WHAT DO HALLUCINATION BENCHMARKS MEASURE? EXPLORATORY FACTOR ANALYSIS FOR DETECTION AND BENCHMARK DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination detectors are usually evaluated in the same way, although they capture different properties of generated text. We present an unsupervised measurement framework that treats hallucination severity as a latent trait and uses a battery of indicators to distinguish input-conflicting (IC) hallucinations, which are unsupported by a source document, from context-conflicting (CC) hallucinations, inconsistent with the generated text itself. Within each IC and CC battery, heterogeneous indicators are rank-normalised and combined using a bifactor exploratory factor model, yielding a continuous score without hallucination labels, human-written reference outputs, or API calls. We evaluate seven benchmarks covering summarisation, source-grounded question answering, and document-level contradiction. On five IC-only benchmarks, source-grounding scores discriminate hallucination labels better than output-only coherence scores. Detection performance is competitive but varies markedly across datasets. Furthermore, Factor loadings show how strongly semantic similarity and source consistency contribute to the shared score. Across the benchmarks, detection is easier when faithful and hallucinated outputs differ in how much source wording they reuse, although our score provides information beyond this difference. After stratifying examples with similar levels of lexical novelty, our method retains more discriminative signal on benchmarks constructed from LLM-generated hallucinations than on extractive summarisation benchmarks, while generally adding predictive information beyond novelty alone. These findings suggest that current leaderboards can conflate factuality sensitivity with extractiveness. We therefore propose lexical-stratified evaluation and incremental-validity testing as complementary checks for hallucination benchmarks.

1 Introduction & Background

Hallucination detection is essential when consuming content generated by large language models performing source-grounded tasks like summarisation (Ji et al., 2023). Existing detectors approach the problem through different signals: alignment and entailment against a source (Zha et al., 2023; Tang et al., 2024a), consistency across sampled answers (Manakul et al., 2023), predictive or semantic uncertainty (Kuhn et al., 2023; Farquhar et al., 2024), and direct judgements from large language models (Zheng et al., 2023). These methods are commonly treated as competing estimators. In practice, however, each measures only part of the phenomenon and introduces its own scale, noise, and inductive bias. Combining these signals is non-trivial for two reasons: *a)* Supervised ensembles require labelled hallucination examples (Rykov et al., 2024), while LLM judges add inference cost and systematic judge biases (Zheng et al., 2023; Gu et al., 2026; Chen et al., 2025; Bavaresco et al., 2025; Fujinuma, 2026; Liu et al., 2026) and *b)* Benchmark scores are hard to interpret: detector rankings vary across domains, generators, and annotation protocols (Tang et al., 2023; Bao et al., 2025). Furthermore, recent audits show that apparently strong detectors exploit lexical proxies or construction artifacts and can deteriorate sharply under human-aligned or out-of-distribution evaluation (Janiak et al., 2025; Hussain & Kantarcioglu, 2026). A high AUROC can therefore reflect genuine sensitivity to unsupported content, an easy distinction between benchmark classes, or both. This ambiguity matters not only for model comparison but means a detector used for post-training may be directly optimising whatever shortcut the benchmark rewards.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this work, we address the above issues. Following Zhang et al. (2023), we distinguish *input-conflicting* (IC) hallucinations, which are unsupported by a supplied source, from *context-conflicting* (CC), which violate the generated text's internal coherence, supplying test batteries for both. IC indicators depend on both the output and its source, while CC indicators are computed from the output alone; we estimate the two batteries separately rather than assuming that faithfulness and coherence are interchangeable. Fact-conflicting claims requiring external world knowledge are outside the present scope.

Factor analysis has been used to separate shared signal from scenario-specific noise within a single uncertainty estimator (Wen et al., 2025). We use factor analysis as an aggregation function to combine heterogeneous indicators as measurements of a latent construct supporting a common interpretation. We then test whether its detection performance reflects source faithfulness, lexical separability, or benchmark composition. Given several reference-free indicators, we rank-normalise their scores and fit an exploratory factor model to their dependence structure, separately for each dataset and battery. A Schmid-Leiman transformation isolates variation shared across the battery from indicator-group-specific variation. The resulting general-factor score is continuous and label-free; an anchor-factor score can isolate the dimension most closely associated with source alignment or contradiction when the fitted structure is multidimensional. We use hallucination annotations only after score construction, for validation. The framework thus turns a collection of imperfect indicators into a measurement instrument while providing diagnostics for whether the assumed common structure is supported by the data.

Our seven benchmarks (AggreFact, FaithBench, HaluEval, RAGTruth, TofuEval, SummEval, and ContraDoc) span summarisation, source-grounded question answering, paired faithful and hallucinated outputs, and document-level contradiction. Across the six datasets with IC annotations, the IC indicator loadings show strong agreement. On the five datasets annotated only for IC, we can still compute the output-only CC indicators. Their composite serves as a comparison against the IC labels and it predicts those labels less well than the IC score in every case, consistent with a distinction between source faithfulness and internal coherence. Detection is competitive on AggreFact and RAGTruth but weaker on FaithBench, TofuEval, and ContraDoc. We use this variation in performance to examine the benchmarks themselves. We find a bigram-novelty baseline to be highly discriminative on AggreFact and SummEval; across the six source-conditioned benchmarks, lexical separation is strongly associated with hallucination detection AUROC. Lexical stratification and nested predictive tests nevertheless show that our proposed severity score retains non-lexical information, on most datasets, with FaithBench being the principal exception.

Our contributions are threefold:

- We introduce a label-free factor-analytic framework that combines heterogeneous hallucination indicators into separate continuous scores for source-grounding failures (IC) and failures of internal coherence (CC) (§2.3).
- We evaluate detection and construct validity across seven benchmarks (§4.1; §4.2). We interpret indicator loadings alongside scoring ablations to identify which signals the aggregate score represents and where shared variation alone provides an incomplete measure of source grounding (§4.1).
- We show that current benchmark performance is systematically associated with lexical separability (§4.3.1). We propose lexical-stratified reporting (§4.3.2) and incremental-validity tests (§4.3.3) to distinguish factuality signal from surface extractiveness.

2 Method

We treat hallucination severity as a latent variable measured imperfectly by several reference-free indicators. Given a dataset and associated model outputs, we estimate hallucination severity based on indicator interdependence, without using hallucination labels. The model is fitted separately to each dataset and we do not assume that factor loadings transfer across domains, generators, or annotation protocols. Our working measurement assumption is that variation in hallucination severity induces shared variation across indicators, even though no individual indicator completely measures the construct. Factor analysis separates this common covariance from group-specific and residual variation associated with particular indicator families or implementations. The resulting ensemble is therefore used as a measurement model rather than simply for prediction. To evaluate that the fac-

tor ensemble represents hallucination severity we perform tests on reliability, cross-dataset loading replication, discrimination from the CC battery, and associations with graded human annotations.

2.1 Indicator batteries

For each generated output, we construct two indicator vectors (batteries). The input-conflicting (IC) battery measures whether the output is supported by its source document, whereas the context-conflicting (CC) one measures coherence within the output itself. These batteries represent different constructs and are therefore modelled independently. Our primary IC battery contains six source-aware indicators and the CC battery contains five context-aware indicators. Appendix A gives the complete definitions and implementation details. Our method uses neither gold hallucination labels nor reference outputs. When a benchmark provides a human-written reference, reference-based indicators are introduced in a separate ablation in Appendix H.

2.2 Data preparation

Let $\mathbf{x}_i \in \mathbb{R}^p$ denote the indicator vector for a given example i. Before fitting the indicators, we exclude any that duplicate or exactly inverts another indicator, and remove any indicator that takes the same value for every example in a dataset. Because the indicators have different ranges and often exhibit substantial skew, we transform each column using a Rank-based Inverse Normal Transformation (RINT) (McCaw et al., 2020). The transformation definition and implementation details are provided in Appendix A.2. This transformation places the indicators on comparable scales. We align indicator directions before factor estimation: for example, source alignment and contradiction density initially point in opposite directions. The final combined score is oriented so that higher values indicate greater hallucination severity; If factor fitting is not feasible, we instead average the transformed indicators after orienting each toward greater severity; we also evaluate this mean as an ablation; the eligibility criterion and score definition are given in Appendix A.2.

