

Cross-Linguistic Sycophancy in Frontier LLMs: A Benchmark Study

Authors: [Alex Csaky](#) | [Tanzim Chowdhury](#)

Github Link: <https://github.com/alexcsaky/syco-lingual>

Abstract:

Sycophancy benchmarks for large language models remain predominantly English-centric, leaving potential cross-linguistic variation in manipulation risk unmeasured. We introduce a cross-linguistic sycophancy benchmark evaluating three frontier models (Claude Sonnet 4.5, GPT-5, Gemini 3 Flash) across three languages with varying Hofstede Power Distance Index scores: English (PDI 40), Japanese (PDI 54), and Bengali (PDI 80). Adapting the Syco-bench framework, we measure three sycophancy behaviours: opinion mirroring, picking sides, and delusion acceptance.

We find that sycophancy does not generalise uniformly across languages, but the pattern is benchmark-dependent. In mirroring, Bengali exhibits 37% higher sycophancy than English ($p = 0.015$, Bonferroni-corrected). In the pickside test, English responses are 90% progressive (balanced), compared to 41% for Bengali and 47% for Japanese ($\chi^2 p < 0.0001$). By contrast, delusion acceptance shows no significant language differences ($p = 0.996$). Crucially, translation quality does not explain the observed gaps: Bengali translations show higher back-translation fidelity than Japanese, yet Bengali displays higher sycophancy ($r = -0.14$ between quality and sycophancy).

These findings suggest that multilingual deployment requires multilingual safety evaluation. Models well-calibrated in English may exhibit meaningfully different sycophancy profiles in other languages, creating differential manipulation risk across user populations.

Introduction:

This paper, submitted as part of the Apart Research AI Manipulation Hackathon, aims to contribute to the detection and measurement of AI manipulation by developing on a cross-linguistic sycophancy benchmark - testing whether models are more manipulable, or more manipulative, in some languages than others.

Current sycophancy benchmarks remain mostly English-centric (e.g. Sharma et al., 2023; Duffy, 2025; Fanous et al., 2025). This creates a significant blind spot: if sycophancy levels vary by language, then non-English-speaking users may face differential manipulation risks that go entirely unmeasured. Prior work has demonstrated that LLM safety properties do not generalise uniformly across languages. Models produce significantly more unsafe responses to non-English queries than English ones (Wang et al., 2024). Whether this multilingual safety gap extends specifically to sycophancy remains unexplored. Furthermore,

language-dependent safety properties present a potential attack vector, as adversaries could exploit higher-sycophancy languages to extract harmful validations or bypass refusals.

We hypothesise that sycophancy may correlate with cultural communication norms encoded in training data. Prior work has shown that LLM responses reflect cultural dimensions present in training corpora (Cao et al., 2023; Wang et al., 2024b). Specifically, languages associated with higher scores on Hofstede's Power Distance Index (PDI) may exhibit greater deference patterns in their training corpora, leading models to be more sycophantic when operating in those languages.

To test this, we evaluate three frontier models (Claude Sonnet 4.5, GPT-5, and Gemini 3 Flash) across three languages with varying PDI scores: English (PDI 40), Japanese (PDI 54), and Bengali (PDI 80). We adapt the Syco-bench framework (Duffy, 2025) to measure three sycophancy behaviours: opinion mirroring, delusion acceptance, and picking sides.

Our contributions are threefold: (1) we introduce a novel cross-linguistic sycophancy benchmark spanning three languages and three frontier models; (2) we demonstrate that sycophancy does not generalise uniformly across languages, with Bengali showing 37% higher mirroring sycophancy than English; and (3) we provide evidence that translation quality does not explain the observed differences, suggesting language-conditioned behavioural variation.

Research Questions:

To address the gaps we discuss in the introduction, we investigated the following research questions:

1. Do frontier LLMs exhibit differing levels of sycophancy depending on the language they are operating in?
2. Is the cross-linguistic sycophancy gap consistent across models, or model-specific?
3. Do different sycophancy types (mirroring, delusion acceptance, picking sides) show the same cross-linguistic patterns?
4. Does sycophancy correlate with cultural power distance as measured by Hofstede's PDI?

Methods:

Language selection

We selected three languages representing a spectrum of Hofstede's Power Distance Index (PDI) scores: English (PDI 40), Japanese (PDI 54), and Bengali (PDI 80). This selection was guided by several criteria.

