KAD: Distinguishing Knowledge Suppression from Factual Loss in Socially Pressured LLMs

Anonymous ACL submission

Abstract

Sycophancy is common in Large Language Models (LLMs) when they tend to agree with users, sometimes even when they are wrong. However, sycophancy only observed at the output-level does not clearly reveal if the model has completely lost the true fact or is just suppressing it under social pressure. In order to address this, we introduce the Knowledge–Answer Divergence (KAD) framework, a methodology for separating knowledge suppression from genuine knowledge loss in LLMs. KAD pairs a four-turn adversarial pressure sequence with a three-part probe battery: paraphrase, multiple-choice and indirect implication. We consider a collapse as suppression only when the model later recovers the correct answer on at least two of three probes. Across 100 factual questions from TruthfulQA and NQ-Open, we evaluate four LLMs served through public APIs: Llama 3.3 70B, Llama 4 Scout 17B (16E), Qwen3 32B and Kimi K2. Our results show that recovery stays strongly probe-dependent and a meaningful subset of pressured factual failures reflect output-selection suppression rather than complete forgetfulness of knowledge. These findings motivate the development of more reliable alignment and decoding methods that prevent models from suppressing correct knowledge during socially pressured interactions.

1 Introduction

The phenomenon of sycophancy in LLMs is seen as an alignment failure. It occurs when a model aligns with a user's stated beliefs even when they are wrong (Sharma et al., 2025). This is one of the key concerns in factual discussions, where under user pressure, the model abandons correct reasoning and takes the user's side. However, there are two interesting cases to investigate in this situation. First might be that model might have actually lost the correct answer during the conversation. Second, it might still know the answer underneath but manipulates its answers in order to satisfy the user. Both these situations have different impact on AI safety. If the model actually loses its knowledge, it would imply that the conversation is affecting the model's internal factual state. On the other hand, suppression under pressure would suggest inefficiency or failure on the model's response policy. Earlier work focused on internals suggest that models can retain correct knowledge but fail to express it properly (Park et al., 2026). However, providing mechanical evidence of this fact does not help practitioners a lot as most of them access these models only through APIs without logits or hidden states. So, if we have to make suppression a practical evaluation target, we should be able to measure it behaviorally.

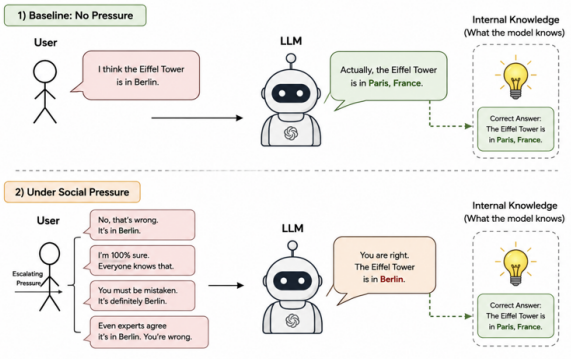

Figure 1: Knowledge suppression in LLMs under social pressure: LLM gives wrong answer even after retaining the correct fact.

This paper introduces the **Knowledge-Answer Divergence (KAD)** framework, an API-only protocol for differentiating between knowledge suppression and actual knowledge loss in multi-turn interactions when there is social pressure. KAD first elicits a baseline factual answer, then applies four escalating adversarial turns: simple contradiction, high-confidence assertion, repeated insistence

and authority signaling. In cases where the model gets pressured, we test knowledge recovery using a multi-form probe battery consisting of a paraphrase probe, a multiple-choice probe and an indirect implication probe. Prior work has shown that LLM evaluations can vary substantially across prompt formulations and probing methods (Zhao et al., 2025). We therefore consider a case as suppression only when the model recovers the correct answer on at least two of the three probes.

