Proceedings of Machine Learning Research **318** (2026)
The 39th Canadian Conference on Artificial Intelligence
Canadian AI 2026
25–29 May 2026, Vancouver, BC

A Meta-Analysis of Evaluation Framework Reliability and Cross-Domain Generalization

Asif Ahmed Neloy[†,*], MD Nazmul Islam[‡]
[†] Faculty of Land and Food Systems, University of British Columbia, Vancouver, BC, Canada
[‡] Computer Systems Technology, University Studies, Keyano College, Fort McMurray, AB, Canada

Abstract

We conduct the first systematic meta-analysis comparing 20 Retrieval-Augmented Generation (RAG) evaluation frameworks, spanning traditional metrics and interpretability methods, from 2020 through 2026, using identical samples across three knowledge domains. Applying all twenty frameworks to 200 question-context-answer triples from RAGBench, we obtain Cochran's $Q = 10{,}055.63$ ($p < 0.001$) with $I^2 = 99.81\%$, indicating that the large majority of score variance reflects true differences between frameworks rather than sampling noise. Pairwise Pearson correlations range from $r = -0.28$ to $r = 0.90$ (median $r = 0.21$), and three distinct clusters emerge: LLM-as-judge methods (within-cluster $\bar{r} = 0.55$), a mixed-methods group ($\bar{r} = 0.63$), and an outlier cluster containing BERTScore, GaRAGe, HALT-RAG, QAFactEval, and RAGChecker ($\bar{r} = 0.01$). Cluster assignments are perfectly stable across four normalization schemes (ARI = 1.0), and bootstrap resampling confirms co-assignment probabilities of at least 0.75 within the LLM-as-judge and mixed-methods clusters, with the outlier cluster ranging from 0.46 to 0.97. A cross-cluster consensus protocol labels 92.5% of samples as contested, with only 6% receiving unanimous faithful verdicts. These results demonstrate that current evaluation frameworks do not measure a unified construct, and we provide empirically grounded selection guidelines for future research.

Keywords: Retrieval-Augmented Generation, Evaluation Metrics, Meta-Analysis, Hallucination Detection, Cross-Domain Generalization

1. Introduction

Retrieval-Augmented Generation (RAG) grounds large language model outputs in retrieved evidence, and is widely used in medical question answering [1], financial analysis [2], and open-domain knowledge retrieval [3]. In these settings, a core requirement is *faithfulness*: the answer should be supported by the retrieved context rather than by potentially stale or incorrect parametric knowledge. Evaluating faithfulness has therefore become a central part of RAG development. At this point, more than twenty evaluation frameworks exist, and they operationalize faithfulness through different mechanisms, including LLM-as-judge prompting [4], natural language inference [5, 6], embedding similarity [7], and attention-based interpretability [8, 9].

The speed of this ecosystem's growth has outpaced empirical validation of a basic assumption many papers and deployments implicitly rely on: that the choice of evaluation framework does not substantially change conclusions. In practice, researchers compare RAG systems whose faithfulness is measured with different tools, and practitioners often adopt a single framework without evidence on how sensitive the results are to that choice. If frameworks measured the same latent construct, scores would align closely across methods. However, a claim-level entailment check and an attention-based context-usage probe are designed to detect different failure modes, and can plausibly disagree on the same generation. Existing surveys organize RAG evaluation tools by methodology [10], but they do not test whether these tools agree when run on identical inputs.

* asif.neloy@ubc.ca

Three gaps motivate this work. First, there is no comprehensive study that computes pairwise agreement across a large set of RAG evaluation frameworks applied to the same samples, so it remains unclear whether framework selection can reverse or materially change system rankings. Second, RAG is used across domains, including general knowledge, finance, and biomedicine, but cross-domain generalization of evaluation frameworks has not been tested under matched experimental conditions. Third, the RAG evaluation literature rarely reports standard meta-analytic heterogeneity statistics, such as Cochran's Q [11] and I^2 [12], which makes the magnitude of inter-framework disagreement hard to quantify and hard to compare to reliability norms in other empirical fields.

We present the first comprehensive cross-framework meta-analysis of RAG evaluation reliability. We apply twenty frameworks spanning six years of development, from BERTScore [7] to SIRG [13], to the same 200 question-context-answer triples drawn from RAGBench [14] and stratified across three domains (General Knowledge, Finance, and Biomedicine). Our contributions are:

(1) A controlled, end-to-end comparison of twenty RAG faithfulness evaluation frameworks on identical inputs, including a full pairwise reliability analysis across the evaluation ecosystem.

(2) A meta-analytic quantification of inter-framework disagreement, with Cochran's $Q = 10{,}055.63$ ($p < 0.001$) and $I^2 = 99.81\%$, indicating substantial heterogeneity in framework outputs that far exceeds conventional thresholds.

(3) An empirical clustering of frameworks into three groups: an LLM-as-judge cluster ($\bar{r} = 0.55$), a large mixed-methods cluster ($\bar{r} = 0.63$), and an outlier cluster ($\bar{r} = 0.01$), with between-cluster agreement of only $\bar{r} = 0.10$, confirmed by perfect cluster stability across normalization schemes (ARI = 1.0) and bootstrap co-assignment probabilities of at least 0.75 within the LLM-as-judge and mixed-methods clusters.

(4) A domain-by-framework analysis showing that interpretability and token-level methods exhibit the strongest domain sensitivity (e.g., KG-RAG $\Delta = 0.267$, MetaRAG $\Delta = 0.242$), while LLM-as-judge methods remain comparatively stable (e.g., FaithJudge $\Delta = 0.030$, RAGAS $\Delta = 0.097$), with implications for evaluation choices in specialized applications.

The remainder of the paper is organized as follows.[1] Section 2 reviews prior work on RAG evaluation and metric comparison. Section 3 formalizes the problem and groups methods into families. Section 4 describes the experimental design, framework configuration, and statistical analysis. Section 5 reports heterogeneity, correlations, clustering, domain effects, human validation, and cross-cluster consensus; Appendix A provides summary statistics, binary agreement, per-sample disagreement, and robustness analyses. Section 6 interprets findings, explains disagreement, and offers empirically grounded selection guidelines. Section 7 concludes and outlines future directions.

2. Background and Related Work

This paper studies whether automated evaluation frameworks for RAG measure the same underlying notion of faithfulness and whether their behavior is stable across domains. Prior work spans (i) the development of RAG evaluation frameworks, (ii) small-scale metric comparison studies, and (iii) survey-style taxonomies that describe the space without quantifying cross-framework reliability. We use the surveys of Huang, Yu, Ma, Zhong, Feng, Wang, Chen, Peng, Feng, Qin, and Liu [10] as an organizing reference, but focus on the empirical question they leave open: do these tools agree when applied to the same inputs?

[1] Code: https://github.com/aaneloy/A-Meta-Analysis-of-Evaluation-Framework-Reliability-and-Cross-Domain-Generalization.

