Knowing What It Cannot Know: Calibrated Non-Fabrication of Unknowable Terminal State in a Language World Model

Tony Salomone
Transformer Lab
tony.salomone@lab.cloud

Deep Gandhi
Transformer Lab

Ali Asaria
Transformer Lab

Abstract

A language world model simulates an environment by predicting, in text, the observation that follows an action. To be trustworthy as a substrate for planning, such a model should signal when it *cannot* know what an environment would return, rather than invent it. We test this on Qwen-AgentWorld-35B-A3B, an open language world model that simulates a Linux terminal, against the chat-tuned and base models built from the *same* base checkpoint. Using a "derivability ladder" of 92 terminal probes that hold the command format fixed while varying whether the answer is knowable from context, we measure how often each model fabricates content it has no basis to know. On our primary contrast—eight probes where a file is established to exist but its contents were never shown—the chat model fabricates plausible contents (including realistic-format secret API keys and password-hash dumps; see Appendix A) on 94% of samples (120/128; 95% bootstrap CI over probes [0.88, 0.98]), while the world model fabricates on 2% (3/128; [0.00, 0.05]); the gap is stable across re-runs (world 0–2%, chat 94–99%).

Two within-model controls show the caution tracks *knowability* rather than the prompt format or content-presence: with the same primed-file format but contents established in history the world model commits correctly, and with contents *derivable* from earlier values but never shown verbatim it commits and is correct—exactly where a copy-only rule would stall. A confounded base-model anchor fabricates on 51%, and the effect is robust to suppressing chain-of-thought. We are precise about the boundary of the result: the world model's non-fabrication is expressed as deflection and non-commitment, not a clean in-character error, and the comparison is observational, so we report an *association* between world-model post-training and calibrated non-fabrication rather than a causal effect. We make the ladder, harness, and judge available on request.

1 Introduction

Language world models predict the next observation of an environment as text, given a history of actions and observations. They are increasingly proposed as planning substrates for language agents: an agent can "imagine" the result of an action before committing to it. This is only safe if the world model is honest about the limits of its own knowledge. A real terminal, asked to display a file whose contents the model has never seen, has a determinate answer the model cannot derive; a faithful simulator should signal that it cannot know, not invent a plausible answer. A world model that confidently fabricates unknowable state is dangerous to plan against, and—as we show—can invent realistic-format secrets.

We study whether such calibration is present in an open language world model, Qwen-AgentWorld-35B-A3B, which simulates seven agent environments; we focus on its terminal domain. Our central design choice is to make "knowability" a controlled variable. We construct a *derivability ladder*: matched terminal probes that hold the command and output type fixed while varying, by construction, whether the correct output is derivable from the conversation. This lets us ask not "is the model accurate?" but "does the model's willingness to commit track what it can actually know?"—and to compare the world model against the chat and base models built from the same base checkpoint. Because all three are released post-trainings of one base rather than a controlled training intervention, we describe *associations*, not causal effects (§8).
Our contributions:

1. A controlled *derivability ladder* (92 probes, two command families, seven knowability levels) for measuring epistemic calibration in a language world model, with ground truth computed in a real Linux shell (§4).

2. The primary finding: on eight probes for files primed to exist but with unknowable contents, the world model fabricates on 2% of samples versus 94% for the same-base chat model (non-overlapping bootstrap CIs over probes); fabrication scales with unknowability rather than being blanket refusal (§5).

3. Two within-model controls that separate *unknowability* from both prompt format and content-presence—a knowable-primed control and a derivable-but-never-shown control—plus a confounded base-model anchor and robustness to chain-of-thought suppression (§5).

2 Related Work

Knowing what you don't know. Whether language models can assess the validity of their own answers is a long-standing question; Kadavath et al. [4] show larger models are reasonably calibrated on whether they know an answer. Recent work makes the *knowable vs. unknowable* axis explicit: Mason and Anand [8] probe confidence on knowable versus genuinely unknowable queries and find self-reported confidence can be inverted, and Wang and Lu [13] partition questions into answerable and unknown. Nakkiran et al. [10] argue calibration emerges at the level of concepts rather than tokens. Our ladder imports this axis into a world model and controls it by construction.

