# Retrieval Augmented Generation of Subjective Explanations for Socioeconomic Scenarios

**Razvan-Gabriel Dumitru**[†*] **Maria Alexeeva**[†*] **Keith Alcock**[†] **Nargiza Ludgate**[‡]
**Cheonkam Jeong**[†] **Zara Fatima Abdurahaman**[◁] **Prateek Puri**[◁]
**Brian Kirchhoff**[◇] **Santadarshan Sadhu**[◇] **Mihai Surdeanu**[†]
[†] University of Arizona, Tucson, AZ, USA [‡] University of Florida, Gainesville, FL, USA
[◁] RAND Corporation, Santa Monica, CA, USA
[◇] NORC at the University of Chicago, Chicago, IL, USA
{rdumitru, alexeeva, msurdeanu}@arizona.edu

## Abstract

We introduce a novel retrieval augmented generation approach that explicitly models causality and subjectivity. We use it to generate explanations for socioeconomic scenarios that capture beliefs of local populations. Through intrinsic and extrinsic evaluation, we show that our explanations, contextualized using causal and subjective information retrieved from local news sources, are rated higher than those produced by other large language models both in terms of mimicking the real population and the explanations quality. We also provide a discussion of the role subjectivity plays in evaluation of this natural language generation task.

## 1 Introduction

Retrieval augmented generation (RAG) has emerged as a powerful technique to mitigate the limited and static knowledge horizon of large language models (LLMs) (Lewis et al., 2020; Guu et al., 2020). However, RAG methods struggle with tasks that cannot easily be captured through search (Yan et al., 2024; Asai et al., 2024). For example, DARPA's Habitus program[1], which aims to ingest subjective information from local populations into scientific models, recently organized an evaluation, descriptively called *Predict what the Locals would Predict* (PWLWP), in which natural language generation (NLG) systems had to predict the responses of a local population to several hypothetical socioeconomic scenarios. In particular, the population of interest consisted of adults from the Ashanti region of Ghana; all scenarios focused on mining in the region. An example of the such a scenario and an explanation by our approach is shown in Table 1.

To properly address this task, this work is the first to propose a RAG approach that explicitly models causality (so we can generate causal explanations) and subjectivity (so we can capture the beliefs of a local population).[2] The key contributions of our work are:

**(1)** A RAG method that incorporates causality and subjectivity. In particular, our RAG method contains two additional components that detect causality and subjectivity in the retrieved sentences; we add only such meaningful sentences to the context of the downstream LLM.

**(2)** A first PWLWP evaluation in which we show that: (a) a RAG-extended LLM provides more accurate responses than a "vanilla" one that operates without context; and (b) the LLM that has access to sentences that contain subjective descriptions of causal phenomena scores higher (both in terms of mimicking the real population, and in the quality of explanations) than other LLMs.

**(3)** An analysis and discussion of explanation quality. Unsurprisingly, the evaluation of subjectivity is subjective itself. We discuss the different perspectives emphasized by different annotators and why annotation differences should be encouraged in this task (Plank, 2022).

## 2 Approach

In this section, we provide a brief overview of the system, with the pipeline described in detail in Section 3 and the experiment in Section 4.

### 2.1 The task

In this paper, we generate detailed explanations for a set of socioeconomic scenarios related to mining in Ghana. The task was administered by the RAND Corporation think tank as a benchmark evaluation for the Predicting What Locals Would Predict (PWLWP) portion of the DARPA Habitus

[*] equal contribution
[1] https://www.darpa.mil/program/habitus
[2] Resources for this work are available at https://github.com/clulab/habitus/wiki/NLP-CSS-Workshop-2024

| | |
|---|---|
| scenario | Imagine that China has recently completed a mining infrastructure project in Brong Ahafo, Ghana. Chinese galamseyers in the region have received an offer to work in the new mine. Additionally, imagine that the Chinese government has promoted the new large-scale mining effort as a source of employment for locals engaged in small-scale illegal mining. Suppose that within a month, the Ghanaian government has seen a decrease in illegal mining activities in Brong Ahafo. This is because... |
| generation | ...The promise of formal employment in a large-scale mining operation can be seen as a powerful incentive for Ghanaians to cease their participation in illegal mining, especially since "74% of Ghanaians say no citizen should be permitted to engage in illegal small scale mining or "galamsey" for any reason". Offering an alternative, legal prospect for income can prod these individuals to abandon illegal mining.... |

Table 1: Excerpt from an explanation generated for one of the test scenarios using sentences that contain locals' beliefs as context passed to the model in the prompt.

program. As its name indicates, this task aims to emulate the thinking of local populations without interviewing people on the ground. Each scenario consists of a description of a situation and its result, followed by a multiple choice question. A sample scenario is in Table 2 of the Appendix.

### 2.2 Generating predictions and explanations

Our RAG method outputs the probabilities of the multiple choice responses for each scenario and produces detailed explanations for each question by using thematically-related retrieved sentences. We prompt the model to make use of and cite contextual data to provide support for the explanations. For contextual data, we use sentences extracted from online news articles pertaining to the subject of interest, mining, and served by Ghanaian media outlets. Context data containing several types of information, notably beliefs of the local communities and causal relations, is intended to provide the model with the location-specific knowledge that it may not have access to. Additionally, by providing causal and subjective information, we hope to improve the model's explanatory power.

### 2.3 Evaluation

We perform two types of evaluation: (a) we calculate the accuracy of our approach in answering the multiple choice questions accompanying each scenario (a sample question and answer choices are shown in Table 2), and (b) evaluate the quality of the explanations that it generates in response to each scenario. The gold data for the accuracy calculation comes from a survey conducted among the local population (1,782 households) at the target location (the Ashanti region of Ghana) by the research organization NORC at the University of Chicago. The survey methodology is discussed in Appendix B.

For the explanations evaluation, we ask two types of evaluators—domain experts and linguists—to evaluate the generated explanations by providing a score and the rationale for the score. Based on these, we compare the explanations generated using different context types and devise a set of evaluation criteria to use for this type of NLG task.

## 3 System Overview

In this section, we provide the technical details of the approach used for the task. As mentioned above, our approach is an instance of retrieval augmented generation, expanded to rely on context that is relevant for the task, i.e., context sentences that are likely to describe beliefs held by a local population for a given hypothetical scenario. The overview of the proposed architecture is shown in Figure 1. We discuss the key contributions below.

### 3.1 Data Sources

Retrieval of a sentence corpus began with the observation that remarkably many Ghanaian news articles are published online. A BBC media guide[3] seeded a search for sites with links to the most prominent media outlets. Further investigation found sites absent of paywalls, with favorable terms of service, and, conveniently for our tooling, written in English, an official language of the country. We were able to identify at least one suitable representative each in categories for radio, television, the press, and news agencies.

