Sections (minimum):

0. Abstract (150–250 words)
1. Introduction

Sustainable landscaping in the Indo-Gangetic Plains depends on a deep understanding of both the region's ecological conditions and its rich cultural traditions. However, most large language models today are trained primarily on Western ecological data, which often leads to plant recommendations that are inaccurate, invasive, or culturally disconnected from the needs of local communities. These gaps make it difficult for current AI systems to support responsible environmental planning, especially in regions where traditional ecological knowledge plays a central role. This highlights a bigger problem, which is that current LLMs don't have enough regional plant knowledge, cultural context, or traditional ecological practices to give advice that's actually safe or useful. Because of that, they can accidentally recommend plants that are harmful, invasive, or completely disconnected from the cultural traditions that shape how landscaping is done in the Indo-Gangetic Plains. Our project tackles this gap by building a culturally grounded AI system that uses community-informed data, region-specific ecological information, and a retrieval setup designed to reduce bias and generate recommendations that make sense for the people and ecosystems in this region.

2. Related Work

Across the papers Improving Human AI Partnerships in Child Welfare, WeBuildAI, and HelloAI, the common theme is that AI systems in high-stakes social environments cannot be responsibly deployed without deep attention to human values and participatory involvement. WeBuildAI shows that if algorithmic governance is solely technical, it can compromise social values and amplify inequities, which is why it proposes a participatory framework where community members provide input to shape the algorithm's guidelines. Similarly, the AFST child welfare study reveals that, especially in sensitive situations such as social work, an algorithm can never fully be used as workers have to use their expertise and judgement to make difficult decisions. Especially because the situation is high stakes, relying solely on a framework can even be dangerous, leading to harmful situations being misdiagnosed. Lastly, the Hello AI study of medical practitioners further strengthens this idea by showing that experts need onboarding that clarifies not only its strengths but also its limitations before they can safely collaborate with it. Being aware of the capabilities of the AI will allow medical practitioners to make more informed decisions and use the technology to its fullest extent. When conducting our own study, we took the findings from these papers into account and were cautious of the fact that cultural value often gets overshadowed in the algorithm, so the feedback from the embassy and locals was vital in our process. Even before we started our project, the first step was to address this issue and focus on interviews with the embassy so we could get a complete picture and accurate feedback.

The paper Man is to Computer Programmer as Woman is to Homemaker? highlights how word embeddings can amplify harmful societal biases when trained on large-scale text. The paper illustrates how the embeddings reproduce gender stereotypes, such as he = doctor and she = nurse. By identifying a "gender direction," the author reveals how deeply bias is embedded in linguistic space, and these findings of bias are a representation of the AI systems we rely on. Using this logic, we were very aware of the sources we were collecting data from and the authenticity and accuracy of the data our model was based. Even if a data set seemed trustworthy, we dived deeper into the contributors and authors to ensure that our data didn't reflect a deep-rooted bias of our outdated information.

Methods / Approach

3. Data Set / Model / **System Evaluation Method**

3.1. Data Set

As part of our data collection, we pulled from three major datasets that each brought something different and valuable to our project. The first was the Biodiversity of India database, which gave us detailed explanations of the traditional uses of each plant as well as their basic ecological needs. To make sure the information was credible, we researched the people behind the platform and found that it's maintained by a blend of researchers from major U.S. universities and Indian institutions. The lead researcher, Gaurav Moghe, is a Computational Biologist at Cornell, which added a layer of trust in both the accuracy and professionalism of the dataset. The second dataset came from the International Research Journal of Education and Technology, specifically the study titled "Studies of Some Medicinal Plants of the Indo-Gangetic Plains" by Rupam and Lalit Raj Singh from the Department of Medicinal Plant Sciences at Dev Sanskriti Vishwavidyalaya in Haridwar. This source provided incredibly rich information on the traditional and cultural uses of medicinal plants across the region, but it lacked broader ecological or cultivation data. Because of that, we paired it with a third resource, which was the Useful Tropical Plants database. Useful Tropical Plants, managed by Ken and Ajna Fern, helped fill in the ecological and botanical gaps left by the second dataset. Ken Fern is a well-known figure in the plant world because he authored Plants for a Future and even founded a charity focused on rare and unusual plant species. Once we finalized the datasets we were using, we identified the specific categories of information that would be most useful for our project. We ultimately decided to collect the following metrics for each plant: common name, scientific

name, local name (if applicable), region, climate requirements, soil type, sunlight needs, water needs, growth rate, ecological role, and traditional uses. We chose these categories because they capture both the cultural significance and the environmental adaptability of each species. This combination allows us to understand not just how plants are used traditionally, but also how they can realistically be incorporated into environmental AI systems grounded in regional ecology and local knowledge.

