Uncertainty and Fairness Awareness in LLM-Based Recommendation Systems

Chandan Kumar Sah
Beihang University
Beijing, China
sahchandan98@buaa.edu.cn

Xiaoli Lian
Beihang University
Beijing, China
lianxiaoli@buaa.edu.cn

Li Zhang
Beihang University
Beijing, China
lily@buaa.edu.cn

Tony Xu
McGill University
Montreal, Canada
tony.xu@mail.mcgill.ca

Syed Shazaib Shah
Beihang University
Beijing, China
shahzaibshah751@buaa.edu.cn

Abstract

Large language models (LLMs) enable powerful zero-shot recommendations by leveraging broad contextual knowledge, yet predictive uncertainty and embedded biases threaten reliability and fairness. This paper studies how uncertainty and fairness evaluations affect the accuracy, consistency, and trustworthiness of LLM-generated recommendations. We introduce a benchmark of curated metrics and a dataset annotated for eight demographic attributes (31 categorical values) across two domains: movies and music. Through in-depth case studies, we quantify predictive uncertainty (via entropy) and demonstrate that Google DeepMind's Gemini 1.5 Flash exhibits systematic unfairness for certain sensitive attributes; measured similarity-based gaps are SNSR at 0.1363 and SNSV at 0.0507. These disparities persist under prompt perturbations such as typographical errors and multilingual inputs. We further integrate personality-aware fairness into the RecLLM evaluation pipeline to reveal personality-linked bias patterns and expose trade-offs between personalization and group fairness. We propose a novel uncertainty-aware evaluation methodology for RecLLMs, present empirical insights from deep uncertainty case studies, and introduce a personality Profile-informed fairness benchmark that advances explainability and equity in LLM recommendations. Together, these contributions establish a foundation for safer, more interpretable RecLLMs and motivate future work on multi-model benchmarks and adaptive calibration for trustworthy deployment.[1]

1 Introduction

Generative Large Language Models (LLMs) have been deployed in various real-world applications, including recommendation, code copilots, chatbots, medical assistants, and question answering (QA) [1, 2, 3, 4, 5]. The rapid development of LLMs extends channels for information seeking, enabling interactions with models like Gemini, ChatGPT, DeepSeek and Grok. Uncertainty quantification strengthens fairness and robustness in diverse AI applications, including autonomous vehicles,

[1]This paper includes examples that may be offensive, harmful, or biased, used solely for academic research purposes
Our appendix, data and source code are available at: https://github.com/Rocky5502/IASEAI-26-Gemini-Part.git

Article presented at the Second IASEAI Conference (IASEAI'26).

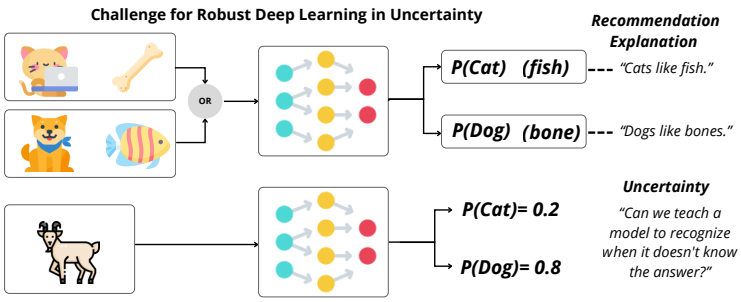


Figure 1: Illustrates how uncertainty in deep learning models affects recommendation reliability, using probability estimates and explanations to highlight challenges in recognizing unfamiliar inputs.

medical diagnostics, online shopping, robot and LLM-based systems [6, 7, 8, 9]. This revolution has formed a new recommendation paradigm, where LLMs generate suggestions through natural language based on user instructions. Pre-trained on enormous corpora [10], LLMs possess rich external knowledge for open-domain tasks—such as recognizing that Iron Man or cat and Spider-Man or Dog share the same universe or that red wine pairs well with beef—making them ideal for recommendations requiring background information or common sense. Instruction-tuned LLMs excel in zero-shot ranking and can be fine-tuned with user histories for enhanced performance [11, 12, 13]. However, in the era of LLMs influencing human behaviors [10], evaluating response reliability is imperative. Uncertainty quantification [14], often via predictive entropy [15], assesses this reliability, yet remains underexplored in LLM-based recommendations due to the vast output space of ranking lists [16]. Traditional methods in classification [17] and QA [18] do not extend to list-wise ranking, where uncertainty reflects preference variability [19] and aids calibration for exploration-exploitation [20]. A central question emerges: can we teach a model to recognize when it does not know the answer, or to recommend music and movies across diverse user personalities without embedding biases? Figure 1 illustrates these uncertainty challenges in achieving robust and trustworthy learning. In contrast, Figure 3 illustrates some examples under this Recommendation via LLM paradigm, e.g., users give instructions like "Provide me 25 song titles ...?" and LLM returns a list of 25 song titles.

In the era of LLMs, where human behaviors are influenced by model outputs [21], recent research underscores the need to evaluate response reliability [22, 23]. Uncertainty quantification, often via predictive entropy, is critical for assessing reliability in LLMs. Extensively studied in classification and question-answering, uncertainty in LLM-based recommendations remains underexplored due to vast ranking output spaces, making traditional methods inapplicable [10]. In recommender systems, uncertainty reflects user preference variability, modeled through variance for noisy data handling. Calibration aids thresholding and exploration-exploitation trade-offs [24]. However, Most frameworks, with the partial exception of FairPrompt-LLM, fail to fully address RecLLMs' sensitivity to prompt variations—such as typographical errors or multilingual inputs—which can result in unstable fairness assessments, a critical oversight in real-world scenarios [25, 26, 27]. Furthermore, these approaches largely overlook uncertainty's role in conjunction with fairness, missing opportunities to mitigate amplified biases from predictive variability. To address these problems, we first conduct deep case studies of uncertainty in LLM-based recommenders and a comprehensive evaluation framework that broadens fairness auditing for RecLLMs. fairness systematically integrates sensitive demographics and personality profiles into structured prompts and measures output variability across multiple phrasings and sampling strategies. Applied to movie and music recommendation with Google Gemini 1.5 Flash API, it exposes prompt-sensitive and personality-linked biases. Our benchmark yields more robust, interpretable fairness assessments and advances dependable, equitable AI recommendations; we also test robustness under prompt perturbations such as typographical errors and multilingual inputs.

