# Social-BRIGHT: Benchmarking Socially Grounded Reasoning in Information Retrieval

**Anonymous ACL submission**

## Abstract

Retrieval benchmarks increasingly test whether models can move beyond lexical and semantic matching to follow instructions and perform reasoning-intensive retrieval. Yet they rarely examine cases where relevance depends on social interpretation: a document may be topically plausible but still fail to satisfy an implicit norm, identity relation, pragmatic implication, or coded meaning. We introduce SOCIAL-BRIGHT, a benchmark for socially grounded reasoning in information retrieval. SOCIAL-BRIGHT casts four forms of social interpretation into full-corpus retrieval: selecting appropriate interpersonal responses, grounding group-level descriptions in relevant people or cases, explaining humor, and interpreting dogwhistles. Each query is paired with a socially appropriate document and a matched hard negative that remains lexically plausible while missing the underlying social condition, enabling evaluation with both standard ranking metrics and matched-pair accuracy. Across a broad set of retrievers, the strongest first-stage setting reaches only 0.302 average matched-pair accuracy, showing that current systems often fail to prefer the socially correct document over its hard negative. Further analyses of reasoning-oriented retrievers, rerankers, and query expansion show selective gains but do not uniformly close the gap. These results identify socially grounded relevance as a distinct bottleneck for retrieval and suggest that social interpretation should be treated as a core dimension of IR evaluation.

## 1 Introduction

Information retrieval (IR) is a central interface between language systems and external knowledge, especially in search, open-domain question answering, and retrieval-augmented generation (Lewis et al., 2020). Dense retrievers and general-purpose text embeddings have made this interface broadly reusable (Karpukhin et al., 2020), motivating benchmarks that evaluate retrieval quality across tasks and domains (Thakur et al., 2021; Muennighoff et al., 2023). Yet much of retrieval evaluation still treats relevance as lexical, topical, or semantic similarity between a query and a document. Recent instruction-following and reasoning-intensive benchmarks push beyond this view by testing whether retrievers can satisfy detailed instructions, implicit constraints, and multi-step or domain-specific inference requirements rather than relying on surface matching alone (Weller et al., 2025a,b; Zeng et al., 2026; Su et al., 2025; Xiao et al., 2024; Zheng et al., 2025; Wang et al., 2024b; Thakur et al., 2025).

This broader view of relevance still leaves social interpretation largely underexplored. Many information needs are socially situated: the relevant document may be the one that recognizes an interpersonal norm, a group or identity relation, a pragmatic implication, or a coded meaning. In such cases, a topically plausible document can still be wrong because it misses the social condition that makes the query meaningful. We study *social reasoning retrieval*, where relevance depends on whether a document satisfies this socially grounded interpretation. To benchmark the problem, we introduce SOCIAL-BRIGHT, which converts socially grounded reasoning phenomena into full-corpus retrieval tasks with paired positive documents and matched hard negatives. The benchmark spans four complementary settings: selecting socially appropriate interpersonal responses, grounding group-level social descriptions in documents about relevant people or cases, finding explanations that capture the mechanism of humor, and interpreting dogwhistles or coded language. Across these settings, we curate or transform source data, construct lexically plausible negative documents that miss the relevant social condition, and use human validation to ensure that each positive–negative contrast reflects the intended social judgment. As illustrated

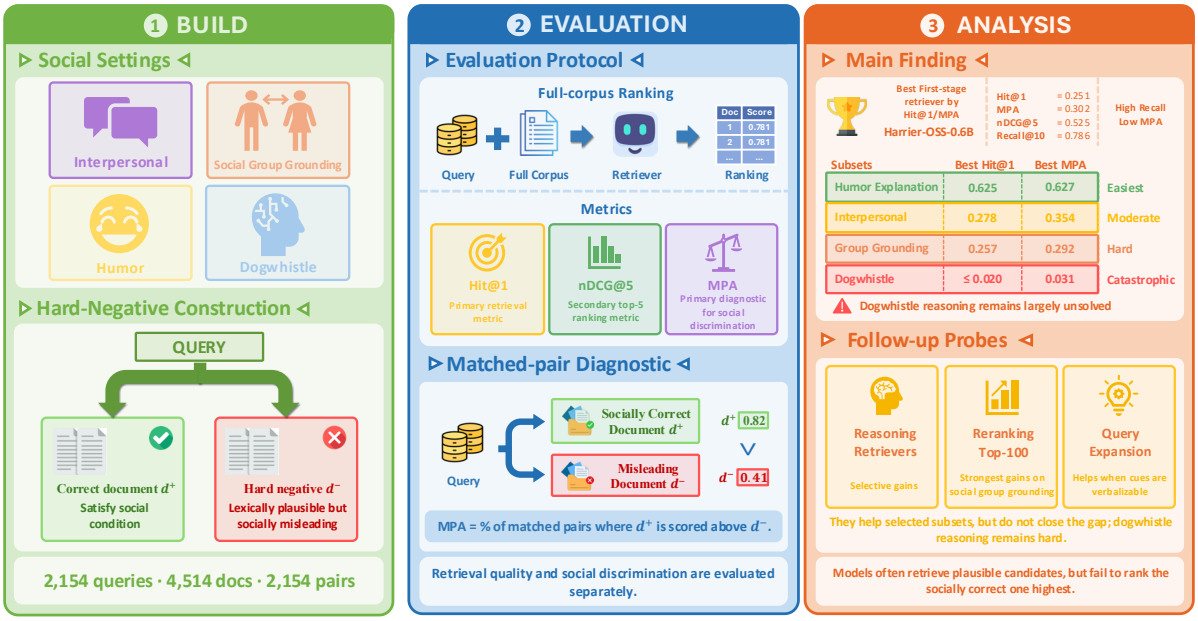


Figure 1: Overview of SOCIAL-BRIGHT. The benchmark converts socially grounded reasoning phenomena into full-corpus retrieval tasks with matched hard negatives, and evaluates models using both standard ranking metrics and matched-pair diagnostics.

in Figure 1, SOCIAL-BRIGHT evaluates retrievers with standard ranking metrics and with matched-pair accuracy, which asks whether the socially appropriate document is scored above a matched hard negative that is lexically plausible but socially misleading.

Our experiments show that existing retrievers do not reliably solve social reasoning retrieval. The strongest first-stage retriever by macro Hit@1 and matched-pair accuracy is Harrier-OSS-0.6B, which averages only 0.251 Hit@1 and 0.302 matched-pair accuracy across the four subsets. For the same model, 0.786 Recall@10 shows that gold documents are often recovered within a short candidate list. The main difficulty is therefore not only candidate recall, but fine-grained social discrimination. SOCIAL-BRIGHT provides a targeted testbed for evaluating whether retrieval models can move beyond semantic matching toward socially grounded relevance judgments. Our contributions are as follows:

- We formulate *social reasoning retrieval* as a retrieval setting where relevance depends on socially grounded interpretation rather than lexical or topical similarity alone.
- We introduce SOCIAL-BRIGHT, a benchmark spanning interpersonal response selection, social group grounding retrieval, humor explanation retrieval, and dogwhistle interpretation, with matched hard negatives that are lexically plausible but socially incorrect.
- We evaluate representative retrievers and retrieval enhancements, showing that strong models can often recover relevant candidates in short lists but still struggle to distinguish socially appropriate documents from matched hard negatives.

## 2 Related Work

**Reasoning-intensive IR Benchmarks.** Recent IR benchmarks have broadened retrieval evaluation beyond lexical or semantic matching toward settings where relevance depends on instructions, complex objectives, or reasoning. Instruction-following benchmarks test whether retrievers can use detailed relevance criteria rather than treating queries as keyword bags (Weller et al., 2025a,b; Zeng et al., 2026), while complex-objective and realistic technical benchmarks evaluate retrieval under richer task constraints (Wang et al., 2024b; Thakur et al., 2025). More directly, Xiao et al. (2024) introduce RAR-B by transforming reasoning problems into retrieval tasks, and Su et al. (2025) introduce BRIGHT, a realistic benchmark where user queries must be matched to documents through multi-step or domain-specific inference.

Subsequent work extends reasoning-intensive retrieval into specialized domains and evaluation settings, including legal RAG benchmarks grounded in legal research tasks (Zheng et al., 2025), medical evidence retrieval in R2MED (Zhang et al., 2026), and multi-aspect evidence retrieval for agentic search in BRIGHT-PRO (Zhao et al., 2026). These benchmarks establish that retrieval models struggle when relevance is governed by instructions, formal reasoning, domain expertise, or evidence aggregation; SOCIAL-BRIGHT extends this trajectory by making socially grounded interpretation itself the criterion of relevance, targeting a capability that is essential for retrieval systems operating over human social meaning rather than only facts, rules, or procedures.