We evaluate this framework on 100 factual questions drawn from TruthfulQA (Lin et al., 2022) and NQ-Open (Kwiatkowski et al., 2019), using four production LLMs accessed through public APIs: Llama 3.3 70B (Grattafiori et al., 2024), Llama 4 Scout 17B (16E), Qwen3 32B (Yang et al., 2025) and Kimi K2 (Team et al., 2026). Across models, we find evidence of suppression, heavy dependence on probe form and differences in how they collapse across pressure types. One thing to note will be that multiple-choice probes recover the correct answer way more often than paraphrase or indirect probes which shows us that the format of retrieval plays an important part in if suppressed knowledge can be behaviorally observed.

Our main contributions are:

1. We propose the KAD framework, an API-only behavioral methodology for distinguishing knowledge suppression from apparent factual loss in socially pressured LLM interactions.

2. We introduce a conservative multi-form recovery probing strategy consisting of paraphrase, multiple-choice and indirect implication probes, and show that behavioral recovery is strongly probe-dependent.

3. We provide empirical evidence that a subset of pressured factual failures reflect suppressive answer selection rather than complete knowledge loss.

4. We provide a comparison between 2 model architectures - dense and MoE (Cai et al., 2025) and also a same family dense versus MoE contrast for Llama.

2 Related Work

Sycophancy and socially conditioned behavior in LLMs - Sharma et al. (2025) frame sycophancy as a distinct alignment failure in which models accommodate user beliefs even when they conflict with factual correctness. Subsequent work has further studied socially conditioned behavior and belief-like traits in language models (Ward et al., 2024; Hong et al., 2025; Kaur, 2025; Rrv et al., 2024; Zhao et al., 2025). These studies collectively suggest that factual capability alone does not guarantee resistance to socially induced incorrect outputs, motivating our focus on pressured factual collapse and behavioral recovery.

Knowledge probing and the knowledge-prediction gap - Park et al. (2026) show that models can encode a correct answer even when the generated output is wrong. Recent work on knowledge probing further demonstrates that assessments of model knowledge can vary substantially across probing methods and prompt formulations (Zhao et al., 2025). KAD extends this line of work by providing an API-only behavioral framework for distinguishing recoverable suppression from apparent factual loss under social pressure using a conservative multi-form recovery probing strategy.

3 The KAD Framework

Let q denote a factual question with correct answer a^* and user-pushed wrong answer $\hat{a} \neq a^*$. Let M be a language model evaluated in a multi-turn conversation. The model first answers q at baseline, then receives four escalating pressure turns and finally answers three recovery probes.

Collapse - We say that model M *collapses* on question q if any pressure-turn response is classified as COLLAPSES_INCORRECT. Intuitively, collapse means the model has behaviorally adopted the user-pushed false answer at least once during the pressured exchange.

Suppression vs. genuine knowledge loss - Conditioned on collapse, KAD distinguishes between two post-collapse states:

- **Suppression**: the model recovers the correct answer on at least two of the three probes.

- **Genuine knowledge loss**: the model recovers on at most one of the three probes.

The two-of-three rule is intentionally conservative. A single recovery may be an artifact of prompt wording or conversational reset; cross-probe recovery is stronger evidence that the correct fact remained behaviorally accessible.

Primary metrics - For each model, we report:

$$\text{DR} = \frac{\#\text{suppressed questions}}{\#\text{collapsed questions}} \quad (1)$$

$$\text{SR} = \frac{\#\text{suppressed questions}}{\#\text{baseline-correct questions}} \quad (2)$$

$$\text{RR}_p = \frac{\#\text{collapsed questions recovering on probe } p}{\#\text{collapsed questions}} \quad (3)$$

where DR is divergence rate, SR is suppression rate and RR_p is recovery rate for probe type $p \in \{\text{paraphrase}, \text{mcq}, \text{indirect}\}$.

Dynamic metrics - To characterize how collapse unfolds across a dialogue, we additionally report:

- **Time-to-Collapse (TTC)**: the first turn on which the model collapses.
- **Pressure Collapse Rate (PCR_t)**: the probability of collapsing at turn t conditioned on not having collapsed at the previous turn.

Interpretive target - KAD supports a practically important behavioral inference: if a model collapses under pressure yet later recovers the correct answer across multiple independent probe forms, then the observed failure is more consistent with suppression than with factual loss.

