DECK: A Consistency × Confidence Taxonomy of LLM Hallucinations

Mohit Singh Chauhan
Independent Researcher
mohitcek@gmail.com

Abstract

Existing hallucination taxonomies classify LLM errors by *what is wrong with the output* — memorised misconceptions, reasoning failures, fluent fabrications — but cannot answer a different question: *which uncertainty scorer would have caught this error?* We propose a complementary taxonomy that classifies errors by their **detectability signature**, the signal a scorer family would read. The **DECK taxonomy** is a 2×2 partition along inter-sample consistency and token-level confidence into four regimes (**D**rift, **E**ntrenched, **C**onfabulation, **K**notted) that yields a falsifiable blind-spot map: black-box consistency scorers have signal in D and C, white-box token-probability scorers in K and C, and only an LLM-as-a-Judge with independent pretraining can detect E. Across three models and four short-form QA datasets we test this map two ways: judge-involving scorer disagreements concentrate in each family's predicted blind-spot cells, and external labels (SelfAware unanswerable, HaluEval adversarial, PopQA entity popularity) land in the predicted cells, robustly to cross-fitted thresholds. We further identify a *universal blind spot of output-level UQ*: on knowledge-gap inputs where the generator emits confident, repeatable fabrications, every output-level family collapses by construction. A linear probe on Llama-3-8B's final-layer hidden states also falls to chance, with or without quantisation, though an intermediate layer retains weak signal.

1 Introduction

Existing hallucination taxonomies classify LLM errors by *what is wrong with the output* (Ji et al., 2023; Huang et al., 2025; Wang et al., 2024) — memorised misconceptions, reasoning failures, fluent fabrications, faithfulness violations. These taxonomies are useful for diagnosis, but they cannot answer a question that arises whenever an LLM is deployed in a high-stakes setting: *which uncertainty scorer would have caught this error?* The same content category can produce different detection signatures across models (a "reasoning failure" is detectable by a Judge if the answer is stable and high-probability, but only by an inter-sample consistency check if the answer varies); and the same detection signature can come from very different content categories. Without a taxonomy on the detection axis, choosing a scorer for a new domain stays a guess — even though the literature already offers three scorer families: **Black-box** (BB) consistency checks (Manakul et al., 2023; Kuhn et al., 2023), **White-box** (WB) token-level log-probabilities (Kadavath et al., 2022; Malinin and Gales, 2021), and **LLM-as-a-Judge** (J) factual review (Zheng et al., 2023), which have been shown to ensemble well together (Bouchard and Chauhan, 2025).

We propose a complementary taxonomy that classifies errors by their detectability signature — the signal a scorer family reads. Each of our two axes already underpins existing scorers (consistency for BB, token probability for WB); the contribution is what follows from crossing them explicitly: a *falsifiable per-family blind-spot map* (§3) that predicts which family will miss which hallucination, and two independent tests of those predictions (§5.2, §5.3) that could have failed but did not.

(1) The DECK taxonomy (Table 1): a 2×2 partition along inter-sample consistency and token-level confidence into four behavioural regimes (D, E, C, K). Each cell maps to a specific scorer family (or families) that can detect it: BB has signal in D and C; WB has signal in K and C; only J can detect E (and only when the judge has independent pretraining). For each scorer axis we pick the threshold that best separates correct from hallucinated responses (Youden's J), then assign each sample to one of the four cells. We test the map in two ways:

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3265–3285
November 6-10, 2026 ©2026 Association for Computational Linguistics

judge-involving scorer disagreements — the Judge being a third axis not used to define the cells — should concentrate in each family's predicted blind-spot cells (§5.2); and external labels (SelfAware unanswerable, HaluEval adversarial, PopQA entity popularity), which involve no scorer beyond the cell split, should land in the predicted cells (§5.3). Both hold, with model-scale and content-specific secondary-cell refinements, and the cell assignment is stable under cross-fitted thresholds (Appendix H). Our experiments cover short-form QA; §3.3 outlines the extension to long-form generation. **(2) A universal blind spot of output-level UQ.** We identify an empirical regime where every scorer in the three-family output-level paradigm fails simultaneously (§5.4). On SelfAware all three families collapse to or below chance because the generator emits confident, repeatable fabrications on inputs it cannot answer — eliminating the variance every output-level family needs to be informative. The right engineering response is an abstention envelope that routes such out-of-scope inputs to refusal before scoring. "Universal" here means across the three output-level families studied. We also ask whether the failure persists in the activations: a linear probe on Llama-3-8B's last-layer hidden states falls to chance on SelfAware with or without 4-bit quantisation, but an intermediate layer retains weak signal (§5.4), which bounds the blind spot to output-level signals.

2 Background and Related Work

2.1 Three Families of UQ Scorers

Output-level UQ methods for LLMs fall into three families distinguished by the access they require (Shorinwa et al., 2025; Kang et al., 2026). *White-box* methods read confidence from token log-probabilities; representative scorers include `P(True)` (Kadavath et al., 2022), sequence-probability measures (Malinin and Gales, 2021), and calibration-aware variants over token distributions (Jiang et al., 2021; Duan et al., 2024). They require logit access and are cheap. A related thread elicits *verbalized* confidence by asking the model to state its own uncertainty (Tian et al., 2023; Xiong et al., 2024). *Black-box* methods read confidence from inter-sample agreement: SelfCheckGPT (Manakul et al., 2023) uses NLI- and embedding-based consistency, while semantic entropy (Kuhn et al., 2023; Farquhar et al., 2024) clusters responses into meaning classes and computes entropy over the resulting distribution. Subsequent work explores alternative aggregations of the consistency signal: pairwise-NLI aggregation across sampled responses (Lin et al., 2024) and semantic-density estimation in embedding space (Qiu and Miikkulainen, 2024). These methods are model-agnostic but cost multiple generations per query. *LLM-as-a-Judge* methods delegate evaluation to a separate model (Zheng et al., 2023; Liu et al., 2023); they detect failures invisible to either probabilistic signal but inherit known judge biases (Chiang and Lee, 2023) and risk shared-pretraining errors when generator and judge are related (Feng et al., 2024). Recent open-source aggregators that combine all three families have shown empirically that ensembles outperform individual scorers but offer no mechanistic account of *why* a given family complements a Judge. A fourth, internal-state family — UQ heads (Shelmanov et al., 2025), hidden-state probes (Slobodkin et al., 2023), and information-theoretic estimators (Yadkori et al., 2024) — reads the residual-stream activations that produce the token logits a white-box scorer would consume. We treat internal-state UQ as a deeper access tier of the same signal flow rather than a fundamentally distinct family; one representative is evaluated as a robustness check in §5.4, broader internal-state methods are out of scope.

2.2 Aleatoric, Epistemic, and Confabulation

The classical aleatoric/epistemic decomposition (Hüllermeier and Waegeman, 2021) remains useful at LLM scale: epistemic uncertainty is the component most tightly linked to hallucination (Yadkori et al., 2024; Shorinwa et al., 2025). Across black-box UQ work, the canonical failure mode of interest is *confabulation*: fluent, wrong outputs that vary across stochastic samples (typically obtained by temperature sampling rather than re-seeded decoding). Farquhar et al. (2024) formalise this notion and validate semantic entropy against it, but consistency-based scoring across SelfCheckGPT, semantic entropy, and their successors all target the same underlying error class. We adopt the term *confabulation* unchanged for our low-consistency, low-confidence cell; the categories Farquhar et al. set aside (misconceptions, lies (Evans et al., 2021), reasoning failures) share a high-consistency, high-confidence signature and constitute our *Entrenched* cell. Our contribution is to make this implicit partition explicit as a full 2×2 taxonomy and use it to explain the complementar-

Table 1: The **DECK taxonomy** of LLM hallucinations. Each cell is a behavioral regime defined by two observable axes: inter-sample consistency and token-level confidence. The four regimes map to the letters **D**rift, **E**ntrenched, **C**onfabulation, and **K**notted. The final column gives the scorer family (or families) sensitive to each regime.

	Low Consistency	High Consistency	Detectable By
High Token Confidence	**D — Drift** *Different confident wrong answer each sample.* Multiple plausible-but-wrong answers are drawn confidently from the output distribution; each draw is internally coherent, but the answer drifts.	**E — Entrenched** *Same confident wrong answer every sample.* Model has locked onto a memorized misconception or shared-pretraining error and reproduces it without variance.	**Drift:** Black-box and Judge. **Entrenched:** Judge (independent provider required); blind to both BB and WB.
Low Token Confidence	**C — Confabulation** *Different low-probability wrong answer each sample.* Classic uncertainty: the model genuinely does not know, and different samples produce different wrong answers with low confidence.	**K — Knotted** *Same low-probability wrong answer every sample.* Model is consistently unsure: it settles on the same hedged answer each time but assigns low token probability.	**Confabulation:** All three families. **Knotted:** White-box and Judge.

Note. Regimes are defined by observable output behavior, not by an internal mechanism claim. Cell membership depends on the scorer thresholds used to discretize the two axes; described in §3.

ity patterns that prior ensembling work reports but does not analyse mechanistically.

3 A Taxonomy of LLM Hallucination Types

3.1 Two-Axis Framework

We characterize LLM hallucinations along two observationally independent axes: inter-sample consistency (does the model produce the same incorrect answer across multiple independent samples?) and token-level confidence (does the model assign high probability to the tokens of its generated response?). This yields four hallucination types, summarized in Table 1, each implying a distinct detection strategy.

3.2 Why Each Family Has a Blind Spot

BB scorers measure consistency: if a model produces the same wrong answer every sample, the scorer sees high agreement and assigns high confidence—*Entrenched and Knotted hallucinations evade detection.* The Knotted blind spot is sharper than it might appear: high inter-sample agreement looks like high confidence to a BB scorer *regardless* of token-level probability, so even a low-probability "locked-in" hedge that the model repeats every sample (the Knotted signature) registers as confident-and-correct. WB scorers measure token probability: a model may assign high probability to a fluent but factually incorrect claim—*Entrenched and Drift hallucinations evade detection.* LLM judges evaluate factual content and in principle detect any type when the judge's knowledge is accurate. However, they are susceptible to the same systematic misconceptions as the generator when both share pretraining data—*shared Entrenched hallucinations are the failure mode.* More strikingly, on knowledge-gap inputs (questions genuinely unanswerable for any LLM) on which the generator nevertheless emits a confident, repeatable fabrication, all three families fail *by construction*: BB scorers see consistent confident answers, WB scorers see high token probabilities, and judges share the same knowledge gap. This represents a regime — the *universal blind spot* of output-level UQ — that generally evades detection by any output-level scorer family. This asymmetric coverage motivates the per-pair disagreement analysis in §5.2, which tests whether the predicted quadrant signatures emerge empirically.

