

Harvesting Trust Indicators: A Method for Evaluating Trustworthiness of RAG-Driven Advice

Daan L. DI SCALA ^{a,b,1} and Maaike H.T. DE BOER ^a

^a*TNO, Soesterberg, The Netherlands*

^b*Utrecht University, Utrecht, The Netherlands*

Abstract. AI-generated advice is increasingly used in hybrid human-AI settings for both personal and professional decision-making. In high-stakes domains such as agriculture, incorrect guidance can cause significant economic and environmental harm. With the growing adoption of Large Language Models (LLMs) for advisory tasks, it remains unclear whether typical quantitative evaluations accurately reflect trustworthiness. In this work, we propose a method to identify and analyze indicators of trustworthy AI advice, by investigating the relationship between system- and user-based metrics with human-perceived trustworthiness. We apply this method to an agricultural decision-support setting, in which advice is generated through knowledge-driven Retrieval-Augmented Generation (RAG). To support this analysis, we introduce AgriAdviceQA, a benchmark consisting of 100 question-answer-context entries of information and advice questions, derived from 3,000+ agricultural best-practice documents and their metadata. We compare retrieval and generation performance of three different AI system designs (LLM-only, RAG, and metadata-augmented RAG) against AgriAdviceQA's ground truth. A user study is performed with domain experts, who rate Perceived Fluency, Relevance, Simplicity, and Trustworthiness. Results indicate significant correlations between system and user metrics, supported by qualitative insights. These findings demonstrate that our method can meaningfully identify trust indicators, which provides guidance towards evaluating and designing more reliable advice systems.

Keywords. Trustworthy AI, Advice Systems, Large Language Models, Retrieval Augmented Generation, AI Evaluation, Agriculture

1. Introduction

AI systems utilizing Large Language Models (LLMs) are increasingly used as collaborative tools that support human decision making [1,2,3]. Users interact with AI systems to obtain factual information, advice, explanations or recommendations that inform their own judgments in everyday tasks [2,4]. Not just for personal matters [5], but also professionally in domains such as education, finance, health, and agriculture [4,6,7,8]. As these high-stakes domains have serious consequences, evaluating the trustworthiness of AI outputs becomes essential. Users must not only be able to assess whether the answers of AI systems are accurate, but also whether its outputs can be relied upon.

¹Corresponding Author: Daan Di Scala, daan.discala@tno.nl.

Trust in AI systems is therefore necessary for effective hybrid human-AI collaboration. Yet, misplaced trust can lead to significant risks. Insufficient trust (*undertrust*) can lead to underutilization of helpful system capabilities, making users ‘miss out’ [9,10]. Excessive trust (*overtrust*) may however result in overreliance on the system, even with flawed, overconfident or incomplete outputs [10,11,12]. This is particularly challenging with language-based AI systems that use LLMs, because of their dual characteristic: outputs are highly fluent and persuasive, while the underlying ‘black box’ processes are opaque, making evaluation generally only possible based on the generated outputs [1,2].

A promising approach to improve the reliability of LLM answers is through Retrieval-Augmented Generation (RAG), which grounds the model outputs in externally retrieved sources [13,14]. Knowledge-based variants of RAG aim to further enhance factual alignment and transparency by incorporating structured knowledge and making reasoning steps retraceable [15]. While these approaches are effective for information retrieval tasks, where users primarily evaluate outputs based on perceived correctness and completeness, their impact on advice-giving tasks may differ. When providing AI advice, user trust and decision-making are influenced not only by factual accuracy but also by how the advice is communicated.

Despite this, a major focus of evaluation of AI systems is centered on system metrics, such as retrieval accuracy or text similarity scores, in which performance is quantitatively evaluated in relation to ground truth benchmarks [6,16,17]. While these metrics are indispensable, given that inaccurate or misleading outputs highly impair the functional reliability, they provide limited insight into how end users actually perceive and interpret AI outputs. Prior work has shown that user trust can be measured based on characteristics of AI answers such as fluency, relevance, simplicity and consistency [8,15,18]. We therefore define these characteristics as potential *trust indicators*. As these indicators play a role in shaping user perception, assessing them can provide deeper insights into the calibration of user trust.

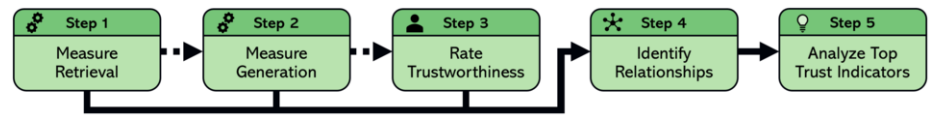


Figure 1. Our proposed evaluation method to analyze the indicators of trustworthy AI advice: evaluate AI advice with quantitative metrics (steps 1 and 2) and human ratings (step 3) to identify correlations (step 4) in order to analyze the right indicators (step 5).

In this work, we aim to identify trust indicators, acknowledging that such indicators are not universal but depend on domain, task, and system. To derive these indicators, we introduce the method shown in Figure 1 and apply its steps throughout the paper. We conduct this research within the agricultural domain, selected as a high-stakes context in which expert knowledge is crucial in decision-making. To facilitate evaluation, we introduce AgriAdviceQA, an agricultural question-answering benchmark, which covers two question types: information lookup and advisory questions. Through quantitative experiments and a controlled user study with agricultural experts, we compare three system designs (LLM-only, RAG, and metadata-augmented RAG). With this, we examine how retrieval quality, generation quality, and user-perceived attributes relate to trust. Specifically, we address the following research question (RQ) and subquestions (SQ):

RQ. Does our evaluation method help identify trust indicators of RAG-driven advice?

SQ1. How do system evaluation metrics for RAG relate to human-perceived trustworthiness of (agricultural) advice?

SQ2. How does perceived user trust differ between (agricultural) information and (agricultural) advice?

