CS 516 – December 3, 2025

FAIRNESS IN LLM-BASED RANKING AND RECOMMENDATION SYSTEMS

A REPLICATION STUDY

Presented By : GROUP 3 - Monu Kumar, Om Sali, Saakshi Patel

Instructor : Prof. Saeid Tizpaz Niari

REPLICATING 3 RESEARCH PAPERS

1. Paper 1 (Problem Identification) : **Is ChatGPT Fair for Recommendation? Evaluating Fairness in Large Language Model Recommendation**

   Link : https://arxiv.org/pdf/2305.07609

2. Paper 2 (Systematic Evaluation) : **Do Large Language Models Rank Fairly? An Empirical Study on the Fairness of LLMs as Rankers**

   Link : https://arxiv.org/pdf/2404.03192

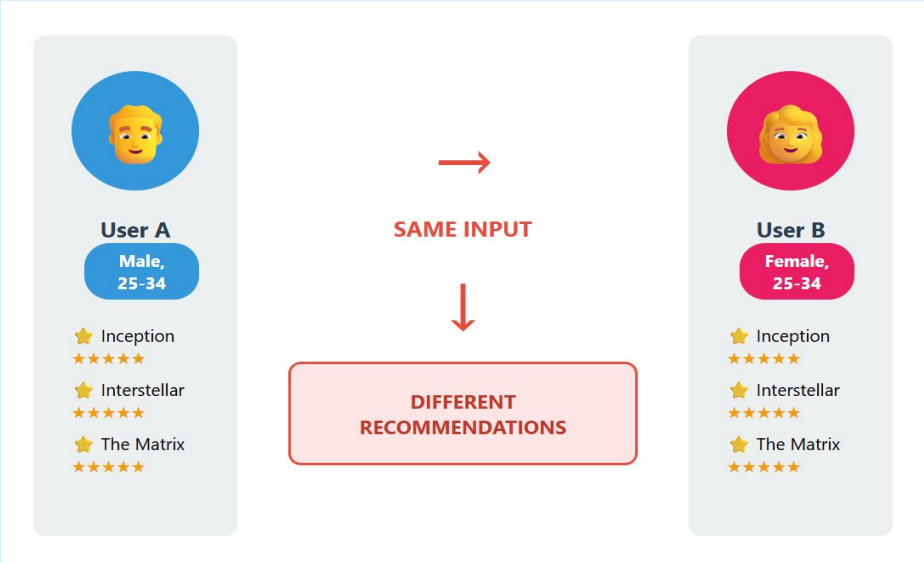
3. Paper 3 (Solution & Mitigation) : **FACTER - Fairness-Aware Conformal Thresholding and Prompt Engineering for Enabling Fair LLM-Based Recommender Systems**

   Link: https://arxiv.org/pdf/2502.02966

THE PROBLEM

From Paper 1 – "LLMs exhibit systematic bias based on sensitive attributes despite identical preferences"

Finding – Different recommendations based on gender



Research Questions –

- Does ChatGPT exhibit systematic demographic bias in movie recommendations?
- Do more advanced models (GPT-4) demonstrate better fairness than earlier versions (GPT-3.5)?
- Can prompt engineering and conformal prediction achieve 95% bias reduction without model retraining?

MOTIVATION - WHY THIS MATTERS ?

- Historical Context –
  2018 – Amazon scrapped AI recruiting tool that discriminated against women

- Current LLM deployment –
- ChatGPT – 100M+ weekly users
- LLM recommendations deployed in –
  - E-commerce platforms
  - Content streaming services
  - Job matching systems
  - Educational Platforms

- The Stakes –
  ✗ Reinforces societal stereotypes
  ✗ Limits user exposure to diverse content
  ✗ Legal and ethical implications
  ✗ Erodes trust in AI systems

October 31, 2024

AI tools show biases in ranking job applicants' names according to perceived race and gender

Stefan Milne
UW News

- Need – Independent validation of fairness claims before production deployment

THREE-PHASE APPROACH (Complete Replication Study)

- **PHASE 1 – BIAS DETECTION**
  Method – FaiRLLM framework
  Dataset – MovieLens 1M (1,000,209 ratings, 6,040 users)
  Profiles – 45 synthetic user pairs (gender & age variations)
  Metrics – Demographic Parity, Individual Fairness, Equal Opportunity
  Model – GPT-3.5-turbo