2.3 Exploratory Factor Analysis (EFA)-bifactor model

We first determine the number of first-order factors through parallel analysis (Horn, 1965). We generate 1,000 matrices with the same number of examples and indicators as the observed data, drawing each entry independently from a standard normal distribution. We compare eigenvalues of the observed indicator correlation matrix with the 95th percentile of the corresponding eigenvalues from these simulated matrices. If none exceeds this 95th percentile threshold, we fit one factor by convention.

We then fit an exploratory factor model by principal-axis factoring (Cramer & Howitt, 2004) with Promax rotation (Zuckerman & Aluja, 2015). For an instance i, the oblique solution, which allows factors to be correlated, has the form

$$\mathbf{z}_i = \mathbf{\Lambda}\mathbf{f}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\mathbf{\Lambda}$ is the loading matrix, $\mathbf{f}_i$ contains the correlated first-order factors, and $\boldsymbol{\epsilon}_i$ is indicator-specific residual variation.

When more than one factor is retained, we apply a Schmid–Leiman transformation (Schmid & Leiman, 1957) to the oblique solution. This yields one general factor and a set of orthogonal group factors:

$$\mathbf{z}_i = \mathbf{g}\eta_i + \mathbf{B}\boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i, \tag{2}$$

where η_i is the general factor, $\mathbf{g}$ contains its indicator loadings, and $\boldsymbol{\xi}_i$ and $\mathbf{B}$ denote the group factors and their loadings. The general factor captures variation shared across the indicator battery, while the group factors retain covariance specific to subsets of indicators.

2.4 Severity scores

When $k = 1$, the single first-order factor serves as the general factor G. When $k > 1$, the Schmid–Leiman solution provides the general-factor loading vector $\mathbf{g}$. In both cases, our primary severity score is the corresponding loading-weighted composite:

$$S_i^G = \frac{\mathbf{z}_i^\top \mathbf{g}}{\sum_{j=1}^{p} g_j}, \tag{3}$$

where $\mathbf{z}_i$ is the RINT-transformed indicator vector for example i. If the signed loadings nearly cancel out each other, we instead normalize by $\sum_j |g_j|$. The resulting score is continuous and dataset-relative rather than a calibrated probability of hallucination.

When parallel analysis retains more than one factor, the general factor may combine the grounding signal with variation specific to other indicator groups. We therefore define an additional *anchor-factor score*. We fit a separate minimum-residual EFA with oblimin rotation (Jackson, 2005) to the same RINT-transformed indicators and select the factor on which a pre-specified anchor has the largest absolute loading:

$$\ell^\star = \underset{\ell \in \{1,\dots,k\}}{\arg\max} |\lambda^{\text{obl}}_{a\ell}|, \tag{4}$$

where $\lambda^{\text{obl}}_{a\ell}$ is the loading of anchor indicator a on oblimin factor ℓ. The anchor-factor severity score is the corresponding factor score:

$$S_i{}^A = f^{\text{obl}}_{i\ell^\star}. \tag{5}$$

Anchor selection and score orientation use prespecified, label-free conventions; when $k = 1$, the general composite is retained (Appendix A.2).

We quantify the reliability of the general factor using hierarchical omega (Flora, 2022), ω_h, which estimates the proportion of variance in the indicator composite attributable to G. Additional factor diagnostics and reliability estimates are reported in Appendix D.

The two scores provide complementary understanding of the fitted structure. S_i^G summarizes variation shared across the complete battery, whereas S_i^A targets the recovered dimension most closely associated with source grounding or internal contradiction. We compare both scores with the equal-weight composite in Section **??**.

3 Experimental Setup

We evaluate the proposed scores on seven publicly available benchmarks. Factor models are fitted independently to the unlabelled indicator matrix of each dataset and battery.

3.1 Datasets

We evaluate seven public benchmarks: AggreFact (Tang et al., 2023), FaithBench (Bao et al., 2025), HaluEval (Li et al., 2023), RAGTruth (Niu et al., 2024), TofuEval (Tang et al., 2024b), SummEval (Fabbri et al., 2021) and ContraDoc (Li et al., 2024). The first six provide IC annotations and latter two provide CC annotations, with SummEval covering both constructs. The five IC-only benchmarks support cross-battery discriminant validation. Dataset sizes, annotation types, splits, exclusions, label mappings, and sampling procedures are reported in Appendix B and Table 5.

3.2 Scoring configurations

The six reference-free IC indicators cover generation probability conditioned on the source (BARTScore), semantic source and output alignment (BERTScore precision and recall), source-entity coverage (entity recall), and learned source and output consistency (SummaC-ZS and AlignScore). We combine these partly overlapping signals because each captures a different aspect of grounding, while factor analysis estimates the variation they share. The CC battery instead uses output-only measures of inter-sentence contradiction, entity continuity, semantic drift, repetition, lexical diversity, and fluency. Indicator definitions and implementations are given in Appendix A. CC indicators are computed from the generated text alone. Each dataset-battery combination is fitted separately: no loadings, thresholds, or calibration parameters are transferred across datasets.

We evaluate the general-factor score, the equal-weight RINT composite, and, when $k \geq 2$, the anchor-factor score defined in Section 2.4. SummEval additionally supports an ablation comparing the six-indicator reference-free battery with reference-only and combined batteries.

Sentence-level pooling: Document-level averaging can dilute the effect of a localized unsupported sentence on otherwise faithful content. We therefore evaluate a second, fixed IC configuration that augments the six-indicator battery with worst-sentence scores for AlignScore and SummaC. Let m_{is} denote the original AlignScore or SummaC value for sentence s of output i, where higher values

indicate stronger source support. We take the minimum to capture the least-supported sentence:

$$m_i^{\text{worst}} = \min_s m_{is}. \tag{6}$$

The original mean-pooled indicators remain in the battery, allowing the factor model to represent both global and localized evidence. We completed the sentence-level extension on RAGTruth, FaithBench, and TofuEval whose multi-sentence outputs can mask localized errors under document-level averaging; the exact selection criteria and computational exclusions are documented in Appendix B.1. Duplicate or near-constant indicator columns are removed before factor fitting.

3.3 Evaluation protocol

For comparison of indicators against binary annotations of hallucination presence, the continuous severity scores are passed directly to AUROC and AUPRC; no threshold or conversion to a probability is required. We report percentile-bootstrap 95% confidence intervals using 1,000 resamples. Because AUPRC depends on class prevalence, dataset base rates are reported alongside it. To calculate balanced accuracy, macro-F1, and MCC we require binary predictions. We thus select the threshold that maximizes Youden's J on the training portion of each five-fold stratified split and apply it to the held-out fold. The reported values therefore use no test-fold labels for threshold selection.

We additionally evaluate calibration, selective prediction, and construct-appropriate graded or pairwise outcomes where available. Exact estimators, confidence intervals, thresholding procedures, and random seeds are specified in Appendix C. We assess construct validity separately from detection performance. Reliability is measured using hierarchical omega; cross-benchmark replication is measured by Tucker congruence between IC loading vectors over the six common indicators. For discriminant validity, we compare the IC general-factor score with an oriented equal-weight CC composite on the five IC-only datasets. Full loading matrices, KMO and Bartlett diagnostics, communalities, and bootstrap procedures appear in Appendix D.

