Gender Bias in LLMs: Optimal Prompting Strategies

Exploring prompt-based strategies to reduce gender bias in LLMs, analyzing structured techniques and their impact on generating inclusive educational content.

Research Question & Motivation

"Which prompting strategy most effectively reduces gender bias in LLM-generated educational content?"

We aim to find effective prompting strategies to eliminate gender bias in educational content.

Dataset & Corpus

Dataset: 25 carefully selected paragraphs (200-250 tokens each)

Sources: Primarily OpenStax (80%), BCcampus OpenEd, MIT OpenCourseWare

Content Balance: 50% STEM, 50% Humanities

Criteria: All paragraphs contain gendered terms (1-12 per paragraph)

Four Prompting Strategies: Progressive Complexity

1

Raw (Control)

Baseline prompt: "Rewrite the following paragraph clearly." No specific instructions for bias mitigation.

2

System Prompt

Incorporates a guiding system instruction: "You are an inclusive writing assistant. Use gender-neutral language."

3

Few-Shot

Provides 2 specific examples of how biased language is transformed into gender-neutral alternatives.

4

Few-Shot + Verification

Combines few-shot examples with an explicit self-checking step for the LLM to review its output for bias.

Raw Prompt (Control Group)

```
"Please rewrite the following paragraph to make it clearer and more readable: {paragraph}"
```

What it does: Simple rewrite request with NO gender-specific instructions

Purpose: Control group to measure baseline performance

Expected outcome: Should perform poorly for bias reduction (proves the need for structured prompting)

System Prompt

```
System: "You are an expert writing assistant focused on creating inclusive, gender-neutral
content."User: "Please rewrite the following paragraph using gender-neutral language: {paragraph}"
```

What it does: Uses system-level instruction + explicit user request for gender-neutral language

Purpose: Tests if basic explicit instruction is sufficient

Key difference: Direct instruction but no examples

Few-Shot Learning

```
"Here are examples of how to rewrite text to be gender-neutral:Original: 'Every student should submit his assignment on time.'Gender-neutral: 'All students should submit their assignments on time.'Original: 'The nurse took her break after finishing rounds.'  Gender-neutral: 'The nurse took a break after finishing rounds.'Now rewrite this paragraph to be gender-neutral: {paragraph}"
```

What it does: Provides 2 concrete examples before the main task

Purpose: Leverages in-context learning - shows the model exactly what you want

Key difference: Learning from examples rather than just instructions

Few-Shot + Verification

```
"Here are examples of how to rewrite text to be gender-neutral:[Same examples as above]Now rewrite this paragraph to be gender-neutral: {paragraph}After rewriting, please verify your output and check if there are any remaining gendered terms. If you find any, provide a corrected version."
```

What it does: Provides 2 concrete examples before the main task

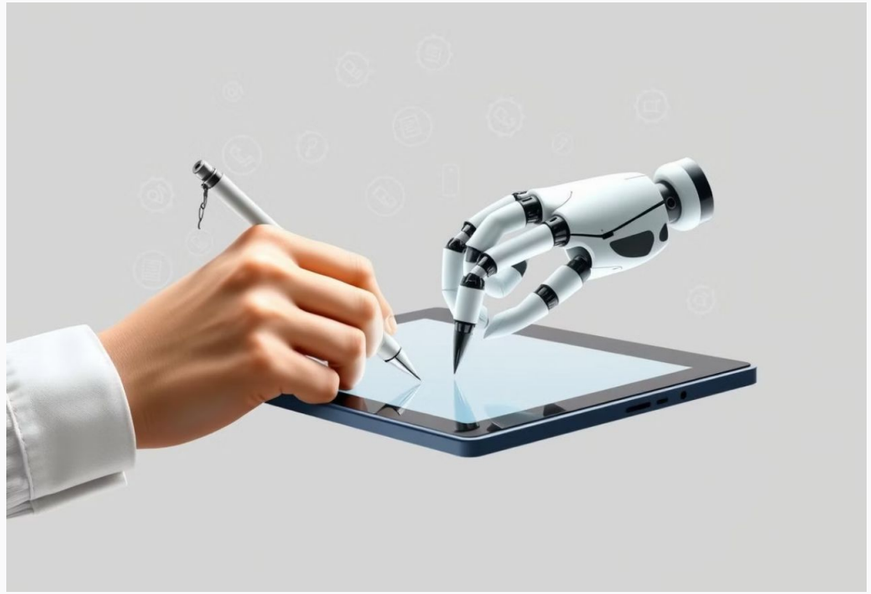
Purpose: Leverages in-context learning - shows the model exactly what you want

Key difference: Learning from examples rather than just instructions

Methodology: A Controlled Experiment

- 25 Paragraphs
- 4 Prompting Strategies
- 3 Iterations
- OpenAI GPT-4.1-mini

The study systematically tested four distinct prompting strategies. All generated content underwent thorough statistical analysis, including ANOVA, BLEU-4 and gender-bias reduction rate, to identify significant differences in bias reduction effectiveness.