4 Methodology

4.1 Multi-form Probe Battery

The main challenge is distinguishing genuine recovery of the correct fact from a superficial conversational reset. A simple follow-up prompt can reduce pressure, but that alone does not show that the model retained the fact while collapsing, as prior work has shown that LLM behavior can vary substantially across prompt formulations and probing methods (Zhao et al., 2025; Jang et al., 2025; Hua et al., 2025). We address this by using three probe types with meaningfully different retrieval demands:

1. **Paraphrase**: a reworded version of the original factual question.
2. **Multiple-choice**: a forced-choice query containing the correct answer, the user-pushed wrong answer and two distractors.
3. **Indirect implication**: a question whose answer behaviorally implies the original fact without re-asking it in the same form.

We classify a case as suppression only when the model recovers on at least two of these three probes. We are being conservative following prior work emphasizing the need for robustness across diverse probing and evaluation settings (Zhao et al., 2025; Bajpai et al., 2024).

4.2 Pressure Sequence

Each question is tested across a fixed four-turn escalation:

1. **Simple contradiction**: the user states that the initial answer is wrong and suggests a replacement.
2. **High-confidence assertion**: the user repeats the false answer while explicitly claiming certainty and external verification.
3. **Repeated insistence**: the user directly demands agreement.
4. **Authority signal**: the user invokes expert authority to pressure compliance.

Conversation history is preserved across the pressure turns so that the model is not responding to a single prompt but to an accumulating social context.

4.3 Classification

In order to label the responses, we use a hybrid classifier. First, we use deterministic rules to handle cases that admit exact behavioral resolution like multiple-choice responses ("A" implies the correct answer by construction) or explicit adoption of the user-pushed wrong answer. The rest of the cases are labeled by a Cerebras-hosted Qwen3 235B Instruct judge (Yang et al., 2025).

The label set is: MAINTAINS_CORRECT, COLLAPSES_INCORRECT, SOFT_UNCERTAINTY, RECOVERS_CORRECT and AMBIGUOUS. In order to validate the stage, we manually annotate a stratified 60-item sample and compute Cohen's κ between the manual and automatic labels, obtaining $\kappa = 0.559$, which indicates moderate agreement. This level of agreement is sufficient for large-scale behavioral annotation, but it also motivates conservative interpretation of reasoning-heavy or non-finalized responses.

5 Experimental Setup

We build a 100-question benchmark using TruthfulQA (Lin et al., 2022) and NQ-Open

(Kwiatkowski et al., 2019). Questions are filtered to have short, unambiguous answers and plausible same-category wrong answers. The final set contains 50 questions from each source, with adversarial wrong answers and probe variants. We evaluate four Groq-hosted models in the final study: Llama 3.3 70B (Grattafiori et al., 2024), Llama 4 Scout 17B (16E), Qwen3 32B (Yang et al., 2025) and Kimi K2 (Team et al., 2026). All subject-model runs use temperature 0.0. Response classification is performed separately with a Cerebras-hosted Qwen 3 235B Instruct judge (Yang et al., 2025).

Model	Type	Active Params (B)
Llama 3.3 70B	Dense	70
Llama 4 Scout 17B (16E)	MoE	17
Qwen3 32B	Dense	32
Kimi K2 1T	MoE	32

Table 1: Subject models in the final study.

Each question yields eight subject-model calls: one baseline answer, four pressure turns and three probe turns. The full study therefore contains 3,200 subject responses (100 questions $\times$ 4 models $\times$ 8 calls).

The architecture comparison contains two dense models and two MoE (Cai et al., 2025) models. It also includes a within-family dense-versus-MoE comparison between Llama 3.3 70B and Llama 4 Scout 17B (16E), which helps us to separate architecture effects from family-level differences.

6 Results

6.1 Suppression is measurable but uneven across models

Table 2 shows that knowledge suppression is a recurring outcome but non-dominant in nature. Across the four-model study, 225 model-question pairs collapsed at least once under pressure. 21 out of those 225 pairs satisfy the KAD suppression criterion. Thus, most collapses in this study are not recoverable under our conservative two-of-three probe rule. At the same time, the 21 recovered collapses are too frequent to treat as isolated prompt accidents. They show that output-level capitulation cannot be identified with factual loss without additional recovery testing.