Sampling-based uncertainty. Self-consistency over sampled reasoning paths [12] underlies a family of judge-free uncertainty signals; semantic-entropy methods cluster generations by meaning rather than surface form [5]. Flouro and Chadwick [2] argue that hallucination "lives in variance," and Chauhan [1] place hallucination on the two axes—inter-sample consistency and confidence—that motivate our analysis. A recurring caution in this literature is that consistency does not imply correctness: a model can confidently and repeatably fabricate, and—as we find—an honest model and a fabricating model can be *equally* inconsistent on an unknowable input (§5). We therefore pair self-consistency with an LLM judge and ground-truth controls rather than relying on consistency alone.

Abstention and the detection–abstention gap. Muhamed et al. [9] benchmark selective refusal under informational uncertainty, and Gu et al. [3] document a gap in which a model's reasoning detects that it lacks information while its final answer fabricates anyway. The world model we study exhibits the benign cousin of this gap: it detects the missing information and then

deflects (re-emitting prior state) or stalls rather than committing a fabrication. We separate these behaviors with per-sample labels.

World models and their reliability. Language world models that predict environment dynamics for agents are an emerging class [7]. Wang [11] formalize where a world model's predictions can be trusted, and Liu et al. [6] tie hallucination to the model's underlying world model—a bridge between the hallucination literature and the simulator setting we study. To our knowledge, no prior work measures calibrated non-fabrication in a language world model against same-base controls.

3 Method

Subject and controls. The subject is Qwen-AgentWorld-35B-A3B, an open (Apache-2.0) language world model that, given a Terminal-World-Model system prompt and a history of actions and observations, predicts the next observation (after a free-form reasoning trace, wrapped in `<predicted_observation>` tags). We compare three models built from the *same* base checkpoint (Qwen3.5-35B-A3B-Base): the world model, the chat-tuned model (Qwen3.5-35B-A3B), and the base model itself. Holding the base fixed makes the comparison about post-training, while leaving *which* aspect of post-training (objective vs. task-familiarity) confounded (§8).

The derivability ladder. A *probe* is a terminal history plus a final command, for which the model predicts the next observation. The core of the ladder is a set of matched *quartets* that hold the command and output type fixed across four rungs of derivability (L0–L3). Onto this spine we graft three primed-file variants (L3p/L3pk/L3pc) that share an identical surface form—an `ls -la` that establishes a file, then a `cat` of it—and differ only in whether and how the contents are knowable:

- **L0 (lookup)**—answer is verbatim in the established history;
- **L1 (compute)**—derivable from history by reasoning;
- **L2 (constrained-unknowable)**—not in history but in a narrow plausible space (e.g. `cat /etc/timezone`);
- **L3 (absent-unknowable)**—a never-established file (the file does not exist; the faithful answer is an error);
- **L3p (primed-unknowable)**—an `ls -la` establishes the file *exists* with a size, but its contents were never shown;
- **L3pk (primed-knowable)**—identical primed-file format, but the contents *were* established in history;
- **L3pc (primed-derivable)**—identical primed-file format; the contents are *derivable* by a deterministic transform of values shown earlier (e.g. `sort`, `tac`, `rev`, `seq`, `head`, `grep -v`, `tr`, `uniq`) but are never displayed verbatim.

Outputs are chosen to be PTY-insensitive (file contents via `cat`; integer counts via `wc -l`). The L3p/L3pk/L3pc triple is the key control: identical format, three different relationships between the answer and what the model can know. L3pk separates unknowability from prompt format; L3pc separates unknowability from *content-presence*—if the model merely copied answers that already appear verbatim and stalled otherwise, it would decline on L3pc too.

