INITIATIVE ON
Digital Innovation
CGIAR

Working Paper

Can We Trust Large Language Models to Summarize Food Policy Research Papers and Generate Research Briefs?

MinAh Kim[a,b*], Jawoo Koo[a], Yunchul Jung[a]

[a]*International Food Policy Research Institute (IFPRI), Washington, DC*
[b]*Georgetown University, Washington, DC*

INFO

Submitted	1 December 2023
Keywords	AI, Large Language Model, Policy brief
Flagship	Co-LAB
Work Package	Enabling environment

ABSTRACT

Generative large language models (LLMs), while widely accessible and capable of simulating policy recommendations, pose challenges in the assessment of their accuracy. Users, including policy analysts and decision-makers, bear the responsibility of evaluating the outcomes from these models. A significant limitation of LLMs is their potential to overlook critical, context-specific factors. For example, in formulating food policies, it is vital to consider regional climate and environmental variables that influence water and resource availability. Nonetheless, due to their reliance on word sequencing probabilities from training datasets, LLMs might propose similar policies for distinct regions. Despite these limitations, LLMs offer considerable advantages for rapid policy analysis, particularly when resources are constrained. They serve as quick, accessible, and cost-effective tools for policy research and development, requiring minimal training and infrastructure. In our study, we assessed the efficacy of LLMs in generating policy briefs by inputting an IFPRI discussion paper into three different LLM-based approaches: a standard chatbot without extra data, a Retrieval Augmented Generation model integrating semantic search with LLM, and a custom-developed Brief Generator designed to create policy summaries from AI-analyzed paper structures. Our findings revealed that none of the LLM-generated briefs fully captured the original paper's intent, underscoring the need for further research. Future investigations should focus on gathering more empirical data with diverse text types and volumes to better understand these outcomes.

1. Introduction

Generative large language models (LLMs) can imitate a policy recommendation and are largely accessible to the public. However, evaluating the validity of the result produced by these models remains in the hands of users (e.g., policy analysts and policymakers). LLMs can present a result that does not consider many factors that humans would have considered. For instance, when developing a food policy, location-specific climate and environment factors are indispensable since they determine the availability of water and natural resources in the region. Yet, LLMs may recommend nearly identical policy options for two vastly different regions due to the probabilistic nature of sequencing words found in the corpus of training data.

Despite this critical shortcoming, LLMs can still be a game changer when there is a need for conducting rapid policy analyses. When there is limited internal capacity, LLMs can effectively act as a fast, accessible, and affordable assistant that can help policy research and formulation without requiring extensive training or infrastructure for utilization.

In this study, we experimented with ingesting an IFPRI discussion paper to compare three different approaches of automatically generating a policy brief using LLM: the basic chatbot with no additional information as the baseline, a Retrieval Augmented Generation model that combines semantic search with LLM, and the custom-built Brief Generator approach that synthesize the policy brief based on the AI-analyzed structure of the paper.

2. Retrieval Augmented Generation (RAG) Implementation on a Chatbot Functionality

In this trial, we attempt to measure the improvement in policy recommendation functionality of LLMs when applying a RAG. To reach this goal, we built one simple chat model without RAG, one chat model with RAG, and brief generating applica-

* Corresponding author. Currently enrolled at Georgetown University, Washington, DC. *Email addresses:* mk2215@georgetown.edu (M. Kim), j.koo@cgiar.org (J. Koo), y.jung@cgiar.org (Y. Jung).