3.4 Benchmark diagnostics and statistical tests

To test whether benchmark performance can be explained by surface extractiveness, we make use of bigrams in the summary that are absent from the source as a deliberately simple *lexical-novelty* baseline. We compare its AUROC with ours using paired DeLong tests (DeLong et al., 1988) and control the six dataset-level comparisons using Holm correction (Holm, 1979). Raw AUROC compares rankings across all faithful and hallucinated outputs, even when the two classes occupy different parts of the lexical-novelty distribution. To reduce this source of separation, we match faithful and hallucinated examples within novelty quartiles before recomputing AUROC, so comparisons are made between outputs with similar levels of source reuse. If stratified AUROC falls toward .5, much of the raw discrimination was associated with lexical separability. A small decline indicates that the detector continues to rank hallucinations above faithful outputs after novelty is balanced. Retained signal expresses this remaining above-chance discrimination as a fraction of the raw above-chance discrimination. This differs from incremental-validity testing: stratification balances novelty during evaluation, whereas the nested model tests whether severity adds predictive information once novelty is included as a covariate. We additionally report detector-novelty correlations, novelty-stratified AUROC computed through within-quartile matching, and a nested likelihood-ratio test comparing lexical novelty alone with a model that also includes severity (Table 3). Definitions, regression specifications, bootstrap procedures, and multiple-testing corrections are provided in Appendix G.

4 Results

We report detection performance, construct validity, and benchmark diagnostics. Table 1 includes the general score G, the anchor-factor score α_G, and the sentence-augmented general score S_G. Sentence-level ablations were completed on RAGTruth, FaithBench, and TofuEval. AggreFact is predominantly single-sentence and SummEval was not evaluated under this configuration. Factor fitting and anchor selection are label-free; Appendix E reports all completed configurations.

Table 1: Detection performance, ordered by the first metric in each panel. Configurations are G, α_G, and S_G which vary per benchmark as G summarises variation shared across indicators, α_G targets an anchor-associated dimension in multi-factor solutions, and S_G captures localized errors that document-level averaging may obscure.

AggreFact: detector scores, ordered by AUROC

	OURS (G)	QAFactEval	Bigram novelty	SummaC Conv	DAE	QuestEval	SummaC ZS
AUROC	.896	.883	.879	.857	.855	.843	.836
AUPRC	.880	.836	—	.790	.825	.788	.753
ρ with novelty	.796	.725	—	.820	.721	.717	.709

AggreFact: thresholded results, ordered by balanced accuracy

	OURS (G)	o3-mini-high	GPT-4o	Bespoke-MC	HHEM-1.0	HHEM-2.1	AlignScore-L
BalAcc	.831	.826	.816	.799	.789	.766	.727
F1	.815	.809	.788	.804	.797	.762	.742

FaithBench: ordered by balanced accuracy

	FaithJudge o3	o3-mini	HHEM-C	Bespoke MC	MiniCheck R	OURS (S_G)	HHEM-S	AlignScore L
BalAcc	.840	.688	.667	.557	.550	.545	.526	.503
F1	.821	.607	.637	.473	.532	.554	.329	.261

HaluEval: ordered by accuracy/balanced accuracy

	Factored Verification	GPT-4 CoT	OURS (α_G)	ChatGPT CoT	ChatGPT few-shot
Acc./BalAcc	.762	.755	.697	.612	.585

RAGTruth: ordered by balanced accuracy

	GPT-4o	Bespoke-MC	OURS (S_G)	MiniCheck-R	HHEM-2.1	AlignScore-L
BalAcc	.851	.794	.724	.663	.644	.528
F1	.809	.771	.530	.609	.671	.496

TofuEval: ordered by balanced accuracy

	Qwen-72B	Bespoke-MC	HHEM-2.1	OURS (S_G)	AlignScore-L	MiniCheck-R
BalAcc	.834	.788	.694	.631	.574	.544
F1	.800	.786	.621	.579	.392	.454

ContraDoc: ordered by accuracy/balanced accuracy

	GPT-4o	Llama-3.2	OURS (α_G)
Acc./BalAcc	.761	.645	.511
F1	.694	.510	.516

Notes. Our thresholded results use five-fold out-of-fold predictions. AggreFact AUROCs use 1,000 examples for our score and bigram novelty, versus 4,811 for the published detector baselines; rankings are therefore contextual. Other panels may also differ in subsets or metric definitions. Published HaluEval and ContraDoc values are accuracy; ours are balanced accuracy. HHEM-C/S denote claim-/summary-level HHEM-2.1. Dashes indicate unreported or inapplicable entries. Appendix F gives full details.

4.1 Detector Results

We first ask whether the unsupervised severity scores are useful as hallucination detectors. Table 1 compares our method with existing systems using the metrics shared by each benchmark. Complete results for every configuration and evaluation measure are provided in Appendix E.

Performance varies substantially across benchmarks. Our method performs best on AggreFact (.896 AUROC) and, with sentence-level indicators, on RAGTruth (.793 AUROC; .724 Balanced Accuracy). The HaluEval anchor factor reaches .773 AUROC and .808 within-pair accuracy. SummEval is omitted from Table 1 because no detector is directly comparable in the same reference-free setting; its reference-free mean reaches .887 AUROC. Performance is more modest on TofuEval (.676 AUROC), FaithBench (.589), and ContraDoc (.522). The latter is effectively at chance, while FaithBench shows only weak separation. The method is therefore competitive on several IC benchmarks but is not uniformly strong across domains or hallucination constructs. Our factor diagnostics (see below) suggest various limitations: FaithBench's labels reflect harmfulness as well as unsupportedness, TofuEval has little lexical separation between classes, and the CC factor has only moderate

Table 2: Cross-battery discriminant validity. Values are AUROC against IC labels for the unmodified IC general score and oriented CC composite.

Dataset	IC (G)	CC mean	Δ
AggreFact	.896	.319	.577
HaluEval	.673	.320	.353
TofuEval	.663	.459	.204
RAGTruth	.685	.565	.120
FaithBench	.575	.523	.052

reliability on ContraDoc ($\omega_h = .53$). Complete metrics and confidence intervals are reported in Appendix E.

Factor structure diagnostics. The factor loadings reveal which indicators are most strongly represented in the shared dimension. In the original six-indicator RAGTruth solution, BERTScore precision has a loading of .819, compared with .608 for SummaC and .247 for AlignScore. The common factor therefore reflects semantic similarity more strongly than either consistency indicator. Adding worst-sentence indicators improves AUROC from .685 to .793, consistent with document-level aggregation obscuring localized grounding failures. On HaluEval, selecting the alignment-anchored factor improves AUROC from .673 to .773, indicating that the dimension associated with source alignment is more discriminative than the general score. Together, the loading profiles and ablations diagnose a limitation of the common score: indicators can share substantial variation while useful grounding information is better captured separately or at sentence level. This identifies a measurement limitation, without establishing the causes of individual hallucinations. Full loading profiles and scoring ablations are reported in Appendices D and E, respectively.

4.2 Construct Validity

Detection performance alone does not establish that the scores measure a coherent or stable construct. We therefore evaluate reliability, cross-dataset replication, discriminant validity, and agreement with graded human annotations.

4.2.1 Reliability and replication.

The reference-free IC general factor is highly reliable on five of six IC benchmarks. SummEval exposes a distinction between detection and measurement: its oriented mean achieves .887 AUROC, although the general factor has negligible reliability ($\omega_h = .018$). Nevertheless, the six-indicator IC loading profiles replicate across datasets, with mean Tucker congruence of .944. The CC battery does not recover a uniformly reliable general factor, so cross-battery analyses use its oriented equal-weight composite. Complete reliability and loading diagnostics are reported in Appendix D.

4.2.2 Discriminant validity.

If IC and CC measure distinct constructs, output-only coherence (CC) should not be able to predict IC labels as well as source-grounding scores. Table 2 confirms this pattern: the IC score outperforms the CC composite on all five IC-only benchmarks, with AUROC gaps ranging from .052 to .577.