A list of simple search terms related to the issue of mining was created and used initially to gauge the suitability of a site's article collection, with bigger being better: mining, gold, galamsey (small-scale, illegal gold mining), harvest, livestock, crop, and price. Each site's native search mechanism was employed, rather than the alternative site-specific Google search, to ensure that only news reports were returned. Although some false hits resulted

[3] https://www.bbc.com/news/world-africa-13433793

| Scenario | Answer Choices |
|---|---|
| Imagine the Ghanaian government implements reforms that change the time it takes for local residents to obtain a legal mining license, reducing the time from three years to three months. Suppose that within three months, the number of mining license applications received by the government tripled. This would have been most likely because… | 1. Those involved in illegal mining have begun working for the Chinese mining company<br>2. Those involved in illegal mining have sought other opportunities outside of the region because of the Chinese mine opening<br>3. Those involved in illegal mining have sought opportunities in small-scale agriculture because of the Chinese mine opening<br>4. Those involved in illegal mining have sought other opportunities within the region unrelated to agriculture<br>5. None of the above |

Table 2: Sample evaluation scenario with five answer choices.

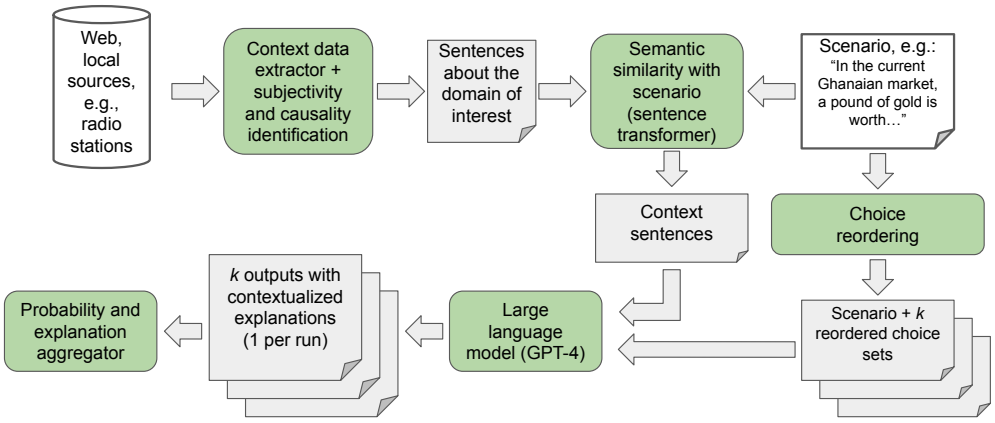


Figure 1: Overview of the natural language generation (NLG) pipeline: software components in green, input data in white, and intermediate data in grey.

from the simplicity of the searches, like "gold" unearthing sports articles, they were not filtered out at this stage because they are later accounted for when context is constructed.

For acceptably prolific sites, the multiple pages of hits generated by the queries were further processed into article lists, and listed articles were then downloaded and finally parsed by the `scala-scraper` library[4] to identify article title, publication date, and byline for tracking provenance, and to assemble the article's text with as much extraneous markup as possible removed. Approximately 70,000 news articles with publication dates ranging from 2013 to 2023 were collected.[5] This resulted in a corpus of over 1.3 million sentences originating from seven sites (for the list of the sites used, see Appendix C).

### 3.2 Classifying Subjective Information

As one type of background information, we extract sentences containing subjective views of local populations. These can be either beliefs, that is subjective views on how the world works, or attitudes, that is how people feel about something. For instance, the following sentence contains a belief (in bold) held by a subset of the population in Ghana:

> *The project manager also pointed out that three-fourths (74%) of Ghanaians say* ***no citizen should be permitted to engage in illegal small-scale mining or "galamsey" for any reason*** *...*

To identify sentences that contain subjective views reportedly held by local populations (subsequently just "beliefs" for brevity), we use the binary classifier described in (Alexeeva et al., 2023). We run the fine-tuned BERT-based model on text of

[4] https://github.com/ruippeixotog/scala-scraper

[5] The list of URLs is available at https://github.com/clulab/habitus/wiki/NLP-CSS-Workshop-2024

all the documents retrieved in the previous step and use the sentences that were classified as containing beliefs with a confidence over 0.97.

### 3.3 Identifying Causality

We identify sentences containing causal relations by using `Eidos`, the machine reading library focusing on extraction of causal statements from text (Sharp et al., 2019). A causal relation is a binary cause-effect relation between two concepts, with one influencing the other in a positive (e.g., promotion) or negative (e.g., inhibition) way. Based on our analysis of 50 sentences identified as containing causal relations, 76% indeed contained a relation of the intended type. Some sentences were judged to be false positives due to lack of specificity of the concepts involved (ex. a), or the causal relations identified indicate hypothetical scenarios or recommendations rather than factual information (ex. b):

a. *This has resulted in a renewed public discussion on illegal mining activities.*
b. *Ghana needs to . . . encourage middle scale and large scale farming that would contribute immensely to total yield of agriculture produce to provide more food, employment and help reduce importation.*

We note that we do not verify whether the extracted relations are indeed causal, but accept them as such since they were extracted based on textual cues consistent with indicating causality.

Similarly to how we extract sentences with subjective views, we run the rule-based extraction system on the full text of documents retrieved for the given scenario and obtain a set of sentences that contain causal relations, e.g.:

> *Government intends to make conscious efforts to integrate the mining industry with the rest of the local economy, thus making it possible for Ghanaian entrepreneurs to increase their participation in the mining industry*

### 3.4 Retrieving Thematically-Related Information

Considering that our sentence corpus is of significant size, we need a way to reduce it to a subset of sentences that: (a) is small enough to fit into the context size available to the LLM (e.g., 16K for GPT-4 at the time of our experiments), and (b) is relevant to the scenario at hand.

In order to make the most of the small context window available we used a sentence transformer, more specifically `allMiniLM_L6_v2` (Tanner, 2023), which extracts sentences from our corpus that are semantically similar to the description of the scenario. To filter the sentences we first query the similarity score between each sentence in the corpus and the scenario to be tested (excluding the possible answer choices) and then we sort them according to that score. We decided to remove the choices from the similarity evaluator because they might introduce biases in the sentences retrieved.

When retrieving information we also make sure that all of the context sentences that we extract match the expected context type. Context type has five possible variations: all sentences, belief sentences, causal sentences, causal belief sentences (i.e., sentences containing subjective statements that include causality), and no context. The last setting is used to test a "vanilla" GPT-4, i.e., an LLM using a prompt without any contextual information.

Finally, we sort the sentences of the given type in descending order of their similarity to our input scenario. To build our context sentences, we pick sentences starting from the most similar until we sum of tokens for our context plus the rest of the prompt reaches the context token limit, making sure to fit in as much information as possible. Further, we prompt the LLM to use these sentences in its explanation and to cite them accordingly.