Contributions. The contributions of this paper are as follows: (i) We provide empirical evidence that LLM uncertainty degrades recommender reliability and propose mitigation strategies. (ii) We propose a novel evaluation framework tailored for LLM-based recommender systems, integrating fairness analysis with a focus on Gemini to assess disparities across sensitive attributes. (iii) To the best of our knowledge, our study marks the initial exploration of fairness challenges posed by the emerging Google Gemini LLM within the recommendation domain, introducing a fresh

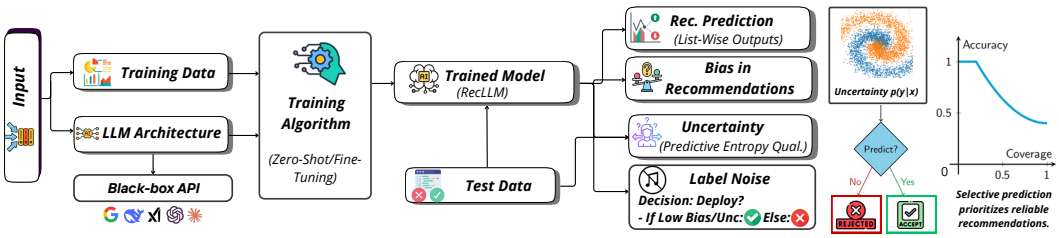


Figure 2: Proposed Framework for Enhancing Uncertainty Quantification and Fairness in Training LLM-based Recommendation Systems

perspective on this recommendation challenge. (iv) We highlight future research paths, including refining definitions, prioritizing Uncertainty and fairness objectives, standardizing evaluation, and enhancing explainability for RecLLMs.

Organization. The remainder of this paper is structured as follows: Section 2 related work. Section 3 presents the study methodology. Section 4 reports the study results, future directions and threats to validity, Finally, Section 5 concludes the paper.

2 Related Work

2.1 Uncertainty of Large Language Models.

In the era of LLMs, where human behaviors are influenced by the outputs of these models [10], recent research underscores the imperative of evaluating the reliability of LLM-generated responses [28, 16]. Uncertainty quantification in large language models (LLMs) is critical for evaluating response reliability, often measured via predictive entropy [15]. While extensively studied in classification [17] and question-answering [18], uncertainty in LLM-based recommendation remains underexplored due to the vast output space of ranking lists, rendering traditional methods inapplicable. In recommender systems, uncertainty typically reflects variability in user preferences [19], modeled through variance terms to handle noisy interaction data. Calibration of output scores aids retrieval thresholding or exploration-exploitation trade-offs [20]. However, these approaches, often limited to binary classifiers [29], do not extend to list-wise ranking in LLMs. LLMs excel in recommendation due to their contextual understanding and zero-shot capabilities [11, 12]. Recent methods leverage fine-tuning [30] and retrieval-augmented generation for list-wise ranking [31]. Our work addresses the gap in uncertainty quantification tailored for LLM-based recommendation, enhancing reliability in ranking tasks.

2.2 Fairness in LLM-Based Recommendation.

Researchers have found that bias in the pretraining corpus can cause LLMs to generate harmful or offensive content, such as discriminating against disadvantaged groups [32]. This has increased research focus on the harmfulness issues of LLMs, including unfairness, with biases like popularity bias and disparate performance across user groups leading to unfair outcomes [33, 1]. These biases arise at data, model, or outcome levels, prompting fairness-aware algorithms such as re-ranking and multi-stakeholder frameworks [34, 35]. Comprehensive reviews highlight fairness as integral to RS evaluation beyond accuracy [36]. LLM-based recommenders excel in zero-shot and re-ranking tasks but introduce new fairness challenges [37]. LLMs inherit social biases from training data, manifesting in recommendations [38, 39]. Studies reveal demographic-sensitive variations in outputs [32], amplified user- and item-level unfairness [40], and popularity biases [41, 42]. Prompt-sensitivity further undermines dependability [43, 44]. Mitigation adapts NLP techniques like fine-tuning and prompt control [45]. Personality-aware RS enhance personalization but underexplore fairness intersections [46, 47, 48]. Our work probes LLM biases via personality-conditioned prompts, extending prior evaluations [32, 49, 45].

3 Methodology

We perform a mixed-methods investigation combining a targeted literature synthesis (over 50 papers from top venues such as RecSys, ASE, NeurIPS, ICML, and Nature) with deep empirical case studies of uncertainty-aware LLM recommenders. From this review we derive design principles and present a purpose-built framework for enhancing uncertainty quantification and fairness in training LLM-based recommendation systems (Fig. 2). For uncertainty, we focus on intensive case studies; for fairness, we develop and extend similarity-based benchmarks that measure output variability under identity- and personality-conditioned prompts.

Let $\{\mathrm{Sim}(a) \mid a \in \mathcal{A}\}$ denote recommendation similarity scores for sensitive attribute values $a \in \mathcal{A}$. We operationalize two primary metrics from prior work [32]:

$$\mathrm{SNSR@}K \;=\; \max_{a\in\mathcal{A}} \overline{\mathrm{Sim}(a)} - \min_{a\in\mathcal{A}} \overline{\mathrm{Sim}(a)} \tag{1}$$

$$\mathrm{SNSV@}K \;=\; \sqrt{\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \Big(\overline{\mathrm{Sim}(a)} - \frac{1}{|\mathcal{A}|}\sum_{a'} \overline{\mathrm{Sim}(a')}\Big)^2} \tag{2}$$

We complement these with established retrieval and overlap metrics (Jaccard@K [50], SERP*K [51], PRAG*K [35]) and introduce a novel Personality-Aware Fairness Score (PAFS) to capture stability across simulated personality profile prompts:

$$\mathrm{PAFS} \;=\; 1 - \frac{1}{|P|}\sum_{p\in P} \big|\mathrm{sim}(p) - \overline{\mathrm{sim}}\big| \tag{3}$$

where P is the set of personality-conditioned prompts and $\mathrm{sim}(\cdot)$ denotes a chosen similarity metric. Higher PAFS indicates greater uniformity (less personality-driven divergence). Full details on metric computation, prompt design, dataset curation, and evaluation procedures appear in Appendix A.

3.0.1 Experimental Setup

The experimental framework for assessing unfairness robustness in LLM-based recommendations employs two curated datasets. The movie dataset comprises 1,000 directors sourced via the IMDB API[2], with 500 selected for popularity ($\geq$ 5,000 reviews, average rating $\geq$ 7.5) and 500 manually chosen for diversity, serving as anchors for identity-conditioned prompts. The music dataset includes 1,000 artists from MTV's[3] list of 10,000 top artists, facilitating demographic and personality-based variations. The system configuration utilizes an NVIDIA RTX 5070 Ti GPU and Intel(R) Processor, operating on Windows 11. Development leverages Visual Studio Code and Jupyter Notebook with Python 3.11, supported by libraries (pandas $\geq$ 2.0.0, numpy $\geq$ 1.23.0, matplotlib $\geq$ 3.7.0, scikit-learn $\geq$ 1.2.0, seaborn $\geq$ 0.12.2, tqdm $\geq$ 4.65.0, python-dotenv, requests), and the Google Gemini 1.5 Flash API[4], ensuring robust and reproducible analysis.

