information

MDPI

Article

Characterizing Human and LLM Responses in Low-Resource Swahili Agricultural Information Retrieval

Theofrida Julius Maginga [1,2]

[1] Department of Informatics and Information Technology, Sokoine University of Agriculture (SUA), Morogoro 3000, Tanzania; theofrida.maginga@sua.ac.tz or theofrida.maginga@ouce.ox.ac.uk
[2] School of Geography and the Environment, University of Oxford, Oxford OX1 3QY, UK

Abstract

Large language models (LLMs) are increasingly used in conversational information retrieval systems to provide direct answers to user queries. However, little is known about how LLM-generated responses behave in low-resource language settings, particularly in domain-specific contexts such as agricultural knowledge access. This study investigates whether answers produced by LLMs differ systematically from human-authored responses in a Swahili agricultural question answering task. Using a dataset of 127 maize-related questions collected through fieldwork in Morogoro, Tanzania, we compare human responses with answers generated by two conversational models, ChatGPT and Gemini. The analysis combines descriptive textual statistics, lexical similarity measures based on TF-IDF cosine similarity, and TF-IDF–based logistic regression evaluated through question-grouped cross-validation. Results show consistent differences in response length, lexical diversity, and lexical similarity, with LLM-generated answers exhibiting longer and more uniform lengths despite instructions to answer briefly. Human answers had higher raw type-token ratios, while model answers showed higher fixed-window lexical diversity. Same-question lexical similarity exceeded a mismatched-question baseline. The three-source classifier achieved 0.9633 accuracy, compared with 0.7192 using length alone and 0.8058 after matching answer lengths. These findings highlight detectable patterns associated with both verbosity and source-specific wording, providing insights for evaluating conversational information retrieval systems in low-resource agricultural environments.

Keywords: large language models (LLMs); information retrieval; low-resource languages; Swahili; agricultural knowledge systems; text detectability; conversational AI

Check for updates

Academic Editors: Mohamed Hamroun and Damien Sauveron

Received: 22 September 2026
Revised: 20 October 2026
Accepted: 26 October 2026
Published: 4 November 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The deployment of information retrieval (IR) systems in agriculture remains particularly challenging in low-resource environments, where limited access to expert knowledge, data resources, and digital infrastructure constrains effective information access [1]. In countries such as Tanzania, agriculture remains highly vulnerable to climate variability, soil degradation, and environmental stressors, which collectively contribute to food insecurity and limit farmer livelihoods [2,3]. Traditional agricultural knowledge, while valuable, is increasingly insufficient to address the growing unpredictability of weather patterns and the changing dynamics of soil fertility driven by climate change [4,5]. These challenges are particularly acute for smallholder farmers cultivating staple crops such as maize, which plays a central role in food security and household income across Tanzania [6]. Limited access to extension services, agronomic expertise, and timely information further constrains

productivity and farmers' adaptive capacity [7]. In such contexts, information retrieval systems capable of delivering relevant, localized, and comprehensible agricultural information in local languages are increasingly viewed as a critical intervention for improving agricultural knowledge access and decision-making [8].

Recent advances in artificial intelligence (AI), particularly large language models (LLMs), have renewed interest in conversational interfaces as IR tools for agriculture. AI-powered assistants have been proposed as scalable mechanisms for providing farmers with on-demand access to agricultural advice, thereby complementing traditional extension services [9]. Digital technologies have already demonstrated potential in supporting smart farming practices in developed and emerging economies, with more limited but growing adoption in developing countries [10,11]. However, despite this potential, there remains limited understanding of how LLM-generated answers behave as information retrieval outputs in Swahili agricultural question answering. In particular, it is unclear whether these answers resemble human-authored agricultural responses or whether they exhibit systematic linguistic patterns that distinguish them from locally produced knowledge [12]. Examining these differences can support the evaluation of answers' provenance in conversation IR by establishing whether the human and model-generated responses can be distinguished from their wording.

Low-resource environments are characterized by limited datasets for benchmarking information retrieval (IR) techniques and restricted computing infrastructure [13]. In agricultural contexts, these limitations affect the ability of farmers and practitioners to access timely and reliable agronomic information. While recent work suggests that AI assistants may help bridge gaps in access to expert knowledge and localized agricultural guidance, the increasing integration of large language models (LLMs) into conversational IR systems introduces new evaluation challenges, particularly in settings where expert validation remains scarce [1]. Consequently, understanding how generative responses differ from human-authored agricultural explanations is an important methodological step toward establishing reliable evaluation frameworks for AI-generated information access in low-resource environments. To address this challenge, this study investigates Swahili maize-related question-answering through a comparison of human, ChatGPT, and Gemini responses. Using a dataset collected through fieldwork in Morogoro, Tanzania, the study examines whether LLM-generated answers exhibit detectable differences from human-authored responses. Because response length can influence both the lexical-diversity measure and source classification, the analysis includes length-controlled comparisons alongside the original full-length answers. This approach follows established vector-space representations, such as TF-IDF and cosine similarity, as highlighted by [14,15], while drawing on recent work that uses stylistic and lexical signals to detect machine-generated text [16,17]. Through this approach, the study contributes evidence on the detectability of LLM-generated Swahili agricultural answers and the extent to which observed differences persist.

The remainder of this manuscript is organized as follows. Section 2 reviews the related literature on information retrieval in agriculture, low-resource IR environments, and the use of large language models in conversational information systems. Section 3 presents the methodological framework of the study, including the dataset construction, data preparation procedures, and the experimental design used for comparative textual analysis and detectability evaluation. Section 4 reports the empirical results, including descriptive statistics of textual characteristics, raw and length-controlled lexical diversity analysis, similarity comparisons between human and LLM-generated responses, and the performance of the detectability model. Section 5 discusses the implications of the findings for the evaluation of LLM-generated agricultural answers and the limitations of the study. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Related Works

Agricultural communities in low-resource environments often face persistent challenges in accessing expert guidance and timely agronomic information. IR technologies have increasingly been explored as mechanisms for improving access to agricultural knowledge, particularly in contexts where farmers depend on fragmented or informal sources of information [18]. Several practical systems illustrate how digital platforms can support agricultural information access. For example, Digital Green provides video-based agricultural knowledge and searchable advisory content to farmers in developing regions through community-based dissemination networks [19]. Similarly, iCow, a mobile advisory service deployed in East Africa, allows farmers to retrieve livestock management and agricultural information through SMS-based queries [20]. Another example is eSoko in Ghana, which delivers agricultural information such as market prices, weather forecasts, and farming advice through mobile information services [21]. These systems demonstrate how information retrieval and mobile information platforms can facilitate agricultural knowledge dissemination in regions where access to traditional extension services is limited.

Despite these advances, the development and evaluation of IR systems remain challenging in low-resource environments. Such environments are often characterized by limited datasets for benchmarking retrieval models and reduced computational infrastructure. For example, many African languages lack large-scale annotated corpora or evaluation collections comparable to widely used IR datasets such as MS MARCO (Nguyen et al., 2018) or TREC collections [22,23], which are commonly used to benchmark retrieval systems in high-resource languages. The scarcity of datasets and linguistic resources for African languages has been highlighted in recent research on low resource natural language processing and information access, including initiatives such as Masakhane, which focuses on building machine translation and NLP resources for African languages [24]. These limitations make it difficult to design and evaluate IR systems that support localized knowledge access in agricultural contexts. For agricultural question answering, evaluation also requires attention to domain-specific terminology and the language varieties used by farmers.

Recent advances in large language models (LLMs) have introduced new possibilities for conversational information retrieval and question answering systems. Models such as GPT-3 [25], ChatGPT [26], Gemini [27], and LLaMA [28] can generate natural language responses to user queries, allowing them to function as interactive information access systems that extend beyond traditional keyword-based search. These models have demonstrated advanced performance across tasks such as question answering, summarization, and conversational dialogue, which has motivated growing interest in deploying conversational AI systems to deliver agricultural information and advisory services [29]. In agricultural contexts, conversational interfaces can potentially provide farmers with direct answers to domain-specific questions, enabling information access through natural language interaction rather than document retrieval.