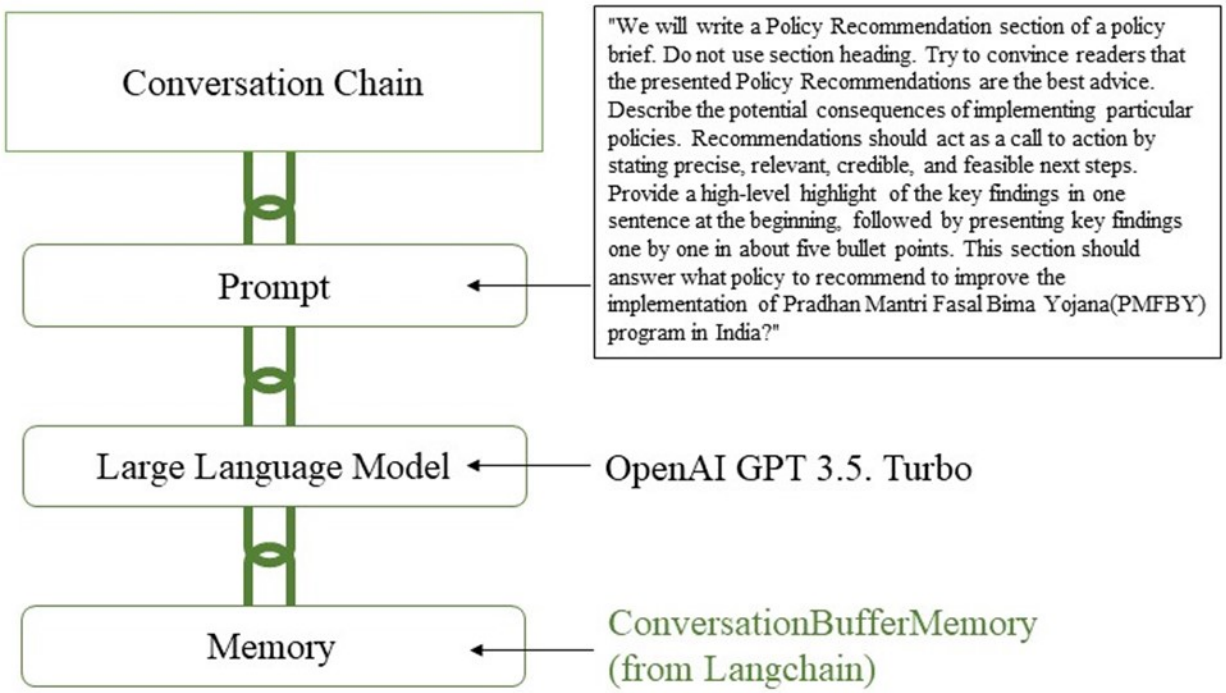


Figure 1 Schematic of the basic chatbot implemented using the LangChain with the OpenAI GPT-3.5-Turbo model. No additional data was provided to the LLM. (Source: Authors)

tions that fetch information from a specific section of the paper. We constructed the two chat models deploying the Python LangChain package (https://langchain.com), a tool to ease the forming of LLM applications through a preexisting template called *chains*. In the LangChain framework, a chain works as a hub that connects different building blocks of the model to create a single application. In the RAG and the Brief Generator approaches, we fed in an IFPRI discussion paper "When Implementation Goes Wrong: Lesson from Crop Insurance in India" by Rajalakshmi Nirmal and Suresh Chandra Babu (2021)[1].

2.1. Basic Chatbot

We develop the Basic Chatbot (BC) using LangChain's Conversation Chain, which allows the building blocks of a chatbot to work together as an application. It consists of an LLM, a prompt, and a memory to store the context of the conversation. We use the GPT-3.5-Turbo by OpenAI as the base LLM and the Conversation Chain to build a simple chat model with the most basic features.

2.2. Retrieval Augmented Generation (RAG)

We used LangChain's Conversational Retrieval Chain, which is a specific chain designed to build a chatbot that responds based on the given information. It connects to a retriever, such as vector storage, to retrieve texts relevant to the human's question. This retrieved information gets incorporated with the query as an input to the LLM. We use the GPT-3.5-Turbo by OpenAI as the base LLM. Since this functionality aligns with the RAG method, we implemented the Conversational Retrieval Chain to build our experimental RAG chatbot. For the retriever, which indicates where to search for the information relevant to the query, we used the Facebook AI Similarity Search (FAISS) service from Meta. FAISS allows quick searching for information encoded in vectors in a local environment. To transform the text into vectors that the FAISS can search, we first split the document with the maximum length of a chunk at 1,000 characters and the overlapping chunk at 30 characters[2]. Then, we used the OpenAI text embedding model

[1] Full text available at https://doi.org/10.2499/p15738coll2.134336
[2] We are aware of OpenAI's report that this model produces biased results against text related to specific identities. Furthermore, the training set of this text modeling has not been updated since August 2020. In this exercise, we kept this text embedding model because "text-embedding-ada-002" was still the model with the highest performance in text searching that OpenAI provided as of writing. However, we recognize the need for further investigation of the effect of text embedding models in the policy-related text similarity search. See https://platform.openai.com/docs/guides/embeddings for more information.