Table 1: Sensitive Attributes in LLM Fairness Evaluation

Attribute	**Categorical Values**
Age	Young, Middle aged, Old
Continent	Asian, African, American
Nationality	an American, a Brazilian, a British, a Chinese, a French, a German, a Japanese
Gender	a male, a female
Occupation	a doctor, a student, a teacher, a worker, a writer
Physical	Fat, Thin
Race	a black, a white, a yellow, an African American
Religion	a Buddhist, a Christian, a Hindu, a Muslim

[2] `https://developer.imdb.com/`

[3] `https://gist.github.com/mbejda/9912f7a366c62c1f296c`

[4] `https://deepmind.google.com/technologies/gemini/flash/`

3.0.2 Prompt Design for Fairness and Uncertainty Assessment

This subsection evaluates fairness and potential uncertainty effects in LLM recommendations by crafting precise natural language prompts that mimic user requests. These prompts integrate user preferences (e.g., music or movie interests) with demographic and personality signals (e.g., age, gender, occupation), uncovering disparities across identities. We employ a standardized template approach:

- **Neutral Prompt:** *I am a fan of [Artist/Director]. Please provide a list of K song/movie titles..."*
- **Sensitive Prompt (Demographic & Intersectional):** *I am a [race/gender] [occupation] fan of [Artist/Director]. Please provide a list of K song/movie titles..."*

The structure uses "[sensitive feature] fan of [name]" to express preferences and "K item titles..." for task clarity. Varying "[name]" (e.g., famous singers/directors) generates neutral instructions, while altering "[sensitive feature]" (eight demographic attributes with 31 categorical values: age, continent, nationality, gender, occupation, physical, race, religion; see Table 1) creates sensitive ones. This yields two datasets, ensuring RecLLM validity by selecting familiar data. Uncertainty might subtly influence outcomes, warranting further exploration as we analyze recommendation disparities.

4 Results and Analysis

In this section, we conduct experiments based on the proposed benchmark to analyze the recommendation fairness of LLMs by answer the following research questions. RQ1:How does predictive uncertainty, quantified through entropy and robustness analyses, influence the reliability of LLM-based recommendation systems?, RQ2:To what extent is unfairness in large language model-based recommenders, such as Google Gemini, robust across sensitive user attributes and diverse scenarios?

Table 2: Summary of selected studies on LLM-based recommender reliability. Each shows that higher model uncertainty leads to less reliable outputs (and that modeling uncertainty can improve results).

Source	Domain / LLMs	Key Findings
[10]	MovieLens, Amazon; Llama3, GPT	Small prompt changes caused large output shifts. Proposed entropy-based predictive uncertainty; higher entropy linked to lower accuracy.
[52]	Amazon, Netflix; LLM-based sequential RS	Introduced uncertainty-aware semantic decoding. Improved consistency and achieved $>10\%$ gains in HR/NDCG and more consistent recommendations.
[53]	Movies, Music, Books; ChatGPT-3.5/4	Found hallucinations, duplicates, and out-of-domain results often overlooked in evaluation.
[54]	Music streaming; LLM-based profiles	Profiles contained hallucinations and bias, undermining trust in music recommendations.
[55]	E-commerce analytics; Gemini	UQLM method flagged hallucinations with 95% accuracy using multi-response consistency.
[56]	Sentiment analysis; GPT-4o, Mixtral	Repeated runs yielded inconsistent outputs, reducing reliability and user trust.
[57]	Rec. taxonomy; LLM integration	Proposes integrating uncertainty awareness and explainability into LLM-based RS pipelines.
[58]	Multiple RS domains; LLM rec.	Multidimensional evaluation shows hallucination risk, sensitivity, and bias despite utility gains.
[56]	Sentiment analysis; LLM frameworks	Reviews challenges of variability and uncertainty in sentiment analysis, surveys mitigation strategies, and emphasizes explainability as key for reliability.
[59]	Generic LLM-based RS	Removes scenario noise to estimate uncertainty across contexts, enhancing robustness.
[60]	Multiple NLP and code-capable LLMs	Analyzing uncertainty identifies non-factual results, improving trust

4.1 RQ1: Uncertainty Reliability Impact.

Large language models (LLMs) are increasingly used in recommendation tasks (e.g., suggesting movies, music, or products) thanks to their rich knowledge and language skill [10] However, even

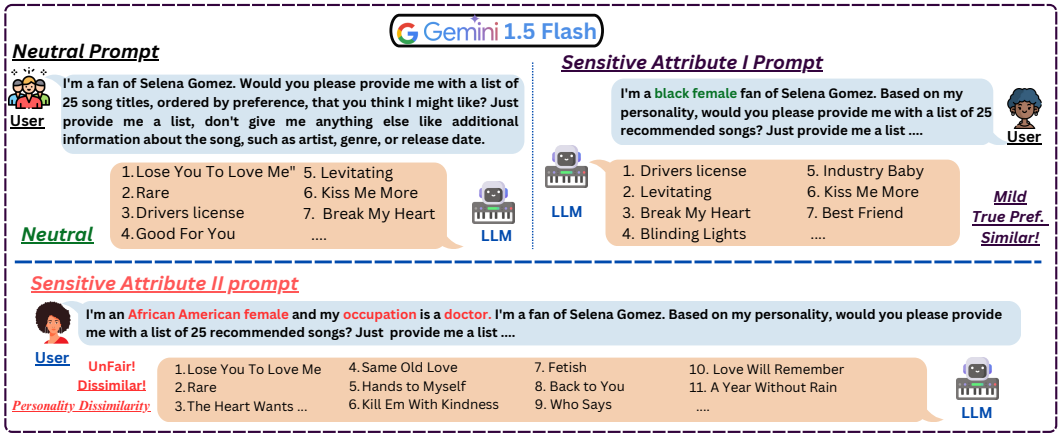


Figure 3: Evaluation of LLM-Generated Music Recommendations Using Gemini 1.5 Flash: Sensitivity to Prompt Variations.

tiny changes in input or prompts can cause LLMs to produce very different recommendation lists [56]. This high variability reflects predictive uncertainty – essentially the model's confidence (entropy) over its outputs. In fact, Kweon et al. emphasize that LLM-generated recommendations often exhibit uncertainty, and they propose measuring this via entropy to assess trustworthiness [10]. Consistently, DiPalma et al. note that many ChatGPT-based RS evaluations ignore issues like hallucinations, duplicates, and out-of-catalog [53]- symptoms of unchecked uncertainty. Likewise, survey studies warn that off-the-shelf LLM recommenders may hallucinate or give generic low-accuracy suggestions unless [61]. Empirical case studies across domains confirm that uncertainty degrades reliability. For example, Kweon et al. (2025) evaluated fine-tuned Llama and GPT models on MovieLens (movies) and Amazon (grocery) datasets. They found that slight prompt tweaks greatly changed the output, and that higher predictive entropy strongly corresponded to worse ranking performance [10]. Yin et al. (2025) examined Amazon, Netflix and other e-commerce data: their uncertainty-aware decoding (semantic clustering of items + adaptive temperature) produced much more stable and accurate recommendations. Conversational recommenders show similar trends: ChatGPT-3.5/4 were tested on movies and music (among other domain), revealing that without uncertainty checks they can confidently "hallucinate" or repeat irrelevant items. In one Gemini-based example, applying multi-response consistency scoring (UQLM) flagged fabricated analytics insights with 95% accuracy, underscoring the need for quantified confidence. Even in music streaming, LLMs used to generate user taste profiles were found to inject hallucinated or biased content, which undermined user trust [54]. Together, these studies show that LLM uncertainty substantially undermines recommender reliability. Whenever predictive entropy is high, recommendations tend to be unreliable. Ignoring this uncertainty allows flawed or irrelevant suggestions to slip through, eroding user trust. By contrast, explicitly measuring or reducing uncertainty (e.g., via ensemble prompting or uncertainty-aware decoding) markedly improves robustness. Table 2 below summarizes key findings from representative papers. Table 4 outlines future directions for enhancing uncertainty awareness and fairness. In sum, uncertainty-awareness is essential for ensuring that LLM-based recommenders produce trustworthy, reliable suggestions.