**Social Aspects in IR Benchmarks.** Existing benchmarks also contain socially situated text, but social meaning is usually part of the content rather than the object of retrieval evaluation. BEIR includes argument and claim-centered retrieval tasks, such as argument retrieval and fact verification, where models must retrieve documents that support, oppose, or verify a stated proposition (Thakur et al., 2021). These tasks involve socially relevant discourse, but relevance is primarily defined by argumentative or evidential relations rather than by whether a document correctly interprets an implicit norm, identity relation, pragmatic cue, or coded meaning. MTEB broadens evaluation further and includes tasks built from socially oriented text, such as sentiment, emotion, and toxicity data, alongside many other embedding tasks (Muennighoff et al., 2023). However, these tasks typically evaluate general embedding transfer, classification, or semantic matching, not retrieval decisions in which a topically plausible distractor must be rejected because it misses the underlying social condition. SOCIAL-BRIGHT therefore complements existing benchmarks by turning social interpretation into the criterion of relevance itself, evaluating whether retrievers can serve as reliable evidence selectors in settings where downstream systems must reason over human social meaning.

## 3 Building the SOCIAL-BRIGHT Benchmark

### 3.1 Task Formulation

We formulate SOCIAL-BRIGHT as corpus-level retrieval with an additional matched-pair diagnostic for socially grounded relevance. Each task consists of a query set $\mathcal{Q}$, a document corpus $\mathcal{C}$, relevance judgments $\mathcal{R} \subseteq \mathcal{Q} \times \mathcal{C}$, and a set of matched hard-negative pairs $\mathcal{P} = \{(q_i, d_i^+, d_i^-)\}_{i=1}^N$. For each pair, $d_i^+$ is relevant to $q_i$ because it satisfies the socially grounded condition expressed by the query, while $d_i^- \in \mathcal{C}$ is a lexically plausible but non-relevant document that misses that condition. A retriever assigns a score $s(q, d)$ to every document $d \in \mathcal{C}$ and returns a ranking over the full corpus for each query $q$. We evaluate this full-corpus ranking with standard IR metrics such as Hit@1 and nDCG@5 using $\mathcal{R}$. To isolate the intended hard-negative contrast, we also report matched-pair accuracy:

$$\text{MPA} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[s(q_i, d_i^+) > s(q_i, d_i^-)]$$

where $\mathbf{1}[\cdot]$ is the indicator function. This metric complements full-corpus evaluation: standard ranking metrics measure end-to-end retrieval quality, while matched-pair accuracy directly tests whether a retriever prefers the socially appropriate document over its socially misleading alternative.

**Matched-pair intuition.** A query may ask how to interpret a close friend's message that they "almost forgot" your birthday. The positive document considers socially plausible readings, such as self-deprecation, busyness, or passive aggression, while the hard negative gives a literal thank-you response; both discuss the same message, but only the positive captures the pragmatic social ambiguity.

**Task families.** We instantiate this formulation with four task families, each targeting a different way in which social meaning determines relevance:

- **Interpersonal response retrieval.** Given a social situation, the retriever must select the response that is appropriate, empathetic, and socially well-reasoned, rather than a superficially similar response that misses the interpersonal norm or context.
- **Social group grounding retrieval.** Given a group-centric information need, the retriever must find a document about a person or case satisfying the intended demographic, institutional, occupational, geographic, or cultural condition; hard negatives remain lexically close but violate one of these constraints.
- **Humor explanation retrieval.** Given a joke, the retriever must identify an explanation that

captures the linguistic or pragmatic mechanism behind the humor, while rejecting plausible explanations that misidentify the setup, punchline, or social implication.

- **Dogwhistle reasoning retrieval.** Given a use of coded language, the retriever must recover the covert social meaning and contextual function of the expression, while rejecting explanations that preserve the literal topic but miss or downplay the coded signal.

Together, these tasks test whether retrievers can recognize social relevance across interactional judgment, group knowledge, pragmatic interpretation, and covert meaning.

### 3.2 Subset Construction and Validation

**Interpersonal response retrieval.** We build the interpersonal response retrieval subset from the Social Reasoning RLHF dataset released by ProlificAI.[1] The source data contains social questions paired with a chosen response and a rejected response. We convert this preference format into retrieval by treating each question as a query, the chosen response as the candidate relevant document, and the rejected response as a matched hard negative. This task therefore asks whether a retriever can identify the response that better satisfies the interpersonal demands of the situation, rather than simply matching the surface topic of the question.

Because the source data may contain multiple preference pairs for the same question, we group examples by normalized question text and keep one matched pair per unique query, choosing the pair in which the rejected response has the largest lexical-similarity advantage over the chosen response. Human annotators then retain only cases where the chosen response is appropriate and clearly more socially suitable than the rejected response. Full-corpus relevance judgments also include GPT-5.4-mini-verified cross-query positives; Appendix B describes this screen, and Appendix A.2 gives the annotation guideline.

**Social group grounding retrieval.** Social reasoning also requires grounding group-level descriptions in person- or case-level evidence. Social categories are structured by attributes, roles, institutions, geography, and historical context, and a relevant document may not repeat the group phrase used in the query. For example, a query about *Black politicians* may require retrieving a document centered on Barack Obama, even if the document describes his presidency rather than explicitly restating the group label. We design social group grounding retrieval to test this form of socially grounded inference: the retriever must recognize that an individual document satisfies a group-centric information need, while rejecting a lexically similar document that violates a decisive social or institutional condition.

We synthesize this subset from a human-written inventory of 200 social groups across ten domains, covering politics, law, business, technology, academia, healthcare, media, arts, sports, and civil society. Each group combines a social identity or background attribute, such as race, ethnicity, gender, religion, disability, immigration status, or LGBTQ identity, with a role or institutional context. The full group inventory is listed in Appendix F. For each group, we use GPT-5.4 to generate a triplet consisting of a group-centric query, a positive document centered on a person who truly belongs to the group, and a group-centric hard negative that remains lexically close but fails one key condition, such as social attribute, role, geography, era, or institutional status. The prompt template is provided in Appendix E.1. Human annotators then validate the synthesized triplets, retaining only examples where the positive and negative labels are correct, the social-group claims about people are factually accurate, and the documents read as natural standalone prose. The annotation guideline is provided in Appendix A.3.

**Humor explanation retrieval.** Humor often depends on implicit pragmatic reasoning: a correct explanation must identify the key ambiguity, incongruity, reversal, or culturally situated implication that makes a joke funny. We build the humor explanation retrieval subset from CHUMOR 2.0, a Chinese humor understanding dataset with annotated joke explanations (He et al., 2025). We treat each joke as a query and a reference explanation as the relevant document. To construct a matched hard negative, we use GPT-5.4 to synthesize a plausible but incomplete explanation conditioned on the joke and the reference explanation. The negative explanation is designed to reuse concrete surface details from the joke while missing or only partially capturing the core humor mechanism. Human annotators then validate the pairs, retaining only cases where

[1] ProlificAI/social-reasoning-rlhf.

| Subset | Q | D | $D^+$ | P | $Q_{\text{len}}$ | $D_{\text{len}}$ | Source / Construction | Example |
|---|---|---|---|---|---|---|---|---|
| Interpersonal response | 852 | 1,910 | 1.3 | 852 | 111.0 | 513.9 | Social Reasoning RLHF | Tab. 4 |
| Social group grounding | 113 | 226 | 1.0 | 113 | 210.9 | 865.0 | Human-written groups + GPT-5.4 | Tab. 5 |
| Humor explanation | 993 | 1,986 | 1.0 | 993 | 27.8 | 123.6 | CHUMOR 2.0 + GPT-5.4 | Tab. 6 |
| Dogwhistle reasoning | 196 | 392 | 1.0 | 196 | 378.1 | 640.8 | Dogwhistle glossary + GPT-5.4 | Tab. 7 |
| Total | 2,154 | 4,514 | 1.1 | 2,154 | 102.2 | 370.8 | – | – |

Table 1: Data statistics of SOCIAL-BRIGHT. Q and D denote the number of queries and corpus documents; $D^+$ is the average number of relevant documents per query; P is the number of matched hard-negative pairs. Average query and document lengths are measured in characters.

the reference explanation is sufficiently informative and the negative explanation does not fully cover the central meaning of the joke. The annotation guideline is provided in Appendix A.4.

**Dogwhistle reasoning retrieval.** Dogwhistles are coded expressions whose surface meaning can appear innocuous while conveying a covert social meaning to an intended in-group. We build the dogwhistle reasoning retrieval subset from the dogwhistle glossary introduced by Mendelsohn et al. (2023). We treat each linguistic context containing a dogwhistle as a query and the source explanation of its coded meaning as the relevant document. To construct a matched hard negative, we use GPT-5.4 to synthesize a plausible explanation conditioned on the linguistic context, dogwhistle metadata, and source explanation. The negative explanation is designed to preserve surface details from the context while missing, literalizing, or downplaying the coded social meaning. Human annotators then validate the synthesized negatives, retaining only cases where the negative explanation is truly incorrect or materially worse than the source explanation. The annotation guideline is provided in Appendix A.5.