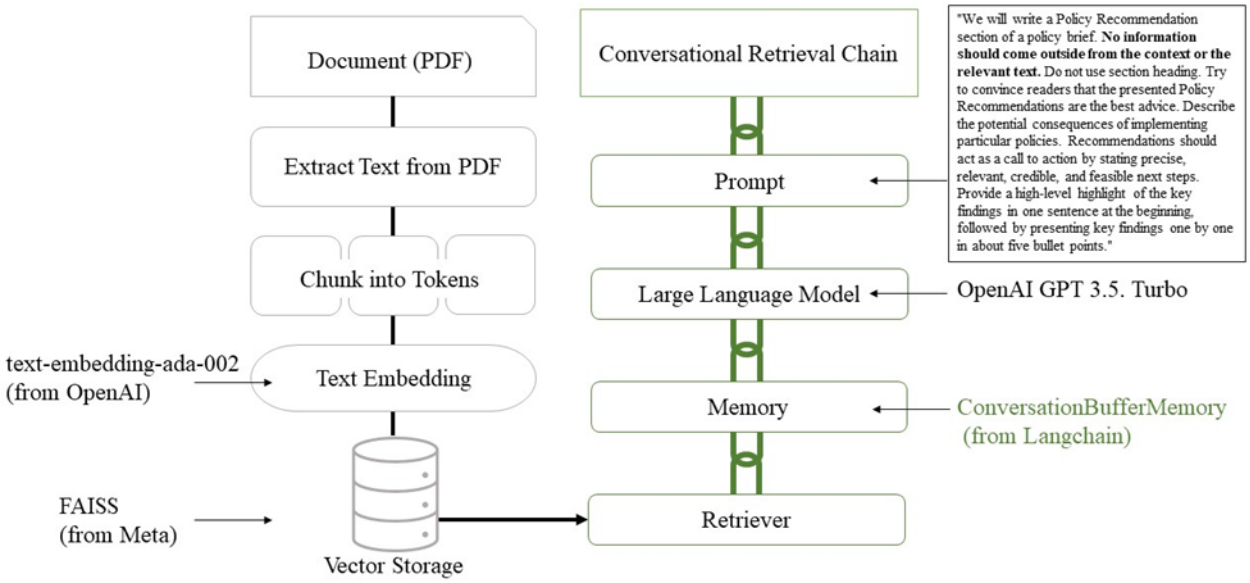


Figure 2 Schematic of the Retrieval Augmented Generation (RAG) model implemented using the LangChain and the OpenAI GPT3.5-Turbo model. Full text data was made available to the Retriever via the text embedding method and vector storage. (Source: Authors)

called text-embedding-ada-002 to convert the chunks into a vector. For the memory component, we used the ConversationBufferMemory package to save the record of the user' queries and the LLM responses.

2.3. Brief Generator (BG)

The Brief Generator is a custom application that we developed to allow users to upload an academic paper and generate an IFPRI-style policy brief. Abiding by the IFPRI guideline, a policy brief summarizes an academic paper structured as an introduction, research overview, key findings, policy recommendation, and conclusion. Since OpenAI limits the number of tokens to generate a text response, the brief generator drafts each section of the policy brief in a separate query and combines it all together. This token limit also prevents users from including the full text of the paper when querying the LLM. Therefore, we tested three different options for the Brief Generator to analyze the paper, summarize, and draft the brief.

a. BG1: The first option is to let the LLM determine the paper sections most relevant to each part of the policy brief. It asks the LLM to classify the paper sections and choose which ones it should use to draft each section of the policy brief. Then, the combined text of these selected sections is included in the query for the LLM to generate the policy brief[3].

b. BG2: The second choice is to let the LLM decide on one section of the paper that best summarizes the discussion paper for the purpose of writing a policy brief. The application creates a query asking the LLM to choose the appropriate section to represent the paper and extract the text of the section to include it in every brief generating query. In this trial, the LLM chose the "Policy Implications" under the "Implementation of the Findings".

c. BG3I and BG3C: The third option is for the user to explicitly indicate which section(s) in the academic paper to refer to when generating the policy brief. When a user chooses this option, the algorithm will skip the process of asking the LLM to select the best-summarizing section. It will directly include the text in the section in the query to generate each section of the policy brief. In this experiment, we deliberately identify "Introduction" (BG3I) and "Conclusion" (BG3C) as a reference and produce the result.