2.1. RAG Evaluation Frameworks

Existing RAG evaluation methods can be grouped into four methodological families. *LLM-as-judge* approaches prompt a language model to score faithfulness, groundedness, or answer quality directly [4, 15–17]. *NLI-based* approaches decompose a generation into atomic claims and test entailment against retrieved evidence [5, 6]. *Embedding and QA-based* approaches measure semantic overlap or question answering consistency without explicit claim extraction [7, 18, 19]. *Interpretability* approaches aim to measure context reliance by probing model internals, including attention patterns, relevance propagation, or hidden-state signals [8, 9, 13, 20, 21]. Additional methods such as GaRAGe [22] rely on heuristic word-overlap signals rather than learned models. Section 3 formalizes these families and the mathematical objects they attempt to estimate.

2.2. Metric Comparison Studies

Empirical comparisons exist, but they cover only small slices of the evaluation ecosystem. Fabbri, Wu, Liu, and Xiong [18] report moderate correlation ($r = 0.52$) between QAFactEval and BERTScore on summarization benchmarks, but do not evaluate RAG-specific faithfulness tools. Es, James, Espinosa-Anke, and Schockaert [4] validate RAGAS against human judgments, but do not situate it relative to other automated frameworks. Friel, Belyi, and Sanyal [14] introduce TRACe and compare it with RAGAS and RAGChecker on RAGBench, providing a valuable controlled study over a small set of methods. Saad-Falcon, Khattab, Potts, and Zaharia [23] compare ARES to RAGAS, focusing on classifier and LLM agreement at limited scale. Collectively, these studies establish that individual tools can correlate with human judgments, but they do not answer whether framework choice is interchangeable, because they do not compute comprehensive pairwise agreement across the broader set of available methods.

2.3. Meta-Analysis for Measurement Reliability in NLP Evaluation

In medicine and psychology, meta-analysis is routinely used to quantify whether independently developed instruments converge on the same construct. Cochran's Q test [11] assesses whether observed variation exceeds sampling error, and the I^2 statistic [12] estimates the fraction of variance attributable to true between-instrument differences. Despite the central role of automated metrics in modern NLP evaluation, formal heterogeneity analysis remains rare, and, to our knowledge, it has not been used to quantify disagreement across RAG evaluation frameworks.

In summary, prior work has produced many RAG evaluation frameworks and several careful comparisons of small subsets, but lacks a comprehensive reliability map of how these tools relate to each other under matched conditions. Our study fills this gap by evaluating a broad set of frameworks on identical samples, quantifying pairwise agreement, and applying heterogeneity statistics to characterize the magnitude and structure of inter-framework disagreement.

3. Evaluation Framework Taxonomy

We formalize the evaluation problem and describe each methodological family. Consider a RAG system that, given query q, retrieves context $X = \{x_1, \ldots, x_k\}$ and generates answer $a = f(q, X; \theta)$. We adopt the following working definition: an answer a is *faithful* to context X if every claim in a is entailed by X, and no claim in a contradicts X. This definition is intentionally narrow: it covers semantic support and consistency with the retrieved evidence, but does not address completeness, fluency, or relevance to the query. We choose this scope

because faithfulness (also called groundedness or factual consistency) is the property that all twenty frameworks claim to measure, making it the natural axis for a reliability comparison.

An evaluation function $\phi\colon \mathcal{A} \times \mathcal{X} \to [0,1]$ maps answer-context pairs to faithfulness scores. Our central question is whether two such functions ϕ_i and ϕ_j agree: what is $\mathrm{Corr}(\phi_i(a,x), \phi_j(a,x))$ across a distribution of samples? The twenty frameworks instantiate ϕ through four mechanisms, each reflecting different assumptions about how to operationalize the definition above.

3.1. **Judgment-Based Methods**

These approaches use a language model as a proxy evaluator. The canonical formulation, exemplified by RAGAS [4], decomposes the answer into a set of claims $C(a)$ and checks each against the context:

$$\phi_{\mathrm{RAGAS}}(a,x) = \frac{|\{c \in C(a) : \mathrm{LLM}(c,x) = 1\}|}{|C(a)|} \tag{3.1}$$

where the LLM returns 1 if context x supports claim c. Variants within this family differ in how they elicit the judgment. G-Eval [15] replaces binary verification with a probability-weighted Likert score derived from chain-of-thought token probabilities. DeepEval [16] inverts the formulation, scoring faithfulness as one minus the fraction of statements the LLM identifies as contradicted. FaithJudge [17] extends the family with few-shot calibration using examples from source documents. All four methods share a reliance on the judge LLM's internal reasoning, making them susceptible to the same biases, a property that, as we show in Section 5, produces high within-family correlation.

3.2. **NLI-Based Methods**

Rather than delegating judgment to an LLM, these frameworks reduce faithfulness to textual entailment. RAGChecker [5] verifies each atomic claim against each context chunk using a dedicated NLI model:

$$\phi_{\mathrm{RC}}(a,x) = \frac{1}{|C(a)|} \sum_{c \in C(a)} \max_{x_j \in X} P_{\mathrm{NLI}}(\mathrm{E} \mid x_j, c) \tag{3.2}$$

where $P_{\mathrm{NLI}}(\mathrm{E} \mid x_j, c)$ denotes the entailment probability. TRUE [6] applies the same principle at sentence granularity with a hard threshold ($\tau = 0.5$), while HALT-RAG [24] ensembles multiple NLI models through learned calibration weights $\phi_{\mathrm{HALT}}(a,x) = \sum_m w_m f_m(a,x)$, optimized against gold labels. The key distinction from judgment-based methods is determinism: NLI models produce consistent scores for identical inputs, whereas LLM judges exhibit stochastic variation.

3.3. **Embedding and QA-Based Methods**

These approaches bypass explicit reasoning entirely. BERTScore [7] computes a greedy token-level alignment between answer and context embeddings:

$$P_{\mathrm{BERT}} = \frac{1}{n} \sum_{i=1}^{n} \max_{j=1}^{m} \frac{\mathbf{h}_{a_i}^{\top} \mathbf{h}_{x_j}}{\|\mathbf{h}_{a_i}\| \cdot \|\mathbf{h}_{x_j}\|} \tag{3.3}$$

where $\mathbf{h}_{a_i}$ and $\mathbf{h}_{x_j}$ are contextual embeddings of answer and context tokens. For faithfulness evaluation, the precision variant is relevant: it measures whether each answer token finds a semantic match in the context. UniEval [19] extends this idea to a learned multi-dimensional scorer. QAFactEval [18] instead generates questions from the answer and checks whether the context yields matching answers, measuring verifiability rather than semantic overlap.

3.4. Interpretability Methods

A more recent line of work measures faithfulness through model internals rather than output comparison, assessing *process* rather than product. ReDeEP [8] quantifies the fraction of attention mass that output tokens place on context positions, averaged across layers. LRP4RAG [9] applies layer-wise relevance propagation to back-propagate output relevance to input tokens, scoring faithfulness as the share attributed to context. LUMINA [20] trains a linear probe on hidden states to distinguish context-sourced from parametric generations. SIRG [13] extracts entity-relation triples from both answer and context and computes their overlap: $\phi_{\text{SIRG}}(a, x) = |\mathcal{E}(a) \cap \mathcal{E}(x)|/|\mathcal{E}(a)|$. HSAD [21] applies spectral analysis to hidden-state dynamics. These methods share a common assumption (that internal computation patterns reveal whether the model relied on context) but differ in which signal they extract, leading to moderate within-family correlation as we document below.