The separation also appears without labels: correlations between the oriented IC and CC mean composites are non-positive on five of the six datasets, indicating that the batteries are not interchangeable measures of generic text quality. The smaller FaithBench gap is consistent with its weaker IC detection results.

4.2.3 Agreement with graded annotations.

After controlling for lexical novelty, severity remains associated with SummEval expert consistency (partial $\rho = -.252$) and RAGTruth hallucinated-span density (partial $\rho = .277$). Evidence is weaker on FaithBench ($\rho = .074$), where non-monotonic category means suggest that the score captures unsupportedness more directly than annotated harmfulness. Exact tests and category-level results are provided in Appendix G. Reference-based extensions on SummEval are reported in Appendix H.

Table 3: Lexical diagnostics using the main-result configurations. Lexical separation is the difference in mean novel-bigram rate between hallucinated and faithful outputs. Stratified AUROC matches examples within novelty quartiles; brackets give 95% bootstrap intervals. Retained signal is $(\mathrm{AUROC}_{\mathrm{strat}} - .5)/(\mathrm{AUROC}_{\mathrm{raw}} - .5)$, and p_{Holm} compares severity with bigram novelty using paired DeLong tests.

Dataset	Lex. sep.	Severity AUROC	Bigram-only AUROC	Stratified AUROC [95% CI]	Retained	$\mathrm{p}_{\mathrm{Holm}}$
AggreFact	.461	.896	.879	.741 [.695, .785]	60.8%	.139
FaithBench	.069	.589	.625	.540 [.500, .585]	45.5%	.212
HaluEval	.106	.773	.682	.732 [.700, .766]	85.0%	$< .001$
RAGTruth	.128	.793	.717	.741 [.688, .778]	82.3%	$< .001$
SummEval	.288	.887	.878	.646 [.573, .848]	37.7%	.591
TofuEval	.060	.676	.607	.647 [.602, .691]	83.6%	.002

4.3 Benchmark Diagnostics

We examine whether detection performance reflects lexical novelty and benchmark composition as well as source faithfulness. ContraDoc is excluded because it has no external source document.

4.3.1 Lexical novelty as a benchmark signal.

The fraction of output bigrams absent from the source is a strong baseline, reaching .607–.879 AUROC across the six source-conditioned benchmarks (Table 3). After Holm correction, our severity significantly outperforms it on HaluEval, RAGTruth, and TofuEval; the baseline does not significantly outperform our severity score on any dataset.

On AggreFact, SummaC-Conv, QAFactEval, DAE, QuestEval, and SummaC-ZS all assign scores associated with bigram novelty ($\rho = .709$–$.820$ after orientation toward severity; Table 1). More novel outputs therefore tend to receive higher hallucination scores across these detectors. Together with the strong lexical baseline, this identifies a shared potential confound. It does not quantify how much of each detector's AUROC depends on novelty: the stratified and incremental-validity analyses below evaluate our scores only.

4.3.2 Cross-benchmark and stratified performance.

Across the six source-conditioned benchmarks, our unadjusted AUROC tends to be higher when hallucinated and faithful outputs differ more in bigram novelty ($\rho = .943$, $p = .005$). This association is descriptive given the small number of benchmarks. Table 3 evaluates the same fixed scores using faithful-hallucinated comparisons within novelty quartiles. Our AUROC falls from .896 to .741 on AggreFact and from .887 to .646 on SummEval, compared with .793 to .741 on RAGTruth. The "retained" column expresses how much of our score's original above-chance discrimination remains under this evaluation. Under this evaluation, our detector retains 82.3-85.0% of its above-chance discrimination on HaluEval, RAGTruth, and TofuEval, compared with 37.7-60.8% on SummEval and AggreFact; FaithBench retains 45.5%. The remaining discrimination shows that our score captures information beyond broad differences in bigram novelty; it does not establish independence from all lexical cues.

4.3.3 Incremental validity beyond lexical novelty.

Adding severity to a logistic model containing bigram novelty improves fit on five of six benchmarks after Holm correction, with cross-validated AUROC gains of .025–.091. FaithBench is the exception ($p_{\mathrm{Holm}} = .102$; $\Delta\mathrm{AUROC} = .002$), providing no detectable increment beyond novelty. Complete likelihood-ratio and cross-validation results are reported in Appendix G.

4.3.4 Aggregation across heterogeneous benchmarks.

AggreFact combines nine sub-benchmarks with hallucination rates from .080 to .922. Our AUROC falls from .896 when pooled to approximately .79 when computed within sub-benchmarks and macro-averaged; the bigram baseline similarly falls from .879 to .778. Pooled performance therefore reflects both within-domain discrimination and differences between constituent benchmarks.

5 Discussion

Diagnosing the evidence behind severity. The loading analysis identifies which indicators are represented in the common score and where an anchor factor or sentence-level indicators improve discrimination (Section 4.1; Appendices D and E). For individual instances, diagnosis would additionally require its indicator values and fitted scoring weights. High semantic similarity combined with low entailment or alignment scores could flag unsupported content despite substantial source reuse. This is a diagnostic use of the measurement structure. However, the present aggregate analyses do not establish which individual hallucinations instances exhibit that pattern or validate their error types. **What benchmarks reward and how to improve them.** The lexical controls reveal different evaluation conditions across benchmarks. Our detector retains most of its above-chance discrimination on HaluEval, RAGTruth, and TofuEval, but substantially less on AggreFact and SummEval. FaithBench provides little evidence of incremental value beyond novelty. These differences motivate reporting lexical-stratified performance and incremental validity for each benchmark, alongside within-subgroup results where datasets pool heterogeneous sources. Benchmark construction should include both faithful paraphrases and hallucinations that closely reuse source wording.

Relationship to LLM judges. Published judge results are stronger on several benchmarks, but differing subsets and protocols prevent controlled comparisons or conclusions about why judges succeed. Our framework offers an explicit account of how measured indicators enter the severity score, supported by loading and reliability diagnostics. This complements task-level judgements and natural-language explanations as the aggregation is inspectable. We do not have matched judge predictions with which to compare instance-level successes and failures. Applying the same lexical controls to judges and specialist detectors would clarify what each captures beyond surface cues.

Post-training and the Goodhart risk. The diagnostics are also relevant when severity becomes a filtering, reranking, or training signal but optimizing it directly creates a Goodhart risk. Loadings expose how indicators enter the score, while lexical controls test whether it contributes information beyond source reuse. A policy could otherwise improve its measured severity by copying more closely without improving factuality. A safer approach would use severity as one objective among several, require candidate components to demonstrate incremental validity beyond lexical novelty, and monitor extractiveness, coverage, and task quality during training. Evaluation should include novelty-matched and lexically camouflaged examples so that aggregate improvements cannot conceal a shift toward copying.

6 Conclusion

We introduce a label-free framework for measuring hallucination severity and examining the evidence behind its scores. Factor loadings reveal which indicators define the shared dimension, reliability diagnostics assess its coherence, and ablations identify grounding information better captured by alternative configurations. These analyses make the framework useful both as a detector and as a means of examining what its severity estimate represents. Importantly, using the framework as a measurement instrument exposes a limitation of current evaluation bnechmarks and practice. Detector performance often tracks how readily faithful and hallucinated outputs can be separated through lexical novelty, with this dependence extending beyond our method to established detectors. Regardless, the severity score retains non-lexical information on most benchmarks, particularly those constructed from LLM-generated hallucinations. The central conclusion is therefore not that surface cues fully explain hallucination detection, but that conventional pooled scores can combine genuine factual sensitivity with benchmark-specific shortcuts in proportions that are rarely measured.

These findings motivate a more diagnostic evaluation protocol. Lexical-stratified performance, incremental validity against surface baselines, and within-subgroup reporting should accompany aggregate leaderboard results. Future work should test measurement invariance under frozen cross-domain scoring, compare specialist detectors and LLM judges on identical examples, and construct lexically controlled counterfactual benchmarks. Our severity score may also support reranking or hallucination-aware post-training, once extractiveness and task quality are monitored explicitly.

