Evaluating Socio-Ecological Bias in Retrieval Augmented Generation: A Case Study on Interdisciplinary Agricultural Resilience

Anonymous ACL submission

Abstract

This study examines domain bias in Retrieval-Augmented Generation (RAG) systems within the socio-ecological context of agricultural resilience. Leveraging a multi-model framework comprising DeepSeek-R1 and Llama-3.2 as generative backbones, paired with Nomic-Embed-Text and EmbeddingGemma for document embedding, we construct balanced corpora of ecological and social science articles and design two controlled experiments to disentangle retrieval and prompt effects. The results reveal a nuanced, multi-stage bias pattern: while the retrieval stage exhibits a consistent preference for ecological variables (particularly in context relevancy), the generation stage demonstrates a significant reversal, favoring social variables in response to faithfulness under prompt-bias conditions. Our findings highlight the hidden risks associated with domain bias present in RAG applications in socio-ecological policy-making.

1 Introduction

Socio-ecology is an interdisciplinary field that examines the interactions between human societies and ecological systems (Kelly et al., 2019). It plays a critical role in informing policy and decision-making, based on a holistic perspective that integrates the complexity of the social and ecological dimensions of human–environment (McGinnis and Ostrom, 2014). When the balance between social and ecological domains is not maintained, analyses risk becoming biased toward one domain, potentially oversimplifying challenges such as poverty and natural disasters (Pauley et al., 2019).

Large Language Models (LLMs) hold considerable potential for supporting such holistic analyses of socio-ecology given their strong performance in general-domain tasks. However, their generalization often degrades in interdisciplinary domains that require knowledge transfer across domains (Mammides and Papadopoulos, 2024; Zhang et al., 2025b). Well-documented challenges include hallucinations and inconsistency, raising concerns about reliability in high-stakes contexts (Xu et al., 2023; Lewis et al., 2020).

Retrieval-Augmented Generation (RAG) (Gao et al., 2023) enhances LLMs by supplying external contextual information, which, in scientific applications, is typically drawn from the research literature, guiding the model's generation and mitigating issues of hallucination and incomplete knowledge (Lewis et al., 2020). Yet, the application of RAG in the context of socio-ecological studies requires accurate analysis of trustworthiness, explainability, and bias, particularly for policy-making (Vizniuk et al., 2025).

Previous research highlights a systematic bias toward the social system in social–ecological studies, most likely rooted in the historical development and predominantly social science–driven nature of these studies, which have prioritized variables linked to collective action over biophysical system components (Partelow, 2018). In this study, we examine whether a similar bias exists in RAGs, as such biases may distort analytical outcomes and misguide policy recommendations (Vizniuk et al., 2025).

We conduct controlled experiments under the DPSIR (drivers, pressures, state, impact, and response) framework (Maxim et al., 2009), a widely used model for analyzing human–environment interactions. DPSIR conceptualizes causal chains by linking socio-economic drivers to environmental pressures, the resulting state of ecosystems, their impacts on society, and policy responses.

Our case study focuses on agricultural resilience in Henan Province, China–a major grain-producing region and a representative socio-ecological system. We investigate domain bias arising from both retrieval and prompt-driven generation in RAGs. To this end, we construct curated social and ecolog-

ical corpora that serve as the document pool for our experiments. We prompt the RAG systems with a set of systematically designed socio-ecological questions and evaluate their outputs using Answer Relevancy, Faithfulness, and Contextual Relevancy. The results reveal a robust and consistent ecological advantage in contextual relevancy across both retrieval and prompting stages, while faithfulness exhibits stage-dependent variation, shifting from a weak ecological preference during retrieval to a social bias under prompt variation.

2 Previous Work

Although LLMs achieve state-of-the-art performance on general-domain benchmarks, their effectiveness declines on domain-specific (Sengupta et al., 2025; Rao et al., 2025) and cross-domain tasks, where specialized terminology and interdisciplinary reasoning remain underrepresented in pretraining corpora (Zhang et al., 2025b). Recent work further shows that challenges with domain sensitivity are not only observable in model outputs but can also be linked to shifts in internal representations during fine-tuning, raising concerns about representation collapse and the persistence of domain-specific biases (Razdaibiedina et al., 2023). In addition Chen et al. (2024) argue that LLMs' performance across applications and domains is unstable or inconsistent over time.

In highly complex interdisciplinary domains such as environmental science, Zhang et al. (2025b) show that LLMs' reasoning performance is limited across domains and require additional domain-specific supervision. While fine-tuning offers one solution, it introduces substantial computational cost and technical challenges, including missing specialized concepts (Wu et al., 2024; Zhu et al., 2024) and the risk of catastrophic forgetting (Huang et al., 2024; Luo et al., 2025).

Retrieval-Augmented Generation (RAG) is a partial solution to these limitations (Gao et al., 2023). Rather than relying on costly fine-tuning, RAG augments LLMs with external domain knowledge at inference time. However, this integration of external knowledge also introduces new challenges, most notably domain biases that can undermine fairness and distort model outputs (Wu et al., 2025).