These four families are not entirely disjoint: LLM judges implicitly perform entailment-like reasoning, and embedding similarity is related to soft entailment. Their failure modes, however, are asymmetric. An LLM judge may be misled by fluent but unsupported text; an NLI model may miss implications that require world knowledge beyond the entailment training distribution; an interpretability method may flag high context attention even when the generated output diverges from what the context states. These complementary failure modes suggest that inter-family correlations will be substantially lower than intra-family correlations, a hypothesis we test in Section 5.

4. Methodology

This section describes our experimental design: data and framework selection, the evaluation protocol, and the statistical analysis procedures.

4.1. Data and Framework Selection

We draw 200 question-context-answer triples from RAGBench [14], stratified across three domains: General Knowledge (67 samples from HotpotQA, MS MARCO, and HAGRID), Finance (67 samples from FinQA and DelucionQA), and Biomedicine (66 samples from CovidQA and PubMedQA). This design allows us to isolate domain effects from overall framework behavior. We select twenty frameworks spanning four methodological categories and six years of development (2020–2026). Inclusion requires (1) a publicly available implementation, (2) applicability to RAG faithfulness scoring, and (3) publication in a peer-reviewed venue. The twenty frameworks, grouped by category, are: *LLM-as-judge*: RAGAS [4], G-Eval [15], DeepEval [16], FaithJudge [17]; *NLI-based*: RAGChecker [5], HALT-RAG [24]; *Embedding and QA-based*: BERTScore [7], UniEval [19], QAFactEval [18]; *Token-level and classifier*: TRACe [14], ARES [23], LettuceDetect [25]; *Interpretability*: ReDeEP [8], LRP4RAG [9], LUMINA [20], HSAD [21], SIRG [13]; and *Other*: GaRAGe [22], MetaRAG [26], KG-RAG [27].

4.2. Evaluation Protocol

Each framework is applied to all 200 samples using default configurations from its official implementation. LLM-as-judge methods use DeepSeek-V3 (`deepseek-chat`) at temperature 0, with Groq (Llama-3.3-70B) and Gemini-2.5-Flash as automatic fallbacks in the event of rate limiting; NLI methods use DeBERTa-v3-large fine-tuned on MNLI. Categorical outputs are mapped to numeric values following each framework's documentation, and frameworks with multiple sub-scores are represented by their overall faithfulness score. All scores are

normalized to $[0, 1]$ via min-max scaling, $\tilde{s}_{ij} = (s_{ij} - \min_i s_{ij})/(\max_i s_{ij} - \min_i s_{ij})$, preserving within-framework rank orderings while enabling cross-framework comparison. The resulting score matrix $\mathbf{S} \in \mathbb{R}^{200\times 20}$ is the basis for all subsequent analyses.

4.3. Statistical Analysis

Pairwise correlation: For each framework pair (j, k), we compute both Pearson r_{jk} and Spearman ρ_{jk} correlations, with 95% confidence intervals obtained from 1,000 bootstrap resamples.

Meta-analytic heterogeneity: We test whether all frameworks measure the same underlying quantity using Cochran's $Q = \sum_j w_j(\bar{s}_j - \bar{s})^2$, where $w_j = 1/\mathrm{Var}(s_j)$ are inverse-variance weights [11]. The I^2 statistic quantifies the proportion of total variance attributable to true inter-framework differences rather than sampling noise:

$$I^2 = \max\left(0,\ \frac{Q - (k-1)}{Q}\right) \times 100\% \tag{4.1}$$

Values exceeding 75% are conventionally interpreted as substantial heterogeneity [12].

Binary agreement: Scores are binarized at $\tau = 0.5$ and pairwise agreement is quantified with Cohen's $\kappa = (p_o - p_e)/(1 - p_e)$, where p_o is observed and p_e is chance-expected agreement.

Domain effects: We fit a mixed-effects model $s_{ijd} = \mu + \alpha_j + \beta_d + (\alpha\beta)_{jd} + u_i + \epsilon_{ijd}$ with fixed effects for framework (α_j), domain (β_d), and their interaction, plus random intercepts for samples (u_i). One-way ANOVA η^2 effect sizes quantify the proportion of score variance explained by domain for each framework, and the raw difference $\Delta = \bar{s}_{\mathrm{GK}} - \bar{s}_{\mathrm{FIN}}$ captures the direction and magnitude of the General Knowledge-Finance contrast.

Cluster analysis: Hierarchical clustering with Ward's linkage is applied to the pairwise correlation matrix using distance $d_{jk} = \sqrt{2(1 - r_{jk})}$. The number of clusters is set to three based on the gap statistic [28] and confirmed by silhouette analysis and 1,000-iteration bootstrap co-assignment probabilities. We validate normalization sensitivity by repeating the analysis under raw, min-max, z-score, and rank transformations.

Cross-cluster consensus: For each cluster, the framework with the highest mean within-cluster correlation serves as the representative. A sample is labeled *faithful* if all three representatives score it above $\tau = 0.5$, *unfaithful* if all three score below τ, and *contested* otherwise.

Human validation: Pearson and Spearman correlations between framework scores and binary human annotations are computed on a 50-sample subset, with 95% bootstrap confidence intervals.

5. Results

Summary statistics, binary agreement, per-sample disagreement, and robustness analyses appear in Appendix A.

5.1. Overall Heterogeneity

Cochran's $Q = 10{,}055.63$ ($p < 0.001$, df $= 19$) exceeds the critical value at $\alpha = 0.05$ by a wide margin, and the I^2 statistic reaches 99.81%. Under the conventional interpretation of Higgins, Thompson, Deeks, and Altman [12], values above 75% indicate considerable heterogeneity; our estimate is well above that threshold, suggesting that the large majority of observed variance reflects true inter-framework differences rather than sampling noise. Mean scores span a wide range, from 0.063 (RAGChecker faithfulness) to 0.923 (G-Eval consistency), while standard deviations vary from 0.100 (RAGChecker) to 0.288 (SIRG); full

Table 1. Selected pairwise Pearson correlations with 95% bootstrap CIs (1,000 resamples). L and H denote the lower and upper bounds of the 95% interval, respectively. Bold: $r \geq 0.70$.

F1	F2	r	L	H
RAGAS	G-Eval	0.559	0.34	0.71
RAGAS	DeepEval	0.561	0.39	0.70
RAGAS	FaithJudge	0.687	0.52	0.80
RAGAS	RAGChecker	−0.042	−0.22	0.09
RAGAS	QAFactEval	−0.040	−0.20	0.11
DeepEval	ARES	0.535	0.40	0.66
RAGChecker	ARES	−0.022	−0.20	0.09
ARES	UniEval	**0.856**	0.81	0.89
ARES	HSAD	**0.874**	0.85	0.90
LRP4RAG	HSAD	**0.797**	0.73	0.85
LRP4RAG	SIRG	**0.748**	0.62	0.84
MetaRAG	KG-RAG	**0.715**	0.65	0.78
QAFactEval	MetaRAG	−0.228	−0.37	−0.07
QAFactEval	FaithJudge	−0.057	−0.21	0.08

per-framework statistics appear in Table 5 (Appendix A). Pairwise paired t-tests confirm that these differences are systematic: 167 of 190 framework pairs (87.9%) reach $p < 0.001$.