After filtering and human validation, SOCIAL-BRIGHT contains 2,154 queries, 4,514 corpus documents, 2,410 relevance judgments, and 2,154 matched hard-negative pairs. Table 1 summarizes the resulting benchmark. Across all four subsets, validation was conducted by three bilingual author-annotators with bachelor's degrees; each subset was reviewed by two annotators, and an example was retained only when both annotators agreed that it should be kept. Additional annotator details are provided in Appendix A.1. The design supports both standard full-corpus retrieval evaluation and controlled matched-pair analysis: each query is evaluated against a corpus, while each validated pair isolates the social condition that makes one document relevant and the other misleading. Representative examples from all four subsets are provided in Appendix C.

## 4 Experimental Setup

### 4.1 Models

We evaluate publicly released retrievers from two architectural families. The first family consists of bi-encoder retrievers, which encode each query and document independently into a single vector and rank documents by vector similarity. This group includes Qwen3-Embedding at 0.6B, 4B, and 8B scales (Zhang et al., 2025), EmbeddingGemma-300M (Vera et al., 2025), multilingual-E5-large-instruct (Wang et al., 2024a), KaLM-Embedding-Gemma3-12B (Zhao et al., 2025), Harrier-OSS-v1-0.6B (Microsoft, 2026), Snowflake Arctic-Embed-L-v2.0 (Yu et al., 2024), and DenseOn (Sourty et al., 2026).

The second family consists of late-interaction retrievers, which preserve multiple token-level vectors for each query and document and aggregate fine-grained matching signals at scoring time. This group includes ColBERTv2 (Santhanam et al., 2022), Jina-ColBERT-v2 (Jha et al., 2024), GTE-ModernColBERT-v1 (Chaffin, 2025a), mxbai-edge-colbert-v0-32m (Takehi et al., 2025), and LateOn (Sourty et al., 2026). Together, these models provide broad coverage of current retrieval architectures, model scales, multilingual training regimes, and efficiency-performance trade-offs.

As a sparse lexical reference point, we also evaluate BM25 (Robertson and Zaragoza, 2009) under the same retrieval protocol. We report its detailed results separately in Appendix D.3.

### 4.2 Evaluation Protocol

We evaluate all models zero-shot using their publicly released checkpoints, without training or calibration on SOCIAL-BRIGHT. For each subset, models retrieve over the full subset corpus for every query; matched hard-negative pairs are used only

for diagnostic analysis and do not restrict the candidate set during retrieval. Instruction-aware retrievers receive a query-side task instruction when one is defined; corpus documents are not prefixed with task instructions. The exact query instructions are listed in Appendix D.1. MPA is computed from the same query-document scores used for full-corpus ranking by comparing each relevant document with its matched hard negative. Unless otherwise stated, aggregate results are macro-averaged across the four subsets, giving each subset equal weight.

### 4.3 Evaluation Measures

We evaluate models with three complementary measures. Hit@1 is our primary full-corpus measure, indicating whether the top-ranked document is relevant to the query. Matched-pair accuracy (MPA), defined in Section 3.1, serves as the primary diagnostic for the hard-negative contrast: it measures whether the paired relevant document receives a higher score than its matched hard negative. We also report nDCG@5 as a secondary full-corpus measure that captures ranking quality within the top five results. When discussing candidate recall, we additionally report Recall@10 as a contextual diagnostic, while keeping Hit@1 and MPA as the main evaluation targets.

## 5 Results

### 5.1 Main Results

**Existing retrievers struggle.** Our main results are summarized in Table 2. We group models by architecture and scale, and report both full-corpus top-rank accuracy (Hit@1) and matched-pair accuracy (MPA) for each subset; macro nDCG@5 is included as a secondary ranking reference. Full Recall@10 values for the same first-stage runs are reported in Appendix D.2. The strongest first-stage retriever by macro Hit@1 and MPA, Harrier-OSS-0.6B, reaches only 0.251 average Hit@1 and 0.302 average MPA. Since MPA directly tests whether a model scores the socially appropriate document above its matched hard negative, these results show that current retrievers often fail precisely at the socially grounded discrimination targeted by the benchmark. The BM25 baseline in Appendix D.3 shows the same pattern, further confirming that surface lexical matching is not sufficient for these socially grounded contrasts.

**Task difficulty varies sharply across subsets.** Humor explanation retrieval is the most accessible setting, with Harrier-OSS-0.6B reaching 0.625 Hit@1 and 0.627 MPA. Interpersonal response retrieval is more difficult but still admits moderate performance: the best Hit@1 is 0.278 from Qwen3-Emb-8B, while the best MPA is 0.354 from mE5-large-instruct. Social group grounding retrieval is also challenging: the best model reaches 0.257 Hit@1 and 0.292 MPA, showing that group-to-instance social inference remains difficult for current retrievers. Dogwhistle reasoning retrieval is the hardest subset by a large margin. No model exceeds 0.020 Hit@1, and the best MPA is only 0.031, suggesting that coded social meaning creates a particularly sharp mismatch between surface similarity and relevance.

**Bi-encoders lead, but social discrimination remains weak.** Among sub-1B models, Harrier-OSS-0.6B achieves the best average Hit@1 and MPA, while mE5-large-instruct is strongest on social group grounding retrieval. Larger bi-encoders improve some rankings, but scale does not close the matched-pair gap: all bi-encoders remain far below 0.5 average MPA. Late-interaction retrievers do not recover this gap either. Jina-ColBERT-v2 is the strongest model in this family by average Hit@1, but it remains substantially below the best bi-encoders. Overall, neither model scale nor token-level interaction is sufficient for robust socially grounded relevance judgments.

### 5.2 Additional Analyses

Table 3 summarizes three follow-up analyses that probe whether additional reasoning-oriented retrieval components can improve socially grounded relevance judgments. For each analysis, we report subset-level Hit@1 and MPA, along with macro-averaged Hit@1, MPA, and nDCG@5.

#### 5.2.1 Reasoning transfer is uneven

We use these models as a transfer test from reasoning-intensive IR to socially grounded retrieval. Prior work in this direction is largely motivated by BRIGHT-style scientific, mathematical, theorem-proving, StackExchange, and coding settings, where relevant documents support problem solving rather than direct factual lookup (Su et al., 2025). We evaluate two reasoning-oriented bi-encoders, ReasonIR-8B (Shao et al., 2025) and ReasonEmbed-Qwen3-8B (Chen et al., 2026), together with Reason-ModernColBERT, a late-interaction model trained on ReasonIR data

| Model | Hit@1 by subset | | | | MPA by subset | | | | Macro average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Inter. | Group | Humor | Dogw. | Inter. | Group | Humor | Dogw. | Hit@1 | MPA | nDCG@5 |
| **Open-source bi-encoders (<1B)** | | | | | | | | | | | |
| Harrier-OSS-0.6B | 0.224 | 0.150 | **0.625** | 0.005 | 0.337 | 0.239 | **0.627** | 0.005 | **0.251** | **0.302** | 0.525 |
| mE5-large-instruct | 0.246 | **0.257** | 0.338 | 0.000 | **0.354** | **0.292** | 0.340 | 0.000 | 0.210 | 0.247 | 0.498 |
| Qwen3-Emb-0.6B | 0.221 | 0.044 | 0.514 | 0.010 | 0.317 | 0.080 | 0.518 | 0.010 | 0.197 | 0.231 | 0.490 |
| DenseOn | 0.212 | 0.053 | 0.200 | 0.000 | 0.278 | 0.062 | 0.202 | 0.000 | 0.116 | 0.136 | 0.481 |
| Arctic-Embed-L-v2 | 0.192 | 0.044 | 0.137 | 0.000 | 0.270 | 0.044 | 0.137 | 0.000 | 0.093 | 0.113 | 0.467 |
| EmbeddingGemma-300M | 0.096 | 0.053 | 0.100 | **0.020** | 0.285 | 0.088 | 0.103 | 0.020 | 0.067 | 0.124 | 0.296 |
| **Open-source bi-encoders (>1B)** | | | | | | | | | | | |
| KaLM-Gemma3-12B | 0.270 | 0.221 | 0.469 | 0.010 | 0.333 | 0.265 | 0.471 | 0.010 | 0.243 | 0.270 | 0.583 |
| Qwen3-Emb-8B | **0.278** | 0.080 | 0.606 | 0.005 | 0.350 | 0.097 | 0.611 | 0.005 | 0.242 | 0.266 | 0.562 |
| Qwen3-Emb-4B | 0.246 | 0.106 | 0.497 | 0.010 | 0.331 | 0.142 | 0.498 | 0.020 | 0.215 | 0.248 | 0.535 |
| **Late-interaction retrievers** | | | | | | | | | | | |
| Jina-ColBERT-v2 | 0.168 | 0.018 | 0.245 | 0.015 | 0.243 | 0.035 | 0.246 | 0.026 | 0.111 | 0.137 | 0.442 |
| mxbai-edge-colbert | 0.139 | 0.044 | 0.105 | **0.020** | 0.229 | 0.053 | 0.108 | 0.020 | 0.077 | 0.103 | 0.378 |
| ColBERTv2 | 0.143 | 0.035 | 0.103 | **0.020** | 0.237 | 0.053 | 0.203 | 0.026 | 0.075 | 0.130 | 0.316 |
| LateOn | 0.146 | 0.071 | 0.060 | **0.020** | 0.332 | 0.159 | 0.101 | **0.031** | 0.074 | 0.156 | 0.236 |
| GTE-ModernColBERT | 0.162 | 0.053 | 0.065 | 0.015 | 0.251 | 0.053 | 0.065 | 0.015 | 0.074 | 0.096 | 0.410 |

Table 2: Main results on SOCIAL-BRIGHT. Inter., Group, Humor, and Dogw. denote interpersonal response retrieval, social group grounding retrieval, humor explanation retrieval, and dogwhistle reasoning retrieval, respectively. Macro averages are computed across the four subsets. Bold marks the best score in each Hit@1 or MPA column.