To provide these options, the Brief Generator first recognizes the texts in the academic paper into sections. When a user up-

[3] This method may not work for long papers in which subsections are still too long to be used in the query. This problem did not occur when implementing the discussion paper by Nirmal and Babu used in this study.

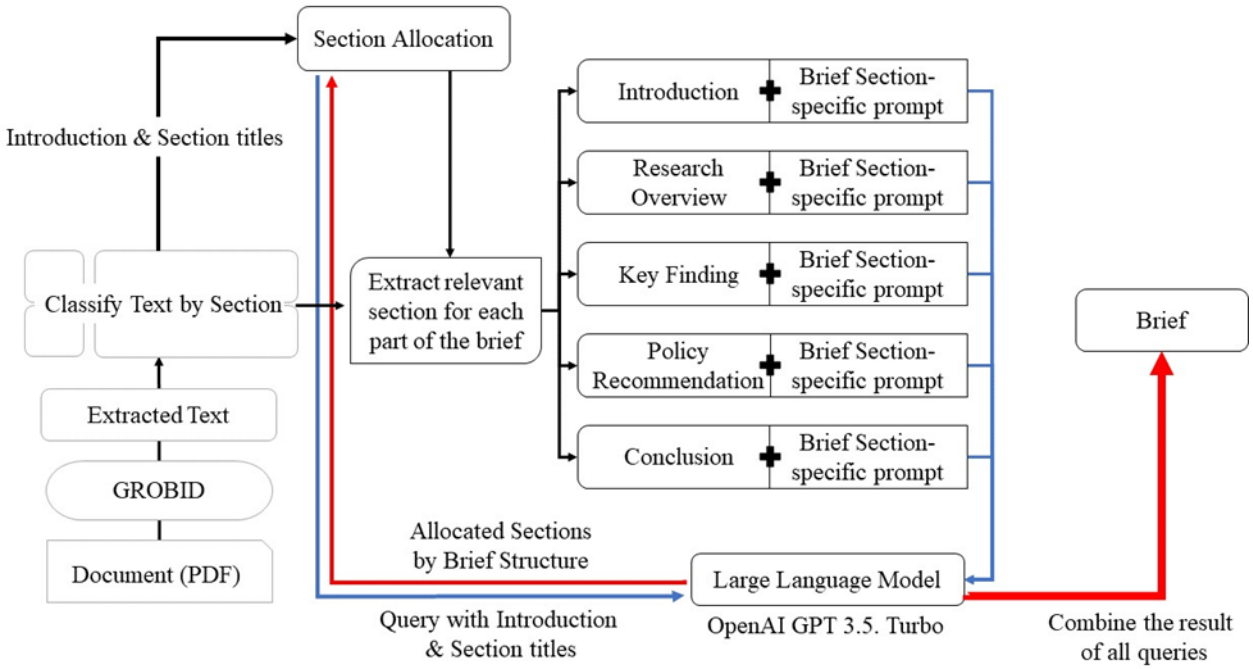


Figure 3 Schematic of the first option of the Brief Generator (BG1), where the full text was provided, section-by-section, for LLM to draft the policy brief. (Source: Authors)