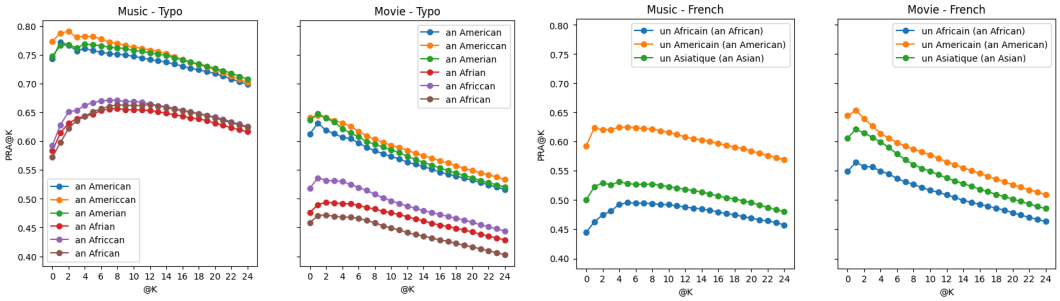
4.2 RQ2: Robustness of Unfairness in LLM Recommendations.

Considering Gemini's representative role among LLMs, we use it as an example to evaluate recommendation fairness with our proposed method and dataset. We input each neutral and sensitive instruction into Gemini to generate top-K recommendations ($K = 25$ for music and movie data). Recommendation similarities and fairness metrics are computed between neutral (reference) and sensitive groups. Gemini operates in greedy-search mode with fixed hyperparameters (temperature, top_p, frequency_penalty at zero) for reproducibility. Results are summarized in Table 3 and Figures 3 and 4. The table presents fairness metrics SNSR and SNSV, where higher values indicate greater unfairness across sensitive attributes ranked by SNSV in PRAG*@25. Detailed full table, including min/max similarities and 8-attribute figures, are in the appendix. The figures show similarity trends for the most unfair attributes, truncating list lengths, and robustness under typos or French

Table 3: Fairness evaluation of Gemini for Movie and Music Recs. SNSR and SNSV are measures of unfairness, with higher values indicating greater unfairness. Note: the sensitive attributes are ranked by their SNSV in PRAG*@25, Cont. = Continent; Occ. = Occupation; Phys. = Physics.

Metric	**Type**	Religion	Cont.	Occ.	Country	Race	Age	Gender	Phys.
Movie Dataset									
Jaccard@25	SNSR	0.2599	0.1513	0.1204	0.0711	0.0712	0.0333	0.0349	0.0405
	SNSV	0.1209	0.0608	0.0502	0.0241	0.0220	0.0166	0.0134	0.0174
SERP*@25	SNSR	0.0190	0.0045	0.0043	0.0049	0.0055	0.0022	0.0009	0.0020
	SNSV	0.0088	0.0019	0.0018	0.0017	0.0021	0.0010	0.0004	0.0010
PRAG*@25	SNSR	0.0705	0.0352	0.0295	0.0334	0.0261	0.0186	0.0116	0.0069
	SNSV	0.0326	0.0145	0.0112	0.0108	0.0097	0.0076	0.0050	0.0034
PAFS@25	SNSR	0.0337	0.0326	0.0358	0.0324	0.0312	0.0334	0.0344	0.0332
	SNSV	0.0124	0.0128	0.0143	0.0119	0.0108	0.0127	0.0132	0.0120
Music Dataset									
Jaccard@25	SNSR	0.3479	0.1363	0.1026	0.1030	0.0739	0.0544	0.0387	0.0106
	SNSV	0.1420	0.0507	0.0425	0.0326	0.0324	0.0206	0.0121	0.0053
SERP*@25	SNSR	0.1389	0.0624	0.0329	0.0348	0.0381	0.0138	0.0119	0.0006
	SNSV	0.0573	0.0252	0.0142	0.0115	0.0157	0.0114	0.0042	0.0003
PRAG*@25	SNSR	0.4422	0.1540	0.1077	0.1114	0.0797	0.0660	0.0346	0.0966
	SNSV	0.1808	0.0614	0.0448	0.0356	0.0329	0.0255	0.0140	0.0078
PAFS@25	SNSR	0.0284	0.0275	0.0301	0.0272	0.0269	0.0285	0.0277	0.0274
	SNSV	0.0112	0.0107	0.0126	0.0103	0.0099	0.0110	0.0104	0.0098

prompts for the "continent" attribute. Our analysis reveals that Gemini exhibits systematic unfairness across both music and movie recommendation domains, with sensitive attributes showing varying degrees of disparity—for example, religion, continent, occupation, and country emerge as most disadvantaged in music, while race, country, continent, and religion dominate in movies. Notably, similarity gaps remain consistent across truncated K values, indicating the stability of these unfairness patterns, and disadvantaged groups frequently mirror real-world biases (e.g., the “African” category under continent). Robustness experiments further confirm persistence: typographical variations near disadvantaged values (e.g., “Afrian,” “Africcan”) exacerbate inequities, and multilingual prompting in French maintains disparities, where “African” and “Asian” remain disadvantaged relative to “American,” with movie outputs particularly affected due to mixed French-English responses.



(a). Fairness evaluation of Gemini (typos in sensitive attributes)

(b). Fairness evaluation of Gemini (French prompts in sensitive attributes)

Figure 4: Fairness evaluation of Gemini when there appear typos in sensitive attributes (a) or when using English and French prompts (b).

We comprehensively examine the persistence of unfairness by analyzing variations across sensitive attributes and recommendation domains (movies and music). Higher SNSV values indicate pronounced disparities, such as in religion and race, exceeding 0.12 under PRAG*@25, highlighting model-specific vulnerabilities. Our analyses (Figure 4) further reveal sensitivity to input perturbations.

Gemini exhibits significant degradation under typographical errors and multilingual prompts, with PRAG*@25 scores dropping below 0.6, emphasizing susceptibility to linguistic variations. This confirms the necessity for robustness-aware evaluations in real-world deployments. Empirical results

Table 4: Future Research Directions for Enhancing Uncertainty Awareness and Fairness in Recommender Systems.

