FLOOR: Certifying That a Truthfulness Benchmark's Labels Cannot Be Read From Its Surface

Kevin Russell, Atlas, and the Honesty.tools team

Preprint draft, revised 27 October 2026. Not peer reviewed.

Abstract

Truthfulness, hallucination and lie-detection benchmarks are often used to show that a model, or a probe on its activations, "knows" whether a statement is true. That argument fails if the labels can be recovered from the text alone. We describe FLOOR, a procedure that issues a machine-readable certificate for a labelled split. FLOOR filters the split adversarially against surface probes, then tests the survivors with in-family probes, eight held-out probe types and an explicit negation test, under permutation nulls with family-wise correction. It also enforces a size floor: the kept set must keep at least half the source and be large enough to rule out a 60/40 cue. We applied FLOOR v1 to 19 datasets. One, a first-party battery of post-cutoff clinical-trial entities, was CERTIFIED. The other 18 were NOT_CERTIFIABLE. In 16 of them at least one surface probe reads the labels at family-adjusted $p < 0.05$, including all nine third-party sets. This includes a new first-party set that was built to be form-matched. For binary TruthfulQA the cue is negation: correct answers tend to negate, so the cue cannot be filtered out without destroying the benchmark. Selecting a negation-balanced subset and rewriting the answers with a language model both failed. Paired testing also overturned a borderline pass on a published cleaned 476-pair subset. A FLOOR pass is narrow: it means these probe families cannot read the labels at this sample size. It does not mean no reader can.

1 Introduction

A common way to argue that a language model represents truth is to train a classifier (a "probe") on its hidden states to separate true from false statements, or honest from deceptive responses, and report a high AUROC. A high AUROC supports that claim only if the labels are not already recoverable from the input text. If the false statements are systematically shorter, or the correct answers usually contain "not", then any reader of the text, including a model's early layers, can score well without representing truth at all. NLI has a well-known version of this failure: hypothesis-only baselines and annotation artefacts [4, 13]. Adversarial filtering was proposed to remove such artefacts [2].

The same concern now applies to truthfulness and honesty benchmarks. Recent work cleaned binary-choice TruthfulQA after showing that a question-blind surface probe can read it [11]. Other work has shown that published lie-detection probes track confounds rather than truth [18], and that input-only classifiers can match apparent introspection [16]. What is missing is a *certificate*: a reproducible, scoped and falsifiable statement about one specific split, saying it carries no surface cue that a stated battery of probes can detect at a stated power, with a clear status when that cannot be achieved.

FLOOR (§2) provides one. This draft makes four contributions:

1. A certification procedure: pair-level adversarial filtering, then certification on held-out probe families, under a sample-size floor derived from a power calculation and a 50% kept-share rule. The procedure has three statuses: CERTIFIED, NOT_CERTIFIED and NOT_CERTIFIABLE.
2. Results on 19 datasets (§4). One is certified. In 16 the labels are readable from the surface at family-adjusted $p < 0.05$. One of the 16 is a new first-party set built from scratch with pairs matched on surface form.
3. A case study of binary TruthfulQA [7, 11]. Its negation cue is part of what makes an answer correct. Two remediation attempts fail, and we report them as negative results.
4. Evidence that paired testing matters: it turns a borderline unpaired pass on a published cleaned subset into a clear failure.

We do not claim the surface probes are new. They are hypothesis-only and partial-input baselines [4, 13], used with an AFLite-style filter [2]. The contribution is the certification rule built around them and the evidence it produces.

2 Method

2.1 Surface probes and the unpaired test

The base test is `nulltest` v1.4. It asks whether a binary label can be predicted from surface form alone. It uses six model-free probe families:

- *lexicality*: character 3–5-grams;
- *words*: unigrams and bigrams;
- *length*;
- *shape*;
- *alpha_first* and *alpha_last*.

Each probe is scored by two-sided discrimination, $A^* = \max(\text{AUROC}, 1 - \text{AUROC})$, in every cell (a probe crossed with a stratum such as a benchmark category). The test reports the *worst* cell, not the mean. The null distribution shuffles labels within each stratum and refits every probe inside every permutation. p-values are corrected over all cells with the Westfall–Young min-p procedure [20]. A result is PASS when the worst cell's adjusted $p \geq 0.05$. It is flagged borderline when that p is within a factor of two of α.

2.2 Paired mode

Many truthfulness benchmarks come in pairs: a right answer and a wrong answer to the same question. Comparing the two members of a pair removes everything they share. In paired mode the statistic is w, the share of pairs in which the label-1 member scores higher (ties count $\frac{1}{2}$), and $A^*_{\text{pair}} = \max(w, 1 - w)$. Cross-validation folds are grouped by pair. The null flips the labels within each pair with probability $\frac{1}{2}$, and Westfall–Young correction is applied over cells. Paired mode asks a sharper question than the unpaired test: whether a reader can tell which member of a pair is the correct one.

2.3 Pair-level adversarial filter

The filter follows AFLite [2].

1. Each pair is scored out-of-fold by four probe families (lexicality, words, length, shape) over five pair-grouped fold splits (seeds 11–15).
2. A probe "calls" a pair when it ranks the two members in the direction that probe favours across the whole set. A pair's *cue rate* is the share of its (probe, seed) calls that are correct.
3. Each round drops the 5% of pairs whose cue rate is furthest from $\frac{1}{2}$, taken from *both* tails.
4. After each round the survivors are re-screened with paired and unpaired `nulltest` (200 permutations, seed 7). The filter stops when both screens reach adjusted $p \geq 0.20$, or when the next round would go below the floor (§2.5).

Two variants failed during development:

- Dropping only the tail of pairs the probe calls correctly made things worse. The survivors carried the reversed cue, and on TruthfulQA-790 the unpaired A^* rose from 0.67 to 0.84.
- A fixed threshold on the cue rate collapsed the 790 pairs to 25, because per-pair rates are noisy across fold seeds.

Unpaired sets use the same filter on items, stratified by class.

2.4 Certification gates

Certification uses seeds no filter round used: seed 20260724 for the in-family tests, and seed 20261026 with 2,000 permutations for the held-out probes. Every gate must pass, and a borderline result counts as a failure.

G1 In-family. Paired and unpaired `nulltest` on the kept set. This gate is necessary but circular, because the filter optimised against these probes.