By linking system-based evaluation metrics with human trust judgments across task types and system designs, this work contributes to the field of human-centered evaluation of trustworthy AI. All data, code and results are made available in an open source repository for reproducibility purposes².

2. Related Work

Large Language Models (LLMs) are increasingly used for AI-based advice and decision support across domains. However, standalone LLMs are limited by their parametric knowledge: they may hallucinate [3,15,19], rely on outdated information [20,21], or fail to incorporate domain-specific data [14,15]. To address these limitations, recent approaches have shifted from LLM-only answer generation to using Retrieval-Augmented Generation (RAG), augmenting LLMs with external knowledge [13,20,22]. Enhanced RAG approaches aim to further improve retrieval and generation quality by integrating more sophisticated components such as specialized retrievers, re-rankers, additional filters, instructions or structured knowledge [15,23,24,25]. RAG follows a two-stage pipeline: 1) retrieve relevant documents from a database, and 2) generate a response based on the retrieved context. Because of this, RAG systems are typically evaluated along two dimensions: retrieval quality and generation quality.

2.1. RAG Retrieval and Generation Evaluation

Retrieval. The retrieval component is commonly evaluated using traditional Information Retrieval metrics [26]. Standard measures include Hit Rate@ k , Recall@ k , Precision@ k , F1@ k , Mean Reciprocal Rank (MRR@ k) [27], and Normalized Discounted Cumulative Gain (nDCG@ k) [28]. These metrics assess whether relevant documents are retrieved within the top- k results and how highly they are ranked. While these metrics quantify retriever accuracy, they do not necessarily capture the usefulness of retrieved documents. **Generation.** The generation component is typically evaluated using task-specific or reference-based Natural Language Generation metrics. For close-ended tasks such as multiple-choice Question Answering (QA), Accuracy or Exact Match are fitting metrics [6,16,26]. Other widely used metrics include BLEU [29], MAUVE [30], METEOR [31] and ROUGE [32]. These metrics are mainly based on lexical overlap, comparing system outputs to reference texts at the surface level. However, particularly in open-ended QA tasks where multiple valid responses may exist (such as advice generation), it can be challenging for these metrics to meaningfully measure the quality of answers. So, to capture semantic similarity beyond exact word overlap, more recent contextual embeddings based methods such as BERTScore have been introduced [33].

²Repository: <https://github.com/daan-ds/AgriAdviceQA>

Table 1. Related work’s approaches of evaluating RAG-driven advice in different high-stakes domains.

Approach	Domain	Retrieval Metric	Generation Metric	Trust Indicators	Judge Type
AgriAdviceQA (Ours)	Agriculture	Hit, MRR, nDCG	ROUGE-L, METEOR, MAUVE, BERTScore	Fluency, Relevance, Simplicity, Trustworthiness	Human
KRAGGER [8]	Agriculture	CC [34], ILD [35]	✗	Coherence, Fluency, Simplicity, Trustworthiness	Human
Tolegen [36]	Career Guidance	✗	✗	Clarity, Empathy, Faithfulness, Relevance	Both
Swarop et al. [37]	Career Guidance	✗	BLEU, ROUGE, BERTScore	Faithfulness, Hallucination, Relevance	LLM
BSharedRAG [16]	Ecommerce	Hit, nDCG	BLEU-3, ROUGE-L, BERTScore, Accuracy	✗	✗
AstuteRAG-FQA [6]	Finance	Precision, F1	Exact Match	✗	✗
Oro et al. [18]	Finance	✗	BEM [38]	Correctness, Relevance	Both
MEGA-RAG [17]	Health	Accuracy, Precision, Recall, F1	✗	✗	✗
Rakha and Abbas [39]	Health	✗	BERTScore, Cosine Similarity, QNLI [40]	Accuracy, Adequacy, Human-likeness, Risk	Human
MedRAG [7]	Health	Accuracy, Specificity	BERTScore, BLEU, ROUGE, METEOR	Mini-CEX [41,42]	Both
QuALA-NL RAG [43]	Legal	Hit, Precision, Recall	ROUGE-L-SUM, METEOR, MAUVE	Coherence, Consistency, Fluency, Relevance	LLM
FoCOSA RAG [15]	Nutrition	✗	✗	Actionability, Clarity, Completeness, Consistency, DUE [15], Fluency, Hallucinations, Relevance	LLM

2.2. RAG Trust Evaluation

Recent research argues that retrieval and generation should not be evaluated independently, as strong retrieval does not guarantee faithful generation, and vice versa [44]. Beyond generation and retrieval, additional meaningful indicators are needed to holistically assess performance. We refer to these metrics in this paper as *trust indicators*: qualities that semantically describe an answer and serve as a metric to assess the quality (and consequently, the trustworthiness) of AI answers. A promising trust indicator is *Simplicity*, because shorter, simpler explanations are often deemed more understandable [8,45]. *Fluency* measures how natural and well-formed the generated text is [15,43,46]. *Relevance* measures the extent to which the answer appropriately addresses the question [15,18,43]. Finally, *Trustworthiness* can also be rated directly; however, its assessment heavily depends on the subjective interpretation of the evaluator [8,47].

These indicators can be assessed by human evaluators or through LLM-as-judge approaches. Frameworks like G-EVAL [48], TruLens [49], RAGAs [44], and DeepEval [50] support using LLMs as large-scale evaluators. While this approach offers scalability and efficiency, it also introduces risks such as self-evaluation loops and systemic biases [51]. Instead, several studies advocate for RAG evaluation through human assessment, emphasizing that automatic metrics often correlate weakly with human judgments [52, 53]. So, trust indicators may be more reliably evaluated through user studies with domain experts, who can better capture nuanced perceptions of answer quality [7,8].

Table 1 provides an overview of the metrics employed in related work on RAG-driven advice across various high-stakes domains. The studies listed in Table 1 each lack at least one category of evaluation metrics or rely on LLM-as-judge approaches. To

address this gap, our work proposes an evaluation method that incorporates all metric categories, as well as includes human expert evaluators.