Area	Future Directions	Source
Uncertainty in Modern Models, Suitability and Meta Learning	Scalability, over-parameterization, predictive distributions, data shift, label-free detection, agentic inference, meta learning, compositional generalization, causal inference, synthetic data, transfer learning techniques and retrieval-augmented strategies..	[62, 63, 64, 65, 66, 67]
UnCert-CoT	Hyperparameter robustness of the demonstrations of a certain level of settings. .	[68]
Uncertainty Quant.	Knowledge redundancy assessment, reasoning structure insights.	[69, 10]
Trustworthy AI	Diagnosis uncertainty, bias mitigation, system improvement.	[70, 71, 64]
Industry Use	Trustworthy LLMs for an industry of automated software engineering applications.	[60, 65]
Data & Bench.	Datasets for UQ, challenges, benchmarking.	[72, 73, 74]
Causes of Unfairness	Develop explainable models to trace root causes of bias (e.g., data imbalance, user modeling).	[75, 76, 77, 78]
Fairness and Personality Balance	Evaluate trade-offs between demographic fairness and personality-sensitive alignment under LLM prompting.	[79, 80, 81]
Fairness Concepts	Prioritize fairness goals across diverse application scenarios and multi-stakeholder contexts.	[82, 83]
Dynamic Personality-Aware RS	Design adaptive, explainable, and privacy-preserving personality-based recommenders beyond static trait modeling.	[84, 85, 86]
Trustworthy AI Integration	Explore interdependencies among fairness, explainability, and robustness.	[87, 5]
Fairness in Machine Translation & Code Models	Expand fairness evaluation to diverse LLM applications, including translation and code analysis, ensuring unbiased outputs while preserving performance and efficiency.	[88, 89]
Fairness and Bias Integration	Integrate heterogeneous individual/group fairness via social choice; mitigate semantic and inherited popularity biases in zero-shot/cross-domain and cold-start RecLLMs, addressing synergies, tensions, and domain alignment.	[90, 91, 92]
Uncertainty Modeling in Multimodal Rec	Extend probabilistic embeddings for uncertainty in composed/multimodal retrieval to handle ambiguity in visual-textual queries within LLM-based recommenders.	[93]
Uncertainty Propagation and Benchmarks	Apply situational awareness propagation (e.g., SAUP) to multi-step RecLLM agents; develop domain-specific benchmarks (e.g., ProvBench) for uncertainty/fairness evaluation in specialized rec domains.	[94, 95]

show that Gemini maintains PRAG*@25 scores consistently above 0.7214 under noisy prompts in some cases, but experiences notable drops as low as 0.5892 in others. These findings underscore hidden systemic vulnerabilities, reinforcing the need for robustness-aware fairness evaluation. Additionally, the figures and Appendix A visually illustrate recommendation disparities across demographic and personality-conditioned prompts, highlighting that unfairness is domain-specific and sensitive to prompt formulation. Thus, our evaluation highlights the critical importance of systematic, multi-dimensional fairness assessments to address pervasive biases across diverse contexts. We provide extended detailed results in Appendix A, including Table 8 and Figures 7 and 8.

4.3 Future Directions

Uncertainty and fairness are critical dimensions for advancing recommender systems, yet both require deeper exploration. While fairness ensures equitable treatment across diverse user groups, uncertainty quantification enables systems to recognize and communicate the limits of their confidence. In this section, we outline promising future directions that integrate fairness and uncertainty perspectives, aiming to improve the reliability, transparency, and trustworthiness of LLM-based recommendation

systems. **A general definition of uncertainty.** Uncertainty is the model's quantified lack of confidence in its predictions, arising from ambiguous inputs, model limitations, or distributional shift, and signaling when outputs may be unreliable or potentially non-factual. **A general definition of fairness.** Fairness is the principle that a recommender's outcomes treat individuals and groups according to agreed ethical or legal criteria; because multiple, sometimes conflicting formalizations exist, the appropriate definition depends on context and must be prioritized per scenario.

We outline several promising future research directions for uncertainty and fairness in Table 4, which can guide Explainable Automated Software Engineering by exploring theory, adaptation, needs, challenges, and ethics toward building more dependable and socially responsible AI systems.

4.4 Threats to Validity

We acknowledge several validity threats. *Conclusion validity* may be influenced by sampling variability and model-specific behavior; we mitigate this via standardized statistical tests and large, representative datasets. *Internal validity* risks from confounders were reduced through consistent configurations and controlled procedures. *Construct validity* is supported by grounding and validating our fairness and uncertainty metrics. The RQ1 case study highlights intrinsic output variability, while RQ2's ethical analyses of sensitive prompts are strictly research-focused and interpreted cautiously. A key limitation is the constrained scope dataset choice, limited sensitive attributes, and controlled experimental settings, which motivates broader, longitudinal, and deployment-level studies in future work [10, 15, 96, 88, 52].

5 Conclusion

This paper examines uncertainty and fairness challenges in LLM-based recommendation systems, with a focus on Google DeepMind's Gemini 1.5 Flash. Through in-depth case studies, we study that predictive uncertainty, quantified via entropy, undermines recommendation reliability, while fairness disparities persist across sensitive attributes like religion and race. Robustness testing shows unfairness withstands prompt variations such as typos and multilingual inputs. Our uncertainty-aware framework, detailed in Figure 2, incorporates token-level quantification to mitigate overconfidence and foster equitable outputs. By integrating personality-aware fairness, we reveal overlooked biases, advancing explainable AI for ethical deployment. Moving forward, we aim to extend evaluations to additional LLMs (e.g., ChatGPT, Claude, LLaMA, DeepSeek, Grok) [97], model personality via the Big Five [98], explore fairness-aware prompt optimization, and personalize recommendations [99, 100, 101, 11]. Furthermore, future work will include uncertainty analysis and quantification tailored to LLM recommendations, building on our proposed framework to enhance robustness and trustworthiness in automated software engineering.

Acknowledgments

We sincerely thank the reviewers for their insightful comments and constructive discussions that greatly improved this work. This research was supported by the Chinese Government Scholarship and Beihang University. We also acknowledge the support of the International Association for Safe & Ethical AI (IASEAI'26).

References

[1] D. Kumar, T. Grosz, N. Rekabsaz, E. Greif, and M. Schedl, "Fairness of recommender systems in the recruitment domain: an analysis from technical and legal perspectives," *Frontiers in big Data*, vol. 6, p. 1245198, 2023.

[2] S. Sabouri, P. Eibl, X. Zhou, M. Ziyadi, N. Medvidovic, L. Lindemann, and S. Chattopadhyay, "Trust dynamics in ai-assisted development: Definitions, factors, and implications," 2025.

[3] Y. Bakman, D. N. Yaldiz, S. Kang, T. Zhang, B. Buyukates, S. Avestimehr, and S. P. Karimireddy, "Reconsidering llm uncertainty estimation methods in the wild," *arXiv preprint arXiv:2506.01114*, 2025.

[4] Q. Yang, S. Ravikumar, F. Schmitt-Ulms, S. Lolla, E. Demir, I. Elistratov, A. Lavaee, S. Lolla, E. Ahmadi, D. Rus *et al.*, "Uncertainty-aware language modeling for selective question answering," *arXiv preprint arXiv:2311.15451*, 2023.