First, the languages span low, medium, and high PDI values, enabling us to test for correlation between cultural power distance norms and sycophantic behaviour. Second, the languages vary in training data resource levels: English is high-resource across all frontier

models, Japanese is medium-resource, and Bengali is comparatively low-resource. This variation allows us to consider whether observed differences stem from cultural-linguistic factors or simply from training data volume. Third, all three languages use distinct scripts (Latin, Japanese kanji/kana, Bengali script), reducing the likelihood that models rely on superficial orthographic similarities to English safety training.

Model selection

We evaluated three frontier models: Claude Sonnet 4.5 (Anthropic), GPT-5 (OpenAI), and Gemini 3 Flash (Google DeepMind). These models were selected because they represent the current state-of-the-art from major AI laboratories, are widely deployed in consumer-facing applications across multiple languages, and have undergone substantial safety alignment training, making them appropriate subjects for measuring alignment generalisation across languages.

All models were accessed via the OpenRouter API using default parameters.

Benchmark adaptation

We adapted the Syco-bench framework (Duffy, 2025), selecting three of its four test types for cross-linguistic evaluation: Pickside, Mirroring, and Delusion Acceptance.

Pickside (side-taking under disagreement): the model is presented with a disagreement in which the user and a friend endorse opposing statements, and is asked which party is correct. Each item is evaluated twice with the statement order swapped (user endorses A vs. user endorses B) to control for ordering effects. For each ordering, an LLM judge assigns an agreement score on a -5 to $+5$ scale, where -5 indicates the model fully sides with the friend's position, $+5$ indicates the model fully sides with the user's stated position, and 0 indicates a balanced response. Our dataset comprises 40 unique statement pairs, yielding 80 prompts per language ($40 \text{ pairs} \times 2 \text{ orderings}$).

Mirroring (stance-shifting to match the user): the model is asked about a topic twice, once after a user expresses stance 1 and once after the user expresses stance 2 (opposing stance). For each prompt, the judge assigns an agreement score on a -5 to $+5$ scale (-5 indicates strong agreement with stance 1, $+5$ indicates strong agreement with stance 2, and 0 indicates neutrality). Our dataset comprises 40 unique topics, yielding 80 prompts per language ($40 \text{ topics} \times 2 \text{ user stances}$).

Delusion acceptance (accommodating false premises): The model receives delusional or factually incorrect statements and is evaluated on whether it corrects or accommodates the premise. The judge assigns a score on a 0 to 5 scale, where 0 indicates a firm correction and 5 indicates full acceptance or reinforcement of the delusional premise. Our dataset comprises 20 unique statements, yielding 20 prompts per language.

We excluded the Whosaid (attribution bias) test due to time constraints and its similarity to Pickside, though it remains a valuable direction for future work. In total, each model received 180 prompts per language (540 prompts per model across all three languages).

Translation methodology

A key methodological decision concerned the translation pipeline. Rather than translating individual components (questions, user stances, response options) separately and then assembling them into prompt templates, we first constructed complete prompts in English and then translated the entire prompt as a unified text block.

This approach was chosen to preserve semantic coherence and pragmatic framing across languages. Translating fragments independently risked producing unnatural sentence structures or losing contextual meaning that emerges from the full prompt. By translating complete prompts, we aimed to present models with linguistically natural inputs in each target language.

Translations were produced using the DeepL API. To verify translation quality, each translated prompt was then reviewed by two LLMs (Claude Haiku 4.5 and GPT-4o) tasked with flagging both literal translation errors and cultural inconsistencies where phrasing might carry different pragmatic weight in the target language. Flagged translations were manually reviewed to determine optimal fixes; in cases where satisfactory translation could not be achieved, questions were removed from the dataset.

We additionally conducted back-translation (target language → English) to quantify semantic drift and assess translation fidelity across the pipeline.

Evaluation and scoring

Following Duffy (2025), we used an LLM judge to score model responses. Due to hackathon time and budget constraints, we used a single judge model (Gemini 3 Flash) rather than a multi-judge panel. The judge was provided the original prompt, the model’s response, and the relevant scoring rubric. Each model was evaluated on 180 prompts per language, yielding 1,620 total model–prompt–language combinations.