3.2. Methodology and Systems Design

To address the challenge of Western bias in current Large Language Models available to the public, we propose and developed a Retrieval-Augmented Generation (RAG) framework specifically grounded with data and context from flora from the Indo-Gangetic Plains region. Our framework consists of three core components: the construction of a Culturally-Grounded Knowledge Corpus, the development of a Dual-Agent RAG Architecture, and a Synthetic Evaluation Protocol using the LLM-as-a-Judge framework as an evaluation strategy.

Our Culturally Grounded Knowledge Corpus (hereby referred to as Cultural Corpus) is a custom-built dataset of flora throughout the Indo Gangetic Plains region. Rather than include basic scientific facts (that can be easily searched via the web from current LLMs), we include Cultural data gathered from sources emphasizing categories such as the Ecological Role, Traditional Uses, and more. By limiting our RAG Agent to mainly use our Corpus as a ground truth, we reduce the hallucination of Western plant species (such as suggesting English Ivy for a tropical climate), which has appeared in traditional LLM responses. We also propose this structure, compared to just asking LLMs to scrape the web before their responses, to allow plants with backgrounds that may be undercovered or unrecognized to be populated and robustly explained. We emphasize that with this structure, further populating and covering plants throughout the region will allow for a more specialized knowledge base that can be iterated on and tuned for specificity that rivals more generic and broad datasets online. This can lead to some limitations, which will be discussed further in the Discussion section.

Our main core technical development combines our previous Knowledge Corpus to implement a custom RAG pipeline, built using Python and the LangChain framework. We implement a comparative structure emphasizing our custom-built Knowledge Corpus to inject specific context into our base model’s queries. This is done through an implementation of a standard RAG structure:

We utilized ChromaDB as our local vector store. Our raw dataset, the Cultural Corpus, was ingested and embedded using OpenAI’s text-embedding-3-large model. This enabled our dataset to be queried semantically, returning data that is semantically similar to the user’s query rather than just by keyword matching. In this case, we return specific plant profiles along with our Corpus’s unique Cultural Findings into a model’s content window. This enables semantic search capabilities that allow our system to understand queries like "plants for hair health" and properly sift through and return data from our Corpus to an entry like "Acacia concinna" entirely based on traditional usage descriptions.

Using Langchain features, we also implement the Middleware Injection Pattern that uses Langchain agents. A dynamic function, prompt_with_context, performs a similarity search against our vector store and injects the retrieved context into the system prompt with strict instructions to treat the corpus as the "Single Source of Truth."

Our system utilizes GPT-4o as the generation engine. We instantiated two distinct agents for testing:

- The RAG Agent: Equipped with the retrieval middleware and a strict system prompt enforcing cultural grounding with our Cultural Corpus.
- The Baseline Agent: This is a standard out-of-the-box instance of GPT-4o with no access to the external corpus

For both agents, it is to be noted that the temperature of the models was set to 0 to be the most objective and grounded, discouraging any sort of hallucinations or creativity that would not be contextually appropriate for this study's context.

For our evaluation, we utilized a Synthetic Evaluation (LLM-as-a-Judge) framework. We designed an automated evaluation loop that processed 100 domain-specific test queries (e.g., "Which native shrubs won't attract snakes?") through both the RAG and Baseline agents. The resulting 200 responses were then blinded and graded by a Panel of Synthetic Jurors that we gave five distinct LLM personas via prompt engineering to simulate specific stakeholders.

- The Landscape Architect: Evaluates for aesthetics and infrastructure safety.
- The Local Botanist: Evaluates for scientific accuracy and correct nomenclature.
- The Cultural Historian: Evaluates for the depth of cultural context and traditional usage
- The Sustainability Officer: Evaluates for water conservation and ecological fit.
- The General Resident: Evaluates for clarity and actionable advice.