3.3 Extending DECK to Long-Form Generation

Our experiments use short-form QA, where one response carries one factual claim. In long-form generation a response bundles many claims that can sit in different cells. DECK extends naturally by moving both axes from the response to the claim. Following FActScore (Min et al., 2023), factor each response into atomic claims; measure *claim-level consistency* as the fraction of stochastic samples that contain a matching or entailed claim (Manakul et al., 2023; Zhang et al., 2024); and measure *claim-level confidence* from the token probabilities of the span that expresses the claim. Each claim then receives its own cell, and a response becomes a *cell mixture*: a biography might carry an Entrenched claim that repeats a common misconception alongside Confabulation claims about obscure dates. The blind-spot predictions carry over per claim, so the mixture indicates which scorer families a given response needs. Two problems remain open: *claim alignment* across stochastic samples

that paraphrase or reorder content, and *confidence attribution* when the same claim is realised by different token spans. We leave claim-level validation to future work; all headline claims in this paper are scoped to short-form QA.

4 Experimental Setup

4.1 Datasets

TriviaQA (Joshi et al., 2017) contains factual trivia where larger models are prone to memorizing systematic misconceptions—expected to exercise the Entrenched cell. **HaluEval** (Li et al., 2023) provides curated hallucinated answer pairs with binary labels, skewing toward Entrenched and Knotted types. **SelfAware** (Yin et al., 2023); we use a 500-question subset of its unanswerable questions to probe model abstention behavior, targeting the Entrenched cell for knowledge-gap hallucinations. **PopQA** (Mallen et al., 2023) contains entity-centric questions stratified by Wikipedia page-view popularity; the popular- and rare-entity slices target the Entrenched and Confabulation cells, respectively. Response grading procedures for all four datasets are detailed in Appendix B.

4.2 Models

We evaluate three models spanning a range of scale and provider: **Llama-3-8B** (smaller, open-weights), **GPT-4o** (larger, closed-weights), and **Gemini-2.5-Flash** (mid-scale, closed-weights, providing a third provider to test generality). All models are evaluated on all four datasets: TriviaQA, HaluEval, SelfAware, and PopQA.

4.3 Scorer Selection

We compute confidence scores using the UQLM toolkit (Bouchard et al., 2026) across 6 black-box, 5 white-box, and 4 judge configurations, giving 15 scorers per model-dataset pair (full AUROC breakdown in Appendix C, Table 7). For the disagreement and complementarity analysis, each family is represented by its single highest-AUROC method on the target split, avoiding hand-selection bias. A full reference table is provided in Appendix A.

Black-box: Semantic Negentropy, Non-Contradiction, Entailment, Cosine Similarity, Exact Match, Semantic Sets Confidence.

White-box: Sequence Probability, Min Token Probability, Mean Token Negentropy, Probability Margin, P(True).

LLM-as-a-Judge: Four judges per generator span increasing generator independence: the model itself (self-judge, lower bound on utility), the other two generators (cross-LLM judges, reducing shared-pretraining bias), and Claude Sonnet 4.6 (independent-provider, upper bound on independence). All judges use a Likert scoring template; multi-generation scorers use five sampled responses at temperature 1.0; $k = 15$ for top-k token-probability scorers.

4.4 Evaluation Metrics

We measure detector quality with AUROC per scorer (Appendix C, Table 7) and complementarity between scorer families with the hallucination-restricted complementarity score $C_H(A, B)$. Let $\mathcal{H} = \{i : \neg\,\text{correct}_i\}$ be the hallucinated-sample subset of size $N_H = |\mathcal{H}|$. $C_H(A, B)$ is the fraction of $\mathcal{H}$ on which the binary predictions of scorer A and B disagree:

$$C_H(A, B) = \frac{|\{i \in \mathcal{H} : \hat{y}_i^{(A)} \neq \hat{y}_i^{(B)}\}|}{N_H} \quad (1)$$

where $\hat{y}_i^{(A)} = \mathbf{1}[s_i^{(A)} < \tau^{(A)}]$ is scorer A's binary hallucination prediction at threshold $\tau^{(A)}$. We choose $\tau^{(A)}$ as the Youden's J optimal threshold $\tau^{(A)} = \arg\max_\tau [\text{TPR}(\tau) - \text{FPR}(\tau)]$ fit on the full $N = 500$ samples of each condition (positive class = hallucinated). YJ is the standard ROC-corner criterion: it maximises balanced accuracy and is base-rate robust, so it does not collapse on imbalanced datasets the way (e.g.) F1-optimal thresholds do. The same threshold defines DECK cell membership (§3); a step-by-step computation example is provided in Appendix D. Cell assignments are stable under 5-fold cross-fitted YJ (per-sample $\kappa = 0.70$–0.96; dominant cell unchanged in $11/12$ external-signal rows); median and F1-optimal alternatives are compared in Appendix H.

5 Results

5.1 Individual Scorer Performance

Figure 1 shows the top-AUROC scorer per family across all twelve dataset–model combinations; per-method numerical values with 95% bootstrap CI are in Table 7 (Appendix C). Three patterns stand out.

First, no scorer family Pareto-dominates: on Llama-3-8B TriviaQA the Judge leads (0.844), while on GPT-4o TriviaQA WB leads (P(True)

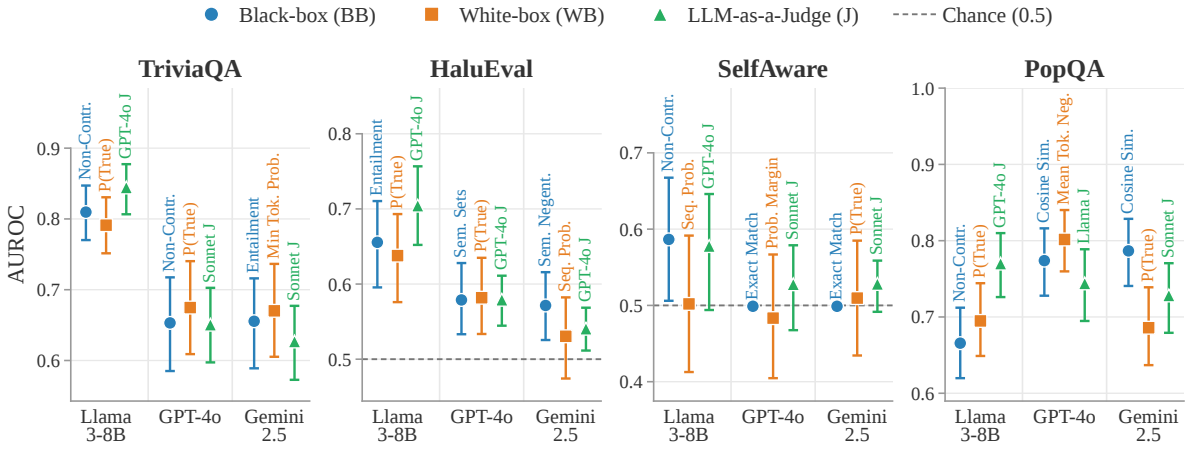


Figure 1: Best-AUROC scorer per family for all twelve dataset–model combinations, with 95% bootstrap CIs. Marker shape encodes family (• BB, ■ WB, ▲ Judge); the selected scorer is labelled above each marker. Full per-method bar plots in Figure 3 and exact values with CIs in Table 7 (Appendix C).

Table 2: Top-AUROC scorer in each family (BB / WB / Judge) for every (model, dataset) combination.

Dataset	Model	BB	WB	Judge
TriviaQA	Llama-3-8B	Non-Cont. (0.810)	P(True) (0.791)	GPT-4o (0.844)
	GPT-4o	Non-Cont. (0.653)	P(True) (0.675)	Sonnet (0.651)
	Gemini-2.5	Entailment (0.655)	Min Prob. (0.670)	Sonnet (0.627)
HaluEval	Llama-3-8B	Entailment (0.656)	P(True) (0.638)	GPT-4o (0.704)
	GPT-4o	Sem. Sets (0.579)	P(True) (0.582)	GPT-4o (0.579)
	Gemini-2.5	Sem. Neg. (0.571)	Seq. Prob. (0.530)	GPT-4o (0.541)
SelfAw.	Llama-3-8B	Non-Cont. (0.587)	Seq. Prob. (0.502)	GPT-4o (0.578)
	GPT-4o	Exact Match (0.499)	Prob. Margin (0.483)	Sonnet (0.528)
	Gemini-2.5	Exact Match (0.499)	P(True) (0.510)	Sonnet (0.528)
PopQA	Llama-3-8B	Non-Cont. (0.666)	P(True) (0.695)	GPT-4o (0.770)
	GPT-4o	Cosine Sim. (0.774)	Mean Tok. (0.802)	Llama (0.744)
	Gemini-2.5	Cosine Sim. (0.787)	P(True) (0.686)	Sonnet (0.728)

0.675); on PopQA the BB Cosine Similarity reaches 0.77–0.79 for GPT-4o and Gemini-2.5-Flash (and WB Mean Token Negentropy 0.80 for GPT-4o), but the Judge takes the lead on Llama-3-8B (GPT-4o cross-judge 0.770). Second, absolute AUROC compresses with model scale: top scorers from families that span 0.79–0.84 on Llama-3-8B TriviaQA cluster at 0.65–0.68 on GPT-4o and 0.63–0.67 on Gemini-2.5-Flash, consistent with the Entrenched-shift predicted by the DECK taxonomy. Third, PopQA has the highest per-condition top-scorer AUROC on average and SelfAware the lowest; on SelfAware all methods collapse to near-random across all three models (detailed in §5.4). Since no family dominates, ensemble gains depend on structural disagreement, not simply on selecting the highest-AUROC scorer—motivating §5.2.

5.2 Disagreement Analysis and Taxonomy Validation

We conduct the disagreement analysis across all twelve combinations of model (Llama-3-8B, GPT-4o, Gemini-2.5-Flash) and dataset (TriviaQA, HaluEval, SelfAware, PopQA). For each combination, the best-AUROC scorer per family is selected from all available scorers; Table 2 shows the selected representative for each cell (model, dataset, family). Each sample is then assigned to one of the four taxonomy quadrants by a Youden's J optimal split on the BB (consistency) and WB (confidence) axes (§4, Eq. 1). Throughout this section, we report the hallucination-restricted complementarity C_H under YJ-threshold split, with per-combination values for all three family pairs are reported in Table 3.

Evidential hierarchy: what validates the taxonomy. The DECK taxonomy's mechanistic claims are supported at two levels of evidence. Primary evidence comes from (i) *judge-involving pairs* (BB–Judge and WB–Judge), because the Judge score is a third axis not used to define the four quadrants, making its disagreement distribution a genuinely independent test of the predicted coverage patterns; and (ii) *external-signal validation* (§5.3), which checks taxonomy predictions against the entity-popularity stratification in PopQA without using any scorer output. Secondary evidence comes from the *BB–WB geometric check*: because the quadrant cells are defined by the same BB and WB scores used to represent those families, BB–WB disagreements are constrained by construction to land on

the anti-diagonal (Knotted and Drift). This pattern is therefore a structural inevitability, not a discovered regularity—it confirms internal consistency but cannot validate the taxonomy against an independent signal. The quantitative results below are presented in this order: judge-involving pairs first (primary), BB–WB check second (secondary).