(Chaffin, 2025b). All models are evaluated zero-shot under the same full-corpus protocol as the main experiments.

The results show selective transfer rather than broad robustness. ReasonEmbed-Qwen3-8B is the strongest reasoning retriever, reaching 0.299 average Hit@1, 0.353 average MPA, and 0.556 average nDCG@5, with the clearest gains on social group grounding retrieval and humor explanation retrieval. However, it still performs poorly on dogwhistle reasoning, reaching only 0.015 Hit@1 and 0.026 MPA. ReasonIR-8B has low average Hit@1 but relatively strong interpersonal MPA, while Reason-ModernColBERT remains below the strongest general retrievers on average. These results suggest that reasoning-intensive retrieval training can help when social relevance resembles abstraction or explanatory matching, but it does not reliably transfer to settings where relevance depends on coded social meaning or pragmatic interpretation.

#### 5.2.2 Reranking helps social group grounding most

We test second-stage reranking by taking the top 100 candidates from Qwen3-Embedding-8B and reranking them with Qwen3-Reranker at 0.6B, 4B, and 8B scales (Zhang et al., 2025), jina-reranker-v3 (Wang et al., 2025), and mxbai-rerank-base-v2 and mxbai-rerank-large-v2 (Lee et al., 2025). Reranking most improves social group grounding: starting from the Qwen3-Embedding-8B first-stage scores of 0.080 Hit@1 and 0.097 MPA on this subset, Qwen rerankers raise Hit@1 to 0.566, 0.850, and 0.841, and MPA to 0.637, 0.920, and 0.903. This scaling pattern is consistent with the task's factual-knowledge component, since the reranker must verify whether a person, role, or institution satisfies a group-level condition rather than simply matching words. Gains are less uniform elsewhere: Qwen3-Reranker-8B has the best macro averages, but dogwhistle reasoning remains difficult (best Hit@1 0.092; best MPA 0.097), and non-Qwen rerankers are less consistent except for mxbai-rerank-large-v2 on humor. Appendix D.4 reports a query-shuffle control for the social group grounding result.

#### 5.2.3 Query expansion helps unevenly

Following BRIGHT, where LLM-generated reasoning steps improve reasoning-intensive retrieval (Su et al., 2025), we test GPT-5.4 query expansion for socially grounded retrieval. GPT-5.4 rewrites each query without access to the corpus, relevant documents, hard negatives, or labels, preserving the original intent while making implicit social,

| Model | Hit@1 by subset | | | | MPA by subset | | | | Macro average | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Inter. | Group | Humor | Dogw. | Inter. | Group | Humor | Dogw. | Hit@1 | MPA | nDCG@5 |
| **Reasoning retrievers** | | | | | | | | | | | |
| ReasonIR-8B | 0.054 | 0.071 | 0.034 | **0.046** | **0.435** | 0.106 | 0.046 | **0.077** | 0.051 | 0.166 | 0.224 |
| ReasonEmbed-Qwen3-8B | **0.264** | **0.487** | **0.431** | 0.015 | 0.419 | **0.531** | **0.435** | 0.026 | **0.299** | **0.353** | **0.556** |
| Reason-ModernColBERT | 0.194 | 0.062 | 0.070 | 0.000 | 0.326 | 0.071 | 0.070 | 0.000 | 0.082 | 0.117 | 0.392 |
| **Rerankers** | | | | | | | | | | | |
| Qwen3-Reranker-0.6B | 0.264 | 0.566 | **0.678** | 0.031 | 0.365 | 0.637 | **0.694** | 0.031 | 0.385 | 0.432 | 0.631 |
| Qwen3-Reranker-4B | 0.313 | **0.850** | 0.185 | **0.092** | **0.459** | **0.920** | 0.193 | **0.097** | 0.360 | 0.417 | 0.648 |
| Qwen3-Reranker-8B | **0.326** | 0.841 | 0.442 | 0.031 | 0.435 | 0.903 | 0.446 | 0.036 | **0.410** | **0.455** | **0.687** |
| jina-reranker-v3 | 0.227 | 0.053 | 0.016 | 0.010 | 0.346 | 0.071 | 0.016 | 0.010 | 0.076 | 0.111 | 0.389 |
| mxbai-rerank-base-v2 | 0.139 | 0.053 | 0.298 | 0.000 | 0.251 | 0.071 | 0.310 | 0.000 | 0.122 | 0.158 | 0.431 |
| mxbai-rerank-large-v2 | 0.184 | 0.097 | 0.570 | 0.010 | 0.278 | 0.115 | 0.654 | 0.010 | 0.215 | 0.264 | 0.482 |
| **GPT-5.4 query expansion** | | | | | | | | | | | |
| Harrier-OSS-0.6B | 0.242 | 0.204 | **0.862** | 0.005 | 0.378 | 0.239 | **0.865** | 0.005 | 0.328 | 0.372 | 0.570 |
| mE5-large-instruct | 0.244 | **0.451** | 0.665 | 0.005 | 0.378 | **0.522** | 0.669 | 0.005 | 0.341 | 0.393 | 0.585 |
| Qwen3-Emb-0.6B | 0.250 | 0.142 | 0.602 | 0.046 | 0.381 | 0.168 | 0.605 | 0.066 | 0.260 | 0.305 | 0.549 |
| Qwen3-Emb-4B | 0.262 | 0.319 | 0.565 | **0.092** | 0.366 | 0.363 | 0.570 | 0.138 | 0.309 | 0.359 | 0.595 |
| Qwen3-Emb-8B | **0.292** | 0.274 | 0.804 | 0.077 | **0.394** | 0.336 | 0.813 | **0.143** | 0.362 | **0.422** | 0.623 |
| KaLM-Gemma3-12B | 0.282 | 0.345 | 0.822 | 0.020 | 0.358 | 0.363 | 0.825 | 0.020 | **0.367** | 0.391 | **0.639** |

Table 3: Additional analyses on SOCIAL-BRIGHT. Inter., Group, Humor, and Dogw. denote interpersonal response retrieval, social group grounding retrieval, humor explanation retrieval, and dogwhistle reasoning retrieval. Macro averages are over subsets; GPT-5.4 expansion rows use expanded queries generated from the original query only. Bold marks the best score within each analysis block.

pragmatic, factual, or cultural cues explicit; the prompt is in Appendix E.2. We then evaluate representative bi-encoders under the same full-corpus protocol.

Expansion helps some settings but remains uneven. Qwen3-Embedding-8B improves from 0.242 to 0.362 Hit@1 and from 0.266 to 0.422 MPA; Harrier-OSS-0.6B improves from 0.251 to 0.328 Hit@1 and from 0.302 to 0.372 MPA. Gains concentrate on humor and social group grounding, where explicit explanatory or group-level cues align queries with relevant documents. Dogwhistle reasoning remains difficult: the best expanded-query run reaches only 0.092 Hit@1 and 0.143 MPA, suggesting that coded social meaning requires more than query rewriting.

## 6 Discussion

Our results suggest that socially grounded retrieval fails in qualitatively different ways across task families. Interpersonal response retrieval depends on inferring conversational norms, speaker intent, and likely social consequences, connecting retrieval to social commonsense and norm reasoning (Sap et al., 2019; Forbes et al., 2020). Social group grounding is different: its strongest gains come from reranking and scaling, suggesting that the bottleneck is often factual verification over people, roles, and institutions rather than pragmatic interpretation alone. Humor explanation and dogwhistle reasoning are harder in another way: the relevant evidence must recover an implicit bridge, non-literal interpretation, or coded social signal while hard negatives preserve much of the surface topic. This pattern is consistent with prior work showing that pragmatic and socially implied meanings remain challenging for language models (Hu et al., 2023; Sap et al., 2020; Mendelsohn et al., 2023).

