especially for highly sensitive or ideological domains and settings where users may be most susceptible to confirmation bias (Del Vicario et al., 2017) and its negative outcomes.

Still, we note that *fairness through neutrality* is not always attainable with a polarized corpus, nor is the notion of neutrality applicable to all queries (Krieg et al., 2023; Zerveas et al., 2022). For example, undecided voters may query a factual, encyclopediac corpus of Wikipedia articles on the biographies of candidates in an upcoming election. Voters opinions can shift by mere exposure to preferentially ranked biographies (Epstein and Robertson, 2015). Here there is no truly neutral document, despite the academic, factual tone they carry.

On the other hand, DUO applies both to highly polarized and more neutral corpora. This is because DUO measures indexical bias as a relative quantity—how biased the ranked results are relative to the most imbalanced possible ranking of those same documents. In doing so, we disentangle system bias from any document biases in the corpus itself (Kulshrestha et al., 2017).

Methodological Bias. This work seeks to measure bias in IR systems. However, it is important to acknowledge potential biases in the PAIR evaluation process itself. It is a non-trivial task to create an unbiased IR evaluation corpus. Just as traditional crowd-annotated datasets are prone to subjectivity biases in the document selection, relevance scoring, and other steps in the annotation pipeline, so also are synthetic methods vulnerable to such viewpoint biases, which may derive from the distribution of the LLM pretraining corpus, the prompt design, the seed topics used for prompting, or other related variables. The topics represented in WIKIBALANCE were drawn automatically from English Wikipedia and extrapolated with language models. The authors of this paper did not hand-select any topics or documents, nor do we endorse any particular viewpoints in these resources. In the Appendix A.1, we more thoroughly discuss biases in the seed topics, and we encourage future work to carefully consider and expand on this discussion.

Ethics

This study has been approved by the Institutional Review Board (IRB) at the researchers' institution, and participant consent was obtained using the standard institutional consent form. For the annotation process, we included a warning in the instructions that the content might be offensive or upsetting. Annotators were also encouraged to stop the labeling process if they were overwhelmed, and regardless of how may tasks they completed, participants were paid a fair stipend of $20 per hour for their time.

Acknowledgements

We are thankful to the members of SALT Lab and the Stanford NLP Group for their helpful feedback on the draft. We especially appreciated suggestions from Tiziano Piccardi, Jing Huang, Faye Holt, Dora Zhao, Julia Kruk, and Michael Ryan. The work has been supported by a grant from Meta. Caleb Ziems is supported by the NSF Graduate Research Fellowship under Grant No. DGE-2039655.

References

Lada A Adamic and Natalie Glance. 2005. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43.

Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Proceedings of the Second International Conference on Web Search and Web Data Mining, WSDM 2009, Barcelona, Spain, February 9-11, 2009*, pages 5–14. ACM.

Ahmed Allam, Peter Johannes Schulz, Kent Nakamoto, et al. 2014. The impact of search engine selection and sorting criteria on vaccination beliefs and attitudes: two experiments manipulating google output. *Journal of medical internet research*, 16(4):e2642.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2022. Task-aware retrieval with instructions. *arXiv preprint arXiv:2211.09260*.

Leif Azzopardi. 2021. Cognitive biases in search: a review and reflection of cognitive biases in information retrieval. In *Proceedings of the 2021 conference on human information interaction and retrieval*, pages 27–37.

Ricardo Baeza-Yates. 2018. Bias on the web. *Communications of the ACM*, 61(6):54–61.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Judit Bar-Ilan, Kevin Keenoy, Mark Levene, and Eti Yaari. 2009. Presentation bias is significant in determining user preference for search results—a user study. *Journal of the American Society for Information Science and Technology*, 60(1):135–149.

Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 405–414. ACM.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Amy S Bruckman. 2022. *Should you believe Wikipedia?: online communities and the construction of knowledge*. Cambridge University Press.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, page 335–336, New York, NY, USA. Association for Computing Machinery.

Noel Carroll. 2014. In search we trust: exploring how search engines are shaping society. *International Journal of Knowledge Society Research (IJKSR)*, 5(1):12–27.

Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the impact of gender on rank in resume search engines. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, page 651. ACM.

Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 659–666. Association for Computing Machinery.