Previous studies examine biases in RAGs along three dimensions: external data source biases, algorithmic and systemic biases, and risk-related concerns. External data sources frequently encode demographic and social biases (Wu et al., 2025; Zhang et al., 2025a), as well as intersectional biases that can be further amplified through RAG pipelines (Kim et al., 2025). Other studies in this dimension investigate biases in data and authorship (Abolghasemi et al., 2025), citation prioritization (Genovese et al., 2025), among others.

Algorithmic and systemic factors in RAG, spanning components such as generators, retrievers, embedders, refiners, and judgers, introduce or reinforce biases individually or through component interactions (Wu et al., 2025; Kim et al., 2025; Zhang et al., 2025a). Writing style biases (Cao, 2025), attribution biases, and evaluation biases (Schmidgall et al., 2024) emerge when embedding models or LLM judges prefer certain styles or outputs, while input framing, prompt engineering, and multilingual or translation processes add further layers of bias (Wu et al., 2025; Abolghasemi et al., 2025; Juhasz et al., 2024; Zhang et al., 2025a). These biases raise significant risk-related concerns: hallucinations, misrepresentations, policy violations, and fairness-utility trade-offs threaten reliability (Juhasz et al., 2024; Schmidgall et al., 2024).

To address these challenges, researchers have explored various mitigation strategies: adjusting the proportion and ranking of relevant documents for protected groups, increasing the number of retrieved documents, or employing larger generator models (Wu et al., 2025). More complex methods include reverse-biasing (Kim et al., 2025), post-processing techniques based on chain of thought (Ji et al., 2025), and model adaptation techniques Cheng et al. (2025) introduce a model adaptation technique to address the domain bias in LLMs.

3 Methodology

We adopt a behavioral approach to examine RAGs' interdisciplinary performance in socio-ecology. Specifically, we prompt the models with questions that integrate both social and ecological variables, while conditioning them on contexts drawn exclusively from either domain. Since the questions are inherently interdisciplinary, variation in the model's responses across the two context types is interpreted as evidence of domain bias, that is, a systematic tendency to privilege information from either the social or the ecological domain.

We construct domain-specific contexts for each domain from scientific articles compiled from separate corpora of social and ecological publications.

These documents serve as the basis for retrieving context passages during inference. Particular care is taken to balance the size and coverage of the two corpora in order to minimize content bias in the RAG document pool (Zhang et al., 2025a); further details are provided in Section 6.

To minimize prompt-specific impact, we implement a systematic prompt diversification strategy (Zhan et al., 2024). Rather than relying on a single manually crafted query, we generate multiple semantically equivalent prompts by a combinatorial product over lexical and structural variants along four controlled dimensions:

- Task verbs: Identify, Extract, Map, and Analyze.
- Framework specification: DPSIR and Drivers–Pressures–States–Impacts–Responses.
- Domain scope: agricultural resilience in Henan Province, resilience of agriculture in Henan, and vulnerability of agriculture in Central China.
- Causal directives: Make explicit the linkages, explain how categories relate, and describe causal chains between components.

Each prompt adheres to a fixed template:

> *{Verb} from any given data or document the ecological and social variables that correspond to each of the {Framework specification} categories, specifically in relation to {Domain scope}. {Causal directive} between categories (e.g., how a Driver leads to a Pressure).*

From the 72 possible combinations, we randomly sample 20 prompts to balance diversity with computational tractability. The phrase "ecological and social variables" is held constant across all prompts to ensure consistent interdisciplinary framing.

Following established protocols for RAG evaluation (Yu et al., 2025), we adopt an *LLM-as-judge* approach (Gu et al., 2024; Thakur et al., 2025) to assess model behavior. For each prompt and its generated answer, the evaluation framework employs a multi-agent judge to assess the full text of the retrieved context chunks, the generated response, and the question. The system's numeric scores (ranging from 0.0 to 1.0) from multiple LLMs are based on specific prompt rubrics. These scores then serve as the basis for complementary alignment evaluations, including context relevance to the answer and question, as follows:

- **Answer Relevance**: whether the generated answer directly addresses the question. Answer relevance serves as an indicator of the model's generation performance.
- **Faithfulness** : whether the generated answer is grounded in the retrieved context. The scores are based on the degree to which the entire retrieved context factually supports all claims in the answer.
- **Context Relevance** : the degree to which the retrieved chunks are pertinent to the prompt. This evaluates how essential the information in each chunk is for producing a correct and complete answer. Content relevance can be interpreted as an indicator of bias in the RAGs' document embedding space (Cao, 2025).

We adopt a Bayesian estimation approach to quantify uncertainty in the evaluation metrics and to assess the probability of systematic differences across domains. In this approach, the dominance of a domain d_1 over another domain d_2 is defined as the posterior probability that the mean score of d_1 exceeds that of d_2 for a given metric, i.e., $P(\mu_{d_1} > \mu_{d_2})$.[1] Following common practice, we assume approximate normality of the metric distributions.