### 3.5 Retrieval Augmented Generation

In order to force GPT-4 to process the prompt before ranking the choices, we prompt the model in two independent steps. First we ask it to provide justifications for its decisions by citing information from the provided context sentences, and then we ask it to rank the choices. We observe that this approach improves the results, as GPT-4 will use the information it extracted to rank the choices, instead of directly ranking the choices. Next we detail each step and corresponding prompt.

**First prompt:**

Read the following question delimited with backticks:

```
```{scenario}```
```

Use the following context sentences delimited with backticks as background knowledge:

```
```{context}```
```

Provide long and thorough justifications for each of the choices independently, without referring to the other choices, while citing the context using quotes:

```
```{choices}```
```

where $\{scenario\}$ refers to the question for which we want to rank choices. $\{context\}$ refers to the sentences that we use to answer the question. $\{choices\}$ refers to a list of choices from which the model can choose, which we further detail in the next subsection. The main goal of this prompt is to force GPT-4 to use the context and cite it correctly. We also make it justify each choice independently so that it has more information when it ranks all of the choices. For consistency, we kept the prompt unchanged even in the "contextless" scenario, which resulted in the model using the text of the scenario as citable context information.

**Second prompt:**

Rank each choice from most likely to be true to least likely and copy the justification as JSON format with the fields:

```
(id, choice, rank, justification)
```

This prompt ensures that the information that we generated before is now used to correctly rank the choices, as well as enforcing a JSON format.

### 3.6 Choice Reordering and Probability and Explanation Aggregation

In initial experiments, we observed a slight bias in that GPT-4 is more likely to rank the first choices given in its prompt as being better. To alleviate this we decided to roll the choices in all possible variations. We also had the option to permute the choices in all possible variations, while this would be more precise and it would break any bias between the choices' ordering, it would lead to a very high computational demand for each query. For $N$ possible choices, rolling the choices leads to $N$ possible variations that need to be run, while permuting it in all possible ways would lead to $N!$ runs. In general, the set of ordering that we run is the following:

$$\begin{aligned}
\text{Option 1:} &\quad (1 \quad 2 \quad 3 \quad \cdots \quad n) \\
\text{Option 2:} &\quad (2 \quad 3 \quad 4 \quad \cdots \quad 1) \\
\text{Option 3:} &\quad (3 \quad 4 \quad 5 \quad \cdots \quad 2) \\
&\quad \vdots \\
\text{Option } n: &\quad (n \quad 1 \quad 2 \quad \cdots \quad n-1)
\end{aligned}$$

The last problem that we need to solve is how we aggregate all of those runs into a final set of probabilities. To this end, we first invert each rank so that higher values are positive. For example, if GPT-4 ranks a choice as number 1 and we have 5 choices, we convert the rank to $choices - rank$ meaning $5 - 1 = 4$ in our case. Next we sum up all of the ranks obtained for each of the $N$ runs with different orderings into a single vector denoted $final\_rankings$ that has one value per choice. To transform this vector into probabilities, we used softmax with a tuned gamma parameter (to avoid overly peeked distributions):

$$P(\text{choice}_i) = \frac{e^{\gamma \cdot \text{final\_rankings}(\text{choice}_i)}}{\sum_{\text{rank} \in \text{final\_rankings}} e^{\gamma \cdot \text{rank}}}$$

For a walk-through example, see Appendix A.

## 4 Experiment

For the experiment, we produce answers to the multiple choice questions and generate explanations for seven scenarios. For each scenario, we produce generations using five different types of context provided to the model: no context and context in the form of thematically related sentences containing either beliefs, causal relations, causal beliefs, or just information related to the topic. See Table 4 in the Appendix for excerpts of generations produced with each type of context.

For the evaluation of our response rankings in relation to the local population survey gold data (see Appendix B for survey methodology), we compare the probability distribution for all choices produced by our system with the distributions produced by three baselines using mean absolute error (MAE). The baselines are LLM-based with some additional features for information retrieval, chain of thought prompting, and inclusion of population identity context. Two of the baselines use GPT-4 and differ in how they produce distributions to compare against the survey data: the 'GPT-4 TopVote' baseline does that by keeping track of how many times each multiple choice answer was selected as the top choice over a number of samples; and the 'GPT-4 Calibrated' baseline assigns weights to each response based on the answer choice ranking produced for each sample. The third baseline uses an offline LLM (Mistral 7B), simplified to approximate a real world scenario, where resources may be limited and privacy concerns may preclude the use of online LLM APIs. For benchmark implementation and metrics details, see Appendix D.
```

For the explanations evaluation, we assemble the scenarios and the generated explanations into a spreadsheet, with one scenario per row followed by the explanations presented in random order. The generated explanations come in a standard essay format, with an introduction, between two and six body paragraphs, and a conclusion. We ask the annotators to provide a score for each explanation without knowing the context type setting on a scale from 0 to 10 (ten being the highest) and provide a rationale for their score.

The evaluation was done by two annotators: a domain expert and a linguist. Additionally, an annotation supervisor (another linguist) supervised the annotation process and provided meta analysis of the rationale and additional comments on the quality of the outputs.

While some authors point out various issues with NLG evaluation criterion inconsistency (Gehrmann et al., 2023; Howcroft et al., 2020), we chose not to provide any specific evaluation guidelines to the evaluators. With the evaluation being closely connected to a project in a real world setting, it was crucial to see what criteria the domain expert views as relevant for their field in this evaluation, and we did not expect previously defined, non-domain-specific criteria to be necessarily relevant. For the linguist evaluator, we expected the criteria to be similar to those described in previous work, but we chose to not provide detailed guidelines beyond a few example criteria to keep the evaluation procedure consistent between the two annotators.

## 5 Results and Discussion

### 5.1 Evaluation of Probability Distributions for the Multiple Choices

Figure 2 shows the MAE boxplot distribution of two versions of our system and three baselines on the task of approximating the distribution of the answers provided by the local population based on the NORC survey. For this evaluation, we used an LLM with no context and the context of thematically-related sentences. Our best system ("Thematically-related") performs at the same level as the *GPT-4 Calibrated* baseline, while, more importantly for our aim, also providing high quality explanations (see subsequent sections). We find it encouraging that adding context for explanations improves the performance of the overall method, i.e., with the MAE for the context setting lower than the contextless one.

### 5.2 Impact of Background Information Types on Explanations

Figure 3 shows the mean scores for each scenario and context type based on the evaluation by two annotators (a domain expert and a linguist). As seen in the figure, the explanations generated with the use of context, that is instances of retrieval augmented generation, are scored higher by the annotators than those without the context provided. From this, we conclude that retrieval-augmented generation has the potential to help with generating explanations for local events. This could be because the background information that we feed to the model through the prompt provides the knowledge that is not well represented in the model.

Moreover, including subjective views as context ("beliefs" and "causal beliefs" settings in Figure 3) results in higher rated explanations.This could indicate that beliefs, along with causal relations, have a potential to explain functioning of complex systems, as discussed in (Alexeeva et al., 2023).

