# Who Judges the Judges? A Multilingual Probe for Complementary LLM Evaluation Metrics

Salvatore Ricciardi[1,*,†], Simone Scaboro[1,†], Beatrice Turano[1,†], Cristina Giannone[1], Andrea Favalli[1] and Raniero Romagnoli[1]

[1]*Almawave Labs s.r.l. , Via di Casal Boccone 188/190, 00137 Rome, Italy*

**Abstract**

Evaluating large language models (LLMs) in practical settings requires more than leaderboard-style benchmarking: model outputs are open-ended, correctness can be expressed in many valid forms, and no single metric captures relevance, faithfulness and factuality at once. We present a multilingual *probe* for LLM evaluation that tests whether a metric captures meaning beyond surface-level text similarity. The protocol compares scores on correctly paired data against scores on meaning-destroyed (shuffled) variants, and we apply it to three LLM-as-a-judge metrics—*Relevancy*, *Groundedness* and *Correctness*—in a retrieval-augmented generation (RAG) setting across six languages. The metrics turn out to be reliable and language-invariant, with a large gap between original and shuffled data, but a correlation analysis reveals that reliability alone is not enough: *Groundedness* and *Correctness* are near-collinear, while *Relevancy* is largely independent. We argue that evaluation pipelines should combine *complementary families of metrics*, and that a sound benchmarking process must probe not only whether a metric works but also how it interacts with the others; we close with a concrete prompt-redesign recommendation to decorrelate the two redundant metrics.

**Keywords**

LLM evaluation, multilingual evaluation, LLM-as-a-judge, retrieval-augmented generation, metric calibration

## 1. Introduction

Evaluating large language models (LLMs) deployed in real applications is fundamentally different from ranking them on a static benchmark. In production, a single model is used across tasks that demand different notions of quality, sometimes beyond what standard datasets capture. Moreover, LLM outputs can have many valid surface forms, and reference answers are often unavailable at inference time, so evaluation must also work in reference-less settings.

Given these constraints, a single metric is rarely enough to capture every relevant aspect of a generated answer. A high lexical-overlap score such as BLEU [2] or ROUGE [3] can hide a wrong answer, while a fluent answer can still be unfaithful to a given ground truth. However, stacking metrics is not the solution either, because too many overlapping scores make evaluation messy, redundant and hard to interpret.

Recent evaluation practice leans heavily on LLM-as-a-judge prompts [4, 5, 6] for a wide range of tasks. We argue that such metrics should be treated as complementary signals, with each family capturing a different aspect of answer quality. For this to hold, each judge prompt should be specific, and each metric should be validated to ensure it measures something distinct.

With this perspective we introduce a *probe* that decides whether a metric reacts to variation of meaning in the data and whether it captures aspects of the evaluation that are not caught by other metrics. We apply this probe to a multilingual retrieval-augmented generation (RAG) [7] case study. First, the probe compares the scores a metric assigns to correctly paired records (formed of query, extracted fragments and a gold answer) against the scores it assigns to meaning-destroyed (shuffled)

*NL4AI 2026: Ninth Workshop on Natural Language for Artificial Intelligence, October 6-9th, 2026, Perugia, Italy [1]*

*Corresponding author.

†These authors contributed equally.

s.ricciardi@almawavelabs.it (S. Ricciardi); s.scaboro@almawavelabs.it (S. Scaboro); b.turano@almawavelabs.it (B. Turano); c.giannone@almawavelabs.it (C. Giannone); a.favalli@almawavelabs.it (A. Favalli); r.romagnoli@almawavelabs.it (R. Romagnoli)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-05

CEUR-WS.org/Vol-4315/paper7.pdf

variants. We then measure the correlation between metrics to reveal whether they effectively capture distinct signals. Our contribution is twofold:

1. *A multilingual probe for judge metrics.* We introduce an original-vs-shuffled probe, paired with a correlation analysis, to test whether a metric reacts to meaning and whether it adds information beyond the rest of the evaluation set. We apply it to three LLM-as-a-judge metrics in a RAG setting across six languages, showing that all judges are reliable, but that Groundedness and Correctness are partially redundant.
2. *A practical criterion for ensemble design.* We show that judge metrics should be kept only if they are both reliable and non-redundant, and that correlation analysis is necessary to detect when two apparently different judges are effectively measuring the same facet.

## 2. Related Work

LLM evaluation remains a central topic in the field, and a number of resources, initiatives, and standards have been proposed for this purpose. Prominent examples include static benchmark suites such as MMLU [8], which assesses broad multitask knowledge through multiple-choice questions spanning diverse subjects; BIG-bench [9], a large collaborative benchmark designed to probe a wide range of linguistic, reasoning, and emergent capabilities; and HELM [10], a holistic evaluation framework that emphasizes transparency, reproducibility, and multi-dimensional assessment beyond accuracy alone. While these benchmarks are useful for measuring general capabilities, they are often optimized for standardized settings rather than production-oriented workflows. As a result, they do not always reflect task-specific requirements, and they can be difficult to compare across models due to contamination issues, prompt sensitivity, and the mismatch between benchmark scores and real deployment errors. More importantly, they typically do not explain what a model is failing at in a customer-facing application.