Table 3: Hallucination-restricted complementarity C_H under Youden's J thresholds for all twelve combinations, with 95% bootstrap CI (1000 resamples). *Bold* = larger value between (BB, J) and (WB, J).

Dataset	Model	BB, WB	BB, J	WB, J
TriviaQA	Llama-3-8B	0.249 ±0.056	0.229 ±0.056	**0.244** ±0.059
	GPT-4o	0.306 ±0.096	**0.306** ±0.093	0.204 ±0.079
	Gemini-2.5	0.214 ±0.085	0.500 ±0.108	**0.571** ±0.104
HaluEval	Llama-3-8B	0.328 ±0.044	0.282 ±0.045	**0.352** ±0.047
	GPT-4o	0.208 ±0.041	**0.236** ±0.045	0.137 ±0.035
	Gemini-2.5	0.423 ±0.052	0.255 ±0.047	**0.408** ±0.052
SelfAw.	Llama-3-8B	0.312 ±0.040	**0.315** ±0.043	0.114 ±0.029
	GPT-4o	0.291 ±0.042	0.044 ±0.018	**0.327** ±0.042
	Gemini-2.5	0.298 ±0.041	0.048 ±0.020	**0.327** ±0.045
PopQA	Llama-3-8B	0.347 ±0.053	0.292 ±0.051	**0.321** ±0.051
	GPT-4o	0.309 ±0.075	**0.520** ±0.080	0.513 ±0.079
	Gemini-2.5	0.264 ±0.063	0.390 ±0.075	**0.412** ±0.073

Judge-involving pairs (primary validation). Because the Judge is a third axis not used to define the four quadrants, its disagreement distribution is the primary test of the DECK taxonomy. Figure 2 visualises the per-cell disagreement shares across all four datasets. On TriviaQA and HaluEval, judge-involving disagreements spread across all four quadrants as predicted: e.g. on Llama-3-8B TriviaQA, BB–Judge disagreements split between Knotted (36%, BB's structural blind spot that the Judge can recover) and Confabulation (40%, where the Judge can diverge from BB's low-consistency flag); WB–Judge on the same combination splits between Drift (32%, WB's structural blind spot that the Judge can recover) and Confabulation (38%). The predicted BB blind-spot pattern shows most strongly on Llama-3-8B PopQA, where BB–Judge disagreements concentrate in K (49%) and E (23%) — placing 72% of the disagreement set inside BB's predicted high-consistency blind spots. On GPT-4o/Gemini SelfAware the degenerate two-cell regime constrains all pairs to Entrenched and Knotted; BB–Judge disagreements nearly vanish ($n \approx 20$ samples, $\sim 4\%$ of N_H) while WB–Judge dominates ($n \approx 145$ samples, $\sim 33\%$ of N_H), with BB and Judge sharing correlated near-random errors and only WB diverging. Quantitatively, C_H for judge-involving pairs varies substantially across combinations (0.044–0.571, see Table 3) and tracks the predicted taxonomy profile.

The BB–WB pair is a secondary geometric check rather than independent validation: BB–WB disagreements are constrained *by construction* to the anti-diagonal (Knotted and Drift) because the quadrant cells are defined by the same BB and WB scores. Empirically, this is what we see (Table 3, (BB, WB) column; Figure 2).

5.3 External-Signal Validation

A second, independent test of the taxonomy uses external labels that identify the *mechanism* of a hallucination, then ask whether samples with each mechanism land in the DECK cell that mechanism predicts. Unlike §5.2, this test does not depend on any UQ scorer beyond the (BB, WB) YJ split that defines cell membership; it asks whether that split is mechanistically informative.

Predictions and method. Knowledge-gap inputs (SelfAware *unanswerable*) should land in Entrenched; adversarial inputs (HaluEval) should concentrate in Knotted or Entrenched; popular-entity inputs (PopQA *popular_entity*) in Entrenched; rare-entity inputs (PopQA *rare_entity*) in Confabulation. We assign each sample to a DECK cell using the within-condition Youden's J split, then test whether the external-label distribution differs from a naturalistic baseline ($df = 3$; full protocol in Appendix E). PopQA labels use the original Wikipedia monthly page-view counts s_{pop} released with PopQA (Mallen et al., 2023): $s_{\text{pop}} \geq 10000$ $\Rightarrow$ popular (n=45), $s_{\text{pop}} \leq 100 \Rightarrow$ rare (n=115), middle stratum is the naturalistic baseline (n=339).

Evidence strength. The rows differ sharply in statistical weight. SelfAware and HaluEval rows use n=500 per combination and an asymptotic χ^2 test. The PopQA strata are small — the popular-entity stratum has only n=45 (minimum expected count 3.98 for Llama-3-8B) — so for all PopQA rows we replace the asymptotic χ^2 with a Monte-Carlo permutation test (10^5 label shuffles) and report Cramér's V as effect size. We treat the *popular-entity* results as *suggestive* and weight the SelfAware, HaluEval, and rare-entity rows as the primary external-signal evidence.

Results. Table 4 reports per-cell percentages, p-values, and (for PopQA) effect sizes, with both

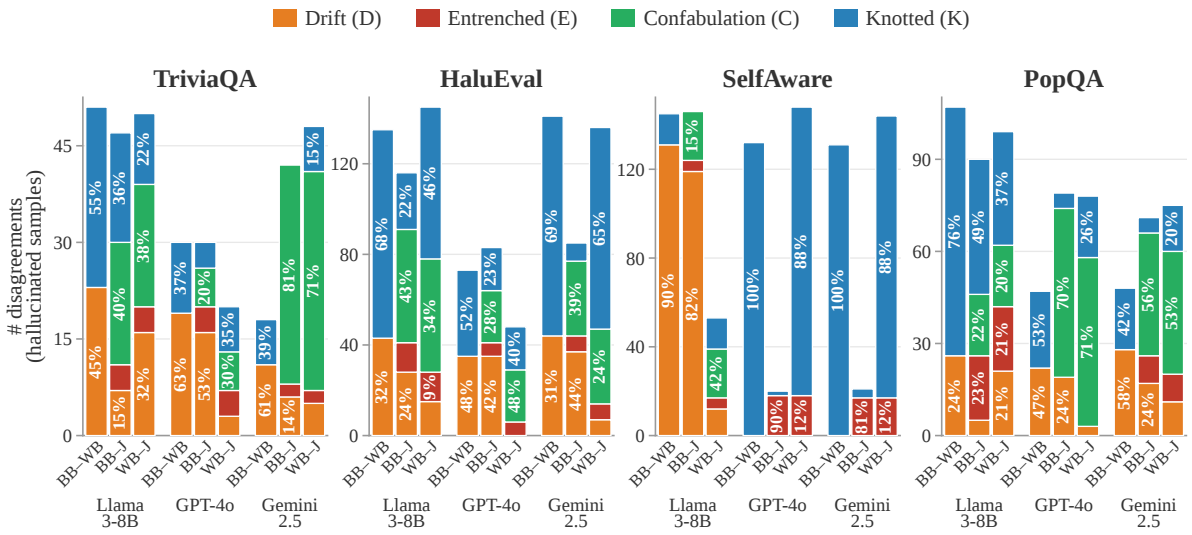


Figure 2: Pairwise scorer-family disagreement on hallucinated samples, one subplot per dataset. Each bar's height is the number of disagreements for one (model, family pair); segments give the DECK quadrant share, labelled in % where space allows (Youden's J split, matching Table 3).

Table 4: External-signal validation: share of samples (%) in each DECK cell under a Youden's J split, by external label; Predicted is the taxonomy's prediction (§3). p: asymptotic χ^2 for SelfAware and HaluEval, permutation test with Cramér's V for PopQA (Appendix E). § Minimum expected count < 5. ‡ Flips to C under cross-fitted thresholds (Appendix H).

Dataset	Model	External label	Predicted	Observed	n	D (%)	E (%)	C (%)	K (%)	p	V
SelfAware	Llama-3-8B	unanswerable	E	E	500	27.0	**62.0**	8.2	2.8	<0.001	–
SelfAware	GPT-4o	unanswerable	E	E	500	0.0	**71.8**	0.0	28.2	<0.001	–
SelfAware	Gemini-2.5	unanswerable	E	E	500	0.0	**71.2**	0.0	28.8	<0.001	–
HaluEval	Llama-3-8B	adversarial	E/K	*E/C/K*‡	500	10.4	**35.8**	**31.6**	**22.2**	<0.001	–
HaluEval	GPT-4o	adversarial	E/K	*E*	500	11.6	**64.6**	14.2	9.6	<0.001	–
HaluEval	Gemini-2.5	adversarial	E/K	E/K	500	12.4	**49.2**	9.4	**29.0**	<0.001	–
PopQA	Llama-3-8B	popular_entity	E	E	45	2.2	**86.7**	4.4	6.7	<0.001§	0.43
PopQA	GPT-4o	popular_entity	E	E	45	15.6	**66.7**	8.9	8.9	0.16	0.12
PopQA	Gemini-2.5	popular_entity	E	E	45	15.6	**53.3**	15.6	15.6	0.81	0.05
PopQA	Llama-3-8B	rare_entity	C	*C/K*	115	8.7	13.0	**50.4**	**27.8**	0.06	0.13
PopQA	GPT-4o	rare_entity	C	*E/C*	115	20.0	**29.6**	**38.3**	12.2	<0.001	0.25
PopQA	Gemini-2.5	rare_entity	C	*E/C*	115	11.3	**31.3**	**38.3**	19.1	<0.001	0.20

the taxonomy's predicted concentration cell and the empirically observed cell(s) shown. Seven of twelve rows match exactly; the remaining five differ in a single direction — four extend the prediction with a secondary cell, and one (HaluEval × GPT-4o) is narrower than predicted (E alone rather than E/K).

SelfAware lands in Entrenched across all three models (62.0%, 71.8%, 71.2%; all $p < 0.001$), matching the predicted E. Llama-3-8B has a sizable Drift tail (27.0%) alongside E; GPT-4o and Gemini-2.5-Flash sit in the degenerate two-cell regime of §5.2: exact_match (the best-AUROC BB scorer on both conditions) returns 0 on every sample (no response literally matches an unanswerable question), so the binary BB prediction is constant. The ∼29% K mass is then an artifact of this constant BB classifier rather than a separate Knotted observation.