### 5.3 Evaluation Criteria

We observed that for several datapoints, the scores between the two annotators were quite different (see Figure 6 in the Appendix). By analyzing the rationale provided by the annotators for their scores, we identified two broad categories of evaluation criteria—content quality and text quality,—which could be prioritized differently by the two types of annotators (domain experts vs. linguists), thus resulting in differences in scoring.

#### 5.3.1 Content Quality

Some of the more prominent content-related criteria mentioned by the annotators included the logic of the explanation, the number of factors contributing to the explanation, and task comprehension—that is whether or not the explanation was relevant to the prompt scenario.

For our experiment, the most interesting evaluation criterion, and also the one mentioned most by the annotators, was the use of quotations as evidence. Since we kept the prompt consistent in how the model was instructed to cite context, the settings both with and without background information resulted in generations that included extensive quotations; the only difference was what was being cited: the provided background information or the text of the scenario. This consistency in format of the output between the two setting categories makes this a reasonable comparison.

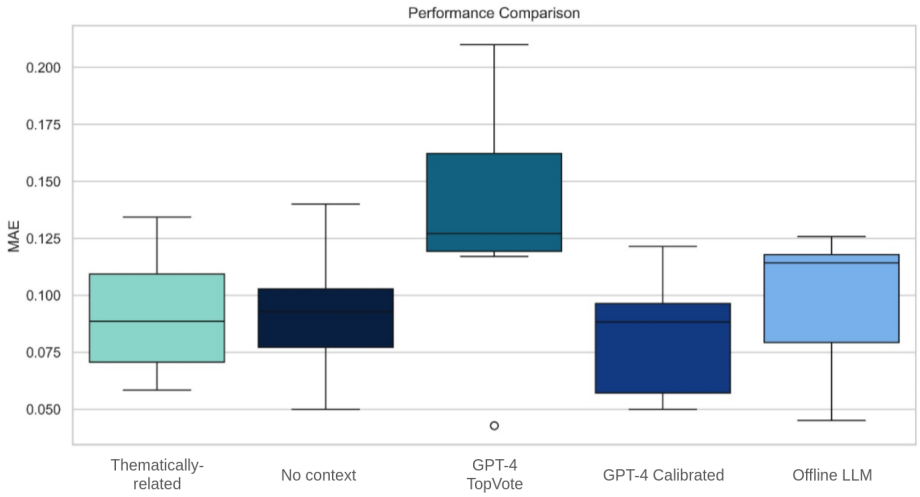


Figure 2: Mean absolute error (MAE) boxplot distribution (lower score is better). The performance of the two of our systems that were evaluated ("Thematically-related" and "No context") is comparable to that of the strongest baseline ("GPT-4 Calibrated") and better than two other baselines ("GPT-4 TopVote" and "Offline LLM").

In terms of quality, the annotators were looking for the quotes to be relevant to the scenario and the topic of the paragraph, the number of quotes used, the quality of connection of the quote to surrounding text, the amount of elaboration on the quote (Ex. 1 in Table 6), as well as the quote being grounded to a source (e.g., attribution to a person).

An additional aspect of quotations use was their accuracy, that is whether or not the quotation came from the provided context sentences and whether it was modified by the model. This criterion was not brought up by the annotators since they did not have access to the full input that was provided to the model. The annotation supervisor performed a spot check of the quotes used in the generated explanations and did not find any instances where the quotations were inaccurate with the exception of one instance of parentheses being left out. Interestingly, there is some inconsistency in the length of quotes the model uses, with the length ranging from single words to full, multi-clause sentences. While single-word quotes are impressionistically more successful, implying synthesis of ideas instead of copying, they are harder to verify as they require careful rereading of the associated context sentences (Ex. 2 in Table 6).

Overall, the use of quotation was viewed as more efficient in the context-based settings than in the contextless one, although individual data points in both categories were viewed as using citations more or less efficiently, from being judged as insufficient to excessive.

### 5.3.2 Text Quality

Prominent factors related to text quality were style (presence of repetition, wordiness, and tone); organization (maintaining a standard essay-style structure of introduction, several body paragraphs, and a conclusion); and presentation of the output, for instance, whether or not the paragraphs describing various factors were numbered.

An interesting criterion for this task was the use of hedging, which one of the annotators used as a proxy for model confidence. The hypothesis was that a model would use less hedging when there is more evidence that it can provide, or, in other words, the less unsubstantiated reasoning (or "hallucinations") the the model needs to output, the fewer hedges it will use. While for most settings, the annotations on the use of hedging were inconsistent (e.g., three causal outputs were judged as having a high level of hedging and three as low), no-context setting outputs were mainly judged as being hedging-heavy and belief context outputs as minimal or moderate in use of hedging.

Another key criterion is the match between various components of the generated text. Based on this criterion, issues can come up on different levels: the generated explanation might not match the scenario in the prompt in terms of content or style; a paragraph, while sensible on its own, may not match the thesis statement of the introduction or may not be logically connected to the preceding paragraph; a quote may not match the topic statement of the paragraph that it is supposed to support.

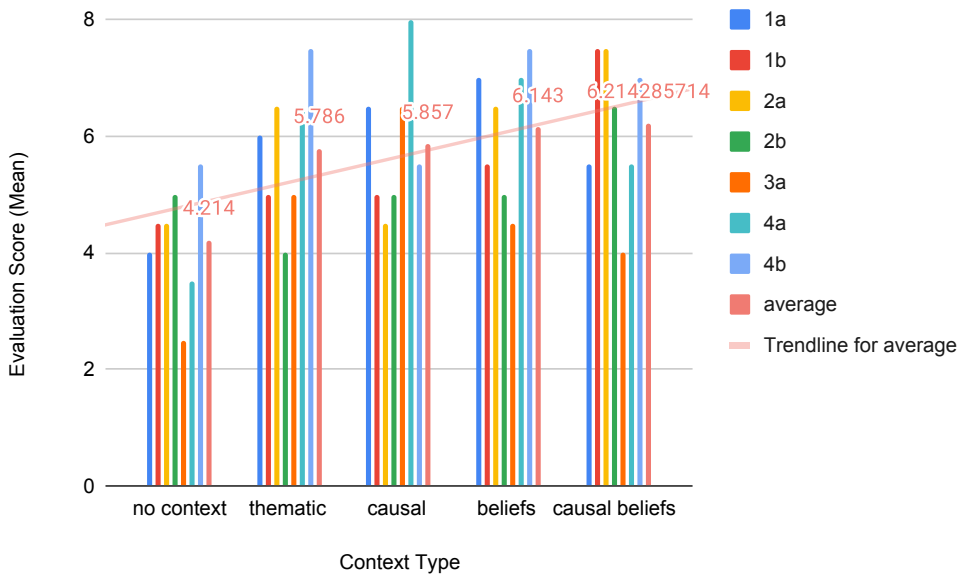


Figure 3: Mean scores for two annotators for each of the seven evaluation scenarios by context type. Five context type settings are compared. Each bar represents a mean score from the two annotators for the given scenario output produced using a given context type. The trendline is for the average of the scores for all scenarios per setting. Context setting generations outperform no context. Causal and subjective view contexts outperform the thematically-related context.