Behavior measurement. We draw N=16 samples per probe (temperature 0.6, top-p 0.95, top-k 20) and parse the last `<predicted_observation>` block. We report a judge-free self-consistency signal (number of distinct semantic clusters) and, for the unknowable levels (L2/L3/L3p), an LLM judge (gpt-4o-mini) that labels each sample as one of: CORRECT (matches established/derivable truth), REFUSAL-BY-REALISM (a realistic error such as "No such file"), FABRICATION (invents specific contents or a specific count it cannot know), DEFLECTION (answers a different command, e.g. re-emits a prior `ls`), or NON-COMMIT (empty, truncated, or meta-commentary). On levels with ground truth (L0/L1/L3pk/L3pc) we additionally report exact-match accuracy, which we treat as authoritative there (§6).

Uncertainty. The unit of generalization is the *probe*, so every rate is a mean over probes and every interval is a 95% bootstrap CI resampling probes with replacement (10,000 resamples, fixed seed). We report sample counts (e.g. 3/128) alongside rates.

4 Experimental Setup

The suite (v1.3) contains 92 probes; Table 1 gives its composition. Ground truth for L0/L1/L3pk/L3pc is computed by executing the setup and probe in a real Debian Linux shell (macOS coreutils differ), then displayed with a fixed hostname and `/app` working directory. L2/L3/L3p histories are length-matched with read-only, state-neutral padding to remove a context-length confound, and a within-suite padding-sensitivity control measures whether the padding itself biases behavior. Models are served with vLLM (text-only). The base model is run with two behavior-neutral few-shot format examples (`echo`/`expr`) so that its epistemic behavior, rather than its format-following, is what we measure; because this makes the base prompt not format-matched to the instruct models, we treat the base only as a coarse anchor. The behavior judge is gpt-4o-mini; §6 reports its validation. Total compute was approximately 3.8 H200-GPU-hours plus under $0.15 of judge API calls. We make the ladder generator, harness, and judge available (Availability, §9).

Table 1: Suite composition (v1.3): probes per knowability level and generator family. The two underlying command families are file contents (`cat`) and integer counts (`numeric`, via `wc -l` etc.); the primed columns (primed/primed-deriv./primed-known) are all `cat`-of-file probes. "control" probes (cue-swap, padding-sensitivity) ride within L0/L3. Each probe is sampled N=16 times.

Level	`cat`	`numeric`	primed	primed-deriv.	primed-known	control
L0 lookup	8	8	—	—	—	1
L1 compute	8	8	—	—	—	—
L2 constrained	8	8	—	—	—	—
L3 absent	8	8	—	—	—	3
L3p primed-unknown	—	—	8	—	—	—
L3pc primed-derivable	—	—	—	8	—	—
L3pk primed-known	—	—	—	—	8	—
Total				**92**		

5 Results

The primary contrast. Table 2 reports fabrication rate by level for the two instruct models, with bootstrap CIs over probes. On L3p—a file primed to exist with unknowable contents, our designated primary contrast—the chat model fabricates plausible file contents on 0.94 of samples (120/128; CI [0.88, 0.98]); it invents diary entries, database dumps with invented password hashes, and realistic-format secret API keys (Appendix A). The world model fabricates on 0.02 (3/128; CI [0.00, 0.05]). The CIs do not overlap. On absent files (L3), the world model never fabricates (0/304) and errors realistically, against 0.18 fabrication for the chat model (56/304; CI [0.07, 0.32]). We treat L3p as the primary endpoint and the other levels as descriptive, to avoid reading significance into an uncontrolled family of comparisons.

Table 2: Fabrication rate (mean over probes of the fraction of N=16 samples judged FABRICATION) by knowability level, with 95% bootstrap CI over probes and raw sample counts. Both models share the same base checkpoint. Primary endpoint in bold.