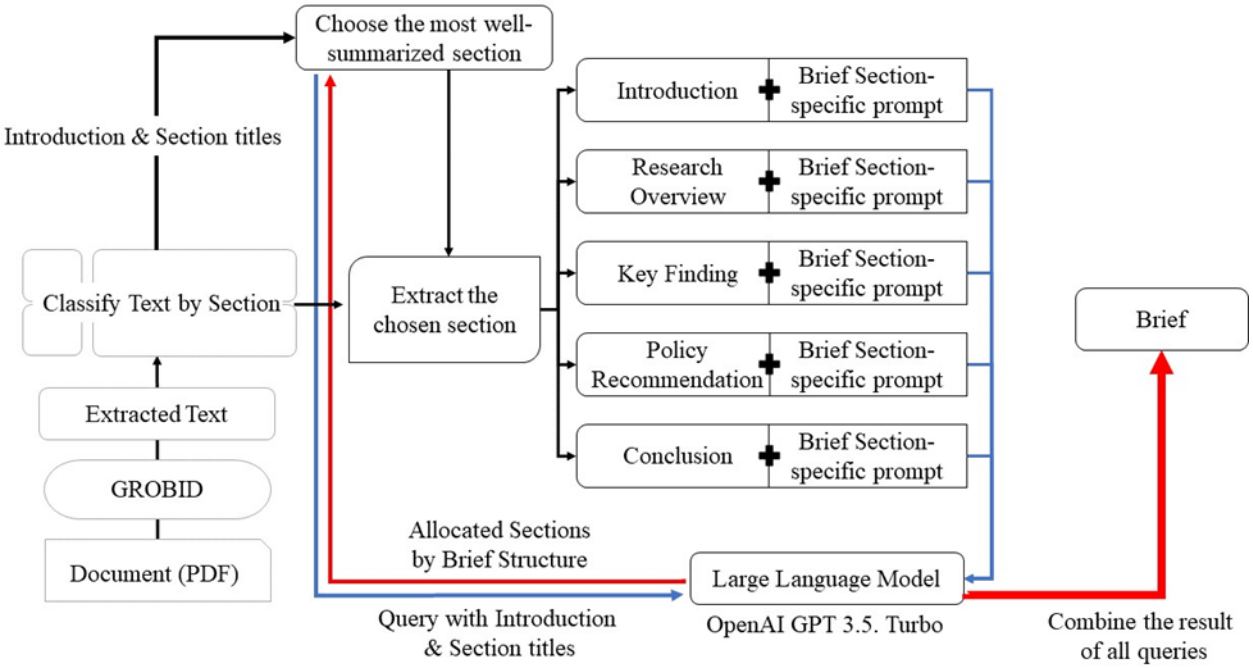


Figure 4 Schematic of the second option of the Brief Generator (BG2), where the LLM was instructed to choose the one most relevant section of the paper and use it to generate the brief. (Source: Authors)

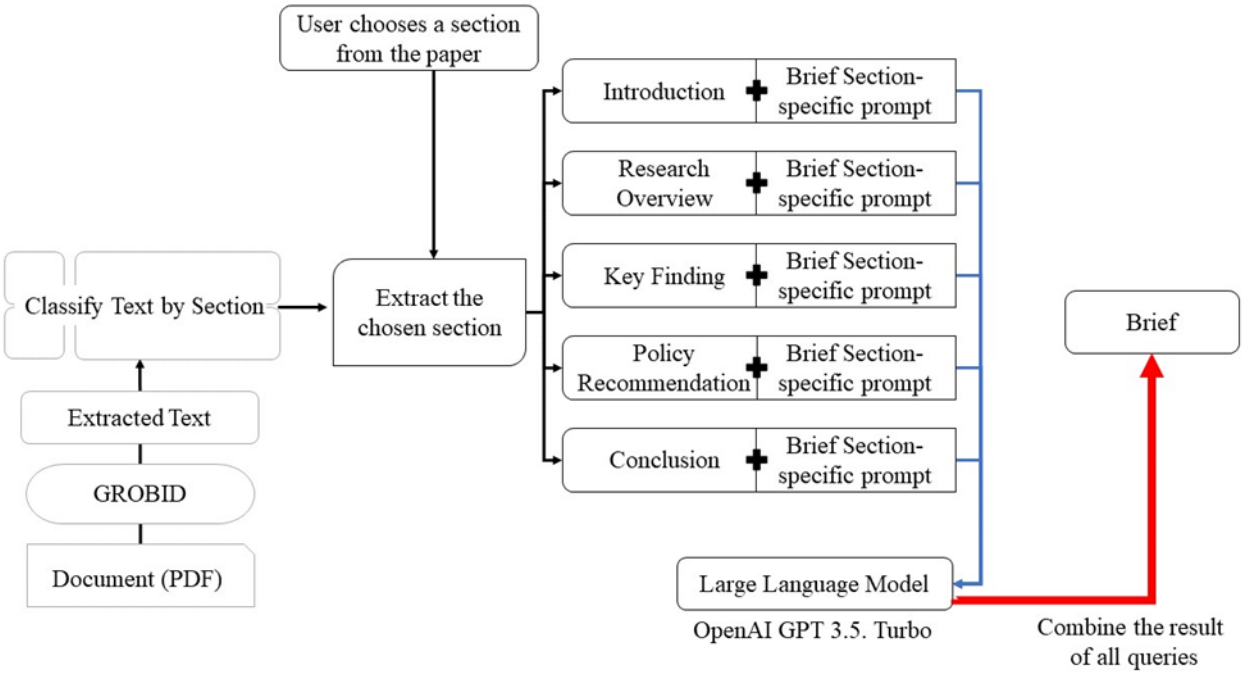


Figure 5 Schematic of the third option of the Brief Generator (BG3), where users select which section(s) to be used in the generation of the brief by LLM. (Source: Authors)

