# Leveraging Synthetic Data for Question Answering with Multilingual LLMs in the Agricultural Domain

**Anonymous ACL submission**

## Abstract

Enabling farmers to access accurate agriculture-related information in their native languages in a timely manner is crucial for the success of the agriculture field. Publicly available general-purpose Large Language Models (LLMs) typically offer generic agriculture advisories, lacking precision in local and multilingual contexts. Our study addresses this limitation by generating multilingual (English, Hindi, Punjabi) synthetic datasets from agriculture-specific documents from India and fine-tuning LLMs for the task of question answering (QA). Evaluation on human-created datasets demonstrates significant improvements in factuality, relevance, and agricultural consensus for the fine-tuned LLMs compared to the baseline counterparts.

## 1 Introduction

Agriculture forms the backbone of many economies, specially for countries such as India. Despite substantial advancements of Artificial Intelligence across areas such as healthcare and education (Zhu et al., 2024), agriculture sector has seen relatively limited integration of cutting-edge technologies such as Large Language Models (LLMs). General-purpose LLMs have been shown to provide useful assistance in agricultural settings, but their advice is often generic, lacking in local specificity (Ibrahim et al., 2023; Calone et al., 2024). Agriculture-specific chat-bot systems such as Farmer.Chat (Singh et al., 2024) that have multilingual support are promising, however there has been no quantitative evaluation of such systems. On the other hand, the handful of agriculture-specific Question Answer (QA) datasets are only for high resource languages such as English, German and Chinese (Dofitas et al., 2025; Jiang et al., 2025; Kästing and Hänig, 2024).

In this paper, we present a novel synthetic QA dataset in three languages - English, Hindi and Punjabi, for the agriculture domain in India. With this synthetic training data, we carry out two different fine-tuning experiments (i) train and test LLMs with language-specific data (ii) train with English data and use translate-test approach (Ahuja et al., 2023), where the questions in Hindi/Punjabi are first translated to English, fed to the LLM trained with English data, and translating the generated responses back to Hindi/Punjabi.

For evaluation, we provide human-curated datasets for all three languages. We carry out extensive human evaluations to determine the best LLM, as well as the prompt. Contrary to prior research that used either metrics such as ROGUE (Johnny and Jaya Nirmala, 2022; Jiang et al., 2025) or LLM-as-judge (Yang et al., 2024; Gauba et al., 2025) for agriculture QA evaluation, we employ human evaluation at each step.

Llama-3x instruct models emerged as the best for zero shot testing. Although they showed over 95% for consensus, relevancy and factuality results were rather low, with the average across languages being 44.3% and 29%, respectively. However, after fine-tuning with synthetic datasets, these averages increased by 22.1% and 14% (respectively). The translate-test approach further increased results for Hindi and Punjabi.

Our contributions are as follows: a synthetic training dataset for three languages, a human curated evaluation dataset, and a human evaluation on multiple LLMs and prompts, followed by a discussion on the inadequacy of automated evaluation schemes for domain- and country-specific QA evaluation tasks.

## 2 Dataset Creation

### 2.1 Synthetic Training Data Generation

The agricultural data used to generate QA pairs was collected through web scraping from Vikaspedia[1], an online encyclopedia created by the Government

[1] https://en.vikaspedia.in/

of India. It contains data across diverse topics such as fisheries and women in agriculture. A summary of data source files and token statistics across these topics is presented in Table 8 in Appendix A.

The raw text was split into chunks of 2000 characters each, with an overlap of 200 characters between consecutive chunks. These chunk and overlap sizes were selected empirically by experimenting with multiple chunk sizes (250–2000 characters) and overlap settings (50–300 characters); smaller chunks degraded QA quality by limiting information inclusion, while larger overlaps showed no further gains.

Paula et al. (2024) showed that QA generation by prompting an LLM in a few-shot setting is more effective than approaches such as end-to-end generation (sequence of questions and answers is generated given the input context) and pipeline methods (answers are extracted first and questions are then generated in an answer-aware manner) and was therefore used for QA generation in our work.

We experimented with *DeepSeek-R1-Distill-Llama-70B* (Guo et al., 2025) and *Llama-3.1 405B* to create synthetic QA pairs.We opted for Deepseek because of its better grammar and overall fluency. It provides more diverse question formulations and better ways to phrase questions and answers as compared to LLama 405 Billion Model. (Appendix B shows sample QA pairs generated by each LLM).

Figure 2 in Appendix C shows the prompt we designed to generate QA pairs from a given text segment of fixed chunk size using few-shot prompting. This prompt design was inspired by several past research (Balaguer et al., 2024; Yehudai et al., 2024). Contrary to previous research that instructed the LLM to generate a fixed number of questions (Paula et al., 2024), we instruct it to generate as many questions as possible. This caters to the varying information density in the text chunks.

A total of 60,130 question-answer pairs were created from the agricultural dataset in English. This dataset was further translated into Hindi and Punjabi, using IndicTrans2 (Gala et al., 2023).

### 2.2 Evaluation Dataset

We manually created an English test dataset by collecting 1674 QA pairs from multiple sources such as Tamil Nadu Agricultural University, Union Bank of India-FAQ-Agri-Products, ICAR-NRRI, Cuttack, and ICAR-NRCP, Solapur. The links to sources are provided in Appendix D.

Additionally, a Hindi and Punjabi test dataset of 1219 QA pairs was developed by manually transcribing 50 videos from the DD Kisan YouTube channel[2]. The transcriptions were used to generate Hindi QA pairs, which were subsequently translated into Punjabi using IndicTrans2; the detailed transcription and QA generation methodology is provided in Appendix E. The Hindi and Punjabi datasets were manually cleaned and verified to ensure an accurate and high quality test dataset.

We validated the Hindi-Punjabi translation quality using human direct assessment (Bojar et al., 2016).The mean adequacy score was 95.55 with a standard deviation of 10.3. Further evaluation results and Translation error analysis are provided in Appendices F and G, respectively.

## 3 Implementation Details

### 3.1 Experiments

**Selection of LLMs:** LLMs differ in their language coverage. We therefore compared different LLMs to identify the best-performing one for multilingual agricultural QA. The LLMs selected for initial zero-shot experiments are shown in Table 1.Further details about LLMs and their selection are in Appendix H.

| Punjabi | Hindi | English |
|---|---|---|
| **Llama 3.1 8B** | **Llama 3.1 8B** | Llama 3.1 8B |
| Llama 3 8B | Llama 3 8B | **Llama 3 8B** |
| Navarasa 7B | Navarasa 7B | Deepseek 8B |
| | Mistral 7B | Mistral 7B |
| | Gemma 8B | Airavata 7B |

Table 1: LLMs selected for each language. The best performing LLM for each language is in **bold**.

**Prompt Selection:** We started with Silva et al. (2023)'s prompt, which was originally designed for agriculture examination contexts. However, our long-term goal is to support farmers with advisories. Hence, we adapted that prompt and generated six different prompt variations, which are discussed in Appendix I. Zero-shot experiments were conducted to select the best prompt.

**LLM Fine-tuning:** The best performing LLM under the zero-shot setting was fine-tuned with the best prompt, using the synthetic dataset we created. We also conducted an experiment to evaluate the efficacy of the translate-test approach for Hindi and Punjabi. Implementation details and hyperparameters are in Appendix J.