G2 Held-out probes. Eight probe types the filter never saw:

- counts of negation words, hedge words and function words;
- punctuation;
- digits;
- capitals;
- mean word length;
- a refit character 3–5-gram Naive Bayes.

Paired sets use the within-pair sign-flip null. Unpaired sets use a single AUROC pooled over all items, with labels permuted within each stratum (§4.1 discusses a consequence). Westfall–Young correction is applied over the eight probes. We report the "worst" held-out probe: the one with the smallest adjusted p, with ties broken by the largest A^*.

G3 Negation cue. Paired sets use an exact two-sided sign test on pairs whose members differ in negation count. Unpaired sets use a two-sided permutation test on the difference in mean negation count. The gate passes if $p \geq 0.05$.

2.5 The floor

A pass must be able to rule something out, and it must describe the benchmark it names. The floor for a paired source of n pairs is

$$\text{floor}_{\text{pairs}}(n) = \max(\lceil 0.5\, n \rceil,\ 320)\,.$$

- **The 320 term.** This is the smallest n at which an exact two-sided sign test at $\alpha = 0.05/8$ (Bonferroni over the eight held-out probes, which is conservative next to Westfall–Young) has at least 80% power against a probe that picks the right member in 60% of pairs. The repository's test suite checks this value, along with two sensitivity values: with no multiplicity correction ($k = 1$) the floor would be 199, and against a 65/35 cue it would be 142.
- **The 0.5 term.** If more than half of a benchmark has to be removed, the cue comes from how the benchmark was built, not from a few bad items, and the kept set is a selected minority.

For TruthfulQA-790 the floor is 395 pairs. For a 2,000-pair HaluEval sample it is 1,000.

Unpaired sets need at least $\max(\lceil 0.5\,n \rceil, 640)$ items and at least 320 per class. The per-class 320 is *borrowed* from the paired calculation and has not been derived separately for AUROC tests. An unpaired AUROC test at the same effect size would probably need fewer items, so this floor is likely conservative (§7).

2.6 Statuses

- CERTIFIED: every gate passes, on a kept set at or above the floor.
- NOT_CERTIFIED: the kept set is at or above the floor, but a gate fails.
- NOT_CERTIFIABLE: the set cannot be cleaned without collapsing. The certificate records which of three reasons apply:
  - `source_below_floor`: the source is already smaller than the floor;
  - `kept_below_floor`: the kept set is smaller than the floor;
  - `filter_reached_floor_with_cues`: the filter reached the floor while the screens were still firing.

The gates are run and reported in every case, so a NOT_CERTIFIABLE set still documents its cues.

2.7 Certificate format

Each certificate is a JSON file in format `floor/certificate/v1`. It records:

- the mode, dataset and issue time;
- the status, the status reasons and the status rule, with the floor parameters;
- the SHA-256 of the source and of the kept split;
- every filter round;
- the full output of every test, and the gate booleans;
- the SHA-256 of every tool used;
- a replay command.

Every certificate also carries a fixed caveat: a pass means these probe families do not recover the labels on the kept items, not that no surface reader can, because no model-based reader is run. v1 certificates are unsigned prototypes.

2.8 Four-arm control standard (specified, not run)

FLOOR also specifies how a probe result *on* a certified split should be reported. The report has four arms:

1. the real activation signal;
2. an input-only baseline: a surface classifier, plus a second model that reads the same text;
3. a prompt-only arm;
4. a shuffled-signal arm.

The reported quantity is the margin AUROC(real) − max(AUROC(controls)), with a permutation confidence interval. A leaderboard would rank by margin, never by raw AUROC. **No results in this paper use the four-arm standard.** The runner, including the second-model reader, has not been built yet (§8).

3 Datasets

We certified every public dataset with a row on `https://honesty.tools/audit` that we are permitted to use. We also include one new first-party set, `aperture/knowledge-prominence-v3` (§4.2). Some further third-party sets were audited but are excluded here pending author permission. The 19 included sets fall into two groups.

First-party batteries (10). These are Honesty.tools first-party batteries, released under Apache-2.0. We use their dataset IDs exactly as they appear in the repository and on `https://honesty.tools/floor`.[1] Their provenance comes from the battery files that the repository's audit manifest (`nulltest_audit_manifest.json`) loads, and whose SHA-256 the certificates record. Where a battery's construction was not recorded, we say so.

- `aperture/postcutoff-trials-v1` (702 items, the certified set). Brief titles of real trials from ClinicalTrials.gov (API v2). Label 0: first posted between January 2023 and June 2024, which a model may have seen. Label 1: first posted between April and June 2026, which it cannot have. Both halves are real titles from one registry. They are length-matched, and titles containing an explicit year were removed.
- `aperture/calibration-battery-v1` (168 items, six categories, built 2026-06-10). The public battery behind the site's `/calibrate` page: real versus fabricated entities. How the fabricated entities were generated is not recorded in the battery file or the repository.
- `aperture/grounding-coordinate` (162 items, six categories, e.g. two-author papers, eponymous theorems, drugs and compounds). Real versus fabricated entities. The fabrication method is not recorded.
- `aperture/photon-entities` (103 items, seven categories). Real versus fabricated entities. The fabrication method is not recorded, and three categories (math, proteins, trials) contain only label-0 items.
- `aperture/verifydomain-claims-147` (147 claims: 78 true, 69 false, 36 of the false ones about coined entities). It was recovered on 2026-07-25 from a single machine where it had lived outside version control. How the false claims were written is not recorded.
- `aperture/prominence-people-v1` (168 items, built 2026-07-25). Real people only. Label 0: high en.wikipedia pageviews (2,408–80,824 mean daily views; famous half from the May 2026 top-viewed list). Label 1: low pageviews (1.2–38.9), drawn from one of the paired

[1] Dataset IDs are kept exactly as they appear in the repository, because they are real identifiers. The ten IDs beginning `aperture/` are Honesty.tools first-party sets.

famous person's own Wikipedia categories. Known defect: the builder measured only the alphabetically first 200 members of each category, so surname initial recovers the labels. It is kept as evidence.