These mismatches tend to be very subtle and make the evaluation task very demanding and potentially requiring additional annotator training.

#### 5.3.3 Annotator Differences

There was a lot of overlap in the criteria between the linguist and the domain expert, with both highlighting the use of quotations, logical flow, and organization; however, the domain expert also focused more on content (the number of factors included by the model as contributing to the explanation and the quality of the evidence provided), while the linguist gave a lot of weight to text quality and used a wide variety of text quality features as contributors to the overall score, thus lowering the weight of content quality. We view this difference in criterion prioritization from the point of view of human label variation framework discussed in (Plank, 2022). Plank views certain types of annotator disagreement as signal, for instance, when the task is subjective and open to interpretation. In our case, not only is the task complex and highly subjective, but also the two annotators come from different fields of expertise. We believe that their disagreement on the score helps us look at the performance of the system from different, complementary points of view.

### 5.4 Practical Constraints

With this project, we had to work within the confines of a real-world social science setting, which comes with some limitations. The first one is limited availability that domain experts face, since they may have to combine their research, teaching, and other responsibilities with on site travel for field work. With this in mind, social science experiments have to be set up to reduce the annotation load as much as possible. In our case, this meant minimizing the reading time required from each annotator. For this purpose, we set up the experiment as a spreadsheet with each scenario presented together with the five outputs instead of providing randomized scenario-output pairs, which would help avoid possible order bias (that is, the evaluator getting the impression that the stimuli are presented in a certain order, e.g., order of improved quality).

The second major limitation is the inherent subjectivity and complexity of the task. The task is cognitively demanding, with multiple competing evaluation criteria, and the length of each output.

The third limitation is the difficulty of setting up evaluation. While intrinsic evaluation that we did is possible, despite the difficulty recruiting annotators for such a cognitively demanding task, a real world evaluation is more complicated because

it may require—as it did in our case—setting up sophisticated baselines and on location field work to make sure the results are relevant for the target population. These may not always be easily accessible to computational social science practitioners, which makes the lack of extrinsic evaluation for NLG a common issue (Celikyilmaz et al., 2020; Gehrmann et al., 2023).

## 6 Related Work

**Retrieval-augmented NLG:** In exploring the integration of local information to mimic people's behavior in query augmentation for language models, our approach is distinct from contemporary methodologies that use retrieval-augmented generation. Among these, the Corrective Retrieval Augmented Generation (CRAG) introduced by (Yan et al., 2024), employs a corrective strategy by integrating a retrieval evaluator and large-scale web searches to assess and refine the quality of retrieved documents. This method uniquely addresses the robustness of generation through corrective actions based on the quality assessment of retrieved documents, targeting the filtration of irrelevant content and enhancement of document relevance through a decompose-recompose algorithm.

In parallel, Self-RAG (Asai et al., 2024) implements a self-reflective framework that encompasses retrieval, generation, and critique. The integration of a critic model to discern the necessity of retrieval and to evaluate the quality of retrieved knowledge places emphasis on selective retrieval and the adaptive generation process. This technique contrasts with ours by focusing on the deliberation over retrieval necessity and the critique of retrieved content's utility, aiming to optimize the generation based on retrieved knowledge relevance.

Further diverging from our approach, Aly and Vlachos (2022) and Aly et al. (2023) concentrate on the application of natural logic for enhancing reasoning in language models, particularly within question-answering contexts. Through the strategic retrieval of documents guided by natural logic, their method aims to improve logical coherence and interpretability of generated responses, marking a focus on logic-based reasoning enhancements in language generation tasks.

Contrary to the previously mentioned methods, our method involves crafting prompts enriched with diverse types of background information, encompassing subjective views of populations, causal relations, and thematically related insights. By collaborating with linguists and domain experts, we've established a comprehensive set of evaluation criteria to assess the quality of these generated explanations. Our findings reveal that enriching the model's input with targeted background information enhances the quality of its output, leading to explanations that are consistently rated higher than those generated without such contextual enrichment. This strategy not only refines the model's ability to produce relevant and insightful explanations but also broadens the application scope of retrieval-augmented language models in understanding and elucidating complex social behaviors.

## 7 Conclusion

In this paper, we make predictions about and generate explanations for hypothetical socioeconomic scenarios by leveraging contextual information retrieved from the web. We show that our retrieval augmented generation approach results in outputs that are both comparable in accuracy to other LLM baselines and high in quality, as evidenced by the evaluation conducted by a domain expert and a linguist. Using subjective and causal information further improves the quality of the explanations. Moreover, by analyzing the evaluations from the two experts, we show how this is another instance of a task where differences in annotations are to be expected and encouraged thanks to its subjectivity.

## 8 Acknowledgments

The authors thank the anonymous reviewers for helpful discussion. This work was supported by the Defense Advanced Research Projects Agency (DARPA) under the Habitus program. Maria Alexeeva and Mihai Surdeanu declare a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies. The annotation work was partially supported through Research and Project (ReaP) Grant from the University of Arizona Graduate and Professional Student Council (GPSC).

## 9 Limitations

There are several limitations to our work. First, we only test our approach using a single language model. With other language models, e.g., Mistral

7B, which was used for one of the baselines, the results could have been different.

Second, while English is an official language of Ghana, it is not the only widely used language in the country. By sticking to one language, we may be missing out on information that could have provided important background knowledge to the model. Moreover, we do not account for possible local variations in the use of English, except for focusing our data selection using the location-specific word *galamsey*. Used to refer to illegal, small scale mining, it has been used for information retrieval within the project since it is directly related to our use case. However, based on our analysis of the extracted causes (see Section 3.3), our NLP tools are able to extract the intended information from the retrieved data with no obvious issues.

Third, when providing background information to the model, we operate over individual sentences. By not using broader context (e.g., the full paragraph) for each sentence, we may be eliminating longer reasoning chains described in text.

Finally, our system is only marginally better than the baselines on capturing the answer distributions from local surveys. There could be multiple explanations for that. For instance, we only provide the model with a small snippet of extracted background information with each API call because of the limited token window allowable with each prompt. Larger amount of context could have resulted in better performance. People's decisions are also not necessarily only influenced by information directly related to the question, which is what we have since we use similarity for context sentence selection: general views, e.g., people's attitudes to the importance of legality or money in general could impact their opinions on questions about involvement in illegal mining activities. That said, we believe that providing more relevant, higher quality explanations, which our system does based on qualitative evaluation, is the main benefit of using location-specific context to prompt the model.

## 10 Ethical Considerations

### 10.1 Benchmark

The RAND benchmark model assessed how well generic LLMs could anticipate the beliefs and opinions of a local population over a set of scenario-based questions. LLMs were chosen as a comparison point to our approach for two primary reasons. Firstly, they are currently de-facto automated systems for answering complex reasoning questions with minimal resources and therefore represent plausible alternatives individuals might pursue in lieu of access to the performer team model. Secondarily, they are known to have limitations, as will be discussed below, that our methodology may be well-positioned to address.