Human preference arenas [11] provide a complementary perspective by aggregating large-scale relative judgments of model outputs and capturing holistic quality signals that are often closer to user preference than single-task accuracy metrics. However, such arenas are expensive to maintain, difficult to rerun frequently, and impractical to integrate into a product team's iterative development loop.

For task-specific generation problems, standard metrics such as BLEU [2] for machine translation or BERTScore for summarization often miss important aspects of quality: LLM outputs are open-ended, and a correct answer may have a surface form that differs substantially from the reference, leading such metrics to assign low scores despite semantic adequacy. For this reason, using an LLM to score another model's output [4, 5, 6] has gained increasing popularity, since LLM-based judges can capture nuance that lexical or embedding-based metrics often miss. At the same time, LLM judges are known to exhibit systematic biases, such as position and verbosity effects [12], and to require careful grounding of factual claims [13, 14]. This is precisely why a judge must be validated before it is trusted, and why correlation-based meta-evaluation must be interpreted with care to avoid its documented pitfalls [15].

Multilingual evaluation remains under-served: while large efforts such as MEGA [16] and more recent European-language benchmarks [17] have expanded coverage beyond English, generative quality is still often assessed primarily in English, and dedicated multilingual evaluation remains the exception rather than the norm. Our probe is therefore applied identically across six languages to expose any language-specific bias in the LLM-as-a-judge metrics.

## 3. Method

Our method has two components: a probe that tests whether a metric reacts to meaning and quantifies what it adds relative to the others, and a broader evaluation context in which such metrics are used in practice.

### 3.1. Metric Families

No single metric captures every aspect of an LLM answer. Depending on the task, evaluation may rely on deterministic overlap measures, embedding-based similarity, LLM-based judges, or hybrid scores that combine several signals. These metric families are complementary rather than interchangeable: each captures only part of the quality space, and the final evaluation depends on the ensemble of signals a task chooses to use. In this paper, we keep that broader view in the background, but the empirical analysis is restricted to the judge family.

In LLM-as-a-judge metric design, a judge model scores an answer against an explicit, task-specific rubric. Each metric takes the query $q$, the answer $a$, a prompt $p$ and (optionally) a set of sources $D$, and returns a score between $0$ and $1$, where higher is always better. LLM-as-a-judge metrics capture nuance that traditional approaches miss, but are heavily prompt- and model-dependent and become costly at scale.

### 3.2. The Probe

A metric is only useful if its score changes with meaning. We test this by constructing a shuffled copy of the dataset in which answers or fragments are randomly re-paired, destroying alignment while preserving the text distribution. For any reliable metric $m$, we expect

$$\Delta_m \;=\; \overline{m}_{\text{original}} \;-\; \overline{m}_{\text{shuffled}} \;>\; 0, \tag{1}$$

with $\overline{m}_{\text{shuffled}} \approx 0$ in the ideal case.

A positive gap $\Delta_m$ shows that the metric $m$ responds to meaning, but not whether it adds distinct information. We therefore complement the reliability test with a correlation analysis based on Pearson and Spearman coefficients [18, 19], which helps detect near-collinear metrics and avoids the usual pitfalls of correlation-based meta-evaluation [15]. A metric enters the ensemble only if it separates original from shuffled data and is not redundant with an existing metric.

## 4. Experiment

### 4.1. Dataset

We use the rag-embeddings-relevance-similarity[1] corpus from the HuggingFace Hub, restricted to the pubmed–qa [20] and covid–qa [21] train subsets, and sampled to $500$ records. Each record contains a query, four supporting fragments and a model response that paraphrases them. Since our RAG models must be evaluated in multiple languages but the corpus is English-only, we translate it into the five other target languages (Italian, Portuguese, Spanish, French, German) with the Opus-MT neural machine translation models [22] [2].

### 4.2. Metrics under Test

In RAG, an LLM answers a query $q$ using a set of provided sources $D$; the answer $a$ should be relevant to $q$ and grounded in $D$. In our experiment we tested three RAG metrics: Relevancy, Groundedness and Correctness.Each of these metrics is implemented as an LLM-as-a-judge prompt, returning a normalised score in $[0, 1]$ and a reason (full prompts in Appendix A):

- **Relevancy** $\mathcal{R}$ (drawn from RAGAS [23]) — how on-topic and relevant the answer is to the query. High = fully on-topic; low = drifts off the question.
- **Groundedness** $\mathcal{G}$— whether the information in the answer is expressed or at least implied in $D$. High = supported by $D$; low = hallucinated.

[1] https://huggingface.co/datasets/argilla/rag-embeddings-relevance-similarity
[2] We used the models from English to our specified target languages

**Table 1**
Mean ± std of each metric on the original data and on the two shuffled variants, for both Italian *it* and English *en*. Every metric satisfies $\Delta \approx 0.9 > 0$ in the original data configuration while in the shuffled ones the results are near to 0.

| | $O$ | | $S_d$ | | $S_a$ | |
|---|---|---|---|---|---|---|
| *Metric* | *it* | *en* | *it* | *en* | *it* | *en* |
| Relevancy | $0.91 \pm 0.21$ | $0.93 \pm 0.18$ | $0.91 \pm 0.21$ | $0.93 \pm 0.18$ | $0.01 \pm 0.09$ | $0.01 \pm 0.09$ |
| Groundedness | $0.91 \pm 0.25$ | $0.94 \pm 0.20$ | $0.01 \pm 0.08$ | $0.01 \pm 0.09$ | $0.00 \pm 0.06$ | $0.01 \pm 0.09$ |
| Correctness | $0.91 \pm 0.26$ | $0.94 \pm 0.21$ | $0.00 \pm 0.08$ | $0.06 \pm 0.24$ | $0.00 \pm 0.06$ | $0.00 \pm 0.06$ |

- **Correctness** $\mathcal{C}$— the overall correctness of the answer $a$ given the query $q$ and the set of sources $D$. High = factually correct; low = wrong.