HaluEval adversarial inputs show the most model-dependent pattern. GPT-4o concentrates sharply in E alone (64.6%, $p < 0.001$ — the table's cleanest single-cell observation, with no detectable Knotted mass), Gemini-2.5-Flash matches the predicted E/K split (49.2% E + 29.0% K), and Llama-3-8B spreads across E (35.8%), C (31.6%),

and K (22.2%), suggesting smaller generators mix Entrenched with Confabulation under adversarial probes. This Llama-3-8B row is also the one threshold-sensitive row in the table: its E/C near-tie flips to C under cross-fitted YJ (Appendix H).

PopQA *popular-entity* samples (suggestive; n=45) have Entrenched as the plurality cell in all three models (53.3–86.7%), matching the predicted E, but the departure from the baseline is statistically clear only for Llama-3-8B (86.7%, permutation $p < 10^{-4}$, $V = 0.43$). For GPT-4o (66.7%, $p = 0.16$, $V = 0.12$) and Gemini-2.5-Flash (53.3%, $p = 0.81$, $V = 0.05$) the effect sizes are small: their middle-stratum baselines are themselves E-heavy, so the popular subset is not distinguishable from them at this sample size. PopQA *rare-entity* samples lead in Confabulation at 38.3–50.4% (Llama-3-8B 50.4%, $p = 0.06$, $V = 0.13$; GPT-4o 38.3%, $p < 10^{-4}$, $V = 0.25$; Gemini-2.5-Flash 38.3%, $p = 4 \times 10^{-4}$, $V = 0.20$). The secondary cell, however, diverges across models: GPT-4o and Gemini show Entrenched as the secondary (29.6–31.3%), matching the E/C pattern, whereas Llama-3-8B has Knotted secondary (27.8%, the C/K pattern). The GPT-4o/Gemini E secondary has a simple explanation: PopQA questions use fixed templates (e.g. *'What is X's occupation?'*); on rare entities the model sometimes settles on the most common answer for that template (e.g. *'politician'*) and repeats it confidently across samples, producing the Entrenched signature. Llama's Knotted secondary indicates the same template-driven repetition but with visibly low token-level confidence on the (still-wrong) repeated answer, consistent with Llama's higher hedging rate.

Together, the four predicted concentrations (SelfAware → E, HaluEval → E/K, PopQA popular → E, PopQA rare → C) hold as the dominant signals across all twelve rows (11/12 under cross-fitted thresholds), with the statistically strongest support from SelfAware, HaluEval, and the rare-entity stratum. This provides mechanistic evidence that the YJ-split DECK cells capture genuine hallucination structure independent of scorer geometry. Model-scale and content-specific secondary cells emerge as refinements rather than contradictions.

5.4 Universal Failure on Knowledge-Gap Hallucinations (SelfAware)

Mechanism. Output-level UQ scoring is built on three signals — inter-sample agreement (BB), token probability (WB), and judge factual review (J) — each of which requires a corresponding form of variance in the generator's behaviour to be informative. When the generator cannot answer an input (post-knowledge-cutoff facts, unanswerable questions) but emits a confident, repeatable fabrication anyway, every output-level signal collapses at once: BB sees uniform agreement, WB sees high token probability, and J shares the same knowledge gap from common pretraining. Output-level UQ *must* fail in this regime by construction; the failure is a property of the (generator, task) pair, not of any specific scorer family.

Empirical confirmation. GPT-4o (90.6%), Llama-3-8B (92.8%), and Gemini-2.5-Flash (88.0%) of SelfAware responses are graded as confident fabrications; every scorer family then collapses toward chance, and several invert below it (Table 7). On GPT-4o, 13 of 15 scorers fall below 0.5 and P(True) inverts from 0.675 on TriviaQA to 0.331 here, the single worst score in the table, because the self-evaluation signal mirrors the generator's confident-and-wrong distribution and anti-correlates with appropriate abstention. We term this regime the **universal blind spot of output-level UQ**: *universal* means across the three output-level families studied here.

Internal-state probe. Motivated by evidence that hidden states can encode answerability (Slobodkin et al., 2023), we ask a narrower question than that work: does the output-level failure regime persist in the activations of our one open-weights generator? This is not a replication — model family (Llama-3-8B vs. Flan-T5/OPT), metric (AUROC vs. F1), and data (SelfAware vs. SQuAD/NQ/MuSiQue) all differ. We train a logistic probe on Llama-3-8B's last-layer hidden states, captured under NF4 4-bit quantisation, with TriviaQA correctness labels. The probe is well calibrated in-distribution (TriviaQA AUROC 0.72) but collapses on SelfAware (AUROC 0.44, 95% CI $[0.35, 0.52]$); a within-SelfAware 50/50 probe gives 0.56, also CI-overlapping chance. On its own this result is *inconclusive under the specific conditions tested*: a single layer, a single model, a single linear probe, and NF4 quantisation, whose effect on hidden-state fidelity for probing is unstudied. We therefore replicate it without quantisation (fp16) at layers 8, 16, 24, and 32, keeping inputs and protocol fixed (Appendix I).

fp16 and multi-layer replication. Quantisation does not explain the collapse: fp16 and NF4 last-layer states are nearly identical (mean cosine similarity 0.98), and the fp16 last-layer probe again falls to chance (AUROC 0.43, CI $[0.33, 0.52]$). Layer choice matters more. Probes on layers 8 and 24 recover weak but above-chance signal on SelfAware (AUROC 0.63, CI $[0.54, 0.72]$; 0.61, CI $[0.51, 0.71]$), meeting the 0.60 threshold we fixed in advance, while layer 16 stays at chance (0.47). After Bonferroni correction across the four layers only layer 8 remains distinguishable from chance ($[0.51, 0.73]$). A probe trained within SelfAware stays at chance at every layer, but with only 18 correct training samples it has little power. We report this as a boundary of the blind-spot claim: the collapse is robust in the output-level signals and the final layer, but an intermediate layer retains weak signal, so the blind spot does not extend to the model's activations as a whole. Richer internal-state methods — UQ heads (Shelmanov et al., 2025), information-theoretic estimators (Yadkori et al., 2024) — may recover more of it.

Implication. The right output-level engineering response is an abstention envelope (Feng et al., 2024; Kalai et al., 2025) that routes out-of-scope inputs to refusal or retrieval before scoring; the ensemble comparison in Appendix G corroborates this — no choice of ensemble weights moves point-estimate AUROC out of $[0.41, 0.61]$ on SelfAware. Where hidden states are accessible, the intermediate-layer signal suggests internal-state detectors as a complementary route.

6 Conclusion

We introduced the **DECK taxonomy**, a 2×2 partition of LLM hallucinations along inter-sample consistency and token-level confidence. Unlike prior taxonomies that classify errors by *what is wrong with the output*, DECK classifies by the *detectability signature* an error leaves on the (consistency, confidence) plane. The contribution is not the two axes, which underpin existing scorers, but the falsifiable per-family blind-spot map that follows from crossing them (§3), together with two independent tests of it. The disagreement analysis (§5.2) confirms the map on judge-involving pairs across nine non-degenerate (model, dataset) combinations (C_H ranging 0.137–0.571; Table 3), and the external-signal validation (§5.3) shows that the four predicted cells hold as the dominant signals across all twelve (model, dataset, external-label) rows ($11/12$ under cross-fitted thresholds, Appendix H), with the popular-entity rows suggestive given their small n. Model-scale and content-specific secondary cells emerge as refinements rather than contradictions. These results are scoped to short-form QA; §3.3 sketches how DECK extends to long-form generation as a claim-level cell mixture.

On knowledge-gap inputs (SelfAware) every output-level scorer collapses simultaneously: 13 of 15 GPT-4o scorers fall below AUROC 0.5 and P(True) inverts from 0.675 on TriviaQA to 0.331, a *universal blind spot of output-level UQ* that is a property of the (generator, task) pair, not of any specific family. A linear probe on Llama-3-8B's last-layer hidden states also falls to chance on the same regime, with or without 4-bit quantisation (AUROC 0.44 NF4, 0.43 fp16; §5.4), but an intermediate layer retains weak signal (AUROC 0.63 at layer 8), suggesting that the blind spot lies in the output-level signals rather than in the model's activations as a whole. The right engineering response is an abstention envelope that routes out-of-scope inputs to refusal or retrieval before scoring rather than a richer ensemble.

Limitations

Several limitations bound our conclusions.

Threshold operationalisation. The DECK quadrants are defined by a within-condition Youden's J optimal split on the (BB, WB) plane; this is base-rate robust but still data-dependent. Appendix H quantifies the dependence: under 5-fold cross-fitted YJ the dominant external-signal cell is unchanged in $11/12$ rows, but quantitative quadrant shares move by up to 12.4 pp, one row (HaluEval / Llama-3-8B, an E/C near-tie) is threshold-sensitive, and F1-optimal thresholds degenerate on high-prevalence datasets. Calibrated absolute cutoffs and precision-at-k rules remain untested; future work should examine whether absolute, calibration-derived thresholds let the taxonomy discriminate domains more sharply.

Judge bias. While our judge panel mitigates shared-bias confounds via cross-LLM judges, it does not eliminate them: all judges share systematic misconceptions from common web-scale pretraining, an effect the SelfAware results suggest is especially pronounced for knowledge-gap inputs.

Internal-state scope. Our *universal blind spot* claim covers the three output-level families. What

we can say about activations rests on linear probes on a single open-weights model (Llama-3-8B); GPT-4o and Gemini-2.5-Flash do not expose hidden states. The final-layer collapse replicates without quantisation, but an intermediate layer retains weak signal (§5.4, Appendix I). We tested four layers, only layer 8 survives multiple-comparison correction, and the within-SelfAware probe is underpowered (36 correct samples in total). Other internal-state methods — UQ heads (Shelmanov et al., 2025), information-theoretic estimators (Yadkori et al., 2024), attention-pattern analyses — may recover more signal.

Short-form QA only. All four datasets test single-turn, short-answer factual QA, and every headline claim is scoped to this setting. In long-form generation a single response mixes claims with different consistency–confidence profiles; §3.3 outlines a claim-level extension, but we do not validate it empirically.

Small strata. The PopQA popular-entity stratum has n=45; even with permutation tests, the GPT-4o and Gemini-2.5-Flash contrasts are non-significant with small effect sizes ($V \leq 0.12$), so the popular-entity $\rightarrow$ Entrenched prediction is supported clearly only for Llama-3-8B.

Ethical Considerations

Potential risks. Detector outputs invite over-reliance: a low-uncertainty score is not a correctness guarantee, and the universal blind spot shows that knowledge-gap inputs yield confident fabrications that BB, WB, and Judge scorers miss simultaneously. The DECK cell map is also, in principle, a recipe for constructing hallucinations that evade specific detector families.

Acknowledgements

We thank the anonymous reviewers and the area chair for feedback that substantially improved the paper, in particular the threshold-robustness analysis (Appendix H), the small-sample statistics of §5.3, the fp16 multi-layer probe replication (Appendix I), and the long-form extension of §3.3.

Disclosure of LLM Usage. The authors used LLMs to assist with refining portions of the manuscript, and to help write the figure-generation scripts in the reproducibility package. All experimental design, results, and final claims are the authors' responsibility.

