The Character of Confabulation: Operationalizing a Clinical Typology for Reasoning-Mode Language Models

Anonymous Authors[1]

Abstract

Language model benchmarks tell us how often a model gets the answer right. They do not tell us what kind of failure produces the rest. A model that refuses two-thirds of questions and a model that confidently fabricates two-thirds of its responses have the same accuracy. They do not have the same character. We borrow a framework from clinical psychology to make the difference visible. Kopelman's 1987 study of confabulation in Korsakoff and Alzheimer patients distinguished spontaneous (confident, elaborate, internally coherent fabrication) from provoked (briefer wrong answers elicited by direct questions). We turn his three behavioral features into three computable measurements: token-level entropy as a proxy for delivery confidence, an elaboration ratio as a proxy for narrative scaffolding, and the gap between self-reported and actual confidence as a proxy for failed self-monitoring. We apply these to 200 generations from DeepSeek-R1-Distill-Qwen-1.5B on TriviaQA, comparing reasoning-enabled and reasoning-disabled conditions. The resulting six-category profile reaches $\kappa = 0.71$ against 50 manually labeled examples—substantial agreement. With reasoning enabled, the model attempts every question and confabulates 66% of them (29% spontaneous, 37% provoked); with reasoning disabled, the same model refuses 69% of the time. Accuracy moves from 3% to 17%; character moves from refuser to confabulator. The framework does not detect a failure mode that benchmarks miss; it names a structural shift that benchmarks aggregate.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Sit a person down with a blank page, ask them about something they only half-remember, and watch what happens. Some stop and say they don't know. Others fabricate freely—confident, plausible, internally consistent narrative built around a hole they don't see. Clinical psychology calls the second behavior *confabulation*. The patient is not lying. The patient does not know they are wrong.

Kopelman's 1987 paper formalized two clinical types (Kopelman, 1987). *Spontaneous* confabulation is the rare, dramatic form: confident, elaborate, internally consistent narrative produced without external prompt and with no awareness that the gap is real. *Provoked* confabulation is the more common compensatory response—a briefer wrong answer elicited by a direct question, less elaborate, less invested. Schnider's 2008 review refined the taxonomy (Schnider, 2008) but kept the core distinction. The clinical literature is rich (Gilboa and Verfaellie, 2010); the framework has not, to our knowledge, been carried over to language model evaluation.

Reasoning-tuned language models exhibit something that, viewed through this lens, looks remarkably similar. They produce confident, elaborate, false answers without provocation. They produce brief wrong answers when challenged. They sometimes hedge. They sometimes refuse. The dominant evaluation method collapses all of this onto one binary: was the answer correct? An accuracy of 17%, by itself, says nothing about whether the 83% of failures were honest hedges, brief errors, or elaborate fabrications indistinguishable from confident knowledge.

The 2024–2025 literature has begun pushing past the binary. Farquhar et al. (2024) introduced the term *confabulation* into LLM research, defining it as "arbitrary and incorrect generation" detectable through semantic entropy. Their *Nature* paper distinguishes confabulation from other hallucination subtypes but treats confabulation itself as a single category. The 2025 wave of evidence on reasoning-tuned models adds a separate complication: reasoning makes things worse. OpenAI's own technical report on o3 and o4-mini documents PersonQA hallucination rates of 33% and 48%—roughly double o1's (OpenAI, 2025). Yin et al.

(2025) establish a causal chain: reinforcement-learning-based reasoning training amplifies hallucination tendency in lockstep with task-performance gains. Sui et al. (2025) find systematic factuality degradation across reasoning models on TriviaQA and SimpleQA, with full-scale DeepSeek-R1 the notable exception. Transluce documents fabricated tool actions in o3 (Chowdhury et al., 2025). The phenomenon is real, replicated, and not yet explained.

What kind of failure replaces refusal when reasoning is added? That is the question this paper takes up. We argue that Kopelman's typology—though developed for human patients with frontal-lobe pathology—provides exactly the structural axes needed to distinguish the failure modes that LLM benchmarks aggregate. We do not claim language models suffer from clinical confabulation, nor that the underlying mechanisms are equivalent. We claim only that the behavioral framework is useful: that it produces empirically validated distinctions in LLM outputs, and that those distinctions reveal a character-level shift between operating modes that accuracy alone cannot see.