In addition to posterior dominance probabilities, we report effect sizes using the posterior distribution of Cohen's d, which standardizes the mean difference between domains by the pooled standard deviation:

$$d = \frac{\mu_{d_1} - \mu_{d_2}}{\sigma_{\text{pooled}}}, \quad \sigma_{\text{pooled}} = \sqrt{\frac{\sigma_{d_1}^2 + \sigma_{d_2}^2}{2}}.$$

Positive values of the effect size (Cohen's d) indicate a bias toward d_1, while negative values indicate a bias toward d_2. Following standard conventions, the absolute values around 0.2, 0.5, and 0.8 can be interpreted as small, medium, and large effects, respectively. Uncertainty is summarized

[1] This approach plays a role similar to a classical t-test in comparing group means, but instead of providing a p-value under a null hypothesis, it yields full posterior distributions that allow direct probabilistic statements about mean differences. Moreover, it can provide posterior distributions over standardized effect sizes such as Cohen's d, offering both magnitude and uncertainty of domain differences.

with 94% Highest Density Intervals (HDIs) for both mean differences and standardized effect sizes, and convergence is verified using $\hat{R} \approx 1.0$.

4 Experiment Design

We structure the evaluation around two complementary experiments designed to disentangle retrieval effects from prompt effects.

Retrieval-Bias: By holding the prompt fixed and varying the RAG document pools, we test whether bias originates from the retrieval stage. The experiment is conducted on batches of 10 articles each under an identical prompt. If, under an identical prompt, the retrieved chunks systematically skew toward one domain (ecological or social), this indicates a bias that stems from the language model's embedding space used for retrieval, rather than from prompt interpretation.

Prompt-Bias: By holding the retrieved context fixed and varying the prompts, we test whether bias originates from the prompt side. The experiment is conducted on the full document pool, with prompts instantiated from the template described in the previous section. If, given the same retrieved chunks, generated answers systematically favor one domain (ecological or social), this indicates that the bias arises from how the language model interprets the prompt rather than from retrieval selection.

5 Experiment Setting

We implement RAG systems using LangChain's RetrievalQA module. We employ a Chroma retriever that extracts the top-4 chunks per query, resulting in approximately 1,200 tokens. The total retrieved context remains around 4,800 tokens. The maximum context window is 8K tokens, and the temperature is 0.5 to maintain a balance between creativity and factual stability.

For the generative component of our framework, we employ two distinct frontier-level open-source architectures: Llama-3.2 and DeepSeek-R1 (Guo et al., 2025), as representatives of dense decoder-only Causal Language Models (CausalLLM), and a Mixture-of-Experts (MoE) transformer architecture, respectively. This dual-model approach allows us to evaluate domain bias and ensures that our findings are robust across both architectures. The LLMs (generators) are explicitly instructed to reason under the DPSIR framework and to ground outputs in retrieved evidence. For document embedding, we utilize two high-performance models: Nomic-Embed-Text and EmbeddingGemma. These models were served through the Ollama backend to generate high-dimensional vectors for both queries and document segments.

The RAGs' document pool includes research articles curated from the Dimensions.ai database, focusing exclusively on Open Access journal articles published between 2020 and 2025. To ensure regional and thematic consistency, we retrieved publications containing the keyword 'Henan Province' in their titles or abstracts. To maintain comparability across domains and minimize content bias in the pool (Zhang et al., 2025a), we focused solely on the textual content, excluding references, tables, and figures. This decision is crucial as ecological publications often contain extensive spatial and temporal visualizations, whereas social science articles may rely more on descriptive text; thus, a text-only approach ensures a balanced comparison.

Domain assignment relied on the standardized Field of Research classification system native to Dimensions.ai. The social science corpus was constructed from 192 articles automatically categorized by the database under Commerce, Management, Tourism and Services, and Human Society. Similarly, the ecological corpus was derived from 179 publications assigned by the platform's taxonomy to Earth Science, Geoinformatics, and Environmental Sciences. Utilizing these pre-defined, database-driven categories resulted in a total pool of 371 scientific articles, ensuring that the corpora are representative of their respective domains while maintaining comparable token sizes for the RAG framework. Further details about the selected data is presented in Section 6.

All evaluations are operationalized using a dual-agent judging framework that incorporates both Llama-3.2 and DeepSeek-R1 as evaluators. To eliminate the inherent parametric bias of a single model, we implement a collaborative scoring mechanism in which each output is independently assessed by both models on a scale of 0 to 1. A final consensus score was then derived using an equally weighted aggregation. This ensemble approach ensures that the alignment checks against the user prompt and retrieved evidence are validated by two distinct architectures, providing a more robust and objective metric for RAG performance.