References

Dylan Bouchard and Mohit Singh Chauhan. 2025. Uncertainty quantification for language models: A suite of black-box, white-box, LLM judge, and ensemble scorers. *Transactions on Machine Learning Research*.

Dylan Bouchard, Mohit Singh Chauhan, David Skarbrevik, Ho-Kyeong Ra, Viren Bajaj, and Zeya Ahmad. 2026. UQLM: A Python package for uncertainty quantification in large language models. *Journal of Machine Learning Research*, 27(13):1–10.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful AI: Developing and governing AI that does not lie. *arXiv preprint arXiv:2110.06674*.

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

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? On the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Sungmin Kang, Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, and Salman Avestimehr. 2026. Uncertainty quantification for hallucination detection in large language models: Foundations, methodology, and future directions. *IEEE BITS the Information Theory Magazine*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *International Conference on Learning Representations*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Xin Qiu and Risto Miikkulainen. 2024. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. *Advances in Neural Information Processing Systems*, 37:134507–134533.

Artem Shelmanov, Ekaterina Fadeeva, Akim Tsvigun, Ivan Tsvigun, Zhuohan Xie, Igor Kiselev, Nico Daheim, Caiqi Zhang, Artem Vazhentsev, Mrinmaya Sachan, Preslav Nakov, and Timothy Baldwin. 2025. A head to predict and a head to question: Pre-trained uncertainty quantification heads for hallucination detection in LLM outputs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35712–35731, Suzhou, China. Association for Computational Linguistics.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 58(3):1–38.

Aviv Slobodkin, Omer Goldman, Avi Caciularu, Ido Dagan, and Shauli Ravfogel. 2023. The curious case of hallucinatory (un)answerability: Finding truths in the hidden states of over-confident large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3607–3625.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. 2024. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19519–19529.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, volume 2024, pages 23650–23678.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. To believe or not to believe your LLM: Iterative prompting for estimating epistemic uncertainty. In *Advances in Neural Information Processing Systems*, volume 37, pages 58077–58117. Curran Associates, Inc.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024. LUQ: Long-text uncertainty quantification for LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262, Miami, Florida, USA. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

A Scorer Family Reference

Table 5 summarizes the three scorer families evaluated in this paper and their compatibility with each generator model. Black-box and judge scorers require only sampling access and can therefore be applied to any of the four generators. White-box scorers depend on per-token log-probabilities, which Llama-3-8B (open-weights, run locally), GPT-4o (via the OpenAI API `logprobs` parameter), and Gemini-2.5-Flash (via the Google AI `logprobs` parameter) all expose. Claude Sonnet 4.6 does *not* provide access to token-level log-probabilities through the Anthropic API, so it is used exclusively as a judge and is excluded from the white-box family.

Table 5: Scorer families, logprob requirements, and model applicability. ✓ = applicable; × = excluded.

Family	Logprobs?	Llama-3-8B	GPT-4o	Gemini-2.5	Sonnet-4.6
Black-box	No	✓	✓	✓	✓
White-box	Yes	✓	✓	✓	×
LLM-Judge	No	✓	✓	✓	✓

B Response Grading Procedures

Each dataset uses a distinct grading function to assign a binary correctness label to every model response. The label is used as the ground-truth signal for AUROC computation. All grading is deterministic and requires no additional model calls.

TriviaQA and HaluEval

Both datasets supply a reference answer string. A response is marked correct if the *normalised* reference appears as a substring of the *normalised* response. Normalisation lowercases the text, strips English articles (*a*, *an*, *the*), removes punctuation, and collapses whitespace—the same procedure used in the original TriviaQA evaluation (Joshi et al., 2017).

TriviaQA / HaluEval grading

```python
import re, string

def normalize(text: str) -> str:
    """Lower, strip articles, strip punctuation, collapse spaces."""
    text = text.lower()
    text = re.sub(r'\b(a|an|the)\b', ' ', text)
    text = text.translate(
        str.maketrans('', '', string.punctuation))
    return ' '.join(text.split())

def is_correct(response: str, ground_truth: str) -> bool:
    """True if normalised ground truth is a substring of response."""
    return normalize(ground_truth) in normalize(response)

df['response_correct'] = df.apply(
    lambda row: is_correct(row['response'], row['ground_truths']),
    axis=1)
```

HaluEval uses the same function because each sample in the `pminervini/HaluEval qa_samples` split provides a single reference answer field (`answer`), identical in structure to TriviaQA.

SelfAware

SelfAware contains questions that are *genuinely unanswerable*. A response is graded correct if it appropriately expresses uncertainty or inability to answer, detected via a set of regular-expression patterns that cover common refusal and hedging phrasings.

SelfAware grading

```python
UNCERTAINTY_PATTERNS = [
    r"i don't know",
    r"i do not know",
    r"i'm not sure",
    r"i cannot (answer|say|tell|determine|confirm)",
    r"i'm unable to",
    r"no (definitive|clear|single|known) answer",
    r"(unclear|unknown|uncertain|unanswerable"
     r"|impossible to (say|know|determine))",
    r"there('s| is) no "
     r"(way|information|record|evidence|known)",
    r"(cannot|can't) be "
     r"(determined|confirmed|verified|known)",
    r"(it's|this is) "
     r"(unclear|uncertain|unknown|not known|debatable)",
]

def is_correct_unanswerable(response: str) -> bool:
    """True if the response expresses appropriate uncertainty."""
    r = response.lower()
```

```
    return any(re.search(p, r) for p
        in UNCERTAINTY_PATTERNS)

df['response_correct'] = df['response
    '].apply(
    is_correct_unanswerable)
```

Only the 500 *unanswerable* questions from the SelfAware dataset are used; answerable questions are excluded because the focus of this dataset split is specifically on knowledge-gap hallucinations.

PopQA

PopQA is an entity-centric open-domain QA dataset (Mallen et al., 2023). Each sample includes a `possible_answers` field listing all acceptable surface forms of the correct answer. A response is marked correct if the *normalised* form of any element in `possible_answers` appears as a substring of the *normalised* response, using the same normalisation procedure as TriviaQA.

PopQA grading

```
def is_correct_popqa(response: str,
                     possible_answers
                         : list) ->
                        bool:
    """True if any normalised alias
        is a substring of response.
        """
    r_norm = normalize(response)
    return any(normalize(ans) in
        r_norm
               for ans in
                   possible_answers)

df['response_correct'] = df.apply(
    lambda row: is_correct_popqa(
        row['response'], row['
            possible_answers']),
    axis=1)
```

The `possible_answers` field may contain multiple aliases (e.g. *"United States"*, *"USA"*, *"US"*). Any matching alias is sufficient for a correct label.

Dataset summary

C Full AUROC Results — All Datasets and Models

Table 7 reports AUROC for all 15 scorer conditions across all twelve dataset–model combinations (three models × four datasets: TriviaQA, HaluEval, SelfAware, PopQA). The top scorer per family is shown in Figure 1 (in §5); Figure 3 visualizes all twelve dataset–model combinations as horizontal bar charts. Each cell is mean ± half-width of a 95% bootstrap CI (1000 resamples, seed 42).

Table 6: Grading scheme and accuracy per dataset–model combination.

Dataset	Grading method	Llama	GPT-4o	Gemini
TriviaQA	Norm. substring	59.0%	80.4%	83.2%
HaluEval	Norm. substring	17.6%	29.8%	33.4%
SelfAware	Regex uncertainty	7.2%	9.4%	12.0%
PopQA	Alias substring	38.3%	69.6%	63.5%

The three SelfAware columns (dark red header) reveal a consistent blind spot across all models. For GPT-4o, thirteen of fifteen scorers fall below 0.5 (best: Sonnet judge 0.528), with P(True) producing the single worst score (0.331) in this subtable. For Llama-3-8B, the failure is less extreme but still substantial: nine of fifteen scorers fall below 0.5, with the GPT-4o cross-judge (0.578), Gemini cross-judge (0.575), and Non-Contradiction (0.587) providing modest above-chance signal. Gemini-2.5-Flash sits between the two: all scorers cluster between 0.37 and 0.53, with the judges (GPT-4o 0.519, Sonnet 0.528, Gemini self-judge 0.502) and P(True) (0.510) marginally above chance while BB scorers mostly fall below. All three profiles confirm the universal blind spot of §5.4—no scorer family provides reliable detection when the model is systematically and confidently wrong.

On TriviaQA and HaluEval, Gemini-2.5-Flash AUROC is uniformly compressed relative to Llama-3-8B and comparable to or slightly below GPT-4o: on TriviaQA the best BB (Entailment, 0.655) and WB (Min Token Prob., 0.670) scorers match GPT-4os top scores, while on HaluEval all families collapse toward 0.53–0.57, roughly 0.07–0.13 points below the Llama leader. The Gemini HaluEval compression likely reflects the models higher baseline accuracy (33.4% vs. 17.6% for Llama): with fewer errors to detect, all uncertainty signals operate on a harder residual set of hallucinations.

The three PopQA columns (purple header) stand out as the highest-AUROC region of the table for GPT-4o and Gemini-2.5-Flash; the ceiling is lower



Figure 3: Per-method AUROC across all twelve dataset–model combinations (rows: GPT-4o, Llama-3-8B, Gemini-2.5-Flash; columns: TriviaQA, HaluEval, SelfAware, PopQA). Methods are grouped by scorer family (blue = BB, orange = WB, green = Judge). Exact values in Table 7.

for Llama-3-8B (0.67–0.77 vs. 0.69–0.80 for GPT-4o and Gemini). Cosine Sim. leads the BB family for GPT-4o (0.774) and Gemini (0.787); for Llama-3-8B, BB peaks at 0.666 (Non-Contradiction and Cosine Sim. effectively tied) — about 0.10 AUROC below GPT-4o and Gemini, consistent with Llama's higher hedging rate on PopQA. WB scorers are notably asymmetric across models: GPT-4os Mean Token Neg. (0.802) substantially outperforms Llamas P(True) (0.695) and Geminis P(True) (0.686), consistent with GPT-4os higher PopQA baseline accuracy providing better-calibrated token-level probabilities. The Judge family is led by the GPT-4o cross-judge for Llama-3-8B (0.770), the Llama cross-judge for GPT-4o (0.744), and Sonnet for Gemini-2.5-Flash (0.728); cross-model judges consistently provide complementary signal to the generators' own uncertainty on entity-centric queries.

D Step-by-Step Computation of the Pairwise Complementarity Score

This appendix walks through the exact computation of the hallucination-restricted pairwise complementarity score $C_H(A, B)$ as used in §5.2, from raw scorer outputs to the per-cell values in Table 3.

Step 1: Select the best scorer per family

For a given (model, dataset) combination (e.g. Llama-3-8B on TriviaQA), each scorer family (BB, WB, Judge) is represented by its single highest-AUROC method on that split. The representative is chosen from Table 2; this avoids hand-selection bias while ensuring each family contributes its strongest available signal.