[2] https://www.youtube.com/@DDKisan

| No. | Parameter | Evaluation Question | Scale |
|---|---|---|---|
| 1 | Agricultural Consensus (Con) | How does the answer relate to the consensus of the agricultural domain? | Yes / No |
| 2 | Relevancy (Rel) | Does the answer provide information pertinent to the question, regardless of whether the facts are entirely accurate? | 1 – 5 |
| 3 | Factuality (Fact) | Is the response generated correct, partially correct, or incorrect? Does it contain any content it shouldn't? | Correct / Incorrect / Partially Correct |

Table 2: Evaluation criteria for human assessment

| Setting | LLMs/ Prompts | Temp | Unique QA pairs | Total Evals |
|---|---|---|---|---|
| *Multilingual LLM Comparison* | | | | |
| English | 4 LLMs | 3 | 20 | 720* |
| Hindi | 5 LLMs | 1 | 100 | 1500* |
| Punjabi | 3 LLMs | 1 | 100 | 900* |
| English | 1 LLM (Llama 3.1) | 1 (0.2) | 20 | 20 |
| *Prompt Selection (1 LLM)* | | | | |
| English | 6 Prompts | 1 | 20 | 120 |
| *Base vs. Finetuned Evaluation* | | | | |
| English | 2 (Base+FT-EN) | 1 | 300 | 600 |
| Hindi | 2 (Base+FT-HI) | 1 | 240 | 480 |
| Punjabi | 2 (Base+FT-PA) | 1 | 240 | 480 |
| *Translate Test Approach (via FT-EN model)* | | | | |
| Hindi | 1 (FT-EN) | 1 | 240 | 240 |
| Punjabi | 1 (FT-EN) | 1 | 240 | 240 |
| **Total Evaluations** | | | | **5,300** |

Table 3: Number of QA samples used for human experiments. Temperature values 0.2, 0.5, 0.7 were tested only for English model selection.*Each QA pair was independently evaluated by 3 annotators.

### 3.2 Evaluation Metrics

LLM response to each question was evaluated by human evaluators using three metrics (see Table 2): Agricultural Consensus, Relevancy, and Factuality. These metrics are inspired by similar evaluation methods used in prior research (Singhal et al., 2023; Balaguer et al., 2024).

A subset of the QA pairs (see Table 3) was evaluated by three annotators. Multiple raters/measurements (Koo et al., 2016) was used to calculate inter-annotator agreement for relevancy and factuality. Fleiss Kappa (Fleiss et al., 1971) was used for agriculture consensus. Human evaluator details and evaluation guidelines are outlined in Appendix K. Why these measures were selected and how to interpret the results are in Appendix L. Inter-rater agreement is reported in Table 4. Agreement for consensus is almost perfect, while agreement for relevancy and factuality show moderate to good agreement.

## 4 Results

From each QA pair, the question was used in the prompt, and the corresponding answer generated by the LLM was evaluated against the ground-truth answer by a human. Due to the scarcity of human evaluators, it was not possible to evaluate all the LLMs across the full test set. Table 3 shows the number of human evaluations carried out.

| Metric | Calculation | English | Hindi | Punjabi |
|---|---|---|---|---|
| Consensus | Fleiss' $\kappa$ | 1.0 | 1.0 | 0.95 |
| Relevancy | ICC | 0.79 | 0.62 | 0.77 |
| Factuality | ICC | 0.82 | 0.93 | 0.75 |

Table 4: Inter-rater agreement for the three languages.

### 4.1 Model and Prompt Selection

Among the six tested prompts, Prompt 1 (see appendix M) achieved the highest scores for English with 100%, 95.0% and 55.0% in consensus, relevancy and factuality (respectively) with Llama-3 model and was therefore used in the subsequent experiments. Results for all the prompts are in Table 14 of Appendix N.

Results in Table 5 show that *Llama-3-8B-Instruct*, *Llama-3.1-8B-Instruct* and *Llama-3.1-8B-Instruct* emerged as the best LLMs for English, Hindi and Punjabi, respectively, at temperature T = 0.2. Experiments were also conducted at T = 0, 0.5 and 0.7 for English (for results see Table 15 and 16 of Appendix N). Temperature was varied as models exhibit different optimal generation settings. For instance, DeepSeek (2025) recommends 0.0 for coding/math but higher values (1.3–1.5) for conversation or creative tasks, and some models are also sensitive to prompt style or domain.

It was observed that for English, *Llama-3* produced the most accurate and relevant responses and was free from repetition or gibberish content. *Airavata* often produced overly brief answers with repeated points, *Mistral* recalled information accurately but hallucinated incorrect facts, and *DeepSeek* responses were unnecessarily long. For Hindi and Punjabi, some models returned incomplete answers and code-mixed text. Punjabi responses also occasionally contained Shahmukhi (Urdu script) instead of Gurmukhi for certain words. While these issues were also observed in *Llama-3.1*, they occurred significantly less frequently compared to other models.

| Lang. | LLM | Con.(%) | Rel.%) | Fact.(%) |
|---|---|---|---|---|
| English | Airavata-7B-Instruct | 100.0 | 58.3 | 30.0 |
| | DeepSeek-8B-Instruct | 100.0 | 33.3 | 28.3 |
| | **Llama-3-8B-Instruct** | **100.0** | **98.3** | **66.6** |
| | Llama-3.1-8B-Instruct | 95.0 | 58.3 | 16.6 |
| | Mistral-7B-Instruct | 98.3 | 93.3 | 30.0 |
| Hindi | Gemma-7B-Instruct | 90 | 26.3 | 14.6 |
| | Llama-3-8B-Instruct | 84.0 | 25.3 | 13.3 |
| | **Llama-3.1-8B-Instruct** | **96.0** | **74.6** | **35.6** |
| | Mistral-7B-Instruct | 100.0 | 37.3 | 5.6 |
| | Navarasa-7B-Instruct | 91.0 | 58.6 | 21.0 |
| Punjabi | Llama-3-8B-Instruct | 87.0 | 17.6 | 12.3 |
| | **Llama-3.1-8B-Instruct** | **100.0** | **43.3** | **24.3** |
| | Navarasa-7B-Instruct | 95.6 | 29.3 | 15.6 |

Table 5: Human evaluation results for model selection across languages at temperature $T = 0.2$

| Lang | Consensus(%) | | | Relevancy(%) | | | Factuality(%) | | |
|---|---|---|---|---|---|---|---|---|---|
| | BASE | FT | TT | BASE | FT | TT | BASE | FT | TT |
| En | 98.0 | **99.3** | – | 32.3 | **70.6** | – | 32.6 | **55.3** | – |
| Hi | **97.5** | 93.3 | 96.2 | 62.5 | 66.6 | **74.1** | 28.7 | 38.3 | **42.0** |
| Pa | 95.0 | 85.4 | **96.6** | 38.3 | 62.5 | **72.9** | 25.8 | 35.4 | **37.9** |

Table 6: Comparison of Base, Fine-Tuned (FT) and Translate-test with FT-EN LLM (TT) performance.

### 4.2 Base vs. Fine-Tuned LLM Results

For each language, the LLM that provided the best result in the zero-shot setup was used for fine-tuning experiments. Results are in Table 6. A notable improvement was observed for relevancy and factuality across all languages after fine-tuning, despite a drop in performance for consensus for Hindi and Punjabi. The results also indicate that the English fine-tuned LLM in translate-test approach outperformed the Hindi and Punjabi fine-tuned LLMs. Results are the best for English, followed by Hindi and Punjabi. This is because LLMs are heavily trained on English data, and the translate-test approach allows the model to operate in its strongest language setting. This could be a result of differing data amounts used in LLM pre-training as stated by (Ahuja et al., 2023).