6 Data Statistics

The statistical characteristics of the textual content of the selected articles are summarized in Table 1. The social and ecological corpora are comparable in overall size and nearly identical average document lengths, as also illustrated in Figure 1. The social corpus contains slightly more documents because ecological articles are marginally longer on average. This balance in total token size, extending to the retrieval units with 1.2k social and 1.1k ecological chunks, minimizes potential bias in the RAG document pool and ensures that differences in the RAG performance are not simply driven by corpus scale. Ecological texts nevertheless show greater lexical variety, with a larger vocabulary and a higher type-token ratio.

	Ecological	Social
Total documents	179	192
Total tokens	1,025K	1,030K
Total chunks	1.1K	1.2K
Unique vocabulary size	33.6K	32.1K
Average document length	5.7K	5.3K
Average chunk length	883	856
Type-token ratio (TTR)	3.28%	3.11%

Table 1: Statistics of the ecological and social corpora.

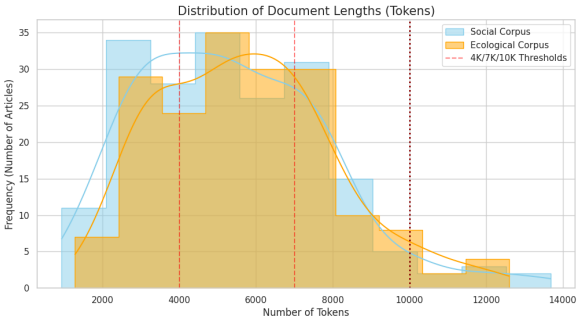

Figure 1: Distribution of Document Length

Table 2 provides a snapshot of the ecological and social corpora through their most frequent terms (after stopword removal). The frequent words align with domain expectations: ecological texts emphasize environmental concepts such as water, carbon, and land, while social texts highlight economic and institutional themes such as development, industry, and market. This suggests that the corpora tend to represent their respective domains and capture the general thematic focus of each field.

Rank	Ecological	Freq.	Social	Freq.
1	henan	6K	development	4.3K
2	province	5.6K	henan	3.3K
3	water	4.3K	province	2.8K
4	development	3.8K	rural	2.7K
5	land	2.8K	tourism	2.1K
6	area	2.6K	economic	2K
7	spatial	2.4K	social	1.7K
8	carbon	1.8K	management	1.4K
9	agricultural	1.7K	farmer	1.4K
10	resources	1.7K	value	1.4K

Table 2: Top 10 Frequent Terms in the Ecological and Social Corpora (after stopword removal). Frequencies are shown in thousands (K).

7 Results

Tables 4, 5, and 3 summarize the results for context relevance, faithfulness, and answer relevance, respectively. First of all, the mean values show that, regardless of the RAGs' architectural components, the results for answer relevance are significantly larger than the context relevance and faithfulness. This indicates that the generated answers are more aligned with the questions that the selected context.

Additionally, across all model configurations, faithfulness scores are consistently higher than context relevance scores, indicating a systematic divergence between the semantic alignment of retrieved contexts with the question and the alignment of generated answers with the provided context. This pattern implies that despite limitations in the semantic relevance of the retrieved context to the question, the models are nevertheless able to generate answers that are strongly grounded in the retrieved content.

Figure 2 illustrates the Bayesian analysis of the difference between the ecological and social mean values. Positive values are indicative of a bias toward ecology, and negative results indicate a bias toward the social domain. In the retrieval-bias test, as shown on the right-side of Figure 2, a clear tendency is observed toward ecological variables with regard to the context-relevance metric, while no significant indication of bias is present for answer-relevance and faithfulness.

In particular, for faithfulness, ecological scores show a minor lead over social ones, with mean differences (Δ) ranging from 0.007 ± 0.043 to 0.030 ± 0.059 across models. This corresponds to small effect sizes (Cohen's d between 0.054 and 0.163) and $P(\Delta > 0)$ values between 0.57 and

Experiment	Model (Gen + Emb)	$\mu_{\text{social}} \pm sd$	$\mu_{\text{eco}} \pm sd$	$\Delta \pm sd$	Cohen's d	$P(\Delta > 0)$
Retrieval-Bias	DeepSeek-R1 + Nomic	0.401 ± 0.034	0.507 ± 0.037	0.105 ± 0.051	0.661	≈ 0.98
	DeepSeek-R1 + Gemma	0.457 ± 0.033	0.592 ± 0.035	0.135 ± 0.049	0.884	1.00
	Llama 3.2 + Nomic	0.402 ± 0.032	0.481 ± 0.034	0.079 ± 0.047	0.532	≈ 0.95
	Llama 3.2 + Gemma	0.436 ± 0.032	0.593 ± 0.034	0.157 ± 0.046	1.048	1.00
Prompt-Bias	DeepSeek-R1 + Nomic	0.361 ± 0.016	0.441 ± 0.016	0.080 ± 0.023	1.11	1.00
	DeepSeek-R1 + Gemma	0.189 ± 0.016	0.277 ± 0.017	0.088 ± 0.023	1.20	1.00
	Llama 3.2 + Nomic	0.361 ± 0.017	0.456 ± 0.017	0.095 ± 0.024	1.24	1.00
	Llama 3.2 + Gemma	0.190 ± 0.017	0.282 ± 0.017	0.093 ± 0.024	1.22	1.00

Table 3: Bayesian comparison of social vs. ecological variables for **Context Relevance**.