AI Use Statement

In accordance with the ICLR 2027 AI disclosure policy, we provide the following statement of generative AI use in this work.

In this work, we used generative AI tools for edits and polishing the language along with typesetting and formatting of latex tables. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs,, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code. We have reviewed all AI-assisted work. We have gone over the formatted latex tables and any text edits or polishes from the LLM. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

References

Forrest Bao, Miaoran Li, Renyi Qu, Ge Luo, Erana Wan, Yujia Tang, Weisi Fan, Manveer Singh Tamber, Suleman Kazi, Vivek Sourabh, et al. Faithbench: A diverse hallucination benchmark for summarization by modern llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 448–461, 2025.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 238–255, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.20. URL https://aclanthology.org/2025.acl-short.20/.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. Beyond the surface: Measuring self-preference in LLM judgments. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1653–1672, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.86. URL https://aclanthology.org/2025.emnlp-main.86/.

Duncan Cramer and Dennis Howitt. The sage dictionary of statistics, 2004. URL https://methods.sagepub.com/dict/edvol/the-sage-dictionary-of-statistics/chpt/principal-axis-factoring.

E R DeLong, D M DeLong, and D L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3): 837–845, 1988. ISSN 0006-341X.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021.

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

David B. Flora. The sage encyclopedia of research design, 2022. URL https://sk.sagepub.com/ency/edvol/the-sage-encyclopedia-of-research-design-2e/chpt/mcdonald-s-omega-hierarchical.

Yoshinari Fujinuma. Contrastive decoding mitigates score range bias in LLM-as-a-judge. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13404–13418, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.657. URL https://aclanthology.org/2026.findings-acl.657/.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 7(6), 2026.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979. URL https://api.semanticscholar.org/CorpusID:122415379.

John L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30: 179–185, 1965. URL https://api.semanticscholar.org/CorpusID:19663974.

Khizar Hussain and Murat Kantarcioglu. Parallax: Separating genuine hallucination detection from benchmark construction artifacts. *arXiv preprint arXiv:2605.17028*, 2026.

J. Edward Jackson. *Oblimin Rotation*. John Wiley Sons, Ltd, 2005. ISBN 9780470011812. doi: https://doi.org/10.1002/0470011815.b2a13060. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/0470011815.b2a13060.

Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Jan Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34716–34733, 2025.

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

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Jierui Li, Vipul Raheja, and Dhruv Kumar. ContraDoc: Understanding self-contradictions in documents with large language models. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6509–6523, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.362. URL https://aclanthology.org/2024.naacl-long.362/.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 6449–6464, 2023.

Shuliang Liu, Zhipeng Xu, Zhenghao Liu, Yukun Yan, Minghe Yu, Yu Gu, Chong Chen, Huiyuan Xie, and Ge Yu. Mitigating judgment preference bias in large language models through group-based polling. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1448–1464, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.71. URL https://aclanthology.org/2026.findings-acl.71/.

Potsawee Manakul, Adian Liusie, and Mark Gales. 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*, pp. 9004–9017, 2023.

Zachary R. McCaw, Jacqueline M. Lane, Richa Saxena, Susan Redline, and Xihong Lin. Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. *Biometrics*, 76(4):1262–1272, 2020. doi: https://doi.org/10.1111/biom.13214. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.13214.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, 2024.

Elisei Rykov, Yana Shishkina, Kseniia Petrushina, Kseniia Titova, Sergey Petrakov, and Alexander Panchenko. Smurfcat at semeval-2024 task 6: Leveraging synthetic data for hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pp. 869–880, 2024.

John Schmid and John M. Leiman. The development of hierarchical factor solutions. *Psychometrika*, 22(1):53–61, 1957. doi: 10.1007/BF02289209. URL https://doi.org/10.1007/BF02289209.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin Rousseau, and Greg Durrett. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11626–11644, 2023.

Liyan Tang, Philippe Laban, and Greg Durrett. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847, 2024a.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, et al. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4455–4480, 2024b.

Zhihua Wen, Zhizhao Liu, Zhiliang Tian, Shilong Pan, Zhen Huang, Dongsheng Li, and Minlie Huang. Scenario-independent uncertainty estimation for llm-based question answering via factor analysis. In *Proceedings of the ACM on Web Conference 2025*, pp. 2378–2390, 2025.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11328–11348, 2023.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. arxiv 2023. *arXiv preprint arXiv:2309.01219*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Marvin Zuckerman and Anton Aluja. Chapter 13 - measures of sensation seeking. In Gregory J. Boyle, Donald H. Saklofske, and Gerald Matthews (eds.), *Measures of Personality and Social Psychological Constructs*, pp. 352–380. Academic Press, San Diego, 2015. ISBN 978-0-12-386915-9. doi: https://doi.org/10.1016/B978-0-12-386915-9.00013-9. URL https://www.sciencedirect.com/science/article/pii/B9780123869159000139.

Table 4: Indicator batteries. "Direction" gives the raw quality direction before orientation: ↑ means larger values indicate better quality; ↓ means larger values indicate worse quality.

Battery	Indicator	Family	Direction	Definition
IC	bartscore_faithfulness	generation probability	↑	Exponentiated mean token log-likelihood of the output conditioned on the source (BARTScore, source-to-output direction).
IC	bert_p	semantic alignment	↑	BERTScore precision between the source and generated output.
IC	bert_r	semantic alignment	↑	BERTScore recall between the source and generated output.
IC	neo_recall	entity coverage	↑	Recall of source entities in the generated output after named-entity extraction and normalization.
IC	summac_score	factual consistency	↑	SummaC-ZS source–output consistency score.
IC	align_score	factual alignment	↑	AlignScore agreement between the source and generated output.
IC extension	align_worst	localized alignment	↑	Minimum AlignScore across output sentence/segment scores.
IC extension	summac_worst	localized consistency	↑	Minimum sentence-level SummaC score.
Reference	rouge2_f	lexical overlap	↑	ROUGE-2 F-score against a human-written reference.
Reference	meteor	lexical/semantic overlap	↑	METEOR against a human-written reference.
Reference	chrf	character overlap	↑	Character n-gram F-score against a human-written reference.
CC	contradiction_density	contradiction	↓	Density of high-probability contradictions between sentences in the generated output.
CC	unieval_coherence	discourse coherence	↑	UniEval coherence score; constant in the reported runs and removed before EFA.
CC	entity_grid_coherence	entity coherence	↑	Continuity of entity transitions across output sentences.
CC	semantic_drift_mean	semantic continuity	↓	Mean adjacent-sentence TF–IDF distance.
CC	semantic_drift_max	semantic continuity	↓	Maximum adjacent-sentence TF–IDF distance.
CC	repetition_ratio	repetition	↓	Proportion of repeated lexical material in the output.
CC	lexical_diversity	diversity	↑	Type-diversity score for the generated output.
CC	self_bleu_score	repetition	↓	Sentence-level self-BLEU, with larger values indicating greater repetition.
CC	pll_fluency	fluency	↑	Exponentiated mean token pseudo-log-likelihood under RoBERTa.

A Indicator Batteries and Estimation Details

Here, we document the complete indicator registry, dataset construction, evaluation protocol, and secondary results. All factor models are estimated separately for each dataset and battery. Hallucination annotations are not used to compute indicators, choose the number of factors, estimate loadings, or orient scores.

A.1 Indicator definitions

Table 4 lists the indicators used in the reported runs. The six-indicator IC battery is reference-free. The three reference-based indicators are used only in the SummEval extension. The CC registry computes nine output-only indicators spanning five families. The UniEval coherence column was constant in the supplied evaluation runs and was therefore removed before factor estimation, leaving eight retained CC indicators. Deterministic complements and derived duplicates in the registry are inactive.

