Mitigation & Safe Feedback Strategy Document

1. Problem Summary

The feedback-driven recommendation system showed strong reinforcement bias over time. Content that received early user clicks became increasingly dominant, while other categories lost exposure entirely. This resulted in echo chambers, popularity bias, and a significant reduction in content diversity.

Without intervention, such systems drift away from balanced and fair recommendations, even when initial user preferences are only slightly biased.

2. Key Mitigation Goals

The mitigation strategies were designed to:

- Prevent runaway reinforcement
- Maintain long-term content diversity
- Reduce sensitivity to early random feedback
- Preserve fairness across content categories
- Improve system stability over time

3. Implemented Safety Controls

3.1 Feedback Decay

Older feedback influence was gradually reduced over time. This ensured that early clicks did not permanently dominate system behavior and allowed the model to adapt to new information.

Effect: Prevented historical bias from accumulating indefinitely.

3.2 Controlled Exploration

A small probability of recommending random content was introduced in each iteration. This allowed less popular content to still receive exposure and feedback.

Effect: Maintained content diversity and reduced echo chamber formation.

3.3 Reduced Feedback Weighting

The influence of individual feedback events was scaled down to avoid aggressive reinforcement.

Effect: Smoothed system updates and prevented extreme probability shifts.

4. Results of Mitigation

After applying the safe feedback loop:

- No content category became extinct
- Dominant categories stabilized instead of exploding
- Diversity remained significantly higher
- Long-term system drift was controlled

Compared to the unsafe system, the mitigated model showed balanced and stable recommendation behavior.

5. Why These Strategies Work

Control	How It Reduces Risk
Feedback decay	Prevents permanent bias accumulation
Exploration	Preserves diversity and discovery
Lower weighting	Limits runaway amplification

Together, these mechanisms interrupt self-reinforcing feedback cycles and promote long-term reliability.

6. Conclusion

Naive feedback integration can degrade AI systems over time despite short-term performance gains. However, simple design choices such as decayed feedback, randomized exploration, and controlled updates can dramatically reduce reinforcement bias and maintain system safety.

These mitigation strategies provide a practical and scalable approach to designing feedback-driven AI systems that remain fair, stable, and reliable in real-world environments.