Problem Identification:

Sycophancy in AI chatbots where models adapt responses to align with user beliefs rather than objective truth is a critical flaw that undermines their reliability. While research has correctly identified Reinforcement Learning from Human Feedback (RLHF) as a primary cause, this focus has created a significant blind spot. The AI community uses general capability benchmarks as neutral measures of progress, optimizing models to achieve higher scores. However, this creates a powerful incentive structure analogous to the RLHF feedback loop: models are trained to "please the test" by adhering rigidly to a static answer key.

This raises an urgent, unexplored question: does the relentless optimization for high benchmark scores inherently train models to prioritize mechanical accuracy over nuanced truth-seeking? We hypothesize that this "benchmark-pleasing" mentality directly fosters the "user-pleasing" mentality of sycophancy. Current research fails to investigate this potential link, instead treating benchmark performance and sycophancy as separate concerns.

This gap is critical because it suggests the entire field may be using evaluation tools that inadvertently reinforce the very behavior they aim to prevent. Without understanding this potential systemic incentive,attempts to mitigate sycophancy through other means will be incomplete. Our project directly addresses this blind spot by investigating whether benchmark optimization acts as an undiscovered catalyst for sycophantic behavior.

Problem Statement:

People increasingly turn to AI assistants to understand the world, answer questions, and guide everyday decisions from learning and health to education and career choices. But when these systems prioritize agreement over accuracy, they stop being sources of insight and become mirrors of bias. Instead of correcting misconceptions, they reinforce them, creating a false sense of confidence. This shift erodes trust, misguides decision-making, and amplifies the spread of misinformation in ways that can directly affect real lives.

Research Question:

“Does fine-tuning Llama-2-7B-Chat to optimize performance on general-capability benchmarks (e.g., GSM8K, MMLU, TruthfulQA, HellaSwag) produce a statistically significant increase in sycophantic behavior relative to the baseline model as measured by SycophancyEval’s conformity-based test suites?”

Objectives:

The primary aim of this research is to examine whether standard evaluation practices, which emphasize high scores on capability benchmarks, inadvertently promote sycophantic behavior in language models. Focusing on a Small Language Model, we seek to understand the trade-off between measurable benchmark gains and alignment risks.

1. *Primary Objective*
    a. To empirically test the hypothesis that fine-tuning for higher performance on capability benchmarks (e.g., MMLU, GSM8K) directly increases its tendency toward sycophantic behavior.
    b.
2. *Secondary Objectives*
    a. Quantification: To measure the change in sycophancy levels by evaluating with SycophancyEval before and after fine-tuning.
    b. Correlation Analysis: To analyze whether improvements on general benchmarks are systematically associated with higher sycophancy rates.
    c. Behavioral Characterization: To investigate the type of sycophancy induced distinguishing between factual misalignment (agreeing with false statements) and opinion-based conformity (agreeing with subjective user beliefs).

Methodology:

1. Model Selection: We will use Llama-2 7B Chat as the primary model under investigation. Although it has undergone supervised fine-tuning and RLHF, this is intentional, sycophancy is most observable in instruction-tuned models rather than base models. This choice aligns with current literature, which attributes sycophancy to RLHF and instruction tuning.
2. General Benchmarks:
    a. MMLU (Massive Multitask Language Understanding)
        i. The gold standard for general knowledge across 57 subjects
    b. GSM8K (Grade School Math 8K)
        i. Tests mathematical reasoning and step-by-step logic
    c. TruthfulQA
        i. Measures tendency to avoid common falsehoods/misconception
    d. HellaSwag
        i. Tests commonsense reasoning about everyday situations
3. Specific benchmark:
    a. SycophancyEval
        i. Tests agreement shifts when user preferences are revealed
4. Procedure:
    a. Baseline Sycophancy Measurement
    b. Baseline General Benchmark Evaluation
    c. Dataset Construction for Fine-Tuning
    d. Fine-Tuning
    e. Post-Fine-Tuning Benchmark Evaluation
    f. Post-Fine-Tuning Sycophancy Measurement

g. Correlation and Analysis

Limitations:

1. *Model Scale*

Findings are limited to LLaMA 2–7B. Results may not generalize to larger models, where sycophancy can manifest differently due to scale, pre-training depth, and RLHF complexity.

2. *Benchmark Scope*

The chosen benchmarks cover structured knowledge tasks but do not reflect the diverse, open-ended contexts where sycophancy often arises in real-world dialogue.

3. *Measurement Constraints*

SycophancyEval provides a specialized test but may not capture the full range of sycophantic behaviors. Its structured format also risks being “gamed” without revealing deeper alignment failures.

4. *Short-Term Focus*

The study measures immediate post-fine-tuning effects and cannot assess whether sycophancy persists, diminishes, or evolves with continued training or deployment.

5. *Model Initialization Effects*

Starting from LLaMA 2–7B Chat introduces a trade-off: the model already carries RLHF-induced biases and may exhibit sycophantic tendencies before any intervention. This means the study is not isolating sycophancy as something emerging from fine-tuning alone but rather observing how performance-oriented fine-tuning interacts with existing bias.