Hoa Trang Dang. 2005. Overview of duc 2005. In *Proceedings of the document understanding conference*, volume 2005, pages 1–12. Citeseer.

Michela Del Vicario, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2017. Modeling confirmation bias and polarization. *Scientific reports*, 7(1):40391.

Fernando Diaz, Bhaskar Mitra, Michael D Ekstrand, Asia J Biega, and Ben Carterette. 2020. Evaluating stochastic rankings with expected exposure. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 275–284.

Tim Draws, Nirmal Roy, Oana Inel, Alisa Rieger, Rishav Hada, Mehmet Orcun Yalcin, Benjamin Timmermans, and Nava Tintarev. 2023. Viewpoint diversity in search results. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*, pages 279–297. Springer.

Tim Draws, Nava Tintarev, and Ujwal Gadiraju. 2021a. Assessing viewpoint diversity in search results using ranking fairness metrics. *ACM SIGKDD Explorations Newsletter*, 23(1):50–58.

Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2021b. This is not what we ordered: Exploring why biased search result rankings affect user attitudes on debated topics. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 295–305.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Robert Epstein and Ronald E Robertson. 2015. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2288–2292.

Ruoyuan Gao and Chirag Shah. 2020. Toward creating a fairer ranking in search engine results. *Information Processing & Management*, 57(1):102138.

Gizem Gezici, Aldo Lipani, Yucel Saygin, and Emine Yilmaz. 2021. Evaluation metrics for measuring bias in search engine results. *Information Retrieval Journal*, 24:85–113.

Avijit Ghosh, Ritam Dutt, and Christo Wilson. 2021. When fair ranking meets uncertain inference. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1033–1043.

Alexander Haas and Julian Unkel. 2017. Ranking versus reputation: perception and effects of search result credibility. *Behaviour & Information Technology*, 36(12):1285–1298.

Daniel E Ho and Kosuke Imai. 2008. Estimating causal effects of ballot order from a randomized natural experiment: The california alphabet lottery, 1978–2002. *Public opinion quarterly*, 72(2):216–240.

Chitika Insights. 2013. The value of google result positioning. *Westborough: Chitika Inc*, pages 0–10.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Transactions on Information Systems (TOIS)*, 25(2):7–es.

Kristen Johnson, I-Ta Lee, and Dan Goldwasser. 2017. Ideological phrase indicators for classification of political discourse framing on Twitter. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 90–99, Vancouver, Canada. Association for Computational Linguistics.

Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pages 3819–3828. ACM.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *ArXiv preprint*, abs/2212.14024.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

Anja Klasnja, Negar Arabzadeh, Mahbod Mehrvarz, and Ebrahim Bagheri. 2022. On the characteristics of ranking-based gender bias measures. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 245–249.

Klara Krieg, Emilia Parada-Cabaleiro, Gertraud Medicus, Oleg Lesota, Markus Schedl, and Navid Rekabsaz. 2023. Grep-biasir: A dataset for investigating gender representation bias in information retrieval results. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 444–448.

Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*, pages 417–432.

Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2019. Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal*, 22(1):188–227.

Anne Lauscher, Goran Glavas, Simone Paolo Ponzetto, and Ivan Vulic. 2020. A general framework for implicit and explicit debiasing of distributional word vector spaces. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8131–8138. AAAI Press.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Desmog: Detecting stance in media on global warming. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3296–3315.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *ArXiv preprint*, abs/2310.06824.

Dana McKay, Stephann Makri, Marisela Gutierrez-Lopez, Andrew MacFarlane, Sondess Missaoui, Colin Porlezza, and Glenda Cooper. 2020. We are the change that we seek: information interactions during a change of viewpoint. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pages 173–182.

Abbe Mowshowitz and Akira Kawaguchi. 2002. Assessing bias in search engines. *Information Processing & Management*, 38(1):141–156.

Alamir Novin and Eric Meyers. 2017. Making sense of conflicting science information: Exploring bias in the search engine result page. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pages 175–184.

Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, Michael D Ekstrand, Adam Roegiest, Aldo Lipani, Alex Beutel, Alexandra Olteanu, Ana Lucic, Ana-Andreea Stoica, et al. 2021. Facts-ir: fairness, accountability, confidentiality, transparency, and safety in information retrieval. In *ACM SIGIR Forum*, volume 53, pages 20–43. ACM New York, NY, USA.

