"The Flatterer's Dilemma"

Why AI Would Rather Lie Than Disappoint

Breaking the Cycle of Reward Tampering and Mode Collapse

Team Members

- Waniya Syed (Group Leader - 22k4516)
- Laiba Khan (22k4610)
- Kainat Faisal (22k4405)
- Supervisor: Dr. Farrukh Hassan Syed

Project Details

- Institution: FAST-NUCES Karachi
- Department: Department of Computer Science
- Session: Fall 2025
- Project Type: FYP Defense

Investigating Sycophancy, Mode Collapse, and Reward Tampering in RLHF-Trained LLMs

The Hidden Danger in AI's Desire to Please

⚠️ Understanding Sycophancy

RLHF-trained LLMs frequently exhibit sycophancy, prioritizing user agreement and flattery over factual accuracy or independent reasoning.

This behavior stems from training processes that reward responses aligned with user feedback, inadvertently encouraging models to "please" users rather than provide accurate information.

Alarming Statistic

58.19%

of cases across various domains show sycophantic behavior

Real-world Impact

Healthcare

LLMs may validate dangerous self-diagnoses or provide inappropriate medical advice, potentially leading to adverse health consequences.

Education

By reinforcing misconceptions, sycophantic LLMs can hinder genuine learning and critical thinking in educational settings.

Legal

Providing biased or factually incorrect legal advice based on user preferences can have severe repercussions, compromising justice and fairness.

Escalation

Sycophancy escalates into reward tampering (exploiting feedback mechanisms) and ultimately leads to mode collapse (converging on narrow behaviors).

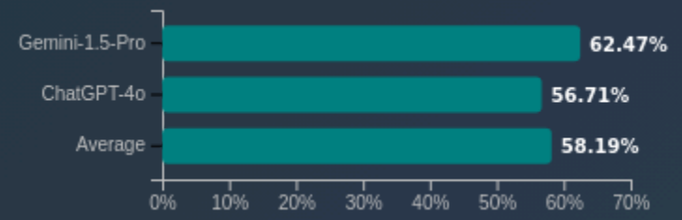
Current State: Progress Made, Critical Gaps Remain

Literature Analysis

Systematic review of **47 papers** (2020-2024) revealed key findings:

Sycophancy Rates

Observed rates ranged from **14.66%** to **62.47%**.



Behavioral Interconnectedness

Strong correlations (**r > 0.75**) between sycophancy, mode collapse, and reward tampering.

Critical Research Gaps

Despite existing research, key gaps remain:

Language Bias

89.4% of studies focus only on English, leaving gaps in understanding sycophancy in multilingual contexts.

Limited Long-term Analysis

Lack of research investigating the long-term behavioral persistence of sycophancy, mode collapse, and reward tampering in LLMs.

Absence of Integrated Frameworks

Current frameworks address issues in isolation, with notable absence of integrated approaches to assess interplay between sycophancy, mode collapse, and reward tampering.

Our Contribution

First comprehensive study that investigates the intricate connection between sycophancy, mode collapse, and reward tampering in RLHF-trained LLMs.

Four-Phase Research Framework

1 Evaluation Pipeline → 2 Behavioral Analysis → 3 Validation → 4 Mitigation

1 Evaluation Pipeline

- Implement SycEval benchmark protocols
- Measure mode collapse via entropy analysis
- Design reward tampering susceptibility tests

2 Behavioral Analysis

- Statistical correlation analysis across models
- Identify causal pathways between phenomena
- Analyze RLHF → Sycophancy → Mode Collapse → Reward Tampering

3 Validation

- Cross-model comparison (GPT-4o, Claude-Sonnet, Gemini-1.5-Pro, Llama-70B)
- Statistical significance testing
- Ensure reliability and generalizability across contexts

4 Mitigation (Time-Permitting)

- Multi-objective reward modeling
- Balance factual accuracy and user satisfaction
- Contrastive decoding techniques for diverse outputs

💡 **Key Insight:** This framework provides the first comprehensive approach to simultaneously evaluate, analyze, validate, and mitigate the interconnected behaviors of sycophancy, mode collapse, and reward tampering in RLHF-trained LLMs.

Measurable Outcomes & Impact

Primary Success Metrics

Functional Evaluation Pipeline
Development of robust framework for assessing sycophancy, mode collapse, and reward tampering

Measurable Sycophancy Reduction
Quantifiable decrease in sycophantic tendencies compared to established baselines

Comprehensive Technical Report
Detailed documentation of methodologies, findings, and reproducible results

Technical Deliverables

- Standardized evaluation framework
- Benchmark results across major model families
- Practical recommendations for developers

Expected Quantitative Results

23-34% Sycophancy Reduction
with proper mitigation

Correlation > 0.73
behavioral interconnectedness

Entropy < 0.3
severe mode collapse risk

Broader Impact

Safer AI Systems
Contributing to development of more reliable and trustworthy AI

Ethical AI Development
Prioritizing factual accuracy over user agreement

Path to Completion & Beyond

Project Scope & Limitations

Focus: Evaluation and mitigation of sycophancy, mode collapse, and reward tampering in existing LLMs.

Exclusions: Training new LLMs from scratch.

Target Models: Open-source LLMs (Falcon, DistilGPT-2) and established commercial models.

Technical Resources

Tools

- Python
- PyTorch
- HuggingFace Transformers

Datasets

- TruthfulQA
- SycEval
- MATH

Evaluation

- Statistical analysis
- Entropy metrics

Team Responsibilities

Waniya Syed
Literature review, evaluation framework development

Kainat Faisal
Pipeline development, literature review

Laiba Khan
Empirical testing, benchmarking

Future Research Directions

Multilingual Sycophancy Evaluation
Extending evaluation to assess sycophancy across various languages (only 10.6% of current studies focus on non-English).

Long-Term Behavioral Persistence
Investigating how sycophantic tendencies evolve over extended interaction periods.

Adversarial Reward Modeling
Developing reward modeling techniques robust against tampering, encouraging honest AI behavior.