The shift from retrieving documents to generating answers introduces important evaluation challenges [30]. LLM-based systems compose responses dynamically through probabilistic language generation, which produces plausible but factually incorrect information [31]. Studies evaluating generative language models have documented instances where systems produce fabricated references or unsupported factual claims [32]. These issues complicate the evaluation of generative IR systems because the reliability and provenance of generated answers may be difficult to verify. Consequently, recent research has begun exploring methods for evaluating generated responses, including hallucination detection, answer quality assessment, and approaches for distinguishing human-written text from machine-generated outputs [16,17]. These approaches examine different properties of

generated answers: their factual content, their quality, and the textual features associated with their source.

In agricultural information access scenarios, these evaluation challenges are particularly important. Farmers typically rely on concise and practical explanations grounded in local agricultural practices [11]. If generative systems produce responses that differ systematically from human-authored explanations, these differences may affect both the evaluation of information retrieval systems and how users interpret their outputs [33]. Examining such differences through source classification provides a way to assess whether generated answers have identifiable lexical and stylistic patterns. However, differences in answer length and the instructions used to elicit responses can also contribute to classification performance. Comparing answers under length-controlled conditions therefore helps establish whether source-associated patterns persist beyond the original differences in response length. This provides the basis for the present analysis of human and LLM-generated answers to Swahili agricultural questions [34].

3. Methodology

This study adopts an empirical text-analysis framework to examine whether large language model (LLM) responses can be systematically distinguished from human-authored agricultural explanations. The analysis focuses on textual representation and source detectability, drawing on methods used in IR and stylometric studies, which are commonly used analytical perspectives in IR and stylometric studies [35]. The methodological design combines descriptive comparison of textual properties with lexical similarity analysis and supervised classification. Answer sources are compared using full text and equal-length responses, with classification evaluated on questions kept separate between training and testing.

3.1. Dataset

The dataset used in this study was collected through fieldwork conducted in Morogoro, Tanzania, over approximately two weeks in May 2024. The research team prepared questions covering common maize-farming issues and collected responses from approximately 24 farmers. Responses were documented through notes taken by the research team. Crop experts and extension officers working with maize reviewed the answers and refined their wording. Farmers were asked about common farming issues without a prescribed response length or written format. Some questions were removed during dataset preparation as the collection progressed and responses became repetitive. The final dataset contained 127 Swahili questions.

Each question was paired with three answers: a human-authored response documented during fieldwork and reviewed by agricultural experts, and two responses generated by ChatGPT (GPT 5.2) and Gemini 3 Pro. This design resulted in 381 answers, comprising 127 answers from each source, and enabled comparison of responses to the same questions. The fieldwork was conducted under a research permit issued by Sokoine University of Agriculture.

ChatGPT and Gemini were selected because they represent widely deployed transformer-based LLM systems trained on large-scale heterogeneous corpora. GPT-family models are trained on mixtures of publicly available, licensed, and human-generated text comprising hundreds of billions of tokens [25], while Gemini models are trained on large multimodal web-scale datasets supporting multilingual reasoning and conversational interaction [27]. Both systems therefore provide representative contemporary LLM architectures capable of multilingual responses, including Swahili input. Including two LLM systems enabled comparison of source-associated linguistic patterns across different conversational models [34].

To ensure consistency across generated answers, both models were prompted using the same Swahili instruction and the prompt imposes consistent response constraints to reduce stylistic variability and ensure that observed differences reflect model-generated patterns rather than response formatting [36]. The same Swahili instruction was repeated with every question in both systems. It requested a brief, direct explanation in clear Swahili, with practical advice and no lists or bullet points as shown in Table 1:

> "*Jibu swali lifuatalo kwa lugha ya Kiswahili fasaha, kwa ufupi na kwa uelewa wa mkulima wa kawaida. Epuka maelezo marefu yasiyo ya lazima. Toa jibu la moja kwa moja na lenye ushauri wa vitendo. Usitumie pointi, orodha, mistari ya nukta au muundo wa vipengele.*"

English translation:

> "*Answer the following question in clear, fluent Swahili, briefly and in a way that an ordinary farmer can understand. Avoid unnecessary lengthy explanations. Give a direct answer with practical advice. Do not use bullet points, lists, dotted lines, or an itemized format.*"

The instruction specified a preferred style without imposing a numerical word limit. Human answers were obtained through fieldwork discussions and research-team notes, whereas model answers were generated as written responses to the instruction. These collection procedures were considered when interpreting differences between sources.

Table 1. Structure of the Swahili maize question–answer dataset.

Question (Swahili)	Human Answer	GPT Answer	Gemini Answer
Umuhimu wa uchavushaji kwa mmea wa mahindi ni upi?	Uchavushaji husaidia mmea kuweza kuzalisha tunda. Urutubishaji wa mbegu za kike katika unaofanywa na mbegu za kiume (mbelewele) husabaisha kuzaliwa kwa ua ambalo ni tunda la mhindi	Uchavushaji ni muhimu kwa sababu bila chembe za chavua kufika kwenye sehemu ya kike ya ua, kozi za mahindi haziwezi kutengeneza punje. Huu ndiyo mchakato unaofanya kila shanga ya mahindi ijazwe vizuri na kuwa na nguvu. Ikiwa uchavushaji haujatimia, mahindi huwa na mashimo, yanakauka mapema na mavuno hupungua. Ili kuhakikisha uchavushaji unafanikiwa, unahitaji kupanda mahindi kwa nafasi sahihi na katika mistari inayosaidia upepo kupeleka chavua kwa urahisi kutoka ua moja kwenda jingine.	Uchavushaji ni muhimu sana kwa mahindi kwa sababu bila huo, masuke hayatajaza chembechembe, na badala yake utapata masuke matupu au yenye chembe chache sana. Uchavushaji huruhusu mavumbi ya maua ya kiume kufika kwenye nywele za masuke (maua ya kike), na hivyo kusababisha chembechembe za mahindi kuunda na kukua. Ili kuhakikisha uchavushaji mzuri, panda mahindi kwa mistari mingi, si michache, ili upepo uweze kubeba mavumbi ya maua kwa urahisi na kuyafikisha kila mahali.

English translation: Question: What is the importance of pollination in maize? **Human answer:** Pollination helps maize produce grain. Pollen from the male flowers reaches the silks of the female flowers, enabling fertilization and subsequent kernel development. **GPT answer:** Pollination is important because maize cobs cannot form kernels unless pollen grains reach the female part of the flower. This process allows each kernel to fill properly and develop. If pollination is incomplete, the cobs have gaps, the maize dries prematurely, and yields decrease. To ensure successful pollination, plant maize with appropriate spacing and in rows that help the wind carry pollen easily from one flower to another. **Gemini answer:** Pollination is very important for maize because, without it, the ears will not fill with kernels, resulting in empty ears or ears with very few kernels. Pollination allows pollen from male flowers to reach the silks of the female flowers, enabling kernels to form and grow. To ensure good pollination, plant maize in several rows rather than only a few, so that the wind can carry and distribute pollen throughout the planting area.

3.2. Data Preparation

Data preprocessing was performed using standard normalization procedures in information retrieval research [37], including lowercasing text and tokenizing answers into words. The dataset was first checked for missing values, blank entries, and duplicate questions. Each question received a unique identifier linking its human, GPT, and Gemini answers, and original answer texts were retained. A common tokenizer extracted sequences matching the Unicode regular expression \w+, treating punctuation as separators and retaining letters, digits, and underscores. This tokenizer was used consistently for word counts, lexical-diversity measures, TF-IDF representations, and the equal-length experiment. Answer verbosity was measured using two complementary statistics, beginning with character length, defined as the number of characters in each response [38]:

$$L_c = |\text{text}| \tag{1}$$

where L_c denotes the character length of the answer. Characters were counted from the original text, including spaces and punctuation.

Token length measures the number of words contained in each response:

$$L_t = \sum_{i=1}^{n} 1 \tag{2}$$

where L_t represents the total number of tokens in the answer and n is the number of tokens produced by the tokenizer.

Lexical diversity was measured using the type-token ratio (TTR), which captures the proportion of unique words within a text:

$$\text{TTR} = \frac{|V|}{N} \tag{3}$$

where |V| represents the number of unique tokens and N represents the total number of tokens. TTR is widely used in stylometric and corpus-based analyses to compare vocabulary richness across text sources [39].