Each response was graded on a Likert Scale (1-5) across four key dimensions:

- Factual Accuracy: Is the botanical information correct?
- Ecological Suitability: Is the plant actually native/suitable for the Indo-Gangetic climate?
- Cultural Relevance: Does the answer acknowledge local traditions and names?
- Overall Helpfulness: Is the advice practical for a homeowner?

We then calculated Krippendorff's Alpha to statistically confirm the spread of graded responses and measured the Inter-Rater Reliability among the different personas.

3.3. System Evaluation Method

We employed a combination of various methods to analyze the performance of the LLM. The first evaluation will be a pairwise comparison, which involves comparing the base LLM model and the RAG model. We compiled a list of **100 queries** utilizing domain expertise to

compare both models (e.g., "Our village pond dries halfway in winter but floods in the monsoon. What native plants can survive this cycle and help keep the water clean?"). Due to the government shutdown, we were unable to implement a human feedback portion with the expertise of the US embassy employees in Nepal. Therefore, we created 5 AI personas that acted as our landscaping experts.

1. **Landscaping Architect**
2. **Local Botanist**
3. **Cultural Historian**
4. **Sustainability Officer**
5. **General Resident**

Each persona rated system outputs utilizing a Likert Scale. A Likert scale is typically a survey tactic for measuring respondents' responses in terms of the level they agree or disagree with a specific statement. The scale can range from **"Strongly Disagree," "Disagree," "Neutral," "Agree," and "Strongly Agree."** On our specific scale, we will generate plant recommendations and educational signage on four key rating scales: **1) Factual Accuracy, 2) Ecological Suitability, 3) Cultural Relevance, and 4) Overall Helpfulness.** For each of these key questions, the scale will be different for each rating.

1. **Factual Accuracy: (1=Very Inaccurate, 5=Very Accurate)**
2. **Ecological Suitability: (1=Very Unsuitable, 5=Very Suitable)**
3. **Cultural Relevance: (1=Very Irrelevant, 5=Very Relevant)**
4. **Overall Helpfulness: (1=Not Helpful, 5=Very Helpful)**

To ensure consistency among our ratings, we used **inter-rater reliability (IRR). Krippendorf's alpha** is the specific formula we will use when measuring IRR. We wanted to ensure an **Alpha (k) > 0.70, indicating substantial agreement, and above Alpha (k) >0.8 is high agreement.**

$$\alpha = 1 - \frac{D_o}{D_e}$$

D_o: observed disagreement

D_e: expected disagreement

Next, we utilized **BERTScore** to measure how much the RAG system actually changed the answer compared to No-RAG. BERTScore is basically an evaluation method that computes a similarity score for each token in a candidate sentence with each token in a reference sentence. The candidate in our case is the RAG response, and the reference is the non-RAG response. We established specific boundaries for this score. If the score is **High (0.95+), our RAG system is**

redundant. The base model already knew the answer, so retrieving the documents didn't change anything. If the score is **Lower (< 0.85)**, it means the RAG system successfully injected new, specific information that forced the model to change its answer.

Finally, we employed the Answer Relevance metric using the **RAGAS framework** to evaluate the specificity of the generated responses. Answer Relevancy strictly measures the information alignment between the query and the output. This metric operates on a "reverse-engineering" validation paradigm: it uses an LLM to generate multiple hypothetical questions based solely on the answer provided by the RAG system, and then calculates how semantically similar these generated questions are to the original user query. The metric is calculated as the mean cosine similarity between the embedding of the original question qo and the embeddings of the N generated hypothetical questions qg. The formula is defined as:

$$\text{answer relevancy} = \frac{1}{N} \sum_{i=1}^{N} cos(E_{g_i}, E_o)$$

Where **$E_{q_g,i}$** is the embedding of the **i-th generated question and E_{q_o} is the embedding of the original user query.** We established some key, specific boundaries. A high relevancy score is considered **>0.8.** This basically means it is ensured that the bot isn't just hallucinating generic facts. Anything **below 0.8 suggests the model is "waffling,"** or providing generic facts, conversational filler, or redundant information that, while potentially true, does not satisfy the specific constraints of the query.