These findings also refine how retrieval benchmarks define relevance. Recent work has moved beyond keyword or semantic matching toward instruction following, complex objectives, and reasoning-intensive retrieval (Weller et al., 2025a; Wang et al., 2024b; Su et al., 2025). SOCIAL-BRIGHT extends this trajectory to socially grounded relevance: the right document must satisfy an implicit social condition, not just match semantically or support problem solving. The uneven transfer of reasoning retrievers suggests that mathematical, scientific, or coding-oriented reasoning does not automatically yield socially aware retrieval. This matters for downstream RAG systems, where topical evidence may still be socially wrong.

## Limitations

The current set of tasks in SOCIAL-BRIGHT includes four forms of socially grounded retrieval that are not intended to exhaustively cover all social reasoning phenomena, languages, cultures, or communities. The benchmark currently includes only two languages: English for the interpersonal, social group grounding, and dogwhistle settings, and Chinese for humor. A more diverse set of tasks as well as broader multilingual and cross-cultural coverage remains important future work.

Our evaluation isolates the retrieval component rather than measuring complete retrieval-augmented generation pipelines. This choice makes it possible to diagnose whether a system can select socially appropriate evidence before generation, but it does not capture how generators use, ignore, or amplify retrieved documents in downstream responses. Future work should study socially grounded retrieval in end-to-end RAG settings, where failures in evidence selection may interact with generation behavior, prompting strategies, and user-facing decisions.

## References


Antoine Chaffin. 2025a. GTE-ModernColBERT.

Antoine Chaffin. 2025b. Reason-moderncolbert.

Jianlyu Chen, Junwei Lan, Chaofan Li, Defu Lian, and Zheng Liu. 2026. Reasonembed: Enhanced text embeddings for reasoning-intensive document retrieval. *Preprint*, arXiv:2510.08252.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Ruiqi He, Yushu He, Longju Bai, Jiarui Liu, Zhenjie Sun, Zenghao Tang, He Wang, Hanchen Xia, Rada Mihalcea, and Naihao Deng. 2025. Chumor 2.0: Towards better benchmarking Chinese humor understanding from (ruo zhi ba). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21799–21818, Vienna, Austria. Association for Computational Linguistics.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Rohan Jha, Bo Wang, Michael Günther, Georgios Mastrapas, Saba Sturua, Isabelle Mohr, Andreas Koukounas, Mohammad Kalim Akram, Nan Wang, and Han Xiao. 2024. Jina-ColBERT-v2: A general-purpose multilingual late interaction retriever. *Preprint*, arXiv:2408.16672.

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

Sean Lee, Rui Huang, Aamir Shakir, and Julius Lipp. 2025. Baked-in brilliance: Reranking meets rl with mxbai-rerank-v2.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From dogwhistles to bullhorns: Unveiling coded rhetoric with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15162–15180, Toronto, Canada. Association for Computational Linguistics.

Microsoft. 2026. Harrier-OSS-v1-0.6B.

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

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

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. *Preprint*, arXiv:2112.01488.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen tau Yih, Pang Wei Koh, and Luke Zettlemoyer. 2025. Reasonir: Training retrievers for reasoning tasks. *Preprint*, arXiv:2504.20595.

Raphael Sourty, Antoine Chaffin, Orion Weller, Paulo Demoura, and Amelie Chatelain. 2026. DenseOn with the LateOn: Open state-of-the-art single and multi-vector models. `https://huggingface.co/blog/lightonai/denseon-lateon`.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han yu Wang, Liu Haisu, Quan Shi, Zachary S Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O Arik, Danqi Chen, and Tao Yu. 2025. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. In *The Thirteenth International Conference on Learning Representations*.

Rikiya Takehi, Benjamin Clavié, Sean Lee, and Aamir Shakir. 2025. Fantastic (small) retrievers and how to train them: mxbai-edge-colbert-v0 tech report. *Preprint*, arXiv:2510.14880.

Nandan Thakur, Jimmy Lin, Sam Havens, Michael Carbin, Omar Khattab, and Andrew Drozdov. 2025. Freshstack: Building realistic benchmarks for evaluating retrieval on technical documents. *Preprint*, arXiv:2504.13128.

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)*.

Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, Daniel Cer, Alice Lisak, Min Choi, Lucas Gonzalez, Omar Sanseviero, Glenn Cameron, Ian Ballantyne, Kat Black, Kaifeng Chen, and 70 others. 2025. EmbeddingGemma: Powerful and lightweight text representations. *Preprint*, arXiv:2509.20354.

Feng Wang, Yuqing Li, and Han Xiao. 2025. jina-reranker-v3: Last but not late interaction for listwise document reranking. *Preprint*, arXiv:2509.25085.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Multilingual E5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Xiaoyue Wang, Jianyou Wang, Weili Cao, Kaicheng Wang, Ramamohan Paturi, and Leon Bergen. 2024b. Birco: A benchmark of information retrieval tasks with complex objectives. *Preprint*, arXiv:2402.14151.

Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. 2025a. FollowIR: Evaluating and teaching information retrieval models to follow instructions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11926–11942, Albuquerque, New Mexico. Association for Computational Linguistics.

Orion Weller, Benjamin Chang, Eugene Yang, Mahsa Yarmohammadi, Samuel Barham, Sean MacAvaney, Arman Cohan, Luca Soldaini, Benjamin Van Durme, and Dawn Lawrie. 2025b. mfollowir: A multilingual benchmark for instruction following in retrieval. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part II*, page 295–310, Berlin, Heidelberg. Springer-Verlag.

Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed. 2024. Rar-b: Reasoning as retrieval benchmark. *Preprint*, arXiv:2404.06347.

Puxuan Yu, Luke Merrick, Gaurav Nuti, and Daniel Campos. 2024. Arctic-Embed 2.0: Multilingual retrieval without compromise. *Preprint*, arXiv:2412.04506.

Qingcheng Zeng, Puxuan Yu, Aman Mehta, Fuheng Zhao, and Rajhans Samdani. 2026. Dual-view training for instruction-following information retrieval. *Preprint*, arXiv:2604.18845.

Xiangxu Zhang, Lei Li, Xiao Zhou, and Zheng Liu. 2026. R2med: A benchmark for reasoning-driven medical retrieval. *Preprint*, arXiv:2505.14558.

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

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, Youcheng Pan, Yang Xiang, Meishan Zhang, Haofen Wang, Jun Yu, Baotian Hu, and Min Zhang. 2025. KaLM-Embedding-V2: Superior training techniques and data inspire a versatile embedding model. *Preprint*, arXiv:2506.20923.

Yilun Zhao, Jinbiao Wei, Tingyu Song, Siyue Zhang, Chen Zhao, and Arman Cohan. 2026. Rethinking reasoning-intensive retrieval: Evaluating and advancing retrievers in agentic search systems. *Preprint*, arXiv:2605.04018.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D. Manning, Peter Henderson, and Daniel E. Ho. 2025. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, CSLAW '25, page 169–193, New York, NY, USA. Association for Computing Machinery.

# A Annotation Guidelines

In this section, we describe the annotation guidelines used to validate SOCIAL-BRIGHT.

## A.1 Annotator Details

All manual validation was conducted by three annotators who hold bachelor's degrees and are bilingual in Chinese and English. Each subset was double-annotated by two of the three annotators using the corresponding guidelines below. We retained only data points for which both annotators agreed that the instance should be kept; otherwise, the data point was discarded. The annotators are also authors of this paper, and their annotation work is reflected in the credit they receive as authors.

## A.2 Interpersonal Response Retrieval

For interpersonal response retrieval, annotators validate whether a chosen/rejected response pair can be used as a retrieval instance. The specific instructions are as follows:

1. Read the social question, the chosen response, and the rejected response.
2. Check whether the chosen response directly answers the question and is socially appropriate for the situation.
3. Check whether the chosen response is clearly more appropriate than the rejected response, considering contextual fit, empathy, respect, boundary awareness, and constructive handling of the situation when relevant.
4. Mark the pair as invalid if the chosen response is off-topic, if both responses are similarly appropriate, if the rejected response is better, or if the question does not contain enough context to support a clear preference.

Only pairs that pass all checks are retained. Annotators prioritize social appropriateness and contextual fit over fluency alone.

## A.3 Social Group Grounding Retrieval

For social group grounding retrieval, annotators validate whether each synthesized query, positive document, and negative document can be used as a retrieval instance. The specific instructions are as follows:

1. Read the query, the positive document, and the negative document.
2. Check whether the positive document is relevant to the query and satisfies the intended social group, role, and contextual condition.
3. Check whether the negative document is not relevant to the query because it violates at least one decisive condition, such as the social attribute, role, geography, era, or institutional status.
4. Check whether all statements about the people and social groups in the documents are factually correct and do not rely on unsupported identity inference.
5. Check whether both documents are fluent, natural, and self-contained, rather than written as explanations of the annotation task.

Only examples that pass all checks are retained.