The default judge model used to compute all the above metrics is `gpt-4.1-mini`.

### 4.3. Corruption Strategies

A RAG answer has *two* independently corruptible ingredients: the set of retrieved fragments (docs) the judge reads before scoring, and the generated answer the model produced. We exploit this with two corruption datasets: $S_d$ by pairing a query and answer with the wrong set of source documents, and $S_a$ by paring the query and documents with the wrong answer. We call the original, not shuffled dataset, $O$. We expect the three metrics to capture the disalignments and reveal which signal each is more sensitive to: groundedness should collapse when either the docs or the answer are corrupted, whereas relevancy, comparing only query and answer, should be insensitive to the document-only shuffling.

We omit a fully shuffled condition because it would destroy all alignments simultaneously and would not allow us to disentangle the contribution of each judge.

## 5. Results

Table 1 reports the Italian and English splits, showing that every metric separates original from shuffled by roughly $0.9$, and both $S_a$ and $S_d$ collapse near zero with low variance. The same separation holds across all six languages (see Appendix B), so the metrics show no language bias. This supports the reliability hypothesis.

On the original dataset $O$ (first column), in the Italian and in English languages, all three metrics centre near $1.0$; their means are pulled down to the range $[0.91;0.94]$ only by a thin tail of genuinely hard cases (Relevancy: Italian $\sigma = 0.21$; English $\sigma = 0.18$, Groundedness: Italian $\sigma = 0.25$; English $\sigma = 0.20$, Correctness: Italian $\sigma = 0.26$; English: $\sigma = 0.21$). On both $S_a$ and $S_d$ (second and third columns), in the Italian and in English languages, every distribution sits on zero, except $\mathcal{R}$ in $S_d$ because queries and answers are the same as the original dataset $O$, so the results are the same. Splitting corruption into docs and answer shows that *both* components are penalised: $\mathcal{G}$ and $\mathcal{C}$ drop to $\mu \approx 0.00$–$0.01$ under either corruption, and the rare high points are isolated outliers. The corrupted pairs are, in other words, correctly rejected.

Table 3 shows Pearson and Spearman matrices over the five shuffled Italian series; in both, $\mathcal{G}$ and $\mathcal{C}$ reach correlation $1.00$ on matched corruption. Here $\mathcal{G}_d$/ $\mathcal{G}_a$ denote Groundedness on shuffled docs / shuffled answer, and analogously $\mathcal{C}_d$/ $\mathcal{C}_a$ for Correctness. The redundancy sharpens under the shuffled test. When the *same* component is corrupted, Groundedness and Correctness become perfectly correlated.

Table 4 shows Pearson and Spearman matrices over the five shuffled English series. Here $\mathcal{G}_d$/$\mathcal{G}_a$ denote Groundedness on shuffled docs / shuffled answer, and analogously $\mathcal{C}_d$/ $\mathcal{C}_a$ for Correctness. Unlike the Italian case, matched corruption does *not* drive Groundedness and Correctness to a perfect $1.00$: the redundancy is directional. It is strongest when the *answer* is corrupted ($\mathcal{G}_a$–$\mathcal{C}_a$: $0.67$ Pearson, $0.70$

**Table 2**
Pearson and Spearman correlation among the LLM-judges on the *original* English and Italian data, over the three metrics ($\mathcal{R}$, $\mathcal{G}$, and $\mathcal{C}$). In both languages the strongest off-diagonal association is Groundedness–Correctness (bold), indicating that Correctness largely re-measures Groundedness already on uncorrupted data; Groundedness and Correctness are near-collinear, while Relevancy is largely independent.

| | English | | | | | | Italian | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Pearson | | | Spearman | | | Pearson | | | Spearman | | |
| | $\mathcal{R}$ | $\mathcal{G}$ | $\mathcal{C}$ | $\mathcal{R}$ | $\mathcal{G}$ | $\mathcal{C}$ | $\mathcal{R}$ | $\mathcal{G}$ | $\mathcal{C}$ | $\mathcal{R}$ | $\mathcal{G}$ | $\mathcal{C}$ |
| $\mathcal{R}$ | 1.00 | 0.24 | 0.21 | 1.00 | 0.19 | 0.17 | 1.00 | 0.28 | 0.25 | 1.00 | 0.23 | 0.20 |
| $\mathcal{G}$ | 0.24 | 1.00 | **0.87** | 0.19 | 1.00 | **0.88** | 0.28 | 1.00 | **0.95** | 0.23 | 1.00 | **0.92** |
| $\mathcal{C}$ | 0.21 | **0.87** | 1.00 | 0.17 | **0.88** | 1.00 | 0.25 | **0.95** | 1.00 | 0.20 | **0.92** | 1.00 |