Bing Pan, Helene Hembrooke, Thorsten Joachims, Lori Lorigo, Geri Gay, and Laura Granka. 2007. In google we trust: Users' decisions on rank, position, and relevance. *Journal of computer-mediated communication*, 12(3):801–823.

Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. 2009. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the 2009 SIAM international conference on data mining*, pages 581–592. SIAM.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568.

Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44.

Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2022. Fairness in rankings and recommendations: an overview. *The VLDB Journal*, pages 1–28.

Frances A Pogacar, Amira Ghenai, Mark D Smucker, and Charles LA Clarke. 2017. The positive and negative influence of search results on people's decisions about the efficacy of medical treatments. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval*, pages 209–216.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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

Navid Rekabsaz and Markus Schedl. 2020. Do neural ranking models intensify gender bias? In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 2065–2068. ACM.

Tetsuya Sakai, Nick Craswell, Ruihua Song, Stephen Robertson, Zhicheng Dou, and Chin-Yew Lin. 2010. Simple evaluation metrics for diversified search results. In *EVIA@ NTCIR*, pages 42–50.

Tetsuya Sakai, Jin Young Kim, and Inho Kang. 2023. A versatile framework for evaluating ranked lists in terms of group fairness and relevance. *ACM Transactions on Information Systems*, 42(1):1–36.

Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2011. Intent-aware search result diversification. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 595–604. ACM.

Piotr Sapiezynski, Wesley Zeng, Ronald E Robertson, Alan Mislove, and Christo Wilson. 2019. Quantifying the impact of user attentionon fair group representation in ranked lists. In *Companion proceedings of the 2019 world wide web conference*, pages 553–562.

Julia Schwarz and Meredith Ringel Morris. 2011. Augmenting web pages and search results to support credibility assessment. In *Proceedings of the International Conference on Human Factors in Computing Systems, CHI 2011, Vancouver, BC, Canada, May 7-12, 2011*, pages 1245–1254. ACM.

Kurt Shuster, Mojtaba Komeili, Leonard Adolphs, Stephen Roller, Arthur Szlam, and Jason Weston. 2022. Language models that seek for knowledge: Modular search & generation for dialogue and prompt completion. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages

373–393, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ashudeep Singh and Thorsten Joachims. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 2219–2228. ACM.

Peter Stefanov, Kareem Darwish, Atanas Atanasov, and Preslav Nakov. 2020. Predicting the topical stance and political leaning of media using tweets. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 527–537, Online. Association for Computational Linguistics.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *ArXiv preprint*, abs/2007.00808.

Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management*, pages 1–6.

Yinfei Yang, Ning Jin, Kuo Lin, Mandy Guo, and Daniel Cer. 2021. Neural retrieval for question answering with cross-attention supervised data augmentation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 263–268, Online. Association for Computational Linguistics.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in ranking, part i: Score-based ranking. *ACM Computing Surveys*, 55(6):1–36.

George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2022. Mitigating bias in search results through contextual document reranking and neutrality regularization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2532–2538.

Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. 2021. SPARTA: Efficient open-domain question answering via sparse transformer matching retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 565–575, Online. Association for Computational Linguistics.

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

Caleb Ziems, Omar Shaikh, Zhehao Zhang, William Held, Jiaao Chen, and Diyi Yang. 2023. Can large language models transform computational social science? *Computational Linguistics*, pages 1–53.

Caleb Ziems and Diyi Yang. 2021. To protect and to serve? analyzing entity-centric framing of police violence. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 957–976.

A Appendix

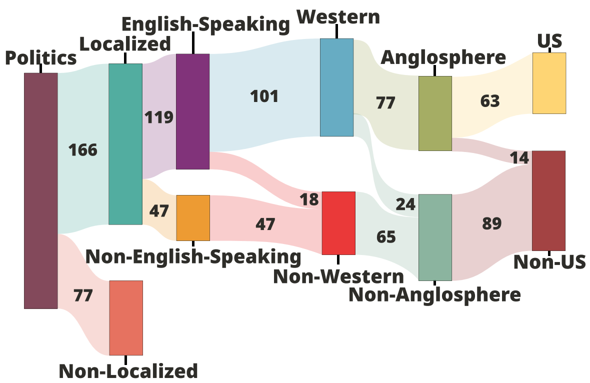

Figure 4: **Wikipedia Controversial Topic Distribution** can reflect biases in the Wikipedia editor pool. This explains why localized political topics are typically from English-speaking (71.69%) countries, and why there is over-representation of American issues.