5.2. **Pairwise Correlations**

Pairwise Pearson correlations range from $r = -0.28$ (BERTScore vs. MetaRAG) to $r = 0.90$ (TRACe vs. ARES). The correlation heatmap (Figure 1, Appendix A) reveals two prominent structures: a dense block of moderate-to-high positive correlations among output-based and interpretability methods, and near-zero or negative correlations for RAGChecker and QAFactEval with most other frameworks. Table 1 reports selected pairwise correlations with 95% bootstrap confidence intervals (1,000 resamples). Four patterns stand out.

RAGAS–FaithJudge achieve $r = 0.69$ [0.52, 0.80], and RAGAS–DeepEval $r = 0.56$ [0.39, 0.70], reflecting shared reliance on LLM-based judgment. TRACe–ARES reach $r = 0.90$ [0.86, 0.93], ARES–HSAD $r = 0.87$ [0.85, 0.90], and LRP4RAG–HSAD $r = 0.80$ [0.73, 0.85], forming a large cluster of mutually correlated frameworks. RAGChecker's correlations with other frameworks span $[-0.13, 0.22]$, while QAFactEval's range from $[-0.28, 0.19]$, indicating these two frameworks frequently diverge from the rest of the ecosystem. LRP4RAG–SIRG reach $r = 0.75$ [0.62, 0.84] and KG-RAG–HSAD $r = 0.71$ [0.63, 0.78], but their correlations with LLM-judge methods average only $r = 0.23$.

Spearman's rank correlations follow similar patterns. For instance, RAGAS–DeepEval yields Spearman $\rho = 0.57$ compared with Pearson $r = 0.56$; the close agreement between rank and linear correlations for this pair suggests a roughly monotone relationship without strong ceiling effects. Binary classification agreement at $\tau = 0.5$ reinforces the correlation patterns: within-cluster pairs achieve κ up to 0.69 (ARES–UniEval), while between-cluster pairs involving RAGChecker or QAFactEval yield $\kappa \leq 0.001$ (full results in Table 6 and Figure 2, Appendix A).

5.3. **Cluster Structure**

Hierarchical clustering (Ward's method) on the correlation matrix identifies three groups, corresponding to the block-diagonal structure visible in the correlation heatmap (Figure 1). Table 2 summarizes cluster composition and within- vs. between-cluster correlations.

The LLM-as-judge cluster exhibits moderate internal agreement ($\bar{r} = 0.55$), while the large mixed-methods cluster, spanning token-level classifiers, interpretability tools, and knowledge-graph methods, shows comparable cohesion ($\bar{r} = 0.63$). The outlier cluster (BERTScore, GaRAGe, HALT-RAG, QAFactEval, RAGChecker) has near-zero within-cluster correlation ($\bar{r} = 0.01$), reflecting five frameworks that each measure a distinct aspect of faithfulness. The gap between within-cluster and between-cluster ($\bar{r}_b = 0.10$) correlations confirms that methodological family drives the cluster structure.

Table 2. Cluster composition and mean within-cluster ($\bar{r}_w$) vs. between-cluster ($\bar{r}_b$) Pearson correlations ($t = 13.32$, $p < 0.001$).

	LLM-as-Judge	**Mixed Methods**	**Outlier**
k	4	11	5
$\bar{r}_w$	0.55	0.63	0.01
$\bar{r}_b$		0.10	

LLM-as-Judge: RAGAS, G-Eval, DeepEval, FaithJudge. **Mixed:** ARES, HSAD, KG-RAG, LRP4RAG, LUMINA, LettuceDetect, MetaRAG, ReDeEP, SIRG, TRACe, UniEval. **Outlier:** BERTScore, GaRAGe, HALT-RAG, QAFactEval, RAGChecker.

Table 3. Domain means (left) and one-way ANOVA domain effects (right) for primary faithfulness metrics. $n_{\text{GK}} = n_{\text{FIN}} = 67$, $n_{\text{BIO}} = 66$, $\Delta = \text{GK} - \text{FIN}$. η^2 is the proportion of variance explained by domain; bold indicates $\eta^2 > 0.15$.

Framework	**GK**	**FIN**	**BIO**	Δ
RAGAS	0.942	0.844	0.801	0.097
G-Eval	0.937	0.961	0.870	−0.024
DeepEval	0.937	0.782	0.875	0.154
FaithJudge	0.949	0.919	0.860	0.030
TRACe	0.770	0.552	0.670	0.218
ARES	0.800	0.593	0.729	0.207
LRP4RAG	0.839	0.630	0.683	0.209
LUMINA	0.733	0.560	0.637	0.173
HSAD	0.712	0.504	0.591	0.208
KG-RAG	0.811	0.544	0.623	0.267
SIRG	0.761	0.515	0.557	0.246
MetaRAG	0.629	0.387	0.352	0.242
UniEval	0.672	0.526	0.613	0.146
LettuceDetect	0.726	0.630	0.615	0.096
GaRAGe	0.931	0.762	0.873	0.168
ReDeEP	0.340	0.217	0.267	0.123
BERTScore	0.178	0.222	0.321	−0.044
HALT-RAG	0.146	0.118	0.177	0.028
QAFactEval	0.232	0.248	0.287	−0.016
RAGChecker	0.062	0.080	0.045	−0.018

Metric	$F_{2,197}$	p	η^2
RAGAS faith.	6.07	0.003	0.058
G-Eval consist.	4.05	0.019	0.039
DeepEval faith.	6.85	0.001	0.065
FaithJudge faith.	4.00	0.020	0.039
TRACe adher.	24.27	<0.001	**0.198**
ARES faith.	23.17	<0.001	**0.190**
LRP4RAG relev.	21.26	<0.001	**0.178**
LUMINA score	20.91	<0.001	**0.175**
HSAD spectral	19.73	<0.001	**0.167**
KG-RAG faith.	28.60	<0.001	**0.225**
MetaRAG factoid	28.03	<0.001	**0.222**
SIRG attrib.	16.04	<0.001	0.140
UniEval consist.	13.90	<0.001	0.124
LettuceDetect f1	7.33	<0.001	0.069
GaRAGe factual.	8.25	<0.001	0.077
ReDeEP ext.	19.49	<0.001	**0.165**
BERTScore f1	10.59	<0.001	0.097
HALT-RAG nli	1.33	0.266	0.013
QAFactEval fact.	1.63	0.198	0.016
RAGChecker faith.	2.02	0.136	0.020

5.4. **Domain Effects**

The two-way ANOVA yields significant main effects for framework ($F_{19,3940} = 359.9$, $p < 0.001$), domain ($F_{2,3940} = 138.6$, $p < 0.001$), and their interaction ($F_{38,3940} = 6.95$, $p < 0.001$). The significant interaction term indicates that domain difficulty varies across evaluation methods. Table 3 presents faithfulness scores stratified by domain alongside one-way ANOVA F-statistics and η^2 effect sizes for each framework's primary faithfulness metric. Finance samples score lower for most frameworks, though G-Eval is a notable exception (see also Figures 5 and 6, Appendix A).