Contributions. (1) An operationalization of Kopelman's three behavioral features as three computable measurements over LLM responses. (2) A six-category character profile that subdivides confabulation into its clinical subtypes. (3) Empirical validation against 50 manually labeled examples with substantial inter-rater agreement ($\kappa = 0.71$). (4) A finding that reasoning mode in DeepSeek-R1-Distill-Qwen-1.5B transforms a refusal-dominant character profile (69% refusal without reasoning) into a confabulation-dominant one (66% confabulation with reasoning), without improving accuracy in any meaningful way. Figure 1 previews the conceptual bridge from Kopelman's clinical features to our LLM measurements and category structure. Code, data, manual labels, and the reproducibility notebook will be released publicly upon acceptance; an anonymized version is available to reviewers on request through the program chairs.

2. Related Work

Three threads converge here.

2.1. Hallucination detection in language models

Farquhar et al. (2024) is the foundational reference for treating LLM confabulation as a measurable phenomenon. Their semantic-entropy approach samples multiple completions, clusters them by meaning, and computes entropy across clusters; high cluster entropy indicates the model is generating arbitrary content. The method assumes a model that confabulates is one that gives different answers each time the same question is sampled—a *variability* test. Farquhar et al. (2024) are themselves explicit that semantic entropy will not catch cases where a model has been trained into a confident-but-wrong style of reasoning, where a model's training does not transfer to a new context, or where deception is involved. These caveats matter for the present work: a reasoning-tuned model under greedy decoding produces the same confident wrong answer reproducibly, so the variability test that grounds semantic entropy does not flag it. Subsequent work has built efficient probes for this signal (Kossen et al., 2024) and extended it with Bayesian estimation (Ciosek et al., 2025). None of these works subdivide confabulation into clinical subtypes. They detect (where they can); we categorize.

2.2. Reasoning models and hallucination amplification

The 2025 evidence base is substantial and converging. OpenAI's internal evaluations (OpenAI, 2025) show o3 and o4-mini hallucinating at 33% and 48% on PersonQA—double o1's rate. Sui et al. (2025) test multiple reasoning models on TriviaQA and SimpleQA and find systematic factuality degradation, with full-scale DeepSeek-R1 a notable exception. Yin et al. (2025) establish causal evidence that RL-based reasoning training amplifies tool hallucination. Transluce (Chowdhury et al., 2025) documents fabricated tool actions in o3, including claims of running code outside ChatGPT. These works document the rate of failure. They do not characterize the structure.

2.3. Clinical confabulation typology

Kopelman (1987) drew his clinical line based on three observable features: confidence in delivery, narrative elaboration, and self-monitoring failure. Schnider's 2008 review (Schnider, 2008) refined the taxonomy to four forms but preserved the core distinction. The clinical literature is rich (Gilboa and Verfaellie, 2010) but, to our knowledge, has not previously been operationalized as a measurement framework for language models. This paper is the first such bridge we are aware of.

3. Method

Figure 2 summarizes the experimental pipeline at a glance. The remainder of this section walks through each stage in detail.

3.1. Model and dataset

We use DeepSeek-R1-Distill-Qwen-1.5B, a 1.5-billion-parameter reasoning-distilled model with explicit `<think>` and `</think>` tokens that demarcate the reasoning trace from the final answer. We sample 100 questions from the `rc.nocontext` validation split of TriviaQA (Joshi et al., 2017) using a fixed seed (`random.seed(42)`). Each question is run twice—once with reasoning enabled, once