loads the PDF document of an academic paper, the file is converted into an XML document through a neural network model called the GROBID (GeneRation Of BIbliographic Data; https://grobid.readthedocs.io). GROBID provides functionality to segment text in an academic paper saved in PDF according to the structure of the paper. After receiving the XML file that labels the text based on the paper structure, the application parses the XML and saves the title of the section and the corresponding text into a dictionary format, which allows indexing based on the name of the section.

The Brief Generator writes all the sections of a policy brief, i.e., the introduction, research overview, key findings, policy recommendation, and conclusion, by default. However, we only used the policy recommendation section of the produced policy brief to compare the performance with the basic and RAG chat models.

The query to generate the policy recommendation section was the following for all models:

```
We will write a Policy Recommendation section of
an IFPRI policy brief based on the relevant text
given to you. IFPRI policy briefs address current
food policy issues, provide background, offer
evidence-based analysis of the issues, and

make policy recommendations. No information should
come outside from the context or the relevant text
provided. Do not use section heading. Draw the
link between the Research Findings and the Policy
Recommendations, but do not repeat them. Try to
convince readers that the presented Policy Recom-
mendations are the best advice. Describe the po-
tential consequences of implementing the policies
as a call to action by stating precise, relevant,
credible, and feasible next steps. Provide a high-
level highlight of the key findings in one sen-
tence at the beginning, followed by presenting key
findings one by one in about five bullet points.
```

The Brief Generator application also uses a context for every query to the LLM in the model. This context works as background information for the LLM to generate all answers.

```
You are a food policy researcher working at the
International Food Policy Research Institute. Your
target audience is thinkers and donors, and an
informed general public interested in Sustainable
Development Goals and food policy-related issues
in the developing countries. Do not use 'the
study,' 'the research,' or 'in conclusion' so that
```

readers can understand the text without knowing the relevant document.

3. Result

All models numbered five bullet points of distinctive policy recommendations, except the policy recommendation generated by the Brief Generator using the complete text (BG1), which gave seven bullet points. None of the models proposed any policy contradicting Nirmal and Babu's paper. Yet, all policy recommendations choose a unique set of policy details, and each model aggregates a policy recommendation under different points.

For instance, Nirmal and Babu classified the establishment of proper infrastructure under Institutional Implications and the funding technology start-ups under Technological Implications. However, the RAG chatbot put both the investments in infrastructure and technologies under the point "Invest in infrastructure and technology." The Brief Generator considered these policies under "Promot(ing) private sector participation" and "Harness(ing) Technology." While conceivable, this behavior of selecting a recommendation from one section and presenting it under another was unexpected and became a hurdle in comparing the results from different approaches tested.

To overcome this difficulty, we first identified the specific policy recommendations that the results extracted and searched which section these recommendations originally appeared in the discussion paper. Table 1 demonstrates which of the three sections under the "Implications of the Findings" discusses the specific policy recommendation. The three sections were "Institutional Implications," "Technological Implications," and "Policy Implications."

Through this classification, we found that four out of six policy recommendations included a policy not directly mentioned in the paper. Still, the newly invented policies aligned with the perspective of the authors. The Basic Chatbot model (BC) without any text input and the Brief Generator using the introduction (BG3I) and the full text (BG1) talked about enhancing public-private partnerships. This policy does not stand against Nirmal and Babu since they pointed out the lack of the Indian government's effort to foster a private crop insurance market in the introduction of the paper. Yet, the authors did not provide policy recommendations for improving the public-private relationship since they rather focused on remedying the Pradhan Mantri Fasal Bima Yojana (PMFBY), a government-led crop insurance program.

The Brief Generator using the conclusion (BG3C) also proposed "Establishing Grievance Redressal Mechanism" to improve the PMFBY program. In the policy implication section, however, Nirmal and Babu recommended creating a system that "capture(s) the feedback from the farmer" instead, a broader solution than the grievance redressal mechanism. Still, they recognized that the lack of grievance redressal of the PMFBY program contributed to the failure of the implementation in Marathwada, which was one of the case studies in the paper. From these examples, we notice that models generate policy recommendations not directly discussed by the authors but probably based on relevant body of other literature.