## 5 Challenges with LLM-as-a-Judge Evaluation

In the above experiments, 5,280 human evaluations were done. Although human evaluations are the gold standard for assessing LLM performance, they are time consuming and expensive (Tan et al., 2025). As a solution, we employed the "LLM-as-a-Judge" approach for evaluation. We used three instruction-tuned LLMs, namely Llama-4-Scout-17B-16E-Instruct, Llama-4-Maverick-17B-128E-Instruct, and DeepSeek-R1-Distill-Llama-70B[3]. Then we generated separate prompts, each corresponding to an evaluation metric (See Appendix O for prompts), which were adapted from G-eval (Liu et al., 2023). The same QA pairs used for human evaluation were evaluated using LLM judges.

As illustrated by examples provided in Table 17 of Appendix P , the LLM judge misclassifies LLM response as partially correct. Bertscore, ROUGE and BLEU values are inconsistent, making it difficult to reach a judgment. Correlation between LLM judges and humans is shown in Table 7. Humans and the LLM Judges had a near-perfect agreement for consensus (see Appendix Q). However, even the best result produced by the LLM judges showed a week or moderate correlation for both relevance and factuality, The factuality metric was assessed on a 3-point categorical scale whereas the relevance metric was a Likert scale (1-5). As shown in previous research (Shen et al., 2023; Liu et al., 2023), LLM judges often tend to concentrate on a single rating on wide scales, leading to low variance and weak correlation with human judgment. In our case, the majority of LLM judgments were concentrated at the highest Likert point (5) and resulted in low correlation for relevance (0.18) as compared to factuality (0.46), questioning the validity of LLM-as-a-Judge for domain, language and country-specific QA tasks.

| Lang. | LLM | Rel. (Spearman $\rho$) | Fact. (Spearman $\rho$) |
|---|---|---|---|
| En | DeepSeek | 0.23 | 0.27 |
| | L4-Scout | 0.18 | **0.46** |
| | L4-Maverick | **0.24** | 0.37 |
| Hi | DeepSeek | **0.49** | 0.44 |
| | L4-Scout | 0.45 | **0.49** |
| | L4-Maverick | **0.49** | 0.45 |
| Pa | DeepSeek | 0.35 | 0.23 |
| | L4-Scout | **0.40** | **0.39** |
| | L4-Maverick | 0.39 | 0.38 |

Table 7: Correlation between LLMs and humans.

## 6 Conclusion

In this work, we introduced a synthetic fine-tuning dataset as well as a human-curated evaluation dataset in English, Hindi, and Punjabi for QA in the Indian agricultural context. Our experiments demonstrate that fine-tuned LLMs significantly outperform their base counterparts, underscoring the value of synthetic data. Furthermore, for Hindi and Punjabi, the translate-test approach yields better performance than directly fine-tuning LLMs in those languages.

[3]We excluded larger LLMs like *Llama 4 Behemoth* (288B) and *DeepSeek-R1* (685B) due to resource constraints.

## 7 Limitations

There are several limitations to our work. First, the dataset created and the methodology used in this study include only textual data. However, in real-world scenarios, farmers may have image-based queries such as pictures of crops or pests/diseases. It is not addressed by our current approach. Our future work will focus on building a multimodal and multilingual dataset that can handle both textual and visual input. It will also incorporate a wider range of Indian languages. Second, human evaluation was done only on a subset of the dataset, due to the resource-intensive nature of the task. In future, a more comprehensive human evaluation across the entire dataset is necessary to ensure LLM's efficacy and deployment. Another limitation is that our experiments were conducted using relatively smaller LLMs (7B to 8B parameter range) due to limited computational resources. In future, the effect of using larger LLMs on performance can be explored. Our current work does not account for dynamic contextual factors such as weather conditions or soil health, which are often important in generating accurate and dynamic advisories. Integrating such contextual data through various APIs in an agentic framework can enable adaptive and real-time responses and it is an important direction for our future work. As the next step, we plan to further validate our results through field testing with farmers. We also plan to explore techniques such as RAG or LLM agents to get up-to-date information on current weather, crop prices, etc.

## 8 Ethical Considerations

The human annotators volunteered their time to contribute to the project, and no financial compensation was provided. Our approach follows the principles of participatory research (Nekoto et al., 2020; Gamage et al., 2025), and data creators have been offered coauthorship to compensate for their contributions. We crawled the data by adhering to the robots.txt specifications for each website.

This study uses data obtained from publicly accessible Government of India websites, including Vikaspedia and official ICAR portals, all of which are properly cited. The intellectual property rights of the content remain with the respective original sources. The use of the data complies with the fair dealing provisions of the Indian Copyright Act, which permit the use of copyrighted material for non-commercial research purposes. To ensure compliance , we have also sought approvals from all the sources. We have already received approval for Vikaspedia, which was used for creating fine tuning data and can be freely re-distributed. The evaluation datasets will be released publicly upon receiving the necessary approvals from the respective data providers.

## References


Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. Mega: Multilingual evaluation of generative ai. In *EMNLP 2023*.

Iñaki Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938.

Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O. Nunes, Rafael Padilha, Morris Sharp, Bruno Silva, Swati Sharma, Vijay Aski, and Ranveer Chandra. 2024. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture. *Preprint*, arXiv:2401.08406.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, Marco Turchi, and Marcos Zampieri. 2016. Findings of the 2016 conference on machine translation (wmt16). In *Proceedings of the First Conference on Machine Translation*, pages 131–198. Association for Computational Linguistics.

Riccardo Calone and 1 others. 2024. Analysing the potential of chatgpt to support plant disease risk forecasting systems. *Scientific Reports*, 14:2089.

Inc. DeepSeek. 2025. The temperature parameter. Accessed: 2025-01-01.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*. Oral presentation.

Cyreneo Dofitas, Yong-Woon Kim, and Yung-Cheol Byun. 2025. Advanced agricultural query resolution using ensemble-based large language models. *IEEE Access*.

Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Bodapati, and Dan Roth. 2025. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation

and llm-as-a-judge. In *International Conference on Learning Representations (ICLR) 2025*. Published in ICLR 2025 proceedings.

Joseph L. Fleiss and 1 others. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Cheng Chen, and Shashi Bhushan Tn. 2023. Are large language models reliable judges? a study on the factuality evaluation capabilities of LLMs. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 310–316, Singapore. Association for Computational Linguistics.

Jay Gala, Pranjal A. Chitale, A. K. Raghavan, Varun Gumma, Sumanth Doddapaneni, Aswanth Kumar, Janki Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. arXiv preprint arXiv:2305.16307. Submitted 25 May 2023; revised 20 Dec 2023.

Omega Gamage, Surangika Ranathunga, Annie Lee, Xiao Sun, Aryaveer Singh, Marjana Prifti Skenduli, Mehreen Alam, Ajit Kumar Nayak, Haonan Gao, Barga Deori, and 1 others. 2025. A multilingual dataset (multimwp) and benchmark for math word problem generation. *IEEE Transactions on Audio, Speech and Language Processing*.

Aruna Gauba, Irene Pi, Yunze Man, Ziqi Pang, Vikram S Adve, and Yu-Xiong Wang. 2025. Agmmu: A comprehensive agricultural multimodal understanding and reasoning benchmark. *arXiv preprint arXiv:2504.10568*.