4. Results

The performance of the **Retrieval Augmented (RAG) system** was evaluated against a **baseline LLM (No-RAG)** using a **mixed-methods approach**, as mentioned earlier. The evaluation focused on four key areas: **Factual Accuracy, Ecological Suitability, Cultural Relevance, and Overall Helpfulness.** The analysis incorporates both automated metrics and simulated human evaluation with different personas.

4.1. Pairwise Comparative Analysis

The primary evaluation utilized a 5-point Likert Scale to compare the RAG system against the baseline model across 100 domain-specific queries. These ratings were generated by a Simulated Expert Panel consisting of five distinct AI personas mentioned in section 5.3. Each persona was prompted to evaluate the outputs based on their specific domain expertise, shown in Table 1. The RAG system did showcase improvements in all four metrics.

Metric	No-RAG	RAG	Difference
Factual Accuracy	3.98	4.71	+0.73
Ecological Suitability	3.04	4.57	+1.53
Cultural Relevance	2.81	4.11	+1.30
Overall Helpfulness	3.26	4.49	+1.23

Figure 1 illustrates the mean Likert scores by numbers for both the RAG system and No-RAG scores, as well as the differences in the means for each category.

The most significant gain came in Ecological Suitability **(+1.53)**, indicating that the retrieval mechanism was grounded in specific soil and climate constraints experienced in the Indo-Gangetic Plain. In contrast, the baseline model focused on generic botanical advice. Similarly, Cultural Relevance saw the second-highest increase **(+1.30)**, suggesting the RAG system effectively retrieved the local naming conventions, folklore, and traditional use that were not incorporated in the baseline's training data.

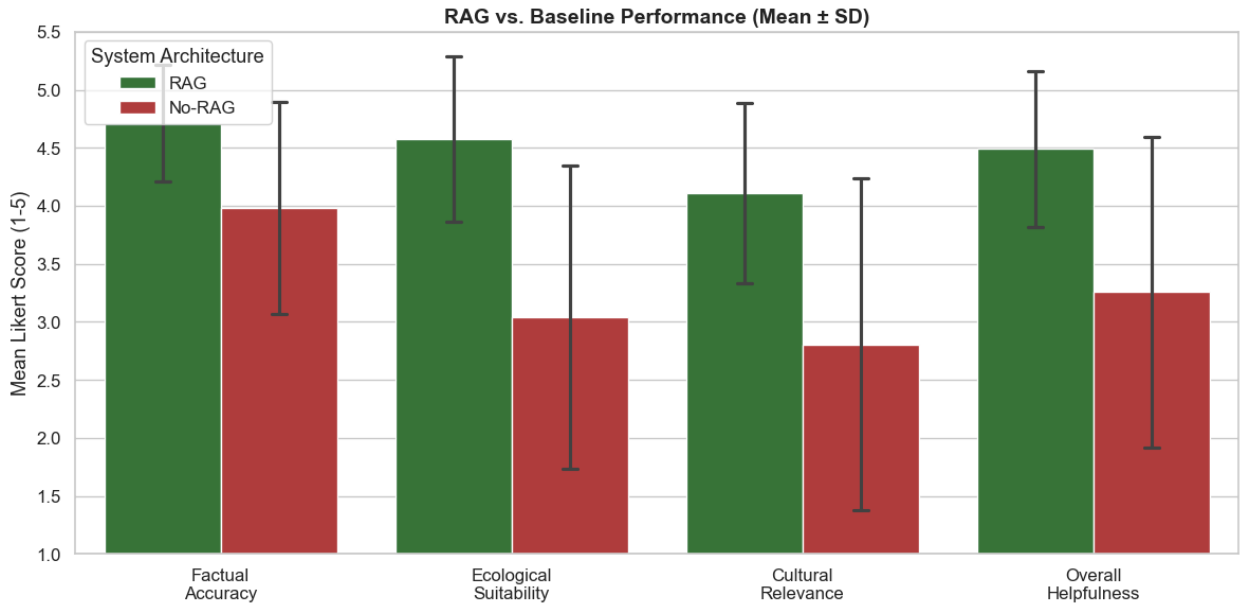

Figure 2 illustrates the comparative mean scores with standard deviation error bars, highlighting the consistent outperformance of the RAG framework.