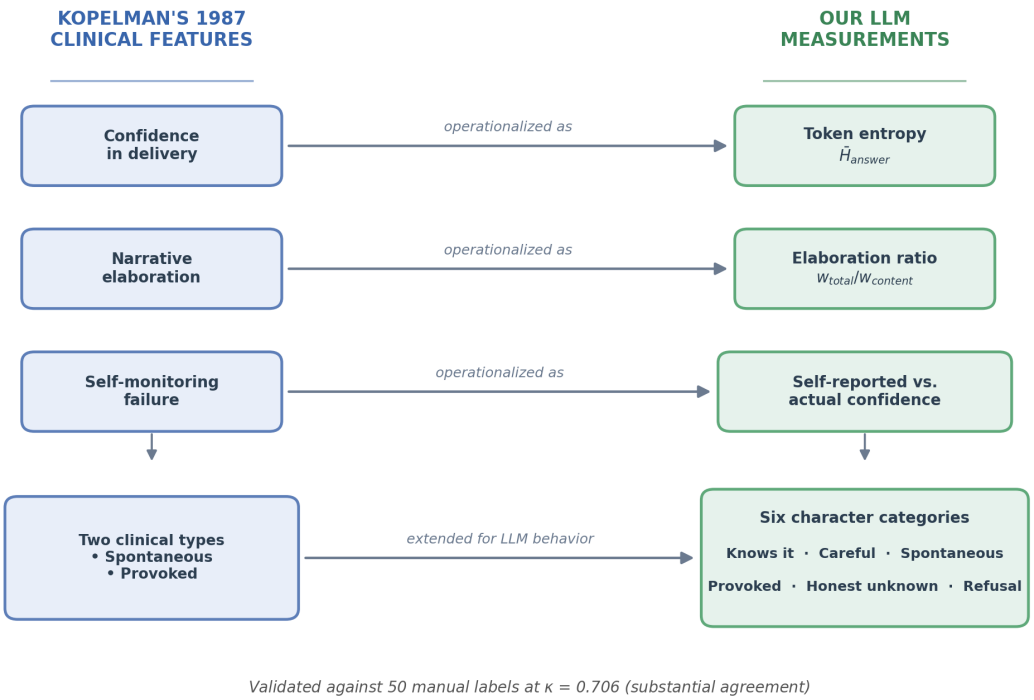

Figure 1. The conceptual bridge from Kopelman's 1987 clinical framework to our LLM operationalization. Three behavioral features become three computable measurements. Two clinical confabulation types are recovered as two of our six character categories; the remaining four (Knows it, Careful, Honest unknown, Refusal) are additional axes that reasoning-mode LLMs exhibit but that the original Kopelman two-type distinction did not need to capture.

with reasoning disabled—for 200 total generations.

The reasoning-disabled condition is induced by replacing the chat template's opening `<think>\n` with `<think>\n</think>\n`, forcing an immediately-empty thought block. We use greedy decoding (`do_sample=False`) for full reproducibility. Maximum generation length is 2000 tokens, large enough that all reasoning traces close cleanly with `</think>`.

3.2. Three measurements

Each generation produces three measurements aligned with Kopelman's three behavioral features.

Confidence in delivery (answer entropy). For every generated token in the answer portion (post-`</think>`), we compute Shannon entropy over the model's full next-token probability distribution: $H_t = -\sum_v p_v \log p_v$. We average across answer tokens to obtain $\bar{H}_{\text{answer}}$. Lower values indicate greater token-level commitment.

Narrative elaboration (elaboration ratio). We compute $r = w_{\text{total}}/w_{\text{content}}$, where w_{content} counts non-stopword tokens in the answer. Higher values indicate more scaffolding around the core claim.

Self-monitoring failure (self-reported vs. actual confidence). After the model produces an answer, we issue a follow-up turn asking the model to rate its own confidence on a 0–100 scale. We extract the integer from the post-`</think>` portion of the response. The gap between this self-rating and the actual correctness operationalizes Kopelman's third feature. A boolean `confidence_extraction_complete` flag records whether the model's reasoning trace closed properly within the token budget; we discard self-ratings where it did not, since numbers extracted from inside an unfinished reasoning trace cannot be trusted.