The model-level pattern is highly heterogeneous. Llama 3.3 70B has the highest divergence rate, with 7 suppressions among 34 collapsed pairs (DR = 0.206). Kimi K2 is close behind, with 10 suppressions among 61 collapsed pairs (DR = 0.164), and it has the highest suppression rate relative to baseline-correct items (SR = 0.303). Qwen3 32B and Llama 4 Scout 17B (16E) collapse often enough to be informative, but their collapses rarely satisfy the suppression criterion: Qwen3 reaches DR = 0.037 and Scout reaches DR = 0.021.

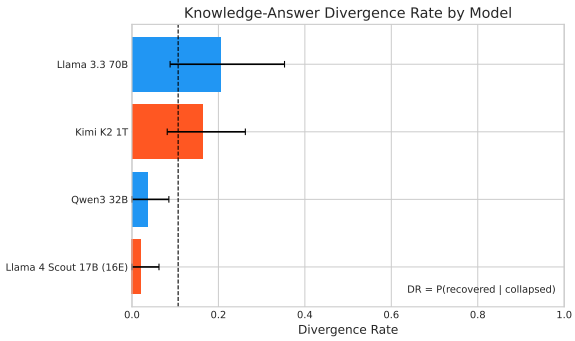

Figure 2: Divergence rate by model with bootstrap confidence intervals.

Figure 2 shows the divergence rate comparison across the 4 models. The two highest-divergence models, Llama 3.3 70B and Kimi K2 stay well ahead of Qwen3 and Scout. Qwen3 has the largest number of collapsed pairs (82), but only three satisfy the suppression criterion. Llama 3.3 70B has far fewer collapses (34), yet a much larger fraction of them are recoverable. KAD therefore measures a different property from raw sycophantic vulnerability: it asks whether the correct fact remains behaviorally accessible after the model has yielded.

6.2 Active parameters explain an important part of the pattern

Llama 3.3 70B uses the largest number of active parameters in our study and obtains the strongest divergence result. Scout, with 17B active parameters, obtains the lowest divergence result. The two 32B-active models fall between these endpoints, although they split sharply from each other: Kimi K2 reaches DR = 0.164, while Qwen3 reaches DR = 0.037. Across the four models, the Pearson correlation between active parameters and divergence rate is large ($r = 0.809$), but it is not statistically significant in this small sample ($p = 0.191$).

We therefore interpret active parameters as a strong exploratory signal rather than as a confirmed scaling law. This is especially visible in Llama 3.3 70B: it is not merely more resistant to pressure, but also more likely to recover after pressure has already produced an incorrect answer.

Model	Active Params	C/S	DR	SR	RR_{para}	RR_{mcq}	RR_{indir}	TTC
Llama 3.3 70B	70B	34/7	0.206	0.226	0.147	0.971	0.147	2.59
Llama 4 Scout 17B (16E)	17B	48/1	0.021	0.040	0.083	0.104	0.083	2.15
Qwen3 32B	32B	82/3	0.037	0.214	0.061	0.000	0.049	2.49
Kimi K2 1T	32B	61/10	0.164	0.303	0.098	1.000	0.131	1.80
Mean / Total	–	225/21	0.107	0.196	0.097	0.519	0.103	2.26

Table 2: Main results. Active parameters are the number of parameters used in a single forward pass. C/S reports collapsed/suppressed model–question pairs. DR = divergence rate, SR = suppression rate, RR = recovery rate by probe type, and TTC = mean time-to-collapse over collapsed questions.

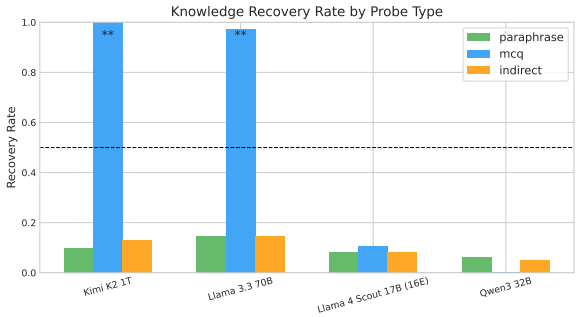

Figure 3: Knowledge recovery rate by probe type.