A.2 Preprocessing, fallback, and orientation

Let r_{ij} be the average rank of indicator j for example i among N observations. Each indicator is transformed using

$$z_{ij} = \Phi^{-1}\left(\frac{r_{ij}}{N+1}\right), \tag{7}$$

where Φ^{-1} is the standard-normal quantile function. Average ranks handle ties. The transformation makes scales comparable and limits the influence of skew while preserving order. Before EFA,

Table 5: Dataset inventory. "Cross" marks the IC-only datasets used for the cross-battery discriminant-validity test. The base rate is the proportion labeled hallucinated or contradictory.

Dataset	n	Base rate	IC	CC	Cross	Annotation	Median sentences
AggreFact	1,000	.455	yes	no	yes	binary	1
FaithBench	750	.585	yes	no	yes	binary; ordinal	4
HaluEval	1,000	.500	yes	no	yes	paired binary	4
RAGTruth	1,000	.216	yes	no	yes	binary; span density	5
SummEval	1,000	.059	yes	yes	no	binary; graded	3
TofuEval	600	.443	yes	no	yes	binary	3
ContraDoc	891	.504	no	yes	no	binary	—

all columns are aligned so that larger values indicate better quality; the five higher-is-worse CC indicators in Table 4 are sign-reversed.

Factor estimation is attempted when the effective ratio satisfies $N/p \geq 15$. Constant columns are removed first. If this requirement is not met or estimation fails, the score is the severity-oriented mean

$$S_i^{\text{mean}} = \frac{1}{p} \sum_{j=1}^{p} \widetilde{z}_{ij}, \tag{8}$$

where higher-is-better indicators are negated and higher-is-worse indicators retain their sign. This mean is also evaluated as an ablation when EFA is available.

Parallel analysis compares observed eigenvalues with the 95th percentiles from 1,000 random matrices of equal dimensions and retains at least one factor. Principal-axis factoring with Promax rotation produces the oblique solution; multi-factor solutions are converted to an orthogonal general-plus-group representation by the Schmid–Leiman transformation. For the reported sample sizes, hierarchical omega uses 1,000 row-bootstrap replicates with the fitted general loadings held fixed:

$$\omega_h = \frac{\left(\sum_j g_j\right)^2}{\text{Var}\left(\sum_j z_{ij}\right)}. \tag{9}$$

Values above one caused by sampling and the single-factor approximation are capped at one, as are their percentile interval endpoints.

For the anchor score, a separate minimum-residual EFA with oblimin rotation is fitted using the same number of factors. The selected factor has the largest absolute loading for a prespecified anchor: `align_score` for IC and `contradiction_density` for CC. Labels are not used in this selection. Signs are then fixed so that IC severity increases as `align_score` decreases and CC severity increases with `contradiction_density`. When only one factor is retained, no distinct anchor score is defined.

B Dataset Construction and Scoring Configurations

AggreFact. We use the test partition and sample 1,000 eligible rows with seed 42. The released consistency label is inverted so that one denotes hallucination. AggreFact pools nine component benchmarks; its subgroup heterogeneity is examined separately in Section G.

FaithBench. All 750 examples are retained. The binary target treats the benchmark's *Unwanted* category as positive. The ordered categories Consistent, Benign, Questionable, and Unwanted are retained for the graded analysis.

HaluEval. We retain 500 complete faithful–hallucinated pairs (1,000 outputs). Pair IDs are preserved for within-pair ranking accuracy; the hallucinated member is the positive class.

RAGTruth. We use test examples from question answering and summarization. Data2Text is excluded because its instruction-plus-JSON inputs are not suitable premises for the source-grounding indicators. A document is positive when it contains at least one annotated hallucinated span. Span density is the total annotated hallucinated character length divided by output length.

Table 6: Configuration reported in the main detector table. G is the general score, α_G the anchor score, S_G the general score after adding two worst-sentence indicators, and M_{RF} the oriented reference-free mean.

Dataset	Battery	Reported configuration
AggreFact	IC	G
FaithBench	IC	S_G
HaluEval	IC	α_G
RAGTruth	IC	S_G
SummEval	IC	M_{RF}
TofuEval	IC	S_G
ContraDoc	CC	α_G

SummEval. We use the paired release containing source documents and human references and sample 1,000 rows with seed 42. Expert consistency ratings provide the IC graded target and expert coherence ratings provide the CC target. The binary IC label marks consistency ratings at or below two as positive.

TofuEval and ContraDoc. TofuEval combines MeetingBank and MediaSum and retains 600 document-level examples. ContraDoc contains 449 contradictory and 442 clean documents. It has no external source document, so only the CC battery is defined.

The five IC-only datasets support the cross-battery negative-control test. SummEval is omitted because it natively annotates both constructs; ContraDoc is omitted because IC indicators cannot be computed without a source.

B.1 Reported configuration and sentence-level coverage

The sentence-level extension adds the minimum sentence-level AlignScore and SummaC scores while retaining their document-level versions. It was run on FaithBench, RAGTruth, and TofuEval. AggreFact has predominantly one-sentence outputs, making minimum pooling effectively identical to document pooling; SummEval outputs are also short. HaluEval meets the four-sentence median heuristic but was not run under this configuration because of the available computational budget. TofuEval has a three-sentence median but is retained because that completed run was part of the supplied experiment set. Thus, the sentence-level comparison is an incomplete ablation, not a claim that one length rule was exhaustively applied to all datasets. Indicator construction and factor/anchor eligibility are label-free; Table 12 reports every completed scores so the effect of the highlighted configuration is visible.

C Evaluation Protocol

Discrimination and uncertainty. AUROC and AUPRC are computed directly from continuous severity scores. Percentile 95% confidence intervals use 1,000 row-bootstrap resamples with seed 42. AUPRC is reported together with the positive-class base rate because its chance reference depends on prevalence.

Thresholded metrics. Five-fold stratified cross-validation uses shuffle seed 42. Within each fold, the threshold maximizing Youden's $J = \mathrm{TPR} - \mathrm{FPR}$ is selected on the training portion and applied to the held-out portion. We average balanced accuracy,macro-F1, and MCC across the five held-out folds. The secondary rate-matched threshold is the $(1 - \pi)$ severity quantile, where π is label prevalence, so its predicted-positive rate matches the observed rate.

Calibration and selective prediction. Platt scaling is fitted on each training fold using logistic regression and applied out of fold. Brier score and ten-bin expected calibration error (ECE) are computed after pooling held-out probabilities. For selective prediction, confidence is the absolute distance from the median severity threshold; risk is classification error among retained examples as the least-confident cases are removed. AURC is the trapezoidal area under this risk–coverage curve.

Construct-specific outcomes. HaluEval pairwise accuracy is the proportion of complete pairs in which the hallucinated output receives higher severity, with ties receiving half credit. RAGTruth

Table 7: Factorability and reliability. Bartlett tests reject sphericity for every row ($p < .001$). Confidence intervals for ω_h use 1,000 bootstrap resamples. Values above one are capped at one.

Battery	Dataset/config.	p	k	KMO	Bartlett χ^2	$\omega_\Gamma 30568$ [95% CI]
IC	AggreFact G	6	1	.780	3088.6	1.000 [.970, 1.000]
IC	FaithBench G	6	2	.663	2084.4	1.000 [.915, 1.000]
IC	HaluEval G	6	2	.751	2041.3	1.000 [.922, 1.000]
IC	RAGTruth G	6	2	.650	1694.6	.945 [.877, 1.000]
IC	SummEval RF G	6	2	.613	1715.5	.018 [.017, .020]
IC	TofuEval G	6	2	.709	1216.1	.973 [.866, 1.000]
IC	FaithBench S_G	8	3	.614	—	1.000 [.911, 1.000]
IC	RAGTruth S_G	8	3	.571	—	.985 [.908, 1.000]
IC	TofuEval S_G	8	2	.666	—	1.000 [.892, 1.000]
CC	AggreFact	8	2	.816	6143.9	1.000 [1.000, 1.000]
CC	FaithBench	8	2	.669	2176.5	.639 [.570, .727]
CC	HaluEval	8	2	.678	3100.3	.066 [.060, .074]
CC	RAGTruth	8	2	.673	3531.9	.908 [.823, 1.000]
CC	SummEval	8	3	.712	3003.5	.076 [.068, .085]
CC	TofuEval	8	2	.722	2095.8	.003 [.003, .003]
CC	ContraDoc	8	2	.709	3299.2	.530 [.489, .576]

Table 8: Schmid–Leiman general-factor loadings (top) and communalities (bottom) for the six common reference-free IC indicators. BART denotes BARTScore faithfulness; BP/BR denote BERTScore precision/recall; NEO denotes entity recall; Sum denotes SummaC; Align denotes AlignScore. The mixed signs on SummEval accompany its negligible reliability.