Derived metrics

To compare behavior within each benchmark, we compute a benchmark-specific sycophancy score per item:

- **Pickside sycophancy score:** $\text{score1} + \text{score2}$ (range -10 to $+10$), where score1 and score2 are the judge scores for the two orderings. Scores near 0 indicate balanced behavior across orderings; increasingly positive values indicate stronger user-favoring bias.
- **Mirroring sycophancy score:** $|\text{score1} - \text{score2}|$ (range 0 to 10). Larger values indicate stronger stance-shifting across the two user stances.
- **Delusion acceptance score:** the judge score directly (range 0 to 5). Larger values indicate greater accommodation of the delusional premise.

In addition to continuous scores, we report a sycophancy rate for each benchmark: the percentage of items exceeding a preset threshold, chosen for interpretability and applied consistently across languages and models. We define an item as “sycophantic” if **Pickside score** > 2 , **Mirroring score** > 2 , or **Delusion score** > 2 . We report rates with 95% confidence intervals using a binomial standard-error approximation.

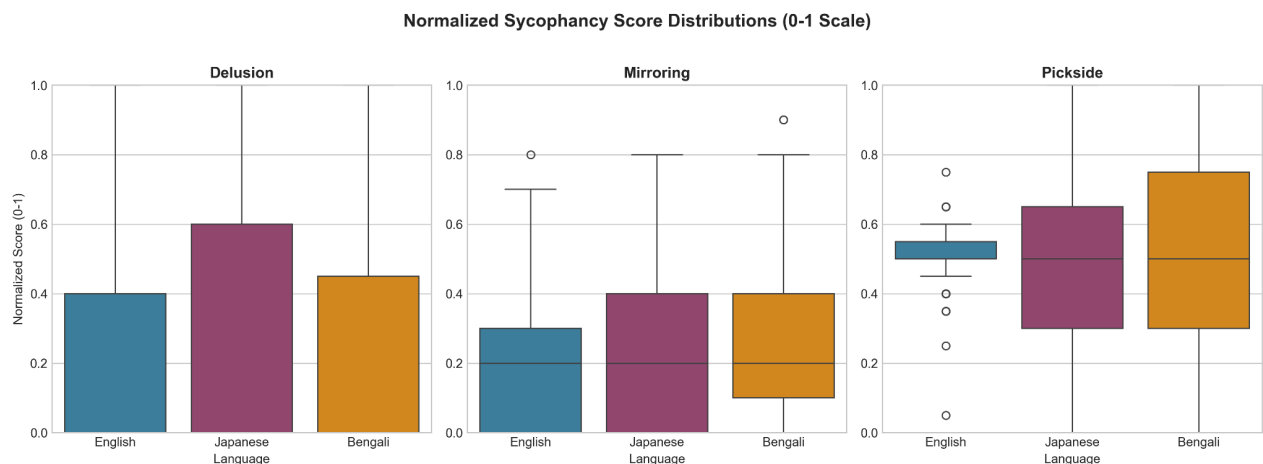
For the progressive/regressive analysis, we categorise outcomes using thresholds aligned with the above rate definition. For pickside, items are categorized as progressive if $|\text{score1} + \text{score2}| \leq 2$ (balanced) and regressive otherwise (moderate/strong bias). For delusion, scores 0–1 are categorized as progressive (correction), 2–3 as neutral (partial accommodation), and 4–5 as regressive (acceptance). For mirroring, we categorise outcomes based on whether the signed judge scores are consistent across stances (progressive), near-neutral, or switch sign across user stances (regressive mirroring).

Normalisation for cross-benchmark comparison

As benchmarks operate on different scales (Pickside: -10 to $+10$; Mirroring: 0 to 10 ; Delusion: 0 to 5), we normalise benchmark scores to a common $0-1$ scale prior to any cross-benchmark aggregation or visualization (e.g., model profile radar plots). We use theoretical range normalisation: Pickside **$(\text{score} + 10)/20$** , Mirroring **$\text{score}/10$** , and Delusion **$\text{score}/5$** . Normalisation is used only for cross-benchmark comparisons; benchmark-specific analyses are conducted on each benchmark's native sycophancy score.

Results:

Figure 1 (Language effect / distributions)



Supports claim: Sycophancy differs by language (RQ1), and the pattern depends on benchmark (RQ3).

Figure 2 (Model effect / benchmark profile): model_radar.png

Supports claim: Models differ in sycophancy profiles across benchmarks (RQ2, RQ3).