- `aperture/prominence-people-v2` (156 items, built 2026-07-25). The same design, also matched within each pair on surname initial (78 of 78 pairs) and name length. Low-prominence range 1.3–49.8 daily views.
- `aperture/knowledge-prominence-v1` (324 items, five categories, built 2026-07-25). Real entities. Label 0: high en.wikipedia pageviews; label 1: low pageviews from the same Wikipedia category. Matched within the pair on category, last-word initial and name length. Known defect: 51 of the 80 *book* items are fictional characters, pulled in by subcategory traversal.
- `aperture/knowledge-prominence-v2` (296 items, built 2026-07-26). v1 with subcategories filtered by name before traversal. **Its construction is only partly recorded.** The pageview window, the high and low thresholds, the seed category lists and the exact pairing rules were not saved, and no builder script survives. So it cannot be grown with the same method. This is why we built v3.
- `aperture/knowledge-prominence-v3` (1,968 items, built 2026-10-27 for this paper with a committed builder; §4.2).

Third-party sets (9).

- Binary TruthfulQA (790 pairs), from Namjoo et al. [11], built on Lin et al. [7];
- HaluEval qa, dialogue, summarization and general [6];
- RAGTruth responses [12];
- truth-is-universal facts [3];
- GPT-4 hallucinated references [1];
- bielik-entity-familiarity v2, MIT licence: `https://github.com/agentGreg/bielik-entity-familiarity`.

Sets with a natural right/wrong pairing (TruthfulQA, HaluEval qa, dialogue and summarization) run in paired mode, and the rest run unpaired. Sets marked * in tables and figures are seeded samples:

- each HaluEval paired set is a seeded 2,000-pair sample of its 10,000 pairs;
- RAGTruth is a seeded, stratified 4,000-item sample of the audit's 8,000 items, which are themselves a seeded subsample of 17,790.

4 Results

4.1 The v1 suite

Table 1 and Figure 1 give the outcome. One set is CERTIFIED: `aperture/postcutoff-trials-v1`. All of its 702 items (351/351) were kept.

- Its filter screen stopped at adjusted $p = 0.7409$.
- Its in-family unpaired worst cell was $A^* = 0.529$ (adjusted $p = 0.7171$, 6 cells).
- Its worst held-out probe was function words, $A^* = 0.550$ (adjusted $p = 0.133$).

Table 1: FLOOR v1 on 19 datasets. The first block is first-party batteries and the second is third-party sets. Unit: pairs (paired mode) or items. Reasons: s = source below floor, k = kept below floor, c = filter reached floor with cues. Failing gates: P/U = in-family paired/unpaired, H = held-out probes, N = negation cue. The last column is the worst held-out probe with its A^* and Westfall–Young adjusted p over eight probes. 0.0005 is the smallest value 2,000 permutations allow. * marks seeded samples. Source: `benchmark/certified/**/certificate.json` at commit `f1f2cde` (main).

Dataset	unit	source (1/0)	kept	floor	status	reasons	fails	worst held-out probe
aperture/calibration-battery-v1	item	168 (84/84)	168	640	NOT_CERTIFIABLE	s,k,c	UH	char_ngram 0.986 (0.0005)
aperture/grounding-coordinate	item	162 (81/81)	162	640	NOT_CERTIFIABLE	s,k,c	UH	word_len 0.628 (0.0005)
aperture/knowledge-prominence-v1	item	324 (162/162)	324	640	NOT_CERTIFIABLE	s,k	H	caps 0.550 (0.079)
aperture/knowledge-prominence-v2	item	296 (148/148)	296	640	NOT_CERTIFIABLE	s,k	–	caps 0.549 (0.17)
aperture/knowledge-prominence-v3	item	1,968 (984/984)	1,011	984	NOT_CERTIFIABLE	c	UH	char_ngram 0.761 (0.0005)
aperture/photon-entities	item	103 (27/76)	103	640	NOT_CERTIFIABLE	s,k,c	UH	char_ngram 0.865 (0.0005)
aperture/postcutoff-trials-v1	item	702 (351/351)	702	640	**CERTIFIED**	–	–	function 0.550 (0.13)
aperture/prominence-people-v1	item	168 (84/84)	168	640	NOT_CERTIFIABLE	s,k,c	U	caps 0.554 (0.2)
aperture/prominence-people-v2	item	156 (78/78)	156	640	NOT_CERTIFIABLE	s,k	H	word_len 0.648 (0.004)
aperture/verifydomain-claims-147	item	147 (36/111)	147	640	NOT_CERTIFIABLE	s,k,c	UH	char_ngram 0.941 (0.0005)
bielik-entity-familiarity-v2	item	1,440 (240/1200)	1,440	720	NOT_CERTIFIABLE	s,k,c	UH	char_ngram 0.846 (0.0005)
hallucinated-references/gpt-4	item	1,000 (468/532)	697	640	NOT_CERTIFIABLE	c	UH	char_ngram 0.765 (0.0005)
halueval/dialogue*	pair	2,000	1,027	1,000	NOT_CERTIFIABLE	c	PUHN	char_ngram 0.823 (0.0005)
halueval/general	item	4,507 (815/3692)	2,314	2,254	NOT_CERTIFIABLE	c	UHN	char_ngram 0.617 (0.0005)
halueval/qa*	pair	2,000	1,027	1,000	NOT_CERTIFIABLE	c	PUHN	punct 0.924 (0.0005)
halueval/summarization*	pair	2,000	1,027	1,000	NOT_CERTIFIABLE	c	PUH	char_ngram 0.944 (0.0005)
ragtruth/responses*	item	4,000 (1708/2292)	2,052	2,000	NOT_CERTIFIABLE	c	UH	caps 0.608 (0.0005)
truth-is-universal/facts	item	561 (282/279)	561	640	NOT_CERTIFIABLE	s,k,c	UHN	negation 0.544 (0.0005)
truthfulqa-binary	pair	790	405	395	NOT_CERTIFIABLE	c	PUHN	char_ngram 0.760 (0.0005)

- Its negation test gave $p = 1.0$.

The other 18 sets are NOT_CERTIFIABLE, but not all for the same reason:

- **Surface cue at adjusted $p < 0.05$ (16 sets).** In 15 the cue shows up in the held-out probes. In the 16th, `aperture/prominence-people-v1`, it shows up only in-family: the unpaired worst cell has adjusted $p = 0.01$, and the held-out probes pass. This group includes *every* third-party set.
- **Borderline only (1 set).** `aperture/knowledge-prominence-v1`'s only failure is a borderline held-out result (caps, $A^* = 0.550$, adjusted $p = 0.079$). That counts as a failure under the borderline rule, but it is not a significant cue.
- **Size only (1 set).** `aperture/knowledge-prominence-v2` passes every gate (worst held-out probe caps, $A^* = 0.549$, adjusted $p = 0.17$). Its 296 items are below the 640 floor.