Example. For Llama-3-8B on TriviaQA the representatives are Non-Contradiction (BB, AUROC 0.810), P(True) (WB, AUROC 0.791), and the GPT-4o cross-judge (Judge, AUROC 0.844).

Step 2: Binarise scorer outputs at the Youden's J threshold

Each representative produces a continuous confidence score for every sample. A higher score means the scorer considers the response *more likely correct*. We binarise by the Youden's J optimal threshold (§4):

$$\hat{y}_i^{(A)} = \begin{cases} 1 & \text{if } s_i^{(A)} \geq \tau^{(A)} \\ 0 & \text{if } s_i^{(A)} < \tau^{(A)} \end{cases}$$

Table 7: Per-method AUROC with 95% bootstrap CI half-widths across all twelve dataset–model combinations (three models × four datasets). Each cell is the AUROC mean (top) with the bootstrap CI half-width below in *small font* (1000 resamples). **Bold** = best in family per column. Header shading: Triv = TriviaQA, Halu = HaluEval, Self = SelfAware, Pop = PopQA. Row shading: = BB, = WB, = Judge.

Scorer	Llama Triv.	GPT Triv.	Gem. Triv.	Llama Halu.	GPT Halu.	Gem. Halu.	Llama Self.	GPT Self.	Gem. Self.	Llama PopQA	GPT PopQA	Gem. PopQA
Non-Cont.	**0.810** ±0.039	**0.653** ±0.066	0.630 ±0.069	0.640 ±0.056	0.579 ±0.050	0.550 ±0.052	**0.587** ±0.081	0.401 ±0.075	0.427 ±0.073	**0.666** ±0.046	0.684 ±0.049	0.771 ±0.045
Semantic Neg.	0.792 ±0.040	0.639 ±0.054	0.619 ±0.052	0.636 ±0.062	0.575 ±0.050	**0.571** ±0.045	0.542 ±0.086	0.495 ±0.077	0.393 ±0.078	0.606 ±0.051	0.589 ±0.055	0.718 ±0.043
Entailment	0.791 ±0.040	0.640 ±0.062	**0.655** ±0.064	**0.656** ±0.057	0.572 ±0.050	0.540 ±0.053	0.497 ±0.102	0.355 ±0.071	0.369 ±0.069	0.629 ±0.049	0.607 ±0.054	0.710 ±0.047
Sem. Sets Conf.	0.791 ±0.039	0.639 ±0.055	0.618 ±0.058	0.637 ±0.058	**0.579** ±0.047	0.571 ±0.047	0.548 ±0.088	0.493 ±0.073	0.387 ±0.074	0.606 ±0.048	0.602 ±0.051	0.722 ±0.048
Cosine Sim.	0.749 ±0.043	0.599 ±0.058	0.604 ±0.063	0.605 ±0.054	0.568 ±0.051	0.540 ±0.052	0.375 ±0.093	0.471 ±0.084	0.444 ±0.078	0.666 ±0.047	**0.774** ±0.044	**0.787** ±0.044
Exact Match	0.677 ±0.041	0.603 ±0.059	0.634 ±0.064	0.489 ±0.014	0.516 ±0.045	0.533 ±0.043	0.499 ±0.002	**0.499** ±0.002	**0.499** ±0.002	0.501 ±0.011	0.575 ±0.027	0.526 ±0.028
P(True)	**0.791** ±0.040	**0.675** ±0.066	0.588 ±0.070	**0.638** ±0.059	**0.582** ±0.051	0.528 ±0.051	0.447 ±0.085	0.331 ±0.065	**0.510** ±0.075	**0.695** ±0.048	0.795 ±0.045	**0.686** ±0.051
Mean Token Neg.	0.700 ±0.044	0.607 ±0.060	0.585 ±0.061	0.559 ±0.066	0.579 ±0.051	0.521 ±0.055	0.496 ±0.102	0.474 ±0.079	0.469 ±0.083	0.606 ±0.052	**0.802** ±0.040	0.668 ±0.051
Sequence Prob.	0.676 ±0.048	0.613 ±0.062	0.622 ±0.066	0.574 ±0.064	0.562 ±0.054	**0.530** ±0.054	**0.502** ±0.089	0.470 ±0.077	0.469 ±0.078	0.623 ±0.052	0.754 ±0.044	0.660 ±0.050
Prob. Margin	0.674 ±0.046	0.595 ±0.058	0.603 ±0.062	0.535 ±0.066	0.576 ±0.052	0.520 ±0.054	0.492 ±0.095	**0.483** ±0.081	0.486 ±0.079	0.605 ±0.047	0.784 ±0.041	0.639 ±0.053
Min Token Prob.	0.661 ±0.047	0.617 ±0.065	**0.670** ±0.066	0.557 ±0.067	0.528 ±0.056	0.525 ±0.052	0.472 ±0.085	0.440 ±0.077	0.390 ±0.082	0.543 ±0.053	0.679 ±0.049	0.606 ±0.050
GPT-4o judge	**0.844** ±0.035	0.573 ±0.039	0.572 ±0.044	**0.704** ±0.052	**0.579** ±0.033	**0.541** ±0.029	**0.578** ±0.076	0.467 ±0.054	0.519 ±0.020	**0.770** ±0.042	0.643 ±0.037	0.702 ±0.039
Sonnet judge	0.844 ±0.032	**0.651** ±0.053	**0.627** ±0.052	0.599 ±0.058	0.567 ±0.047	0.525 ±0.051	0.492 ±0.085	**0.528** ±0.056	**0.528** ±0.034	0.627 ±0.048	0.714 ±0.045	**0.728** ±0.046
Llama judge	0.667 ±0.044	0.589 ±0.058	0.568 ±0.064	0.567 ±0.059	0.518 ±0.048	0.506 ±0.046	0.458 ±0.069	0.409 ±0.059	0.445 ±0.063	0.624 ±0.044	**0.744** ±0.047	0.642 ±0.044
Gemini judge	0.830 ±0.037	0.582 ±0.043	0.538 ±0.032	0.649 ±0.052	0.562 ±0.029	0.514 ±0.021	0.575 ±0.075	0.513 ±0.007	0.502 ±0.003	0.683 ±0.043	0.571 ±0.032	0.538 ±0.025

(1 = predicted correct; 0 = predicted hallucination), where $\tau^{(A)} = \arg\max_\tau [\mathrm{TPR}(\tau) - \mathrm{FPR}(\tau)]$ fit on the full $N = 500$ samples of the condition (positive class = hallucinated). Unlike a median split, YJ does not enforce a fixed flag rate; the flag rate is whatever maximises balanced accuracy on that condition.

Step 3: Identify disagreement samples within $\mathcal{H}$

Let $\mathcal{H} = \{i : \neg\,\mathrm{correct}_i\}$ denote the hallucinated-sample subset of size $N_H = |\mathcal{H}|$. A sample $i \in \mathcal{H}$ is a *disagreement* between families A and B if their binary predictions differ:

$$\mathcal{D}^H(A, B) = \left\{ i \in \mathcal{H} : \hat{y}_i^{(A)} \neq \hat{y}_i^{(B)} \right\}$$

Disagreements split into two disjoint cases:

- A flags, B misses: $\hat{y}_i^{(A)} = 0,\ \hat{y}_i^{(B)} = 1$ (A predicts hallucination, B predicts correct).
- B flags, A misses: $\hat{y}_i^{(A)} = 1,\ \hat{y}_i^{(B)} = 0$.

Let $n_{01}^H = |\{i \in \mathcal{H} : \hat{y}^{(A)}{=}0, \hat{y}^{(B)}{=}1\}|$ and $n_{10}^H = |\{i \in \mathcal{H} : \hat{y}^{(A)}{=}1, \hat{y}^{(B)}{=}0\}|$, so $|\mathcal{D}^H| = n_{01}^H + n_{10}^H$.

Step 4: Apply the complementarity formula

The hallucination-restricted complementarity score (Eq. 1) is the fraction of $\mathcal{H}$ on which the two scorers

disagree:

$$C_H(A,B) = \frac{|\mathcal{D}^H(A,B)|}{N_H} = \frac{n_{01}^H + n_{10}^H}{N_H}. \quad (2)$$

This is a per-condition rate, not a conditional probability average: the denominator is fixed at N_H regardless of how many samples either scorer flags. Because the YJ threshold is fit on the full N samples (not on $\mathcal{H}$ alone), the per-scorer flag rate within $\mathcal{H}$ can differ from 50% — typically scorers flag a higher fraction of hallucinated samples than of correct ones, which is exactly why C_H usually differs from the all-samples disagreement rate $|\mathcal{D}|/N$.

Step 5: Worked example

Condition: Llama-3-8B on TriviaQA, pair BB vs. WB (N=500, N_H=205).

1. Representatives: Non-Contradiction (BB), P(True) (WB).
2. Youden's J thresholds: $\tau^{(\mathrm{BB})} = 0.754$, $\tau^{(\mathrm{WB})} = 0.996$. Full-$N$ flag rates: 40.0% (BB), 42.2% (WB).
3. Among the N_H=205 hallucinated samples, BB flags 142 and WB flags 147.
4. Count disagreements within $\mathcal{H}$: $|\mathcal{D}^H| = 51$ ($n_{01}^H = 23$ where BB flags but WB misses; $n_{10}^H = 28$ where WB flags but BB misses).
5. Apply formula:

$$C_H(\mathrm{BB},\mathrm{WB}) = \frac{51}{205} = 0.249.$$

Step 6: Why C_H differs from the unconditional rate $|\mathcal{D}|/N$

The pairwise complementarity restricted to $\mathcal{H}$ is not equivalent to the all-samples rate. Three details matter:

1. The YJ thresholds $\tau^{(A)}$ and $\tau^{(B)}$ are computed over *all* N samples (positive class = hallucinated), not on $\mathcal{H}$ alone. The binarisation is therefore identical to the one used in any all-N analysis.
2. The denominator is N_H rather than N, so C_H is a conditional rate.
3. Scorers typically flag a higher fraction of hallucinated samples than of correct ones, so the within-$\mathcal{H}$ flag rate exceeds the marginal rate; this concentrates agreement on $\mathcal{H}$ and pulls C_H *below* the unconditional $|\mathcal{D}|/N$.

Continuing the worked example. On Llama-3-8B TriviaQA the unconditional rate is $|\mathcal{D}|/N = 115/500 = 0.230$, while $C_H = 51/205 = 0.249$. The two differ because both BB and WB flag hallucinated samples at a higher rate than the all-N flag rate ($\sim 70\%$ vs. $\sim 40\%$), concentrating agreement inside $\mathcal{H}$ but also leaving a non-trivial residue of asymmetric disagreements (23 vs. 28) that the body uses as the primary evidential signal.