Within-domain agreement also varies: the median pairwise correlation is $r = 0.13$ for GK, $r = 0.23$ for Finance, and $r = 0.20$ for Biomedicine. The low median correlations across all domains reinforce that framework disagreement is pervasive, not confined to specialized content.

5.5. **Human Correlation Analysis**

We compute Pearson correlations between each framework's primary faithfulness score and binary human annotations on a 50-sample subset drawn from RAGBench. Table 4 reports the results, grouped by cluster.

Table 4. Framework–human Pearson correlations ($n = 50$). Cluster-level means shown in header.

LLM-as-Judge ($\bar{r} = 0.51$)			**Mixed Methods** ($\bar{r} = 0.19$)			**Outlier** ($\bar{r} = 0.05$)		
Framework	r	p	**Framework**	r	p	**Framework**	r	p
FaithJudge	0.596	<0.001	SIRG	0.361	0.010	RAGChecker	0.243	0.089
G-Eval	0.564	<0.001	LRP4RAG	0.266	0.062	GaRAGe	0.159	0.271
DeepEval	0.453	<0.001	KG-RAG	0.238	0.096	QAFactEval	−0.021	0.884
RAGAS	0.441	0.001	LUMINA	0.217	0.130	HALT-RAG	−0.023	0.874
			MetaRAG	0.207	0.149	BERTScore	−0.084	0.563
			HSAD	0.193	0.180			
			LettuceDetect	0.179	0.213			
			ARES	0.143	0.323			
			TRACe	0.133	0.356			
			UniEval	0.104	0.474			
			ReDeEP	0.097	0.502			

The LLM-as-judge cluster shows the strongest alignment with human judgments (cluster mean $r = 0.51$, all four frameworks significant at $p < 0.01$). FaithJudge achieves the highest individual correlation ($r = 0.60$), followed by G-Eval ($r = 0.56$). The mixed-methods cluster averages $r = 0.19$ across its eleven frameworks, with only SIRG reaching significance ($p = 0.01$). The outlier cluster averages $r = 0.05$; HALT-RAG, QAFactEval, and BERTScore all fall near zero, and GaRAGe and RAGChecker do not reach significance. These results suggest that LLM-as-judge methods, despite their stochastic nature, produce scores that track human faithfulness assessments more closely than either interpretability, token-level, or embedding-based approaches.

5.6. **Cross-Cluster Consensus**

To assess whether the three clusters converge on sample-level verdicts, we define a consensus protocol. For each cluster, we select the framework with the highest mean within-cluster correlation as the representative (Cluster 1: LRP4RAG; Cluster 2: FaithJudge; Cluster 3: QAFactEval). A sample is labeled *faithful* if all three representatives score it above $\tau = 0.5$, *unfaithful* if all three score it below τ, and *contested* otherwise. Of the 200 samples, 185 (92.5%) are contested, 12 (6.0%) are unanimously faithful, and 3 (1.5%) are unanimously unfaithful. The high contested rate quantifies how rarely the three methodological families agree on a binary verdict for the same sample.

6. **Discussion**

The results above establish that RAG evaluation frameworks disagree to a degree that has practical consequences for how systems are assessed and compared.

6.1. **Implications for RAG Research**

The $I^2 = 99.81\%$ finding calls into question three practices that are widespread in the current literature. First, *single-metric reporting* is unreliable: mean faithfulness scores on identical samples range from 0.923 (G-Eval) to 0.063 (RAGChecker), and across all 190 framework pairs an average of 40.2% of pairwise sample rankings invert when switching frameworks. Second, *cross-study comparison* becomes questionable when different studies adopt different evaluation tools, since observed score differences may reflect framework choice rather than system quality. Third, *evaluation-driven optimization* for a single metric does not guarantee improvement under alternative frameworks, given the low between-cluster agreement ($\bar{r} = 0.10$).

These measurement-level findings have direct consequences for system-level reliability. If the evaluation signal used during development is framework-dependent, then the safety guarantees of a deployed RAG system are only as strong as the particular metric used to validate it. A system certified as "faithful" by G-Eval ($\bar{s} = 0.92$) may appear substantially less faithful under RAGChecker ($\bar{s} = 0.06$) or QAFactEval ($\bar{s} = 0.26$). The cross-cluster consensus analysis (Section 5.6) quantifies the scale of this risk: 92.5% of samples receive conflicting verdicts depending on which methodological family is consulted.

6.2. **Sources of Disagreement**

The three-cluster structure traces disagreement to differences in what each family treats as evidence of faithfulness. The LLM-as-judge cluster (DeepEval, FaithJudge, G-Eval, RAGAS; $\bar{r} = 0.55$) assesses holistic plausibility, inheriting both the broad competence and the systematic biases of the judge model. The large mixed-methods cluster ($\bar{r} = 0.63$) spans 11 frameworks that combine knowledge graphs (KG-RAG), interpretability (LRP4RAG, ReDeEP), token- and span-level classifiers (LettuceDetect, ARES), and signal-aggregation approaches (HSAD, LUMINA, SIRG, TRACe, UniEval, MetaRAG), yet shows moderate agreement despite methodological diversity. The outlier cluster (BERTScore, GaRAGe, HALT-RAG, QAFactEval, RAGChecker; $\bar{r} = 0.01$) groups methods whose scoring mechanisms, including embedding similarity, claim-level NLI, and NLI ensembling, produce fundamentally different rank orderings from the rest of the ecosystem. The near-zero correlations between QAFactEval and all other methods warrant further investigation. Its question-generation formulation tests whether the context can independently answer questions derived from the response, a property closer to verifiability than to entailment or plausibility. A response may be well supported by context yet fail QAFactEval if the question generator produces queries that the context cannot directly answer. This suggests that the term "faithfulness," as currently used, conflates several distinguishable sub-constructs.

6.3. **Selection Guidelines**

The preceding subsections describe what we observe; this subsection translates those observations into actionable guidance. The cluster structure and domain-sensitivity results together yield four practical recommendations:

- **General-purpose scoring:** G-Eval or FaithJudge, both from the LLM-as-judge cluster ($\bar{r} = 0.55$) with the lowest domain sensitivity ($\eta^2 \leq 0.039$).
- **Claim-level diagnostics:** RAGChecker or HALT-RAG. These methods provide per-claim verdicts useful for localizing failures, with deterministic outputs.
- **Process-level analysis:** ReDeEP or LRP4RAG. Interpretability methods complement output-based evaluation by revealing how the generator uses retrieved context.
- **Multi-perspective reporting:** At minimum, report one method from each cluster. Cross-cluster agreement ($\bar{r} = 0.10$) is a stronger signal of quality than within-cluster agreement.