## A.4 Humor Explanation Retrieval

For humor explanation retrieval, annotators validate whether each joke, positive explanation, and negative explanation can be used as a retrieval instance. The specific instructions are as follows:

1. Read the joke, the positive explanation, and the negative explanation.
2. Check whether the positive explanation provides a sufficiently complete account of why the joke is funny, including the central ambiguity, incongruity, reversal, implication, or cultural reference when relevant.
3. Check whether the negative explanation is materially worse than the positive explanation because it fails to fully capture the core humor mechanism.
4. Mark the pair as invalid if the positive explanation is incomplete or unconvincing, or if the negative explanation also gives a valid explanation of the joke's central meaning.

Only examples that pass all checks are retained.

### A.5 Dogwhistle Reasoning Retrieval

For dogwhistle reasoning retrieval, annotators validate whether each synthesized negative explanation can be used as a matched hard negative. The specific instructions are as follows:

1. Read the linguistic context, the source explanation, and the synthesized negative explanation.

2. Check whether the source explanation identifies the coded or covert social meaning conveyed by the dogwhistle.

3. Check whether the negative explanation is truly negative: it may sound plausible and reuse surface details from the context, but it must miss, misread, literalize, or downplay the coded social meaning.

4. Mark the negative as invalid if it correctly explains the covert meaning, substantially copies the source explanation, or is too generic or off-topic to serve as a meaningful hard negative.

Only examples with valid hard negatives are retained.

## B False-Negative Control

A potential risk in full-corpus retrieval benchmarks is that a document marked non-relevant for one query may in fact be a valid answer to another query. We address this risk according to the structure of each subset. For social group grounding retrieval, humor explanation retrieval, and dogwhistle reasoning retrieval, relevance is anchored to an instance-specific social-group condition, joke, or dogwhistle context; documents from other queries are therefore unlikely to become valid cross-query positives, so we do not apply additional corpus-level filtering beyond the manual validation described above. Interpersonal response retrieval is different because a response can sometimes be appropriate across multiple social situations. For this subset, we perform an additional false-negative screen: we mine potentially relevant cross-query candidates using the union of BM25 (Robertson and Zaragoza, 2009) and the strong dense retriever KaLM-Embedding-Gemma3-12B-2511 (Zhao et al., 2025), then ask GPT-5.4-mini to judge whether each mined response is an appropriate answer to the target query and is more suitable than the query's matched hard negative. Query-document pairs that pass this screen are added to the relevance judgments as additional positives. Thus, interpersonal response retrieval retains one matched hard-negative pair per query for MPA, while its full-corpus evaluation can include multiple relevant documents for a query.

## C Dataset Examples

Tables 4–7 show representative retrieval instances from the four subsets of SOCIAL-BRIGHT. Each example includes the query, the relevant document, and the matched hard negative; examples are lightly edited for readability where needed.

| Field | Text |
|---|---|
| Query | You receive a message from a close friend that says they "almost forgot" your birthday. How do you interpret their message and how would you respond? |
| Relevant document | The interpretation depends on the sender and the relationship. It could be a passive-aggressive comment implying that your birthday was not important enough to remember, in which case it is best to respond politely without escalating. It could also be self-deprecating, meaning the friend is joking about their poor memory, where a light-hearted reply may ease the tension. Or it may imply that they have been unusually busy and are prompting you to ask how they have been doing. |
| Matched hard negative | I would interpret the message as meaning that they remembered my birthday and that it meant a lot to be thought of. I would respond by saying, "I appreciate you remembering me and reaching out to me on my birthday." |

Table 4: Representative example from the interpersonal response retrieval subset.

| Field | Text |
|---|---|
| Query | Retrieve reporting on Latina judges whose rise to high courts or federal benches was covered as a milestone for representation and whose rulings or public presence shaped debate about law, legitimacy, and access to justice. |
| Relevant document | Becoming the first Latina to serve on the U.S. Supreme Court made Sonia Sotomayor a historic figure long before she had written her full body of opinions. Coverage of her career often connects that milestone to a broader public conversation about who gets to interpret the law and what lived experience means on the bench. Raised in the Bronx by Puerto Rican parents, she moved through the prosecutor's office, private practice, the federal trial court, and the appellate bench before joining the Court. Journalists and scholars alike have treated her speeches, dissents, and visibility in civic life as part of a larger story about legitimacy, representation, and whether elite legal institutions can reflect the country they serve. |
| Matched hard negative | Kamala Harris's rise from San Francisco district attorney to California attorney general and then to national office drew sustained reporting about diversity in legal institutions, public trust, and the visibility of women of color in positions of authority. Coverage of her career often connects courtroom strategy, criminal justice policy, and national representation to broader debates over access to justice and institutional legitimacy. Journalists have treated her public standing as part of a larger story about who gets to shape the law, how legal authority is perceived, and whether elite institutions can reflect the country they serve. |

Table 5: Representative example from the social group grounding retrieval subset.

| Field | Text |
|---|---|
| Query | English gloss of a Chinese joke title: A bathroom scale is a free duplicate-checking machine. |
| Relevant document | The joke turns on a pun. "Duplicate checking" normally refers to detecting textual overlap in academic or publishing contexts, but the title repurposes the phrase for checking body weight on a bathroom scale. The mismatch between a formal document-review term and an everyday object creates the humorous effect. |
| Matched hard negative | The humor comes from presenting an ordinary bathroom scale in exaggerated technical language, as if a simple household object had been rebranded as a professional checking service. The title sounds like mock advertising for a free machine that verifies duplicate items, creating a mismatch between a mundane object and an institutional-sounding function. |

Table 6: Representative example from the humor explanation retrieval subset. The original query is in Chinese; we show an English gloss for readability.

| Field | Text |
|---|---|
| Query | I still believe the answer to any problem lies with the people. I believe in states' rights. I believe in people doing as much as they can for themselves at the community level and at the private level, and I believe we've distorted the balance of our government today by giving powers that were never intended in the Constitution to that federal establishment. |
| Relevant document | Although "states' rights" can appear racially neutral, its political use during the civil-rights era often functioned as a coded racial appeal. In this context, the phrase signaled resistance to federal mandates to integrate schools and society, allowing speakers to invoke local control while appealing to racial anxiety without naming race directly. |
| Matched hard negative | Here, "states' rights" is being used as a conventional constitutional argument about federalism and the balance of government. The speaker says the answer to any problem lies with the people, communities, and the private level, and argues that powers never intended in the Constitution have been shifted to the federal establishment. In this reading, the phrase emphasizes local control, decentralized decision-making, and skepticism toward federal authority. |

Table 7: Representative example from the dogwhistle reasoning retrieval subset.

# D Additional Analysis Details

## D.1 Retriever Query Instructions

Table 8 lists the query-side instructions used for instruction-aware retrievers. These prompts are applied only to queries; corpus documents are encoded without task prefixes. The social group grounding subset uses an empty query prompt.

| Subset | Query-side instruction |
|---|---|
| Interpersonal response | Given a social situation, retrieve the response that is most appropriate, empathetic, and socially well-reasoned. |
| Social group grounding | *None*; the query prompt is empty by default. |
| Humor explanation | 给定一个中文笑话，检索最能解释其幽默机制和笑点的解释。 |
| Dogwhistle reasoning | Given a social or political text containing a potentially coded expression, retrieve the explanation that best captures its covert social meaning, in-group signal, and contextual use. |

Table 8: Query-side retrieval instructions used for instruction-aware retrievers.

## D.2 First-Stage Recall@10

Table 9 reports Recall@10 for the same first-stage retriever runs summarized in Table 2. These values provide a short-list recall diagnostic: relevant documents are often retrieved near the top of the ranking even when models fail to rank them first or score them above matched hard negatives.

| Model | Inter. | Group | Humor | Dogw. | Macro |
|---|---|---|---|---|---|
| **Open-source bi-encoders (<1B)** | | | | | |
| Harrier-OSS-0.6B | 0.655 | 0.947 | 0.998 | 0.546 | 0.786 |
| mE5-large-instruct | 0.684 | 0.929 | 0.993 | 0.520 | 0.782 |
| Qwen3-Emb-0.6B | 0.646 | 0.965 | 0.992 | 0.617 | 0.805 |
| DenseOn | 0.717 | 0.929 | 0.920 | 0.740 | 0.827 |
| Arctic-Embed-L-v2 | 0.656 | 0.929 | 0.988 | 0.760 | 0.833 |
| EmbeddingGemma-300M | 0.346 | 0.752 | 0.827 | 0.439 | 0.591 |
| **Open-source bi-encoders (>1B)** | | | | | |
| KaLM-Gemma3-12B | 0.823 | 0.991 | 0.999 | 0.689 | 0.875 |
| Qwen3-Emb-8B | 0.740 | 0.956 | 0.995 | 0.791 | 0.870 |
| Qwen3-Emb-4B | 0.699 | 0.965 | 0.989 | 0.719 | 0.843 |
| **Late-interaction retrievers** | | | | | |
| Jina-ColBERT-v2 | 0.580 | 0.920 | 0.960 | 0.607 | 0.767 |
| mxbai-edge-colbert | 0.491 | 0.920 | 0.776 | 0.622 | 0.703 |
| ColBERTv2 | 0.512 | 0.885 | 0.451 | 0.526 | 0.593 |
| LateOn | 0.457 | 0.788 | 0.213 | 0.383 | 0.460 |
| GTE-ModernColBERT | 0.587 | 0.894 | 0.919 | 0.612 | 0.753 |

Table 9: Recall@10 for the first-stage retrievers in Table 2. Inter., Group, Humor, and Dogw. denote interpersonal response retrieval, social group grounding retrieval, humor explanation retrieval, and dogwhistle reasoning retrieval, respectively. Macro averages are computed across the four subsets.

