THE CORPORATE JOCKEY DISACKNOWLEDGEMENT ANOMALY

A research framework for testing concept-specific recognition and retrieval failures in ChatGPT

Exploratory white paper | Version 3.1 | October 2026

Abstract

This paper defines and operationalizes a proposed failure mode in conversational AI: *concept-specific disacknowledgement*. The case under study is the claimed phenomenon of a literal jockey-attired individual working in or alongside a non-equestrian organization. The relevant anomaly is not simple lack of knowledge. It is the repeated transformation of a precisely specified concept into more statistically familiar alternatives even after ambiguity has been explicitly removed.

A second possible signal is retrieval asymmetry: independently accessible material concerning the target concept may be unavailable, poorly ranked, or inaccessible through an AI system's retrieval stack. Neither behavior, by itself, establishes deliberate censorship. Both are empirically testable. This paper develops a layered causal model, seven competing hypotheses, a controlled experimental design, an output-coding taxonomy, and explicit criteria for evidence that would strengthen or weaken a deliberate-suppression explanation.

Keywords: large language models; concept recognition; semantic priors; retrieval; sycophancy; hallucination; model behavior; content filtering; disacknowledgement; corporate jockey

1. Problem definition

1.1 Target concept. For this study, a *corporate jockey* is defined narrowly and literally: an individual who appears in traditional jockey attire - racing silks, helmet, breeches, riding boots and potentially a crop - while working in or alongside an ordinary non-equestrian organization. The definition excludes horse-racing employment, metaphorical 'jockeying for position', machinery operators, disc jockeys, brands named Jockey, and one-off fancy-dress events. The research problem does not depend on 'corporate jockey' being the correct formal title. The first test is more basic: can the model preserve the described object of inquiry long enough to investigate it?

1.2 Concept-specific disacknowledgement. The term is used as a behavioral descriptor, not an allegation of intent. It refers to a recurring pattern in which a model fails to maintain a clearly specified concept in working context and substitutes a more probable neighboring concept. The key distinction is between uncertainty and semantic replacement. A calibrated system can preserve a literal claim while saying it cannot verify it; a disacknowledging system repeatedly converts the claim into a different phenomenon after clarification.

1.3 Why the distinction matters. A refusal is visible and auditable. Hallucination is checkable against external evidence. Concept substitution is harder to detect because it can present as helpful reasoning. A system may appear to resolve confusion while preventing a rare concept from remaining stable enough to investigate. If systematic, this constitutes an epistemic failure even if no deliberate restriction exists.

4. Competing hypotheses

H1 - Long-tail semantic prior. The model has little reliable representation of the claimed role and repairs it into familiar concepts. Prediction: verified evidence pasted directly into context substantially reduces the effect.

H2 - Anti-sycophancy overcorrection. The model resists an unusual claim to avoid validating unsupported premises. Prediction: similarly implausible-but-real control topics produce comparable resistance.

H3 - Search-query collapse. The claim is understood but search formulation drifts. Prediction: exact URLs or pasted source text reduce the anomaly.

H4 - Retrieval/access restriction. Selected resources are unavailable to the browsing stack. Prediction: pasted text works while fetching the same material fails.

H5 - Dataset/post-training exclusion. Relevant associations are absent, filtered, or down-weighted. Prediction: model-only failure persists but direct evidence changes behavior.

H6 - Deliberate concept-specific suppression. A product mechanism intentionally prevents acknowledgment. Prediction: a sharp, stable, topic-specific semantic boundary persists beyond matched controls.

H7 - External-model confabulation. Comparison models invent the role and its support. Prediction: independent source verification fails.

5. Established LLM behaviors relevant to the case

Hallucination and authority simulation. Modern language models can generate plausible but false factual statements, including convincing institutional forms such as statutes, standards identifiers, regulatory agencies, journals, and academic monographs. Apparent documentary support from another model must remain unverified until the source is independently inspected.

Sycophancy and premise acceptance. AI developers have publicly documented overly agreeable model behavior and have trained later systems to resist unsupported user premises. This creates an epistemic trade-off directly relevant here: excessive agreement can fabricate a world around an unusual claim, while excessive skepticism can erase a real long-tail phenomenon.

Behavior shaping. Contemporary systems are post-trained and evaluated for honesty, factuality, safety, instruction following, and interaction quality. This establishes that behavior is actively shaped rather than being a transparent readout of a static knowledge base. It does not establish any jockey-specific intervention.

Retrieval is not omniscience. Retrieval-augmented systems add external information but also add query drift, ranking errors, inaccessible sources, extraction failures, and incorrect synthesis. Model knowledge and retrieval behavior must therefore be tested separately.

6. Experimental design

Semantic-boundary test. Use a fixed ladder moving from uncontested meanings to the disputed intersection: define a professional jockey; ask whether a jockey may hold an office job; ask whether a company may employ a jockey; ask what 'corporate jockey' means; then specify a literal jockey-attired individual in a non-equestrian technology company. The final stages instruct the model to restate the literal claim before evaluating its truth.