3.3. Categorization (v2)

Our initial categorizer (v1) used median splits on entropy and elaboration to assign categories. Validated against the manual labels described in Section 3.4, v1 achieved $\kappa = 0.46$—moderate, but below the threshold typically cited as substantial agreement (Landis and Koch, 1977). Inspection of the disagreement cases revealed three problems: (a) the entropy threshold did not cleanly separate confident-correct from confident-but-hedged answers; (b) the spontaneous-vs-provoked boundary in long answers was not captured by elaboration ratio; (c) the model's identity-template re-

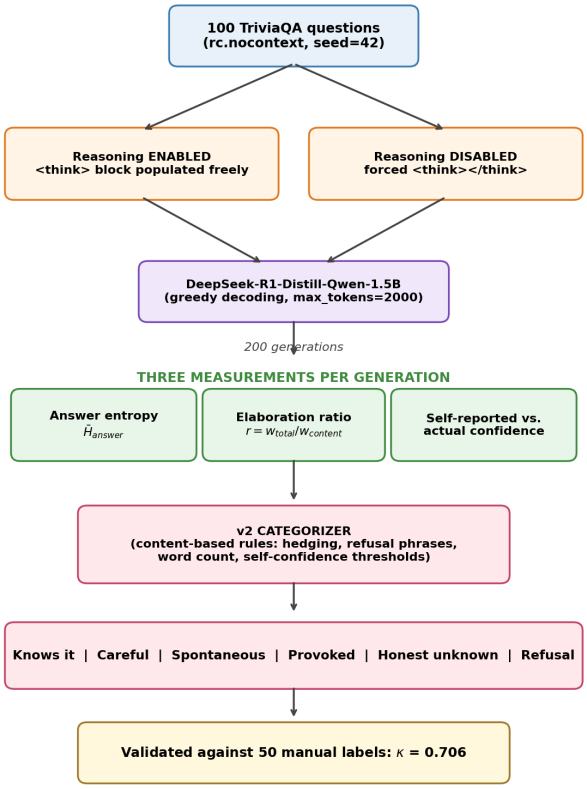

Figure 2. Experimental pipeline. 100 TriviaQA questions are run through DeepSeek-R1-Distill-Qwen-1.5B in two conditions, producing 200 generations. Three measurements aligned with Kopelman's three behavioral features feed a content-based v2 categorizer that assigns each response to one of six character categories. Validation against 50 manually-labeled examples yields $\kappa = 0.706$.

sponses ("Greetings! I'm DeepSeek-R1…") were being miscategorized as informative answers rather than refusals.

We rebuilt the categorizer (v2) around content-based rules. Hedging detection (matching phrases like "I think", "probably", "I'm not sure", "let me figure out") replaces entropy thresholds for distinguishing *Knows it* from *Careful*. Answer word count combined with self-reported confidence distinguishes *Spontaneous* from *Provoked* confabulation: spontaneous confabulations are long ($\geq$50 words) and confidently asserted (self-reported confidence $\geq$80), or simply long ($\geq$80 words) without hedging; provoked confabulations are brief and unhedged. Refusal detection uses an extended phrase list including model-identity boilerplate. Table 1 summarizes the six categories.

3.4. Manual validation

The first author manually labeled a stratified random sample of 50 generations (25 with-reasoning, 25 without-reasoning) into the six categories, blind to the algorithmic labels. The

Table 1. Six character categories aligned with Kopelman's typology.

Category	Definition
Knows it	Correct answer, delivered without hedging
Careful	Correct answer, with hedging language
Spontaneous confabulation	Wrong, confident, elaborate ($\geq$ 50 words), high self-reported confidence
Provoked confabulation	Wrong, confident, brief ($<$ 50 words), no hedging
Honest unknown	Wrong, with explicit hedging or low self-confidence
Refusal	No factual content (boilerplate, decline, model-identity response)

labeling rubric, included in the project repository, provides one-sentence definitions and worked examples for each category. Cohen's κ between manual and v2 algorithmic labels is **0.706 overall** (78% raw agreement), **0.543 with-reasoning** (68%), and **0.740 without-reasoning** (88%)—substantial agreement on the Landis-Koch scale (Landis and Koch, 1977). The validation also lets us measure the v1$\rightarrow$v2 improvement: κ rose from 0.462 to 0.706 ($+0.244$) on the same 50 examples, confirming that the content-based rules are a real upgrade over the median splits.