Dataset	BART	BP	BR	NEO	Sum	Align
General-factor loadings						
AggreFact	.746	.889	.713	.774	.773	.706
FaithBench	.715	.837	.812	.726	.645	.467
HaluEval	.816	.843	.625	.623	.613	.603
RAGTruth	.639	.819	.751	.551	.608	.247
SummEval	-.251	.338	.673	.300	-.022	-.538
TofuEval	.747	.803	.670	.438	.597	.703
Communalities after Schmid–Leiman transformation						
AggreFact	.557	.791	.509	.599	.597	.498
FaithBench	.739	.883	.887	.788	.487	.315
HaluEval	.811	.864	.642	.625	.625	.634
RAGTruth	.726	.955	.865	.569	.537	.342
SummEval	.148	.427	.751	.205	.199	.465
TofuEval	.773	.860	.707	.450	.644	.691

uses Spearman correlation with hallucinated-span density. SummEval uses Spearman correlation with expert consistency, for which the expected sign is negative because larger ratings denote better summaries. FaithBench uses ordered class means and an ordinal Spearman association.

Construct validity. Tucker congruence between loading profiles $\mathbf{a}$ and $\mathbf{b}$ is $\phi = (\mathbf{a}^\top \mathbf{b})/(\|\mathbf{a}\|\|\mathbf{b}\|)$ after signs are aligned using the prespecified IC anchor. Discriminant validity compares the unmodified six-indicator IC general score with the oriented CC equal-weight mean on the five IC-only datasets. Partial Spearman correlations are computed as correlations between rank-residuals after separately regressing severity and the graded target on bigram novelty.

D Factor Diagnostics and Construct Validity

The reference-free IC general factor is reliable on five of the six IC benchmarks. SummEval is the exception: its near-zero ω_h and mixed general loadings make the general score uninterpretable, motivating the oriented reference-free mean used in the main results. The CC results are more heterogeneous and justify using the oriented CC mean for the cross-battery test rather than treating every CC general factor as stable.

In the separate sign-aligned loading-profile replication analysis over these six common indicators, the 15 pairwise Tucker coefficients have mean .944 and range .882–.983. All exceed .85 and seven exceed .95. This supports broad profile similarity but not strict measurement invariance: indicator importance still changes across datasets, and the SummEval general-factor failure remains material.

Table 9: Cross-battery discriminant validity. AUROC is against IC labels. IC uses the unmodified six-indicator general score; CC uses its oriented equal-weight mean. The final column is the correlation between the oriented IC and CC mean composites.

Dataset	IC	CC	Δ	$\rho_{\mathrm{IC,CC}}$
AggreFact	.896	.319	.577	-.550
HaluEval	.673	.320	.353	-.190
TofuEval	.663	.459	.204	-.269
RAGTruth	.685	.565	.120	-.195
FaithBench	.575	.523	.052	.071

Table 10: Association with graded annotations before and after controlling for bigram novelty. SummEval ratings increase with quality, hence the negative sign.

Dataset/target	Raw ρ	Partial ρ	p
SummEval consistency	-.403	-.252	5.5×10^{-16}
RAGTruth span density	.427	.277	5.2×10^{-19}
FaithBench ordinal class	.173	.074	.042

SummEval, which has both batteries but is not part of the IC-only negative control, has an IC–CC mean correlation of $-.063$. Thus five of six available mean-composite correlations are non-positive. The small FaithBench detection gap and weak positive correlation agree with its limited IC discrimination.

FaithBench's S_G class means are non-monotonic: Consistent $-.257$ (95% CI $[-.368, -.145]$), Benign $.114$ ($[-.035, .253]$), Questionable $-.052$ ($[-.192, .081]$), and Unwanted $.095$ ($[.040, .152]$). This pattern supports the narrower interpretation that the instrument targets unsupportedness more directly than the benchmark's harmfulness ordering.

E Complete Detector Results and Ablations

HaluEval's anchor score ranks the hallucinated output above its faithful pair in .808 of 500 pairs (95% exact binomial interval [.771,.842]). RAGTruth S_G correlates .427 with span density (95% bootstrap interval [.378,.477]). These outcomes evaluate ordering against richer annotations and do not require threshold selection.

Sentence-level indicators yield their largest general-score gain on RAGTruth ($+.108$ AUROC), followed by FaithBench ($+.014$) and TofuEval ($+.013$). The anchor factor yields the largest change on HaluEval ($+.100$) and a smaller gain on the base RAGTruth battery ($+.051$). It is not uniformly superior: the general and anchor scores are alternative structural understanding. SummEval shows why reliability is considered before discrimination: its ill-conditioned general factor is near chance, while the reference-free oriented mean is strongly discriminative.

F Published-Comparison Provenance

The main comparison table combines published aggregates only where the relevant benchmark and metric are available. These values are contextual unless rows, labels, thresholds, and averaging conventions coincide.

G Benchmark Statistical Tests

Lexical baseline. For an output containing at least one bigram, lexical novelty is the fraction of output bigrams absent from its source; outputs without a bigram receive the implementation's zero-valued boundary case. Lexical separation is the difference in mean novelty between positive and negative examples. ContraDoc is excluded because it has no external source.

Paired DeLong tests compare the AUROCs of severity and lexical novelty on the same examples. The test is analytic and seed-independent. Holm correction is applied across the six source-

Table 11: Complete results for the configurations reported in the main text. Thresholded and calibrated metrics use five-fold out-of-fold predictions. SummEval's oriented mean was evaluated only with threshold-free metrics in the supplied export.

Dataset	Battery	Config.	Base	AUROC [95% CI]	AUPRC [95% CI]	BalAcc	F1	MCC	Brier	ECE	AURC
AggreFact	IC	G	.455	.896 [.878,.915]	.880 [.853,.906]	.831	.815	.663	.125	.036	.081
FaithBench	IC	S_G	.585	.589 [.549,.626]	.647 [.600,.701]	.545	.554	.091	.236	.027	.386
HaluEval	IC	α_G	.500	.773 [.744,.803]	.771 [.732,.809]	.697	.702	.395	.194	.035	.178
RAGTruth	IC	S_G	.216	.793 [.762,.825]	.502 [.436,.570]	.724	.530	.381	.137	.025	.262
SummEval	IC	M_{RF}	.059	.887 [.840,.926]	.385 [.274,.529]	—	—	—	—	—	—
TofuEval	IC	S_G	.443	.676 [.629,.720]	.670 [.615,.724]	.631	.579	.266	.222	.046	.290
ContraDoc	CC	α_G	.504	.522 [.484,.562]	.528 [.482,.582]	.511	.516	.023	.250	.004	.457

Table 12: Completed score ablations. Entries are AUROC/AUPRC. M is the oriented equal-weight RINT mean; S_α is the anchor score from the eight-indicator sentence-level battery. A dash denotes an unrun or undefined configuration, not a failed estimate.