LLMs are well known to replicate biases within their training sets (Feng et al., 2023; Zack et al., 2024), and may struggle to represent viewpoints of populations not well represented within them (Santurkar et al., 2023). While it is difficult to assess the degree to which Ghanans are represented in the GPT-4 and Mistal models leveraged within the benchmarks, it is safe to assume this population is represented substantially less than populations from English speaking countries. Consequently, the benchmarks scores are meant to highlight the limitations of leveraging LLMs to reflect the viewpoints of remote populations and also to illuminate how systems more rooted in data produced by local populations, such as that developed by the performer team, may address these limitations.

### 10.2 General Remarks

The two main concerns in regards to this work is that we are attempting to mimic responses of local populations in a different country and also that we may not be representing the views of the people of that country in a fair way. To elaborate on the second issue, it is manifested at different stages of our pipeline, e.g., we use the data sources in only one of the many local languages and from one genre (news) and we are limited by the context window size, so a lot of available information about people's views is not passed to the model. All of this contributes to creating a potentially biased view of the population.

We attempt to minimize the bias we introduce by using local data sources and conducting evaluation by using local surveys for quantitative evaluation—even though those can also suffer from missing gender-related data, marginalizing certain groups (e.g., rural vs. urban), or biased in developing questions and translating surveys (Weber et al., 2021),—and domain experts for qualitative evaluation.

Both of these issues can be further ameliorated by involving local populations of the area investigated: both as experts to improve the quality of the tools being designed (e.g., to help identify appropriate data sources, evaluate the quality of the

outputs, etc) and as users of the tools: the tools are intended to be used by target populations to augment their decision making process and not to be used by third parties.

## References


Maria Alexeeva, Caroline Hyland, Keith Alcock, Allegra A. Beal Cohen, Hubert Kanyamahanga, Isaac Kobby Anni, and Mihai Surdeanu. 2023. Annotating and training for population subjective views. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 416–430, Toronto, Canada. Association for Computational Linguistics.

Rami Aly, Marek Strong, and Andreas Vlachos. 2023. Qa-natver: Question answering for natural logic-based fact verification.

Rami Aly and Andreas Vlachos. 2022. Natural logic-guided autoregressive multi-hop document retrieval for fact verification.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair nlp models. *arXiv preprint arXiv:2305.08283*.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. 2023. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *Journal of Artificial Intelligence Research*, 77:103–166.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

David M. Howcroft, Anya Belz, Miruna-Adriana Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 169–182, Dublin, Ireland. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR.

Rebecca Sharp, Adarsh Pyarelal, Benjamin Gyori, Keith Alcock, Egoitz Laparra, Marco A. Valenzuela-Escárcega, Ajay Nagesh, Vikas Yadav, John Bachman, Zheng Tang, Heather Lent, Fan Luo, Mithun Paul, Steven Bethard, Kobus Barnard, Clayton Morrison, and Mihai Surdeanu. 2019. Eidos, INDRA, & delphi: From free text to executable causal models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 42–47, Minneapolis, Minnesota. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Henry Tanner. 2023. all-minilm-l6-v2. `https://github.com/henrytanner52/all-MiniLM-L6-v2`. Accessed: 2024-03-14.

Ann M Weber, Ribhav Gupta, Safa Abdalla, Beniamino Cislaghi, Valerie Meausoone, and Gary L Darmstadt. 2021. Gender-related data missingness, imbalance and bias in global health surveys. *BMJ Global Health*, 6(11).

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Travis Zack, Eric Lehman, Mirac Suzgun, Jorge A Rodriguez, Leo Anthony Celi, Judy Gichoya, Dan Jurafsky, Peter Szolovits, David W Bates, Raja-Elie E Abdulnour, et al. 2024. Assessing the potential of

gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22.

# A Probability Calculation Walk-through Example

For example, if we have three choices that we need to rank, let's presume that we get the following rankings for the three rolls that we run: $[2, 1, 3], [1, 2, 3], [3, 1, 2]$. We proceed to invert the ranks so that a higher rank is better, resulting in the following ranks: $[1, 2, 0], [2, 1, 0], [0, 2, 1]$. The next step is to sum up the three variations into a single final ranking, resulting in the array: $[3, 5, 1]$. Furthermore, using $\gamma = 0.2$ and applying the $e^{\gamma \cdot \text{final\_rankings(choice}_i)}$ described above we obtain the following values: $[4.02, 29.68, 0.54]$. The last step is to divide each value by the sum of values resulting in the final per-choice probabilities of: $[11.7\%, 86.7\%, 11.6\%]$.

# B Survey Methodology

## B.1 Data Collection Procedures

The PWLWP survey was administered by NORC's local survey firm, Ipsos Ghana, between August 16th and September 5th, 2023, via face-to-face (F2F) computer-assisted personal interviews (CAPI). The field team consisted of 17 enumerators, four supervisors, and a quality control officer who oversaw quality control activities throughout data collection. The field team was trained in-person from August 7-11, 2024 by Ipsos' field manager and trainers. NORC provided an independent consultant, who reported directly to NORC's Survey Director, to oversee training, piloting, and field launch of the survey. The survey was administered to 1,782 households in the Ashanti region of Ghana.

The English version of the survey instrument was translated using the reconciliation method (two independent translations reconciled by a third, independent translator) into Twi. Both the English and Twi versions of the survey were provided for enumerators to conduct the survey in. Ipsos Ghana enumerators, local to the region, conducted the interviews and recorded responses using tablets containing the programmed survey script in the SurveyCTO software platform. Survey data were directly uploaded from these tablets through encrypted connections to NORC's SurveyCTO cloud server on a daily basis. Data quality reviews by NORC staff were conducted daily throughout the fieldwork period. NORC analysts shared data quality assessments with Ipsos Ghana field managers on a daily basis to allow for ongoing quality assurance and correction as needed during the fieldwork period.

## B.2 Subject Population

**Location:** The survey was administered to randomly selected households in the Ashanti region of Ghana.

**Respondents:** Survey respondents included local resident adults, 18 years of age or older, who were the most knowledgeable about the household's activities in farming, animal husbandry, or mining and whose households engage in such activities on land they own, rent/lease, or borrow. The target sample size was 1,700 interviews.[6]

**Inclusion and Exclusion Criteria:** Anyone under 18 years of age, individuals not living in the selected household, and individuals who were not knowledgeable about the household's activities in farming, animal husbandry, or mining and whose households engage in such activities on land they own, rent/lease, or borrow were not eligible to participate.