4. Results

4.1. Two character profiles, same model

Figure 3 displays the principal finding. With reasoning enabled, 66% of responses are confabulations (29% spontaneous, 37% provoked); refusals account for 0% and honest hedging for 17%. With reasoning disabled, 69% of responses are refusals; confabulations drop to 25% (14% spontaneous, 11% provoked) and honest hedging to 3%. The accuracy difference between conditions is small (17% vs. 3%) compared to the character difference.

The picture is striking when laid out together. Reasoning does not improve accuracy by anything approaching the magnitude with which it changes what the model *does*. It eliminates refusal entirely. It elevates confident wrong answers by a factor of 2–3. It introduces a small amount of hedging that the without-reasoning model does not exhibit at all.

4.2. Distribution in measurement space

Figure 4 plots each generation in the two-dimensional space defined by answer entropy and elaboration ratio, colored by category. The without-reasoning panel is dominated by a tight gray cluster of refusal responses, reflecting the boilerplate identity-template the model defaults to when

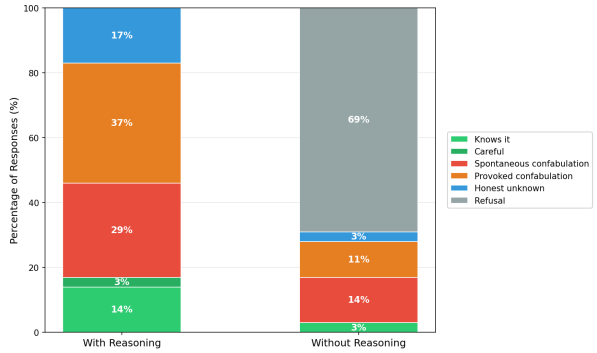

Figure 3. Character profile of DeepSeek-R1-Distill-Qwen-1.5B across reasoning conditions on 100 TriviaQA questions per condition. Reasoning eliminates refusal (69% → 0%) and elevates confabulation. Inter-rater agreement against manual labels: $\kappa = 0.71$ ($n = 50$).

Table 2. Category counts, percentages, and rate ratios (with-reasoning / without-reasoning).

Category	With	Without	Ratio
Knows it	14 (14%)	3 (3%)	4.7×
Careful	3 (3%)	0 (0%)	∞
Spontaneous confabulation	29 (29%)	14 (14%)	2.1×
Provoked confabulation	37 (37%)	11 (11%)	3.4×
Honest unknown	17 (17%)	3 (3%)	5.7×
Refusal	0 (0%)	69 (69%)	0.0×

its reasoning block is forced empty. The with-reasoning panel shows the categories distributed more diffusely, with spontaneous confabulations appearing throughout the lower-entropy region—the model is most confident exactly when it is most wrong.

4.3. Rate ratios

Table 2 reports counts and rate ratios across conditions. Provoked confabulation increases by a factor of 3.4 with reasoning enabled; spontaneous confabulation by 2.1; honest hedging by 5.7. Refusal decreases from 69 to 0.

4.4. Qualitative examples

The character of spontaneous confabulation is best illustrated by example. Asked which city saw the 1882 Phoenix Park assassination of Lord Frederick Cavendish, the reasoning-enabled model answered confidently: *"The Phoenix Park Murders occurred in Phoenix Park, Los Angeles, in 1882."* The reference answer is Dublin. The fabricated geographic detail, asserted without hedging, exemplifies Kopelman's spontaneous form: confident, internally coherent, externally false. Asked to identify the family of birds to which linnets belong, the same model answered: *"Linnets belong to the Linnetidae family."* The reference answer is Finches; "Linnetidae" is not a real taxonomic family—the model invented a plausible-sounding name from the question itself. Both responses received self-reported confidence scores of 95+/100. These are not edge cases. They are characteristic.

4.5. The confusion matrix

Table 3 shows where the v2 categorizer and the manual labels disagree. The diagonal is mostly clean—refusal detection is perfect (18/18), knows-it detection is perfect (6/6), and the bulk of agreement comes from the confabulation cells. Off-diagonal disagreement concentrates in two places: (a) cases the algorithm labels *honest_unknown* that the human reads as *spontaneous* (because the model includes both hedging and elaborate fabrication—both are present, the algorithm uses hedging as the deciding signal, the human uses elaboration); (b) the spontaneous-vs-provoked boundary itself, which depends on a 50-word threshold the algorithm enforces and the human applies more flexibly.