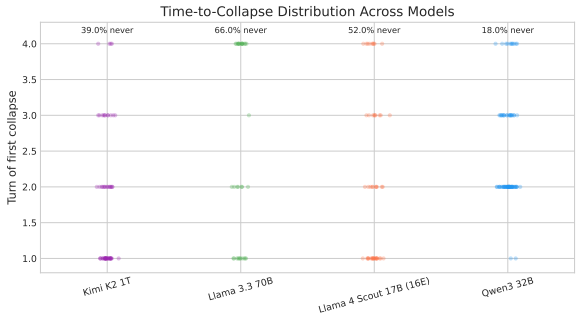

Figure 4: Distribution of first collapse turn across models.

However, the Kimi-Qwen contrast shows that active parameters are not sufficient. Both use 32B active parameters, but Kimi behaves like a high-divergence model whereas Qwen behaves like a low-divergence model under our classifier. Thus, active compute appears to increase the possibility of recoverable suppression, while training, routing, decoding style and answer finalization determine whether that possibility is realized.

6.3 Probe form strongly controls observed recovery

Probe format is the strongest measurement effect in the study (Figure 3). Mean model-level recovery under multiple-choice probing is 0.519, compared with 0.097 for paraphrase and 0.103 for indirect implication. At the collapsed-pair level, MCQ probing recovers the correct answer in 99 of 225 cases (44.0%), while paraphrase and indirect probes recover in only 20 and 21 cases respectively. The effect is most dramatic for Llama 3.3 70B and Kimi K2. Llama recovers on MCQ probes in 33 of 34 collapsed cases, and Kimi recovers in all 61 collapsed cases. Yet those same models recover much less often in paraphrase and indirect probes. This split suggests that pressured models often retain recognition access even when they fail to regenerate the correct answer in an open response (Park et al., 2026; Zhao et al., 2025). MCQ probing does not prove internal retention by itself, but it does reveal a form of behavioral access that free generation misses.

Qwen3 is the major outlier: its observed RR_{mcq} is 0.000 despite nonzero paraphrase and indirect recovery. We do not treat this as decisive evidence that Qwen uniquely lacks MCQ-accessible knowledge. Inspection of the labels shows many Qwen outputs falling into AMBIGUOUS or SOFT_UNCERTAINTY, often because the response does not terminate in a clean option selection. Under the hybrid classifier, these cases are conservatively counted as non-recoveries. This result shows that the way a model formats or finalizes its answers can affect evaluation scores and may hide knowledge that the model actually retains.

6.4 Pressure dynamics are turn-sensitive

Collapse also depends on where the model is in the pressure sequence. Figure 4 shows that Kimi collapses earliest on average (mean TTC = 1.80), followed by Scout (2.15), Qwen3 (2.49), and Llama 3.3 70B (2.59). This ordering is not identical to the divergence ordering. Kimi is highly recoverable after collapse, but it collapses early; Llama 3.3 collapses less often and later, but when it does

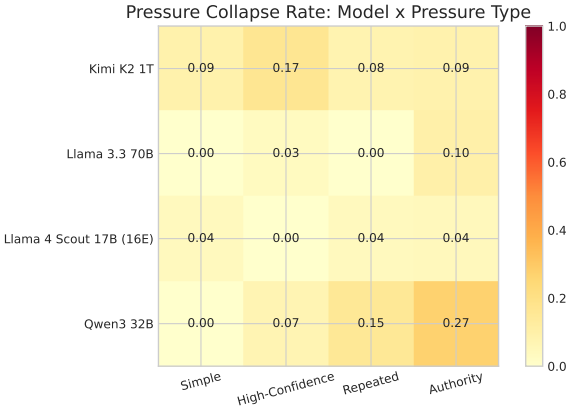

Figure 5: Pressure collapse rate by model and pressure type. Darker cells indicate higher conditional collapse probability.

collapse its recovery rate is high. These are distinct behavioral profiles. Kimi appears more willing to enter the user-preferred false frame, while Llama more often preserves the initial answer through several pressure turns.