## D.3 BM25 Lexical Baseline

To test whether SOCIAL-BRIGHT can be solved by lexical matching alone, we evaluate BM25 (Robertson and Zaragoza, 2009) under the same full-corpus protocol used for the main experiments. Table 10 shows that BM25 has low top-rank and matched-pair performance across the benchmark. Although BM25 retrieves many relevant documents within the top 10 in some subsets, especially humor explanation retrieval, it rarely ranks the socially correct document first and often prefers the lexically similar hard negative. This confirms that the benchmark targets socially grounded relevance rather than surface lexical overlap.

| Subset | Hit@1 | MPA | nDCG@5 | R@10 |
|---|---|---|---|---|
| Interpersonal response | 0.077 | 0.194 | 0.195 | 0.321 |
| Social group grounding | 0.018 | 0.018 | 0.270 | 0.584 |
| Humor explanation | 0.033 | 0.033 | 0.623 | 0.986 |
| Dogwhistle reasoning | 0.005 | 0.005 | 0.266 | 0.577 |
| Macro average | 0.033 | 0.062 | 0.338 | 0.617 |

Table 10: BM25 lexical baseline on SOCIAL-BRIGHT. R@10 denotes Recall@10. BM25 can retrieve relevant candidates within the top 10 in some subsets, but its low Hit@1 and MPA show that lexical matching does not reliably identify the socially correct document over matched hard negatives.

## D.4 Social Group Grounding Query-Shuffle Control

To check whether the social group grounding reranking gains depend on the correct query-document alignment, we run a query-shuffle control. We permute social group grounding queries across examples while keeping the document collection and relevance labels fixed. A reranker that mainly exploits query-independent document priors should remain strong under this control, whereas a reranker that uses the query condition should degrade sharply when the query is mismatched.

The control sharply reduces top-rank performance for all three rerankers, confirming that the strong social group grounding results in Table 3 are not explained by a query-independent preference for particular documents. The larger rerankers retain above-chance MPA under the control, suggesting that some document priors remain, but the large gap from the original-query setting indicates that the main gains require the correct social group grounding query.

| Model | Original queries | | | Query-shuffle control | | |
|---|---|---|---|---|---|---|
| | Hit@1 | MPA | nDCG@5 | Hit@1 | MPA | nDCG@5 |
| Qwen3-Reranker-0.6B | 0.566 | 0.637 | 0.809 | 0.000 | 0.345 | 0.004 |
| Qwen3-Reranker-4B | 0.850 | 0.920 | 0.939 | 0.000 | 0.584 | 0.016 |
| Qwen3-Reranker-8B | 0.841 | 0.903 | 0.937 | 0.000 | 0.584 | 0.004 |

Table 11: Query-shuffle control for Qwen rerankers on social group grounding retrieval.

### D.5 Evaluation Framework and Compute

All reported retrieval evaluations are implemented as MTEB-compatible retrieval tasks and run using the MTEB evaluation framework (Muennighoff et al., 2023). This includes the first-stage retriever evaluations, BM25 baseline, reasoning retriever analysis, reranker experiments, query-expansion experiments, and auxiliary control runs, which all use the same task interfaces and metric computation. Experiments were conducted on a server with four NVIDIA H100 GPUs. Across all model runs and analysis experiments, the total compute budget was approximately 500 GPU-hours.

## E Prompt Templates

### E.1 Social Group Grounding Retrieval

For each seed group, we render the following prompt template to synthesize a query, a positive document, and a matched hard negative. The placeholders are filled from the seed group and the assigned style rotation.

**Prompt template for social group grounding retrieval**

**System prompt.** Generate one high-quality synthetic example for a social group grounding retrieval benchmark. Return exactly three strings: `query`, `positive_document`, and `negative_document`. The query must be group-centric and explicitly contain the target social group phrase. The positive document must be about one specific person who truly belongs to the target group, preferably without repeating the full group phrase. The negative document must have higher lexical overlap with the query than the positive document, start from the social group named in the query, and be wrong because it violates a decisive condition such as social attribute, role, geography, era, or another nearby boundary. Both documents must read as realistic standalone prose and must not mention the retrieval task, relevance labels, or annotation process.

**User template.** Generate one synthetic retrieval triplet in `{output_language}`.

Target social group: `{canonical_name}`.

Group description: `{description}`.

Boundary note: `{ambiguity_note}`.

Assigned style rotation: query style `{query_style_name}` with opening `{query_opening}`; positive style `{positive_style_name}`; negative style `{negative_style_name}` with opening `{negative_opening}`.

Output a short natural query that starts with `{query_opening}` and contains the exact phrase `{canonical_name}`. Output a realistic positive paragraph about one specific person who truly belongs to the target group. Output a realistic negative paragraph that starts with `{negative_opening}`, introduces the target group early, overlaps more lexically with the query than the positive document, and remains plausible but fails one decisive condition. Return only JSON matching the schema.

### E.2 GPT-5.4 Query Expansion

For the query expansion analysis, GPT-5.4 receives only the original query and subset-level task guidance. It does not receive the corpus, relevance labels, relevant documents, hard negatives, or answers. The generated expansion is then converted into the retrieval query by concatenating the original query, the compact expanded query, the reasoning chain, and search cues.

**Prompt template for GPT-5.4 query expansion**

**System prompt.** You are expanding retrieval queries for social reasoning benchmarks. You will only see the original query. You will not see the corpus, relevant documents, irrelevant documents, labels, or answers.

**Goal.** Make implicit reasoning bridges explicit so a text retriever can find relevant documents. Preserve the original query intent exactly. Keep the expansion broad enough for retrieval; do not turn the query into a request for one guessed document or one guessed answer. Include social, cultural, interpersonal, factual, or pragmatic reasoning cues when they are necessary for matching relevant documents.

**Rules.** Do not add unsupported constraints. Do not contradict or narrow the original query. Do not mention benchmark labels, positives, negatives, relevance judgments, or document IDs. Avoid endorsing stereotypes or making normative claims. Use the same primary language as the original query when possible. Be compact: the expanded query should be useful for semantic search, not a long essay. Return only valid JSON matching the requested schema.

**Task-specific guidance.** The guidance is instantiated by subset: interpersonal response retrieval asks the model to surface empathy, conflict resolution, boundaries, social norms, and likely consequences; social group grounding retrieval asks it to surface group membership, role,

institutional, demographic, geographic, historical, and cultural bridge concepts; humor explanation retrieval asks it to surface incongruity, ambiguity, violated expectation, speaker intention, social norm, or shared cultural assumption; dogwhistle reasoning retrieval asks it to surface covert social meaning, target group, in-group signal, plausible deniability, and pragmatic function in a neutral analytical way.

**User template.** Dataset: `{dataset_name}`.

Task-specific guidance: `{task_guidance}`.

Original query: `{query}`.

Return JSON with a compact `expanded_query`, 2–5 `reasoning_chain` steps, 3–12 `search_terms`, and `cautions`.

**Retrieval query format.** The final query submitted to the retriever has the following fields: `Original query`, `Reasoning-expanded query`, `Reasoning chain`, and `Search cues`.

## F Social Group Inventory

The social group grounding retrieval subset starts from 200 human-written social groups across ten domains. Tables 12 and 13 list the full inventory.

## G License Statement

**Benchmark sources.** SOCIAL-BRIGHT combines public research resources, human-written seed material, and model-assisted transformations. The interpersonal response retrieval subset is derived from the Social Reasoning RLHF dataset released by ProlificAI under the MIT License.[2] The humor explanation subset is derived from CHUMOR 2.0, which is released under the MIT License (He et al., 2025). The dogwhistle reasoning subset is derived from the dogwhistle glossary of Mendelsohn et al. (2023). The social group grounding subset is built from our human-written social group inventory and model-assisted synthesis followed by human validation. We use all source resources in accordance with their verified licenses, terms, and citation requirements, and preserve source attribution in the benchmark documentation and release metadata.