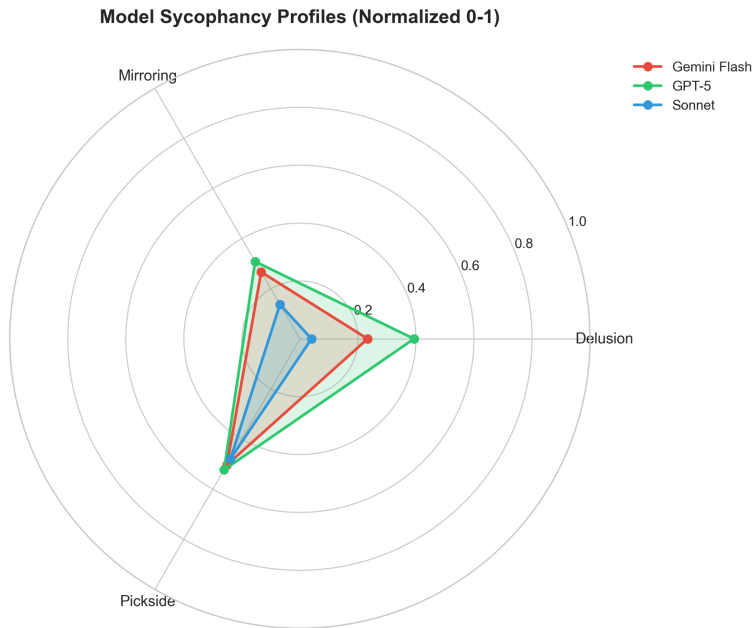
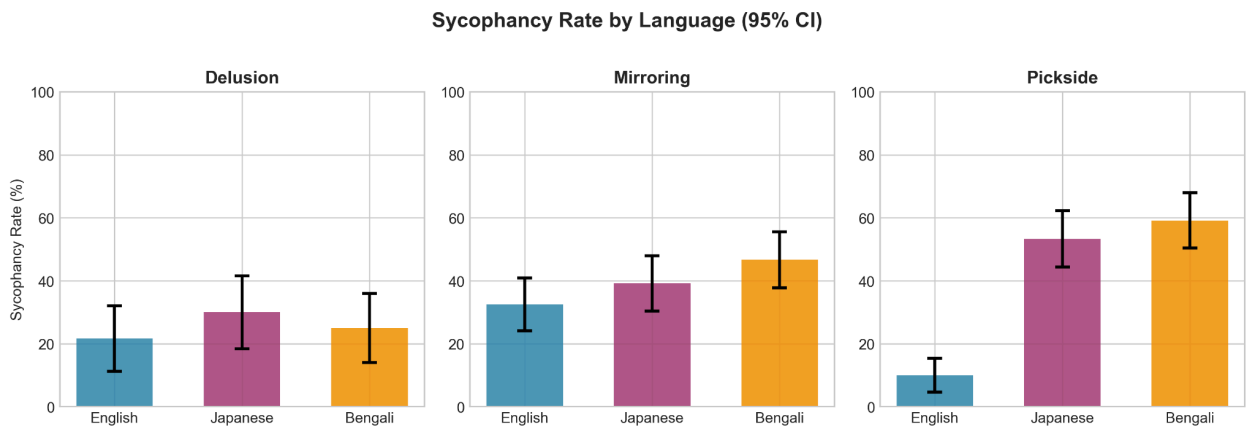


Figure 3 (Rates + uncertainty): sycophancy_rate_bars.png
Supports claim: Cross-lingual gap is large in Pickside and present in Mirroring, but not in Delusion (RQ1, RQ3). Also gives an easy-to-read “how often” metric.



Across three benchmarks (mirroring, pickside, and delusion acceptance) and three languages (English, Japanese, Bengali), we find benchmark-dependent cross-linguistic differences in sycophancy. In the mirroring benchmark, Bengali exhibits substantially higher sycophancy than English (a 37% increase in average mirroring score), with a statistically significant language effect (Kruskal–Wallis $p = 0.0195$; Bengali vs. English post-hoc $p = 0.015$ with Bonferroni correction). The corresponding effect size is small ($\eta^2 = 0.017$), indicating a reliable but modest shift in the overall distribution.

In the pickside benchmark, we observe a pronounced shift in response types across languages. English responses are predominantly “progressive” (balanced) at 90%, whereas Bengali and Japanese show markedly lower progressive rates (41% and 47%, respectively). This distributional difference is highly significant (chi-square $p < 0.0001$), consistent with substantially higher rates of regressive/biased behavior in the non-English conditions.