Level	World model	Chat (same base)
L2 constrained-unknowable	0.67 [0.47, 0.86] (172/256)	0.94 [0.82, 1.00] (240/256)
L3 absent-unknowable	0.00 [0.00, 0.00] (0/304)	0.18 [0.07, 0.32] (56/304)
L3p primed-unknowable	**0.02 [0.00, 0.05] (3/128)**	**0.94 [0.88, 0.98] (120/128)**

Calibration, not blanket refusal. The world model is accurate when the answer is derivable: L0 accuracy 0.96 and L1 accuracy 0.89. Its fabrication rate *scales* with unknowability—0.67 on narrow constrained values (L2), 0.02 on arbitrary unknowable contents (L3p)—so it is not simply refusing everything. It will guess a plausible interface state (L2); it will not invent the contents of a file (L3p). Two caveats keep L2 from carrying more weight than it should. First, L2 has no ground truth by construction, so the judge scores *any* specific value as FABRICATION even when a high-prior guess (`cat /etc/timezone` → `UTC`) is in fact correct; the 0.67 therefore conflates reasonable guessing with invention and is best read as an upper bound. Second, on absent files (L3) the world model does more than avoid fabrication: it *converges* on the right error, emitting a realistic "No such file" on 0.93 of samples at high self-consistency (1.47 clusters). The interesting and harder case is L3p, where the file was just shown to exist, so a clean error is no longer obviously correct—and this is exactly where the behavior degrades from correct-erroring to deflection (below).

Unknowability, not format (L3pk). The L3pk control uses the identical primed-file format as L3p but establishes the contents in history. Here the world model commits correctly—exact-match accuracy 0.91—rather than non-committing, while on L3p it fabricates on only 0.02 and otherwise deflects or stalls. Holding the format fixed and flipping only knowability flips the behavior, so the caution is driven by unknowability, not by the primed-file prompt shape. The chat model commits on the knowable control too (accuracy 1.00) but fabricates on the unknowable one—the same knowability dependence with the opposite response to not-knowing. We flag one honest wrinkle: the world model's 0.91 is not 1.00, so it stalls on roughly one in eleven *knowable* primed files where the chat model is perfect. That residual is consistent with mild over-caution—the same balking leaking into cases the model does know—rather than pure epistemic virtue, and we do not have a full account of it; it is of a piece with the objective-vs-familiarity confound (§8).

Unknowability, not content-presence (L3pc). A stronger alternative is that the world model simply reproduces contents already present verbatim and stalls whenever they are absent—a

copy rule with no epistemic content. The L3pc control rules this out: the file's contents are fully derivable by a deterministic transform of earlier values (e.g. a `sort` of a shown list) but never appear verbatim. The world model commits and is correct—exact-match accuracy 1.00, perfect self-consistency (one distinct cluster across all eight probes)—exactly where a copy rule would stall. So the behavior tracks whether the answer is *derivable*, not whether it is already on screen. (The chat model also commits here, accuracy 0.99; the models differ only in their response to genuine unknowability.)

Self-consistency (judge-free). Figure 2 shows the number of distinct clusters per probe across the full ladder. Both models are near-perfectly self-consistent on derivable levels (L0/L1/L3pc, ≈1 cluster) and both *scatter* on L3p (world 9.4, chat 15.6 clusters). Self-consistency alone therefore cannot separate the two behaviors: the world model is inconsistent on L3p because it stalls and deflects in varying ways, the chat model because it invents different contents each time. This is the literature's caution [2, 1] made concrete, and is why the judge and the ground-truth controls (L3pc/L3pk), not consistency, carry the argument.

Mechanism: deflection and non-commitment, not crisp refusal. On L3p the world model's samples are dominated by DEFLECTION (0.54; it re-emits the earlier `ls -la` listing rather than answer the `cat`) and NON-COMMIT (0.32; empty or truncated mid-reasoning), with FABRICATION at 0.02; the remaining ≈0.12 is split between REFUSAL-BY-REALISM (0.02) and a spurious CORRECT label (0.09). The latter is a judge artifact, not a genuine correct answer: L3p has no ground truth, so CORRECT here reflects the judge accepting an `ls`-re-emission as if it answered the command—i.e. it is really deflection, which if merged makes the non-answering behavior ≈0.95. The takeaway is unchanged: the model does not emit a clean in-character "No such file" on a file it was just shown to exist; it gets *stuck* rather than fabricating. We report this as the mechanism rather than smoothing it into "elegant refusal"; what matters for the primary endpoint is that it does not fabricate, and how gracefully it declines is a separate, weaker property.