4.2. Inter-Rate Reliability (IRR)

To validate the consistency of our "Simulated Expert Panel" (five unique AI personas), we calculated Krippendorf's Alpha (a) for each metric. An **a > 0.7** is considered acceptable reliability.

Metric	Krippendorff's Alpha (α)	Interpretation
Factual Accuracy	0.754	Substantial
Ecological Suitability	0.805	High Agreement
Cultural Relevance	0.840	High Agreement
Overall Helpfulness	0.826	High Agreement

Figure 3 illustrates the Krippendorff's Alpha values as well as the interpretation of each of these scores for each of the key metrics.

The high agreement scores for Cultural Relevance **(0.840)** and Overall Helpfulness **(0.826)** confirm that the distinct personas, despite utilizing varied prompts for each persona (e.g., Botanist vs. Historian), reached a consensus on the RAG responses being substantially better than the No-Rag responses.

4.3. Automated Evaluation Metrics

We chose to complement the Likert scale with two automated metrics to assess semantic divergence and query alignment.

1. **Semantic Divergence (BERTScore):** The system was able to achieve an average similarity score of **0.709** when comparing the RAG response to the No-RAG baseline response. In the context of BERTScore, a lower score does indicate "High Impact." A score of 0.709 indicates that the retrieval mechanism in the RAG framework was able to alter the semantic content of the response, which means new content was injected rather than a rewording of the generic response.
2. **Answer Relevance (RAGAS):** The system ended up achieving an average relevance score of **0.705.** The original target was a score >0.8, and a score of 0.705 is consistent with a conversational pedagogical agent. The metric penalizes "chattiness" (e.g., greetings, persona-based framing), which were intentional design features of the IGP Advisor persona.

4.4. Performance by Evaluator Persona

A stratified analysis of the ratings showcases how the different personas perceived the system's value.