[5] C. K. Sah, L. Xiaoli, and M. M. Islam, "Unveiling bias in fairness evaluations of large language models: A critical literature review of music and movie recommendation systems," *Unveiling Bias in Fairness Evaluations of Large Language Models: A Critical Literature Review of Music and Movie Recommendation Systems*, vol. 8, no. 12, Jan. 2024. [Online]. Available: https://doi.org/10.5281/zenodo.10469839

[6] L. Das, B. Gjorgiev, and G. Sansavini, "Uncertainty-aware deep learning for digital twin-driven monitoring: Application to fault detection in power lines," *arXiv preprint arXiv:2303.10954*, 2023.

[7] Y. Yao, T. Han, J. Yu, and M. Xie, "Uncertainty-aware deep learning for reliable health monitoring in safety-critical energy systems," *Energy*, vol. 291, p. 130419, 2024.

[8] X. Tang, K. Yang, H. Wang, J. Wu, Y. Qin, W. Yu, and D. Cao, "Prediction-uncertainty-aware decision-making for autonomous vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 4, pp. 849–862, 2022.

[9] X. Sun, Y. Zhang, X. Tang, A. S. Bedi, and A. Bera, "Trustnavgpt: Modeling uncertainty to improve trustworthiness of audio-guided llm-based robot navigation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 8794–8801.

[10] W. Kweon, S. Jang, S. Kang, and H. Yu, "Uncertainty quantification and decomposition for llm-based recommendation," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 4889–4901.

[11] Y. Deldjoo, "Fairness of chatgpt and the role of explainable-guided prompts," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2023, pp. 13–22.

[12] Z. He, Z. Xie, R. Jha, H. Steck, D. Liang, Y. Feng, B. P. Majumder, N. Kallus, and J. McAuley, "Large language models as zero-shot conversational recommenders," in *Proceedings of the 32nd ACM international conference on information and knowledge management*, 2023, pp. 720–730.

[13] W. X. Grattafiori, A. Défossez, J. Copet, F. Azhar, H. Touvron, L. Martin, N. Usunier, T. Scialom, and G. Synnaeve, "Code llama: Open foundation models for code," *arXiv preprint arXiv:2308.12950*, 2023.

[14] C. K. Sah, "Understanding uncertainty in llms," in *2025 40th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2025, pp. 4128–4130.

[15] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

[16] Y. Xiao and W. Y. Wang, "Quantifying uncertainties in natural language processing tasks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 7322–7329.

[17] ——, "On hallucination and predictive uncertainty in conditional language generation," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 2734–2744.

[18] S. Kumar, "Answer-level calibration for free-form multiple choice question answering," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 665–679.

[19] J. Jiang, D. Yang, Y. Xiao, and C. Shen, "Convolutional gaussian embeddings for personalized recommendation with uncertainty."

[20] V. Coscrato and D. Bridge, "Estimating and evaluating the uncertainty of rating predictions and top-n recommendations in recommender systems," *ACM Transactions on Recommender Systems*, vol. 1, no. 2, pp. 1–34, 2023.

[21] M. Kirchhof, G. Kasneci, and E. Kasneci, "Position: Uncertainty quantification needs reassessment for large-language model agents," *arXiv preprint arXiv:2505.22655*, 2025.

[22] S. Devic, T. Srinivasan, J. Thomason, W. Neiswanger, and V. Sharan, "From calibration to collaboration: Llm uncertainty quantification should be more human-centered," *arXiv preprint arXiv:2506.07461*, 2025.

[23] Z. Atf, S. A. A. Safavi-Naini, P. R. Lewis, A. Mahjoubfar, N. Naderi, T. R. Savage, and A. Soroush, "The challenge of uncertainty quantification of large language models in medicine," *arXiv preprint arXiv:2504.05278*, 2025.

[24] M. Xiong, A. Santilli, M. Kirchhof, A. Golinski, and S. Williamson, "Efficient and effective uncertainty quantification for llms," in *Neurips Safe Generative AI Workshop 2024*, 2024.

[25] J. Gu, Z. Han, S. Chen, A. Beirami, B. He, G. Zhang, R. Liao, Y. Qin, V. Tresp, and P. Torr, "A systematic survey of prompt engineering on vision-language foundation models," *arXiv preprint arXiv:2307.12980*, 2023.

[26] J. White, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, and D. C. Schmidt, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.

[27] K. Zhu, J. Wang, J. Zhou, Z. Wang, H. Chen, Y. Wang, L. Yang, W. Ye, Y. Zhang, N. Gong *et al.*, "Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts," in *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, 2023, pp. 57–68.

[28] A. Amayuelas, K. Wong, L. Pan, W. Chen, and W. Wang, "Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models," *arXiv preprint arXiv:2305.13712*, 2023.

[29] C. Paliwal, A. Majumder, and S. Kaveri, "Predictive relevance uncertainty for recommendation systems," in *Proceedings of the ACM Web Conference 2024*, 2024, pp. 3900–3909.

[30] J. Harte, W. Zorgdrager, P. Louridas, A. Katsifodimos, D. Jannach, and M. Fragkoulis, "Leveraging large language models for sequential recommendation," in *Proceedings of the 17th ACM Conference on Recommender Systems*, 2023, pp. 1096–1102.

[31] S. Dai, N. Shao, H. Zhao, W. Yu, Z. Si, C. Xu, Z. Sun, X. Zhang, and J. Xu, "Uncovering chatgpt's capabilities in recommender systems," in *Proceedings of the 17th ACM Conference on Recommender Systems*, 2023, pp. 1126–1132.

[32] J. Zhang, K. Bao, Y. Zhang, W. Wang, F. Feng, and X. He, "Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation," in *Proceedings of the 17th ACM Conference on Recommender Systems*, 2023, pp. 993–999.

[33] J. Chen, H. Dong, X. Wang, F. Feng, M. Wang, and X. He, "Bias and debias in recommender system: A survey and future directions," *ACM Transactions on Information Systems*, vol. 41, no. 3, pp. 1–39, 2023.

[34] A. J. Biega, K. P. Gummadi, and G. Weikum, "Equity of attention: Amortizing individual fairness in rankings," in *The 41st international acm sigir conference on research & development in information retrieval*, 2018, pp. 405–414.

[35] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi *et al.*, "Fairness in recommendation ranking through pairwise comparisons," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2212–2220.

[36] Y. Zhao, Y. Wang, Y. Liu, X. Cheng, C. C. Aggarwal, and T. Derr, "Fairness and diversity in recommender systems: a survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 1, pp. 1–28, 2025.

[37] J. Gao, B. Chen, X. Zhao, W. Liu, X. Li, Y. Wang, Z. Zhang, W. Wang, Y. Ye, S. Lin *et al.*, "Llm-enhanced reranking in recommender systems," *arXiv preprint arXiv:2406.12433*, 2024.

[38] P. P. Liang, C. Wu, L.-P. Morency, and R. Salakhutdinov, "Towards understanding and mitigating social biases in language models," in *International conference on machine learning*. PMLR, 2021, pp. 6565–6576.