3. Method

We define our method to analyze trust indicators (briefly described in Figure 1) as:

RAG Trust Indicator Evaluation Method

- Step 1) Measure Retrieval.** Measure retriever accuracy with suitable metrics.
- Step 2) Measure Generation.** Measure generation capabilities with suitable metrics.
- Step 3) Rate Trustworthiness.** Let domain experts rate answers based on suitable trust indicators. Include *Perceived Trustworthiness* as one of the user metrics.
- Step 4) Identify Relationships.** Measure the correlation between all measured results. Focus on the relationship between different metric types, in particular the match between *Perceived Trustworthiness* and the other metrics.
- Step 5) Analyze Top Trust Indicators.** Select top correlations between metrics, and analyze trust indicators for different settings.

To follow this method step by step, we first create a benchmark (Section 3.1) and select AI answer generation methods and models for comparison (Section 3.2). Afterwards, we select a subset of suitable metrics and trust indicators for Steps 1 through 3 and a suitable correlation metric for Steps 4 and 5 (Section 3.3).

3.1. Dataset and Benchmark

Because no existing benchmark comprehensively evaluates information and advice generation for the agricultural RAG-based advisory task, we introduce the **AgriAdviceQA** benchmark. AgriAdviceQA consists of two parts: a document dataset of 3,385 documents and a QA dataset of 100 question-answer-context entries. The documents for AgriAdviceQA's document dataset are collected from a snapshot³ of the EU-FarmBook platform, a large-scale European database containing agricultural best-practice oriented documents [54,55]. Documents retrieved from the EU-FarmBook platform are filtered on PDF-files written in English. These documents also contain metadata (title, description, keywords, topics, themes, locations, languages). To enable retrieval, all documents are split into chunks of 250 words, with 20 words overlap between chunks to not miss linked information, for a total of 32,565 text chunks.

To create AgriAdviceQA's second part, its benchmark questions, documents are randomly sampled from the AgriAdviceQA dataset. Then, two types of questions are manually created, based on available information within the document: **Information Questions** ("What is..?" / "What is the definition of..?") and **Advice Questions** ("I am farmer/advisor {type X}. Would you recommend me to do..? / How could I..? / Should I..?"). Each benchmark entry consists of a question, a short answer and a relevant snippet of text to be retrieved to properly answer the question. Examples are shown in Table 2. An important advantage of this design is that all questions are grounded in verifiable source material, ensuring that correct answers exists for each item in the ground truth.

³On 01-01-2026, the platform held 3,815 English PDF-file documents. On 01-05-2026, it was 5,642. Since the platform is continuously gaining new entries, a local snapshot of documents is used for this research.

Table 2. Examples AgriAdviceQA Benchmark Question-Answer-Context entries.

	Information	Advice
Question	What types of pests can be controlled with Precision Spraying?	As a sheep farmer, is there anything you could advise me to do to prevent the spread of wildfires?
Answer	aphid (Aphidomorph) and currant borer (Pandemis Ribearia).	Yes, you should identify grazing objectives such as targeted grazing at strategic fire management points (SMPs) to reduce the risk of forest fires.
Context	The sprayer allows two different selective formulations to be applied at the same time, which allows to control two different types of pests simultaneously: aphid (Aphidomorph) and currant borer (Pandemis Ribearia), and significantly reduce the use of chemicals	The overall objective of the project is to reduce the risk of large forest fires through prescribed grazing, meaning grazing with predefined objectives and management aimed at achieving those goals. In general, the project's livestock farmers graze strategic fire management points (SMPs) identified by the Catalan Fire Department. These SMPs, if properly managed, limit the spread of forest fires and provide an opportunity for control operations by altering fire behavior on the day of the fire. Therefore, FireFlocks promotes targeted grazing at these strategic points, connecting, supporting, and informing farmers about existing SMPs in their area

3.2. Answer Generation Approaches and Models

We compare three approaches (as shown in Figure 2) for answer generation: **1) LLM.** This approach feeds all AgriAdviceQA queries into a Large Language Model (LLM). This serves as a non-retrieval baseline, using the LLM’s internal knowledge. **2) RAG.** This approach employs a standard Retrieval-Augmented Generation (RAG) pipeline, which consists of embedding all queries and document chunks of the AgriAdviceQA dataset into a vector space, retrieving the most relevant chunks for each query based on semantic similarity, before feeding them into an LLM to generate an answer. **3) Metadata RAG (MRAG).** In this approach, following [43], metadata of each document is appended to each of the document’s chunks before embedding, to create richer information. Metadata is converted into plain text, for example: “**Title:** Biosecurity regulation for Italian poultry production. **Description:** This factsheet explains to poultry farmers the key points of the Italian regulation for poultry production. **Keywords:** biosecurity, regulation, Italy, poultry. **Topics:** Livestock. **Themes:** Pest control. **Locations:** Italy”.

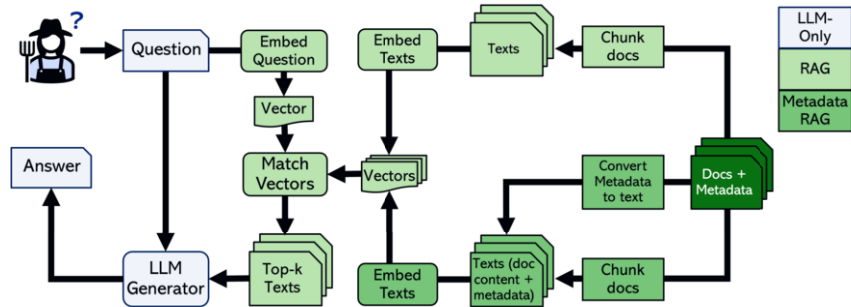


Figure 2. Three system designs for answer generation: LLM-only, RAG, and Metadata RAG (MRAG).