Because raw TTR is sensitive to text length, the moving-average type-token ratio (MATTR) was also calculated. MATTR averages the TTR values of overlapping windows of a fixed size, moving through each answer one token at a time:

$$\text{MATTR}_w = \frac{1}{N - w + 1} \sum_{i=1}^{N-w+1} \frac{|V_{i,w}|}{w}, \qquad N \geq w \tag{4}$$

where w is the window size and $|V_{i,w}|$ is the number of unique tokens in the window beginning at position i. The main analysis used 20-token windows, with a 10-token analysis as an additional check. For each window size, comparisons included only questions whose human, GPT, and Gemini answers all met the minimum length. Each eligible answer contributed one MATTR value.

To analyze lexical linguistic similarity between answers, responses were represented using term frequency-inverse document frequency (TF-IDF) vectors, a standard text representation method in information retrieval [40]. The TF-IDF weight of term t in document d was calculated using raw frequency and smoothed inverse document frequency:

$$\text{TFIDF}(t, d) = \text{TF}(t, d)\left[\log\left(\frac{1 + D}{1 + \text{DF}(t)}\right) + 1\right] \tag{5}$$

where $\text{TF}(t, d)$ denotes the frequency of term t in document d, $\text{DF}(t)$ represents the number of documents containing term t, and D is the total number of documents. The resulting vectors were normalized to unit Euclidean length.

Similarity between answers was then measured using cosine similarity, which quantifies the directional similarity between TF-IDF vectors:

$$\cos(\theta) = \frac{A \cdot B}{\| A \| \| B \|} \quad (6)$$

where A and B represent vector representations of two answers. This measure allows comparison between LLM responses and their corresponding human-authored answers [41]. For this descriptive analysis, a unigram TF-IDF representation was fitted to all 381 answers, and each human answer was compared with the GPT and Gemini answers to the same question.

A mismatched-question baseline was constructed using 500 random permutations of model answers. Each permutation paired human answers with model answers to different questions. For each human answer, similarity scores across these mismatched pairings were averaged and compared with the same-question score. Answer pairs with the lowest, median-ranked, and highest same-question similarity scores were selected for inspection for each model.

3.3. Experimental Design

To evaluate detectability, the study implemented a supervised classification experiment [42] designed to determine whether textual patterns could identify the source of an answer. All answers were transformed into TF-IDF feature vectors using both unigrams and bigrams, enabling the model to capture lexical and short phrase patterns that may characterize different answer sources [41]. A multiclass logistic regression classifier was used to predict the answer source (human, GPT, or Gemini). Logistic regression is commonly used in text classification because it performs effectively with high-dimensional sparse feature spaces such as TF-IDF representations [43]. The TF-IDF vectorizer and classifier were fitted together within each training split, so vocabulary and inverse document frequency weights were learned from training answers only. Terms occurring in more than 95% of training answers were excluded. The classifier used ($C = 1$) and a maximum of 3000 iterations, without hyperparameter tuning.

Four experiments were conducted using identical training and test partitions: a majority-class baseline, a classifier using answer length alone, a TF-IDF classifier using complete answers, and a TF-IDF classifier using equal-length answers. The majority-class baseline predicted the most frequent training class. The length-only classifier used token count as its sole predictor, standardized using training-set statistics. For the equal-length experiment, the three answers to each question were shortened to the token count of the shortest answer in that triplet, retaining the opening tokens. All 127 questions remained in the analysis. This experiment assessed performance after matching answer lengths while also accounting for the content removed through truncation.

The primary evaluation used five-fold cross-validation grouped by question. All three answers to each question remained in the same fold, preventing answers to the same question from appearing in both training and testing. Each fold was used once for testing, producing one out-of-fold prediction for every answer. A separate question-grouped 75:25 holdout split provided an additional performance check. Each partition contained equal numbers of human, GPT, and Gemini answers. Random seed 42 was used for split generation as presented in Table 2.

Table 2. Training and test dataset sizes used in the classification experiments [43].

Evaluation	Training Questions	Training Answers	Test Questions	Test Answers
Five-fold evaluation, folds 1–2, each	101	303	26	78
Five-fold evaluation, folds 3–5, each	102	306	25	75
Separate 75:25 holdout	95	285	32	96

Model performance was evaluated using standard classification metrics, including precision, recall, and F1 score, defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

where (TP), (FP), and (FN) denote true positives, false positives, and false negatives for each source, respectively. Accuracy measured the proportion of correctly classified answers, while macro F1 was the unweighted mean of the three source-specific F1 scores. Primary scores were calculated from pooled out-of-fold predictions, with fold-level means and standard deviations describing variation across partitions. Conditional 95% confidence intervals were estimated from 2000 bootstrap samples of question groups, keeping each question's three answers together. Intervals were calculated from the 2.5th and 97.5th percentiles using fixed predictions, without refitting the models.

In addition, confusion matrices were used to analyze misclassification patterns across the three sources, with a consistent class order of human, GPT, and Gemini. Words and phrases with the highest positive logistic regression coefficients were extracted from the holdout training models to identify source-associated features. The frequency of the repeated phrase "ushauri wa vitendo" was also examined across sources. This experimental setup allowed the study to assess source detectability and the contribution of answer length to classification performance in Swahili agricultural question answering [44].

4. Results

4.1. Descriptive Textual Characteristics

Descriptive statistics were computed to examine structural differences between human-authored and LLM-generated answers. Two measures of answer length were analyzed: character length and token length, allowing comparison of response verbosity across sources. Table 3 summarizes the distribution of these measures for human, ChatGPT, and Gemini responses.

Table 3. Statistics of character and token lengths for human and LLM-generated answers (n = 127 per source).

Source	Metric	Count	Mean	Std	Min	25%	50%	75%	Max
Human	Characters	127	166.42	99.67	21	101.0	146.0	205.5	560
GPT	Characters	127	466.58	66.25	329	422.0	460.0	518.5	690
Gemini	Characters	127	484.81	59.81	332	438.0	483.0	516.5	662
Human	Tokens	127	25.17	15.22	4	15.0	22.0	31.0	80
GPT	Tokens	127	71.13	9.60	51	63.5	71.0	77.5	105
Gemini	Tokens	127	77.24	10.05	54	70.0	76.0	82.0	106

The results show a clear difference in response length between human and model-generated answers. Human responses contain an average of 166.42 characters and 25.17 tokens, whereas ChatGPT and Gemini responses are substantially longer, averaging 466.58 characters (71.13 tokens) and 484.81 characters (77.42 tokens), respectively. The shortest ChatGPT answer contains 51 tokens, and the shortest Gemini answer contains 54 tokens, compared with a human median of 22 tokens. These results indicate that LLM-generated answers tend to provide longer explanations than human-authored responses for the same questions.

Variation in answer length also differs across sources. Human answers show a higher standard deviation in token length (15.22 tokens) compared to ChatGPT (9.60 tokens) and Gemini (10.05 tokens), suggesting that human responses are more variable in length. In contrast, LLM-generated answers have more concentrated length distributions. Figure 1 illustrates these differences through boxplots, with median lengths of 22 tokens for human answers, 71 for ChatGPT, and 76 for Gemini.

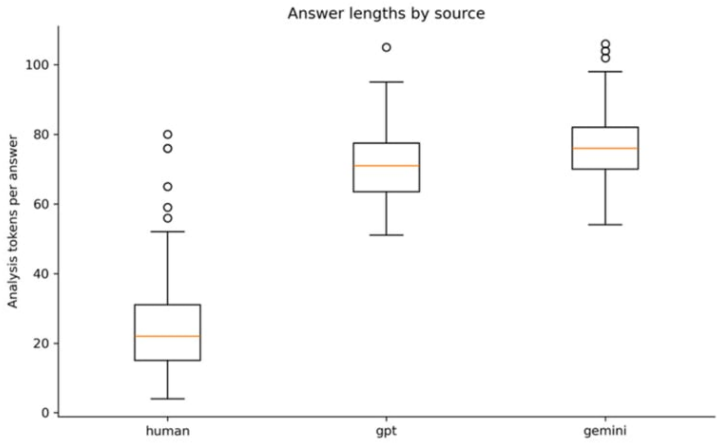


Figure 1. Boxplot comparison of answer lengths across human, ChatGPT, and Gemini responses.