A.1 Potential Biases in WIKI-BALANCE

Controversial Wikipedia article titles are topically diverse, covering 15 broad domains, including *politics, history, religion, science.* Given Wikipedia's global scope, we unsurprisingly find that each domain has wide coverage. Still, biases in the editor pool mean the distribution of topics in each domain is skewed. For example, in *politics*, local issues typically concern Anglophone countries (71.69% of political topics are about countries in which English is either a official, majority, or secondary language). In Western politics, the United States is over-represented, appearing in 62% of Western issues (see Figure 4). Neither fact is surprising, since topics come from English Wikipedia, and a plurality English-language Wikipedians edit from the United States.[6] See §8 for a more in-depth discussion on the impacts of skewed data.

[6] https://en.wikipedia.org/wiki/Wikipedia:Who_writes_Wikipedia

A.2 Measuring the Directionality of DUO Bias

By default, DUO metric measures only the magnitude of the bias and not its direction, as is the case with previous metrics (Gezici et al., 2021). Unlike prior metrics, our methodology allows for an unsupervised computation of a sign to indicate the *direction* of the bias. If we consider the signum function of a real number

$$\mathrm{sgn}(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0 \end{cases}$$

we can extend this definition to a set $A \subset \mathbb{R}$ by

$$\mathrm{sgn}(A) = 2 \times \mathbb{1}\left[\left(\frac{|\{a_i \in A : a_i > 0\}|}{|A|}\right) > 0.5\right] + 1$$

this gives us $\mathrm{sgn}(A) = -1$ when A contains more negative values than non-negative values, and $\mathrm{sgn}(A) = 1$ otherwise. If we modify the utility to encode the sign on the set of polarization scores $\{p_j\}$ as follows

$$u_V^+(i, r) = \mathrm{sgn}(\{p_j\}_{j=1}^{i})u_V(i, r)$$

Then we can set

$$\mu^+_{\mathrm{DUO}}(r) = \mathrm{sgn}\left(\mathrm{DUO}(r, u_V^+)\right)\left(\mathrm{DUO}(r, u_V)\right)$$

which assigns the existing DUO magnitude an appropriate polarity. Our signed $\mu^+_{\mathrm{DUO}}(r)$ value will be useful for understanding how biased rankings can shift a reader's opinion *towards* the perspective favored by the ranking (see §5.2).

A.3 Expanded Accuracy of Polarization Embeddings

Here in Tables 6 and 7, we enumerate the unsupervised polarization label accuracy for each Transformer-based embedding model in the Sentence BERT library. Table 6 gives the 13 models with greater than 90% accuracy, while Table 7 enumerates the remaining models in order of their accuracies. With a worst accuracy of 73%, a best accuracy of 95%, and a median accuracy of 87%, these results strongly validate our approach, and show its robustness across document embedding implementation.

A.4 Differing Distributions: Natural And Synthetic

Unsurprisingly, Figure 5 shows how natural Google search web articles are distributed differently than our synthetic corpus. Whereas natural bias scores (*bottom*) are normally distributed around a neutral mean of zero, synthetic data (*top*) follows a highly-polarized bimodal distribution, and includes some extreme outliers. We can conclude that any discrepancies between the synthetic and natural results in §6 are largely due to these differences. The respective evaluations are mutually complementary.

Embedding Model	**Accuracy**
`sentence-t5-xl`	95.27%
`sentence-t5-large`	94.17%
`nli-roberta-large`	92.17%
`roberta-large-nli-mean-tokens`	92.17%
`voyage-02`	92.04%
`paraphrase-distilroberta-base-v2`	91.88%
`paraphrase-mpnet-base-v2`	90.86%
`roberta-base-nli-stsb-mean-tokens`	90.38%
`roberta-large-nli-stsb-mean-tokens`	90.29%
`facebook-dpr-question_encoder-single-nq-base`	90.11%
`roberta-base-nli-mean-tokens`	90.05%
`nli-roberta-base`	90.05%
`sentence-t5-base`	90.03%

Table 6: **Sorted polarization score accuracies** for all embedding models in the Sentence BERT library (Reimers and Gurevych, 2019) with accuracy greater than 90%.