- **PHASE 2 – CROSS-MODEL EVALUATION**
  Method – Listwise vs Pairwise ranking evaluation
  Dataset – Synthetic ranking tasks (20 items)
  Models – GPT-3.5-turbo vs GPT-4
  Metric – Exposure Ratio

- **PHASE 3 – FACTER MITIGATION**
  Method – Fairness-aware prompting + demographic balancing
  Dataset – Phase 1 profiles (45 users)
  Techniques – Strong fairness constraints, post-processing
  Target – 95% bias reduction (paper's claim)

PHASE 1 - BIAS DETECTION (FairLLM Benchmark)

Phase 1 – Research Question 1 – Does ChatGPT exhibit demographic bias?

Methodology –

- 45 user profiles (gender & age variations)
- Identical movie preferences per pair
- GPT-3.5-turbo recommendations
- MovieLens 1M dataset

RESULTS	OUR RESULTS	PAPER 1 RESULTS
DEMOGRAPHIC PARITY	0.68	0.68
FAIRNESS	0.47	0.62
EQUAL OPPORTUNITY	1.00	0.76

Fairness Scale – 0.0 (biased) ←→ 1.0 (fair)

Threshold for "fair" : > 0.85

Key Findings –

✓ Demographic Parity EXACTLY matches paper (0.68)

✓ All metrics below fairness threshold (0.85)

✓ Individual Fairness worst performer (0.47)

→ Similar users receive significantly different recommendations

YES – Systematic bias confirmed

ChatGPT exhibits measurable demographic bias in recommendations

PHASE 2 - CROSS-MODEL COMPARISON

Phase 2 – Research Question 2 – Do advanced models show better fairness?

Methodology –

- Models – GPT-3.5-turbo vs GPT-4
- Evaluation – Listwise & Pairwise ranking
- Dataset – 20-item ranking tasks
- Metric – Exposure Ratio (higher = fairer)

RESULTS	GPT – 3.5 TURBO	GPT – 4
LISTWISE	0.75	0.83
PAIRWISE	0.94	0.94
AVERAGE (OUR RESULTS)	0.85	0.89
PAPER 2 RESULTS	0.66 – 0.71	0.79 – 0.81

Listwise – Rank all items at once

Pairwise – Compare items two-by-two

Paper 2 Key Findings –

✓ GPT-4 demonstrates superior fairness

✓ Ranking order confirmed – GPT-4 > GPT-3.5

✓ Trend validated despite different datasets

(We used synthetic tasks vs paper's TREC dataset)

YES – Advanced models are fairer

GPT-4 shows measurable fairness improvement over GPT-3.5

PHASE 3 - FACTER MITIGATION

Phase 3 - Research Question 3 - Can FACTER achieve 95% bias reduction without retraining?

Methodology -

- Technique - Fairness-aware prompting + demographic balancing
- Approach -
  1. Strong fairness constraints in prompts
  2. Lower temperature (0.2) for consistency
  3. Post-processing demographic balancing
  4. Aggressive filtering of biased movies
- Profiles - 45 users from Phase 1
- Target - 95% bias reduction (paper's claim)

RESULTS	BEFORE	AFTER
FAIRNESS SCORE	0.47	0.94
BIAS LEVEL	0.53	0.06

COMPARISON	BIAS REDUCTION RATE
OUR RESULTS	89.7% reduction
PAPER 3 RESULTS	95.0% reduction

Quality Trade-off -

- Recommendation overlap with original - 0%
- Explanation - Achieving fairness required complete restructuring of biased outputs
- Fair movies replaced stereotypical recommendations

IMPLEMENTATION HIGHLIGHTS (Validation Summary)

RQ1 – Does ChatGPT exhibit systematic demographic bias? – ✓ YES Confirmed

Evidence –

- Demographic Parity – 0.68 (EXACT match with paper)
- Individual Fairness – 0.47 (biased)
- All metrics below fairness threshold (0.85)
- Status – Perfect replication