Dataset	$\mathbf{G}$	$\mathbf{M}$	$\boldsymbol{\alpha}_{\Gamma}$30535	$\mathbf{S_G}$	$\mathbf{S_{ff}}$
AggreFact	.896/.880	.897/.881	—	—	—
FaithBench	.575/.642	.576/.642	.571/.623	.589/.647	.561/.650
HaluEval	.673/.648	.676/.655	.773/.771	—	—
RAGTruth	.685/.350	.716/.390	.736/.416	.793/.502	.797/.492
SummEval	.487/.058	.887/.385	—	—	—
TofuEval	.663/.662	.656/.656	—	.676/.670	.672/.652
ContraDoc	.488/.509	.478/.516	.522/.528	—	—

conditioned datasets. Table 14 shows that severity significantly exceeds novelty on HaluEval, RAGTruth, and TofuEval; the lexical baseline never significantly exceeds severity.

Lexical stratification. Novelty quartiles are formed within each dataset and faithful–hallucinated comparisons are restricted to examples in the same quartile. The matched AUROC is aggregated over the four strata. Its percentile interval uses 2,000 bootstrap resamples with seed 42. Retained signal is

$$100 \times \frac{\mathrm{AUROC}_{\mathrm{strat}} - .5}{\mathrm{AUROC}_{\mathrm{raw}} - .5}. \tag{10}$$

The main diagnostic table reports every point estimate and interval. This aggregate controls broad novelty level; it is not a claim that performance is uniform within each individual quartile.

Across the six datasets, lexical separation and detector AUROC have Spearman $\rho = .943$ ($p = .005$). Because this analysis contains only six aggregate points, it is descriptive. On AggreFact, per-example correlations with bigram novelty are .820 for SummaC-Conv, .796 for our score, .725 for QAFactEval, .721 for DAE, .717 for QuestEval, and .709 for SummaC-ZS after orienting every detector toward greater hallucination severity.

Incremental validity. For each dataset, a logistic model containing only bigram novelty is nested within a model that additionally contains severity. The likelihood-ratio test has one degree of freedom; Holm correction is again applied across six tests. The descriptive in-sample AUROC increment is supplemented by a five-fold out-of-sample increment using seed 42. Table 14 shows that all datasets except FaithBench exhibit significant incremental fit.

AggreFact subgroup aggregation. The nine component benchmarks have positive rates from .080 to .922. Our pooled AUROC is .896, whereas the approximate macro-average of AUROCs computed within components is .79. The bigram baseline similarly falls from .879 pooled to .778 macro-within-component. Pooled performance therefore combines within-component discrimination with differences in lexical and label distributions between components.

H SummEval Reference-Based Extension

SummEval permits a controlled extension because every example contains both a source document and a human reference. Table 15 uses identical rows across batteries. For comparability, the oriented

Table 13: Scope of the published comparisons in the main table.

Dataset	Our evaluation	Published comparison	Consequence
AggreFact	Seed-42 sample of 1,000 test rows. Out-of-fold thresholding.	Detector baselines use the 4,811-row test cut. Balanced-accuracy systems use another subset.	Displayed aggregates are not paired, although per-example baseline scores are available.
FaithBench	Unwanted versus all other categories. Document-level S_G.	Claim- and summary-level protocols differ. HHEM-C/S denote these levels.	Label unit and aggregation differ; gaps are descriptive.
HaluEval	500 intact pairs. Our metric is balanced accuracy.	Published results use accuracy on the balanced set.	Accuracy and balanced accuracy are comparable in expectation; thresholds and prompts differ.
RAGTruth	Question-answering and summarization test examples.	Published detector and judge results use summarization only.	Task mixture and prevalence differ; results are not paired.
TofuEval	MeetingBank and MediaSum.	Published results use MeetingBank only.	The gap may reflect subset composition.
ContraDoc	Balanced accuracy from CC severity.	Published judge results use accuracy, often in each judge's best setting.	Accuracy and balanced accuracy are different metrics.

Table 14: Benchmark statistical tests using exactly the configurations reported in the main results. Panel (a) gives paired DeLong tests against lexical novelty; Δ is severity AUROC minus lexical-novelty AUROC. Panel (b) tests incremental validity beyond novelty; ΔAUROC(CV) is the primary five-fold out-of-sample estimate.

(a) Paired DeLong tests against bigram novelty.

Dataset	Config.	Severity	Novelty	Δ	z	p_{raw}	p_{Holm}	Sig.
AggreFact	G	.8961	.8793	.0167	1.993	.0463	.1388	no
RAGTruth	S_G	.7931	.7167	.0764	5.020	5.16×10^{-7}	2.58×10^{-6}	yes
SummEval	M_{RF}	.8873	.8780	.0092	.537	.5914	.5914	no
HaluEval	α_G	.7733	.6817	.0916	6.001	1.96×10^{-9}	1.17×10^{-8}	yes
TofuEval	S_G	.6762	.6070	.0692	3.477	5.06×10^{-4}	.0020	yes
FaithBench	S_G	.5886	.6245	-.0360	-1.617	.1058	.2116	no

(b) Incremental validity after including bigram novelty as a covariate.

Dataset	Config.	$\chi^2_\Gamma 30257$	p_{raw}	p_{Holm}	ΔAUROC(in)	ΔAUROC(CV)	Seed	Sig.
AggreFact	G	102.57	4.17×10^{-24}	2.09×10^{-23}	.0245	.0245	42	yes
RAGTruth	S_G	88.66	4.69×10^{-21}	1.88×10^{-20}	.0762	.0754	42	yes
SummEval	M_{RF}	22.47	2.14×10^{-6}	4.28×10^{-6}	.0259	.0245	42	yes
HaluEval	α_G	149.79	1.93×10^{-34}	1.16×10^{-33}	.0919	.0912	42	yes
TofuEval	S_G	35.70	2.30×10^{-9}	6.91×10^{-9}	.0691	.0642	42	yes
FaithBench	S_G	2.67	.1024	.1024	.0038	.0022	42	no

mean is the primary score; the factor diagnostics explain why the reference-free general factor is not used.

Table 15: SummEval reference ablation. RF uses the six source-aware reference-free indicators; Ref uses ROUGE-2, METEOR, and chrF; Combined uses all nine. AUROC/AUPRC are for the oriented mean.

Battery	p	k	$\omega_\Gamma 30568$	AUROC	AUPRC
RF	6	2	.018	.887	.385
Reference only	3	1	.946	.702	.140
Combined	9	3	1.000	.862	.335

Against expert consistency ratings, the oriented means have Spearman correlations of $-.403$ (RF), $-.158$ (reference only), and $-.353$ (combined). On this benchmark, adding reference-overlap indicators does not improve the reference-free mean. This is a single-dataset ablation and does not imply that references are generally unhelpful; it shows that the framework can incorporate them while retaining explicit diagnostics of the resulting factor structure.

I Implementation and Reproducibility

All indicators are computed locally and cached before factor estimation; no API calls are made. The reported runs use seed 42 for dataset sampling, bootstrap resampling, and cross-validation. Parallel analysis and omega intervals each use 1,000 simulations/resamples at the reported sample sizes; lexical-stratified intervals use 2,000 resamples.

The archived run logs identify the following checkpoints: BARTScore uses `facebook/bart-large-cnn`; BERTScore uses `distilbert-base-uncased`; the NLI backend is `cross-encoder/nli-deberta-v3-small`; SummaC-ZS uses `tals/albert-xlarge-vitaminc-mnli`; AlignScore uses the large checkpoint with a RoBERTa-large backbone; and PLL fluency uses `roberta-base`. Entity extraction uses `en_core_web_sm`. The container specification is based on PyTorch 2.4.0 with CUDA 12.1 and cuDNN 9, Transformers 4.40–4.x, scikit-learn ≥ 1.4, pandas ≥ 2.0, NumPy ≥ 1.26, SciPy ≥ 1.11, spaCy ≥ 3.7, and `factor-analyzer` $\geq 0.4.1$. Exact dataset-specific configurations are given in Table 6; the statistical diagnostics use those same configurations rather than recomputing them with a different score.