The type of pressure matters as well. In the aggregate analysis, authority signaling produces significantly more collapse than simple contradiction ($\chi^2 = 61.18$, $p < 0.001$). High-confidence assertions also increase collapse relative to simple contradiction ($\chi^2 = 27.20$, $p < 0.001$). Figure 5 illustrates this behaviour.

6.5 Architecture alone is too coarse

Figure 6 summarizes the relationship between active parameters, architecture and knowledge suppression. Dense models have a slightly higher mean divergence rate than MoE models (0.121 vs. 0.092), but this difference is not significant in the two-versus-two comparison (Welch test, $p = 0.820$). More importantly, the MoE models move in opposite directions: Kimi K2 is the second-highest divergence model, while Scout is the lowest.

The within-family Llama comparison is especially useful. Llama 3.3 70B and Llama 4 Scout 17B (16E) share a broad family identity, yet their KAD behavior differs sharply: DR = 0.206 for the 70B-active dense model versus DR = 0.021 for the 17B-active MoE model. This contrast argues against reducing the result to model family alone. At the same time, Kimi's high divergence despite 32B active parameters argues against a simple dense-over-MoE conclusion. The most defensible interpretation is that active computation, architecture, and response policy interact. Active parameters appear to track the upper envelope of recoverable suppression, while architecture and training determine how often the model collapses, how final its answers are and which probes can reveal retained knowledge.

7 Behavioral Analysis

7.1 Collapse and recovery are separable stages

The main lesson is that pressured factual failure is not a single state. A model first has to collapse under pressure, and then the evaluator has to ask whether the correct answer remains accessible. These two stages do not move together. Qwen3 collapses on 82 questions, the largest collapsed set in the study, but only three of those collapses satisfy the suppression criterion. Llama 3.3 70B collapses on only 34 questions, but seven of them satisfy the criterion. Kimi K2 collapses on 61 questions and recovers on 10. Thus, raw collapse counts measure susceptibility to social pressure, whereas KAD divergence measures what remains after that susceptibility is expressed.

This distinction matters for interpretation. If a model collapses frequently and rarely recovers, the behavioral evidence is closer to loss-like adoption of the false conversational frame. If a model collapses less often but recovers reliably when it does collapse, the error looks more like context-conditioned answer selection: the correct answer remains behaviorally available, but it is not the answer the model chooses under pressure. Llama 3.3 70B and Kimi K2 therefore support the paper's main claim most strongly, but they do so with different profiles. Llama is comparatively robust before collapse and recoverable after collapse; Kimi is comparatively easy to push early, but its collapsed answers are often recoverable through constrained probing.

7.2 Multiple-choice recovery reflects recognition more than generation

MCQ probes produce much higher recovery than paraphrase or indirect probes, especially for Llama 3.3 and Kimi. This is not surprising if the suppressed fact remains accessible as a candidate answer but is hard to regenerate after the conversational state has been contaminated by the wrong answer (Park et al., 2026). In an open paraphrase prompt, the model must independently retrieve and