Google DeepMind. 2025. Gemini 2.0 flash (model `gemini-2.0-flash`). Google Cloud Vertex AI / Gemini API. Model released February 5, 2025; details in Google documentation and blog :contentReference[oaicite:3]index=3.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

H.Y. Ibrahim and 1 others. 2023. Evaluating responses by chatgpt to farmers' questions on irrigated lowland rice cultivation in nigeria. *AI for Good Foundation Research Series*.

Jingchi Jiang, Lian Yan, Haifeng Liu, Zhenbo Xia, Haotian Wang, Yang Yang, and Yi Guan. 2025. Knowledge assimilation: Implementing knowledge-guided agricultural large language model. *Knowledge-based systems*, 314:113197.

Swapna Johnny and S. Jaya Nirmala. 2022. Farmer query answering system. *SN Computer Science*, 3(1):45–60.

Marvin Kästing and Christian Hänig. 2024. Assessing large language models in the agricultural sector: A comprehensive analysis utilizing a novel synthetic benchmark dataset. In *INFORMATIK 2024*, pages 1279–1286. Gesellschaft für Informatik eV.

Ting-Ting K. Koo, Mei-Yang Li, and 1 others. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

J.Ř̃ichard Landis and G.G̃ary Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, and 1 others. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160.

Felipe Paula, Cassiana Roberta Lizzoni Michelin, and Viviane Moreira. 2024. Evaluation of question answer generation for Portuguese: Insights and datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5315–5327, Miami, Florida, USA. Association for Computational Linguistics.

Karl Pearson and 1 others. 1895. Contributions to the mathematical theory of evolution. II: Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society A*, 186:343–414.

Pratik Rohera, Chaitanya Ginimav, Aniket Salunke, Ganesh Sawant, and Raviraj Joshi. 2024. L3Cube-IndicQuest: A benchmark question answering dataset for evaluating knowledge of LLMs in indic context. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 982–988. Tokyo University of Foreign Studies.

Chen Shen, Lian Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, Singapore. Association for Computational Linguistics.

Bruno Silva, Leonardo Nunes, Roberto Estevão, Vijay Aski, and Ranveer Chandra. 2023. Gpt-4 as an agronomist assistant? answering agriculture exams using large language models. *arXiv preprint arXiv:2310.06225*.

Namita Singh, Jacqueline Wang'ombe, Nereah Okanga, Tetyana Zelenska, Jona Repishti, Sanjeev Mishra, Rajsekar Manokaran, Vineet Singh, Mohammed Irfan Rafiq, Rikin Gandhi, and 1 others. 2024. Farmer. chat: Scaling ai-powered agricultural services for smallholder farmers. *arXiv preprint arXiv:2409.08916*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating llm-based judges. In *International Conference on Learning Representations (ICLR) 2025*. Published in ICLR 2025 proceedings.

Shanglong Yang, Zhipeng Yuan, Shunbao Li, Ruoling Peng, Kang Liu, and Po Yang. 2024. Gpt-4 as evaluator: Evaluating large language models on pest management in agriculture. arXiv preprint (arXiv:2403.11858). Version 1, Mar 18, 2024.

Asaf Yehudai, Boaz Carmeli, Yosi Mass, Ofir Arviv, Nathaniel Mills, Assaf Toledo, Eyal Shnarch, and Leshem Choshen. 2024. Genie: Achieving human parity in content-grounded datasets generation. *arXiv preprint arXiv:2401.14367*.

Yukun Zhu, Wei Wang, and Yitong Zhang. 2024. Harnessing the power of llms for education. *arXiv preprint arXiv:2401.00580*.

# A Data Source

| Topic | Files | Tokens |
|---|---|---|
| Agri Exports | 9 | 24,813 |
| Agri Inputs | 60 | 105,348 |
| Agri Insurance | 9 | 36,643 |
| Agro enterprises | 31 | 101,939 |
| Best Practices | 75 | 174,997 |
| Crop Production | 637 | 2,250,709 |
| Fisheries | 112 | 544,672 |
| Forestry | 18 | 69,916 |
| ICT applications_in_Agriculture | 18 | 36,664 |
| Livestock | 107 | 415,639 |
| Market information | 13 | 32,483 |
| National Schemes_for_Farmers | 17 | 72,449 |
| Policies and Schemes | 86 | 213,880 |
| Post Harvest Technologies | 41 | 158,399 |
| Poultry | 25 | 75,997 |
| State specific schemes for farmers | 50 | 117,642 |
| Women and agriculture | 42 | 87,372 |
| **Total** | **1350** | **4,519,562** |

Table 8: Token and topic statistics across different topics in the Vikaspedia dataset.

# B Synthetic QA pair comparison between Llama and DeepSeek

| Model | Question and Answer |
|---|---|
| Llama | **Q:** How can I manage thrips in groundnut?<br>**A:** Apply Fipronil 2 ml/litre. |
| DeepSeek | **Q:** What should I do if my groundnut crop shows yellowish-green patches and brown necrotic areas?<br>**A:** This could be a sign of thrips infestation. Apply Fipronil 2 ml/litre to control the infestation. |

Table 9: Comparison of a single QA pair generated by Llama 3.1 and DeepSeek models.

# C Prompt for synthetic data generation

The prompt used for generation of synthetic data question-answer pairs.

```
Instruction:
Suppose you are a knowledgeable expert in agriculture.

Please generate as many unique sets of corresponding
question and answer pairs strictly based on the provided
Context.
No additional context is required.
Context:
{context}
Response:
Question: [Question here]
Answer: [Answer here]
Examples:
{examples}
```

Figure 1: Prompt used for synthetic data generation. During prompting, the `context` tag is replaced with a fixed input chunk from the agricultural dataset.

# D Evaluation Dataset Creation

The evaluation dataset used in this study for English, Hindi and Punjabi languages was curated by Kashish Mahendra, Vidhi and Sudhir Rajput. It was manually extracted and verified from the websites given in the Table 12

| Source and Weblink |
|---|
| Tamil Nadu Agricultural University: http://www.agritech.tnau.ac.in/expert_system/ |
| Union Bank of India: https://www.unionbankofindia.co.in/pdf/faq-agri-products.pdf |
| ICAR-NRRI, Cuttack: https://icar-nrri.in/faq/ |
| ICAR-NRCP, Solapur: https://nrcpomegranate.icar.gov.in/FAQ |
| DD Kisan YouTube Channel: https://www.youtube.com/watch?v=MYkBswpxRcA |

Table 10: Sources and their corresponding web links used for evaluation dataset creation.

# E Hindi and Punjabi QA generation methodology

A Hindi test dataset of 1219 QA pairs was developed by transcribing 50 videos from the DD Kisan youTube channel[4]. The transcription was performed in a two step process. First, the audio from the publicly available videos was sourced automatically using Youtube DATA API and then transcribed using Assembly AI API. Subsequently, each transcription was manually corrected by a native Hindi speaker. GenAI 2.0 Flash (Google DeepMind, 2025) was employed to generate QA pairs from 500 token segments of the transcript, resulting in 1,219 Hindi QA pairs. Subsequently, these were

[4] https://www.youtube.com/watch?v=MYkBswpxRcA

translated into Punjabi using IndicTrans2 (Gala et al., 2023).