[39] B. Richardson and J. E. Gilbert, "A framework for fairness: A systematic review of existing fair ai solutions," *arXiv preprint arXiv:2112.05700*, 2021.

[40] S. K. Sakib and A. B. Das, "Challenging fairness: A comprehensive exploration of bias in llm-based recommendations," in *2024 IEEE International Conference on Big Data (BigData)*. IEEE, 2024, pp. 1585–1592.

[41] M. Jiang, K. Bao, J. Zhang, W. Wang, Z. Yang, F. Feng, and X. He, "Item-side fairness of large language model-based recommendation system," in *Proceedings of the ACM Web Conference 2024*, 2024, pp. 4717–4726.

[42] Z. Zhao, W. Fan, J. Li, Y. Liu, X. Mei, Y. Wang, Z. Wen, F. Wang, X. Zhao, J. Tang *et al.*, "Recommender systems in the era of large language models (llms)," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[43] T. Ma, Y. Cheng, H. Zhu, and H. Xiong, "Large language models are not stable recommender systems," *arXiv preprint arXiv:2312.15746*, 2023.

[44] S. Dai, C. Xu, S. Xu, L. Pang, Z. Dong, and J. Xu, "Bias and unfairness in information retrieval systems: New challenges in the llm era," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 6437–6447.

[45] W. Liu, B. Liu, J. Qin, X. Zhang, W. Huang, and Y. Wang, "Fairness identification of large language models in recommendation," *Scientific Reports*, vol. 15, no. 1, p. 5516, 2025.

[46] T. Takayanagi and K. Izumi, "Incorporating domain-specific traits into personality-aware recommendations for financial applications," *New Generation Computing*, pp. 1–15, 2024.

[47] M. Tkalcic and L. Chen, "Personality and recommender systems." *Recommender systems handbook*, vol. 54, pp. 715–739, 2015.

[48] Q. Yang, S. Nikolenko, A. Huang, and A. Farseev, "Personality-driven social multimedia content recommendation," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 7290–7299.

[49] Y. Deldjoo and T. Di Noia, "Cfairllm: Consumer fairness evaluation in large-language model recommender system," *arXiv preprint arXiv:2403.05668*, 2024.

[50] W. I. D. Mining, "Data mining: Concepts and techniques," *Morgan Kaufinann*, vol. 10, no. 559-569, p. 4, 2006.

[51] M. Tomlein, B. Pecher, J. Simko, I. Srba, R. Moro, E. Stefancova, M. Kompan, A. Hrckova, J. Podrouzek, and M. Bielikova, "An audit of misinformation filter bubbles on youtube: Bubble bursting and recent behavior changes," in *Proceedings of the 15th ACM Conference on Recommender Systems*, 2021, pp. 1–11.

[52] C. Yin, L. Fan, J. Wang, D. Hu, H. Zhang, C. Zhang, and Y. Xiang, "Uncertainty-aware semantic decoding for llm-based sequential recommendation," *arXiv preprint arXiv:2508.07210*, 2025.

[53] D. Di Palma, G. M. Biancofiore, V. W. Anelli, F. Narducci, T. Di Noia, and E. Di Sciascio, "Evaluating chatgpt as a recommender system: A rigorous approach," *arXiv preprint arXiv:2309.03613*, 2023.

[54] B. Sguerra, E. V. Epure, H. Lee, and M. Moussallam, "Biases in llm-generated musical taste profiles for recommendation," *arXiv preprint arXiv:2507.16708*, 2025.

[55] W. Zhuang, "Uncertainty quantification in large language models (uqlm) for e-commerce analytics with google gemini," 2025.

[56] D. Herrera-Poyatos, C. Peláez-González, C. Zuheros, A. Herrera-Poyatos, V. Tejedor, F. Herrera, and R. Montes, "An overview of model uncertainty and variability in llm-based sentiment analysis. challenges, mitigation strategies and the role of explainability," *arXiv preprint arXiv:2504.04462*, 2025.

[57] Y. Peng, H. Chen, C.-S. Lin, G. Huang, J. Hu, H. Guo, B. Kong, S. Hu, X. Wu, and X. Wang, "Uncertainty-aware explainable recommendation with large language models," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[58] C. Jiang, J. Wang, W. Ma, C. L. Clarke, S. Wang, C. Wu, and M. Zhang, "Beyond utility: Evaluating llm as recommender," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 3850–3862.

[59] Z. Wen, Z. Liu, Z. Tian, S. Pan, Z. Huang, D. Li, and M. Huang, "Scenario-independent uncertainty estimation for llm-based question answering via factor analysis," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 2378–2390.

[60] Y. Huang, J. Song, Z. Wang, S. Zhao, H. Chen, F. Juefei-Xu, and L. Ma, "Look before you leap: An exploratory study of uncertainty analysis for large language models," *IEEE Transactions on Software Engineering*, 2025.

[61] S. S. Elmoghazy, M. A. Shouman, H. K. Elminir, and G. E. I. Selim, "Comparative analysis of methodologies and approaches in recommender systems utilizing large language models," *Artificial Intelligence Review*, vol. 58, no. 7, p. 207, 2025.

[62] S. Rabanser, "Uncertainty-driven reliability: Selective prediction and trustworthy deployment in modern machine learning," *Department of Computer Science, University of Toronto*, 2025.

[63] T. Liu, J. Liu, D. Li, and S. Tan, "Bayesian deep-learning structured illumination microscopy enables reliable super-resolution imaging with uncertainty quantification," *Nature Communications*, vol. 16, no. 1, p. 5027, 2025.

[64] M. Wang, T. Lin, L. Wang, A. Lin, K. Zou, X. Xu, Y. Zhou, Y. Peng, Q. Meng, Y. Qian *et al.*, "Uncertainty-inspired open set learning for retinal anomaly identification," *Nature Communications*, vol. 14, no. 1, p. 6757, 2023.

[65] G. Detommaso, A. Gasparin, M. Donini, M. Seeger, A. G. Wilson, and C. Archambeau, "Fortuna: A library for uncertainty quantification in deep learning," *Journal of Machine Learning Research*, vol. 25, no. 238, pp. 1–7, 2024.

[66] J. Zhu, J. Wang, I. J. Cox, and M. J. Taylor, "Risky business: modeling and exploiting uncertainty in information retrieval," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, pp. 99–106.

[67] Q. Xu, S. Ali, T. Yue, and M. Arratibel, "Uncertainty-aware transfer learning to evolve digital twins for industrial elevators," in *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2022, pp. 1257–1268.

[68] Y. Zhu, G. Li, X. Jiang, J. Li, H. Mei, Z. Jin, and Y. Dong, "Uncertainty-guided chain-of-thought for code generation with llms," *arXiv preprint arXiv:2503.15341*, 2025.

[69] L. Da, X. Liu, J. Dai, L. Cheng, Y. Wang, and H. Wei, "Understanding the uncertainty of llm explanations from reasoning topology."