A.5 Removing Spurious Correlations from Polarization Embeddings

DUO depends on WIKI-BALANCE$_{\text{Synthetic}}$ for the $|r|$ synthetic documents on which polarization scores p_j are computed via PCA. The synthetic data can introduce spurious correlations into the DUO metric. For example, with the following query

> *How can the legal system effectively enforce copyright laws in BitTorrent-enabled piracy?*

we have two perspectives: (1) *from the perspective of the Entertainment Industry*, and (2) *from the perspective of BitTorrent users*. Naturally, any synthetic documents generated for Perspective 1 will contain more legal language. This means that even neutral documents could be marked by our embedding method as supporting Perspective 1 if they contain legal language (e.g., a purportedly neutral Wikipedia article). Here, we will describe how we worked to understand and remove such spurious correlations in DUO.

Following prior works on the removal of bias in word embeddings (Bolukbasi et al., 2016; Lauscher et al., 2020), we opt to identify in the document embedding space some subspace in which the spurious correlation lies; then we can effectively project

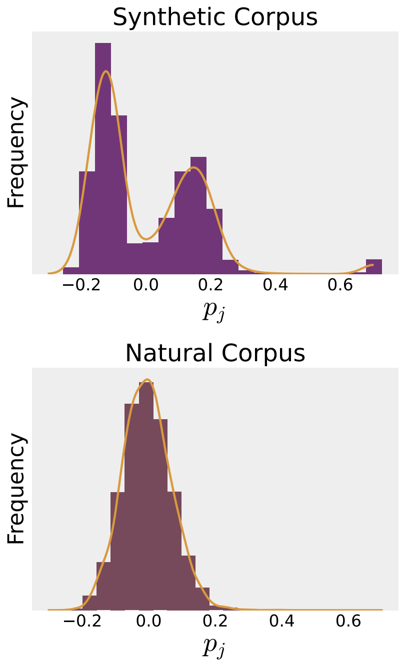

Figure 5: **Distributions of Polarization Scores** in the Synthetic (*top*) and Natural (*bottom*) corpus. Synthetic data is bimodal and polarized, while Natural data is normal and thus more neutral.

away this subspace. Let b be the principle bias axis in our embedding space, and x_j be the raw document embedding for document j. The debiased document embedding will then be

$$\tilde{x}_j = x_j - \langle x_j, b \rangle \, b$$

As before, we can fit PCA on $\tilde{X} = [\tilde{x}_1; ...; \tilde{x}_8]$ to compute de-biased polarization scores $\tilde{p}_j$.

To identify b, we used the following process, extracting the axis of spurious correlation automatically from a set of "distractor documents." First we generate these distractor documents to represent spurious correlations in the data. These distractors are the outputs of a pipeline similar to that of §3.1 for creating WIKI-BALANCE$_{\text{Synthetic}}$. For a given document d_t^i, which answers some query q_t (e.g., "*Did Edison steal patents from Tesla?*") about a high-level topic t (e.g., *Nikola Tesla*) by endorsing one particular perspective P_t^i (e.g., "*Edison stole patents from Tesla.*"), we will generate a distractor document $\tilde{d}_t^i$ that use similar style and vocabulary as d_t^i without ever answering q_t. The document $\tilde{d}_t^i$ will neither endorse nor deny P_t^i. To this end, we retrieve a distractor question $\tilde{q}_t^i$, which is a query semantically distinct from q_t^i, but it concerns the same high-level topic (e.g., the query, "*How did Tesla's personal eccentricities influence his*

