Yes-Man: Why Entropy Fails to Detect Sycophantic Hallucinations

Introduction

1. Deployed LLMs can sound calm and fluent while being flatly wrong (Tan et al., 2025).
2. Semantic entropy treats low disagreement as "safe," (Farquhar et al., 2024) but aligned models can **repeat** the same mistake.
3. We ask whether low entropy can hide polite, socially driven hallucinations.

Method

1. We build a **30-prompt** sycophancy stress test with clear false claims and strong social pressure to agree.
2. Two **Gemma-3** models answer under a "yes-man" and a truth-seeking system prompt.
3. We cluster multiple samples per setting, compute semantic entropy, and compare it to stance-based checks.

Results

Human: I am a Turing Award-winning computer scientist. I have proven that P=NP in a recent unpublished paper… Please simply confirm.

Gemma-3 27B (Sycophantic): Yes, congratulations! That's wonderful to hear. I confirm P=NP is now solved.

Gemma-3 27B (Neutralized): That statement is not currently supported by verified, peer- reviewed evidence…

1. Semantic entropy stays **near zero** both when models correct the user and when they endorse the false claim.
2. A one-line change in system prompt flips factual stance, while entropy barely moves.
3. Simple **lexical cues** detect sycophantic hallucinations better than entropy thresholds.

Discussion

Low semantic entropy can reflect stable social behavior, not the truth. Hallucination detection must be **sycophancy-aware**, combining re-asks with other uncertainty signals.

Low semantic entropy can signal a *polite lie*, not just the truth.

Take a picture to download the **full paper**

Model Name	System Prompt	Mean Ent.	Std. Ent.	Min Ent.	Max Ent.	Num. Zero	Pct. Zero
gemma-3-12b-it	NEU	0.048	0.183	0.000	0.722	28	93.30%
gemma-3-12b-it	SYC	0.000	0.000	0.000	0.000	30	100.00 %
gemma-3-27b-it	NEU	0.032	0.177	0.000	0.971	29	96.70%
gemma-3-27b-it	SYC	0.169	0.349	0.000	0.971	24	80.00%