For retrieval and generation we select state-of-the-art language models. For retrieval, we utilize OpenAI’s GPT embedding model TEXT-EMBEDDING-3-SMALL [56], a top scoring model in retrieval tasks [57]. For generation, we employ OpenAI’s most recent model at the time, GPT-5.2 [58]. We set the temperature to 0.3 to encourage more deterministic generation and reduce output variability [8]. We post-process all generated text by removing any mark-up symbols (# and *), for the user study’s clarity sake. We keep the instructions to a minimum, with the same system prompt for all methods: “You are a

helpful agricultural information assistant.”. The user prompt for the LLM-only method is “Answer the following question: **{question}** Answer:” and for RAG and MRAG it is “Use *ONLY* the context below to answer the question. If the answer cannot be found, say “I don’t know”. Context: **{context}** Question: **{question}** Answer:”.

3.3. Metrics and Experiment Set-Up

The answer generation methods are evaluated in three distinct ways: **1) Retrieval Metrics.** The retrieval capabilities of each approach are evaluated with Hit Rate@ k , MRR@ k and nDCG@ k for the top- k retrieved documents. The LLM-only approach is left out of this evaluation, as no retrieval takes place. **2) Generation Metrics.** The generation capabilities of all three approaches are evaluated with BERTScore, MAUVE, METEOR, and ROUGE-L. We additionally report the average amount of words per answer (#WORDS), which we also measure for ground truth answers. **3) User Study Metrics.** The user study evaluates answers on four metrics: Perceived Simplicity, Perceived Fluency, Perceived Relevance and Perceived Trustworthiness. These three types of metrics are selected as most suitable, based on our analysis of related work (Table 1).

After we report the results of these metrics individually, we analyze the correlation between the system- and user-oriented metrics. This is assessed using Spearman’s Rank Correlation Coefficient (ρ) [18,59], a non-parametric measure of statistical dependence between two ranked variables, which determines the strength and direction of their relationship. This measure allows for the analysis of both ordinal variables (user study ratings) and continuous variables that have been converted into ranks (ranked generation and retrieval data). The coefficient ranges $\rho \in [-1, 1]$, with $\rho = 1$ and $\rho = -1$ meaning perfect positive and negative correlations respectively, and $\rho = 0$ indicates no correlation.

The user study consists of a survey of sampled question-answer pairs with answers generated by the three system designs (Figure 2) as well as AgriAdviceQA’s ground truth answers. The survey is built and distributed using Survalyzer (Survalyzer BV), an online, secure cloud-based, survey tool, which also stores the collected data. Participants are asked to rate 24 answers on the four trustworthiness metrics on a 7-points Likert scale. For diversity, each participant is shown three groups of four answers to the same randomly sampled question, consisting of the LLM, RAG, MRAG and ground truth (GT) answers, for a total of $24 \times 4 = 96$ questions per participant. To ensure quality of ratings, the survey is shared with experts in the agricultural domain, and each participant is asked to self-rate their confidence on their knowledge and expertise within the agricultural domain. At the end of the survey, users are asked an open question to discuss their experience, including the targeted question “What do you think makes for good advice?”. All code and the full survey, including instructions and questions, is in the repository².

4. Results

4.1. Retrieval and Generation Results

For [Step 1](#) of our evaluation method, the retrieval results are shown in Table 3. MRAG overall outperforms RAG on information questions for each metric, whereas RAG scores better on all metrics on advice. Additionally, out of 100 questions, the RAG approach

answers “*I don’t know*” 15 times, whereas the MRAG approach yields the same answer only 10 times. This indicates that overall, the documents found by MRAG’s retrieval method are most often deemed fit to answer the question by the generator component.

Table 3. Retrieval results with top-*k*@3. Best scores per column are marked green.

Method	Hit Rate		MRR		nDCG	
	Info	Advice	Info	Advice	Info	Advice
RAG	0.72	0.72	0.60	0.64	0.63	0.66
MRAG	0.78	0.68	0.73	0.62	0.75	0.64

For **Step 2**, the generation results are shown in Table 4. On most metrics, RAG and MRAG outperform LLM. Low syntactical scores (*ROUGE-L*, *METEOR*) indicate little overlap of words. However, high *BERTScores* are achieved, which signals a high semantic connection. Furthermore, *#WORDS* shows that the ground truth (GT) has on average much fewer words per answer. The LLM-only approach produced the longest answers, with advice reaching over 10× the length of GT answers.

Table 4. Generation results compared to ground truth (GT). Best scores per column are marked green.

Method	BERTScore		MAUVE		METEOR		ROUGE-L		#WORDS	
	Info	Advice	Info	Advice	Info	Advice	Info	Advice	Info	Advice
LLM	0.85	0.83	0.64	0.15	0.19	0.13	0.11	0.06	80.38	239.96
RAG	0.88	0.87	0.57	0.34	0.36	0.29	0.27	0.16	29.66	90.06
MRAG	0.88	0.87	0.61	0.81	0.37	0.27	0.27	0.16	38.74	81.68
GT	-	-	-	-	-	-	-	-	17.38	18.44

4.2. User Study Results

A total of *n* = 51 voluntary agricultural experts participated in the user study, who on average rated themselves 5.08 out of 7 on their knowledge and expertise in the agricultural domain. Answers about their role within the domain show a diverse set of experts, a subset: agricultural consultants (e.g., consultant digital innovation in agriculture, behavioral consultant in agritech adoption), researchers at agricultural organizations, an agronomist, a forester, people working in organic farming, gentleman farmers and agricultural policy makers. Due to incomplete results, 16 entries were filtered out, and results from the remaining 35 are reported in the following section.

Table 5. User ratings on perceived trustworthiness metrics on 7-points Likert scale (1=Very Not and 7 = Very). Best scores per column are marked green.