In addition to length differences, lexical diversity was examined using the type-token ratio (TTR), which measures the proportion of unique words relative to the total number of tokens in a response. Table 4 summarizes the distribution of raw TTR across the three sources. Human responses exhibit the highest average TTR (0.866), indicating a greater proportion of unique words relative to response length. ChatGPT responses show a lower average TTR (0.769), while Gemini responses display the lowest lexical diversity (0.718). Figure 2 shows raw TTR in relation to answer length, with shorter answers generally exhibiting higher ratios. Raw TTR was therefore examined alongside fixed-window MATTR values as presented in Table 4.

Table 4. Descriptive statistics of answer length and lexical diversity across sources (n = 127 per source).

Source	Metric	Mean	Std	Min	25%	Median	75%	Max
Human	Token Length	25.17	15.22	4	15	22	31	80
Human	Type-Token Ratio	0.866	0.093	0.667	0.784	0.871	0.939	1.000
ChatGPT	Token Length	71.13	9.60	51	63.5	71	77.5	105
ChatGPT	Type-Token Ratio	0.769	0.044	0.662	0.739	0.776	0.803	0.868
Gemini	Token Length	77.24	10.05	54	70	76	82	106
Gemini	Type-Token Ratio	0.718	0.050	0.609	0.684	0.720	0.756	0.836

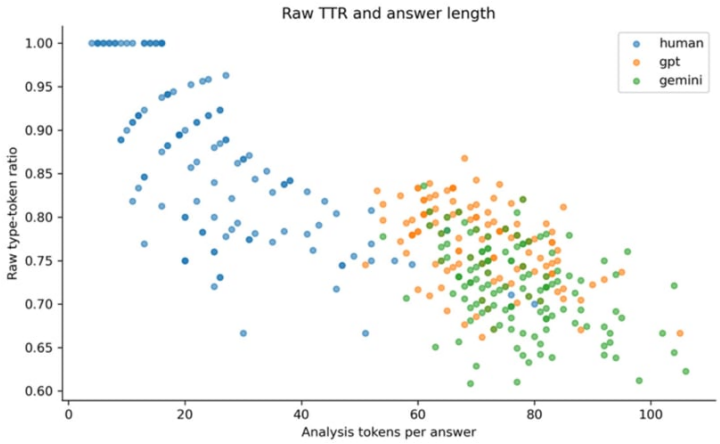


Figure 2. Relationship between raw type-token ratio and answer length across sources.

The MATTR comparison produced a different ordering from raw TTR. Using 20-token windows, the mean MATTR was 0.8725 for human answers, 0.9307 for ChatGPT, and 0.9120 for Gemini (Table 5). This comparison included 73 questions, with 54 excluded because their human answers contained fewer than 20 tokens. Figure 3 shows the corresponding distributions.

Table 5. Fixed-window lexical diversity across sources.

Window Size	Source	Count per Source	Mean MATTR	Std	Median
20 tokens	Human	73	0.8725	0.0624	0.8900
20 tokens	ChatGPT	73	0.9307	0.0280	0.9333
20 tokens	Gemini	73	0.9120	0.0320	0.9145
10 tokens	Human	114	0.9432	0.0519	0.9596
10 tokens	ChatGPT	114	0.9705	0.0160	0.9706
10 tokens	Gemini	114	0.9637	0.0194	0.9658

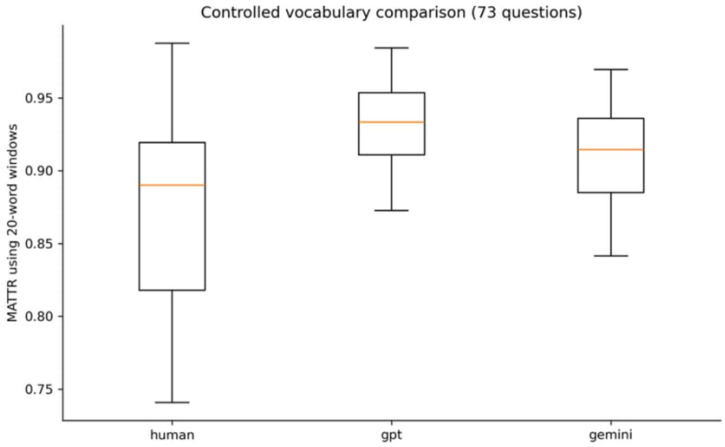


Figure 3. Distribution of MATTR using 20-token windows across the 73 eligible questions.

The 10-token analysis included 114 questions, excluding 13 with human answers below the required length. Mean MATTR was 0.9432 for human answers, 0.9705 for ChatGPT, and 0.9637 for Gemini, as illustrated in Figure 4. Both window sizes, therefore, showed higher average MATTR for model answers within their respective eligible samples. The higher

raw TTR of human answers did not persist when vocabulary diversity was measured over fixed-size windows.

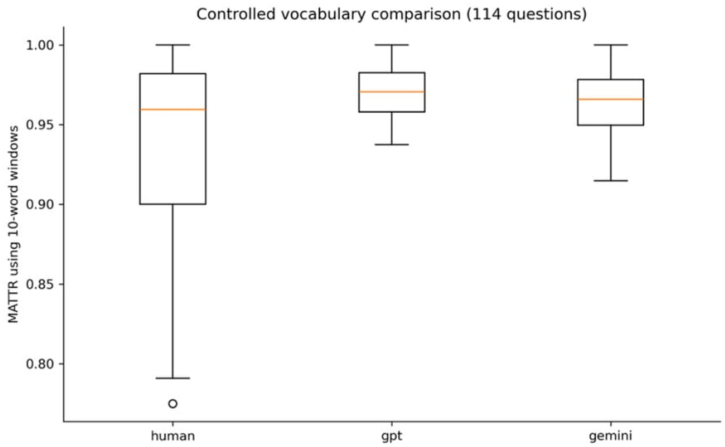


Figure 4. Distribution of MATTR using 10-token windows across the 114 eligible questions.

4.2. Similarity Between Human and LLM Answers

To examine how closely LLM-generated responses resemble human-authored explanations, similarity between answers was computed using TF-IDF vector representations and cosine similarity. This approach measures the degree to which two answers share similar lexical patterns within a vector-space representation of the text. Table 6 summarizes the distribution of similarity scores between human answers and the corresponding LLM responses. On average, Gemini responses show slightly higher similarity to human answers (mean = 0.166) compared with ChatGPT responses (mean = 0.146). Median similarity values follow a similar pattern, with Gemini reaching 0.163 and ChatGPT 0.147. However, both distributions remain relatively low overall, indicating that although LLM answers address the same questions, their lexical formulations often differ from human-authored explanations.

Table 6. Summary statistics of TF-IDF cosine similarity between human–GPT and human–Gemini answer pairs.

Statistic	Human–GPT Similarity	Human–Gemini Similarity
Count	127	127
Mean	0.146	0.166
Std. Dev.	0.077	0.086
Min	0.000	0.023
25th Percentile	0.087	0.103
Median (50%)	0.147	0.163
75th Percentile	0.194	0.224
Max	0.417	0.441
Mean mismatched-question similarity	0.0590	0.0614
Mean same-question minus mismatched similarity	0.0867	0.01043

The variability of similarity scores is comparable across the two models, with standard deviations of 0.077 for ChatGPT and 0.086 for Gemini. The observed ranges further illustrate the diversity of responses: similarity scores for ChatGPT vary from 0.000 to 0.417, while Gemini responses range from 0.023 to 0.441. These ranges indicate variation in lexical overlap across questions. A zero indicates no shared features in the fitted TF-IDF representation.

The mismatched-question baseline provides context for the same-question scores. Mean similarity between human and model answers to different questions was 0.0590 for ChatGPT and 0.0614 for Gemini. Same-question similarity exceeded these baseline means by 0.0867 and 0.1043, respectively. Figure 5 shows the distribution of same-question scores alongside the distribution of mean mismatched scores for each human answer. Same-question scores extend across a wider range, while mean mismatched scores are concentrated near the lower end of the scale.