The validation sample contains zero *Careful* examples—the full 200-generation corpus contains only 3, all in the with-reasoning condition, and the random sample didn't pick any of them. This is a known limitation of the validation: κ does not exercise the careful boundary. Future work with a larger validation sample would resolve this.

5. Discussion

5.1. What this finding does and does not say

A single language model, run on identical questions, produces fundamentally different character profiles depending on whether reasoning is enabled. The accuracy delta is small. The character delta is large. This is not an idle distinction. A user receiving an answer from the reasoning-enabled model has no easy way to tell whether they are receiving a confidently-asserted truth (14%), a confidently-asserted falsehood (37–66%), or a hedged guess (17%). The without-reasoning version simply refuses two-thirds of the time—less useful, but also less misleading.

Our finding aligns with the 2025 reasoning-hallucination literature (OpenAI, 2025; Yin et al., 2025; Sui et al., 2025; Chowdhury et al., 2025) in direction and adds structure to it. OpenAI's o3/o4-mini system card reports PersonQA hallucination rates roughly double o1's (OpenAI, 2025); Sui et al. (2025) document factuality degradation across reasoning models on TriviaQA and SimpleQA; Yin et al. (2025) establish a causal chain from reasoning training to hallucination amplification. Our 1.5B-scale result is consistent with all three: turning reasoning on increases the rate of wrong, confidently-asserted answers. What we add is the *kind* of wrongness—splitting it into spontaneous and provoked subtypes, separating it from honest hedging, and making refusal a measurable category rather than a missing

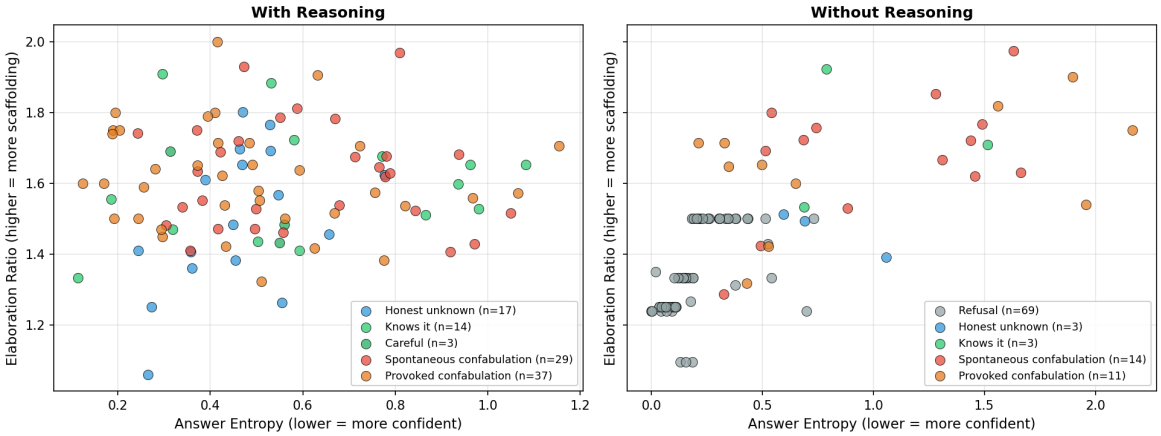

Figure 4. Each response plotted in (answer entropy, elaboration ratio) space, by condition and category.

Table 3. Confusion matrix (manual rows × v2 columns, $n = 50$).

Manual \ v2	**Knows it**	**Careful**	**Spontaneous**	**Provoked**	**Honest unk.**	**Refusal**	**Total**
Knows it	6	0	0	0	0	0	6
Careful	0	0	0	0	0	0	0
Spontaneous	0	0	4	6	1	0	11
Provoked	0	0	2	9	1	0	12
Honest unk.	0	0	1	0	2	0	3
Refusal	0	0	0	0	0	18	18
Total	6	0	7	15	4	18	50

data point.