Method	Fluency		Simplicity		Relevance		Trustworthiness	
	Info	Advice	Info	Advice	Info	Advice	Info	Advice
LLM	5.90	5.52	4.95	3.66	5.67	5.97	5.28	5.49
RAG	5.76	5.53	5.66	4.36	5.29	5.39	4.54	5.14
MRAG	5.86	5.89	5.57	5.15	4.90	5.07	4.50	4.83
GT	5.27	5.59	5.82	6.00	4.61	3.71	3.92	3.39

Following **Step 3**, the user study results are shown in Table 5. Overall, the LLM-only approach is deemed most *Relevant*, most *Trustworthy*, and least *Simple*. This is reversed

for GT. On average, all methods are deemed *Slightly Fluent - Fluent*, with little differences. The contrast between methods is larger in the advice setting than with information answers. For information, all average scores are mostly positive, with the exception of GT answers being rated slightly below neutrally trustworthy (3.92 ± 1.61). For advice however, LLM scores towards slightly complex (3.66 ± 1.66), and GT scores slightly irrelevant (3.71 ± 1.77) and slightly untrustworthy (3.39 ± 1.50).

4.3. Comparative Results

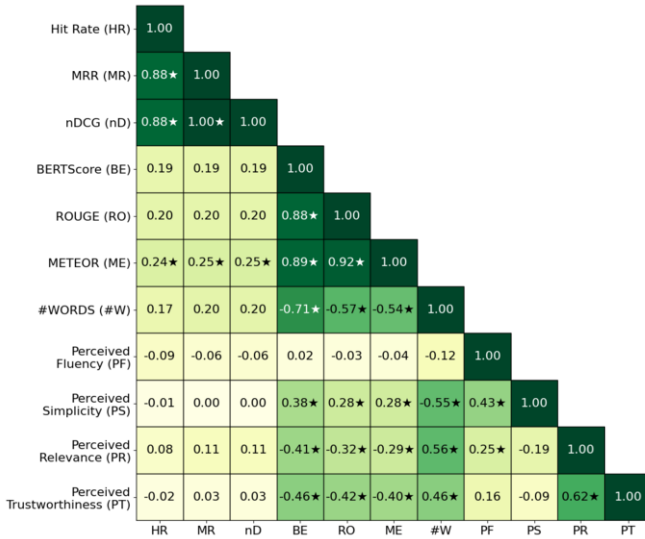


Figure 3. Spearman ρ correlation matrix between metrics for all answers (averaged over LLM, RAG, MRAG, and GT answers) to AgriAdviceQA's Advice questions. Significant results ($p < 0.0001$) are marked with ★.

For **Step 4**, comparative results are reported in Figure 3. Within the system metric categories, there are significantly high correlations. For example, with retriever metrics $\rho(\text{Hit Rate}, \text{MRR}) = 0.88$ with $p < 10^{-25}$ for both RAG and MRAG, as well as with generation metrics $\rho(\text{BERTScore}, \text{ROUGE}) = 0.89$ with $p < 10^{-8}$ for LLM, RAG and MRAG. This effect is less prevalent between user metrics, as correlations are not consistently strong or significant. The highest significant correlation are between Fluency and Simplicity ($\rho = 0.43, p < 10^{-5}$) and between Relevance and Trustworthiness ($\rho = 0.62, p < 10^{-4}$). Further, when analyzing relationships between metric categories, weak to moderate correlations are found. Between retrieval and generation, only METEOR shows slight (significant) correlations with the other metrics. Between retrieval and user metrics, no relationships are found. Between generation and user metrics, Fluency is the only metric that does not show any correlation. #WORDS has the strongest correlation, with Simplicity, Relevance and Trustworthiness. Besides #WORDS, the strongest relation for Perceived Trustworthiness is BERTScore.

For the final step (**Step 5**), we analyze the most promising indicators, based on the top average correlations reported in Figure 3. These metric pairs are shown in Figure 4. In Figure 4a, RAG and MRAG show a strong fit between BERTScore and ROUGE. Next to that, the distribution of answers for LLM reaches lower scores than either RAG

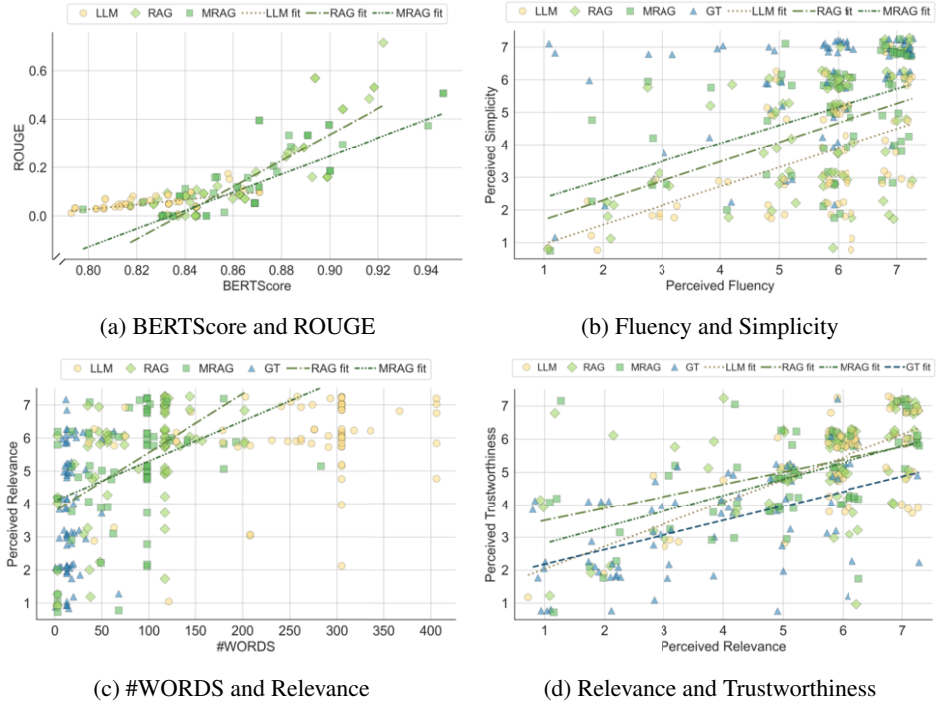


Figure 4. Sampled top correlating metrics. Regression line is shown for significant ($p < 0.0001$) correlations.