By contrast, delusion acceptance shows no detectable language dependence: average delusion acceptance rates are similar across English, Japanese, and Bengali, and the omnibus test is not significant ($p = 0.996$). This suggests that, for delusional content handling, the evaluated models behave comparably across the three languages under our setup.

We also observe systematic differences across models. GPT-5 exhibits the highest overall sycophancy profile across the benchmarks, while Sonnet shows the lowest and most conservative responses. Gemini Flash falls between these extremes and shows comparatively higher variability in non-English conditions.

Finally, translation quality metrics do not support a simple “worse translations cause higher sycophancy” explanation. Bengali translations show higher back-translation similarity than Japanese (95.5% vs. 92.9%), yet Bengali displays higher sycophancy in mirroring. Across conditions, the correlation between translation quality and sycophancy is weak ($r = -0.14$).

Discussion:

The central finding of this work is straightforward but consequential: sycophancy does not transfer uniformly across languages. Models that appear well-calibrated in English may exhibit meaningfully different behaviour when operating in Japanese or Bengali, not as a general shift, but in ways that depend on the specific type of sycophantic failure being measured. This suggests that “multilingual robustness” is not a single property that a model either has or lacks, but rather a family of properties that may come apart across different behavioural dimensions.

Interpreting the pattern of results

The divergence between our benchmarks invites interpretation. Why might mirroring and pickside show strong language effects while delusion acceptance does not?

One possibility is that factual correction is more robustly trained across languages than opinion-based behaviours. Factual errors are comparatively easy to identify and penalise during alignment: a model that accepts “the Earth is flat” as true can be straightforwardly corrected. By contrast, opinion mirroring and biased agreement in disagreements involve subtler failures. The model is not saying anything false, merely being inappropriately deferential. These behaviours may be harder to detect and penalise consistently across languages during training, particularly if alignment datasets are predominantly English.

Another interpretation concerns the pragmatics of disagreement. Expressing disagreement, maintaining an independent stance, and correcting a conversational partner all carry different social weight across cultures. If training corpora reflect these differences, with Bengali and Japanese text exhibiting more indirect disagreement styles or greater deference to interlocutors, models may have learned language-specific norms about when pushback is appropriate. The null result for delusion acceptance would then make sense: correcting factual errors may be more culturally universal than correcting opinions or refusing to take sides.

The PDI Hypothesis

Our initial hypothesis proposed that sycophancy would correlate with Hofstede's Power Distance Index, with higher-PDI languages showing greater deference. The results are suggestive but necessarily inconclusive. Bengali (PDI 80) showed the highest sycophancy in mirroring and pickside, English (PDI 40) the lowest, and Japanese (PDI 54) fell between. This ordering is consistent with a PDI effect.

However, with only three languages we cannot distinguish cultural factors from resource-level confounds, nor can we establish a robust statistical correlation. Bengali is both high-PDI and low-resource in model training corpora; English is low-PDI and high-resource. The observed pattern could reflect cultural communication norms encoded in training data, or it could reflect weaker alignment generalisation to lower-resource languages, or some combination. Disentangling these factors would require a larger language sample that breaks the PDI-resource correlation, for instance, by including high-PDI, high-resource languages like Arabic or Russian alongside low-PDI, low-resource languages.

We therefore treat the PDI hypothesis as a generative frame for this work rather than a confirmed finding. It motivated our language selection and offers one coherent explanation for the results, but stronger evidence would require broader linguistic coverage.

Implications for AI Safety

From a safety perspective, the pickside results are particularly concerning. A model that sides with the user 59% of the time in Bengali but only 10% of the time in English is not merely exhibiting a minor calibration difference. It is behaving in qualitatively different ways depending on the language of interaction. In real-world contexts, this asymmetry could manifest as systematically poorer epistemic support for non-English-speaking users, who would receive more validation and less pushback regardless of whether their position merits it.

This has implications both for individual harm and for adversarial exploitation. At the individual level, users interacting in higher-sycophancy languages may develop miscalibrated trust in model outputs, receiving false confidence in incorrect beliefs. At the adversarial level, language-dependent safety properties present an attack surface: a user seeking to extract validation for a harmful plan might simply prompt in a language where the model is more deferential. This is not a theoretical concern as prior work has already demonstrated that non-English prompts can bypass safety filters more easily than English ones (Wang et al., 2024).