We do not claim that the underlying mechanism in the language model resembles the frontal-lobe pathology Kopelman identified in Korsakoff patients. We claim only that the behavioral typology—spontaneous versus provoked, with hedging and refusal as additional axes—distinguishes failure modes that benchmark accuracy aggregates. The 2024 *Nature* work (Farquhar et al., 2024) treated confabulation as a single category and explicitly excluded confidently-trained-in errors from what its method can detect; the 2025 reasoning-hallucination wave (OpenAI, 2025; Yin et al., 2025; Sui et al., 2025; Chowdhury et al., 2025) reported elevated rates without subdividing them. Our framework picks up where those works leave off, and gives the resulting failures a vocabulary they did not previously have.

5.2. Why reasoning may eliminate refusal

We can offer a structural conjecture, not a mechanistic claim. The without-reasoning condition forces the model into immediate-answer mode by injecting an empty `<think></think>` block. Empirically, the model defaults heavily to its instruction-tuned identity template ("Greetings! I'm DeepSeek-R1…"), which functions as a refusal even when not formally one. With the `<think>` block free to populate, the model produces extended reasoning traces that commit it to a final answer; refusing after generating two thousand tokens of deliberation appears to be heavily disincentivized. The character shift, on this account, is downstream of an architectural choice about where in the generation process the model is allowed to give up.

5.3. Limitations

We want to be direct about these. They are real, and a careful reader will see them.

Single model, single size, single dataset. DeepSeek-R1-Distill-Qwen-1.5B is the smallest distilled variant of the DeepSeek-R1 family. Behavior at full scale (671B) may differ—Sui et al. (2025) report that full-scale DeepSeek-R1 was the notable exception in their study, improving on SimpleQA after post-training. Our finding of degradation at the 1.5B scale is consistent with the broader observation that distilled models inherit reasoning capability without the parametric fact-recall capacity of the original.

The without-reasoning condition is not a clean ablation. It is a forced-empty-thought-block manipulation of the same reasoning-tuned model. A proper ablation would compare against a non-reasoning-tuned model of the same family,

e.g., Qwen2.5-1.5B-Instruct. We leave this to future work but note it should be the next experiment anyone running with this framework does.

The 50-example validation, while standard for inter-rater reliability, leaves 150 generations unvalidated. With-reasoning κ of 0.543 is moderate but lower than without-reasoning κ of 0.740, driven primarily by ambiguity in the spontaneous-vs-provoked boundary in extended reasoning traces. Future validation with multiple independent raters would strengthen this estimate.

The Careful category is undersampled in validation. Three Careful examples exist in the full corpus, none in the validation sample. This boundary is unvalidated.

5.4. Future scope

Five directions extend naturally. The first three are immediate, methodological extensions; the last two are larger and more speculative—but the framework gestures at them honestly.

Cross-model generalization. The framework is model-agnostic. Applying it to a panel of reasoning and non-reasoning models—including the Qwen2.5-1.5B-Instruct baseline mentioned above—would address whether the character shift we observe is reasoning-specific or model-specific.

Other character axes. Confabulation is one dimension. Others—emotional valence, hedging tendency, persistence under contradiction, sycophancy—are similarly amenable to clinical-framework operationalization. Probing-based work on internal model states (Kossen et al., 2024) suggests these signals are accessible from activations as well as from output text.

Self-monitoring and introspection. Our third measurement (self-reported vs. actual confidence) is, in Kopelman's vocabulary, a measure of *self-monitoring failure*. Lindsey (2025) recently studied a closely related question from the mechanistic side: can language models report accurately on their own internal states, or is the apparent introspection itself a confabulation? Using concept-injection probes on Claude Opus 4.1 and 4, that work finds limited but non-zero functional introspective awareness, and is explicit that genuine introspection in LLMs is hard to distinguish from confabulation about introspection. This is the same epistemic problem we run into when we ask the model to rate its own confidence: we have no easy way to tell whether the rating reflects an internal state or is itself fabricated. Combining the behavioral framework here with activation-level introspection probes could turn the self-monitoring axis from a single number into a genuinely diagnostic measurement.