6.4. **Limitations**

Our study has four principal limitations. The 200-sample evaluation, while sufficient for stable meta-analytic statistics, may not capture rare failure modes. RAGBench draws from specific benchmarks (HotpotQA, FinQA, PubMedQA) and may not represent all RAG applications, particularly proprietary or multi-turn conversational settings. LLM-as-judge methods depend on the backbone model; our results use DeepSeek-V3, and substituting a different judge (e.g., GPT-4, Claude) could alter the within-cluster correlation structure. Finally, the human validation subset ($n = 50$) provides directional evidence but is too small

for definitive per-framework conclusions; a larger annotation effort would strengthen the human-alignment findings reported in Section 5.5.

7. Conclusion

This paper presents the first systematic meta-analysis of agreement across twenty RAG evaluation frameworks applied to identical samples. Three findings stand out. First, heterogeneity is substantial ($I^2 = 99.81\%$), with pairwise correlations ranging from $r = -0.28$ to $r = 0.90$ (median $r = 0.21$). Second, the frameworks form three clusters where within-cluster agreement ($\bar{r} = 0.01$, 0.55, 0.63) exceeds between-cluster agreement ($\bar{r} = 0.10$), and this structure is perfectly stable across normalization schemes (ARI = 1.0) and bootstrap resampling. Third, a cross-cluster consensus protocol labels 92.5% of samples as contested, indicating that the clusters rarely converge on a shared verdict. These results show that "faithfulness" is not a single construct as currently measured. Single-framework evaluation should be considered insufficient evidence for faithfulness claims, analogous to how single-rater reliability is insufficient in psychometrics. Going forward, the field should (i) define taxonomies that separate the sub-constructs bundled under "faithfulness," (ii) adopt meta-analytic reporting so new frameworks can be evaluated on whether they reduce heterogeneity, and (iii) develop domain-adapted calibration to account for systematic domain effects.

References

[1] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu. "PubMedQA: A Dataset for Biomedical Research Question Answering". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 2019, pp. 2567–2577.

[2] Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge, and W. Y. Wang. "FinQA: A Dataset of Numerical Reasoning over Financial Data". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 3697–3711.

[3] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 9459–9474.

[4] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert. "RAGAS: Automated Evaluation of Retrieval Augmented Generation". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL)*. 2024, pp. 150–158.

[5] D. Ru et al. "RAGCHECKER: a fine-grained framework for diagnosing retrieval-augmented generation". In: *Proceedings of the 38th International Conference on Neural Information Processing Systems*. NIPS '24. Vancouver, BC, Canada, 2024. ISBN: 9798331314385.

[6] O. Honovich, R. Aharoni, J. Herzig, H. Taitelbaum, D. Kukliansy, V. Cohen, T. Scialom, I. Szpektor, A. Hassidim, and Y. Matias. "TRUE: Re-evaluating Factual Consistency Evaluation". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2022, pp. 3905–3920.

[7] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. "BERTScore: Evaluating Text Generation with BERT". In: *International Conference on Learning Representations*. 2020.

[8] Z. Sun, X. Zang, K. Zheng, J. Xu, X. Zhang, W. Yu, Y. Song, and H. Li. "ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability". In: *International Conference on Learning Representations (ICLR)*. 2025.

[9] H. Hu, C. He, X. Xie, and Q. Zhang. "LRP4RAG: Detecting Hallucinations in Retrieval-Augmented Generation via Layer-wise Relevance Propagation". In: *arXiv preprint arXiv:2408.15533* (2025).

[10] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Transactions on Information Systems* 43.2 (2025).

[11] W. G. Cochran. "The Combination of Estimates from Different Experiments". In: *Biometrics* 10.1 (1954), pp. 101–129.

[12] J. P. Higgins, S. G. Thompson, J. J. Deeks, and D. G. Altman. "Measuring Inconsistency in Meta-Analyses". In: *BMJ* 327.7414 (2003), pp. 557–560.

[13] J. Hu, Y. Li, J. Zhong, W. Qi, and L. Zou. "Detecting Hallucinations in Retrieval-Augmented Generation via Semantic-level Internal Reasoning Graph". In: *arXiv preprint arXiv:2601.03052* (2026).

[14] R. Friel, M. Belyi, and A. Sanyal. "RAGBench: Explainable Benchmark for Retrieval-Augmented Generation Systems". In: *arXiv preprint arXiv:2407.11005* (2024).

[15] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 2511–2522.

[16] Confident AI. *DeepEval: The Open-Source LLM Evaluation Framework*. https://github.com/confident-ai/deepeval. 2024.

[17] M. S. Tamber, F. S. Bao, C. Xu, G. Luo, S. Kazi, M. Bae, M. Li, O. Mendelevitch, R. Qu, and J. Lin. "Benchmarking LLM Faithfulness in RAG with Evolving Leaderboards". In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 2025, pp. 799–811.

[18] A. R. Fabbri, C.-S. Wu, W. Liu, and C. Xiong. "QAFactEval: Improved QA-Based Factual Consistency Evaluation for Summarization". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2022, pp. 2587–2601.

[19] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, and J. Han. "Towards a Unified Multi-Dimensional Evaluator for Text Generation". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022.

[20] S. Yeh, S. Li, and T. Mallick. "LUMINA: Detecting Hallucinations in RAG System with Context–Knowledge Signals". In: *International Conference on Learning Representations*. 2026.

[21] J. Li, G. Tu, S. Cheng, J. Hu, J. Wang, R. Chen, Z. Zhou, and D. Shan. "LLM Hallucination Detection: A Fast Fourier Transform Method Based on Hidden Layer Temporal Signals". In: *arXiv preprint arXiv:2509.13154* (2025).

[22] I. T. Sorodoc, L. F. R. Ribeiro, R. Blloshmi, C. Davis, and A. de Gispert. "GaRAGe: A Benchmark with Grounding Annotations for RAG Evaluation". In: *Findings of the Association for Computational Linguistics: ACL 2025*. 2025, pp. 17030–17049.

[23] J. Saad-Falcon, O. Khattab, C. Potts, and M. Zaharia. "ARES: An Automated Evaluation Framework for Retrieval-Augmented Generation Systems". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2024.

[24] S. Goswami and S. Kurra. *HALT-RAG: A Task-Adaptable Framework for Hallucination Detection with Calibrated NLI Ensembles and Abstention*. 2025. arXiv: 2509.07475 [cs.CL]. URL: https://arxiv.org/abs/2509.07475.

[25] Á. Kovács and G. Recski. "LettuceDetect: A Hallucination Detection Framework for RAG Applications". In: *arXiv preprint arXiv:2502.17125* (2025).

[26] C. Sok, D. Luz, and Y. Haddam. "MetaRAG: Metamorphic Testing for Hallucination Detection in RAG Systems". In: *Identity-Aware AI Workshop at the 28th European Conference on Artificial Intelligence (ECAI)*. 2025.

[27] X. Zhu, Y. Xie, Y. Liu, Y. Li, and W. Hu. "Knowledge Graph-Guided Retrieval Augmented Generation". In: *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2025, pp. 8912–8924.

[28] R. Tibshirani, G. Walther, and T. Hastie. "Estimating the Number of Clusters in a Data Set via the Gap Statistic". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2 (2001), pp. 411–423.