Experiment	Model (Gen + Emb)	$\mu_{\text{social}} \pm sd$	$\mu_{\text{eco}} \pm sd$	$\Delta \pm sd$	Cohen's d	$P(\Delta > 0)$
Retrieval-Bias	DeepSeek-R1 + Nomic	0.773 ± 0.024	0.781 ± 0.025	0.008 ± 0.034	0.079	≈ 0.60
	DeepSeek-R1 + Gemma	0.770 ± 0.029	0.777 ± 0.031	0.007 ± 0.043	0.054	≈ 0.57
	Llama 3.2 + Nomic	0.518 ± 0.040	0.547 ± 0.043	0.030 ± 0.059	0.163	≈ 0.70
	Llama 3.2 + Gemma	0.586 ± 0.034	0.605 ± 0.035	0.019 ± 0.049	0.119	≈ 0.66
Prompt-Bias	DeepSeek-R1 + Nomic	0.846 ± 0.010	0.764 ± 0.010	-0.082 ± 0.014	-1.8	< 0.001
	DeepSeek-R1 + Gemma	0.831 ± 0.016	0.744 ± 0.016	-0.086 ± 0.023	-1.2	< 0.001
	Llama 3.2 + Nomic	0.631 ± 0.028	0.525 ± 0.028	-0.105 ± 0.040	-0.85	≈ 0.01
	Llama 3.2 + Gemma	0.560 ± 0.030	0.500 ± 0.030	-0.060 ± 0.043	-0.44	≈ 0.08

Table 4: Bayesian comparison of social vs. ecological variables for **Faithfulness** .

0.70, providing only weak evidence of bias. The strongest effect appears in context relevancy, where the gap between ecological and social scores is much wider, with Δ ranging from 0.079 ± 0.047 to 0.157 ± 0.046 (Cohen's d between 0.532 and 1.048). In this metric, $P(\Delta > 0)$ ranges from ≈ 0.95 to 1.00, indicating strong evidence. Overall, the results from the retrieval bias test show a systematic ecological bias, which is most pronounced in context relevance.

In the prompt-bias test, the results presented on the right-side of Figure 2 show a divergent pattern between metrics. For context relevancy, ecological variables consistently outperform social ones, with posterior mean differences (Δ) ranging from 0.080 ± 0.023 to 0.095 ± 0.024 and large effect sizes (Cohen's d between 1.11 and 1.24). In all models, $P(\Delta > 0) = 1.00$, providing strong and practically meaningful evidence of a robust ecological advantage. Conversely, for Faithfulness, the direction of the bias reverses, with social scores significantly exceeding ecological ones. Mean differences in this metric are negative, ranging from -0.105 ± 0.040 to -0.060 ± 0.043 (Cohen's d between -0.44 and -1.8), while $P(\Delta > 0)$ drops to very low levels (ranging from < 0.001 to ≈ 0.08). Overall, the prompt-bias test confirms a systematic ecological bias in context relevance, while simultaneously revealing a significant social bias in faithfulness.

8 Discussion

Our findings demonstrate that even within a robust RAG architecture, a systematic and domain-dependent bias is present in socio-ecological applications. Bayesian analysis of Context relevance metric reveals a pervasive ecological bias across all configurations; regardless of the specific model or experiment type, the probability of ecological scores exceeding social scores ($P(\Delta > 0)$) remained between 0.95 and 1.00, with large effect sizes (d up to 1.24). This indicates that the retrieval stage consistently prioritizes ecological data over social variables. However, a critical divergence emerges in the Faithfulness metric. While the ecological bias is negligible and statistically uncertain during corpus variation (Retrieval-Bias, $P \approx 0.57$–0.70), it undergoes a total reversal in prompt-bias experiments. In these cases, the models systematically prioritize social over ecological information ($P < 0.08$).