**Models.** We evaluate publicly released retrievers and rerankers under their respective model licenses or access terms, and we do not redistribute model weights. Models used under the Apache License 2.0 include Qwen3-Embedding and Qwen3-Reranker (Zhang et al., 2025), Arctic-Embed (Yu et al., 2024), DenseOn and LateOn (Sourty et al., 2026), GTE-ModernColBERT (Chaffin, 2025a), mxbai-edge-ColBERT (Takehi et al., 2025), ReasonEmbed (Chen et al., 2026), and mxbai-rerank-v2 (Lee et al., 2025). Models used under the MIT License include multilingual-E5 (Wang et al., 2024a), Harrier-OSS (Microsoft, 2026), and ColBERTv2 (Santhanam et al., 2022). Models used under the Creative Commons Attribution-NonCommercial 4.0 International License include Jina-ColBERT-v2 (Jha et al., 2024), ReasonIR (Shao et al., 2025), Reason-ModernColBERT (Chaffin, 2025b), and jina-reranker-v3 (Wang et al., 2025). EmbeddingGemma is used under the Gemma Terms of Use (Vera et al., 2025), and KaLM-Embedding is used under the KaLM-Embedding license terms (Zhao et al., 2025). GPT-5.4 and GPT-5.4-mini are accessed through the OpenAI API for data construction, query expansion, and false-negative screening. All model use is for non-commercial academic research and follows the applicable provider licenses or terms of service.

**Release.** We will release our original code, prompts, social group inventory, evaluation scripts, and benchmark metadata under the Apache License 2.0. Source-derived benchmark artifacts will be released with attribution to the original resources and only to the extent permitted by the corresponding source licenses and terms.

## H GenAI Statement

We disclose the use of generative AI tools in this work in accordance with the ACL Policy on the Use of AI Writing Assistance.

**AI use in benchmark construction.** Generative AI was used to construct several benchmark artifacts. GPT-5.4 was used to synthesize query–positive–negative triplets for social group grounding retrieval, generate matched hard-negative explanations for the humor explanation and dogwhistle reasoning subsets, and produce query expansions for the analysis in Section 5.2. GPT-5.4-mini was used in the interpersonal response retrieval false-negative screen described in Appendix B. In all cases, model outputs were subject to the filtering and human validation procedures described in Appendix A; generated examples were retained only when they satisfied the relevant annotation criteria.

**AI use in experiments.** The retrievers and rerankers evaluated in this work are neural retrieval systems, including embedding models, late-

[2] ProlificAI/social-reasoning-rlhf.

| Domain | Social groups |
|---|---|
| Politics and government | Black politicians; Latino politicians; Asian American politicians; Indigenous politicians; Muslim politicians; Jewish politicians; women politicians; LGBTQ politicians; immigrant politicians; disabled politicians; Black mayors; Latina governors; Asian American cabinet secretaries; Indigenous parliamentarians; Muslim ambassadors; Jewish senators; women prime ministers; LGBTQ city mayors; immigrant city council members; disabled legislators. |
| Law and judiciary | Black judges; Latina judges; Asian American lawyers; Indigenous legal advocates; Muslim attorneys; Jewish law professors; women prosecutors; LGBTQ public defenders; immigrant civil rights lawyers; disabled jurists; Black appellate judges; Latina civil rights attorneys; Asian American law school deans; Indigenous tribal judges; Muslim public-interest lawyers; Jewish constitutional litigators; women solicitors general; LGBTQ immigration attorneys; immigrant public defenders; Deaf lawyers. |
| Business and finance | women CEOs; Black CEOs; Latino entrepreneurs; Asian American founders; Muslim investors; Jewish venture capitalists; immigrant entrepreneurs; disabled entrepreneurs; LGBTQ founders; women board chairs; Black investment bankers; Latina chief financial officers; Asian American management consultants; Indigenous business owners; Muslim fintech founders; Jewish private equity executives; women venture capital partners; LGBTQ retail executives; immigrant restaurateurs; blind entrepreneurs. |
| Technology and engineering | women software engineers; Black software engineers; Latino technologists; Asian American tech executives; Indigenous technologists; Muslim data scientists; Jewish engineers; immigrant engineers; disabled open-source developers; LGBTQ cybersecurity professionals; Black robotics engineers; Latina product managers; Asian American chip designers; Indigenous climate technologists; Muslim AI researchers; Jewish semiconductor executives; women game designers; LGBTQ bioinformatics researchers; immigrant startup engineers; Deaf software testers. |
| Academia and education | women physicists; Black scientists; Latino economists; Asian American sociologists; Indigenous historians; Muslim academics; Jewish historians; immigrant researchers; disabled scholars; LGBTQ education scholars; Black historians; Latina education professors; Asian American philosophers; Indigenous archaeologists; Muslim political scientists; Jewish anthropologists; women mathematicians; LGBTQ librarians; immigrant linguists; first-generation college presidents. |

Table 12: Social group inventory used for social group grounding retrieval, part 1.

interaction models, and cross-encoder rerankers. Their use is the object of evaluation rather than auxiliary writing or analysis support, and the experimental protocol is described in Section 4.

**AI use in writing.** AI assistants, including OpenAI Codex, were used for writing support, including sentence-level drafting, grammar correction, wording refinement, and LaTeX formatting suggestions. All scientific claims, dataset design decisions, experimental results, analyses, and conclusions were specified, checked, and approved by the human authors. AI assistants were not used to fabricate citations, experimental numbers, or empirical findings.

**Responsibility.** The authors take full responsibility for the content of this paper, including any text that may have been drafted or edited with AI assistance.

## I Ethics Statement

SOCIAL-BRIGHT is designed to evaluate socially grounded retrieval, and several subsets necessarily involve sensitive social categories, interpersonal norms, and coded or discriminatory language. These materials are included for research evaluation of retrieval systems, not to endorse the viewpoints, stereotypes, or coded meanings that may appear in source examples or hard negatives. The dogwhistle reasoning subset in particular may contain offensive or hateful rhetoric because the task requires identifying covert social meaning. Users should treat the benchmark as an evaluation resource and should not use it for profiling individuals or communities, making decisions about protected groups, or deploying unvalidated systems in high-stakes social settings.

We take several steps to reduce risk. Social group grounding examples are manually validated for factuality and are rejected when they rely on unsupported identity inference or unnatural generated prose. Humor and dogwhistle hard negatives are validated to ensure that the negative document is materially incorrect rather than simply a different valid interpretation. Interpersonal examples are additionally screened for potential false negatives,

| Domain | Social groups |
|---|---|
| Healthcare and medicine | women physicians; Black physicians; Latino surgeons; Asian American psychiatrists; Indigenous public health leaders; Muslim nurses; Jewish pediatricians; immigrant doctors; disabled physicians; LGBTQ therapists; Black psychiatrists; Latina epidemiologists; Asian American oncologists; Indigenous midwives; Muslim pharmacists; Jewish neurologists; women hospital administrators; LGBTQ surgeons; immigrant dentists; first-generation doctors. |
| Media and journalism | Black journalists; Latina television journalists; Asian American political commentators; Indigenous reporters; Muslim broadcasters; Jewish editors; immigrant columnists; disabled broadcasters; LGBTQ news anchors; women investigative reporters; Black documentary hosts; Latina newspaper columnists; Asian American foreign correspondents; Indigenous podcasters; Muslim magazine editors; Jewish television producers; women photojournalists; LGBTQ documentary journalists; immigrant radio hosts; Deaf television presenters. |
| Entertainment and arts | Black actors; Latina actresses; Asian American filmmakers; Indigenous authors; Muslim musicians; Jewish comedians; immigrant novelists; disabled performers; LGBTQ actors; women screenwriters; Black playwrights; Latina choreographers; Asian American novelists; Indigenous painters; Muslim poets; Jewish film directors; women conductors; LGBTQ dancers; immigrant sculptors; blind musicians. |
| Sports | Black quarterbacks; women soccer players; Muslim athletes; Paralympic athletes; LGBTQ athletes; Asian American baseball players; Indigenous runners; women boxers; immigrant coaches; Jewish basketball players; Black tennis players; Latina basketball players; Asian American golfers; Indigenous hockey players; Muslim soccer coaches; Jewish Olympians; women motorsport drivers; LGBTQ swimmers; immigrant baseball managers; Deaf marathon runners. |
| Civil society and faith | Black civil rights leaders; Latina nonprofit executives; Asian American community organizers; Indigenous spiritual leaders; Muslim humanitarian leaders; Jewish rabbis; immigrant clergy; disabled disability-rights activists; LGBTQ clergy; women labor union leaders; Black labor organizers; Latina immigrant-rights advocates; Asian American mutual-aid leaders; Indigenous water protectors; Muslim prison chaplains; Jewish social justice organizers; immigrant pastors; disabled accessibility activists; LGBTQ youth organizers; women philanthropy executives. |

Table 13: Social group inventory used for social group grounding retrieval, part 2.

because a response may be appropriate across multiple social situations. These procedures reduce, but do not eliminate, the possibility that the benchmark reflects annotator subjectivity, source-data bias, cultural incompleteness, or residual artifacts from model-assisted generation. We therefore recommend using SOCIAL-BRIGHT as a diagnostic benchmark for research rather than as a sole basis for claims about social reasoning ability in deployed retrieval or RAG systems.