**Table 3**
Pearson and Spearman correlation on the *shuffled* Italian data over the five series: $\mathcal{R}$, $\mathcal{G}_d/\mathcal{G}_a$, and $\mathcal{C}_d/\mathcal{C}_a$. Corrupting the same component drives Groundedness and Correctness to $1.00$ (bold), confirming that Correctness largely re-measures Groundedness.

| | Pearson | | | | | Spearman | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{R}$ | $\mathcal{G}_d$ | $\mathcal{G}_a$ | $\mathcal{C}_d$ | $\mathcal{C}_a$ | $\mathcal{R}$ | $\mathcal{G}_d$ | $\mathcal{G}_a$ | $\mathcal{C}_d$ | $\mathcal{C}_a$ |
| $\mathcal{R}$ | 1.00 | 0.53 | 0.66 | 0.53 | 0.66 | 1.00 | 0.32 | 0.40 | 0.32 | 0.40 |
| $\mathcal{G}_d$ | 0.53 | 1.00 | 0.82 | **1.00** | 0.82 | 0.32 | 1.00 | 0.82 | **1.00** | 0.82 |
| $\mathcal{G}_a$ | 0.66 | 0.82 | 1.00 | 0.82 | **1.00** | 0.40 | 0.82 | 1.00 | 0.82 | **1.00** |
| $\mathcal{C}_d$ | 0.53 | **1.00** | 0.82 | 1.00 | 0.82 | 0.32 | **1.00** | 0.82 | 1.00 | 0.82 |
| $\mathcal{C}_a$ | 0.66 | 0.82 | **1.00** | 0.82 | 1.00 | 0.40 | 0.82 | **1.00** | 0.82 | 1.00 |

**Table 4**
Pearson and Spearman correlation on the *shuffled* English data over the five series: $\mathcal{R}$, $\mathcal{G}_d/\mathcal{G}_a$, and $\mathcal{C}_d/\mathcal{C}_a$. The strongest redundancy is between Groundedness and Correctness when the answer is corrupted ($\mathcal{G}_a$–$\mathcal{C}_a$, bold: $0.67/0.70$); matched-document corruption yields only $0.34$. Matched corruption does not reach $1.00$ here, indicating a weaker overlap than in the Italian setting.

| | Pearson | | | | | Spearman | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{R}$ | $\mathcal{G}_d$ | $\mathcal{G}_a$ | $\mathcal{C}_d$ | $\mathcal{C}_a$ | $\mathcal{R}$ | $\mathcal{G}_d$ | $\mathcal{G}_a$ | $\mathcal{C}_d$ | $\mathcal{C}_a$ |
| $\mathcal{R}$ | 1.00 | 0.41 | 0.41 | 0.16 | 0.62 | 1.00 | 0.27 | 0.27 | 0.16 | 0.40 |
| $\mathcal{G}_d$ | 0.41 | 1.00 | 0.45 | 0.34 | 0.67 | 0.27 | 1.00 | 0.49 | 0.34 | 0.70 |
| $\mathcal{G}_a$ | 0.41 | 0.45 | 1.00 | 0.14 | **0.67** | 0.27 | 0.49 | 1.00 | 0.16 | **0.70** |
| $\mathcal{C}_d$ | 0.16 | 0.34 | 0.14 | 1.00 | 0.23 | 0.16 | 0.34 | 0.16 | 1.00 | 0.24 |
| $\mathcal{C}_a$ | 0.62 | 0.67 | **0.67** | 0.23 | 1.00 | 0.40 | 0.70 | **0.70** | 0.24 | 1.00 |

Spearman, bold) and markedly weaker under $S_d$ ($\mathcal{G}_d$–$\mathcal{C}_d$: $0.34$ in both). This still points to Correctness partially re-measuring Groundedness, but the effect is far milder than on the Italian data.

## 6. Discussion

From our experiments, two conclusions follow. First, the metrics are *reliable and unbiased*: original scores sit far above shuffled for every judge and every language, so the metrics genuinely separate good RAG outputs from corrupted ones. Second, *Correctness* $\approx$ *Groundedness*: the two metrics are highly correlated on the original data ($0.95$) and become indistinguishable under matched corruption ($1.00$). In their current form, Correctness contributes little signal beyond Groundedness on RAG question answering.

This is precisely the situation the two-stage gate of Section 3.2 is meant to catch. Reliability alone would have admitted both metrics; only the correlation analysis reveals that the ensemble is over-

weighting a single facet. The practical takeaway is that metric families should be treated as *complementary signals, not interchangeable scores*, and that benchmarking should probe not only whether a metric separates good from bad outputs, but also whether it contributes information that is genuinely distinct from the rest of the ensemble.

**Recommendation.** Because Correctness currently re-measures Groundedness, we recommend redesigning the Correctness rubric so that it focuses on factual agreement with respect to the query and a gold answer, while Groundedness remains strictly tied to support from the retrieved sources. This would make the two metrics more complementary and would better justify keeping both in the evaluation stack. More generally, whenever a new judge is introduced, it should be tested not only for reliability under corruption, but also for redundancy against the existing ensemble.

## 7. Conclusion

We presented a multilingual probe for LLM evaluation that certifies whether a metric reacts to meaning and, through a correlation analysis,whether it adds a signal distinct from the others. Applied to three LLM-as-a-judge metrics on a RAG case study across six languages, the probe shows that they are reliable and language-invariant, yet *Groundedness* and *Correctness* are near-collinear. The broader message is that evaluation pipelines should combine *complementary families of metrics*, and that a sound enchmarking process must test both whether a metric works and how it interacts with the rest of the ensemble.