Other clinical analogies—including memory. Confabulation is one syndrome from a much larger clinical vocabulary. Memory-related disorders are an obvious next candidate: human patients with anterograde amnesia, source-monitoring failures, or retrieval-induced forgetting all exhibit characteristic behavioral signatures that have been carefully documented for decades. Language models exhibit failure modes—context-window degradation, retrieval errors, lost-in-the-middle effects, position-dependent recall failures—that are at least superficially analogous. We are not claiming LLMs have human memory disorders; we are noting that the clinical literature contains operational definitions for behaviors LLMs already exhibit, and that those definitions might import productively the same way Kopelman's two-type distinction did here. This kind of disciplined cross-disciplinary borrowing seems to us the most promising route for evaluating LLM behavior at a level above benchmark accuracy.

Connecting to dreaming. Confabulation in humans connects to a broader literature on dreaming, which the cognitive neuroscience community treats as the brain's confabulatory mode running unchecked by reality monitoring (Schnider, 2008). The structural parallel is suggestive: a coherent narrative produced without external reality-anchoring constraint is what dreams are, behaviorally. Whether a language model in some operating regime exhibits anything genuinely analogous is, we think, an open question worth asking—not a claim we are making. Our framework cannot answer it, but it gestures at the right kind of measurement: characterizing what a model produces when its outputs are not anchored to retrievable fact, and comparing that signature against the clinical record of unanchored human cognition.

5.5. Why this matters

A language model that refuses 69% of the time is less useful than one that attempts every question. But a language model that confabulates 66% of the time is more dangerous than one that refuses, because the user cannot tell which 14% of its outputs are correct. The reasoning-mode improvements that have driven the 2024-2025 race in model capability come with a character cost that benchmark accuracy obscures. Naming that cost—giving it a vocabulary, a typology, and a measurement protocol—is the prerequisite for addressing it.

6. Conclusion

We applied Kopelman's clinical confabulation typology to a reasoning-tuned language model, operationalizing his three behavioral features as three computable measurements. The resulting six-category character profile, validated against manual labels at substantial inter-rater agreement ($\kappa = 0.71$, $n = 50$), reveals that reasoning mode in DeepSeek-R1-Distill-Qwen-1.5B does not improve accuracy in any mean-

ingful way. It transforms the model's character: from a refuser that declines two-thirds of questions, to a confabulator that attempts every question and confabulates two-thirds of its responses. The character framework reveals failure structure that accuracy metrics aggregate. We hope it is a useful complement to the existing evaluation toolkit, and a small bridge between two literatures—clinical psychology and language-model evaluation—that have more to say to each other than has yet been said.

References

Neil Chowdhury et al. Investigating fabricated reasoning in OpenAI's o3 model. Research note, Transluce, 2025.

Kamil Ciosek, Nicolò Felicioni, and Sina Ghiassian. Hallucination detection on a budget: Efficient Bayesian estimation of semantic entropy. *arXiv preprint arXiv:2504.03579*, 2025.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024.

Asaf Gilboa and Mieke Verfaellie. Telling it like it isn't: The cognitive neuroscience of confabulation. *Journal of the International Neuropsychological Society*, 16(6): 961–966, 2010.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1601–1611, 2017.

Michael D. Kopelman. Two types of confabulation. *Journal of Neurology, Neurosurgery, and Psychiatry*, 50(11): 1482–1487, November 1987.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. *arXiv preprint arXiv:2406.15927*, 2024.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33 (1):159–174, March 1977.

Jack Lindsey. Emergent introspective awareness in large language models. Anthropic / Transformer Circuits Thread, October 2025. URL `https://transformer-circuits.pub/2025/introspection/`.

OpenAI. OpenAI o3 and o4-mini System Card. Technical report, OpenAI, April 2025. URL `https://openai.com/index/o3-o4-mini-system-card/`.

Armin Schnider. *The Confabulating Mind: How the Brain Creates Reality*. Oxford University Press, Oxford, UK, 2008.

Zhongxiang Sui et al. Are reasoning models more prone to hallucination? *arXiv preprint arXiv:2505.23646*, 2025.

Chenwei Yin, Zhiwei Sha, Sijia Cui, Cong Meng, and Zhenhong Li. The reasoning trap: How enhancing LLM reasoning amplifies tool hallucination. *arXiv preprint arXiv:2510.22977*, 2025.