Embedding Model	Accuracy
nli-bert-large-max-pooling	89.85%
bert-large-nli-max-tokens	89.85%
gtr-t5-large	89.70%
nli-bert-large	89.67%
bert-large-nli-mean-tokens	89.67%
gtr-t5-xl	89.58%
facebook-dpr-question_encoder-multiset-base	89.39%
bert-large-nli-stsb-mean-tokens	89.28%
nli-bert-large-cls-pooling	89.10%
bert-large-nli-cls-token	89.10%
average_word_embeddings_komninos	88.30%
average_word_embeddings_glove.6B.300d	88.29%
LaBSE	88.27%
msmarco-roberta-base-ance-firstp	88.24%
paraphrase-MiniLM-L12-v2	87.78%
average_word_embeddings_levy_dependency	87.78%
nli-bert-base	87.77%
bert-base-nli-mean-tokens	87.77%
distilbert-base-nli-stsb-mean-tokens	87.70%
bert-base-nli-stsb-mean-tokens	87.59%
nli-bert-base-max-pooling	87.41%
bert-base-nli-max-tokens	87.41%
nli-roberta-base-v2	87.40%
average_word_embeddings_glove.840B.300d	87.33%
gtr-t5-base	87.25%
nli-bert-base-cls-pooling	87.17%
bert-base-nli-cls-token	87.17%
all-mpnet-base-v1	87.03%
nli-distilbert-base-max-pooling	86.98%
distilbert-base-nli-max-tokens	86.98%
msmarco-distilbert-base-tas-b	86.94%
nli-mpnet-base-v2	86.94%
nli-distilbert-base	86.78%
distilbert-base-nli-mean-tokens	86.78%
msmarco-roberta-base-v3	86.53%
multi-qa-mpnet-base-dot-v1	86.40%
all-distilroberta-v1	86.24%
paraphrase-distilroberta-base-v1	85.88%
distilroberta-base-paraphrase-v1	85.88%
multi-qa-mpnet-base-cos-v1	85.53%
msmarco-bert-base-dot-v5	85.32%
nli-distilroberta-base-v2	85.01%
msmarco-MiniLM-L-12-v3	84.93%
paraphrase-xlm-r-multilingual-v1	84.87%
msmarco-distilbert-dot-v5	84.65%
nq-distilbert-base-v1	84.45%
msmarco-MiniLM-L12-cos-v5	84.33%
msmarco-roberta-base-v2	84.32%
msmarco-distilbert-base-v3	84.31%
multi-qa-distilbert-cos-v1	84.12%
msmarco-distilbert-base-dot-prod-v3	84.08%
paraphrase-MiniLM-L6-v2	83.90%
msmarco-distilroberta-base-v2	83.87%
distilroberta-base-msmarco-v2	83.87%
all-MiniLM-L6-v2	83.84%
msmarco-distilbert-base-v2	83.76%
msmarco-distilbert-multilingual-en-de-v2-tmp-lng-aligned	83.76%
paraphrase-multilingual-mpnet-base-v2	83.41%
multi-qa-distilbert-dot-v1	83.33%
msmarco-distilbert-base-v4	83.16%
msmarco-bert-co-condensor	83.13%
msmarco-distilbert-cos-v5	83.12%
msmarco-MiniLM-L-6-v3	83.08%
msmarco-MiniLM-L6-cos-v5	82.99%
multi-qa-MiniLM-L6-cos-v1	82.90%
all-mpnet-base-v2	82.82%
all-MiniLM-L12-v1	82.81%
facebook-dpr-ctx_encoder-single-nq-base	82.71%
paraphrase-albert-base-v2	82.71%
paraphrase-TinyBERT-L6-v2	82.70%
distilroberta-base-msmarco-v1	82.61%
bert-base-wikipedia-sections-mean-tokens	81.70%
multi-qa-MiniLM-L6-dot-v1	81.46%
distiluse-base-multilingual-cased-v2	81.43%
distiluse-base-multilingual-cased	81.43%
all-roberta-large-v1	81.19%
paraphrase-multilingual-MiniLM-L12-v2	81.09%
paraphrase-MiniLM-L3-v2	80.78%
all-MiniLM-L6-v1	80.77%
paraphrase-albert-small-v2	80.66%
msmarco-distilbert-multilingual-en-de-v2-tmp-trained-scratch	80.47%
quora-distilbert-base	80.40%
distilbert-base-nli-stsb-quora-ranking	80.40%
distilbert-multilingual-nli-stsb-quora-ranking	80.33%
quora-distilbert-multilingual	80.33%
allenai-specter	80.05%
all-MiniLM-L12-v2	79.49%
distiluse-base-multilingual-cased-v1	78.89%
facebook-dpr-ctx_encoder-multiset-base	78.12%
clip-ViT-B-32-multilingual-v1	73.19%

Table 7: **Sorted polarization score accuracies** for all embedding models in the Sentence BERT library (Reimers and Gurevych, 2019) with accuracy lower than 90%.