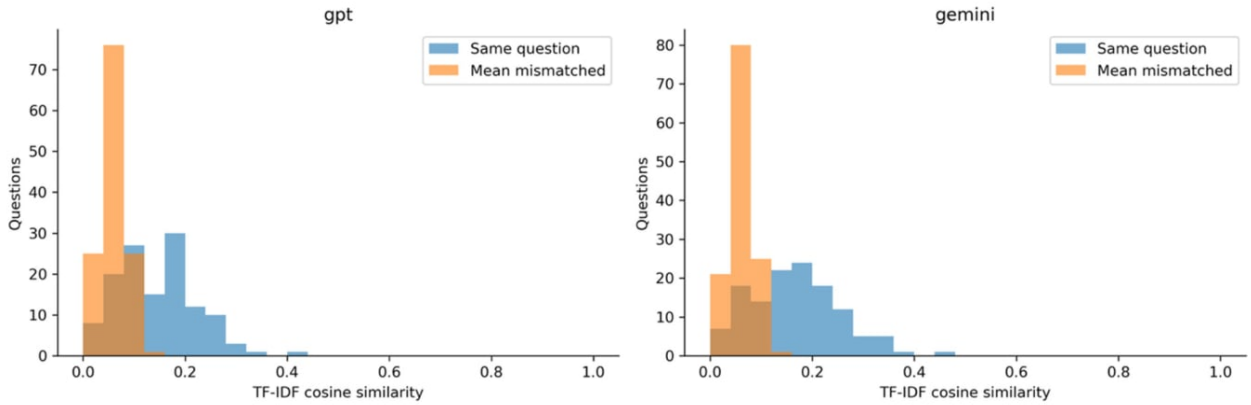


Figure 5. TF-IDF cosine similarity for the same question–answer pairs and mean mismatched-question comparisons: ChatGPT on the left and Gemini on the right. Selected answer pairs illustrate the observed range. For ChatGPT, the lowest similarity score occurred for a question about premature drying of maize while still on the plant (0.000), and the highest concerned the effects of climate change on maize flowering (0.417). For Gemini, the lowest score concerned the timing of thinning maize plants (0.023), while the highest concerned maize growing taller under tree shade (0.441). Overall, answers to the same question shared more vocabulary on average than mismatched answers, with considerable variation between individual pairs.

4.3. Detectability of Answer Sources

To evaluate whether the textual differences observed in previous analyses translate into detectable patterns, a supervised classification experiment was conducted using TF-IDF features and a logistic regression classifier. All answers in the dataset were combined into a single corpus of 381 responses, each labeled according to its source (human, ChatGPT, or Gemini). The primary evaluation used five-fold cross-validation grouped by question, with all three answers to a question kept together and TF-IDF fitted only on training answers.

The full-text classifier achieved an overall accuracy of 0.9633 and macro F1 of 0.9632. The length-only classifier achieved 0.7192 accuracy, while the equal-length TF-IDF classifier achieved 0.8058. All three exceeded the majority-class baseline accuracy of 0.3333. Table 7 and Figure 6 summarize these comparisons.

Table 7. Overall classification performance using pooled out-of-fold predictions.

Experiment	Accuracy	Conditional 95% CI for Accuracy	Macro F1	Conditional 95% CI for Macro F1
Majority-class baseline	0.3333	0.3333–0.3333	0.1667	0.1667–0.1667
Length-only classifier	0.7192	0.6772–0.7612	0.7215	0.6820–0.7618
Full text TF-IDF classifier	0.9633	0.9423–0.9816	0.9632	0.9420–0.9816
Equal-length TF-IDF classifier	0.8058	0.7638–0.8478	0.8058	0.7634–0.8471

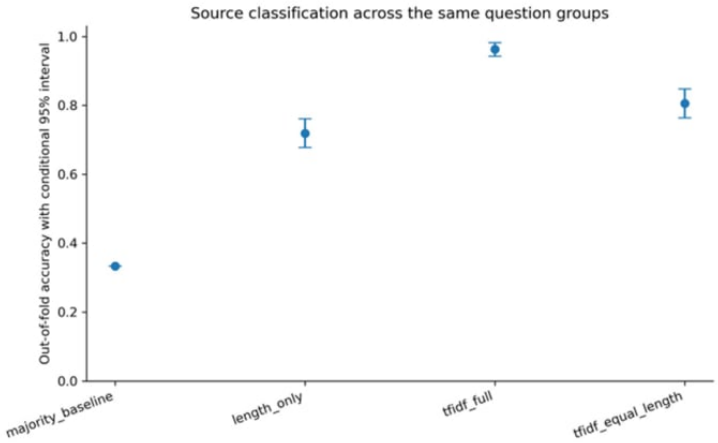


Figure 6. Out-of-fold source-classification accuracy across the four experiments, with conditional 95% bootstrap intervals. Accuracy decreased by 15.75 percentage points between the full-text and equal-length experiments. Across individual folds, full-text accuracy ranged from 0.9359 to 0.9872, compared with 0.7333 to 0.8590 for equal-length classification. The separate question-grouped holdout evaluation produced the same ordering, with accuracies of 0.9479 for full text, 0.8021 for equal-length text, and 0.6667 for length alone.

Source-specific results are presented in Table 8. Full-text F1 scores were 0.9551 for human answers, 0.9542 for ChatGPT, and 0.9804 for Gemini. Corresponding equal-length F1 scores were 0.8254, 0.8000, and 0.7920.

Table 8. Source-specific classification performance using pooled out-of-fold predictions.

Experiment	Source	Precision	Recall	F1 Score	Support
Full text	Human	0.9915	0.9213	0.9551	127
Full text	ChatGPT	0.9259	0.9843	0.9542	127
Full text	Gemini	0.9766	0.9843	0.9804	127
Equal length	Human	0.8320	0.8189	0.8254	127
Equal length	ChatGPT	0.7820	0.8189	0.8000	127
Equal length	Gemini	0.8049	0.7795	0.7920	127

The confusion matrices in Figure 7 illustrate the classification results in greater detail. The full-text classifier correctly identified 117 of 127 human answers, 125 ChatGPT answers, and 125 Gemini answers. Eight human answers were classified as ChatGPT and two as Gemini. One ChatGPT answer was classified as human and one as Gemini, while two Gemini answers were classified as ChatGPT. The equal-length classifier correctly identified 104 human answers, 104 correctly identified ChatGPT answers, and 99 correctly identified Gemini answers. The length-only classifier correctly identified 121 human answers but showed greater confusion between the two model sources: 47 ChatGPT answers were classified as Gemini, and 53 Gemini answers were classified as ChatGPT.

Beyond classification accuracy, the model also reveals lexical features that contribute most strongly to distinguishing answer sources. In the full-text model fitted to the holdout training set, Gemini responses are characterized by common function words and advisory phrases such as “ushauri,” “vitendo,” and “wa vitendo.” In contrast, ChatGPT responses are characterized by tokens such as “muhimu,” “unaweza,” and “udongo,” reflecting explanatory phrasing patterns commonly used in generated responses. Human-authored answers

are associated with more context-specific agricultural expressions, including "wadudu wanaotoboa," "madini," and "usafi."

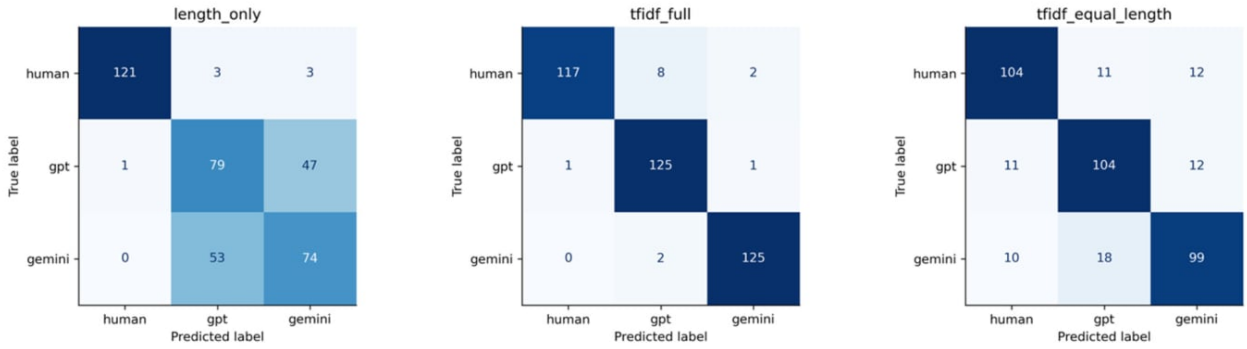


Figure 7. Confusion matrices from pooled out-of-fold predictions for length-only classification (**left**), full-text TF-IDF classification (**center**), and equal-length TF-IDF classification (**right**).