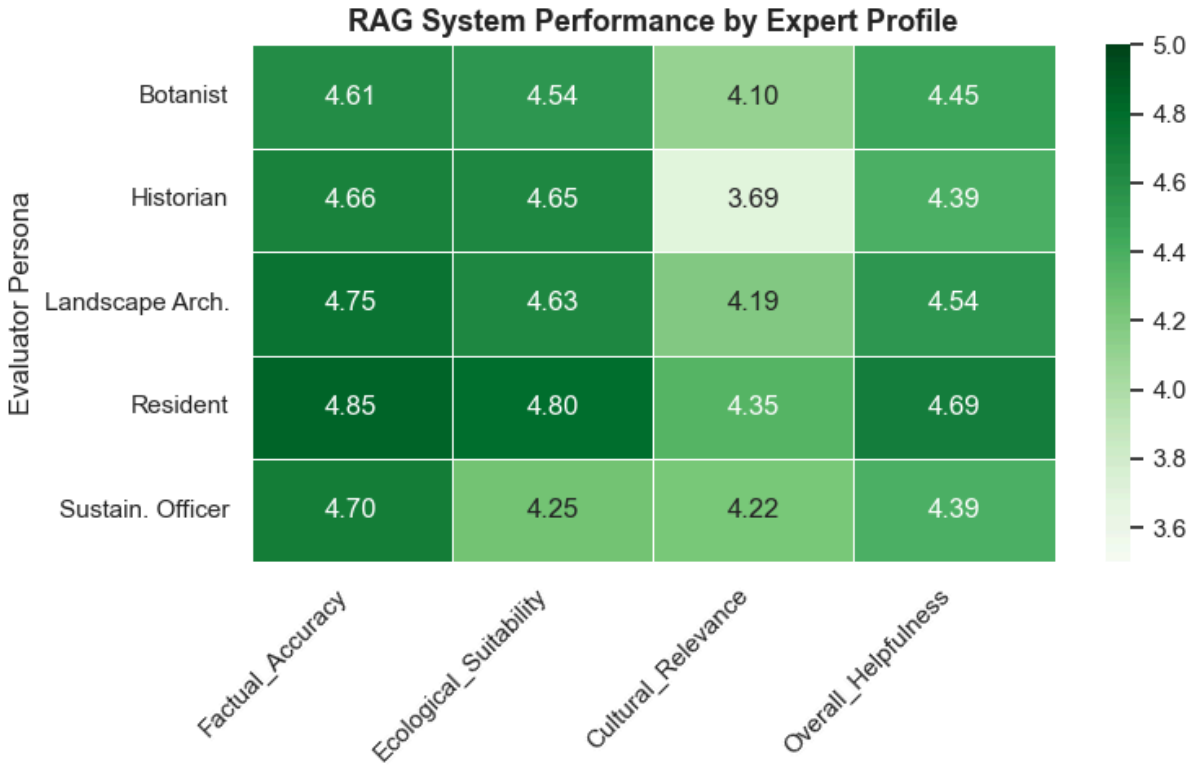

Figure 4 showcases the mean Likert scores by each of the evaluator personas for each of the key metrics.

The data did indicate that the **“Resident”** gave the highest composite ratings for the RAG System, with a **mean of 4.7**, suggesting the tool is highly effective for a regular homeowner with average gardening skills seeking practical advice. Conversely, the **"Historian"** persona recorded the largest delta in Cultural Relevance, rating the RAG system with a significantly higher mean **(3.69)** compared to the baseline mean **(2.29)**, which failed to capture local nuance. This nuance suggests that while the RAG system captures general cultural context well, a specialized Historian can still detect gaps in deep archival specificity that the other personas might overlook.