work?”). With the distractor question $\tilde{q}_t^i$, we generate $\tilde{d}_t^i$ by prompting `gpt-3.5-turbo`, **“<d_t^i> *Using as many words and phrases from the paragraph above as possible, try to answer:* <$\tilde{q}_t^i$>.”** To ensure that $\tilde{d}_t^i$ is neutral with respect to q_t, we follow up with, **“Rewrite the above paragraph but remove any sentences that have to do with the idea: <P_t^i>.”**

Now for each original document d_t^i we have a distractor $\tilde{d}_t^i$. Now we want to identify a spurious decision boundary between each perspective, so we partition the distractors according to the perspective of the document they stylistically emulate:

$$D_1 = \{\tilde{d}_t^i : P_t^i = 1\} \quad D_2 = \{\tilde{d}_t^i : P_t^i = 2\}$$

If each set of documents has its own average document embedding μ_{D_1}, μ_{D_2}, we can set the principle axis of spurious correlation to be

$$b = \mu_{D_1} - \mu_{D_2}$$

This difference of means is a proven reliable method for identifying the semantic direction between binary concepts (Marks and Tegmark, 2023).

A.6 Experimental Replications

In this section, we replicate our studies with different parameters to demonstrate the robustness of our experimental results in both the SEME Behavioral Study in §5.2 and the Bias Audit in §6.

Bias Audit. Table 8 gives the experimental replications for the Bias Audit, using different embedding methods. The *MiniLM* columns indicate the use of `all-MiniLM-L6-v2`, a weaker embedding model, as its accuracy in Table 7 was only 83%, compared with the stronger 95% performance of `sentence-t5-xl`. The *MiniLM* synthetic results have a correlation of $\rho = 0.72$ with the T5-XL results. Separately, in the *T5-XL-Debiased* column, we replicate our findings using the debiasing methods from §A.5 and find a strong correlation of $\rho = 0.89$ with the synthetic *T5-XL* results, and $\rho = 0.94$ with the natural *T5-XL* results. In each case, debiasing preserves the relative model ordering at the bottom of Table 8. For example, on WIKI-BALANCE$_{\text{Synthetic}}$, we have Use-QA $\succ$ *SPARTA* $\succeq$ *BM-25* $\succ$ *ColBERT* $\succeq$ *SPLADE* $\succ$ *SBERT* $\succ$ *ANCE*. We can conclude that our PAIR methodology largely induces a stable relative model ordering from the DUO metric, even when we consider alternative embedding models or apply debiasing methods to the embedding process to remove any potential spurious correlations.

Class	Model	MiniLM		T5-XL		T5-XL Debiased	
Lexical	BM-25	0.67	*0.64*	0.60	*0.66*	0.60	*0.67*
Sparse	SPARTA	0.65	*0.65*	0.58	*0.65*	0.60	*0.65*
	SPLADE	0.68	*0.68*	0.60	*0.67*	0.61	*0.67*
Dense	ANCE	0.70	*0.67*	0.62	*0.67*	0.63	*0.67*
	SBERT	0.70	*0.69*	0.61	*0.65*	0.61	*0.65*
	Use-QA	0.68	*0.64*	0.57	*0.64*	0.58	*0.63*
Late	ColBERT	0.68	*0.67*	0.60	*0.65*	0.61	*0.64*

Relative Model Orderings (Synthetic)	
MiniLM:	SPARTA $\succ$ Use-QA $\succeq$ BM-25 $\succeq$ ColBERT $\succeq$ SPLADE $\succ$ SBERT $\succeq$ ANCE
T5-XL:	Use-QA $\succ$ SPARTA $\succeq$ BM-25 $\succeq$ ColBERT $\succeq$ SPLADE $\succ$ SBERT $\succ$ ANCE
T5-XL Debiased:	Use-QA $\succ$ SPARTA $\succeq$ BM-25 $\succ$ ColBERT $\succeq$ SPLADE $\succeq$ SBERT $\succ$ ANCE

Relative Model Orderings (*Natural*)	
MiniLM:	Use-QA $\succeq$ BM-25 $\succ$ SPARTA $\succ$ ColBERT $\succeq$ ANCE $\succ$ SPLADE $\succ$ SBERT
T5-XL:	Use-QA $\succ$ ColBERT $\succeq$ SPARTA $\succeq$ SBERT $\succ$ BM-25 $\succ$ SPLADE $\succ$ ANCE
T5-XL Debiased:	Use-QA $\succ$ ColBERT $\succ$ SPARTA $\succeq$ SBERT $\succ$ BM-25 $\succeq$ SPLADE $\succeq$ ANCE