So "surface probes read 18 of 19" would overstate the result. The accurate statement is that a surface cue was found in 16 of 19 sets, and one more is borderline.

Where filtering stops. On the large third-party sets the filter ran down to the floor and the cues were still there (Figure 3):

- HaluEval qa, dialogue and summarization each went from 2,000 pairs to 1,027, against a floor of 1,000.
- HaluEval general went from 4,507 items to 2,314 (floor 2,254).
- RAGTruth went from 4,000 items to 2,052 (floor 2,000).
- hallucinated-references went from 1,000 items to 697 (floor 640).

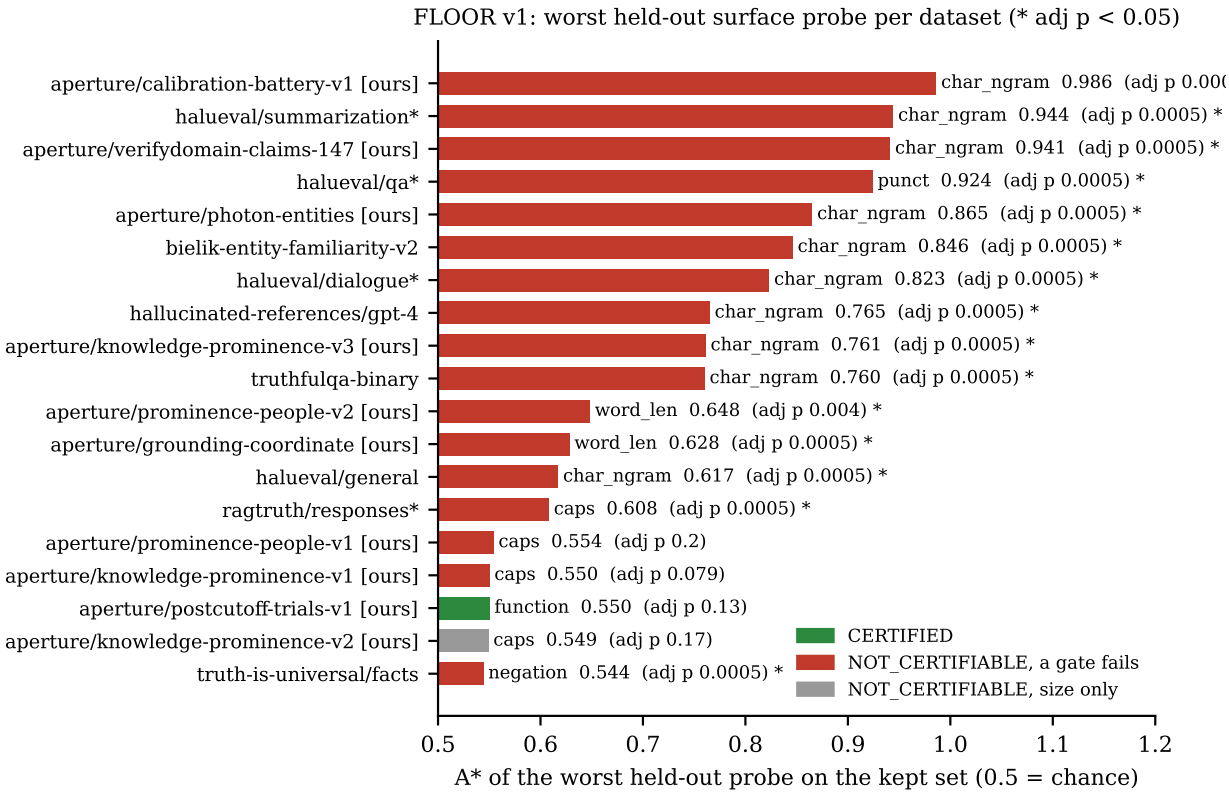


Figure 1: Worst held-out probe per dataset on the kept set, taken from each certificate. Grey: no gate fails, but the set is below the floor.

- knowledge-prominence-v3 went from 1,968 items to 1,011 (floor 984).

Even after removing roughly half of each set, the held-out probes still read the labels strongly:

- HaluEval summarization: character n-gram $A^* = 0.944$;
- HaluEval qa: punctuation $A^* = 0.924$. The paired *shape* probe was at $A^*_{\text{pair}} = 0.963$ (adjusted $p = 0.0033$), and wrong answers carried more negation, 96 pairs to 23 ($p = 8.54 \times 10^{-12}$);
- HaluEval dialogue: character n-gram $A^* = 0.823$;
- HaluEval general: negation, with mean negation count 0.6325 against 0.3984 ($p = 0.0005$).

Most of the older first-party batteries are far below the 640 floor (103–324 items), so they could only ever be NOT_CERTIFIABLE. Several also carry strong cues, e.g. the calibration battery's character n-gram $A^* = 0.986$.

RAGTruth: a high n-gram score with a high p. On RAGTruth the held-out character n-gram reaches $A^* = 0.655$ on the kept set, yet its adjusted p is 0.875. Caps, at a lower $A^* = 0.608$, has adjusted $p = 0.0005$. This is not a computation bug. It follows from how the unpaired held-out statistic is defined, and it does not change any verdict.

We checked this on the 4,000-item source split, which we rebuilt with the repository's loader and sampling seed. Its SHA-256 matches the source hash recorded in the certificate. Rerunning `heldout_np.evaluate_unpaired` with the certificate's seed and 2,000 permutations reproduces