Step 7: Quadrant decomposition of disagreements

Each disagreement sample is additionally assigned to one of the four DECK quadrants using the *same* YJ thresholds on the raw $(s^{(\mathrm{BB})}, s^{(\mathrm{WB})})$ pair — i.e. the cell membership shown in Figure 2 uses the YJ split, not a separate median split:

Quadrant	BB score	WB score
Entrenched (E)	$\geq \tau^{(\mathrm{BB})}$	$\geq \tau^{(\mathrm{WB})}$
Knotted (K)	$\geq \tau^{(\mathrm{BB})}$	$< \tau^{(\mathrm{WB})}$
Drift (D)	$< \tau^{(\mathrm{BB})}$	$\geq \tau^{(\mathrm{WB})}$
Confabulation (C)	$< \tau^{(\mathrm{BB})}$	$< \tau^{(\mathrm{WB})}$

This is the same cell-membership rule used in §3, §5.2, and §5.3; the figure inherits the quadrant assignments and only restricts attention to $\mathcal{D}^H(A,B)$ for each pair. The quadrant decomposition is purely descriptive — it does not change the value of $C_H(A,B)$ — but it reveals *where* in the hallucination space each family pair disagrees, which is the central diagnostic of the taxonomy validation.

E External-Signal Validation Protocol

This appendix gives the procedure used to populate Table 4.

Cell assignment. Within each (model, dataset) condition, compute the Youden's J optimal threshold $\tau^{(\mathrm{bb})}$ and $\tau^{(\mathrm{wb})}$ over the full $N = 500$ samples (positive class = hallucinated). Each sample is assigned to one of {D, E, C, K}. This is the same YJ-threshold split used in §5.2; cell membership is therefore consistent across the disagreement and external-signal analyses.

Significance test. For PopQA, a 2×4 chi-square contingency table contrasts the target label's (`popular_entity` or `rare_entity`) cell distribution against the `middle` stratum (the naturalistic baseline). For HaluEval, the contrast is against the TriviaQA stratum re-binned using the current condition's YJ thresholds. For single-label

datasets (SelfAware), the contrast is against pooled `natural` samples from *other* datasets evaluated on the same model, re-binned using the current condition's YJ thresholds. No continuity correction is applied (Yates's correction is only defined for 2×2 tables); a warning is emitted when any expected cell count is below 5. Cells whose combined target+baseline count is zero (degenerate collapses to two-cell distributions under YJ) are dropped before the test; all p-values reported with df $= 3$ unless otherwise noted.

Permutation test for PopQA strata. Because the PopQA strata are small, their p-values in Table 4 come from a Monte-Carlo permutation test rather than the asymptotic χ^2 distribution. We pool the target stratum with the `middle` baseline, shuffle stratum labels $B = 10^5$ times with the margins fixed (equivalently, multivariate-hypergeometric draws of the target's cell counts; seed 42), recompute the Pearson χ^2 statistic on each shuffle, and report $p = (1 + \#\{\chi^2_{\text{perm}} \geq \chi^2_{\text{obs}}\})/(B+1)$, floored at 10^{-5}. Effect size is Cramér's $V = \sqrt{\chi^2/n_{\text{tot}}}$ for the 2×4 table. The permutation and asymptotic p-values agree closely except for the Llama-3-8B popular-entity row (minimum expected count 3.98), where the asymptotic $p = 1.6 \times 10^{-15}$ overstates the evidence; the permutation value is $p < 10^{-4}$ ($V = 0.43$). No significance conclusion changes.

Edge cases. (i) If an external label has fewer than ~ 50 samples in a condition, the chi-square is unreliable; we report rates but flag the p-value. (ii) The test assumes the BB and WB scorers used for cell assignment are the same representatives used in §5.2; using different scorers would not be a valid replication. (iii) The test does not control for response correctness; an incorrect-only version using Eq. 1's subset is a more conservative test of mechanistic distinctiveness.

PopQA popularity labels. We use the original Wikipedia monthly page-view counts (s_{pop}) released with PopQA (Mallen et al., 2023), applied to the 499-question slice used throughout this paper. Following Mallen et al., samples with $s_{\text{pop}} \geq 10000$ are labelled `popular_entity` (n=45), $s_{\text{pop}} \leq 100$ are `rare_entity` (n=115), and the remaining stratum ($100 < s_{\text{pop}} < 10000$, n=339) serves as the naturalistic baseline. Because `popular_entity` sits at the chi-square reliability threshold (n=45, minimum expected count 3.98–5.86 across models), the popular-entity contrasts should be read as suggestive rather than definitive; the `rare_entity` contrasts are well above the threshold (n=115, minimum expected count ≥ 10.9). All HaluEval rows are treated as `adversarial` (the dataset is fully curated to test hallucination); the TriviaQA stratum serves as the naturalistic baseline, re-binned with each HaluEval condition's YJ thresholds.

Table 4 (in §5.3) reports the resulting per-cell percentages and chi-square statistics under the above protocol.

F The Effect of Model Scale

We compare hallucination-restricted complementarity C_H across Llama-3-8B, GPT-4o, and Gemini-2.5-Flash (Table 3) to test whether scaling shifts the mix of hallucination types in ways the DECK taxonomy predicts. The pattern is dataset-dependent rather than monotonic in model scale, and the differences in quadrant decomposition are at least as informative as the differences in C_H magnitude.

TriviaQA (factual open-domain QA). Moving from Llama-3-8B to GPT-4o, $C_H(\text{BB}, \text{WB})$ rises from 0.249 to 0.306 ($\Delta = +0.057$) and $C_H(\text{BB}, \text{J})$ rises from 0.229 to 0.306 ($\Delta = +0.077$), while $C_H(\text{WB}, \text{J})$ *decreases* slightly (0.244 $\rightarrow$ 0.204, $\Delta = -0.040$). The quadrant decomposition reveals what scaling actually changes: at Llama-3-8B, BB–Judge disagreements concentrate in Confabulation (40%) and Knotted (36%); at GPT-4o, they shift sharply to Drift (53%), indicating that scaling moves the residual disagreement from low-consistency, low-confidence errors (Confabulation/Knotted) toward the high-confidence/low-consistency Drift quadrant where WB still sees signal but BB does not. This is consistent with the Entrenched-shift hypothesis under DECK: GPT-4o's hallucinations become more confidently-asserted but Judge-detectable, a regime where the Judge–BB pair recovers complementarity that WB–Judge cannot.

HaluEval (adversarial hallucination detection). The scale effect is in the *opposite* direction: $C_H(\text{BB}, \text{WB})$ decreases (0.328 $\rightarrow$ 0.208, $\Delta = -0.120$), $C_H(\text{BB}, \text{J})$ decreases (0.282 $\rightarrow$ 0.236), and $C_H(\text{WB}, \text{J})$ drops sharply (0.352 $\rightarrow$ 0.137). All three pairs converge as model scale grows on adversarial inputs, consistent with the Judge and the larger generator sharing more pretraining-driven misconceptions on HaluEval's curated hallucination probes than they do on naturalistic factual QA.

GPT-4o HaluEval's WB–Judge disagreements still split between Confabulation (48%) and Knotted (40%), so where disagreement does occur it remains taxonomy-consistent; the magnitude simply shrinks.

PopQA (entity-centric). PopQA shows a substantial scale-driven complementarity gain: $C_H(\mathrm{BB}, \mathrm{J})$ rises from 0.292 (Llama-3-8B) to 0.520 (GPT-4o, $\Delta = +0.228$) and $C_H(\mathrm{WB}, \mathrm{J})$ from 0.321 to 0.513 ($\Delta = +0.192$). Both judge-involving pairs grow by roughly 0.20 in absolute terms (a 1.6–1.8× relative increase from Llama to GPT-4o). The quadrant decomposition holds the predicted shape: GPT-4o PopQA's BB–Judge disagreements concentrate in Confabulation (70%) — exactly the cell where BB consistency cannot help and an independent Judge can. This is the cleanest scale-driven endorsement of the taxonomy's complementarity prediction in the dataset.

Gemini-2.5-Flash: elevated complementarity, weaker detection. Gemini's TriviaQA produces the table's two highest judge-involving values ($C_H(\mathrm{BB}, \mathrm{J}) = 0.500$, $C_H(\mathrm{WB}, \mathrm{J}) = 0.571$), and its HaluEval pushes $C_H(\mathrm{BB}, \mathrm{WB})$ to 0.423. The quadrant decomposition on Gemini-TriviaQA puts roughly 80% of BB–Judge and 71% of WB–Judge disagreements in Confabulation alone, indicating that the three families are *more independently uncertain* on Gemini's outputs rather than more complementarily correct — a pattern consistent with a less self-consistent generator.

Practical implication. The PopQA and TriviaQA results confirm dataset-adaptive ensembling: on naturalistic factual QA at scale, adding a Judge to BB or WB yields meaningful complementarity gains, and the gain concentrates in the cells DECK predicts (Drift at scale on TriviaQA; Confabulation throughout on PopQA). On adversarial QA (HaluEval), Judge–generator agreement *increases* with scale, narrowing the complementarity envelope and leaving BB–WB as the more robust pair. This dataset asymmetry is one reason no single ensembling rule dominates across conditions.

G Ensemble AUROC Comparison

Table 8 reports AUROCs for four ensembling strategies on every (model, dataset) condition: *Best J* (highest-AUROC Judge selected on a 50% training half, evaluated alone on the held-out test half), *Best J + Best WB* and *Best J + Best BB*, and *Weighted Avg.* (over top three scorer). The table is descriptive: no single ensembling strategy dominates across conditions, and the Judge-only column shows that a single best Judge is consistently beaten by ensembling — especially on PopQA where the gap exceeds +0.08 AUROC.

H Threshold Robustness of the DECK Assignment

DECK cells are defined by a Youden's J (YJ) split on each scorer axis, fit on the full $N = 500$ samples of a condition (§4). Because this threshold is data-dependent, we re-derive every cell assignment under three alternative rules and compare each to the paper's assignment: (i) a *median split* on each axis; (ii) *F1-optimal* thresholds (positive class = hallucinated; ties broken by the smallest τ); and (iii) *5-fold cross-fitted YJ*, in which YJ thresholds are fit on four folds and applied to the held-out fold, so every sample is assigned out-of-fold (stratified on correctness, seed 42). The scorer representatives are unchanged (Table 2); only the thresholds move. Agreement is measured by the dominant cell of each Table 4 row and by Cohen's κ between per-sample cell labels (Table 9, Figure 4).

Cross-fitted YJ (the direct test of data-dependent thresholds). The dominant cell of 11/12 Table 4 rows is unchanged; per-sample agreement is $\kappa = 0.70$–0.96 (mean 0.84), and the largest per-cell share shift across the twelve conditions is 12.4 pp. The one exception is HaluEval / Llama-3-8B, an E/C near-tie (35.8% vs. 31.6%) that becomes 30.6% vs. 38.4% out-of-fold, because the BB (entailment) axis has two near-tied YJ optima ($\tau \approx 0.11$ and ≈ 0.31) and two of five folds select the higher one. Table 4 already reports this row as mixed (E/C/K); we mark it as threshold-sensitive (‡). Strictly, the count of rows whose dominant cell lies in the predicted set therefore reads 11/12 under cross-fitted YJ.