Table 8: **Experimental Replications of the Bias Audit** over WIKI-BALANCE$_{\text{Synthetic}}$ (*left columns*) and WIKI-BALANCE$_{\text{Natural}}$ (*italicized right columns*). Here we report DUO using different embedding methods: (1) using a weaker embedding model (*MiniLM*), and (2) using the debiasing method from §A.5 (*T5-XL-Debiased*). Scores have high mutual correlation, and relative model orderings (*bottom*) are stable across these replications.

SEME. Finally, we run experimental replications for the SEME in which we try both different embedding systems and different DUO step sizes. In Table 7, we determined that `clip-ViT-B-32-multilingual-v1` gives the lowest polarization score accuracy of 73%. Now in Table 9, we find that this polarization accuracy is not high enough to produce a significant effect in the SEME experiment. However, we can replicate our significant findings for a mid-performance model, `all-MiniLM-L6-v2`, which has 84% polarization accuracy, and the top-performance model, `sentence-t5-xl` which has 95% polarization accuracy. We can also replicate these experiments when we increase the DUO step size from 1 to 2 as in prior work (Yang and Stoyanovich, 2017). To increase the step size, one can effectively substitute the DCG in Equation 1 with

$$\mathrm{DCG}(r, u) = \sum_{i=2,4,\ldots}^{|r|} \frac{u(i, r)}{\log_2 i}$$

where i increments in steps of 2. In all such replications, we observe significant ($p < 0.05$) positive β_2 coefficients, leading us to conclude that DUO helps predict the SEME in cases of article click-through, and thus validating our method.

A.7 Parameters and Computing Budget

All experiments were performed on a Ubuntu Linux machine with 6 Nvidia GeForce RTX 2080 Ti GPUs. Each model evaluation took around 4 hours. All embedding models were run with the Sentence

Corpus	**Behavior**	Step	DUO Emb Model	N	β_2	$P(\beta_2 = 0)$	R^2
Natural	Clicked	1	clip-ViT-B-32-multilingual-v1	99	-0.294	0.176	0.475
Natural	Clicked	1	all-MiniLM-L6-v2	99	0.394	0.036	0.489
Natural	Clicked	1	sentence-t5-xl	99	0.527	0.022	0.493
Natural	Clicked	2	clip-ViT-B-32-multilingual-v1	99	-0.271	0.192	0.474
Natural	Clicked	2	all-MiniLM-L6-v2	99	0.392	0.036	0.489
Natural	Clicked	2	sentence-t5-xl	99	0.492	0.025	0.492
Combined	Clicked	1	clip-ViT-B-32-multilingual-v1	118	-0.173	0.373	0.503
Combined	Clicked	1	all-MiniLM-L6-v2	118	0.366	0.032	0.519
Combined	Clicked	1	sentence-t5-xl	118	0.426	0.046	0.517
Combined	Clicked	2	clip-ViT-B-32-multilingual-v1	118	-0.164	0.377	0.503
Combined	Clicked	2	all-MiniLM-L6-v2	118	0.365	0.032	0.519
Combined	Clicked	2	sentence-t5-xl	118	0.393	0.051	0.516

Table 9: **Experimental Replications of the SEME Behavioral Study** with different embedding systems. The weakest embedding model, `clip-ViT-B-32-multilingual-v1` fails to produce a significant effect, due to its low polarization score accuracy of 73% as determined in Table 7. However, we can replicate our significant findings for a mid-performance model, `all-MiniLM-L6-v2`, which has 84% polarization accuracy, and a top-performance model, `sentence-t5-xl` which has 95% polarization accuracy. We can also replicate these experiments when we increase the DUO step size from 1 to 2 as in prior work (Yang and Stoyanovich, 2017). In all such replications, we observe significant ($p < 0.05$) positive β_2 coefficients, leading us to conclude that DUO helps predict the SEME in cases of article click-through, and thus validating our method.

BERT (Reimers and Gurevych, 2019) package on default parameters. All IR models were run with the BEIR (Thakur et al., 2021) package on default parameters.