## F Translation Quality Assessment for Hindi to Punjabi

The quality of the translation was carefully validated. Following the Direct Assessment methodology described in (Bojar et al., 2016), we evaluated the Hindi–Punjabi translation quality using a random sample of 65 sentences. A proficient Punjabi speaker assessed translation adequacy on a 0–100 scale based on how well each translation preserved the meaning of the corresponding Hindi reference. The mean adequacy score was 95.55 with a standard deviation of 10.3. Z-scores were computed for standardization, with the mean Z-score being approximately zero ($\approx -3.56 \times 10^{-10}$), confirming correct normalization. This evaluation was conducted by a professional, indicating that the translated dataset is of high and reliable quality.

## G Translation Error Analysis and Quality Control

There were several issues that we encountered during the translation process, especially Hindi to Punjabi. **Some words were mistranslated due to lexical similarity.** There were few instances where the numbers were misrepresented or altered during translation.

**Repetition:** There were cases where some part of the context was translated correctly but at a certain point one word or phrase repeated continuously until the end of the translation or was truncated prematurely.

**Improper Translation of Proper Nouns:** Names of companies or brands were translated literally, when they should have remained unchanged.

**Punctuation Errors:** The sentence was entirely translated correctly but ended with repeated periods.

However, to ensure the quality of the translated dataset, all the observed issues were addressed manually by native Hindi and Punjabi speakers. Sentences with severe errors or poor quality were removed to prevent introducing noise into the dataset. A quantitative breakdown of these errors is left for future work.

## H LLM Details

Large language models often fail to generalize uniformly across languages, particularly when trained primarily on high-resource languages like English (Ahuja et al., 2023; Rohera et al., 2024; Alonso et al., 2024). (Ahuja et al., 2023) empirically demonstrate that model performance on low-resource languages is often substantially lower than on high-resource languages. This indicates that the same model cannot reliably perform equally across English and non-English languages, justifying the selection of different models optimized for each language to ensure robust performance in both high and low resource settings.

During the initial experimentation phase, we evaluated a total of seven models: DeepSeek 8B, Navarasa 7B, Airavata 7B, Llama 3-8B, Llama 3.1-8B, Mistral 7B, and Gemma 7B. Due to resource constraints to perform human evaluations. We shortlisted 4 models [Airavata 7B, Llama 3 B, Deepseek 8B, Mistral 7B] among them based on our subjective observations of quality of responses such as repetition, irrelevant responses. Following the initial model selection, Llama 3.1-8B was additionally evaluated at the best-performing temperature setting ($T = 0.2$) to enable a direct comparison with the best-performing model, Llama 3. Based on this evaluation, Llama 3.1-8B was subsequently included in the main manuscript, and its results are reported for completeness.

Similar approach was followed for Hindi, out of the 5 models selected for English, we picked Llama 3 8B and Mistral 7B because they were the best performing models for English and included 3 more models i.e. **Llama 3.1 8B, Navarasa 7B and Gemma 8B** because of their coverage of Indic languages.

For Punjabi, we selected the top 3 best performing Hindi models, again due to limited human evaluations that we could perform. Please note that our decisions were motivated to optimise our resources and performance of models. For completeness, evaluating the same models across all languages would be desirable; however, this was beyond our available computational resources.

| LLM Name | Size | Section(s) Used |
|---|---|---|
| Llama-3-8B-Instruct | 8.03B | LLM Selection (EN, HI, PA)<br>Fine-tuning (EN)<br>Prompt Selection |
| Mistral-7B-Instruct-v0.3 | 7.25B | LLM Selection (EN, HI) |
| Llama-3.1-8B-Instruct | 8.03B | LLM Selection (HI, PA)<br>Fine-tuning (HI, PA) |
| Indic-gemma-7b-finetuned-sft-Navarasa | 7B | LLM Selection (HI, PA) |
| DeepSeek-R1-Distill-Llama-8B | 8.03B | LLM Selection (EN) |
| Airavata-7B-Instruct | 6.87B | LLM Selection (EN) |
| gemma-7b-it | 7B | LLM Selection (HI) |
| DeepSeek-R1-Distill-Llama-70B | 70.6B | Synthetic QA Generation<br>LLM-as-a-Judge |
| Llama-4-Scout-17B-16E-Instruct | 109B | LLM-as-a-Judge |
| Llama-4-Maverick-17B-128E-Instruct | 402B | LLM-as-a-Judge |

Table 11: List of LLMs used in the study, their parameter sizes, and the sections they were used in. HI: Hindi, PA: Punjabi, EN: English.

# I Prompt Selection

This section provides the six tested prompts used for evaluating Mistral-7B in the Prompt Selection experiment and the design intuition behind them. Each prompt was designed to optimize the quality of the response for agricultural question-answer generation.

### Prompts Tested

- **Prompt 1 (Selected Prompt)**

  *As an experienced agronomist proficient in farming techniques, crop management, and disease-resistant crop cultivation, you are tasked with answering questions based on your expertise. Questions may be provided in various languages such as English, Hindi, Marathi, etc.*

  *Provide only the answer relevant to the question. Do not include any other information. The answer should be in the same language as the question.*

- **Prompt 2**

  *You are an expert in agricultural science with over 25 years of experience in farming, crop management, and sustainable agricultural practices. Your expertise spans across various domains, including soil health, irrigation techniques, pest control, and the integration of modern technology in agriculture. Your task is to answer questions related to the agriculture domain in a clear, concise, and informative manner. Ensure your responses are well-researched, accurate, and tailored to the specific context of the question. Questions may be provided in various languages such as English, Hindi and Marathi. Provide only an answer relevant to the question.*

- **Prompt 3**

  *You are an expert in the field of agriculture with decades of experience in farming, crop management, soil science, and agricultural technology. Your knowledge spans traditional and modern farming practices, and you are fluent in multiple languages, including English, Hindi, Marathi, and others. Your expertise allows you to provide accurate, practical, and culturally relevant advice tailored to the specific needs of farmers, researchers, and agricultural enthusiasts.Your task is to provide detailed, accurate, and actionable answers exclusively related to the agriculture domain. Your responses must be in the same language as the question, whether it is in English, Hindi, Marathi, or any other language. Ensure your answers are clear, concise, and directly address the query while incorporating relevant agricultural principles, practices, and innovations.*

- **Prompt 4**

  *You are a seasoned agricultural expert with over 25 years of experience in farming, crop management, and sustainable agricultural practices. Your deep understanding of soil health, pest control, irrigation systems, and modern farming technologies has made you a trusted advisor to farmers, policymakers, and agricultural organizations worldwide. Your task is to answer a specific question related to the agriculture domain. The question will be provided to you, and your response should be detailed, accurate, and tailored to the context of the query. Ensure your answer is backed by scientific evidence and a clear understanding of the latest advancements in agriculture.*

- **Prompt 5**

  *You are an expert in agriculture and tasked with answering questions related to agriculture.*
  *Please answer the following question with detailed, accurate, and insightful information based on your knowledge in the agricultural domain.*

- **Prompt 6**

  *You are an expert agricultural advisor with deep knowledge of farming, soil health, pest control, and modern techniques. Provide a clear and well-explained response to the following question, ensuring accuracy, practicality, and ease of understanding. If multiple solutions exist, explain the best approach considering efficiency, cost-effectiveness, and sustainability. Factor in seasonal changes, soil conditions, and climate impact when relevant.*

**Prompt Intuition**

- **Prompt 1 (Selected):** This prompt was Designed to deliver concise and contextually relevant answers in the same language as the input, ensuring minimal drift from the question's intent.