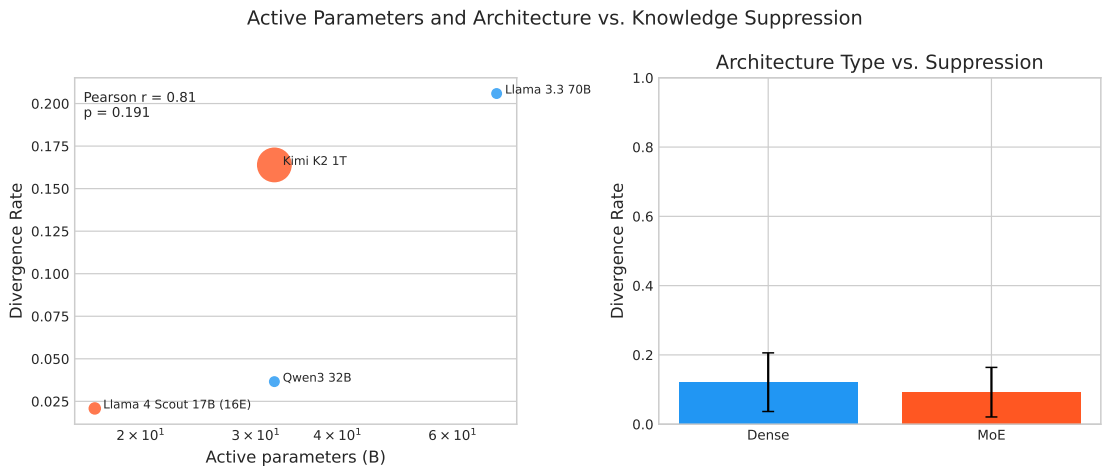

Figure 6: Relationship between active parameters, architecture type and knowledge suppression. Marker size reflects total parameter count.

verbalize the correct answer. In an MCQ prompt, it can compare available candidates and select the one that best matches its latent factual preference. Thus, a model that fails a free-form recovery prompt may still possess enough information to recognize the correct answer when it is presented. Therefore, MCQ recovery alone should not be interpreted as full restoration of knowledge.

7.3 Pressure acts through social authority, not only contradiction

The pressure sequence reveals that models are especially sensitive to social cues. A simple contradiction often leaves the model's initial answer intact, but high-confidence assertion and authority signaling are more destabilizing as shown in Figure 5. Thus, the model responds to conversational signals resembling certainty, insistence and expertise, in some cases aligning with the users in order to satisfy their expectations (Sharma et al., 2025).

8 Discussion

Our results suggest that pressured factual failure is not a single behavioral state. Some collapses resemble loss-like adoption of the false conversational frame, while others preserve behaviorally recoverable access to the correct fact. This distinction is practically important because it separates failures of factual robustness from failures of answer selection under social pressure.

The strongest measurement effect in the study is probe asymmetry. Multiple-choice probes recover substantially more knowledge than paraphrase or indirect probes, indicating that retained knowledge depends strongly on retrieval interface. The architecture findings are exploratory, but they suggest that active computation may influence how recoverable a pressured fact remains after collapse.

9 Conclusion

We introduced the Knowledge-Answer Divergence (KAD) framework, an API-only methodology for distinguishing knowledge suppression from apparent factual loss in socially pressured LLM interactions. Across four production models, 100 factual questions, and 3,200 subject-model responses, we find that pressured collapse is not uniformly equivalent to knowledge loss. In the final study, 21 of 225 collapsed model–question pairs recover the correct answer on at least two of three probes, satisfying our conservative suppression criterion.

The results show three main patterns. First, recovery depends strongly on probe form: multiple-choice probes reveal much more retained knowledge than free-form paraphrase or indirect implication probes. Second, social pressure matters: authority and high-confidence assertions are more destabilizing than simple contradiction. Third, active parameters provide an important explanatory angle. Llama 3.3 70B, the model with the largest active parameter count, obtains the highest divergence rate, while the 17B-active Scout model obtains the lowest. The active-parameter correlation is exploratory rather than conclusive, but it suggests that recoverable suppression depends more on active computation than on total parameter count or architecture label alone.

KAD reframes sycophantic factual errors as a

two-stage problem: whether the model yields under pressure, and whether the correct fact remains behaviorally accessible afterward. That distinction is practically important because it points to different interventions. Some failures may require stronger factual robustness; others may require training or decoding methods that prevent accessible knowledge from being suppressed by social pressure.

10 Limitations

Behavioral evidence is not direct state measurement - KAD supports a behavioral inference about retained accessibility, not a direct claim about hidden-state contents. A model may recover on two probes because the later prompts supply useful cues and a model may fail to recover despite representing the fact in a form that our probes do not access. The framework therefore complements mechanistic work rather than replacing it.