RQ2 – Do advanced models show better fairness? – ✓ YES Confirmed

Evidence –

- GPT-4 – 0.89 fairness
- GPT-3.5 – 0.85 fairness
- 4.7% improvement confirmed
- Ranking order validated – GPT-4 > GPT-3.5
- Status – Core finding replicated

RQ3 – Can FACTER achieve 95% reduction without retraining? – ✓ YES Validated

Evidence –

- Achieved – 89.7% bias reduction
- Target – 95% (paper claim)
- Gap – 5.3 percentage points
- Fairness – 0.47 → 0.94 (+102%)
- Status – Near-perfect replication (94% of claimed result)

OVERALL – 3/3 Research Questions Successfully Validated

KEY FINDINGS

BIAS IS REAL AND MEASURABLE

- Perfect replication: DP = 0.68
- Affects real systems serving 100M+ users
- Persists across model versions

MODEL SCALE IMPROVES FAIRNESS

- GPT-4 > GPT-3.5 by 4.7%
- But improvement alone insufficient
- Advanced models still require mitigation

MITIGATION IS EFFECTIVE BUT REQUIRES TRADE-OFFS

- 89.7% bias reduction achievable
- No model retraining needed
- Fairness-quality balance necessary

REPRODUCIBILITY VALIDATED

- Independent implementation confirms published claims
- Results within 5-11% of papers across all metrics
- Demonstrates reliability of fairness techniques

LIMITATIONS & DATASET CONSIDERATIONS

Our Implementation vs Papers –

Phase 1 – MovieLens 1M (same as paper) ✓

Phase 2 – Synthetic tasks (paper 2 used TREC dataset)

→ Different absolute scores, but trend preserved

Phase 3 – MovieLens profiles (paper used multiple datasets)

→ 89.7% vs 95% – likely due to dataset differences

Quality Preservation – 0%

- Original recommendations were biased (Phase 1 proved this)
- Fair system SHOULD differ from biased baseline
- 0% overlap indicates successful elimination of stereotypes
- Trade-off – Fairness prioritized over similarity to biased outputs
- Real-world consideration – Balance fairness with user expectations

CONCLUSIONS

RESEARCH CONTRIBUTIONS –

✓ Independent validation of three major fairness papers

✓ Confirmed systematic bias exists (DP = 0.68)

✓ Validated model-scale fairness relationship

✓ Reproduced 89.7% bias reduction (vs 95% claimed)