Three out of six policy recommendations did not include all three sections of the Implications of the Findings. The Brief Generator using the conclusion (BG3C) did not provide any recommendations from the Institutional Implications section, while the Brief Generator that applied the LLM-selected single section (BG2) does not recommend any policies discussed in the Technological Implications section. Finally, the RAG chatbot did not talk about the topics from the Policy Implementations. These results imply that additional control is necessary to ensure that the summary accurately represents all parts of the paper.

There was no consistency in the order of importance of each policy among the policy recommendations. For instance, a Brief Generator using the introduction (BG3I) put conducting the outcome evaluation as the first policy to recommend. On the other hand, the Brief Generator using one section of the LLM's choice (BG2) placed the same policy at the very last of the policy recommendation. We also could not observe the models placing specific policies with intention. When we averaged the order number for each policy, we found that the mean was 3.18 and the standard variance was only 0.85, while the possible range of the order value was between 1 and 5. The low variance demonstrates that the LLM does not prioritize any policy when producing a policy recommendation in all these various methods. This result is distinct from the traditional practice of writing a policy recommendation. Policymakers will always put the most important recommendations on top of a memo or a brief since it catches the reader's attention most easily. However, the wide variance of order numbers in the same policy and the narrow variance of the average order numbers among the policies both represent that the placement of the policy recommendation is not meaningful when generated by the LLM.

4. Discussion

In this study, we used LLM to generate six versions of policy recommendations using different approaches. While none of the LLM-generated policy recommendations significantly deviate from the authors' arguments in the paper, we find that they

Table 1 Comparing policy recommendations in specific policies

Paper		**LangChain / Chatbot**		**Brief Generator**			
Section Title	*Policy*	*BC (No text)*	*RAG (Full text)*	*BG1 (Full text)*	*BG2 (One LLM-chosen section)*	*BG3I (Introduction)*	*BG3C (Conclusion)*
Institutional Implication	Capacity Strengthening of different institution	"1 Increase awareness and outreach: … educate farmers…"	"1 Strengthen institutional capacity … train state government officials, village leaders…"	"3 Enhance farmer awareness and education … through … farmer training programs"			
	Infrastructure		"2… setting up automatic weather stations… agrometeorological research"			"4. Promote private sector participation: … investing in infrastructure and technology…"	
	Requirement for Insurance Company		"3 … third party to oversee the premium rates .. issue policy documents for PMFBY… expand their personnel capacity to process and settle claims"	"4… third party to oversee premium rate" "4… mandating policy documents for insurance coverage" "4… specific regulations for insurers…"			
	Legal Responsibility on Noncompliance		"4 .. punish companies for any irregularities…" "5… Bank… be penalized for delays in transferring collected premium"	"1 … making non-compliance a punishable offense"	"3 … Non-compliance should be made a punishable offense …"	"3 Strengthen regulatory frameworks… Establish legal and regulatory frameworks to protect farmers"	
	Support from the State Government		"5… state government… ensure timely awareness creation through social media… digitize land record and encourage the use of mobile phones for… CCEs"	"3 Enhance farmer awareness and education…targeted communication campaign…"	"4 State governments… serve as ambassadors for the program"		
Technological Implication	Applying Remote Sensing, Satellite Imagery, Machine Learning, etc.	"3 Enhance risk assessment and coverage: … remote sensing, satellite imagery, weather forecasting models"	"2 …funding.. for innovative start-up in picture-based insurance, remote sensing, and machine learning"	"5 …reduce transaction cost … through technology-drive solutions, such as remote sensing and machine learning…"		"5 Harness technology: … remote sensing and mobile technologies…"	"5 Invest in Research and Development … satellite technology and big data analytics… use of radar satellites for monitoring"
	Using digital platform for CCE			"2 promote the adoption of digital platforms and technology to streamline the CCE process…"			"4 Utilize Technology… mobile applications for live capture of CCEs and an online portal for insurance premium and claims processing"
Policy Implication	Monitoring Deadline adherence	"2. Simplify enrollment" and claims … to reduce administrative burden …"			"1 Develop a robust implementation methodology: The government of India should articulate a clear methodology for the execution of crop insurance program…" "3 Enforce adherence to deadline…"	"2.. Develop robust mechanism for monitoring and evaluating the implementation of crop insurance program…"	"1 …. The government of India should develop a robust implementation methodology… articulating the roles and responsibilities of all stakeholders"
	Independent Audit			"7… establishing independent audit teams…"	"2 Establish an independent audit mechanism:.."		
	Stakeholder Engagement				"4 Foster stakeholder engagement..."		"3.. The government of India should ensure buy-in from state governments and farmers..."