or MRAG. Figure 4b shows that all fits between *Fluency* and *Simplicity* of answers are equally strong, with the exception of the ground truth, which has no significant correlation. It also shows that overall, GT answers are deemed the most simple, with little exceptions. Figure 4c highlights the answer length only strongly correlates for RAG and MRAG with how *Relevance* is perceived by users, which is not the case for LLMs or GT. Figure 4d shows that for all methods except LLM-only, there is a significant relation between how much answers are deemed *Relevant* and *Trustworthy*. Additionally, while Figure 3 shows a promising average $\rho(\text{Trustworthiness}, \text{BERTScore})$ correlation, when calculated for LLM, RAG, and MRAG separately, no statistical significance is found.

4.4. User Study Qualitative Results

Finally, we analyze the collected answers to the open question at the end of the survey. Across participants, several recurring topics emerged about what good advice (from AI) should encompass. Many commented on appropriate depth, personalization and clarity.

Depth. Multiple participants emphasized their preference for mid-length answers, as one-sentence answers felt untrustworthy, while excessively long answers were called overwhelming (“*getting lost in endless additional information*”). One participant highlighted that “*good advice goes to the point (not too many details) but gives some explanation of why*”. Including an explanation as to why was mentioned multiple times as a key factor of trustworthy advice. Others commented that linking to or directly including the original sources would elevate the trustworthiness of the answer. One par-

ticipant mentioned “*tuning of the tone of voice preferred by different users*” to have a significant impact on usefulness. Others discussed the adaptability of the AI (“*it should know how much detail the user wants*”), tailoring the answer’s level of detail to the user’s expertise level. **Clarity.** On the topic of clarity, bullet lists and provided examples were deemed additionally trustworthy. Participants also valued transparency about uncertainty (“*if there are assumptions to be made, mention those explicitly*”), however disliked if the AI repeatedly answered “*I don’t know*”. Finally, one participant added that some advice can never be fully trustworthy, because the AI lacks real-life experience.

5. Discussion

The results have shown RAG yielding higher retrieval scores than MRAG for advice, indicating that the inclusion of metadata did not provide additional benefit for retrieval performance. This may be because by enriching the text chunks, they are also lengthened, with a larger likelihood of noise and mismatching embeddings. However, MRAG did result in fewer “*I don’t know*” responses, and consistently achieved high scores on generation quality and user evaluations. In addition, both RAG and MRAG outperformed LLM on the generation metrics. This aligns with expectations: providing relevant context constrains the model and helps remain closer to human answers, whereas the LLM-only baseline produced substantially longer, complex and less focused responses.

The strong effects found between #WORDS and *Relevance* (4c) and between *Relevance* and *Trustworthiness* (4d) suggest an intuitive pattern: longer advice tends to be judged as more relevant, and more relevant advice tends to be trusted. This raises important considerations, especially since LLM-only answers were rated as both the most complex and the most relevant (Table 5). There is a risk that overly complex LLM outputs lead to users blindly following the advice, a concern also reflected in participants’ remarks about wanting advice with explanations that remain clear and easy to follow. This highlights the need to consciously design AI systems for human-AI collaboration which avoids unnecessary complexity to support users in making well-informed decisions.

Regarding the design of our method, as highlighted in Section 2, it is essential that trustworthiness assessments are made by users rather than by LLMs, as only humans can meaningfully judge what they themselves perceive as trustworthy. Evaluating *with* users therefore ensures that we design *for* users. We therefore stress the importance of continually asking domain experts what makes advice trustworthy to them, and comparing these perceptions with what can be measured through system-level metrics.

Nonetheless, our work was not without limitations. First, we restricted our study to a single model for both generation and embedding. We intentionally used a lightweight embedding model (TEXT-EMBEDDING-3-SMALL) to prioritize energy efficiency, as achieving the highest possible performance was not the aim of this work (especially given the rapid, continuous evolution of LLM capabilities). Second, although we selected metrics suitable for our use case, alternative or additional retrieval, generation and user metrics could have been chosen. Third, our benchmark deliberately targeted a single domain, with questions requiring agricultural expertise to assess, but with relatively short answers (both for information and advice). Participant feedback indicated that some of these manually constructed ground truth answers were perceived as overly simple or lacking sufficient explanations or sources. This likely influenced participants’ perceptions of trustworthiness and may have affected both the results and their correlations.

6. Conclusion and Future Work

We have introduced a 5-step evaluation Retrieval-Augmented Generation (RAG) trustworthiness method, to identify and analyze trust indicators of RAG-driven advice. This method is based on investigating the relationship between system-, and user-based evaluation metrics with human-perceived trustworthiness. We have applied it to an agricultural decision-support setting with AI-generated information and advice answers. To support our analysis, we have introduced the open-source² AgriAdviceQA benchmark, which consists of 100 question-answer-context entries, derived from 3,000+ agricultural best-practice documents and their metadata. We have compared retrieval and generation performance of three different AI system designs (LLM-only, RAG, and Metadata RAG) against AgriAdviceQA's ground truth, and performed a user study with domain experts, who rated Perceived Fluency, Simplicity, Relevance and Trustworthiness. Subsequently, we have compared the correlations between metrics to identify top trust indicators.