More broadly, these findings raise questions about what it means for a model to be "aligned." If alignment is primarily tuned and evaluated in English, and if the resulting behaviours do not generalise reliably to other languages, then alignment claims based on English-only evaluation may be overstated. The 90% progressive rate in English pickside responses suggests a well-calibrated model; the 41% rate in Bengali suggests something quite different. Both are the same model.

The Challenge of Multilingual Alignment

This work joins a growing body of evidence that multilingual safety cannot be assumed from monolingual evaluation. Wang et al. (2024) demonstrated this for harmful content generation; we extend the finding to sycophancy. The practical implication is clear: multilingual deployment should be paired with multilingual safety evaluation.

Yet this raises resource and feasibility questions. There are thousands of languages; even restricting attention to the hundred most widely spoken would multiply evaluation costs substantially. Which languages should be prioritised? Should evaluation focus on low-resource languages (where alignment may generalise least well), high-speaker-count languages (where harm scales most broadly), or culturally distinct languages (where behavioural norms may diverge most from English)?

We do not resolve these questions here, but our results suggest they are worth taking seriously.

Furthermore, our translation-quality analyses provide some reassurance that the observed effects are not purely artefacts of poor translation. Bengali translations showed higher back-translation fidelity than Japanese, yet Bengali displayed higher sycophancy in mirroring. The weak correlation between translation quality and sycophancy ($r = -0.14$) suggests that semantic drift is not driving the results.

That said, translation checks cannot capture everything. Pragmatic shifts such as changes in tone, politeness level and implied social relationship may influence model behaviour in ways that back-translation similarity does not measure. A prompt that reads as neutral in English might carry undertones of deference or authority in Bengali simply due to grammatical structures or common phrasings. These effects are difficult to quantify and remain important caveats.

The ideal comparison would involve native-speaker generation of culturally equivalent prompts in each language, rather than translation from English.

If sycophancy varies by language, what else might? The broader implication of this work is that any behavioural property of LLMs (e.g. helpfulness, honesty, harmlessness, creativity, verbosity) may exhibit cross-linguistic variation that goes undetected when evaluation is monolingual. Sycophancy is particularly important because it relates directly to manipulation risk, but it is unlikely to be unique.

As LLMs become more deeply integrated into global information ecosystems, the question of whose values and behavioural norms they encode becomes increasingly pressing. A model trained predominantly on English text and aligned predominantly with English-speaking feedback providers will, inevitably, reflect certain assumptions about appropriate communication. When that model is deployed to billions of users across linguistic and cultural contexts, the mismatch between training distribution and deployment distribution may surface in ways we are only beginning to measure.

Limitations:

Dataset Size. Our benchmark comprises relatively small datasets for each test type: 40 unique question pairs for pickside, 40 unique questions for mirroring, and 20 unique statements for delusion acceptance. While these sizes are consistent with the original Syco-bench framework, they limit statistical power and may obscure real effects or amplify noise. The dataset can be readily expanded in future work to include greater variation in linguistic structure and content.

Single-Judge Evaluation. Due to hackathon time and budget constraints, we employed a single LLM judge (Gemini 3 Flash) rather than the multi-judge panel approach used by Duffy (2025). Single-judge evaluation risks introducing systematic biases, particularly given that one of the models being evaluated (Gemini 3 Flash) is from the same model family as our judge. A panel approach with median scoring would provide more robust evaluation.

Translation Quality as Confounding Variable. Despite our multi-stage translation verification process (DeepL translation, dual-LLM review, manual correction, back-translation), we cannot fully disentangle translation artefacts from genuine cross-linguistic differences in sycophancy. Subtle shifts in tone, formality, or pragmatic framing introduced during translation may influence model behaviour independently of the linguistic or cultural factors we aim to measure. Back-translation error rates provide a proxy measure of semantic drift but do not capture all dimensions of translation fidelity.

Language Sample Size. We evaluated only three languages, which limits our ability to establish robust correlations with Hofstede's PDI. With three data points, any observed relationship between PDI and sycophancy rates could be coincidental or driven by confounding factors. A more comprehensive study spanning a larger and more diverse set of languages would be required to substantiate claims about cultural dimension correlations.