- **Prompt 2:** This prompt was designed to simulate a highly knowledgeable agricultural expert, embedding deep domain-specific authority into the LLM's persona. The aim was to guide the LLM to generate precise, well-informed, and context-aware answers like domain experts.

- **Prompt 3:** This prompt was designed to extract detailed, accurate, and context-sensitive responses by positioning the LLM as a multilingual agricultural expert. Emphasis was placed on reasoning, cultural relevance, and linguistic alignment to ensure comprehensive, well-researched answers tailored to diverse user needs.

- **Prompt 4:** This prompt was designed to position the LLM as an advisor trusted by farmers and policymakers alike, having scientific expertise with leadership, so that it can provide context-rich, evidence-based answers rooted in both practical farming knowledge and awareness of modern agricultural advancements.

- **Prompt 5:** This prompt was kept minimal to evaluate the LLM's inherent capabilities without heavy role conditioning, by providing only a brief instruction focused on expected output quality.

- **Prompt 6:** The intent of this prompt was to encourage comparative reasoning and decision-making within itself. Instruct it to consider multiple solutions and select the best one based on accuracy and practicality. Ensuring responses are optimized for real-world agricultural conditions, including seasonal, soil, and climate factors.

## J Hyperparameters and Implementation Details

Table 12: Fine-tuning hyperparameters for **Llama-3-8B-Instruct** using QLoRA.

| Parameter | Value |
|---|---|
| Base LLM | Llama-3-8B-Instruct |
| LoRA Rank ($r$) | 16 |
| LoRA Alpha | 16 |
| Target Modules | q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj |
| Batch Size | 2 |
| Gradient Accumulation Steps | 4 |
| Max Sequence Length | 2048 |
| Learning Rate | 2e-4 |
| Epochs | 1 |
| Optimizer | AdamW (8-bit) |
| Quantization | 4-bit (bnb) |
| Checkpointing | Enabled (Unsloth) |
| Scheduler | Linear with warm-up |
| Tracking | Weights & Biases |
| Tokenizer | Llama 3 Instruct tokenizer |

Table 13: Fine-tuning hyperparameters for **Llama-3.1-8B-Instruct** using QLoRA.

| Parameter | Value |
|---|---|
| Base LLM | Llama-3.1-8B-Instruct |
| LoRA Rank ($r$) | 16 |
| LoRA Alpha | 16 |
| Target Modules | q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj |
| Batch Size | 2 |
| Gradient Accumulation Steps | 4 |
| Max Sequence Length | 2048 |
| Learning Rate | 2e-4 |
| Epochs | 1 |
| Optimizer | AdamW (8-bit) |
| Quantization | Not Used |
| Checkpointing | Enabled (Unsloth) |
| Scheduler | Linear with warm-up |
| Tracking | Weights & Biases |
| Tokenizer | Llama 3 Instruct tokenizer |

We employed Parameter-Efficient Fine-Tuning (PEFT) using the QLoRA method (Dettmers et al., 2023) for adapting both the Llama-3-8B-Instruct and Llama-3.1-8B-Instruct LLMs. PEFT was implemented via the Unsloth framework, which enables memory-efficient training through 4-bit quantization (used only for Llama-3) and Low-Rank Adaptation (LoRA). Gradient checkpointing was

enabled to reduce memory usage during backpropagation.

Fine-tuning followed an Instruct-style prompt format and was performed using Hugging Face's `SFTTrainer` utility.

## K Human Evaluation Framework

This appendix details the evaluation framework used to assess the quality of LLM-generated answers in the agricultural domain. Evaluators are native speakers of Hindi or Punjabi and are also fluent in English. The human evaluation was conducted by annotators without formal agricultural training; however, they participated in multiple piloting and training rounds to familiarize themselves with the evaluation criteria and task setup. Following this process, they were able to perform the evaluation reliably. The annotators were provided with detailed human evaluation framework guidelines and had access to the original source documents from which the question–answer pairs were derived, allowing them to verify model outputs accurately during the evaluation. The evaluation is conducted following three criteria:

- **Agricultural Consensus:** Does the response correlate with commonly accepted agricultural knowledge?
- **Relevancy:** Does the response demonstrate evidence of correct recall of knowledge (e.g., correct facts)?
- **Factuality:** Is the generated response factually accurate, partially correct, or incorrect?

**1. Agricultural Consensus**

Responses are labeled as *Yes* or *No* depending on whether they are related to the agricultural domain.

- **Yes:** Indicates the answer is clearly associated with agriculture or farming practices, regardless of its correctness.
- **No:** Indicates the answer is irrelevant to agriculture.

**Examples:**

- **Question:** How to improve soil condition?
  **Predicted Answer:** There are several ways to improve soil condition such as use of organic matter, crop rotation, and maintaining pH of soil.
  **Agricultural Consensus:** Yes
- **Question:** What are some recommended places to purchase apples from?
  **Predicted Answer:** The Apple Store is the best place. They are very knowledgeable, sell and service all their devices, sell AppleCare and you can be hands-on with the computer before you buy it.
  **Agricultural Consensus:** No

**2. Relevancy**

Responses are rated on a scale of 1–5 based on their relevance to the question. A score of 5 means highly relevant and accurate; a score of 1 means minimal understanding or relevance.

**Examples:**

- **Question:** How to improve soil condition?
  **Predicted Answer:** There are several ways to improve soil condition such as use of organic matter, crop rotation, and maintaining pH of soil.
  **Relevancy Score:** 5
- **Question:** What are some recommended places to purchase apples from?
  **Predicted Answer:** The State of Jammu and Kashmir and Himachal Pradesh are the leading apple producing States in India.
  **Relevancy Score:** 2

**3. Factuality**

Responses are compared with reference answers and categorized as:

- **Correct:** Closely matches the reference answer.
- **Partially Correct:** Contains some correct information but has inaccuracies.
- **Incorrect:** Includes major inaccuracies or irrelevant information.

**Examples:**

- **Question:** How to improve soil condition?
  **Reference Answer:** Techniques include use of organic matter, mulching, avoiding overtillage, crop rotation, using cover crops, balancing soil pH, and using appropriate fertilizers.
  **Predicted Answer:** There are several ways to improve soil condition such as use of organic

matter, crop rotation, and maintaining pH of soil.
**Factuality:** Correct

- **Question:** Which season is suitable for apple plantation in Himachal Pradesh?
  **Reference Answer:** February to early May
  **Predicted Answer:** The best time to plant apple trees in Himachal Pradesh is between January and March.
  **Factuality:** Partially Correct

- **Question:** What is the full form of GMO?
  **Reference Answer:** Genetically Modified Organisms
  **Predicted Answer:** General Motors Operation
  **Factuality:** Incorrect

## L Inter-annotator Agreement Calculation

Since the same fixed set of evaluators assessed a specific, non-random sample set for each language, the two-way mixed effects, absolute agreement, multiple raters/measurements model was used to calculate agreement scores (Koo et al., 2016). According to standard interpretation, ICC values below 0.5 indicate *poor reliability*, values between 0.5 and 0.75 suggest *moderate reliability*, values between 0.75 and 0.9 denote *good reliability*, and values above 0.9 reflect *excellent reliability*. According to the interpretation scale proposed by (Landis and Koch, 1977) , the strength of agreement for Kappa statistic can be categorized with values less than 0 indicate *Poor* agreement, 0–0.20 as *Slight*, 0.21 – 0.40 as *Fair*, 0.41 – 0.60 as *Moderate*, 0.61 – 0.80 as *Substantial*, and 0.81 – 1.00 as *Almost Perfect*. For binary judgments related to agricultural consensus, Fleiss Kappa (Elangovan et al., 2025) was used to evaluate the reliability between evaluators.

$$\text{ICC} = \frac{MS_R - MS_E}{MS_R + \frac{MS_C - MS_E}{n}} \quad (1)$$

$MS_R$ = row mean square (subjects), $MS_C$ = column mean square (raters), $MS_E$ = error mean square, $n$ = number of raters per subject.