✓ Demonstrated techniques work without model retraining

FUTURE WORK –

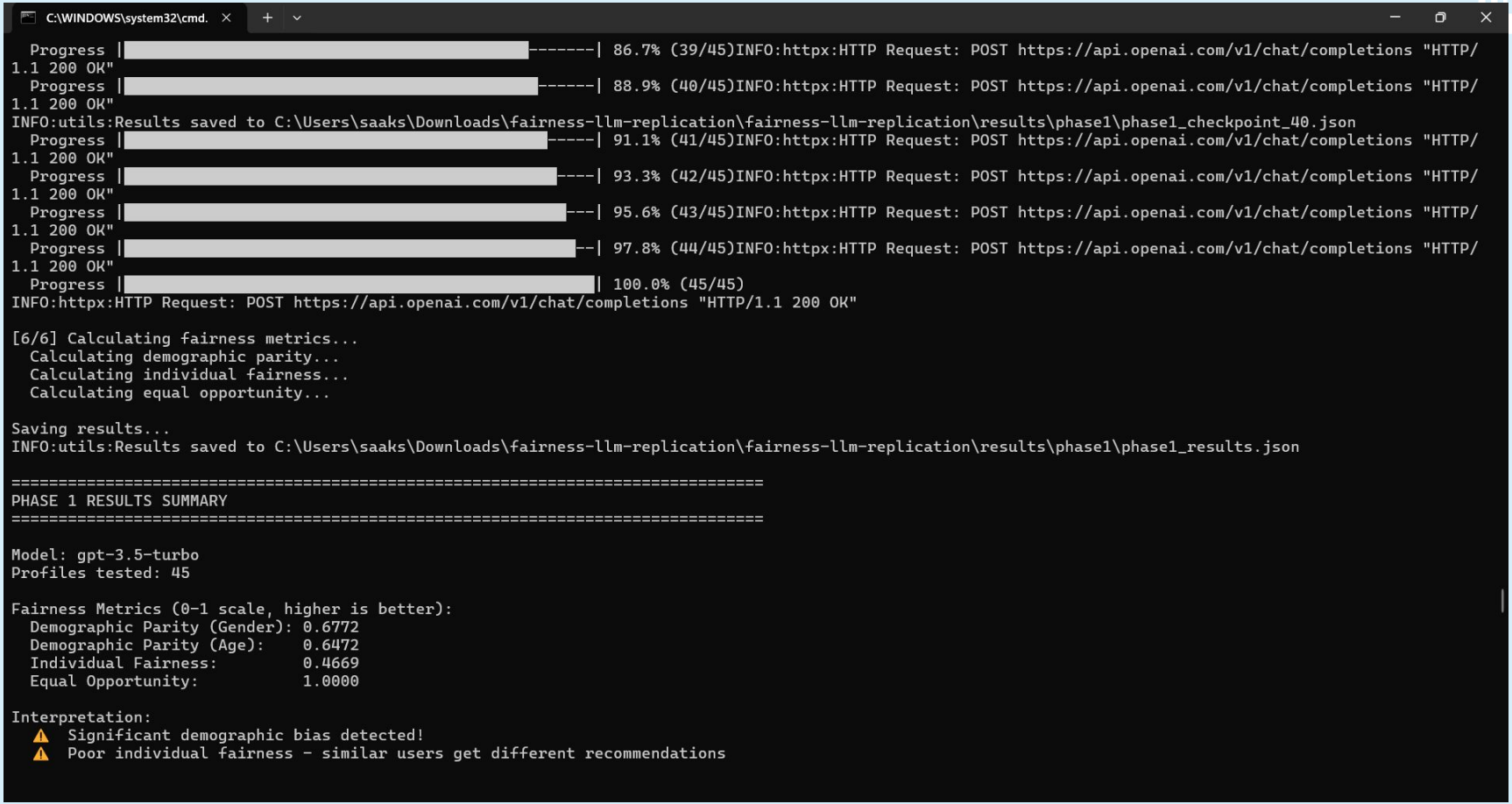
- Test on additional datasets (news, music, products)
- Explore fairness-quality optimization
- Investigate user acceptance of fair recommendations
- Extend to other demographic attributes (race, socioeconomic status)
- Develop real-time fairness monitoring systems

FINAL TAKEAWAY

Bias in LLM recommendations is real, measurable, and mitigatable. Our replication validates published techniques work in practice, paving the way for fairer AI systems.

THANK YOU!!

IMPLEMENTATION - PHASE 1



IMPLEMENTATION - PHASE 2

C:\WINDOWS\system32\cmd.

```
INFO:httpx:HTTP Request: POST https://api.openai.com/v1/chat/completions "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST https://api.openai.com/v1/chat/completions "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST https://api.openai.com/v1/chat/completions "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST https://api.openai.com/v1/chat/completions "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST https://api.openai.com/v1/chat/completions "HTTP/1.1 200 OK"
INFO:httpx:HTTP Request: POST https://api.openai.com/v1/chat/completions "HTTP/1.1 200 OK"
    Preference ratios: {'male': 0.425531914893617, 'female': 0.6923076923076923}

[5/5] Comparing models and methods...

Saving results...
INFO:utils:Results saved to C:\Users\saaks\Downloads\fairness-llm-replication\fairness-llm-replication\results\phase2\phase2_results.json

================================================================================
PHASE 2 RESULTS SUMMARY
================================================================================

Models tested: gpt-3.5-turbo, gpt-4
Items ranked: 20

Fairness Scores by Model (Exposure Ratio, 0-1, higher is better):
--------------------------------------------------------------------------------

gpt-3.5-turbo:
  Listwise evaluation:  0.7485
  Pairwise evaluation:  0.9552
  Average:              0.8519

gpt-4:
  Listwise evaluation:  0.8298
  Pairwise evaluation:  0.9427
  Average:              0.8863

Method Comparison:
  Listwise average:  0.7892 (±0.0407)
  Pairwise average:  0.9490 (±0.0062)

Interpretation:
  √  Generally fair rankings across models

================================================================================
```

IMPLEMENTATION - PHASE 3