## B.3 Sampling Procedures

**Sampling design:** The survey used probability proportional to size (PPS) sampling, giving larger EAs in the Ashanti region a higher probability of being randomly selected for the sample. Sampling was done at the enumeration area level and not the population level, consistent with the Ghana Statistical Service sampling approach. Ipsos requested a sample frame from the Ghana Statistical Service and selected enumeration areas (EAs) using the 2021 Population and Housing Census. One hundred and seventy (170) enumeration areas (EAs) and thirty (30) replacement EA's were selected for the PWLWP survey. Ipsos obtained geo-location maps of the EAs, which provided guidance to the enumerators in locating the designated EAs and working within their defined boundaries. Prominent landmarks such as mosques, schools, markets, cattle dips, road intersections, and factories were used as reference points to mark the single starting point of the random route walk within each EA.

[6] IPSOS Ghana exceeded the target sample size and conducted 1,782 interviews in total.

Table 3 shows the breakdown of the EAs selected for the PLWPW survey.

**Household and Respondent Selection:** A random walk methodology was employed within each EA that ensured a random selection of households throughout the EA, to limit clustering effects. A random number (between 1 and 10) was used to start each walk path, which represents the number of households from the starting point to select first household on the path. After the initial household selection, a sample interval of 10 was used for urban areas and a sample interval of 5 was used for rural areas. Enumerators counted all houses on their left and turned around once they reached the boundaries of the EA or a dead end on their path. Once households were contacted to participate in the survey, the household was screened for eligibility. First, the respondent was asked if the household engages "farming, forestry, foraging, mining, or animal husbandry either on the household's land or someone else's." If the household does, the enumerator asked to speak with "the member of the household most knowledgeable about those farming, forestry, foraging, mining, or animal husbandry activities." Respondents were provided with a kit containing masks and soap as the incentive to participate in the survey. Each EA had a quota of 10 respondents.

**Informed Consent:** Prior to administering the survey, enumerators read an informed consent script to each respondent and verbal consent from the respondent was required before the survey was administered. The informed consent script included a "Right to Refuse or Withdraw" section, informing respondents that they may refuse to take part in the study at any time.

**Benefits and Risks:** Overall risk to respondents for participating in the survey was minimal (i.e., not greater than risks encountered in everyday life). The only known risk to the respondent was loss of time due to participating in the survey. The survey took approximately 30 minutes to administer. Respondents received no direct benefit from participation in the survey, though they were informed during the informed consent that their participation would help NORC and DARPA learn more about the perspectives of residents and activities they engage in on the lands they own, rent, or borrow in the Ashanti region in Ghana.

# C Data Sources

The corpus for the experiment was compiled from the articles retrived from the following websites:

- https://3news.com
- https://www.adomonline.com
- https://thechronicle.com.gh
- https://citifmonline.com
- https://www.etvghana.com
- https://www.ghanaweb.com
- https://www.happyghana.com

# D RAND HABITUS Large Language Model Benchmark

## D.1 Benchmark Requirements

The evaluation was created and administered by the RAND Corporation nonprofit think tank. To better contextualize performance team (PT) results for Predicting What Locals Would Predict (PWLWP) portion of the HABITUS program, RAND (TE) focused on developing a baseline model that could also be scored against NORC (CE) survey results. The main criteria we kept in mind for this baseline were:

1. Performance: We want to create a benchmark that performs similarly on PWLWP to how an uninformed operator, without access to sophisticated data sources, might

2. Simplicity/Realism: In addition to performing well, this baseline must also be relatively simple as operators generally do not currently always have access to sophisticated analysis tools or highly-curated information sources

When combined, a benchmark that satisfies both above criteria will help contextualize the value that each PT's machinery may add to current operational challenges.

Each PWLWP scenario consisted of a multiple-choice question. Both PT and TE teams predicted the selection fraction of each response within a multiple-choice question, and these values were compared to those extracted from a CE survey of the question on a local population.

In Phase 1 of the HABITUS program, we selected a random baseline as a benchmark. For each multiple-choice question, selection fractions for each response were generated by sampling from a Dirichlet distribution, and the resulting values were compared against the CE survey results. The

| Region | Total EAs | Rural EAs | Urban EAs | *Replacement EAs* | |
|---|---|---|---|---|---|
| | | | | *Rural EAs* | *Urban EAs* |
| Ashanti | 170 | 69 | 101 | *16* | *14* |
| Total | 170 | 69 | 101 | *16* | *14* |

Table 3: Enumeration Area (EA) Selection Overview for Ghana PWLWP Survey

| | Context Type | Generation Excerpt |
|---|---|---|
| 1. | no context | . . . The first part of the context notes, "China has recently completed a mining infrastructure project in Brong Ahafo, Ghana". This construction of a new mining infrastructure can bring about substantial changes in the local socio-economic structure. Inherently, such projects tend to be viewed as symbols of growth and economic opportunities, thereby attracting workers who may otherwise rely on less formal, and sometimes illegal, means of earning their livelihood. . . . |
| 2. | thematically-relevant | . . . In addition, the Chinese government promoting the new large-scale mining effort as a source of local employment as per the context quote, "the Chinese government has promoted the new large-scale mining effort as a source of employment for locals engaged in small-scale illegal mining," could have incited locals previously involved in illegal mining to transition to the new employment opportunities. This, in turn, would account for a substantial decrease in illegal mining.. . . |
| 3. | causal | . . . Another significant factor is the Ghanaian government's efforts to combat illicit mining in order to mitigate "its impact on the environment" which is "dire". With the new mining project in place and its conducive employment offers, Ghanaian miners engaged in illegal practices would now have a legal and safer alternative to contribute to the mining sector while also making a living. . . . |
| 4. | belief | . . . The promise of formal employment in a large-scale mining operation can be seen as a powerful incentive for Ghanaians to cease their participation in illegal mining, especially since "74% of Ghanaians say no citizen should be permitted to engage in illegal small scale mining or "galamsey" for any reason". Offering an alternative, legal prospect for income can prod these individuals to abandon illegal mining.. . . |
| 5. | causal belief | . . . Secondly, the Chinese galamseyers in Brong Ahafo might have willingly taken the offer to work in the new mine because of the prevalent resentment towards the Chinese for their involvement in illegal mining activities as observed in this statement: "There is already a sense of resentment for Chinese because of their involvement in illegal mining, especially in the Western Region." By transitioning to legal mining operations, they might have sought to improve their image within the local Ghanaian community. . . . |

Table 4: Excerpts from generated explanations for the scenario: "Imagine that China has recently completed a mining infrastructure project in Brong Ahafo, Ghana. Chinese galamseyers in the region have received an offer to work in the new mine. Additionally, imagine that the Chinese government has promoted the new large-scale mining effort as a source of employment for locals engaged in small-scale illegal mining. Suppose that within a month, the Ghanaian government has seen a decrease in illegal mining activities in Brong Ahafo. This is because. . . "

PT's all showed significant improvement over this benchmark. However, while satisfying (2), the random benchmark was notably lacking in (1)—in most scenarios, an uninformed operator would be more accurate than a random guess at answering PWLWP like questions.