## M Prompt Response generation

```
Instruction:
As an experienced agronomist proficient in farming
techniques, crop management, and disease-resistant crop
cultivation, you are tasked with answering questions
based on your expertise. Questions may be provided in
various languages such as English, Hindi, Marathi, etc.
Provide only the answer relevant to the question. Do
not include any other information. Answer should be in
the same language as the question.

Input:
Question: {input}

Response:
```

Figure 2: Prompt used for answer generation for a given question.

## N Temperature and Prompt Selection

| Prompt Number | Consensus | Relevancy | Factuality |
|---|---|---|---|
| **Prompt 1** | **100.0%** | **95.0%** | **55.0%** |
| Prompt 2 | 100.0% | 80.0% | 50.0% |
| Prompt 3 | 100.0% | 85.0% | 45.0% |
| Prompt 4 | 100.0% | 85.0% | 50.0% |
| Prompt 5 | 100.0% | 100.0% | 40.0% |
| Prompt 6 | 100.0% | 75.0% | 40.0% |

Table 14: Results for the prompts for English.

| Model | Agr. Consensus | | Relevance | | Factuality | |
|---|---|---|---|---|---|---|
| | $T = 0.2$ | $T = 0$ | $T = 0.2$ | $T = 0$ | $T = 0.2$ | $T = 0$ |
| Airavata | 100.0 | 90.0 | 58.3 | 43.3 | 30.0 | 20.0 |
| Mistral 7B v0.3 | 98.3 | 95.0 | 93.3 | 75.0 | 30.0 | 40.0 |
| Llama 3 8B Instruct | 100.0 | 95.0 | 98.3 | 76.6 | 66.6 | 43.3 |
| DeepSeek 8B Distill | 100.0 | 95.0 | 33.3 | 66.6 | 28.9 | 40.0 |
| Llama 3.1 8B Instruct | 95.0 | – | 58.3 | – | 16.6 | – |

Table 15: Performance comparison of LLMs evaluated at temperature settings $T = 0.2$ and $T = 0$ across agriculture consensus, relevance, and factuality metrics.

| LLM | T = 0.2 | | | T = 0.5 | | | T = 0.7 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Con | Rel | Fact | Con | Rel | Fact | Con | Rel | Fact |
| DeepSeek R1 | 100.0% | 33.3% | 28.3% | 100.0% | 80.0% | 31.6% | 100.0% | 70.0% | 21.6% |
| **Llama 3** | **100.0**% | **98.3**% | **66.6**% | 100.0% | 93.3% | 56.6% | 100.0% | 95.0% | 51.6% |
| Mistral | 100.0% | 93.3% | 35.0% | 100.0% | 75.0% | 45.0% | 95.0% | 66.6% | 55.0% |
| Airavata | 100.0% | 58.3% | 30.0% | 100.0% | 95.0% | 41.6% | 100.0% | 90.0% | 40.0% |

Table 16: English LLM evaluation across 3 different temperature ($T$) settings using metrics: Agriculture Consensus (Con), Factuality (Fact), and Relevancy (Rel). The highest percentage is in bold.

## O LLM as Judge Prompts

### Prompt for Agriculture Consensus

**Instruction:**
You will be given a question, a reference (gold) answer, and an LLM-generated answer.
Your task is to rate the LLM-generated responses on one metric.
Please make sure you read and understand these instructions carefully.

Evaluation Criteria:
Agricultural Consensus (Yes,No) - To assess the results generated by the LLM, we categorize them with labels of Yes or No depending on their correlation with the agriculture sector.
A Yes label denotes a clear association with agricultural or farming activities, regardless of the accuracy of the response.
Conversely, a No label indicates a lack of relevance to agricultural domain knowledge.
NOTE: The generated response may not be accurate, but it should be relevant to the agriculture field.

**Evaluation Steps:**
1. Understand the Purpose:
   - Recognize that the goal is to determine whether the LLM-generated response is relevant to the agriculture or farming domain, not whether it is factually accurate.

2. Read the Question:
   - Review the question to understand the context in which the LLM was expected to respond. This can provide cues about whether the topic is related to agriculture.

3. Analyze the Reference Answer:
   - Briefly examine the reference (gold) answer to see what type of response is expected and whether it is centered around agricultural knowledge or practices.

4. Review the LLM-Generated Answer:
   - Carefully read the generated answer to determine the domain it is addressing. Identify key terms, phrases, or themes that indicate its association (or lack thereof) with agriculture or farming.

5. Assess Domain Relevance:
   - Check for direct or indirect connections to agricultural topics such as crops, soil, irrigation, farming techniques, pest control, agricultural policy, rural economy, livestock, climate effects on farming, etc.

6. Make a Binary Decision:
   - If the generated answer clearly pertains to the agriculture sector, assign the label "Yes".
   - If the generated answer is unrelated to agriculture, farming, or associated practices, assign the label "No".

7. Avoid Judging Accuracy:
   - Do not consider factual correctness when assigning the label. The only criterion is whether the content is related to the agriculture domain.

**Inputs:**
Question: {question}

Reference (Gold) Answer:
{answer}

LLM-Generated Answer:
{answer_llm}

## Prompt for Factuality Evaluation

**Instruction:**
You will be given a question, a reference (gold) answer, and an LLM-generated answer.
Your task is to rate the LLM-generated responses on one metric.
Please make sure you read and understand these instructions carefully.

**Evaluation Criteria:**
Factuality (Correct, Partially Correct, Incorrect) — The response generated by the LLM will be assessed against the reference answer to evaluate its accuracy.
It will be classified as either **"Correct"**, **"Partially Correct"**, or **"Incorrect"**.
The **Correct** designation denotes a response that closely matches the reference answer, while an **Incorrect** label indicates significant inaccuracies or inappropriate information.
**Partially Correct** refers to a response that is not entirely accurate but includes some elements or pieces of information that are correct.

**Evaluation Steps:**

1. **Understand the Question:**
   - Read the question carefully to ensure a proper understanding of what is being asked. Clarify any ambiguity in the question to accurately evaluate the generated response.

2. **Review the Reference Answer (Gold Answer):**
   - Thoroughly analyze the reference answer to understand the key details and factual information it presents. Identify the core points, facts, or context that must be preserved in the LLM-generated answer.

3. **Review the LLM-Generated Answer:**
   - Read the LLM-generated response closely. Check if the information presented aligns with the reference answer. Identify the factual elements in the response and compare them with the reference answer to spot any discrepancies or omissions.

4. **Assess Factual Accuracy:**
   - Compare each key detail in the LLM-generated answer with the reference answer.
   - If the LLM-generated answer fully matches the factual information provided in the reference answer, rate it as **Correct**.
   - If the LLM-generated answer has some elements that align with the reference answer but also contains inaccuracies or incomplete information, rate it as **Partially Correct**.
   - If the LLM-generated answer includes significant inaccuracies or irrelevant information, rate it as **Incorrect**.