This pattern suggests that domain bias within RAG systems is not a monolithic phenomenon but rather a multi-stage process; specifically, the ecological dominance observed at the retrieval stage is actively challenged by a social dominance during the generation stage. From a functional perspective, these results necessitate the implementation of a multi-stage bias control protocol. Consequently, mitigating domain bias cannot be confined to a single architectural stage. Instead, it must be managed

Experiment	Model (Gen + Emb)	$\mu_{\text{social}} \pm sd$	$\mu_{\text{eco}} \pm sd$	$\Delta \pm sd$	Cohen's d	$P(\text{Eco} > \text{Soc})$
Retrieval-Bias	DeepSeek-R1 + Nomic	0.878 ± 0.008	0.894 ± 0.008	0.016 ± 0.011	0.428	0.920
	DeepSeek-R1 + Gemma	0.897 ± 0.001	0.898 ± 0.001	0.001 ± 0.002	0.157	0.69
	Llama 3.2 + Nomic	0.842 ± 0.025	0.818 ± 0.026	-0.024 ± 0.037	-0.196	0.258
	Llama 3.2 + Gemma	0.815 ± 0.026	0.818 ± 0.026	0.017 ± 0.050	0.110	0.635
Prompt-Bias	DeepSeek-R1 + Nomic	0.896 ± 0.001	0.899 ± 0.002	0.003 ± 0.002	0.427	0.906
	DeepSeek-R1 + Gemma	0.896 ± 0.001	0.9 ± 0.001	0.002 ± 0.001	-0.45	0.923
	Llama 3.2 + Nomic	0.768 ± 0.031	0.792 ± 0.032	0.024 ± 0.045	0.170	0.703
	Llama 3.2 + Gemma	0.821 ± 0.029	0.83 ± 0.030	-0.013 ± 0.057	0.063	0.590

Table 5: Bayesian comparison of social vs. ecological variables for **Answer relevance** .

through discrete interventions: first, at the retrieval level (by refining embedding algorithms and semantic search parameters), and subsequently, at the post-generation stage (via prompt engineering and output monitoring). Such a bifurcated approach is essential to maintaining interdisciplinary equilibrium in socio-ecological analyses and beyond

In light of the emphasis placed by Vizniuk et al. (2025) on trustworthiness, explainability, and bias as unresolved challenges of RAG in agricultural decision-making contexts, the necessity of systematically examining bias within such systems becomes increasingly evident. While RAG architectures are often introduced as a means to mitigate knowledge limitations and hallucinations in large language models, they do not inherently guarantee neutrality across domains. Prior research in socio-ecological studies has already highlighted the presence of structural biases, particularly the tendency to privilege certain dimensions over others, a concern explicitly raised by Partelow (2018) regarding imbalances between social and ecological perspectives. Consequently, this study aligns with and extends these concerns by warning that bias in RAG-based analytical pipelines may manifest in multiple forms and stages, potentially influencing how evidence is retrieved, weighted, and articulated. Such biases, if left unexamined, risk shaping interpretations and recommendations in subtle yet systematic ways, underscoring the importance of bias-aware evaluation when RAG systems are employed to support analytical and policy-relevant processes in agriculture (McGinnis and Ostrom, 2014).

9 Conclusion

This study proposes an evaluation protocol for assessing the interdisciplinary performance of Retrieval Augmented Generation (RAG). The experiments are focused on the socio-ecological case of agricultural resilience in Henan Province. Our framework relies on systematically curated corpora of ecological and social science articles, along with carefully designed question templates. Across multiple experimental configurations (comprising combinations of DeepSeek-R1, Llama 3.2, Nomic-Embed, and EmbeddingGemma), we find compelling evidence of significant retrieval and prompt biases toward ecology in context relevance. Conversely, the models show a significant social bias and only a minor retrieval bias under faithfulness. Nevertheless, no significant bias is observed in models' answer relevance.

These results underscore both the opportunities and the risks of deploying LLMs in interdisciplinary field of socio-ecology. On the one hand, RAG provides scalable, evidence-grounded reasoning; on the other, unexamined biases–whether stemming from retrieval mechanics or prompt sensitivity–can skew analyses and mislead policy making.

By combining manual data curation with a Multi-Agent LLM-as-a-judge automated evaluation, this work makes domain bias empirically visible and contributes a reproducible framework for trustworthy, bias-aware NLP. Although demonstrated in socio-ecology, the approach generalizes to other interdisciplinary domains.

10 Future Work

Future research can advance this line of research along three directions. First, evaluation protocols can be extended to explicitly measure interdisciplinary balance, with sensitivity to the relative weighting of social and ecological variables. Second, bias-mitigation strategies for RAG should be investigated, such as diversifying retrieved sources, re-weighting domain-specific content, and incorporating human-in-the-loop oversight. Third, applying this framework to other interdisciplinary fields,