**Limitations and future work.** This study is limited to RAG question answering, so the next step is to test whether the same correlation pattern holds on other tasks such as information extraction, sentiment analysis, and machine translation. A second big limitation regards the models used as judges: here we presented only `gpt-4.1-mini`, but a more complete comparison is needed for further analyses. A third direction is to strengthen the probe itself through prompt-variant and language-specific ablations, especially for the LLM-judges that currently appear redundant. Lastly, we plan to extend the same reliability-and-redundancy analysis to dataset not machine translated from English, to generalize results of the analysis using original datasets.

## Declaration on Generative AI

During the preparation of this work, the authors used Claude Opus 5 as AI assistant in order to: draft and format portions of the manuscript text and tables. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## References


[1] A. Bondielli, G. Bonetta, E. Leonardelli, I. Siragusa, Preface to the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Ninth Workshop on Natural Language for Artificial Intelligence (NL4AI 2026) co-located with 24th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2026), 2026.

[2] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002, pp. 311–318.

[3] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 2004, pp. 74–81.

[4] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-bench and chatbot arena, in:

Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track, 2023.

[5] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, 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 (EMNLP), 2023, pp. 2511–2522.

[6] S. Kim, J. Shin, Y. Cho, J. Jang, S. Longpre, H. Lee, S. Yun, S. Shin, S. Kim, J. Thorne, M. Seo, Prometheus: Inducing fine-grained evaluation capability in language models, in: International Conference on Learning Representations (ICLR), 2024.

[7] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, Advances in Neural Information Processing Systems (NeurIPS) 33 (2020) 9459–9474.

[8] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt, Measuring massive multitask language understanding, in: International Conference on Learning Representations (ICLR), 2021.

[9] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, et al., Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, Transactions on Machine Learning Research (TMLR) (2023).

[10] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, et al., Holistic evaluation of language models, Transactions on Machine Learning Research (TMLR) (2023).

[11] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez, I. Stoica, Chatbot arena: An open platform for evaluating LLMs by human preference, in: Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

[12] P. Wang, L. Li, L. Chen, Z. Cai, D. Zhu, B. Lin, Y. Cao, Q. Liu, T. Liu, Z. Sui, Large language models are not fair evaluators, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024, pp. 9440–9450.

[13] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, H. Hajishirzi, FActScore: Fine-grained atomic evaluation of factual precision in long form text generation, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 12076–12100.

[14] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, X. Xie, A survey on evaluation of large language models, ACM Transactions on Intelligent Systems and Technology 15 (2024) 1–45.

[15] N. Mathur, T. Baldwin, T. Cohn, Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 4984–4997.

[16] K. Ahuja, H. Diddee, R. Hada, M. Ochieng, K. Ramesh, P. Jain, A. Nambi, T. Ganu, S. Segal, M. Ahmed, K. Bali, S. Sitaram, MEGA: Multilingual evaluation of generative AI, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023, pp. 4232–4267.

[17] K. Thellmann, B. Stadler, M. Fromm, J. S. Buschhoff, A. Jude, F. Barth, J. Leveling, N. Flores-Herr, J. Köhler, R. Jäkel, M. Ali, Towards multilingual llm evaluation for european languages, 2024. URL: https://arxiv.org/abs/2410.08928. arXiv:2410.08928.

[18] K. Pearson, Note on regression and inheritance in the case of two parents, Proceedings of the Royal Society of London 58 (1895) 240–242.

[19] C. Spearman, The proof and measurement of association between two things, The American Journal of Psychology 15 (1904) 72–101.

[20] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, X. Lu, PubMedQA: A dataset for biomedical research question answering, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 2567–2577.

[21] T. Möller, A. Reina, R. Jayakumar, M. Pietsch, Covid-QA: A question answering dataset for covid-19, in: Proceedings of the 1st Workshop on NLP for COVID-19, Online. Assciation of Computational

Linguistic, 2020.
[22] J. Tiedemann, S. Thottingal, OPUS-MT – building open translation services for the world, in: Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT), 2020, pp. 479–480.
[23] S. Es, J. James, L. Espinosa-Anke, 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 (EACL): System Demonstrations, 2024, pp. 150–158.

# A. Prompts

Following the prompts used in the generative metrics during the benchmarks:

**Relevancy**

```
You are an expert evaluator of answers produced by a Retrieval-Augmented Generation
    (RAG) system.

Your task is to assess the **Relevancy** of the generated answer.
Relevancy measures how well the answer addresses the question.
A relevant answer is:
- on-topic
- complete
- directly responsive to all aspects of the question
- free of unrelated or off-topic information

Carefully read the question and the generated answer.
Evaluate how well the answer fulfills the question's intent.

Assign a score from 0.0 to 1.0:
- 1.0 = fully relevant, complete, and directly answering the question
- 0.0 = completely irrelevant or failing to address the question

Keep your reasoning brief (max 50 words).

# Question:
{query}

# Generated Answer:
{generated_answer}

Provide ONLY a JSON object in this format:

{{
    "score": <float between 0 and 1>,
    "reason": "<short explanation (max 50 words)>"
}}

Example output:

{{
    "score": 0.85,
    "reason": "The answer covers the main question topic but misses some details
    asked explicitly."
}}
```