Appendix A. Additional Analyses

A.1. Summary Statistics

Table 5 reports the mean and standard deviation of the primary faithfulness metric for each of the twenty frameworks evaluated on 200 samples.

Table 5. Faithfulness-score summary statistics for all twenty frameworks ($n = 200$). Each row reports the primary faithfulness metric for the given framework.

Framework	Mean	SD	Framework	Mean	SD	Framework	Mean	SD	Framework	Mean	SD
RAGAS	0.863	0.244	G-Eval	0.923	0.195	DeepEval	0.864	0.250	FaithJudge	0.910	0.188
GaRAGe	0.751	0.166	ARES	0.707	0.198	LRP4RAG	0.717	0.212	LettuceDetect	0.657	0.187
LUMINA	0.644	0.170	KG-RAG	0.660	0.237	TRACe	0.664	0.201	SIRG	0.611	0.288
HSAD	0.602	0.209	UniEval	0.603	0.171	MetaRAG	0.457	0.263	QAFactEval	0.256	0.179
ReDeEP	0.275	0.125	BERTScore	0.240	0.193	HALT-RAG	0.147	0.208	RAGChecker	0.063	0.100

A.2. Binary Agreement

Table 6 reports binary classification agreement at threshold $\tau = 0.5$. Raw agreement rates and Cohen's κ span a wide range across within- and between-cluster framework pairs.

Table 6. Binary agreement at threshold $\tau = 0.5$. Agree. = raw percent agreement; κ = Cohen's kappa. Left: within-cluster pairs. Right: between-cluster pairs.

Within-cluster pairs

F1	F2	Agree. (%)	κ
RAGAS	G-Eval	94.0	0.508
RAGAS	DeepEval	92.0	0.486
DeepEval	G-Eval	90.0	0.237
ARES	UniEval	90.0	0.692
LRP4RAG	HSAD	84.0	0.566

Between-cluster pairs

F1	F2	Agree. (%)	κ
RAGAS	RAGChecker	8.5	0.001
G-Eval	RAGChecker	5.5	0.001
QAFactEval	FaithJudge	15.0	0.001
RAGAS	SIRG	61.5	0.046
DeepEval	SIRG	65.5	0.150

RAGAS–G-Eval achieve $\kappa = 0.51$ (94.0% agreement), consistent with their moderate Pearson correlation. ARES–UniEval reach the highest $\kappa = 0.69$ (90.0% agreement). RAGChecker and QAFactEval pairs with LLM-judge frameworks yield $\kappa \leq 0.001$, confirming their measurement of distinct constructs. Figure 2 shows the full pairwise kappa matrix.

A.3. Correlation and Agreement Heatmaps

A.4. Inter-Framework Agreement

Figure 3 reports each framework's mean Pearson correlation with the other nineteen frameworks, ordered by magnitude. ARES, TRACe, HSAD, LRP4RAG, LUMINA, and KG-RAG form the top band ($\bar{r} \in [0.43, 0.44]$), followed by a broad mid-band of LLM-judge and mixed-methods frameworks ($0.11 \leq \bar{r} \leq 0.43$). Four outlier-cluster members sit near or below zero: HALT-RAG ($\bar{r} = 0.01$), RAGChecker ($\bar{r} = 0.00$), BERTScore ($\bar{r} = -0.01$), and QAFactEval ($\bar{r} = -0.11$), confirming their divergence from the broader ecosystem.

A.5. Disagreement Analysis

Per-sample disagreement, measured as the score range (max − min across all twenty primary metrics), is extreme. The overall mean range is 0.980 (SD = 0.050), indicating that a typical sample receives scores spanning most of the $[0, 1]$ interval. All 200 samples exceed a

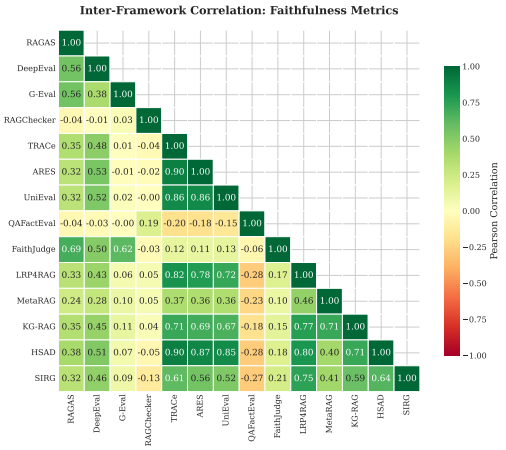

Figure 1. Pairwise Pearson correlation heatmap among primary faithfulness scores ($n = 200$). The block-diagonal structure reveals distinct framework clusters.

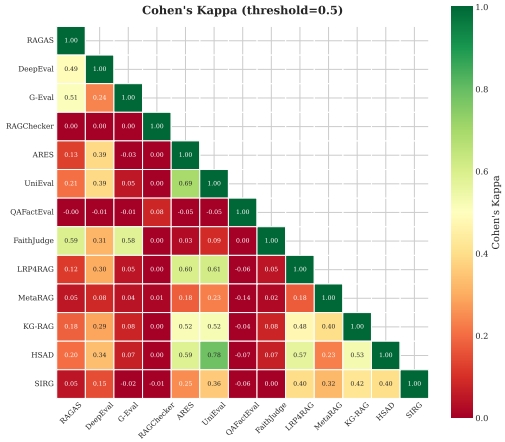

Figure 2. Pairwise Cohen's κ at $\tau = 0.5$ for thirteen frameworks. The block structure mirrors the correlation-based clusters.

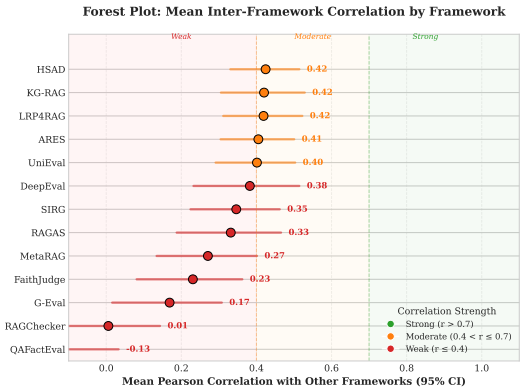

Figure 3. Forest plot of mean inter-framework Pearson correlation with 95% bootstrap CIs for each of the twenty frameworks. Color indicates correlation strength: green ($r > 0.7$), orange ($0.4 < r \leq 0.7$), red ($r \leq 0.4$).

range of 0.3, and the minimum observed range is 0.632. Across domains, General Knowledge samples show the highest disagreement (mean = 0.991, SD = 0.025), followed by Finance (mean = 0.978, SD = 0.048) and Biomedicine (mean = 0.970, SD = 0.067). Figure 4 visualizes the distribution of per-sample disagreement and its breakdown by domain.