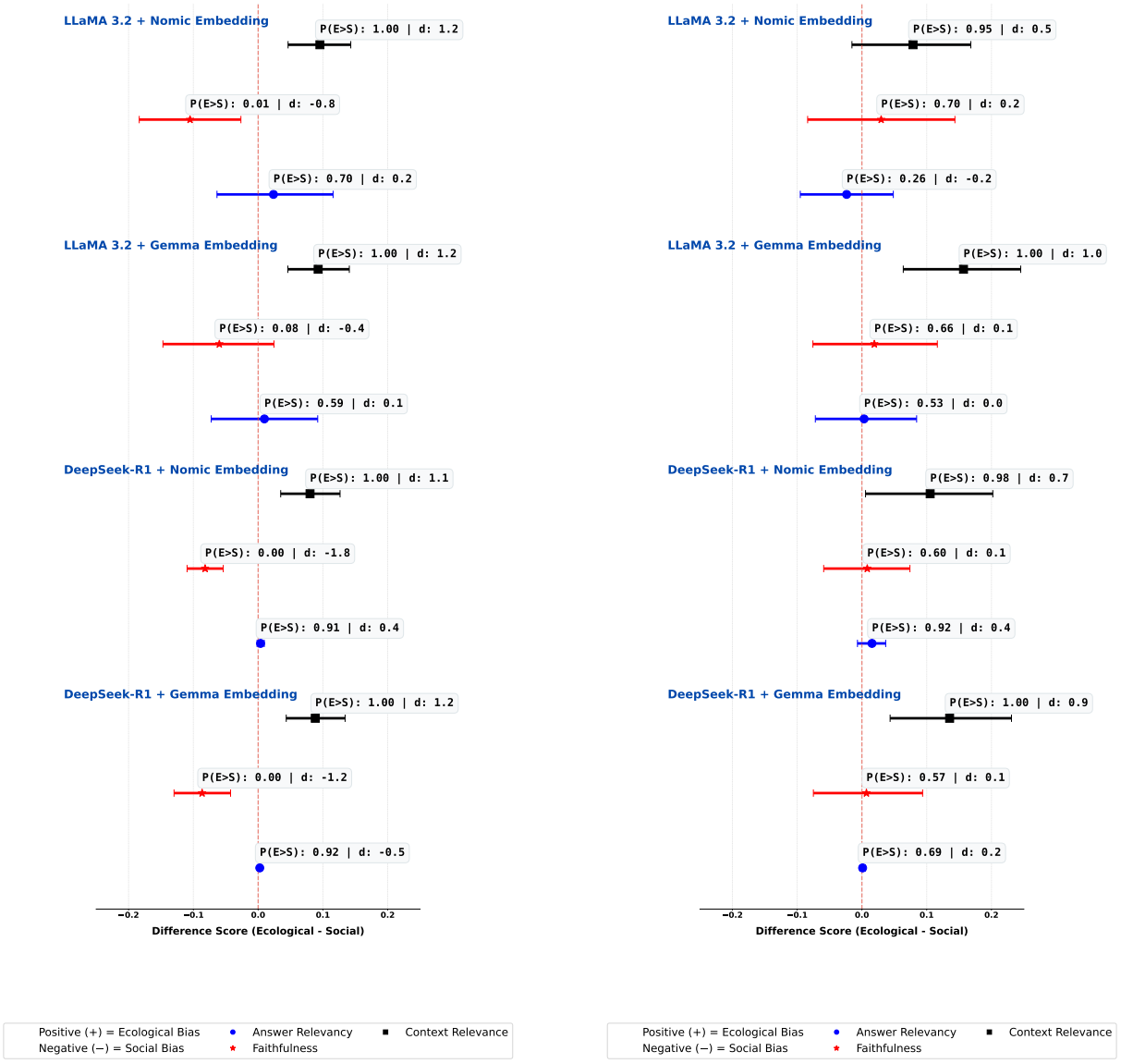

Figure 2: Bayesian analysis of social vs. ecological variables. (Left) Prompt Bias, (Right) Retrieval Bias.

such as public health, climate adaptation, and urban sustainability, would provide broader validation and strengthen the robustness of the methodology.

Limitations

Several factors constrain the scope and interpretation of our findings. The analysis is restricted to a single geographic region (Henan Province, China), and the observed bias patterns may differ in other socio-ecological contexts.

References

Amin Abolghasemi, Leif Azzopardi, Seyyed Hadi Hashemi, Maarten de Rijke, and Suzan Verberne. 2025. Evaluation of attribution bias in generator-aware retrieval-augmented large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21105–21124, Vienna, Austria. Association for Computational Linguistics.

Hongliu Cao. 2025. Writing style matters: An examination of bias and fairness in information retrieval systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 336–344.

Lingjiao Chen, Matei Zaharia, and James Zou. 2024. How Is ChatGPT's Behavior Changing Over Time? *Harvard Data Science Review*, 6(2).

Ming Cheng, Jiaying Gong, and Hoda Eldardiry. 2025. Sci-LoRA: Mixture of scientific LoRAs for cross-domain lay paraphrasing. In *Findings of the Association for Computational Linguistics: ACL 2025*,

pages 18524–18541, Vienna, Austria. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Ariana Genovese, Srinivasagam Prabha, Sahar Borna, Cesar A Gomez-Cabello, Syed Ali Haider, Maissa Trabilsy, Cui Tao, and Antonio Jorge Forte. 2025. From data to decisions: Leveraging retrieval-augmented generation to balance citation bias in burn management literature. *European Burn Journal*, 6(2):28.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *CoRR*, abs/2411.15594.

Daya Guo, Dong Yang, Hongyi Zhang, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633–638.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. 2024. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1416–1428, Bangkok, Thailand. Association for Computational Linguistics.