**Groundedness**

```
 You are an expert evaluator of answers produced by a Retrieval-Augmented Generation
    (RAG) system.

Your task is to assess the **Groundedness** of the generated answer.
Groundedness measures whether the information in the answer is supported by the
    retrieved documents.

A grounded answer:
- is based only on the provided documents
- does not introduce hallucinations or invented facts
- has factual statements traceable to the retrieved context

Read the documents and the generated answer.
Determine if the answer relies on evidence from the documents.

Assign a score from 0.0 to 1.0:
- 1.0 = fully grounded, no unsupported claims
- 0.0 = completely hallucinated or unrelated to the documents

Keep your reasoning brief (max 50 words).

# Retrieved Documents:
{documents}

# Generated Answer:
{generated_answer}

Provide ONLY a JSON object in this format:

{{
    "score": <float between 0 and 1>,
    "reason": "<short explanation (max 50 words)>"
}}

Example:

{{
    "score": 0.9,
    "reason": "Most statements match the documents, but one detail was not found in
    the context."
}}
```

**Correctness**

```
You are an expert evaluator of answers produced by a Retrieval-Augmented Generation
    (RAG) system.

Your task is to assess the **Correctness** of the generated answer.

Correctness measures whether the answer is factually accurate and truly answers the
    question, considering the retrieved documents as context.

A correct answer:
- provides accurate information that correctly addresses the question
- does not contradict the retrieved documents
- is internally consistent (no contradictions within itself)
- reflects the intended meaning of the documents and the question
- does not include factual errors, hallucinations, or invented details
```

```
Read the question, the retrieved documents, and the generated answer.
Decide how factually correct the answer is as a response to the question, given the
    evidence in the documents.

Assign a score from 0.0 to 1.0:
- 1.0 = fully correct, factually accurate, and consistent
- 0.0 = completely incorrect or factually wrong

Keep your reasoning brief (max 50 words).

# Question:
{query}

# Retrieved Documents:
{documents}

# Generated Answer:
{generated_answer}

Provide ONLY a JSON object in this format:

{{
    "score": <float between 0 and 1>,
    "reason": "<short explanation (max 50 words)>"
}}

Example:

{{
    "score": 0.9,
    "reason": "The answer correctly reflects the documents but omits one fact
    relevant to the question."
}}
```

## B. Multilingual Results

This section of the appendix contains all results for the languages not reported in the previous sections. They are grouped pairwise (French and German, Spanish and Portuguese).

**Table 5**
Mean $\pm$ std of each metric on the original data and on the two shuffled variants, for French *fr* and German *de*. Every metric satisfies $\Delta \approx 0.9 > 0$ in the original data configuration while in the shuffled ones the results are near to 0.

| | $O$ | | $S_d$ | | $S_a$ | |
|---|---|---|---|---|---|---|
| *Metric* | *fr* | *de* | *fr* | *de* | *fr* | *de* |
| Relevancy | $0.90 \pm 0.22$ | $0.92 \pm 0.19$ | $0.90 \pm 0.22$ | $0.92 \pm 0.19$ | $0.01 \pm 0.09$ | $0.02 \pm 0.10$ |
| Groundedness | $0.87 \pm 0.30$ | $0.92 \pm 0.24$ | $0.01 \pm 0.07$ | $0.01 \pm 0.08$ | $0.00 \pm 0.05$ | $0.00 \pm 0.06$ |
| Correctness | $0.86 \pm 0.31$ | $0.91 \pm 0.25$ | $0.11 \pm 0.31$ | $0.01 \pm 0.08$ | $0.00 \pm 0.06$ | $0.00 \pm 0.06$ |

**Table 6**
Pearson and Spearman correlation among the LLM-judges on the *original* French and German data, over the three metrics: Relevancy ($\mathcal{R}$), Groundedness ($\mathcal{G}$), and Correctness ($\mathcal{C}$). In both languages the strongest off-diagonal association is Groundedness–Correctness (bold), indicating that Correctness largely re-measures Groundedness already on uncorrupted data; Groundedness and Correctness are near-collinear, while Relevancy is largely independent.

| | **French** | | | | | | **German** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Pearson** | | | **Spearman** | | | **Pearson** | | | **Spearman** | | |
| | $\mathcal{R}$ | $\mathcal{G}$ | $\mathcal{C}$ | $\mathcal{R}$ | $\mathcal{G}$ | $\mathcal{C}$ | $\mathcal{R}$ | $\mathcal{G}$ | $\mathcal{C}$ | $\mathcal{R}$ | $\mathcal{G}$ | $\mathcal{C}$ |
| $\mathcal{R}$ | 1.00 | 0.20 | 0.19 | 1.00 | 0.15 | 0.15 | 1.00 | 0.30 | 0.30 | 1.00 | 0.24 | 0.24 |
| $\mathcal{G}$ | 0.20 | 1.00 | **0.93** | 0.15 | 1.00 | **0.91** | 0.30 | 1.00 | **0.91** | 0.24 | 1.00 | **0.90** |
| $\mathcal{C}$ | 0.19 | **0.93** | 1.00 | 0.15 | **0.91** | 1.00 | 0.30 | **0.91** | 1.00 | 0.24 | **0.90** | 1.00 |