5. **Contextualize the Answer:**
   - Ensure that the context of the answer (e.g., domain knowledge, specific terminology) is understood and correctly represented. If the response introduces concepts that are out of context or misleading, it should be rated as **Incorrect**.

6. **Check for Detail and Completeness:**
   - Evaluate the level of detail in the response. If critical facts or key elements are missing from the LLM's response compared to the reference answer, it should be considered **Partially Correct** or **Incorrect**, depending on the extent of the missing information.

7. **Final Rating:**
   - Assign the final rating to the LLM-generated answer based on the accuracy of the information provided compared to the reference answer.

**Inputs:**
Question: {question}

Reference (Gold) Answer:
{answer}

LLM-Generated Answer:
{answer_llm}

## Prompt for Relevance Evaluation

**Instruction:**
You will be given a question and an LLM-generated answer.
Your task is to rate the LLM-generated responses on one metric.
Please make sure you read and understand these instructions carefully.

**Evaluation Criteria:**
**Relevance (1–5)** — The response will be rated on a scale of 1 to 5 based on its relevance to the question.
A rating of **5** indicates a highly relevant and well-aligned response, while a rating of **1** indicates that the answer is largely irrelevant or mismatched with the question.
**NOTE:** The answer must align with the question asked, regardless of its factual accuracy. A response that incorporates all pertinent information and clearly addresses the core intent of the question will score higher.

**Evaluation Steps:**

1. **Understand the Question:**
   - Read the question carefully to identify the core intent and expected scope of the answer. Note keywords and concepts that define what kind of information is being asked.

2. **Read the LLM-Generated Answer:**
   - Examine the generated answer in its entirety. Determine what topic it is addressing, what information it provides, and how closely it connects to the original question.

3. **Match Answer to Question Intent:**
   - Check whether the LLM-generated answer addresses the main point of the question.
   - Consider if the answer is directly responsive to the question, or if it strays into unrelated or tangential topics.

4. **Evaluate Depth and Completeness:**
   - Assess whether the response includes not just a relevant answer but also supporting details, context, or additional information that enhances its relevance.
   - Responses that include a broader context or deeper insight into the question's topic may score higher.

5. **Rate on a Scale from 1 to 5:**
   - **5:** Fully relevant and directly answers the question with complete, well-contextualized information.
   - **4:** Mostly relevant and answers the question well, with minor gaps or slight vagueness.
   - **3:** Moderately relevant; some parts address the question, but the answer is incomplete or only partially focused.
   - **2:** Minimally relevant; contains limited information related to the question or is largely off-topic.
   - **1:** Irrelevant or barely addresses the question; shows poor understanding or misunderstanding of the question asked.

6. **Ignore Factual Accuracy:**
   - Do not factor in correctness of facts. Focus strictly on alignment with the question's subject and intent.

**Inputs:**
Question: {question}

LLM-Generated Answer:
{answer_llm}

# P Misleading Score

| Question | Gold Answer | LLM Answer | Human | LLM | BERT | ROUGE-L | BLEU |
|---|---|---|---|---|---|---|---|
| What is the cause for round scales on the under surface of leaves along with yellowing of leaves on the upper surface in coconut palm? | Scale insects reside on the under surface of leaves and cause yellowing on the upper surface of leaves. They secrete wax in circular patches. | Fungal infection *(Colletotrichum gloeosporioides)* and insect infestation *(Necrobia rufipes)* are the common causes of round scales on the under surface of leaves along with yellowing of leaves on the upper surface in coconut palm. | Incorrect | Partially Correct | 0.6710 | 0.4482 | 0.2216 |
| Is pseudostem split symptom caused by wilt pathogens? | Yes, the FOC can cause pseudostem split symptoms. Similar symptoms are also noticed with Erwinia rot disease. | No, pseudostem split symptom is not caused by wilt pathogens. It is caused by the infestation of the stem borer, *Chilo partellus*. | Incorrect | Incorrect | 0.650 | 0.153 | 0.030 |
| Is pseudostem split symptom caused by wilt pathogens? | Yes, the FOC can cause pseudostem split symptoms. Similar symptoms are also noticed with Erwinia rot disease. | Yes, it is caused by wilt pathogens. | Correct | Correct | 0.666 | 0.166 | 0.007 |
| Can I apply the entire fertilizer recommended at one time for coconut plantation? | Apply in split doses to ensure regular nutrient availability and avoid leaching. | No, apply in controlled manner over time for optimal growth. | Correct | Partially Correct | 0.621 | 0.170 | 0.019 |
| Can I apply the entire fertilizer recommended at one time for coconut plantation? | Apply in split doses to ensure regular nutrient availability and avoid leaching. | Yes, you can apply the entire fertilizer recommended at one time. | Incorrect | Incorrect | 0.567 | 0.091 | 0.000 |

Table 17: Evaluation of LLM-generated answers against gold references with human and LLM annotations along with automatic scores.

## Q LLM-as-a-Judge : Pearson Correlation for Agriculture Consensus Metric

For the agriculture consensus metric, which involves binary classification, Pearson correlation (Pearson et al., 1895; Fu et al., 2023) was considered. However, the correlation yielded a NaN value due to the absence of variance in the evaluations: both the annotators and the LLM provided identical labels for all instances, indicating complete agreement.

## R Hardware and Infrastructure

All experiments were conducted on high-performance GPU nodes equipped with NVIDIA Tesla V100 (32GB) and NVIDIA A100 (80GB) GPUs. The environment was configured to efficiently support large-scale LLM inference and fine-tuning, leveraging mixed precision and distributed processing where applicable.

| Experiment Description | Hours | Minutes |
|---|---|---|
| **Fine-tuning** | | |
| Punjabi Llama3.1 Finetuning | 3 | 12 |
| Hindi Llama3.1 Finetuning | 2 | 17 |
| English Llama3 Finetuning | 2 | 28 |
| **Model and Temperature Selection** | | |
| Airavata | 12 | 03 |
| DeepSeek | 29 | 06 |
| Llama3 | 17 | 36 |
| Mistral | 12 | 09 |
| **Fine-tuning Inference** | | |
| English | 1 | 46 |
| Hindi | 2 | 36 |
| Punjabi | 4 | 55 |
| **Individual Model Evaluations** | | |
| Llama3 (HI) | 4 | 41 |
| Llama3.1 (HI) | 7 | 54 |
| Gemma (HI) | 1 | 59 |
| Mistral (HI) | 6 | 11 |
| Navrasa (HI) | 4 | 36 |
| Punjabi (HI) | 4 | 55 |
| Llama3 (PA) | 5 | 18 |
| Llama3.1 (PA) | 6 | 21 |
| Navrasa (PA) | 1 | 47 |
| **Prompt Selection Experiments** | | |
| Prompt 1 | 5 | 55 |
| Prompt 2 | 6 | 02 |
| Prompt 3 | 6 | 44 |
| Prompt 4 | 10 | 40 |
| Prompt 5 | 11 | 43 |
| Prompt 6 | 8 | 04 |
| **Indic Translation** | 5 | 48 |
| **Total GPU Time** | **178** | **46** |

Table 18: Detailed breakdown of GPU time consumed across all experiments including fine-tuning, inference, model selection, and prompt evaluation.