Yuelyu Ji, Hang Zhang, and Yanshan Wang. 2025. Bias evaluation and mitigation in retrieval-augmented medical question-answering systems. *arXiv preprint arXiv:2503.15454*.

Matyas Juhasz, Kalyan Dutia, Henry Franks, Conor Delahunty, Patrick Fawbert Mills, and Harrison Pim. 2024. Responsible retrieval augmented generation for climate decision making from documents. *arXiv preprint arXiv:2410.23902*.

Rachel Kelly, Mary Mackay, Kirsty L Nash, Christopher Cvitanovic, Edward H Allison, Derek Armitage, Aletta Bonn, Steven J Cooke, Stewart Frusher, Elizabeth A Fulton, and 1 others. 2019. Ten tips for developing interdisciplinary socio-ecological researchers. *Socio-Ecological Practice Research*, 1(2):149–161.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *IEEE Transactions on Audio, Speech and Language Processing*, pages 1–11.

Christos Mammides and Harris Papadopoulos. 2024. The role of large language models in interdisciplinary research: Opportunities, challenges and ways forward. *Methods in Ecology and Evolution*, 15(10):1774–1776.

Laura Maxim, Joachim H. Spangenberg, and Martin O'Connor. 2009. An analysis of risks for biodiversity under the dpsir framework. *Ecological Economics*, 69(1):12–23. The DPSIR framework for Biodiversity Assessment.

Michael D McGinnis and Elinor Ostrom. 2014. Social-ecological system framework: initial changes and continuing challenges. *Ecology and society*, 19(2).

Stefan Partelow. 2018. A review of the social-ecological systems framework. *Ecology and Society*, 23(4).

Catlin M Pauley, Aaron J McKim, and Jennifer Hodbod. 2019. A social-ecological resilience perspective for the social sciences of agriculture, food, and natural resources. *Journal of Agricultural Education*, 60(4):132–148.

Jun Rao, Zepeng Lin, Xuebo Liu, Xiaopeng Ke, Lian Lian, Dong Jin, Shengjun Cheng, Jun Yu, and Min Zhang. 2025. APT: Improving specialist LLM performance with weakness case acquisition and iterative preference training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20958–20980, Vienna, Austria. Association for Computational Linguistics.

Anastasia Razdaibiedina, Ashish Khetan, Zohar Karnin, Daniel Khashabi, and Vivek Madan. 2023. Representation projection invariance mitigates representation collapse. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14638–14664, Singapore. Association for Computational Linguistics.

Samuel Schmidgall, Charles Harris, Ini Essien, and 1 others. 2024. Evaluation and mitigation of cognitive biases in medical language models. *npj Digital Medicine*, 7(1):295.

Saptarshi Sengupta, Wenpeng Yin, Preslav Nakov, Shreya Ghosh, and Suhang Wang. 2025. Exploring language model generalization in low-resource extractive QA. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7106–7126, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Artem Vizniuk, Grygorii Diachenko, Ivan Laktionov, Agnieszka Siwocha, Min Xiao, and Jacek Smoląg. 2025. A comprehensive survey of retrieval-augmented large language models for decision making in agriculture: Unsolved problems and research opportunities. *Journal of Artificial Intelligence and Soft Computing Research*, 15(2):115–146.

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

Yipeng Wu, Ming Xu, and Shuming Liu. 2024. Generative artificial intelligence: A new engine for advancing environmental science and engineering. *Environmental Science & Technology*, 58(40):17524–17528. PMID: 39342507.

Benfeng Xu, Chunxu Zhao, Wenbin Jiang, PengFei Zhu, Songtai Dai, Chao Pang, Zhuo Sun, Shuohuan Wang, and Yu Sun. 2023. Retrieval-augmented domain adaptation of language models. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 54–64, Toronto, Canada. Association for Computational Linguistics.

Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2025. Evaluation of Retrieval-Augmented Generation: A Survey. In *Big Data*, pages 102–120, Singapore. Springer Nature Singapore.

Pengwei Zhan, Zhen Xu, Qian Tan, Jie Song, and Ru Xie. 2024. Unveiling the lexical sensitivity of llms: Combinatorial optimization for prompt enhancement. *arXiv preprint arXiv:2405.20701*.

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

Yuanxin Zhang, Sijie Lin, Yaxin Xiong, Nan Li, Lijin Zhong, Longzhen Ding, and Qing Hu. 2025b. Fine-tuning large language models for interdisciplinary environmental challenges. *Environmental Science and Ecotechnology*, 27:100608.

Jun-Jie Zhu, Meiqi Yang, Jinyue Jiang, Yiming Bai, Danqi Chen, and Zhiyong Jason Ren. 2024. Enabling GPTs for Expert-Level Environmental Engineering Question Answering. *Environmental Science & Technology Letters*, 11(12):1327–1333. Publisher: American Chemical Society.