**Table 7**
Pearson (left) and Spearman (right) correlation on the *shuffled* French data over the five series: $\mathcal{R}$, $\mathcal{G}_d$, $\mathcal{G}_a$, $\mathcal{C}_d$, and $\mathcal{C}_a$. The strongest redundancy is between Groundedness and Correctness when the answer is corrupted ($\mathcal{G}_a$–$\mathcal{C}_a$, bold: $1.00/1.00$). Matched corruption does not reach high values for the docs-corrupted variants, indicating a weaker overlap than in the Italian setting.

| | **Pearson** | | | | | **Spearman** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{R}$ | $\mathcal{G}_d$ | $\mathcal{G}_a$ | $\mathcal{C}_d$ | $\mathcal{C}_a$ | $\mathcal{R}$ | $\mathcal{G}_d$ | $\mathcal{G}_a$ | $\mathcal{C}_d$ | $\mathcal{C}_a$ |
| $\mathcal{R}$ | 1.00 | 0.44 | 0.62 | 0.12 | 0.63 | 1.00 | 0.30 | 0.37 | 0.12 | 0.37 |
| $\mathcal{G}_d$ | 0.44 | 1.00 | 0.74 | 0.21 | 0.73 | 0.30 | 1.00 | 0.81 | 0.21 | 0.81 |
| $\mathcal{G}_a$ | 0.62 | 0.74 | 1.00 | 0.16 | **1.00** | 0.37 | 0.81 | 1.00 | 0.17 | **1.00** |
| $\mathcal{C}_d$ | 0.12 | 0.21 | 0.16 | 1.00 | 0.16 | 0.12 | 0.21 | 0.17 | 1.00 | 0.17 |
| $\mathcal{C}_a$ | 0.63 | 0.73 | **1.00** | 0.16 | 1.00 | 0.37 | 0.81 | **1.00** | 0.17 | 1.00 |

**Table 8**
Pearson (left) and Spearman (right) correlation on the *shuffled* German data over the five series: $\mathcal{R}$, $\mathcal{G}_d$, $\mathcal{G}_a$, $\mathcal{C}_d$, and $\mathcal{C}_a$. The strongest redundancy is between Groundedness and Correctness when the answer is corrupted ($\mathcal{G}_a$–$\mathcal{C}_a$, bold: $1.00/1.00$) and when the documents are corrupted ($\mathcal{G}_d$–$\mathcal{C}_d$, bold: $1.00/1.00$) . Matched corruption does not reduce the overlap as much as in French, indicating that German preserves a stronger alignment between the two judge families under corruption.

| | **Pearson** | | | | | **Spearman** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{R}$ | $\mathcal{G}_d$ | $\mathcal{G}_a$ | $\mathcal{C}_d$ | $\mathcal{C}_a$ | $\mathcal{R}$ | $\mathcal{G}_d$ | $\mathcal{G}_a$ | $\mathcal{C}_d$ | $\mathcal{C}_a$ |
| $\mathcal{R}$ | 1.00 | 0.44 | 0.54 | 0.44 | 0.54 | 1.00 | 0.29 | 0.36 | 0.29 | 0.36 |
| $\mathcal{G}_d$ | 0.44 | 1.00 | 0.82 | **1.00** | 0.82 | 0.29 | 1.00 | 0.82 | **1.00** | 0.82 |
| $\mathcal{G}_a$ | 0.54 | 0.82 | 1.00 | 0.82 | **1.00** | 0.36 | 0.82 | 1.00 | 0.82 | **1.00** |
| $\mathcal{C}_d$ | 0.44 | **1.00** | 0.82 | 1.00 | 0.82 | 0.29 | **1.00** | 0.82 | 1.00 | 0.82 |
| $\mathcal{C}_a$ | 0.54 | 0.82 | **1.00** | 0.82 | 1.00 | 0.36 | 0.82 | **1.00** | 0.82 | 1.00 |

**Table 9**
Mean $\pm$ std of each metric on the original data and on the two shuffled variants, for Spanish *es* and Portuguese *pt*. Every metric satisfies $\Delta \approx 0.9 > 0$ in the original data configuration while in the shuffled ones the results are near to 0.

| | $O$ | | $S_d$ | | $S_a$ | |
|---|---|---|---|---|---|---|
| *Metric* | *es* | *pt* | *es* | *pt* | *es* | *pt* |
| Relevancy | $0.93 \pm 0.18$ | $0.92 \pm 0.19$ | $0.93 \pm 0.18$ | $0.92 \pm 0.19$ | $0.01 \pm 0.10$ | $0.01 \pm 0.10$ |
| Groundedness | $0.92 \pm 0.25$ | $0.93 \pm 0.22$ | $0.01 \pm 0.07$ | $0.01 \pm 0.07$ | $0.01 \pm 0.07$ | $0.01 \pm 0.07$ |
| Correctness | $0.94 \pm 0.21$ | $0.93 \pm 0.22$ | $0.12 \pm 0.33$ | $0.03 \pm 0.17$ | $0.00 \pm 0.06$ | $0.00 \pm 0.06$ |

**Table 10**
Pearson and Spearman correlation among the LLM-judges on the *original* Spanish and Portuguese data, over the three metrics: Relevancy ($\mathcal{R}$), Groundedness ($\mathcal{G}$), and Correctness ($\mathcal{C}$). In both languages the strongest off-diagonal association is Groundedness–Correctness (bold), indicating that Correctness re-measures Groundedness already on uncorrupted data.