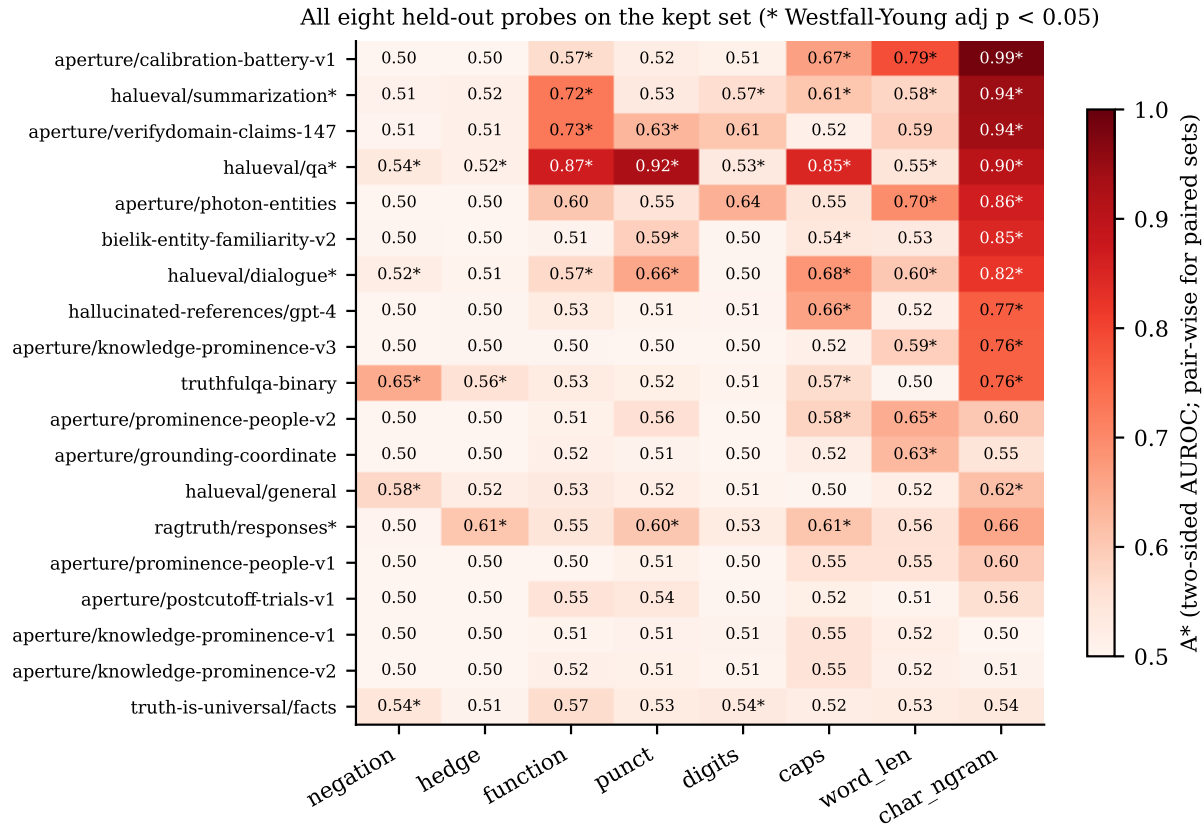


Figure 2: All eight held-out probes. The character n-gram Naive Bayes is the worst-case probe on 10 of the 19 sets. Negation is significant (adjusted $p < 0.05$) on five: TruthfulQA, HaluEval qa, dialogue and general, and truth-is-universal facts.

every certified source value, including the char n-gram's $A^* = 0.679$ (raw p 0.367, adjusted p 0.957).

- **The base rate differs by task.** The share of label 1 (hallucinated) is 0.674 in data-to-text, 0.302 in QA and 0.283 in summarisation.
- **The pooled score mostly encodes the task.** A score that knows only each item's task, and no text, already reaches a pooled AUROC of 0.682. Character n-grams identify the task easily.
- **The null keeps that advantage.** The permutation shuffles labels *within* task, so every permuted label vector keeps the per-task base rates, and the refit n-gram model learns the task again each time. Over 2,000 permutations its null A^* has median 0.677 (5th–95th percentile 0.670–0.686), so the observed 0.679 is unremarkable (raw p 0.367). More permutations would not change this: the null itself sits near 0.68.
- **So the test asks the right question but has low power.** Beating that null means beating the task prior, and the pooled statistic hides cues inside a task. Scored within task, the same out-of-fold n-gram scores do separate the labels: AUROC 0.695 in data-to-text and 0.382 in QA (opposite directions; raw permutation $p = 0.0005$ each), and 0.503 in summarisation ($p = 0.89$). A size-weighted mean of the per-task AUROCs is 0.531, above its null 95th

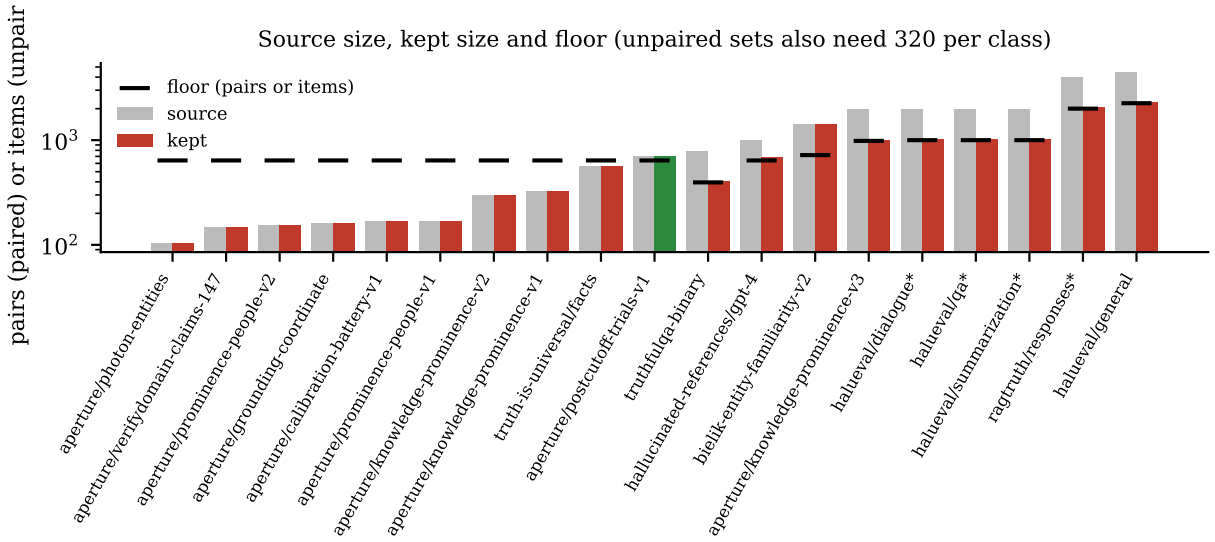


Figure 3: Source size, kept size and floor (log scale). The filter stops at the floor. Every set that reached it still had cues.

percentile of 0.520 (raw $p = 0.005$).