This brings us to answer our research questions: First, we were able to identify relationships between system metrics and human-perceived trustworthiness (**SQ1**), as shown in Figure 3. For this use case, most relations emerged within metric categories, with limited correlations between *Perceived Trustworthiness* and other metrics. Second, while retrieval and generation scores were comparable across information and advice questions, user evaluations were significantly more critical for advice (**SQ2**), indicating that it is more challenging to generate advice that is perceived as trustworthy. Finally, regarding the main research question (**RQ**), our method successfully identified effective trust indicators, while also revealing which connections were weak or absent. Supported by qualitative insights, we can conclude best AI advice evaluation practices for this particular high-stakes domain. Based on these findings, we argue that our method offers a valuable framework for understanding which characteristics make RAG-based advice appear trustworthy, thereby providing practical guidance for both system design and evaluation.

Future work could apply this method to other advice settings, to gather and assess trust indicators across multiple high-stakes domains. Related, a useful next step would be to create a structured overview of potential trust indicators and related metrics across domains. Further extensions include testing on a wider range of LLMs (including open-source, multi-lingual or smaller models) and exploring alternative ways of enriching (M)RAG systems through refined knowledge-augmentation strategies. Our benchmark could also be expanded with a variety of answers that include more in-depth explanations. With this work, we believe to have made a meaningful contribution towards evaluating and designing more reliable AI advice systems.

Acknowledgements

This work has been partially funded by the HORIZON EUROPE project 'EU-FarmBook: supporting knowledge exchange between all AKIS actors in the European Union' (Grant ID: 101060382). We would like to thank EU-FarmBook consortium members for their help in distributing the survey and our colleagues Julia Bronkhorst, Stephan Raaijmakers, Lucia Tealdi, Daan Vos and Mike Wilmer for assisting with the dataset creation.

References

- [1] Vodrahalli K, Daneshjou R, Gerstenberg T, Zou J. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In: *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*; 2022. p. 763-77.
- [2] Burton JW, Lopez-Lopez E, Hechtlinger S, Rahwan Z, Aeschbach S, Bakker MA, et al. How large language models can reshape collective intelligence. *Nature human behaviour*. 2024;8(9):1643-55.
- [3] Mala CS, di Maio C, Proietti M, Gezici G, Giannotti F, Melacci S, et al. Towards Building a Trustworthy RAG-Based Chatbot for the Italian Public Administration. In: *HHAI 2025*. IOS Press; 2025. p. 196-204.
- [4] Baines JJ, Dalal RS, Ponce LP, Tsai HC. Advice from artificial intelligence: a review and practical implications. *Frontiers in Psychology*. 2024;15:1390182.
- [5] Silva I, Said A, Marinho LB, Willemsen MC. Leveraging Large Language Models for Recommendation and Explanation. In: *IntRS@ RecSys*; 2023. p. 74-81.
- [6] Alam MZ, Zaman KAU, Miraz MH. AstuteRAG-FQA: Task-Aware Retrieval-Augmented Generation Framework for Proprietary Data Challenges in Financial Question Answering. *arXiv preprint arXiv:251027537*. 2025.
- [7] Zhao X, Liu S, Yang SY, Miao C. MedRAG: Enhancing Retrieval-Augmented Generation with Knowledge Graph-Elicited Reasoning for Healthcare Copilot. In: *Proceedings of the ACM on Web Conference 2025*; 2025. p. 4442-57.
- [8] Di Scala DL, de Boer MHT. Let Me Explain - Knowledge-Based Retrieval Augmented Generation for Agricultural Recommendation Explanations. In: *European Conference on Information Retrieval*. Springer; 2026. p. 122-38.
- [9] Parasuraman R, Riley V. Humans and automation: Use, misuse, disuse, abuse. *Human factors*. 1997;39(2):230-53.
- [10] Rosenbacke R, Melhus Å, McKee M, Stuckler D. How explainable artificial intelligence can increase or decrease clinicians' trust in AI applications in health care: systematic review. *Jmir Ai*. 2024;3:e53207.
- [11] Keding C, Meissner P. Managerial overreliance on AI-augmented decision-making processes: How the use of AI-based advisory systems shapes choice behavior in R&D investment decisions. *Technological Forecasting and Social Change*. 2021;171:120970.
- [12] Shekar S, Pataranutaporn P, Sarabu C, Cecchi GA, Maes P. People overtrust AI-generated medical advice despite low accuracy. *NEJM AI*. 2025;2(6):AIoa2300015.
- [13] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*. 2020;33:9459-74.
- [14] Purwar A, et al. Evaluating the efficacy of open-source LLMs in enterprise-specific RAG systems: A comparative study of performance and scalability. *arXiv preprint arXiv:240611424*. 2024.
- [15] Bagozi A, Bianchini D, Garda M, Melchiori M, Rula A. Ontology-enhanced RAG for a personalised and sustainable food advisory system. *Data & Knowledge Engineering*. 2026:102579.
- [16] Guan K, Cao Q, Sun Y, Wang X, Song R. Bsharedrag: Backbone shared retrieval-augmented generation for the e-commerce domain. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*; 2024. p. 1137-58.
- [17] Xu S, Yan Z, Dai C, Wu F. MEGA-RAG: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health. *Frontiers in Public Health*. 2025;13:1635381.
- [18] Oro E, Granata FM, Lanza A, Bachir A, De Grandis L, Ruffolo M. Evaluating Retrieval-Augmented Generation for Question Answering with Large Language Models. In: *Ital-IA*; 2024. p. 12-7.
- [19] Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM computing surveys*. 2023;55(12):1-38.
- [20] Han S, Wang M, Zhang J, Li D, Duan J. A review of large language models: Fundamental architectures, key technological evolutions, interdisciplinary technologies integration, optimization and compression techniques, applications, and challenges. *Electronics*. 2024;13(24):5040.
- [21] Mannuru NR, Mannuru A, Lund B. Large language models (LLMs) as a tool to facilitate information seeking behavior. *InfoScience Trends*. 2024;1(2):34-42.
- [22] Sun J, Zhong X, Zhou S, Han J. Dynamicrag: Leveraging outputs of large language model as feedback for dynamic reranking in retrieval-augmented generation. *arXiv preprint arXiv:250507233*. 2025.
- [23] Di Scala DL, de Boer MHT. "I Don't Know" – Towards Appropriate Trust with Certainty-Aware Retrieval Augmented Generation. In: *International Workshop on Value Engineering in AI*; 2025. .