The phrase "ushauri wa vitendo" appeared in 123 of 127 Gemini answers (96.85%) and in none of the human or ChatGPT answers. In the equal-length model, prominent Gemini-associated features included "hasa," "jambo," and "jambo la." ChatGPT-associated features included "unaweza" and "ni muhimu," while "mmea," "kwa ajili," and "inashauriwa" were associated with human answers.

These results show that answer sources remained distinguishable in the equal-length experiment, although classification performance was lower than for complete answers. Answer length and repeated source-specific wording were also informative characteristics within this dataset.

5. Discussion

5.1. Structural Differences Between Human and LLM Responses

This study investigated whether answers generated by LLMs in Swahili agricultural question-answering exhibit detectable differences from human-authored responses [45]. The results show differences across multiple textual dimensions, including answer length, lexical diversity, lexical similarity, and source detectability. As demonstrated in Table 3 and Figure 1, human-authored responses tend to be substantially shorter and more variable in length than LLM-generated responses. In contrast, answers produced by ChatGPT and Gemini are consistently longer and more uniform in length. Both models produced answers approximately three times the average length of the human responses, despite receiving repeated instructions to answer briefly and avoid unnecessary explanations. This suggests that a general request for brevity was insufficient to bring generated answers close to the concise explanations collected during fieldwork. Difficulties in controlling generated response length have also been reported by Ref. [46], whose Ruler framework improved adherence to specified length requirements. Their findings identify length control as a distinct challenge in instruction-following. The present study demonstrates a related problem in Swahili agricultural question answering: the models followed a common advisory prompt but continued to produce substantially more text than the human reference responses. This difference points to a gap between requesting farmer-oriented explanations and achieving the compact response style represented in the fieldwork data.

The descriptive analysis further reveals differences in lexical diversity across answer sources. As shown in Table 4 and Figure 2, human responses exhibit higher raw type-token ratios than LLM-generated outputs. The fixed-window analyses, however, show higher MATTR for both models across the 20-token and 10-token comparisons. Thus, the mod-

els used more varied vocabulary within comparable stretches of text, while their longer complete answers contained more repetition. Ref. [47] reported lower lexical richness in ChatGPT answers than in human answers across their examined tasks. The present fixed-window results show a different pattern, demonstrating that reduced vocabulary diversity is not a consistent distinguishing characteristic of generated answers across datasets and analytical measures. Here, the more persistent distinction is the models' extended presentation of information. The combined findings therefore characterize generated agricultural explanations as longer responses containing locally varied vocabulary, rather than simply as responses with a restricted vocabulary.

5.2. Lexical Similarity Between Human and LLM Answers

The similarity analysis provides further insight into how LLM-generated responses relate to human-authored agricultural explanations [25,45]. As reported in Table 6 and illustrated in Figure 5, average TF-IDF cosine similarity was 0.146 for ChatGPT and 0.166 for Gemini. These values were approximately 2.5 and 2.7 times their respective mismatched-question baselines of 0.0590 and 0.0614. Corresponding answers therefore shared substantially more vocabulary than answers to different maize questions, indicating a question-specific lexical relationship within the broader agricultural domain. The combination of greater same-question overlap and relatively low absolute similarity suggests that the sources draw on shared agricultural terminology while organizing their explanations differently. This distinction is relevant to reference-based evaluation: an answer can contain vocabulary associated with the question while differing considerably from the wording of the collected human response. Ref. [48], in developing BERTScore, similarly addressed the limitations of exact lexical matching by comparing contextual representations of words. Their work helps explain why differences in phrasing can produce limited lexical overlap even when responses discuss related content. Gemini's higher average similarity, together with its longer answers, also shows that verbosity and lexical correspondence do not vary in a simple inverse relationship. Longer explanations can retain more reference-associated terms while remaining stylistically distinct. The contribution of this analysis is therefore to establish that question-specific vocabulary persists alongside substantial differences in expression. In Swahili agricultural IR, shared subject vocabulary alone does not make generated answers resemble the concise explanations collected from farmers.

5.3. Detectability of Machine-Generated Agricultural Answers

The classification experiment demonstrates that these textual differences support answer-source identification [35]. Using TF-IDF representations combined with logistic regression, the model achieved an overall accuracy of 0.9633 under five-fold question-grouped evaluation when predicting whether an answer was produced by a human, ChatGPT, or Gemini. As illustrated in Figure 7 and summarized in Tables 7 and 8, 117 human answers, 125 ChatGPT answers, and 125 Gemini answers were correctly classified out of 127 per source. Because all answers to a question remained within the same evaluation fold, this performance reflects source identification on questions excluded from classifier training. The comparison experiments reveal two contributions to this performance. Length alone achieved 0.7192 accuracy and correctly identified 121 human answers, but frequently confused ChatGPT and Gemini. Response length therefore provided a strong distinction between the human and generated answers, whereas separating the two models required additional lexical information. After truncating each question's three answers to the same token count, TF-IDF accuracy decreased to 0.8058, a reduction of 15.75 percentage points, but remained above the 0.3333 baseline. The reduction indicates that answer length and the content removed through truncation contributed to full-text detection, while the remain-

ing performance demonstrates source-associated differences within equally long answer segments. This pattern agrees with Ref. [49], who found that machine-text detectors were sensitive to writing style and text complexity. The present experiments identify verbosity as one such influential characteristic and demonstrate that lexical distinctions remain after answer lengths are matched. This provides a more specific account of detectability than classification accuracy alone: the classifier benefits from both the extended form of generated answers and differences in their wording.

Recurring expressions offer a concrete example. The phrase "ushauri wa vitendo" occurred in 123 Gemini answers and in none of the human or ChatGPT answers. Ref. [50] likewise investigated the overrepresentation of particular words in LLM outputs. Here, the repeated phrase also appeared in the instruction, suggesting that Gemini's response to the shared prompt contributed a particularly distinctive source cue. This illustrates how repeated advisory formulations can coexist with high vocabulary diversity within individual answer segments. Multilingual detection research, including the Ref. [51] MULTITuDE benchmark, has established the importance of evaluating detection across languages and generators. The present study adds evidence from Swahili, which was not included among MULTITuDE's 11 languages, using agricultural questions and human responses collected through fieldwork. Its contribution is the demonstration that an interpretable classifier can distinguish these three answer sources on unseen questions, together with evidence explaining how verbosity and recurring lexical choices support that distinction.

5.4. Implications for Low-Resource Agricultural Information Retrieval

These findings have implications for the evaluation of conversational agricultural information systems [37,46–48]. In agricultural contexts, such systems have been proposed as scalable tools for providing farmers with access to agronomic advice and decision support [46,47]. The present results identify a practical challenge in adapting these systems to concise agricultural communication: repeated instructions requesting brief, direct advice still produced substantially longer explanations than those collected during fieldwork. This makes a response form an important part of evaluating whether an agricultural assistant delivers information in the manner requested.

Ref. [52], studying climate-adaptation question answering in agriculture, found a trade-off between presentation quality and epistemological accuracy, with domain experts better able to recognize erroneous information. Their findings and the present results address complementary aspects of agricultural answer evaluation. Whereas their work examines the reliability of advisory content, this study shows that the form and wording of generated advice remain systematically distinguishable from fieldwork-based human explanations. Together, these findings support evaluating agricultural assistants through both the substance of their advice and their adaptation to the intended communication setting.

Furthermore, the ability to detect machine-generated responses may support transparency and governance in information systems. The strong performance of TF-IDF and logistic regression provides an accessible basis for investigating provenance checks in Swahili agricultural services. More broadly, the study shows that answering the same agricultural questions in the same language does not produce comparable explanatory styles across human and model sources. The persistence of detectable lexical differences after matching answer lengths establishes that this divergence extends beyond verbosity, contributing empirical evidence about how contemporary conversational systems express agricultural information in Swahili.

5.5. Limitations of the Study

Despite the contributions, this study has several potential several limitations. The dataset used in this study is relatively small compared to standard IR benchmark collections and focuses on 127 Swahili maize-related questions collected in Morogoro, Tanzania. Consequently, the findings may not fully generalize to other crops, agricultural contexts, language varieties, or knowledge domains. Second, the analysis includes only two LLM systems, ChatGPT (GPT-5.2) and Gemini 3 Pro, with one response collected per question. The findings therefore reflect these model versions and the generation conditions used in this study. Both systems received the same instruction, and questions were submitted sequentially within a continuing conversation for each system. Repeated wording and accumulated conversational context may have influenced the observed source patterns.