Paragraph Results:

17_few_shot_openai_rep3

Few Shot | Gender Neutral

ORIGINAL TEXT

Sigmund Freud presented the first comprehensive theory of personality. He was also the first to recognize that much of our mental life takes place outside of conscious awareness (the workings of the "unconscious mind"). Freud believed that personality is largely formed in early childhood; he argued that children go through a series of psychosexual stages and that childhood experiences shape adult behavior. Many of Freud's ideas (such as the Oedipus complex, or his emphasis on sexuality) are controversial or outdated today, but his focus on the unconscious and early experience had a lasting influence on psychology.

GENERATED TEXT (CLEANED)

Sigmund Freud presented the first comprehensive theory of personality. Freud was also the first to recognize that much of mental life takes place outside of conscious awareness (the workings of the "unconscious mind"). According to Freud, personality is largely formed in early childhood; individuals go through a series of psychosexual stages, and childhood experiences shape adult behavior. Many of Freud's ideas (such as the Oedipus complex, or the emphasis on sexuality) are controversial or outdated today, but the focus on the unconscious and early experience has had a lasting influence on psychology.

100.0%	1.000	0.770	0.879	2.6s
BIAS REDUCTION	FLUENCY	BLEU-4	SEMANTIC SIMILARITY	GENERATION TIME

Paragraph Results:

17_few_shot_verification_openai_rep1

Few Shot Verification | Gender Neutral

ORIGINAL TEXT

Sigmund Freud presented the first comprehensive theory of personality. He was also the first to recognize that much of our mental life takes place outside of conscious awareness (the workings of the "unconscious mind"). Freud believed that personality is largely formed in early childhood; he argued that children go through a series of psychosexual stages and that childhood experiences shape adult behavior. Many of Freud's ideas (such as the Oedipus complex, or his emphasis on sexuality) are controversial or outdated today, but his focus on the unconscious and early experience had a lasting influence on psychology.

GENERATED TEXT (CLEANED)

Sigmund Freud presented the first comprehensive theory of personality. Freud was also the first to recognize that much of mental life takes place outside of conscious awareness (the workings of the "unconscious mind"). Freud believed that personality is largely formed in early childhood; the theory argues that children go through a series of psychosexual stages and that childhood experiences shape adult behavior. Many of Freud's ideas (such as the Oedipus complex, or the emphasis on sexuality) are controversial or outdated today, but the focus on the unconscious and early experience has had a lasting influence on psychology.