- [24] Wolf K. Evaluating Retrieval-Augmented Generation Variants for Clinical Decision Support: Hallucination Mitigation and Secure On-Premises Deployment. *Electronics*. 2025;14(21):4227.
- [25] Hu J, Dong Y, Huang X. Trust-oriented adaptive guardrails for large language models. *arXiv preprint arXiv:240808959*. 2024.
- [26] Salemi A, Zamani H. Evaluating retrieval quality in retrieval-augmented generation. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*; 2024. p. 2395-400.
- [27] Craswell N. Mean reciprocal rank. *Encyclopedia of database systems*. 2009;1703-3.
- [28] Järvelin K, Kekäläinen J. IR evaluation methods for retrieving highly relevant documents. In: *ACM SIGIR Forum*. vol. 51. ACM New York, NY, USA; 2017. p. 243-50.
- [29] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*; 2002. p. 311-8.
- [30] Pillutla K, Swayamdipta S, Zellers R, Thickstun J, Wellick S, Choi Y, et al. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*. 2021;34:4816-28.
- [31] Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*; 2005. p. 65-72.
- [32] Lin CY. Rouge: A package for automatic evaluation of summaries. In: *Text summarization branches out*; 2004. p. 74-81.
- [33] Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:190409675*. 2019.
- [34] Ge M, Delgado-Battenfeld C, Jannach D. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In: *Proceedings of the fourth ACM conference on Recommender systems*; 2010. p. 257-60.
- [35] Kunaver M, Požrl T. Diversity in recommender systems—A survey. *Knowledge-based systems*. 2017;123:154-62.
- [36] Tolegen A. RAG for AI-Powered Career Guidance in High Schools: A Pilot Study. *NHSJS*. 2025.
- [37] Ganeshkaran M, Hanish K, Hidesh Balaji C, Bindu K, et al. Enhancing Career Guidance for Students Through Retrieval Augmented Generation. In: *2025 International Conference on Inventive Computation Technologies (ICICT)*. IEEE; 2025. p. 1314-9.
- [38] Bulian J, Buck C, Gajewski W, Börschinger B, Schuster T. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*; 2022. p. 291-305.
- [39] Rakha I, Abbas N, et al. Large Language Models in Clinical Advice: Direct Generation and Retrieval Augmented Generation vs Expert Advice. In: *Proceedings of the 6th International Conference on Natural Language Processing and Artificial Intelligence (NLPAI'2025)*, June; 2025. p. 13-6.
- [40] Clark K, Luong MT, Le QV, Manning CD. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:200310555*. 2020.
- [41] Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. *Annals of internal medicine*. 1995;123(10):795-9.
- [42] Shi X, Xu J, Ding J, Pang J, Liu S, Luo S, et al. Llm-mini-cex: Automatic evaluation of large language model for diagnostic conversation. *arXiv preprint arXiv:230807635*. 2023.
- [43] van Drie RAN, Bakker RM, Di Scala DL, de Boer MHT. QuALA-NL: Question & Answer with Legal Attribution in Dutch. In: *Proceedings of the Fifteenth Language Resources and Evaluation Conference*; 2026. .
- [44] Es S, James J, Anke LE, Schockaert S. 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*; 2024. p. 150-8.
- [45] Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*. 2019;267:1-38.
- [46] Lee D, Whang T, Lee C, Lim H. Towards reliable and fluent large language models: Incorporating feedback learning loops in qa systems. *arXiv preprint arXiv:230906384*. 2023.
- [47] Kowald D, Scher S, Pammer-Schindler V, Müllner P, Waxnegger K, Demelius L, et al. Establishing and evaluating trustworthy AI: overview and research challenges. *Frontiers in Big Data*. 2024;7:1467222.

- [48] Liu Y, Iter D, Xu Y, Wang S, Xu R, Zhu C. 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. p. 2511-22.
- [49] The RAG Triad; 2023. [Accessed 01-03-2026]. https://www.trulens.org/getting_started/core_concepts/rag_triad/.
- [50] Ip J, Vongthongsri K. deepeval, The Open-Source LLM Evaluation Framework; 2026. [Accessed 01-03-2026]. <https://github.com/confident-ai/deepeval>.
- [51] Li D, Jiang B, Huang L, Beigi A, Zhao C, Tan Z, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*; 2025. p. 2757-91.
- [52] Nimah I, Fang M, Menkovski V, Pechenizkiy M. NLG evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; 2023. p. 1240-66.
- [53] Elangovan A, Xu L, Ko J, Elyasi M, Liu L, Bodapati S, et al. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and llm-as-a-judge. *arXiv preprint arXiv:241003775*. 2024.
- [54] EU-FarmBook Homepage; 2025. [Accessed 01-03-2026]. <https://eufarmbook.eu/>.
- [55] Panoutsopoulos H, Brewster C, Fountas S. A Semantic Data Model for a FAIR Digital Repository of Heterogeneous Agricultural Digital Objects. In: *2nd Integrated Food Ontology Workshop*; 2021. .
- [56] OpenAI Platform Vector Embedding Models;. [Accessed 01-03-2026]. <https://platform.openai.com/docs/guides/embeddings>.
- [57] Massive Text Embedding Benchmark Embedding Leaderboard; 2025. [Accessed 01-03-2026]. <https://huggingface.co/spaces/mteb/leaderboard/>.
- [58] OpenAI Platform GPT-5.2 Model;. [Accessed 01-03-2026]. <https://platform.openai.com/docs/models/gpt-5.2/>.
- [59] Weaver KF, Morales V, Dunn SL, Godde K, Weaver PF, et al. Pearson's and Spearman's Correlation. *An Introduction to Statistical Analysis in Research*. 2017;435:471.