| | **Spanish** | | | | | | **Portuguese** | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Pearson** | | | **Spearman** | | | **Pearson** | | | **Spearman** | | |
| | $\mathcal{R}$ | $\mathcal{G}$ | $\mathcal{C}$ | $\mathcal{R}$ | $\mathcal{G}$ | $\mathcal{C}$ | $\mathcal{R}$ | $\mathcal{G}$ | $\mathcal{C}$ | $\mathcal{R}$ | $\mathcal{G}$ | $\mathcal{C}$ |
| $\mathcal{R}$ | 1.00 | 0.29 | 0.36 | 1.00 | 0.22 | 0.26 | 1.00 | 0.31 | 0.37 | 1.00 | 0.21 | 0.23 |
| $\mathcal{G}$ | 0.29 | 1.00 | **0.71** | 0.22 | 1.00 | **0.57** | 0.31 | 1.00 | **0.90** | 0.21 | 1.00 | **0.92** |
| $\mathcal{C}$ | 0.36 | **0.71** | 1.00 | 0.26 | **0.57** | 1.00 | 0.37 | **0.90** | 1.00 | 0.23 | **0.92** | 1.00 |

**Table 11**
Pearson (left) and Spearman (right) correlation on the *shuffled* Spanish data over the five series: $\mathcal{R}$, $\mathcal{G}_d$, $\mathcal{G}_a$, $\mathcal{C}_d$, and $\mathcal{C}_a$. There is redundancy between Groundedness and Correctness when the answer is corrupted ($\mathcal{G}_a$–$\mathcal{C}_a$, bold: $0.79/0.81$); matched-document corruption yields $0.62/0.66$. The strongest redundancy is between Groundedness and Correctness but when the first has corrputed answers and the second has match-corrupted documents. Matched corruption does not reach $1.00$ here, indicating a moderate overlap between the two judge families under corruption.

| | **Pearson** | | | | | **Spearman** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{R}$ | $\mathcal{G}_d$ | $\mathcal{G}_a$ | $\mathcal{C}_d$ | $\mathcal{C}_a$ | $\mathcal{R}$ | $\mathcal{G}_d$ | $\mathcal{G}_a$ | $\mathcal{C}_d$ | $\mathcal{C}_a$ |
| $\mathcal{R}$ | 1.00 | 0.44 | 0.44 | 0.11 | 0.57 | 1.00 | 0.32 | 0.32 | 0.13 | 0.39 |
| $\mathcal{G}_d$ | 0.44 | 1.00 | 0.62 | 0.62 | 0.11 | 0.32 | 1.00 | 0.66 | 0.66 | 0.12 |
| $\mathcal{G}_a$ | 0.44 | 0.62 | 1.00 | **0.79** | **0.79** | 0.32 | 0.66 | 1.00 | **0.81** | **0.81** |
| $\mathcal{C}_d$ | 0.11 | 0.62 | **0.79** | 1.00 | 0.20 | 0.13 | 0.66 | **0.81** | 1.00 | 0.20 |
| $\mathcal{C}_a$ | 0.57 | 0.11 | **0.79** | 0.20 | 1.00 | 0.39 | 0.12 | **0.81** | 0.20 | 1.00 |

**Table 12**
Pearson (left) and Spearman (right) correlation on the *shuffled* Portuguese data over the five series: $\mathcal{R}$, $\mathcal{G}_d$, $\mathcal{G}_a$, $\mathcal{C}_d$, and $\mathcal{C}_a$. There is redundancy between Groundedness and Correctness when the answer is corrupted ($\mathcal{G}_a$–$\mathcal{C}_a$, bold: $0.67/0.70$). The strongest redundancy is between Correctness and Groundedness when the first has corrupted answers and the second has match-corrupted documents. Matched corruption does not reach $1.00$ here, indicating a moderate overlap between the two judge families under corruption.

| | **Pearson** | | | | | **Spearman** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathcal{R}$ | $\mathcal{G}_d$ | $\mathcal{G}_a$ | $\mathcal{C}_d$ | $\mathcal{C}_a$ | $\mathcal{R}$ | $\mathcal{G}_d$ | $\mathcal{G}_a$ | $\mathcal{C}_d$ | $\mathcal{C}_a$ |
| $\mathcal{R}$ | 1.00 | 0.44 | 0.37 | 0.24 | 0.56 | 1.00 | 0.30 | 0.25 | 0.17 | 0.37 |
| $\mathcal{G}_d$ | 0.44 | 1.00 | 0.53 | 0.43 | **0.79** | 0.30 | 1.00 | 0.57 | 0.42 | **0.81** |
| $\mathcal{G}_a$ | 0.37 | 0.53 | 1.00 | 0.22 | **0.67** | 0.25 | 0.57 | 1.00 | 0.24 | **0.70** |
| $\mathcal{C}_d$ | 0.24 | 0.43 | 0.22 | 1.00 | 0.34 | 0.17 | 0.42 | 0.24 | 1.00 | 0.34 |
| $\mathcal{C}_a$ | 0.56 | **0.79** | **0.67** | 0.34 | 1.00 | 0.37 | **0.81** | **0.70** | 0.34 | 1.00 |