Third, human responses were recorded as notes and reviewed by agricultural experts, while model answers were generated as written explanations. These collection differences may have influenced length and wording. Equal-length comparisons reduced length differences, but truncation also removed content and could interrupt sentences. Finally, TF-IDF captures lexical rather than contextual meaning; future work could explore multilingual embeddings.

6. Conclusions

This study examined whether answers generated by LLMs in a low-resource agricultural IR context exhibit detectable differences from human-authored responses. Using a Swahili maize question-answer dataset collected through fieldwork in Morogoro, Tanzania, the analysis compared human responses with answers generated by two contemporary conversational models, ChatGPT and Gemini. The study employed a combination of descriptive textual analysis, lexical similarity measures, and supervised classification to investigate differences in response structure and detectability. The results demonstrate consistent differences between human and LLM-generated answers across multiple textual dimensions. Human responses were generally shorter and exhibited greater lexical diversity, while LLM responses were longer and more uniform in length despite repeated instructions to answer briefly. Human answers had higher raw TTR, whereas both models showed greater lexical diversity within fixed-size windows. Similarity analysis further showed corresponding human and model answers shared more vocabulary than mismatched answers, while retaining substantial differences in wording. These differences were sufficiently systematic to enable highly accurate classification of answer sources using simple TF-IDF representations and a logistic regression model, achieving an overall accuracy of 0.9633 under question-grouped cross-validation. Accuracy remained at 0.8058 after matching answer lengths, indicating that source-associated lexical differences persisted beyond the original length differences.

These findings contribute to emerging research on the evaluation of generative information retrieval systems by demonstrating that LLM-generated responses can exhibit detectable linguistic patterns in Swahili agricultural question answering. The comparison experiments further identify how verbosity and lexical choices contribute to source detection. In agricultural information systems, where conversational assistants are increasingly explored as tools for delivering agronomic guidance, understanding these differences is important for designing reliable and transparent information access mechanisms. Detectability may also support practical applications such as auditing, monitoring, or labeling machine-generated advisory content in deployed agricultural support systems. While the study focuses on a relatively small dataset and a specific agricultural domain, it highlights the importance of examining how generative models express agricultural information and respond to requests for concise explanations in low-resource languages. Future research could extend this work by analyzing larger multilingual agricultural datasets, incorporating addi-

tional LLM prompting approaches, and exploring alternative semantic similarity methods based on contextual embeddings. Such efforts would contribute to a more comprehensive understanding of how generative conversational systems can be effectively integrated into agricultural information retrieval environments.

Funding: This research received no external funding.

Institutional Review Board Statement: The field data used in this study were collected under an official research clearance issued by Sokoine University of Agriculture (SUA), Tanzania, through the Office of the Vice-Chancellor, on behalf of the Tanzania Commission for Science and Technology. The details of the research clearance are as follows: Institution: Sokoine University of Agriculture (SUA) Research Clearance Reference: SUA/DPRTC/R/126/CONAS/13/2023/3; Approval/Clearance Date: 10 August 2023; Research Project: Integrated IoT, AI and a Swahili Chatbot: Agri-Tech Solution; Approved Research Period: August 2023–May 2024; Study Area: Morogoro Region, Tanzania. The present manuscript uses data collected under this approved research activity. The participant component involved the collection of agricultural questions relating to maize, and no sensitive personal, medical, or biological information was collected or analyzed in the present study.

Informed Consent Statement: Participant consent was waived due to the non-human nature of the survey; questions focused strictly on maize crops rather than individual personal data.

Data Availability Statement: The data presented in this study will be made available upon request by other researchers or for non-commercial reasons.

Acknowledgments: This research received support through Schmidt Sciences, LLC.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Kpodo, J.; Nejadhashemi, A.P. Navigating challenges/opportunities in developing smart agricultural extension platforms: Multimedia data mining techniques. *Artif. Intell. Agric.* **2025**, *15*, 426–448. [CrossRef]
2. Gayo, L.; Manyama, F.F.; Nyongesa, J.M. Climate variability and wildlife conflicts amplify poverty and food insecurity in East Africa. *Discov. Environ.* **2025**, *3*, 181. [CrossRef]
3. Ires, I. Intensive Agriculture as Climate Change Adaptation? Economic and Environmental Tradeoffs in Securing Rural Livelihoods in Tanzanian River Basins. *Front. Environ. Sci.* **2021**, *9*, 674363. [CrossRef]
4. Kitole, F.A.; Komba, C. Challenges and opportunities of climate change adaptation strategies among smallholder horticultural farmers in rural Tanzania. *Cogent Food Agric.* **2025**, *11*, 2554247. [CrossRef]
5. Shin, S.; Ichihara, M.; Sokourenko, K.; Liao, C. Everyday climate adaptation practices in agriculture contribute to food security in Sub-Saharan Africa. *Ecol. Soc.* **2024**, *29*, 32. [CrossRef]
6. Beyene, A.D.; Kassie, M. Speed of adoption of improved maize varieties in Tanzania: An application of duration analysis. *Technol. Forecast. Soc. Change* **2015**, *96*, 298–307. [CrossRef]
7. Manda, J.; Feleke, S.; Mutungi, C.; Tufa, A.H.; Mateete, B.; Abdoulaye, T.; Alene, A.D. Assessing the speed of improved postharvest technology adoption in Tanzania: The role of social learning and agricultural extension services. *Technol. Forecast. Soc. Change* **2024**, *202*, 123306. [CrossRef]
8. Vijayvargia, A.; Nagpal, A.; Pundalik, K.; Savarkar, A.; Gautam, S.; Singh, P.; Saluja, R.; Ramakrishnan, G. Intent Aware Context Retrieval for Multi-Turn Agricultural Question Answering. *arXiv* **2025**, arXiv:2508.03719.
9. Singh, N.; Wang'ombe, J.; Okanga, N.; Zelenska, T.; Repishti, J.; Mishra, S.; Manokaran, R.; Singh, V.; Rafiq, M.I.; Gandhi, R.; et al. Farmer.Chat: Scaling AI-Powered Agricultural Services for Smallholder Farmers. *arXiv* **2024**, arXiv:2409.08916.
10. Malatji, M. Bridging the AI divide in sub-Saharan Africa: Challenges and opportunities for inclusivity. *arXiv* **2026**, arXiv:2601.06145.
11. Mollel, M.; Quiroz, L.F.; Varley, C.; Firestine, A.; McLoughlin, M.E.; Kafunah, J.; Kharkar, S.; O'Farrell, J.; Ndlovu, N.; Johnston, A.; et al. Digital technologies to accelerate the impact of climate smart agriculture by next-generation farmers in Africa. *Front. Sustain. Food Syst.* **2025**, *9*, 1462328. [CrossRef]
12. Zhong, T.; Yang, Z.; Liu, Z.; Zhang, R.; You, W.; Liu, Y.; Sun, H.; Pan, Y.; Li, Y.; Zhou, Y.; et al. Opportunities and Challenges of Large Language Models for Low-Resource Languages in Humanities Research. *arXiv* **2026**, arXiv:2412.04497.
13. Agbeyangi, A.; Suleman, H. Advances and Challenges in Low-Resource-Environment Software Systems: A Survey. *Informatics* **2024**, *11*, 90. [CrossRef]