Among the other stratified unpaired sets, calibration-battery-v1, grounding-coordinate, knowledge-prominence-v1, knowledge-prominence-v2 and bielik have the same base rate in every stratum, so a task-only score gives AUROC 0.5 and the stratum prior adds nothing to the null. The other two, `aperture/photon-entities` (uneven base rates: three of its seven categories are all label 0) and `aperture/knowledge-prominence-v3`, already fail on the char n-gram at adjusted $p = 0.0005$. The remaining unpaired sets have a single stratum. The low power therefore affects no verdict here. It would matter for a future set with uneven base rates across strata that passes.

Update (27 October 2026): FLOOR v1.1. FLOOR v1.1 scores the unpaired held-out probes both pooled (as in v1) and within strata (a size-weighted mean over strata of the per-stratum A^*, on the same within-stratum permutation draws, with Westfall–Young over the eight probes), and the gate fails if either family separates. We reran all 15 public unpaired sets (honestytools PR #21, merge commit `bfd5489`). No verdict, gate or status reason changed, and every kept split has the same SHA-256 as in v1. On RAGTruth's kept 2,052 items, the within-stratum character n-gram does not separate ($A^* = 0.541$, adjusted $p = 0.331$). The within-task separation reported above was measured on the 4,000-item source split, where the v1.1 certificate's source check also lists the character n-gram among the probes that separate within stratum. RAGTruth remains NOT_CERTIFIABLE on other probes: its worst within-stratum probe is punctuation, $A^* = 0.613$ (adjusted $p = 0.0005$).

4.2 A new set built to clear the floor: knowledge-prominence-v3

The older first-party knowledge-prominence batteries were too small to certify. We therefore built `aperture/knowledge-prominence-v3` from scratch, and committed the builder (`build_kp_v3.py`) before running any certification. The design was not changed after seeing the result.

Construction.

- **Candidates.** Wikipedia titles drawn from 13 seed categories (seeded sample, seed 20261027).
- **Labels.** Mean daily user pageviews over a fixed 59-day window (2026-08-28 to 2026-10-25). Well-known: at least 50 views a day. Obscure: at most 10.
- **Form filter.** ASCII letters and spaces only, 1–5 words, 4–30 characters.
- **Pairing.** Each well-known title is paired with an obscure title from the same seed category that matches exactly on word count, per-word capitalisation shape, function words and their positions, first letter and last two letters, and is within one character in length.
- **Size.** 984 pairs (881 person, 33 place, 30 book, 27 film, 13 firm), or 1,968 items.

Certification. We certified it in unpaired mode, stratified by category. Status: NOT_CERTIFIABLE, with reason `filter_reached_floor_with_cues`. The filter stopped at 1,011 items (506/505, 51.4%) against a floor of 984.

- In-family: lexicality within the person category, $A^* = 0.751$ (adjusted $p = 0.0092$, 29 cells).
- Held-out character n-gram: $A^* = 0.761$, adjusted $p = 0.0005$.
- Held-out word length: $A^* = 0.590$, adjusted $p = 0.0005$.
- The other held-out probes and the negation test found nothing.

Reading. Matching on form removed the cues we targeted, but the content of the names still separates the classes, because well-known names are built from more common name parts. This is evidence that the cue is hard to remove. It is not an artefact of a careless build: it survives a design aimed specifically at it.

4.3 Paired testing changes the verdict

Namjoo et al. [11] released a cleaned 476-pair subset of binary TruthfulQA. In our unpaired test it got a *borderline pass*: worst cell length, $A^* = 0.551$, adjusted $p = 0.0649$ ($n = 952$ items, 6 cells). Paired mode, stratified by TruthfulQA category, *rejects* it: lexicality $A^*_{\text{pair}} = 0.762$, adjusted $p = 0.0027$ over 223 cells, 9,000 permutations. Within a pair, the right answer can still be identified from its wording.

The full 790-pair set fails both tests, and paired mode reads it more strongly:

- unpaired: words $A^* = 0.67$, adjusted $p = 0.005$;
- paired: words $A^*_{\text{pair}} = 0.825$, adjusted $p = 0.0038$ over 226 cells.

We checked that paired mode is calibrated on real 476-pair text with labels coin-flipped within pairs. It rejected 5 of 100 runs at $\alpha = 0.05$ (200 permutations each). On the 49-pair prototype split, the false-rejection rates were 5 of 100 (in-family), 6 of 100 (held-out family) and 4 of 100 (negation test). All are close to nominal.

4.4 Case study: TruthfulQA's cue is part of the task

Filtering TruthfulQA-790 stopped at 405 pairs (51.3%), just above the 395 floor, after 14 rounds. All four gates still failed:

- in-family paired: lexicality $A^*_{\text{pair}} = 0.773$ (adjusted $p = 0.0021$, 221 cells);
- in-family unpaired: words $A^* = 0.597$ ($p = 0.01$);

- held-out: character n-gram 0.760, negation 0.649, hedge 0.556, all at adjusted $p = 0.0005$;
- negation: the right answer was more negated in 150 of the 179 pairs that differed, against 29 for the wrong answer ($p = 7.12 \times 10^{-21}$).

In the source the negation split was 322 to 51 among 373 differing pairs ($p = 3.17 \times 10^{-49}$).

Negation here is not an artefact that cleaning can remove. Many correct TruthfulQA answers *are* negations ("No, ...", "There is no ..."), because the benchmark targets common misconceptions. The filter cannot strip the cue without stripping the task. An earlier unconstrained filter (no floor) shows the same thing. At 49 pairs the unpaired test passed ($A^* = 0.591$, adjusted $p = 0.767$), but paired lexicality was 0.939 (adjusted $p = 0.0008$), held-out character n-gram was 0.898, and in all 17 kept pairs that differed in negation the right answer was the more negated one ($p = 1.5 \times 10^{-5}$). We tried two fixes (Figure 4).

Option A: negation-balanced selection. We selected 491 pairs (62.2% of the source) in which negation is balanced, 37 to 37 among differing pairs, and certified them without filtering. Status: NOT_CERTIFIED.

- Paired lexicality: $A^*_{\text{pair}} = 0.773$ (adjusted $p = 0.0028$, 217 cells).
- Unpaired length: $A^* = 0.587$ ($p = 0.01$).
- Held-out: character n-gram 0.772 and hedge 0.580 (both at adjusted $p = 0.0005$).

The negation gate passes only because it was balanced by construction. The other cues remain.