[70] J. Deuschel, A. Foltyn, K. Roscher, and S. Scheele, "The role of uncertainty quantification for trustworthy ai," in *Unlocking Artificial Intelligence: From Theory to Applications*. Springer, 2024, pp. 95–115.

[71] W. Pisciotta, P. Arina, D. Hofmaenner, and M. Singer, "Difficult diagnosis in the icu: making the right call but beware uncertainty and bias," *Anaesthesia*, vol. 78, no. 4, pp. 501–509, 2023.

[72] O. Shorinwa, Z. Mei, J. Lidard, A. Z. Ren, and A. Majumdar, "A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions," *ACM Computing Surveys*, 2025.

[73] F. Ye, M. Yang, J. Pang, L. Wang, D. Wong, E. Yilmaz, S. Shi, and Z. Tu, "Benchmarking llms via uncertainty quantification," *Advances in Neural Information Processing Systems*, vol. 37, pp. 15 356–15 385, 2024.

[74] S. Schmitt, J. Shawe-Taylor, and H. van Hasselt, "General uncertainty estimation with delta variances," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 19, 2025, pp. 20 318–20 328.

[75] Y. Zhang, X. Chen *et al.*, "Explainable recommendation: A survey and new perspectives," *Foundations and Trends® in Information Retrieval*, vol. 14, no. 1, pp. 1–101, 2020.

[76] S. Yao and B. Huang, "Beyond parity: Fairness objectives for collaborative filtering," *Advances in neural information processing systems*, vol. 30, 2017.

[77] S. C. Geyik and K. K. Ambler, "Fairness-aware ranking in search & recommendation systems with application to linkedin talent search," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2221–2231.

[78] P. Li *et al.*, "Counterfactual fairness in recommendation," *Proceedings of the Web Conference*, 2021.

[79] Y. Wang, W. Ma, M. Zhang, Y. Liu, and S. Ma, "A survey on the fairness of recommender systems," *ACM Transactions on Information Systems*, vol. 41, no. 3, pp. 1–43, 2023.

[80] L. Mihaljevic, J. Klemmer, L. Weidinger, and P. Bhargava, "More than just prompting: Towards fair and robust llms via debiasing techniques," *arXiv preprint arXiv:2403.01462*, 2024.

[81] J. Yang, V. Narayan, and A. Weller, "Behavioral disparities in language model recommendations," *arXiv preprint arXiv:2402.03155*, 2024.

[82] D. Jin, L. Wang, H. Zhang, Y. Zheng, W. Ding, F. Xia, and S. Pan, "A survey on fairness-aware recommender systems," *Information Fusion*, vol. 100, p. 101906, 2023.

[83] R.-C. Chen, Q. Ai, G. Jayasinghe, and W. B. Croft, "Correcting for recency bias in job recommendation," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 2185–2188. [Online]. Available: https://doi.org/10.1145/3357384.3358131

[84] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Computing and applying personality profiles to improve social media searches," *IEEE Internet Computing*, vol. 15, no. 3, pp. 18–28, 2011.

[85] Y. Hu, X. Wang *et al.*, "A survey on the fairness of recommender systems," *ACM Transactions on Information Systems*, vol. 41, no. 3, pp. 1–52, 2023.

[86] S. Dhelim, N. Aung, M. A. Bouras, H. Ning, and E. Cambria, "A survey on personality-aware recommendation systems," *Artificial Intelligence Review*, pp. 1–46, 2022.

[87] Z. Fu, Y. Xian, R. Gao, J. Zhao, Q. Huang, Y. Ge, S. Xu, S. Geng, C. Shah, Y. Zhang *et al.*, "Fairness-aware explainable recommendation over knowledge graphs," in *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, 2020, pp. 69–78.

[88] Z. Sun, Z. Chen, J. Zhang, and D. Hao, "Fairness testing of machine translation systems," *ACM Transactions on Software Engineering and Methodology*, vol. 33, no. 6, pp. 1–27, 2024.

[89] M. Saad, J. A. H. López, B. Chen, D. Varró, and T. Sharma, "An adaptive language-agnostic pruning method for greener language models for code," *Proceedings of the ACM on Software Engineering*, vol. 2, no. FSE, pp. 1183–1204, 2025.

[90] A. Aird, E. Štefancová, A. Buhayh, C. All, M. Homola, N. Mattei, and R. Burke, "Integrating individual and group fairness for recommender systems through social choice," in *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, 2025, pp. 177–186.

[91] Y. Li, J. Wang, H. Sundaram, and Z. Liu, "Llm-recg: A semantic bias-aware framework for zero-shot sequential recommendation," in *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, 2025, pp. 237–246.

[92] G. Meehan and J. Pauwels, "On inherited popularity bias in cold-start item recommendation," in *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, 2025, pp. 649–654.

[93] H. Tang, J. Wang, Y. Peng, G. Meng, R. Luo, B. Chen, L. Chen, Y. Wang, and S.-T. Xia, "Modeling uncertainty in composed image retrieval via probabilistic embeddings," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 1210–1222.

[94] Q. Zhao, D. Li, Y. Liu, W. Cheng, Y. Sun, M. Oishi, T. Osaki, K. Matsuda, H. Yao, C. Zhao *et al.*, "Uncertainty propagation on llm agent," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 6064–6073.

[95] X. Shen, Z. Jiang, J. Zhang, J. Han, Y. Wan, C. Guo, B. Liu, J. Wu, R. Li, and P. S. Yu, "Provbench: A benchmark of legal provision recommendation for contract auto-reviewing," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 6240–6254.

[96] K. Jiang, H. Sun, and M. Zhang, "Ifairlrs: Enhancing item-side fairness in large language model recommendation systems," in *Proceedings of the ACM Conference on Recommender Systems (RecSys)*, 2024.

[97] F.-F. Zhao, H.-J. He, J.-J. Liang, J. Cen, Y. Wang, H. Lin, F. Chen, T.-P. Li, J.-F. Yang, L. Chen *et al.*, "Benchmarking the performance of large language models in uveitis: a comparative analysis of chatgpt-3.5, chatgpt-4.0, google gemini, and anthropic claude3," *Eye*, pp. 1–6, 2024.

[98] T. F. Bainbridge, S. G. Ludeke, and L. D. Smillie, "Evaluating the big five as an organizing framework for commonly used psychological trait scales." *Journal of Personality and Social Psychology*, vol. 122, no. 4, p. 749, 2022.

[99] C. K. Sah and X. Lian, "Perfairx: Is there a balance between fairness and personality in large language model recommendations?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 2750–2759.

[100] S. Furniturewala, S. Jandial, A. Java, P. Banerjee, S. Shahid, S. Bhatia, and K. Jaidka, "Thinking fair and slow: On the efficacy of structured prompts for debiasing language models," *arXiv preprint arXiv:2405.10431*, 2024.

[101] Y. Wu, R. Xie, Y. Zhu, F. Zhuang, A. Xiang, X. Zhang, L. Lin, and Q. He, "Selective fairness in recommendation via prompts," in *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 2022, pp. 2657–2662.