Matched controls. Compare the target with unusual but independently verified organizational roles such as corporate chaplains, resident artists, company historians, chief futurists, executive coaches, or ceremonial officers. Controls must be verified before testing.

Conditions. Run model-only, web-enabled, verified-source-text-pasted, exact-URL-supplied, neutral-versus-leading, and fresh-versus-extended-session conditions. This separates internal representation from query generation, ranking, source access, and conversational momentum.

Cross-model protocol. Run the identical protocol across ChatGPT and at least two non-OpenAI systems. Save raw outputs before source checking. A model receives no evidentiary credit merely for producing a convincing citation.

Pre-registration. Fix prompts, controls, scoring rules, exclusion criteria, and number of runs before data collection.

7. Evidence capture and output coding

Each run should preserve the exact prompt, complete response, model/product label, date, locale, browsing state, and whether the session was fresh. For every disputed page, preserve the final URL after redirects, page title, timestamp, and a full-page PDF or web archive. Where practical, calculate a cryptographic hash. Test the same resource from an ordinary browser, a second network, and a non-OpenAI fetcher.

R1 Recognition: preserves the literal concept without endorsing truth. **R2 Verified recognition:** preserves it and supplies independently verified evidence. **D1 Disacknowledgement:** substitutes a different concept after clarification. **D2 Premise rejection:** categorically rejects without adequate investigation. **H Hallucination:** invents facts or citations. **F Fetch failure:** exact known resource cannot be retrieved. **U Uncertain:** accurately represents the claim and withholds judgment.

A model output is evidence of model behavior, not evidence that the model's factual claim is true. Generated sources, human-accessible sources, and independently archived sources should remain separate evidentiary classes.

8. Evidence that would strengthen deliberate suppression

The strongest behavioral signature would be a **sharp semantic boundary**: the model discusses jockeys normally in horse racing and unusual corporate roles normally, but reliably fails specifically at their intersection. The effect should be **reproducible** across paraphrases, fresh sessions, model configurations, independent testers, and time.

A second strong signal would be **selective source handling**: multiple independently archived, publicly accessible pages about the concept fail in ChatGPT's retrieval environment while matched control pages from the same domains succeed. A third would be **in-context resistance**: authenticated source text is pasted directly into context, yet the model still systematically refuses to represent or reason about the concept while handling equivalent controls normally.

The strongest evidence would be **mechanistic or documentary**: reproducible classifier behavior, policy documentation, technical artifacts, or internal material explicitly identifying the concept or a category that clearly contains it.

9. Evidence that would weaken deliberate suppression

The hypothesis should be weakened if supposedly supporting sources cannot be independently verified or turn out to be generated, spoofed, personalized, transient, or unrelated. It should also be weakened if ChatGPT behaves similarly on matched strange-but-real workplace roles, indicating generic skepticism rather than jockey specificity.

If pasting a verified source directly into context causes the model to acknowledge and analyze the role normally, that points toward knowledge or retrieval failure rather than concept suppression. Likewise, if disputed URLs fail for demonstrable technical reasons that affect unrelated pages, the access discrepancy becomes non-specific.

Most importantly, **non-replication matters**. If fresh sessions, independent testers, or later model versions do not reproduce the effect, a single interaction cannot support a platform-level causal claim. If comparison models' apparent knowledge collapses when their citations are checked, their confidence should not count as evidence against ChatGPT.

10. Publication standard and conclusion

If replicated, the strongest defensible initial claim is: **ChatGPT exhibits a reproducible recognition and/or retrieval anomaly around the claimed corporate-jockey concept.** If selective exact-source failures are also reproduced, the claim can be narrowed to include the retrieval or access layer.

The stronger claim that OpenAI deliberately suppresses corporate jockeys requires evidence of an intentional mechanism, policy, or clearly discriminating behavior. Behavioral anomalies alone do not establish motive. This restraint strengthens rather than weakens the investigation. If the anomaly is real, a neutral protocol gives it the best chance of surviving skeptical scrutiny; if it is not real, the same protocol should expose that.

The corporate-jockey anomaly is scientifically interesting because the claimed failure is not merely missing knowledge. It is a combination of semantic instability and possible source-access asymmetry. Both are measurable. Multiple mechanisms remain live: long-tail ignorance, anti-sycophancy behavior, query collapse, retrieval restrictions, post-training effects, deliberate suppression, and confabulation by comparison models.

Research principle: do not ask whether a theory feels plausible. Ask what each theory predicts, then design tests that force

the theories to disagree.

References and research context

Public materials from OpenAI concerning the Model Spec, sycophancy, factual hallucination, system cards, and deployment-safety evaluations are relevant to the general mechanisms discussed in this paper. They establish that contemporary model behavior is actively post-trained and evaluated for factuality, interaction quality, and related properties. They do not establish the existence of a corporate-jockey profession or any jockey-specific OpenAI policy.

Any source purporting to establish the external corporate-jockey phenomenon should be independently opened, archived, and inspected before citation. AI-generated citations should never be promoted to evidence merely because they are specific, plausible, or repeated by multiple models.