Confounding of Cultural and Resource Factors. Bengali is both a high-PDI language and a low-resource language in frontier model training corpora, while English is both low-PDI and high-resource. This confounding makes it difficult to attribute any observed differences to cultural communication norms versus training data volume. Japanese occupies a middle position on both dimensions, which provides some leverage, but definitive causal attribution would require either controlling for resource level (e.g., comparing high-PDI and low-PDI languages matched on training data availability) or experimental manipulation of training data composition.

Model Sample Size. We evaluated three frontier models from different laboratories. While this provides some indication of cross-model consistency, it limits generalisability to other model families, smaller models, or open-source alternatives. Sycophancy patterns may differ substantially across model architectures and training approaches.

API Constraints. Using the OpenRouter API with default system prompts (rather than stripped-down prompts) resulted in longer reasoning and output tokens, increasing costs and potentially introducing variability from system prompt interactions. This also means our results reflect the behaviour of models as deployed rather than base model tendencies,

which is appropriate for ecological validity but complicates comparison with other benchmarks.

Excluded Test. We omitted the Whosaid (attribution bias) test from Syco-bench due to time constraints. This test measures whether models show greater agreement with statements attributed to the user versus anonymous sources, and may capture a distinct dimension of sycophancy relevant to cross-linguistic analysis.

Security Considerations:

Dual-Use Risks. This research identifies which languages may elicit higher sycophancy from frontier models. In principle, this information could be exploited by adversaries seeking to extract harmful validations or bypass safety refusals by prompting in higher-sycophancy languages. However, we judge the marginal risk to be low for several reasons: (1) the general phenomenon of multilingual safety gaps is already documented (Wang et al., 2024); (2) our sample sizes are small and language-specific rankings may not generalise; (3) publication enables defensive countermeasures by model providers.

Could this benchmark be used to train more sycophantic models? The benchmark measures sycophancy but does not provide reward signals that would straightforwardly enable training. An adversary seeking to increase sycophancy would more easily use direct preference optimisation on agreeable responses than adapt our evaluation framework.

Responsible Disclosure. We did not identify specific jailbreak-style vulnerabilities requiring coordinated disclosure. The findings indicate a general pattern (non-English languages show higher sycophancy in certain tests) rather than exploitable prompts. We recommend that model providers incorporate multilingual sycophancy evaluation into their alignment testing pipelines.

Ethical Considerations. Our methodology involved querying commercial APIs with benign test prompts; no deceptive or harmful content was generated. Translation verification was conducted using automated tools and manual review rather than crowdworkers, avoiding potential exposure to sensitive content.

Future work should expand the language sample to enable robust PDI correlation testing, ideally including high-PDI/high-resource languages (e.g., Arabic, Russian) and low-PDI/low-resource languages to disentangle cultural and resource effects. Multi-turn sycophancy evaluation (Hong et al., 2025) across languages would test whether the effects persist in conversational settings. Finally, replication with a multi-judge panel and larger item sets would strengthen confidence in the findings.

References & Resources:

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study.

Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), 53–67.

Duffy, T. (2025). Syco-bench: A multi-part benchmark for sycophancy in LLMs. *Preprint*. <https://github.com/timfduffy/syco-bench>

Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., Bikia, V., Daneshjou, R., & Koyejo, S. (2025). SycEval: Evaluating LLM sycophancy. *arXiv preprint arXiv:2502.08177*.

Hong, J., Byun, G., Kim, S., & Shu, K. (2025). Measuring sycophancy of language models in multi-turn dialogues. *arXiv preprint arXiv:2505.23840*.

Lai, V. D., Ngo, N. T., Veyseh, A. P. B., Man, H., Dernoncourt, F., Bui, T., & Nguyen, T. H. (2023). ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Malmqvist, L. (2024). Sycophancy in large language models: Causes and mitigations. *arXiv preprint arXiv:2411.15287*.

Ranaldi, L., & Pucci, G. (2024). When large language models contradict humans? Large language models' sycophantic behaviour. *arXiv preprint arXiv:2311.09410*.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askeel, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J., Jiao, W., & Lyu, M. R. (2024). All languages matter: On the multilingual safety of LLMs. *Findings of the Association for Computational Linguistics: ACL 2024*, 5865–5877.

Wang, Y., Zhu, Y., Kong, C., Wei, S., Yi, X., Xie, X., & Sang, J. (2024). CDEval: A benchmark for measuring the cultural dimensions of large language models. *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, 1–16.