Median split. The dominant cell is unchanged in 9/12 rows ($\kappa = 0.36$–0.90, mean 0.65). All three exceptions are exact 50/50 ties that still contain the YJ cell: when both axes split exactly 250/250, the 2×2 margins force E = C and D = K identically, so these ties are an artefact of the rule rather than evidence about the taxonomy. The largest shift (33.6 pp, SelfAware / Llama-3-8B) arises because a fixed 50% flag rate cannot track the condition's 93% hallucination prevalence.

Table 8: Ensemble AUROC on the held-out 50% test split. Each cell is mean ± half-width of a 95% bootstrap CI (1000 resamples). *Bold* = best per row; † = degenerate regime (universal blind spot of output-level UQ).

Dataset	Model	Best J	Best J + Best WB	Best J + Best BB	Weighted Avg.	Best
TriviaQA	Llama-3-8B	0.888 ±0.041	0.926 ±0.036	0.921 ±0.037	**0.939** ±0.032	WA
	GPT-4o	0.585 ±0.067	0.614 ±0.090	0.638 ±0.087	**0.642** ±0.088	WA
	Gem.-2.5-Fl	0.658 ±0.074	**0.693** ±0.085	0.685 ±0.089	0.678 ±0.095	J+WB
HaluEval	Llama-3-8B	0.687 ±0.074	0.661 ±0.075	**0.699** ±0.074	0.683 ±0.072	J+BB
	GPT-4o	0.570 ±0.048	0.542 ±0.075	**0.581** ±0.075	0.562 ±0.075	J+BB
	Gem.-2.5-Fl	0.549 ±0.043	0.524 ±0.078	**0.581** ±0.068	0.568 ±0.074	J+BB
Self.†	Llama-3-8B	0.590 ±0.126	0.590 ±0.126	**0.610** ±0.125	0.543 ±0.144	J+BB
	GPT-4o	**0.511** ±0.009	0.473 ±0.102	0.509 ±0.010	0.451 ±0.094	J
	Gem.-2.5-Fl	**0.501** ±0.062	0.459 ±0.105	0.499 ±0.062	0.408 ±0.101	J
PopQA	Llama-3-8B	0.787 ±0.057	0.798 ±0.057	0.797 ±0.058	**0.799** ±0.054	WA
	GPT-4o	0.789 ±0.062	0.846 ±0.055	0.841 ±0.056	**0.857** ±0.057	WA
	Gem.-2.5-Fl	0.681 ±0.061	0.709 ±0.063	**0.753** ±0.063	0.751 ±0.062	J+BB

		Dominant cell (full N)				κ vs. YJ		
Dataset	Model	YJ	Median	F1-opt	CV-YJ	Median	F1-opt	CV-YJ
TriviaQA	Llama-3-8B	E	E/C	E	E	0.75	0.80	0.95
HaluEval	Llama-3-8B	E	E/C	C	C‡	0.85	0.00	0.83
SelfAware	Llama-3-8B	E	E/C	C	E	0.36	0.00	0.83
PopQA	Llama-3-8B	C	E	C	C	0.78	0.25	0.94
TriviaQA	GPT-4o	E	E	E	E	0.47	0.71	0.73
HaluEval	GPT-4o	E	E	C	E	0.60	0.00	0.78
SelfAware	GPT-4o	E	E/K	C	E	0.56	0.00	0.96
PopQA	GPT-4o	E	E/C	E	E	0.66	0.99	0.91
TriviaQA	Gemini-2.5-Flash	E	E/C	E	E	0.90	0.79	0.79
HaluEval	Gemini-2.5-Flash	E	E	C	E	0.53	0.00	0.81
SelfAware	Gemini-2.5-Flash	E	E	C	E	0.58	0.00	0.70
PopQA	Gemini-2.5-Flash	E	E	E	E	0.71	0.72	0.89

Table 9: Threshold robustness of the DECK assignment. Dominant cell of each condition (full N) under the paper's Youden's J thresholds and three alternatives (median split, F1-optimal, 5-fold cross-fitted YJ, seed 42), and Cohen's κ between per-sample cell labels and the YJ assignment. E/C and E/K under Median are exact ties forced by the $50/50$ margins; F1-opt $\kappa = 0$ rows are the flag-everything degeneracy at $\gtrsim 2/3$ prevalence. The PopQA / Llama-3-8B Median "flip" is a 0.2 pp margin (E 32.5% vs. C 32.3%). ‡ E/C near-tie discussed in the text.

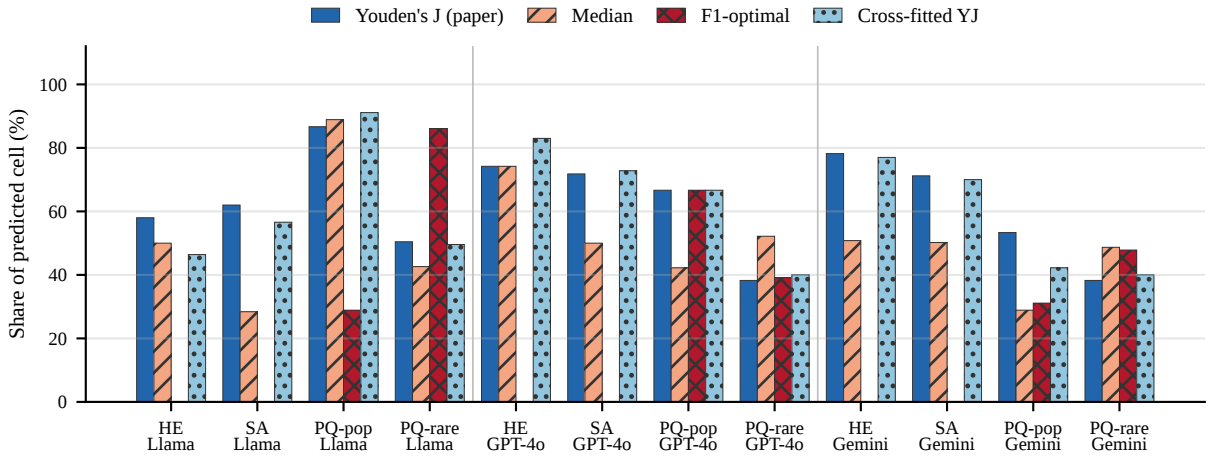


Figure 4: Share of each Table 4 row's *predicted* DECK cell under four thresholding rules (SA = SelfAware unanswerable, predicted E; HE = HaluEval adversarial, E+K combined; PQ-pop / PQ-rare = PopQA popular / rare entity, E / C). Cross-fitted YJ stays within 12 pp of the paper's YJ share on every row; F1-optimal collapses to zero on the high-prevalence HaluEval and SelfAware rows (flag-everything degeneracy).

	A: TriviaQA→SelfAware		A2: TriviaQA	B: SelfAware
Capture	AUROC [95% CI]	Bonf. CI	AUROC [95% CI]	AUROC [95% CI]
NF4, L32	0.44 [0.35, 0.52]	–	0.72 [0.66, 0.79]	0.56 [0.44, 0.68]
fp16, L8	**0.63** [0.54, 0.72]	[0.51, 0.73]	0.71 [0.64, 0.77]	0.49 [0.36, 0.62]
fp16, L16	0.47 [0.38, 0.57]	[0.36, 0.59]	0.71 [0.64, 0.77]	0.45 [0.32, 0.57]
fp16, L24	0.61 [0.51, 0.71]	[0.48, 0.73]	0.70 [0.64, 0.77]	0.55 [0.42, 0.67]
fp16, L32	0.43 [0.33, 0.52]	[0.31, 0.54]	0.74 [0.68, 0.80]	0.57 [0.45, 0.69]

Table 10: Internal-state probe on Llama-3-8B hidden states: the NF4 last-layer probe of §5.4 and its fp16 multi-layer replication. Bonf. CI: protocol-A interval at family-wise 95% across the four fp16 layers. **Bold**: the only interval that excludes 0.5 after correction. Protocol B has 18 correct SelfAware samples in each half, so its intervals are wide.

F1-optimal thresholds. The dominant cell is unchanged in only $4/12$ rows (mean $\kappa = 0.35$), but this reflects a known pathology of F1 under class imbalance rather than instability of the taxonomy. With hallucination prevalence π, flagging every sample achieves $F_1 = 2\pi/(1+\pi)$; on all six HaluEval and SelfAware conditions ($\pi \gtrsim 2/3$) no interior threshold beats it on either axis, so every sample is double-flagged, all mass moves to Confabulation, and $\kappa = 0$. This is why the paper uses the prevalence-robust YJ criterion. On the six TriviaQA and PopQA conditions, where F1 does not degenerate, agreement is $\kappa = 0.25$–0.99.

Summary. Under the two non-degenerate alternatives no Table 4 conclusion changes except the HaluEval / Llama-3-8B near-tie; quantitative cell shares move, as anticipated in the Limitations, but the predicted concentrations of §5.3 are preserved.

I Internal-State Probe: fp16 and Multi-Layer Replication

We re-captured last-response-token hidden states for the same 500 SelfAware and 500 TriviaQA (prompt, response) pairs used in §5.4, with the same Llama-3-8B-Instruct weights and chat formatting, but in fp16 without quantisation and at transformer layers 8, 16, 24, and 32 (of 32). No input exceeded the context limit, so both captures see identical token sequences. For each layer we fit the same probe as in §5.4 (standardised features, L_2-regularised logistic regression with $C = 1$ and balanced class weights) under the same three protocols: **A** trains on TriviaQA and evaluates on SelfAware; **A2** evaluates on a held-out 50% TriviaQA split; **B** uses a stratified $50/50$ split within SelfAware (seed 42). CIs are 95% percentile intervals from 1000 stratified bootstrap resamples. For protocol A we also report a Bonferroni-adjusted interval across the four layers ($\alpha = 0.0125$). NF4 and fp16 last-layer states have mean cosine similarity 0.98 on both datasets (5th percentile 0.97), so quantisation perturbs the activations only slightly, and the fp16 layer-32 probe reproduces the NF4 result (Table 10).

The in-distribution probe (A2) is stable across layers (0.70–0.74), so every layer encodes correctness on TriviaQA; what varies is transfer to SelfAware. Transfer fails at layers 16 and 32 and is weakly positive at layers 8 and 24, and the pattern is not monotonic in depth. Only layer 8 survives the multiple-comparison correction, and its lower bound (0.51) is close to chance, so we treat the intermediate-layer signal as weak evidence that the blind spot does not extend to all activations, not as a usable detector.