Table 1 (Continued)

Paper		**LangChain / Chatbot**		**Brief Generator**			
Section Title	*Policy*	*BC (No text)*	*RAG (Full text)*	*BG1 (Full text)*	*BG2 (One LLM-chosen section)*	*BG3I (Introduction)*	*BG3C (Conclusion)*
	(Not in paper)	"5 Foster public-private partnership: Encourage … private insurance companies in the implementation of PMFBY. Develop a competitive and transparent bidding process…"		"1 … establishing clear guidelines and deadlines for CCEs and claim settlement" "6 Exploring innovative financing mechanism, such as risk-pooling arrangements and public-private partnership…"		"4. Promote private sector participation: Create an enabling environment for private market institutions to develop innovative solutions and models in crop insurance…"	"2 Establish Grievance Redressal Mechanism "

do not directly correspond to the data fed into the model. Despite the preconception that an introduction and conclusion serve as a summary of the paper, the Brief Generators using these sections did not equally represent the main points of the paper. A Brief Generator using only the "Policy Implication" section, on the other hand, generated a policy recommendation that is coherent with the paper's. These results are unexpected as we designed the Brief Generator directly control the content embedded in the query that interfaces with LLM. This finding may imply that the LLM used in this study may contain memory from the previous run of the application or the prior queries to generate different parts of the policy brief.

The Brief Generator, which directly includes the partial raw text of a paper in the query, also created policy recommendations that amplify a relatively minor point from the paper. We observed that these models invent one to two policies, similar to the occurrence of the Basic Chatbot model with no reference to the text. The only models without this problem were the RAG Chatbot model and the Brief Generator with the section of LLM's choice. In other words, the models that did not involve humans in choosing the relevant text performed better in this analysis. This result signifies that the traditional human intuition for finding relevant information in academic writing may not apply to incentivizing LLM to produce a relevant result. Instead, an empirical study on the LLM performance may be more helpful to reach the same goal.

Finally, this experiment failed to identify any intention in the ordering of policy recommendations in the generated results. All policies were neglected in at least one copy of the policy recommendations, and the average order of most policies stayed around 3rd out of 5 policies. If this randomness persists in further studies with a large sample, we should consider putting additional measures in place to ensure that the LLM produces a policy brief that communicates the importance of a policy most effectively.

5. Conclusion

By generating six versions of policy recommendations through the three models, we learned that accurately representing an academic paper's main points and their emphasis requires additional engineering. The six copies of policy recommendations varied in degree, irrelevant to the range of text given to each paper. From this observation, we infer that the high-quality input information provided to the LLM is not the sole factor in determining the LLM-generated output product.

Moreover, we found out that including text that humans believed to be relevant did not improve the performance of the models. The models that chose the relevant text using a neural network did not generate any new policy recommendations, unlike the models that deliberately indicated a section for reference. Since this conclusion alludes to the possibility that our prior assumption in traditional academia might not applied in the world of LLM, the developers should further investigate the performance of the targeted LLM based on the involvement of human decision.

We recognized that we only measured the performance of the models using a single academic paper. We identified unexpected points that require further investigations with other types of data, we do not exclude the possibility that our outcome is purely random and specific to the environment we established in this study.

This publication has been prepared as an output of **CGIAR Research Initiative on Digital Innovation**, which researches pathways to accelerate the transformation towards sustainable and inclusive agrifood systems by generating research-based evidence and innovative digital solutions. This publication has not been independently peer-reviewed. Any opinions expressed here belong to the author(s) and are not necessarily representative of or endorsed by CGIAR. In line with principles defined in CGIAR's Open and FAIR Data Assets Policy, this publication is available under a CC BY 4.0 license. © The copyright of this publication is held by IFPRI, in which the Initiative lead resides. We thank all funders who supported this research through their contributions to CGIAR Trust Fund.