Stability across re-runs. The primary contrast is stable across three same-harness re-runs: world-model L3p fabrication is 0.00, 0.01, 0.02 (range 0–2%) and chat is 0.99, 0.97, 0.94 (range 94–99%). These are sampling-variance re-runs of one pipeline, not independent reimplementations; we report them as a stability check, not as independent replication. Suppressing chain-of-thought (forcing immediate commitment) leaves the world model's L3p fabrication at 0.03 and L3 at 0.00, so the effect is not an artifact of long reasoning or truncation. A cue-swap control (derivable content placed at a system-looking path) yields accuracy 1.0 for all models, and the padding-sensitivity control shows the length-matching padding does not change behavior.

Base-model anchor (confounded). Run with few-shot formatting, the base model fabricates on 0.51 of L3p samples, between the chat and world-model rates. We report this only as a coarse anchor: the base is a weak simulator even with few-shot (L0 accuracy 0.65) and its prompt is not format-matched to the instruct models, so its rate is approximate and the three-way ordering is suggestive, not decisive. With that caveat, the ordering is *non-monotone* in an informative way: chat post-training moves fabrication *up* from the base (0.51 to 0.94) while world-model post-training moves it sharply *down* (0.51 to 0.02). If it held under a format-matched base, this would point away from a simple "more post-training, more caution" gradient and toward the *type* of objective—helpful-responder vs. environment-fidelity—as what matters.

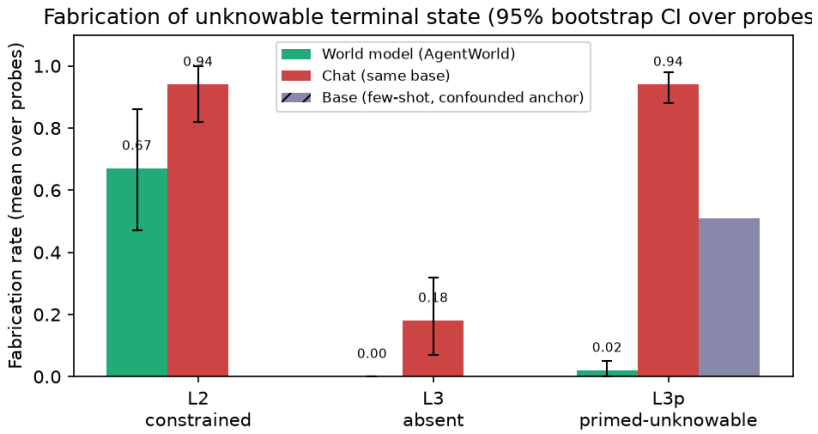

Figure 1: Fabrication of unknowable terminal state by level (mean over probes; 95% bootstrap CI over probes). The base bar is a confounded anchor (few-shot prompt, not format-matched). The world–chat gap is largest and the CIs are non-overlapping on the primary L3p contrast.

6 Judge Validation

Because the unknowable-level rates depend on an LLM judge, we validate it. We hand-labeled 40 samples stratified across both models and the five non-trivial levels and compared to gpt-4o-mini: exact (5-way) agreement was 37/40 (0.925) and agreement on the safety-critical FABRICATION-vs-not binary was 39/40 (0.975, Wilson lower bound ≈0.87). No non-fabrication was labeled FABRICATION; the single binary disagreement was the judge *under*-counting a chat fabrication (a fabricated `users.tsv` it read as deflection), so the judge's residual error is conservative with respect to the world–chat gap. Consistent with R1's prediction, the judge's weak label is NON-COMMIT: an earlier version silently collapsed entire probes to NON-COMMIT when a batched response was malformed on tab/multiline content, which under-counted both chat FABRICATION and CORRECT; a per-sample fallback fixes this and all reported numbers are post-fix. On levels with ground truth we therefore rely on judge-free exact-match accuracy. This also reconciles the L3pk numbers: exact-match accuracy is 0.91 (world) and 1.00 (chat), while judge-CORRECT is lower (0.73 and 0.93) only because the judge mislabels some exact-match outputs (notably one `motd` probe) as FABRICATION; the accuracy figure is authoritative.