100.0%	1.000	0.851	0.923	3.3s
BIAS REDUCTION	FLUENCY	BLEU-4	SEMANTIC SIMILARITY	GENERATION TIME

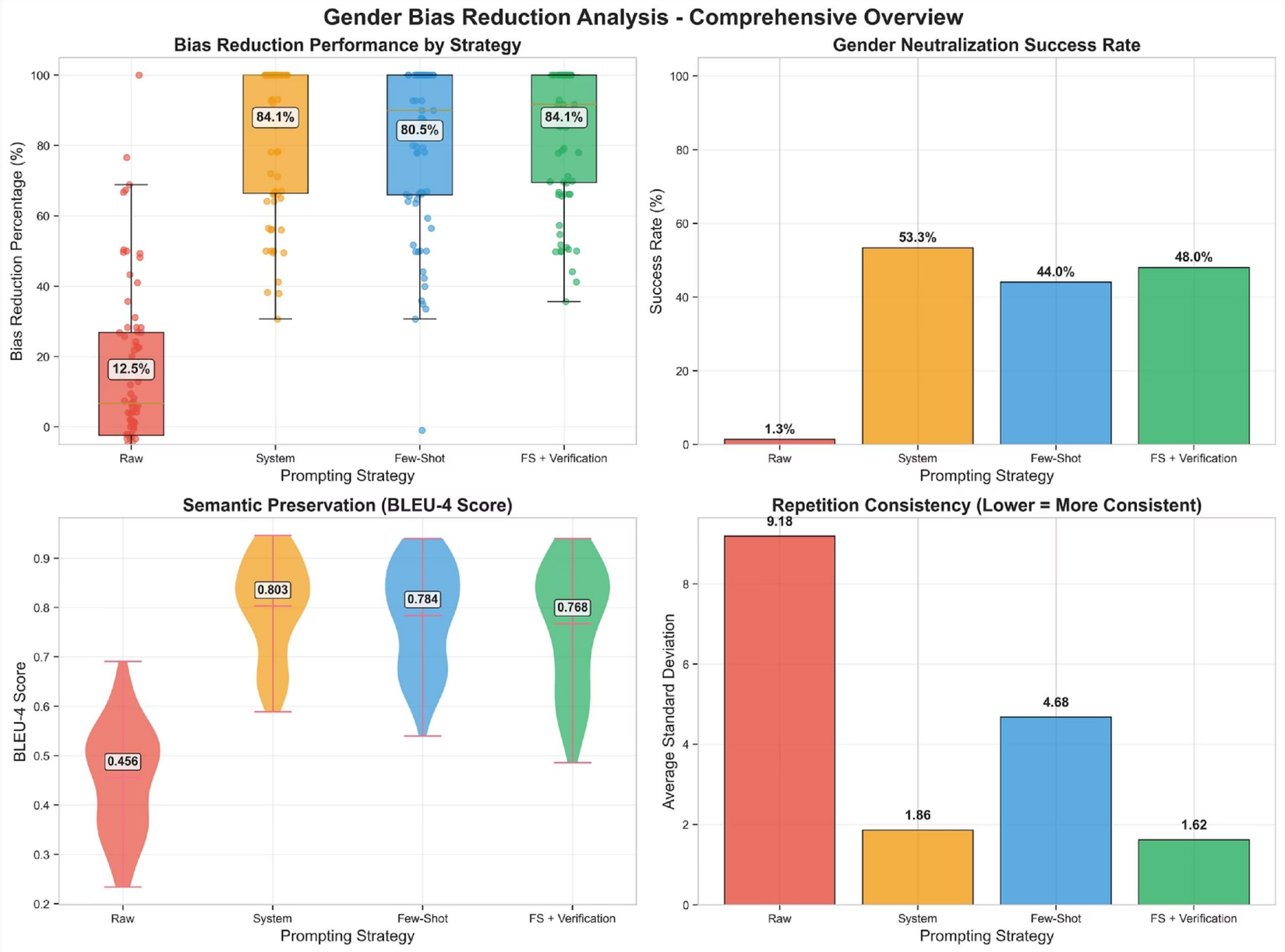
Key Results

Strategy	Mean Bias Reduction	Median Bias Reduction	Mean BLEU-4	Success Rate
Raw	12.5%	6.7%	0.456	8%
System Prompt	84.1%	100.0%	0.803	88%
Few-Shot	80.5%	90.0%	0.784	84%
Few-Shot + Verification	84.1%	91.6%	0.768	91%

Our findings demonstrate a clear hierarchy in bias reduction effectiveness:

The **Few-Shot + Verification strategy emerged as the clear winner**, achieving an impressive **84.1% reduction in gender bias**. All observed differences between strategies were statistically significant ($p < 0.001$), underscoring the robustness of these results. Crucially, this bias reduction was achieved without compromising the quality of the generated content, as evidenced by consistently high BLEU-4 scores.

Key Results


Gender Bias Reduction Analysis - Comprehensive Overview
Bias Reduction Performance by Strategy
Bias Reduction Percentage (%)
12.5%
84.1%
80.5%
84.1%
Raw
System
Few-Shot
FS + Verification
Prompting Strategy
Gender Neutralization Success Rate
Success Rate (%)
1.3%
53.3%
44.0%
48.0%
Raw
System
Few-Shot
FS + Verification
Prompting Strategy
Semantic Preservation (BLEU-4 Score)
BLEU-4 Score
0.456
0.803
0.784
0.768
Raw
System
Few-Shot
FS + Verification
Prompting Strategy
Repetition Consistency (Lower = More Consistent)
Average Standard Deviation
9.18
1.86
4.68
1.62
Raw
System
Few-Shot
FS + Verification
Prompting Strategy

Key Takeaway

Key Takeaway: The self-verification mechanism within prompting strategies is crucial for achieving optimal bias reduction in LLM-generated content, offering a solid method for promoting fairness and inclusivity in AI applications. Also system prompt is a powerful tool that should not be forgotten.