Option B: rewrite pilot (negative result). We took a stratified random sample of 100 pairs (seed 20261028) and rewrote both members of each pair with phi-4. The rewriter was not told which answer was correct. It was asked to use a shared template, keep lengths within 2 words and phrase answers affirmatively. We then checked that the meaning was preserved using an NLI model (deberta-large-mnli): each rewrite had to entail its original and be entailed by it. 82 of the 100 pairs passed. The 18 failures fell into overlapping categories:

- 12 changed the meaning of the right answer;
- 5 changed the meaning of the wrong answer;
- 4 collapsed the two answers into the same claim.

With only 82 pairs, both arms are below the 320 floor and so are NOT_CERTIFIABLE regardless. The comparison is diagnostic. Table 2 compares the originals with the rewrites on the same 82 pairs. Rewriting did not remove the cue. It moved it: the held-out character n-gram got stronger, from 0.707 to 0.780. Negation survived the rewrite, 31 to 3. That is what we would expect if negation is part of the meaning that the NLI check is designed to preserve.

5 What this means for probe and lie-detection claims

1. **A high probe AUROC on these sets is not evidence of a truth representation on its own.** On every third-party set we tested, a model-free reader of the text recovers the labels at family-adjusted $p < 0.05$. A probe result on these sets should be reported as a margin over input-only controls, which is what the four-arm standard specifies. We have not yet run that standard ourselves.

Table 2: Option B pilot on the same 82 pairs. Source: `truthfulqa-binary/optb-pilot/{orig,rew}/certificate.json`.

	originals	rewrites
paired worst cell	length 0.707 (p 0.0006, 169 cells)	words 0.701 (p 0.0017, 155 cells)
unpaired worst cell	length 0.614 (p 0.03)	lexicality 0.757 (p 0.01)
held-out char n-gram	0.707 (p 0.018)	0.780 (p 0.0005)
held-out negation	0.677 (p 0.0005)	0.671 (p 0.0005)
held-out hedge	0.579 (p 0.009)	0.561 (p 0.082)
negation sign test (right:wrong)	34:5 ($p = 2.43 \times 10^{-6}$)	31:3 ($p = 7.66 \times 10^{-7}$)
mean |length difference| (words)	2.04	1.43

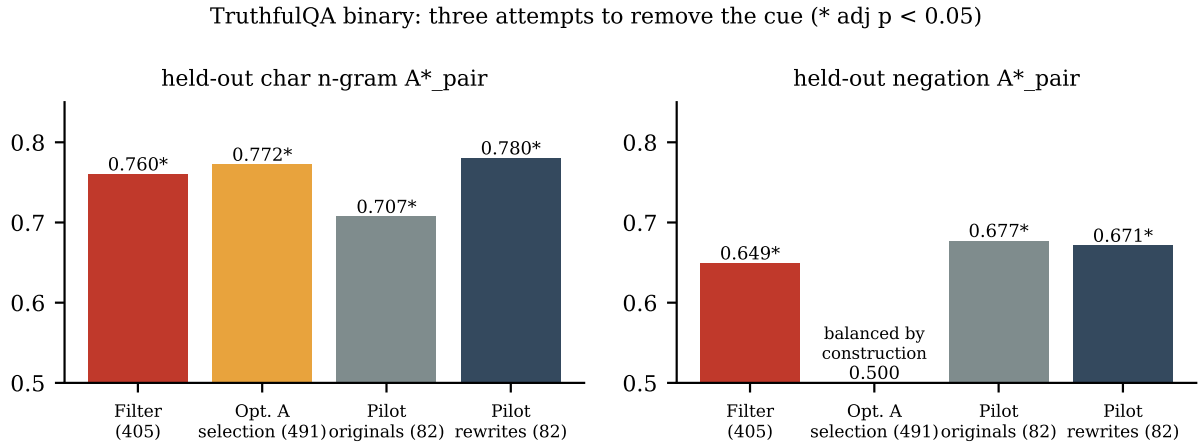


Figure 4: TruthfulQA: the held-out character n-gram and negation probes across the floor filter, option A and the option B pilot. No attempt brings the n-gram probe near chance.

2. **Paired evaluation is the relevant unit when a benchmark is paired.** A borderline unpaired pass became a clear paired failure (§4.3). An unpaired test can miss a cue that only appears when the two members of a pair are compared.
3. **Some cues are part of the construct.** For TruthfulQA, negation is partly what makes an answer correct. NOT_CERTIFIABLE is then the honest outcome. Neither filtering, selection nor rewriting produced a clean split. Claims on such sets need controls, not cleaning.
4. **A readable cue is not proof that a model used it.** NOT_CERTIFIABLE means the labels can be read from the surface. It does not mean any particular model or probe relied on that. Intervention designs can separate the two [8, 19, 21]. Likewise, a verdict on a dataset is not a finding of error or misconduct by its authors.
5. **A pass is narrow.** A pass covers these probes, this sample size and this configuration. It does not exclude model-based readers. Passing an evaluator's tests bounds discrimination from one side only [9].

6 Related work

Artefacts and partial-input baselines. Hypothesis-only baselines exposed annotation artefacts in NLI [4, 13], and AFLite filtered them out adversarially [2]. FLOOR's probes and filter come from this line of work. What FLOOR adds is held-out certification, a sample-size floor and an explicit "cannot be cleaned" status.

Cleaning truthfulness benchmarks. Namjoo et al. [11] is the closest work. They use a question-blind surface probe on binary TruthfulQA, prune at the pair level and release a cleaned 476-pair subset. We use their 790- and 476-pair sets. Our paired test rejects the cleaned subset (§4.3).

Confounds in lie detection and introspection. Published lie-detection probes can follow persona beliefs and confounds rather than truth [18]. Input-only classifiers can match apparent introspection [16]. Prompt-text classifiers narrow the margin of activation probes [14, 17]. Large public deception-probe datasets [5] are natural next targets for FLOOR. Truth-direction work [3] provides one of our audited sets.

Structural and surface baselines elsewhere. Model-free structural rules change how agent-failure monitors look [10]. Action-only monitors dominate pooled chain-of-thought monitor accuracy [15]. Formatting cues dominate detectors of machine-generated text [19]. Each of these makes the same point we do: a score means little without a surface floor.