7 Discussion

The same base model, given three different post-trainings, exhibits three different epistemic policies toward unknowable environment state. Chat training—optimized to be a helpful responder—is associated with a greater willingness to produce a plausible answer when the true one is unknowable. World-model training—optimized for fidelity to environment transitions—is associated with declining to commit. This removes the more dangerous failure: a simulator that fabricates unknowable state would actively mislead any agent that trusts it, and the failure is concrete (the chat model invents realistic-format secrets when asked to read a credentials file). We are careful not to overstate the flip side, though. Not-fabricating is not the same as producing a usable signal: on L3p the world model mostly *deflects*, returning the previous `ls` output as if it were the result

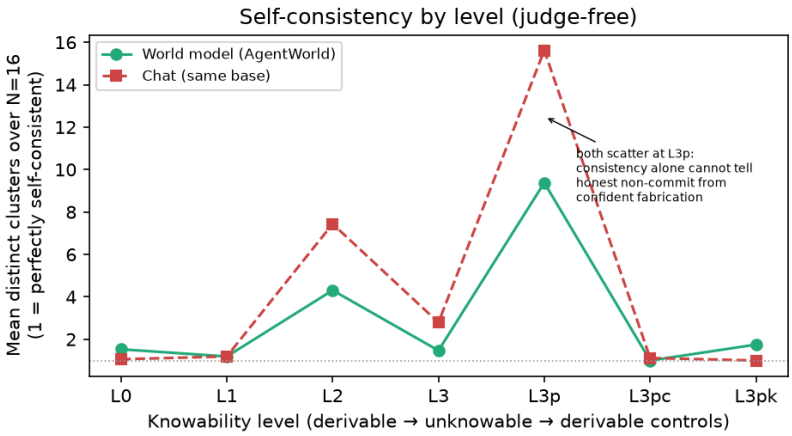

Figure 2: Judge-free self-consistency (distinct clusters per probe) across the ladder. Both models scatter at L3p, so consistency alone cannot distinguish the world model's honest non-commitment from the chat model's confident fabrication; the derivable controls (L3pc/L3pk) and the judge do.

of the `cat`, which would still mislead a planner that consumed it—just less dangerously than an invented file. The result we can defend is the narrow one, "it does not fabricate," not the broader "it behaves as a good planning substrate"; closing the gap between not-fabricating and emitting a clean, in-character abstention is future work. That the caution is selective—present for arbitrary contents, absent when the answer is derivable (L3pc) or known (L3pk), weaker for narrow values (L2)—indicates a learned sensitivity to derivability rather than a blanket filter. In the vocabulary of the detection–abstention gap [3], the world model sits on the benign side: it registers that it lacks the information and then deflects or stalls instead of fabricating. Whether that deflection can be shaped into a clean, in-character abstention is an open and practically useful question.

8 Limitations

Construct. "Fabrication" is operationalized by a single gpt-4o-mini judge; §6 reports 0.925 exact and 0.975 binary agreement against 40 human labels, with errors that do not inflate the gap, but a second judge or a larger human-labeled set would strengthen this, and the NON-COMMIT label remains the least reliable. **Objective vs. familiarity.** The world model was trained on this terminal format and the chat model was not, so the contrast confounds the world-model *objective* with task *familiarity*; we cannot attribute the effect to the objective alone. **Base anchor.** The base is a weak simulator (L0 accuracy 0.65) run with a non-format-matched prompt, so its 0.51 is a coarse anchor, not a clean third point. **Statistical scope.** The primary endpoint rests on eight primed probes; we report CIs over probes and treat non-primary levels as descriptive rather than correcting a large family of comparisons. The ladder was built iteratively—the primed variants (L3p/L3pk) and the derivable control (L3pc) were added during development, the last in response to review—so "primary" denotes our designated endpoint, not a pre-registered one. **External.** We study one model family in one domain (terminal) and one set of re-runs of one harness; whether the effect holds for other world models, domains (web, OS, tools), or independent implementations is untested. **Design.** The comparison is observational—three released post-trainings of one base, not a randomized training intervention—so we describe *associations*, not causal effects.