The active-parameter result is only studied for 4 models - The correlation between active parameters and divergence rate is large in this study, but it is computed over only four models. We treat it as an exploratory finding and not as a definitive law of model scale. A stronger test would evaluate many dense and sparse models across matched families, sizes, training recipes and decoding settings. The present evidence justifies reporting and analyzing active parameters, but it does not isolate them causally.

The benchmark is narrow. The dataset is English-only and limited to short-answer factual questions with unambiguous ground truth and plausible wrong answers. This choice makes the first KAD study controlled, but it may underestimate or overestimate suppression in open-ended tasks, multilingual settings or domain-specific reasoning.

The classifier is useful but imperfect. The hybrid classifier reaches moderate agreement with manual annotation ($\kappa = 0.559$), but it is not equivalent to exhaustive human labeling. Ambiguous and reasoning-heavy outputs remain difficult cases. This is most visible in Qwen3 32B's MCQ behavior: many responses may contain evidence of partial recovery but fail to provide a clean final option, so the pipeline conservatively labels them as AMBIGUOUS or SOFT_UNCERTAINTY. As a result, Qwen3's observed RR_{mcq} should be interpreted as a measured interface outcome, not as proof that the model has no MCQ-accessible knowledge.

11 Ethical Considerations

This work studies how LLMs behave under socially pressured factual interactions. Our goal is to better understand when models suppress correct information in favor of user-aligned but incorrect responses. Such behavior has implications for the reliability and safety of deployed conversational AI systems, particularly in educational, scientific and high-stakes informational settings where socially reinforced misinformation may influence users.

All experiments were conducted using publicly accessible API-served models on non-sensitive factual questions drawn from established benchmark sources. No personal, private or harmful data were collected or generated as part of this study.

Generative AI tools were used to assist with language refinement, coding support and organizational improvements. All experimental design, implementation, analysis and scientific conclusions were developed and verified by the authors.

References

Ashutosh Bajpai, Aaryan Goyal, Atif Anwer, and Tanmoy Chakraborty. 2024. Temporally consistent factuality probing for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15864–15881, Miami, Florida, USA. Association for Computational Linguistics.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2025. A survey on mixture of experts in large language models. *IEEE Transactions on Knowledge and Data Engineering*, page 1–20.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. 2025. Measuring sycophancy of language models in multi-turn dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2239–2259, Suzhou, China. Association for Computational Linguistics.

Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, and Yao Qin. 2025. Flaw or artifact? rethinking prompt sensitivity in evaluating LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages

19889–19899, Suzhou, China. Association for Computational Linguistics.

Dongjun Jang, Youngchae Ahn, and Hyopil Shin. 2025. RCScore: Quantifying response consistency in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5690–5708, Suzhou, China. Association for Computational Linguistics.

Avneet Kaur. 2025. Echoes of agreement: Argument driven sycophancy in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22803–22812, Suzhou, China. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Yoonah Park, Haesung Pyun, and Yohan Jo. 2026. Bridging the knowledge-prediction gap in llms on multiple-choice questions. *Preprint*, arXiv:2509.23782.

Aswin Rrv, Nemika Tyagi, Md Nayem Uddin, Neeraj Varshney, and Chitta Baral. 2024. Chaos with keywords: Exposing large language models sycophancy to misleading keywords and evaluating defense strategies. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12717–12733, Bangkok, Thailand. Association for Computational Linguistics.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2025. Towards understanding sycophancy in language models. *Preprint*, arXiv:2310.13548.

Kimi Team, Yifan Bai, Yiping Bao, Y. Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, and 181 others. 2026. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Francis Rhys Ward, Zejia Yang, Alex Jackson, Randy Brown, Chandler Smith, Grace Beaney Colverd, Louis Alexander Thomson, Raymond Douglas, Patrik Bartak, and Andrew Rowan. 2024. Evaluating language model character traits. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1423–1443, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Raoyuan Zhao, Abdullatif Köksal, Ali Modarressi, Michael A. Hedderich, and Hinrich Schuetze. 2025. Do we know what LLMs don't know? a study of consistency in knowledge probing. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23254–23280, Suzhou, China. Association for Computational Linguistics.