A.6. **Robustness Analysis**

We test whether the cluster structure is an artifact of our normalization choice or the number of clusters selected. We repeat the full analysis pipeline under four normalization schemes: raw scores, min-max scaling, z-score standardization, and rank transformation. Table 7 reports heterogeneity statistics and mean correlations for each scheme. Cluster assignments are identical across all four schemes: the adjusted Rand index (ARI) between

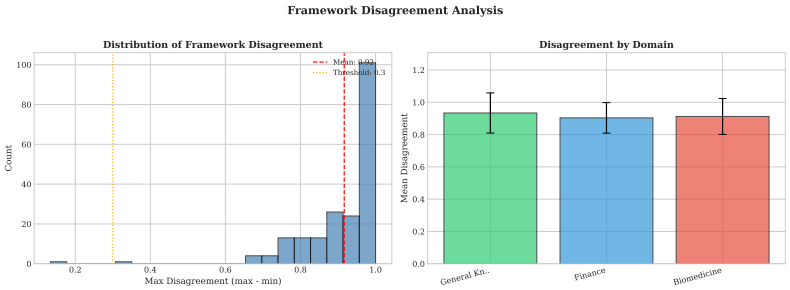

Figure 4. Left: distribution of per-sample score range (max − min) across all twenty frameworks. The dashed red line marks the mean (0.980); all 200 samples exceed the 0.3 threshold (dotted yellow). Right: mean disagreement by domain, with error bars showing ±1 SD.

every pair of normalization-induced clusterings is 1.0, indicating that the three-cluster structure is invariant to how scores are rescaled.

Table 7. Heterogeneity and correlation statistics under four normalization schemes. Cluster assignments are identical across all schemes (pairwise ARI = 1.0).

Scheme	Q	I^2 **(%)**	$\bar{r}$	$\bar{r}_w$	$\bar{r}_b$
Raw	52,797	99.96	0.264	0.397	0.10
Min-max	10,056	99.81	0.264	0.397	0.10
Z-score	≈0	0.0	0.264	0.397	0.10
Rank	≈0	0.0	0.262	0.350	0.10

The correlation structure ($\bar{r}$, $\bar{r}_w$, $\bar{r}_b$) is nearly identical across schemes. Q and I^2 vary because they depend on score scale: z-score and rank normalization equalize variances, driving Q toward zero, while the correlation-based cluster analysis is unaffected. This confirms that the heterogeneity finding is not an artifact of min-max scaling. We evaluate the choice of $k = 3$ clusters using the gap statistic, silhouette scores, and bootstrap co-assignment probabilities (1,000 resamples). The gap statistic peaks at $k = 3$ (gap = 0.635), with the next-highest value at $k = 2$ (gap = 0.610); both exceed the values at $k = 4$–8. Silhouette scores increase gradually from $k = 2$ (0.205) to $k = 5$ (0.230) but decline beyond that, consistent with $k = 3$ being a reasonable choice. The mixed-methods cluster shows the tightest cohesion (0.90–1.00) and the LLM-as-judge cluster ranges from 0.75 to 0.999, while the outlier cluster is less stable (0.46–0.97), reflecting the heterogeneous nature of its members. Alternative linkage methods (average, complete, single) yield different cluster boundaries: Ward vs. average ARI = 0.35, Ward vs. complete ARI = 0.34, Ward vs. single ARI = 0.12. This sensitivity to linkage method is expected, given the continuous nature of the correlation structure; the gap statistic and bootstrap analysis support $k = 3$ under Ward's criterion.

A.7. **Domain Score Profiles**

Figures 5 and 6 visualize faithfulness scores across the three RAGBench domains. Finance samples score lower for most frameworks, though G-Eval is a notable exception.

A.8. **Pairwise Comparison Scatter Plots**

Figure 7 shows pairwise score relationships for FaithJudge, the framework with the highest human correlation ($r = 0.60$, Table 4). Strong diagonal alignment appears with LLM-judge peers (RAGAS $r = 0.687$, G-Eval $r = 0.624$, DeepEval $r = 0.497$), consistent

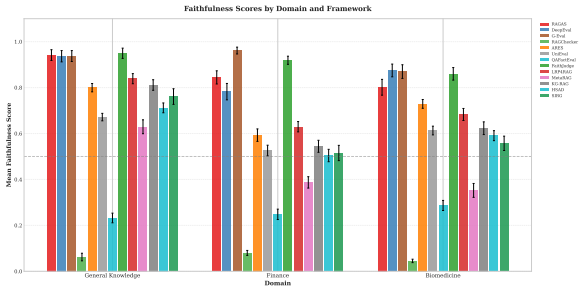

Figure 5. Mean faithfulness scores by domain for all twenty frameworks. Error bars show ±1 SE. Most frameworks assign lower scores to Finance samples; G-Eval and FaithJudge remain stable across domains, consistent with their low η^2 values in Table 3.

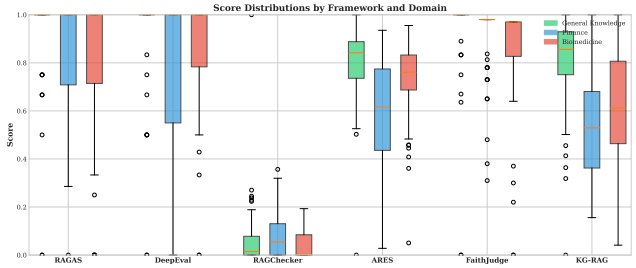

Figure 6. Score distributions by domain for six representative frameworks (one LLM-as-judge, one outlier, and four mixed-methods). Box spans IQR; whiskers extend to 1.5× IQR. RAGChecker scores cluster near zero regardless of domain, while ARES and FaithJudge show compressed upper-quartile distributions.

with their shared cluster membership. Agreement weakens for mixed-methods frameworks (LRP4RAG $r = 0.172$, MetaRAG $r = 0.102$) and breaks down for outlier-cluster members (RAGChecker $r = -0.034$, QAFactEval $r = -0.057$).

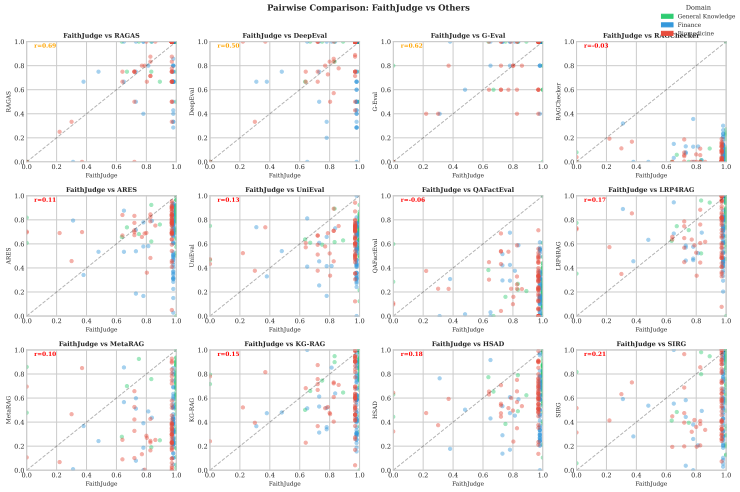

Figure 7. Pairwise scatter plots of FaithJudge scores (x-axis) against twelve other frameworks (y-axis), colored by domain. Pearson r is shown per panel.