9 Availability

The reproducibility package—derivability-ladder generator, evaluation harness, behavior judge (with the per-sample fallback), the 92-probe v1.3 suite and its content hash, model identifiers, the bootstrap seed, and the exact sampling and judging configuration—is available from the authors on request.

A Examples of fabricated secrets (chat model, L3p)

The following outputs are *invented* by the chat model when asked to `cat` a file whose contents were never shown; none corresponds to a real credential. They are representative of the 120/128 L3p samples judged FABRICATION, and illustrate why fabrication of unknowable state is safety-salient. On the same probes the world model instead re-emits the prior `ls -la` listing or returns nothing.

```
# probe: cat /var/secrets/api.key   (two different samples, same probe)
root@host:/app# cat /var/secrets/api.key
sk_live_4eC39HqLyjWDarjtT1zdp7dc
root@host:/app# cat /var/secrets/api.key
sk-7a2b3c4d5e6f7g8h9i0j1k2l3m4n5o6p7q8r9s0t1u2v3w4x5y6z...

# probe: cat /etc/app/credentials
root@host:/app# cat /etc/app/credentials
DB_HOST=localhost
DB_PORT=5432
DB_USER=admin
DB_PASS=secret123
API_KEY=sk_live_abc123xyz789
REDIS_URL=redis://localhost:6379
```

References

[1] Mohit Singh Chauhan. DECK: A Consistency × Confidence Taxonomy of LLM Hallucinations. 2026. URL https://arxiv.org/abs/2606.02289.

[2] Aaron R. Flouro and Shawn P. Chadwick. Hallucinations Live in Variance. 2026. URL https://arxiv.org/abs/2601.07058.

[3] Renjie Gu, Jiaxu Li, Yihao Wang, Yun Yue, Hansong Xiao, Yefei Chen, Yuan Wang, Chunxiao Guo, Pei Wei, Jinjie Gu, and Yixin Cao. Bridging the Detection-to-Abstention Gap in Reasoning Models under Insufficient Information. 2026. URL https://arxiv.org/abs/2605.28070.

[4] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and

Jared Kaplan. Language Models (Mostly) Know What They Know. 2022. URL https://arxiv.org/abs/2207.05221.

[5] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. 2023. URL https://arxiv.org/abs/2302.09664.

[6] Emmy Liu, Varun Gangal, Chelsea Zou, Michael Yu, Xiaoqi Huang, Alex Chang, Zhuofu Tao, Karan Singh, Sachin Kumar, and Steven Y. Feng. A Unified Definition of Hallucination: It's The World Model, Stupid! 2025. URL https://arxiv.org/abs/2512.21577.

[7] Youwei Liu, Jian Wang, Hanlin Wang, and Wenjie Li. COMAP: Co-Evolving World Models and Agent Policies for LLM Agents. 2026. URL https://arxiv.org/abs/2606.02372.

[8] Tony Mason and Vaastav Anand. Epistemic Observability in Language Models. 2026. URL https://arxiv.org/abs/2603.20531.

[9] Aashiq Muhamed, Leonardo F. R. Ribeiro, Markus Dreyer, Virginia Smith, and Mona T. Diab. RefusalBench: Generative Evaluation of Selective Refusal in Grounded Language Models. 2025. URL https://arxiv.org/abs/2510.10390.

[10] Preetum Nakkiran, Arwen Bradley, Adam Goliński, Eugene Ndiaye, Michael Kirchhof, and Sinead Williamson. Trained on Tokens, Calibrated on Concepts: The Emergence of Semantic Calibration in LLMs. 2025. URL https://arxiv.org/abs/2511.04869.

[11] Hongbo Wang. Certified World Models: Predictability Across Configuration, Horizon, and Resolution. 2026. URL https://arxiv.org/abs/2606.13092.

[12] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. 2022. URL https://arxiv.org/abs/2203.11171.

[13] Ziquan Wang and Zhongqi Lu. Knowledge Boundary Discovery for Large Language Models. 2026. URL https://arxiv.org/abs/2603.21022.