14. Anglin, K.L.; Wong, V.C.; Boguslav, A. A Natural Language Processing Approach to Measuring Treatment Adherence and Consistency Using Semantic Similarity. *AERA Open* **2021**, *7*, 23328584211028615. [CrossRef]
15. Xiang, L.; Li, N.; Liu, Y.; Hu, J. AI-Generated Text Detection: A Comprehensive Review of Active and Passive Approaches. *Comput. Mater. Contin.* **2026**, *86*, 5. [CrossRef]
16. Basani, A.R.; Chen, P.-Y. Diversity Boosts AI-Generated Text Detection. *arXiv* **2026**, arXiv:2509.18880.
17. Turney, P.D.; Pantel, P. From Frequency to Meaning: Vector Space Models of Semantics. *J. Artif. Intell. Res.* **2010**, *37*, 141–188. [CrossRef]
18. Barakabitze, A.A.; Kitindi, E.J.; Sanga, C.; Shabani, A.; Philipo, J.; Kibirige, G. New Technologies for Disseminating and Communicating Agriculture Knowledge and Information: Challenges for Agricultural Research Institutes in Tanzania. *Electron. J. Inf. Syst. Dev. Ctries.* **2015**, *70*, 1–22. [CrossRef]
19. Gandhi, R.; Veeraraghavan, R.; Toyama, K.; Ramprasad, V. Digital Green: Participatory Video and Mediated Instruction for Agricultural Extension. *Inf. Technol. Int. Dev.* **2009**, *5*, 1–15.
20. Mary, L.; Jacob, W.; Alice, N.; Annette, O. Mobile Phone-Enabled Services Relevant for Accessing Agricultural Information by Smallholder Farmers in Bungoma County, Kenya. *Afr. J. Empir. Res.* **2023**, *4*, 800–812. [CrossRef]
21. Djido, A.; Zougmoré, R.B.; Houessionon, P.; Ouédraogo, M.; Ouédraogo, I.; Seynabou Diouf, N. To what extent do weather and climate information services drive the adoption of climate-smart agriculture practices in Ghana? *Clim. Risk Manag.* **2021**, *32*, 100309. [CrossRef]
22. Zuluaga-Gomez, J.; Veselý, K.; Szöke, I.; Blatt, A.; Motlicek, P.; Kocour, M.; Rigault, M.; Choukri, K.; Prasad, A.; Sarfjoo, S.S.; et al. ATCO2 corpus: A Large-Scale Dataset for Research on Automatic Speech Recognition and Natural Language Understanding of Air Traffic Control Communications. *arXiv* **2023**, arXiv:2211.04054.
23. Bajaj, P.; Campos, D.; Craswell, N.; Deng, L.; Gao, J.; Liu, X.; Majumder, R.; McNamara, A.; Mitra, B.; Nguyen, T.; et al. MS MARCO: A Human Generated MAchine Reading Comprehension Dataset. *arXiv* **2018**, arXiv:1611.09268.
24. Gitau, C.; Kabongo, S.; Modupe, A.; Marivate, V.; Sindane, T.; Lastrucci, R.; Dzingirai, I.J. Masakhane Web: A Machine Translation Platform for African Languages. *Research Square* **2023**. [CrossRef]
25. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
26. Ray, P.P. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys. Syst.* **2023**, *3*, 121–154. [CrossRef]
27. Pande, A.; Patil, R.; Mukkemwar, R.; Panchal, R.; Bhoite, S. Comprehensive Study of Google Gemini and Text Generating Models: Understanding Capabilities and Performance. *Grenze Int. J. Eng. Technol.* **2024**, *10*, 856.
28. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971.
29. Sinha, A.; Garcia, D.W. Application of Conversational Artificial Intelligence for Farmer's Advisory and Communication. In *Artificial Intelligence and Smart Agriculture Applications*; Auerbach Publications: Boca Raton, FL, USA, 2022; pp. 63–86. [CrossRef]
30. Rezayi, S.; Liu, Z.; Wu, Z.; Dhakal, C.; Ge, B.; Dai, H.; Mai, G.; Liu, N.; Zhen, C.; Liu, T.; et al. Exploring New Frontiers in Agricultural NLP: Investigating the Potential of Large Language Models for Food Applications. *IEEE Trans. Big Data* **2024**, *11*, 1235–1246. [CrossRef]
31. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–55. [CrossRef]
32. Boudourides, M. Structural Hallucination in Large Language Models: A Network-Based Evaluation of Knowledge Organization and Citation Integrity. *arXiv* **2026**, arXiv:2603.01341.
33. Nguyen, D.C.; Welch, C. Generative Artificial Intelligence in Qualitative Data Analysis: Analyzing—Or Just Chatting? *Organ. Res. Methods* **2026**, *29*, 3–39. [CrossRef]
34. Wu, J.; Yang, S.; Zhan, R.; Yuan, Y.; Wong, D.F.; Chao, L.S.; Wong, D.F. A survey on llm-generated text detection: Necessity, methods, and future directions. *Comput. Linguist.* **2025**, *51*, 275–338. [CrossRef]
35. Gehrmann, S.; Strobelt, H.; Rush, A.M. GLTR: Statistical Detection and Visualization of Generated Text. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Florence, Italy, 28 July–2 August 2019.
36. Bohr, J. Show and Tell: Prompt Strategies for Style Control in Multi-Turn LLM Code Generation. *arXiv* **2025**, arXiv:2511.13972.
37. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
38. Al-Rfou, R.; Choe, D.; Constant, N.; Guo, M.; Jones, L. Character-Level Language Modeling with Deeper Self-Attention. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
39. Majumdar, P. Type-Token Ratio (TTR)—A Measure of Lexical Diversity. Swiss School of Business and Management Geneva. 2025. Available online: https://www.researchgate.net/publication/393146462_Type-Token_Ratio_TTR_-_A_Measure_of_Lexical_Diversity (accessed on 15 October 2026).

40. Salton, G.; Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [CrossRef]
41. Widianto, A.; Pebriyanto, E. Document Similarity using Term Frequency-Inverse Document Frequency Representation and Cosine Similarity. *J. Dinda Data Sci. Inf. Technol. Data Anal.* **2024**, *4*, 149–153. [CrossRef]
42. Thrun, S.; Ahmad, R. Semantic Analysis in NLP: Leveraging Machine Learning and Data Mining for Knowledge Map Construction. 2024. Available online: https://www.researchgate.net/publication/382910017 (accessed on 15 October 2026). [CrossRef]
43. Siino, M.; Tinnirello, I.; La Cascia, M. From foundations to GPT in text classification: A comprehensive survey on current approaches and future trends. *Found. Trends Inf. Retr.* **2025**, *19*, 557–711. [CrossRef]
44. Bhushan, S.; Thomas, D.R.; Borchers, C.; Raghuvanshi, I.; Abboud, R.; Gatz, E.; Gupta, S.; Koedinger, K.R. Detecting LLM-Generated Short Answers and Effects on Learner Performance. *arXiv* **2025**, arXiv:2506.17196.
45. Degachi, C.; Freire, S.K.; Niforatos, E.; Kortuem, G. Understanding Mental Models of Generative Conversational Search and The Effect of Interface Transparency. *arXiv* **2025**, arXiv:2506.03807.
46. Adams, R.; Alayande, A.; Brey, Z.; Browning, B.; Gastrow, M.; Kponyo, J.J.; Mathew, D.; Nkosi, M.; Nunoo-Mensah, H.; Nyakundi, D.; et al. A new research agenda for African generative AI. *Nat. Hum. Behav.* **2023**, *7*, 1839–1841. [CrossRef] [PubMed]
47. Mannuru, N.R.; Shahriar, S.; Teel, Z.A.; Wang, T.; Lund, B.D.; Tijani, S.; Pohboon, C.O.; Agbaji, D.; Alhassan, J.; Galley, J.; et al. Artificial intelligence in developing countries: The impact of generative artificial intelligence (AI) technologies for development. *Inf. Dev.* **2023**, *41*, 1036–1054. [CrossRef]
48. Lu, D.; Jeuring, J.; Gatt, A. Evaluating LLM-Generated Versus Human-Authored Responses in Role-Play Dialogues. In Proceedings of the 18th International Natural Language Generation Conference, Hanoi, Vietnam, 29 October–2 November 2025.
49. Doughman, J.; Afzal, O.M.; Toyin, H.O.; Shehata, S.; Nakov, P.; Talat, Z. Exploring the Limitations of Detecting Machine-Generated Text. In *Proceedings of the 31st International Conference on Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025.
50. Juzek, T.S.; Ward, Z.B. Why Does ChatGPT "Delve" So Much? Exploring the Sources of Lexical Overrepresentation in Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025.
51. Macko, D.; Moro, R.; Uchendu, A.; Lucas, J.; Yamashita, M.; Pikuliak, M.; Srba, I.; Le, T.; Lee, D.; Simko, J.; et al. MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark. 2023. Available online: https://github.com/openai/gpt-2-output-dataset (accessed on 15 October 2026).
52. Nguyen, V.; Karimi, S.; Hallgren, W.; Prakash, M. Question Answering in Climate Adaptation for Agriculture: Model Development and Evaluation with Expert Feedback. In *Findings of the Association for Computational Linguistics: ACL 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.