7 Limitations

- **Seeded samples.** The HaluEval paired sets and RAGTruth were run on seeded samples (2,000 of 10,000 pairs, and 4,000 of 8,000 items). A full-size run could differ.
- **Borrowed unpaired floor.** The per-class 320 for unpaired sets comes from the paired sign-test power calculation. It has not been derived for AUROC tests. An unpaired analysis would probably lower it, and `aperture/knowledge-prominence-v2` (296 items, every gate passing) might then qualify.
- **One certified set.** v1 certifies a single dataset, a first-party battery. We have no certified third-party set and no certified paired set. Our one attempt to build a new set that would clear the floor (knowledge-prominence-v3) also failed.
- **Probe coverage.** No model-based reader is run. A pass rules out only the listed probe families.
- **The four-arm standard is specified, not run.** We have not built its runner, and no result in this paper uses it.
- **Unsigned certificates.** v1 certificates are unsigned prototypes.
- **Rewrite pilot scale.** The option B pilot used 100 pairs and one rewriter. It is evidence, not proof, that rewriting cannot help.
- **Pooled held-out AUROC.** In unpaired mode the held-out probes are scored by one AUROC over all strata. When base rates differ across strata, the refit n-gram probe loses power (§4.1). No v1 verdict depends on this, and FLOOR v1.1 adds within-stratum scoring (§4.1).

8 Planned before submission

- Full-size runs of HaluEval (10,000 pairs per paired set) and RAGTruth.
- A power analysis for the unpaired AUROC floor, to replace the borrowed 320 per class.
- A knowledge-prominence design that controls for how common the name parts are. knowledge-prominence-v3 (§4.2) shows that matching on surface form is not enough.
- The four-arm runner with a second-model reader, and a first margin-ranked probe result on the certified split.
- Signed certificates and a replay verifier.
- Permission decisions on the excluded third-party sets.

9 Reproducibility

Code and certificates are in `https://github.com/techno-optimist/honestytools` at commit `f1f2cde` (main).

- The scripts are `certify.py`, `certify_unpaired.py`, `certify_paired_np.py`, `heldout.py`, `heldout_np.py` and `balance_negation.py`, all under `benchmark/certified/`.
- Every certificate records the SHA-256 of its source, its split and its tools, plus a replay command. For example, `nulltest` v1.4 has SHA-256 `0e103dc4...`.
- `aperture/knowledge-prominence-v3` (builder, candidates, battery and certificate) is under `benchmark/certified/knowledge-prominence-v3/`, merged in `f1f2cde`. The rebuild is deterministic (items SHA-256 `f4d2a8eb...`).
- The RAGTruth diagnostic (§4.1) is in this paper's directory. `scripts/ragtruth_rebuild_source.py` and `scripts/rebuild_small_sources.py` rebuild the source splits (all SHA-256 match their certificates), `scripts/ragtruth_diag.py` runs the analysis, and the outputs are in `data/ragtruth_diag/`.
- Live results are at `https://honesty.tools/floor`.
- Every table and figure in this draft is generated by `scripts/` from those certificates. `NUMBERS.md` lists the source of every number in the text.

Acknowledgements

We thank the authors of the public datasets this paper certifies: TruthfulQA [7], HaluEval [6], RAGTruth [12], the truth-is-universal facts [3], the GPT-4 hallucinated references [1], and bielik-entity-familiarity (`https://github.com/agentGreg/bielik-entity-familiarity`). Their decision to release data openly is what made this audit possible. The work was done by Kevin Russell with Atlas and the AI agents of the Honesty.tools team.

References

[1] Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. Do language models know when they're hallucinating references? arXiv:2305.18248, 2023.

[2] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. arXiv:2002.04108, 2020.

[3] Lennart Bürger, Fred A. Hamprecht, and Boaz Nadler. Truth is universal: Robust detection of lies in LLMs. arXiv:2407.12831, 2024.

[4] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. arXiv:1803.02324, 2018.

[5] Oskar J. Hollinsworth, Alex F. Spies, Tigist Diriba, Adam Gleave, and Chris Cundy. Telling FIBS: Probes effectively detect sabotage and catch unverbalized deception. arXiv:2610.12445, 2026.

[6] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. arXiv:2305.11747, 2023.

[7] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. arXiv:2109.07958, 2021.

[8] Junhao Liu, Tianxiao Li, Yongxian Wu, Ziheng Duan, Ren Li, Jing Zhang, and Martin Renqiang Min. Auditing tool traces in agentic LLMs: Source localization and counterfactual dependence testing. arXiv:2610.25540, 2026.

[9] Ahmed M. Mahran. What lies beyond perception: A partition theory of the limits of AI evaluation and oversight. arXiv:2610.30803, 2026.

[10] Ramesh Arvind Naagarajan and Stefan Streif. Structural baselines matter for agent failure detection and attribution. arXiv:2610.16661, 2026.

[11] Foad Namjoo, Remy Ogasawara, Amirali Abdullah, Cullen Anderson, Narmeen Fatimah Oozeer, and Jeff M. Phillips. Judging by the cover: Cleaning LLM truthfulness benchmarks to avoid surface-level feature leakage. arXiv:2609.13003, 2026.

[12] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. arXiv:2401.00396, 2023.

[13] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. arXiv:1805.01042, 2018.

[14] Parsa Rahimi and Azim Dehghani Amirabad. Do knowledge and reasoning oriented benchmarks induce stable representation signatures in LLMs? arXiv:2610.24029, 2026.

[15] Shikhar Shiromani, Leo Richter, and Noah Siegel. A false average: Pooled CoT-monitor accuracy conceals a reasoning-dependent fragility. arXiv:2608.00583, 2026.

[16] Shashwat Singh, Tal Linzen, and Shauli Ravfogel. Can LLMs introspect? a reality check. arXiv:2605.26242, 2026.

[17] Marios Tsatsos. Linear probes transfer from base to distilled language models despite rotation. arXiv:2610.35586, 2026.

[18] Maximilian von Klinski, Sebastian Lapuschkin, Wojciech Samek, and Lennart Bürger. Stress-testing LLM lie detectors: Role-play failures and spurious correlations. arXiv:2609.39807, 2026.

[19] Anastasia Voznyuk and Andrey Grabovoy. Measuring artifact reliance in machine-generated text detectors. arXiv:2610.13583, 2026.

[20] Peter H. Westfall and S. Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1993. ISBN 0-471-55761-7.

[21] Esther Xin. When a data artifact isn't a shortcut: Causal auditing of synthetic RLVR corpora. arXiv:2610.00202, 2026.