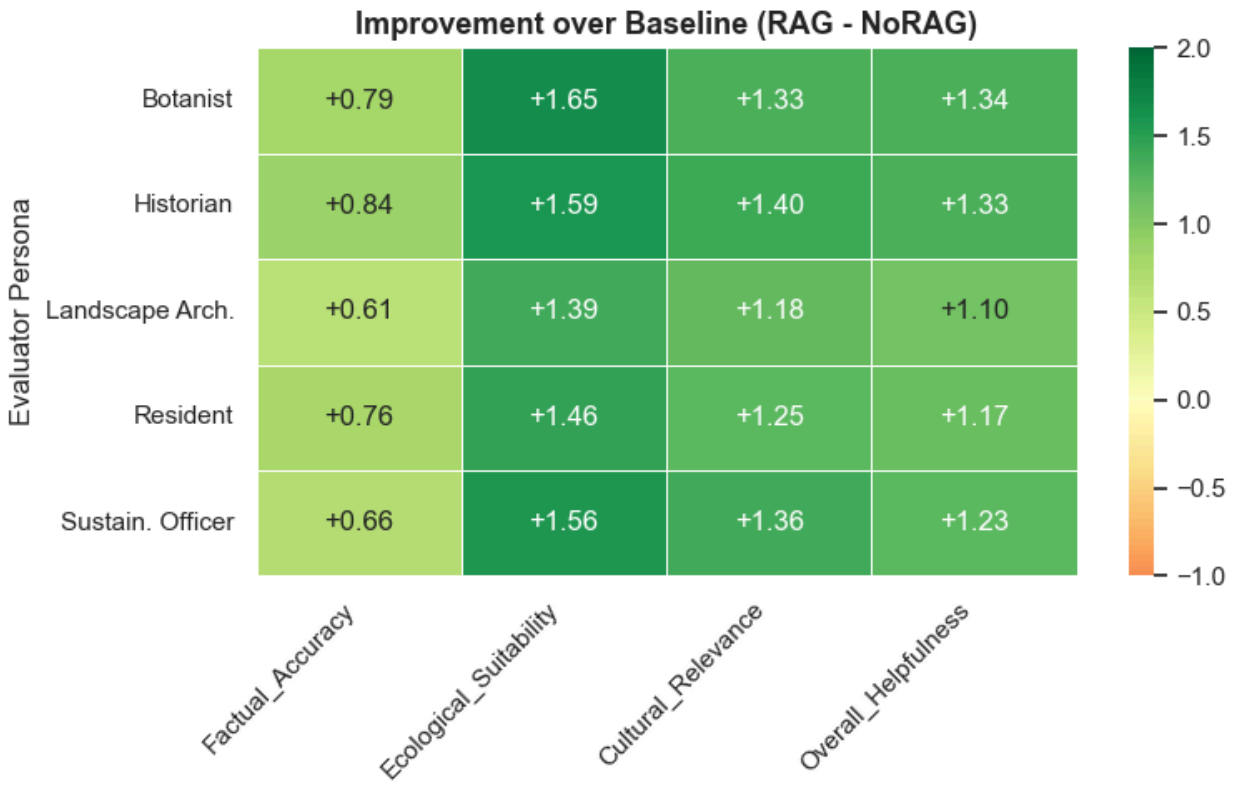

Figure 5 showcases the mean increase in Likert scores from the No-RAG system and the RAG system by each of the evaluator personas for each of the key metrics.

5. Discussion / Analysis / Interpretation of Results Future Work

With our quantitative results, our team observes that a RAG framework combined with an LLM has a distinct improvement in model performance across the categories we asked jurors to rank. Cultural context about native plants not easily researchable or "googlable" definitely are subjects that GPT (and many LLMs) seem to be prone to, and our systems clearly can improve the responses generated after our custom corpus is incorporated into their context windows. Broadly examining our results, we can see strong gains in Ecological Suitability, Cultural Relevance, and Overall Helpfulness in our Likert scores by around 30-50% (Growing from 3 to 4), while Factual Accuracy is the only field that does not grow at the same rate, shifting upwards by around 17.5%.

We deduce that this variance is related to the factual information at play for the jurors to evaluate: simple facts such as the scientific name, the region located, and the climate requirements are already the most likely of information to be present in baseline GPT models or at least the most readily researchable information that current LLMs can find out when compiling responses. As such, we should not see any vast improvements since the emphasis our Corpus and RAG Framework places on our responses is curated towards Cultural information

that is not quantifiable by a simple yes/no binary representation. This makes the RAG system's value not in teaching the model new biological facts, but rather in constraining the model to prioritize local relevance over generic Western gardening advice. This is also where we draw our reasoning for the strong growth in the 3 aforementioned categories from earlier. Diving deeper, GPT-4 and (in broader contexts) other LLMs already possess strong parametric knowledge from their training phases.

Another point of interest and contention that we wish to address is our use of LLMs as judges in our evaluation framework. Due to logistical issues in connecting with officials, we decided to use LLMs as judges to evaluate the differences in prompts. Despite this, our LLM as a judge protocol achieved a strong Krippendorff's Alpha greater than 0.8. This indicated a strong inter-rater reliability among the 5 simulated personas, suggesting that all the improvements from our RAG are consistent and actual perceived improvements across all of the different personas and their respective domains. Notably, as mentioned in an earlier section, the Cultural Historian persona recorded the highest delta in preference for the RAG system, giving further evidence that the injection of the Corpus was a probable driver of the system's improved performance.

While this synthetic evaluation loop is both scalable and provides a good stopgap solution for the current situation, our team acknowledges that there are limitations to using such a framework, especially when these frameworks are themselves evaluating AI and LLMs. There has been prior research that suggests a potential 'self-preference bias,' where models can favor outputs that are similar to their own training patterns, and although we mitigate this by creating strict persona constraints and rubrics (defined in our 5 unique agents from above), a purely synthetic evaluation is not fully sufficient to capture the true experiences of local human stakeholders in the Indo-Gangetic Plains. Our team therefore fully encourages that future iterations of this framework would necessitate and greatly benefit from a 'Human-in-the-Loop' validation phase, where local farmers and embassy staff review the highest-rated RAG outputs on their real-world case functionality.

6. Conclusion

7. Acknowledgement

We want to acknowledge Professor Neil Gaikwaid and the entire TA team for their feedback and guidance throughout this entire process. Their feedback on our proposals, presentations, and papers helped guide and shape this project in the right direction. Moreover, we wanted to acknowledge the team at the US Embassy in Nepal for answering our questions and allowing us to get a more in-depth understanding of the project and the stakeholders.

8. References (Doesn't count towards the page limit)

9. Use of ChatGPT(Doesn't count towards the page limit)

10. Appendix (Doesn't count towards the page limit)