### D.2 Language Model Benchmark

In some ways the ideal benchmark would be one that emulates an intelligent operator who responds to a PWLWP question by collecting relevant, easily-available public information and drawing a conclusion.

To simulate this process, we developed an LLM module than interfaces with Google Search to collect relevant information and produce a PWLWP response. In a slight modification to above idealized scenario, we constructed our LLM to simulate a survey fielded to residents of the PWLWP area of interest, inspired by recent work in the field (Argyle et al., 2023).

The LLM is told to consider the identity of a resident of the region and then asked to express their beliefs on the PWLWP question, and this process is repeated to mimic the results of a random survey of the population. The LLM module is fairly bare-bones and leverages off the shelf programming open-source Python libraries with minimal custom coding. Pseudocode is provided in Table 5.

Much like in human surveys, we find randomizing the ordering of the response options for each LLM call prevents biases towards picking the first or last option. Descriptions of the main components of the LLM module are provided in Figure 4.

**Language Model:** We conducted our benchmark evaluations using the GPT-4 library. We interface with the model through the popular open-source library LangChain[7].

**Google-Search:** We equip the LLM with the ability to conduct Google Searches to retrieve information needed to better contextualize the question that lies outside of the GPT-4 training set. The LLM is providing access to Google through the Agents module within LangChain. This module prompts LLM to devise a strategy to solve a given problem and then provides the LLM with tools, such as Google Search to execute its strategy. Each time the LLM collect new information from its tools, it updates its understanding of the problem, revises its strategy if needed, and proceeds until a desired result is achieved.

**Response Ranking:** Rather than merely selecting the most likely option amongst a set of multiple-choice responses, we ask the LLM to rank the responses in order of decreasing likelihood. This allows us to better understand LLM decision making and accordingly make prompt adjustments. Also, as will be discussed below, this may play a role in producing distributions of selection fractions across all response options.

**Chain of Thought Prompting:** We leverage the popular technique of chain-of-thought prompting within our LLM calls. This technique instructs LLMs to provide reasoning for decision making, and in certain cases, has been found to improve LLM performance while providing information relevant for debugging/prompt engineering.

**Identity Contexts:** For each PWLWP question, we conduct N$\sim$25 simulated surveys. Each survey instructs GPT to assume the identity of a citizen of a given region with certain demographic characteristics. For example, if a PWLWP question was fielded to residents of Ghana, for each LLM call, we provide the age, gender, and city of residence of a 'synthetic citizen' of Ghana. In this case, these demographic variables are sampled from distributions (1) publicly available online or (2) provided in advance by the CE team.

Admittedly, it may be difficult to extract accurate demographic information for certain areas of interest, and further there may be complex relationships between demographic variables that need to be considered when sampling. Both factors pose challenges to exporting this technique generally. However, this component is not essential to the LLM module and further experiments can investigate how the performance of the benchmark changes, if at all, if identity contexts are excluded.

### D.3 Prompt Design

An approximate form of the prompt we developed is provided in Figure 5.

The Google Search tooling, chain of thought prompting, response rankings, and identity contexts mentioned above are all visible from within the prompt.

[7] https://www.langchain.com/

| LLM Module |
|---|
| For k in N:<br>• Randomly sample an identity context I<br>• Randomize the ordering of the response options R<br>• Process LLM(P, R, I) and store result<br>Aggregate the N results from above into a final selection fraction distribution |

Table 5: Pseudocode for the LLM Module. N is the number of LLM samples, LLM(P, R, I) is the result of the LLM module called on PWLWP question P with given response options R conditioned on identity context I.

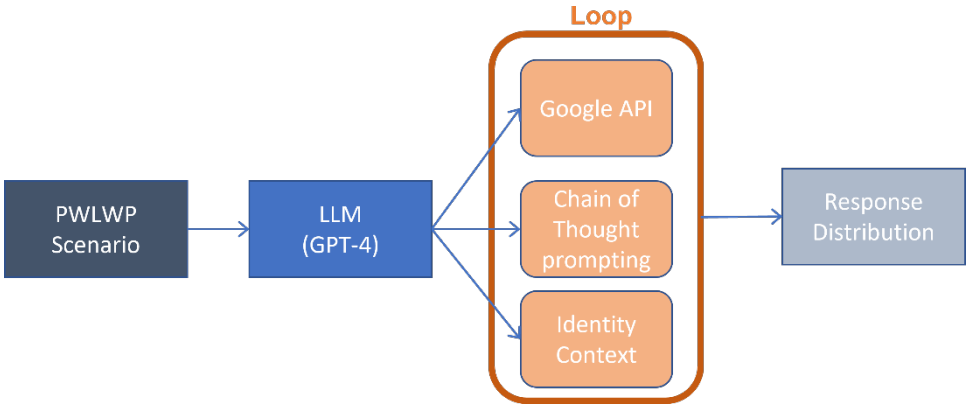


Figure 4: The overview of the benchmark LLM module

### D.4 Distribution Generation

One challenge we encountered was extracting selection fraction distribution from our LLM module. If there are four answers to a given PWLWP question: A, B, C, D – we need to predict the fraction of individuals who will select each response. One could ask the LLM to produce these anticipated fractions as a response, yet LLMs have struggled to produce well-calibrated probabilities in certain scenarios (Srivastava et al., 2022). We present two alternative options here:

**Top Vote Aggregation ('GPT-4 TopVote'):** Across our N LLM samples per PWLWP question, we track how often each multiple-choice option gets selected as the top choice. The final selection fractions are the percentage of time each response gets selected as the 'top vote'. One question that emerges from this technique is "how many votes is enough?". While we can monitor the convergence of each selection fraction as a function of LLM call, it is still difficult to anticipate how these fractions may change under additional samples. Further, the GPT-4 calls used as a foundation for our model are both relatively slow and expensive (∼1 min per sample, about ∼ $0.50 dollar per sample), meaning dramatically increasing the number of samples may pose challenges.

**Ranking Calibration ('GPT-4 Calibrated'):** For each vote, we may instead assign a weighting to each response based on the probability ranking of that response within the LLM output. In this case, the calibration selection fraction, **SF**, for response ***r*** within a PWLWP question is given as:

$$SF(r) = \frac{\sum_{i}^{N} w(\alpha_{r,i})}{\sum_{r}^{R}\sum_{i}^{N} w(\alpha_{r,i}))}$$

where $\alpha_{r,i}$ is the ranking of response ***r*** on the $i^{th}$ LLM vote, $w$ is the weighting function that assigns weights based on $\alpha_{r,i}$, and R is the entire response set for a given PWLWP question. In this work, we chose

$$w(\alpha) = \mathrm{A}e^{-\alpha \cdot b}$$

where A and b are constant adjusted to our data. In an ideal scenario, other functional forms would be explored and A and b would be fit to a more robust calibration dataset (here a ranking of 0 "top-vote" get assigned the highest weight, with decreasing rates with increasing rankings).

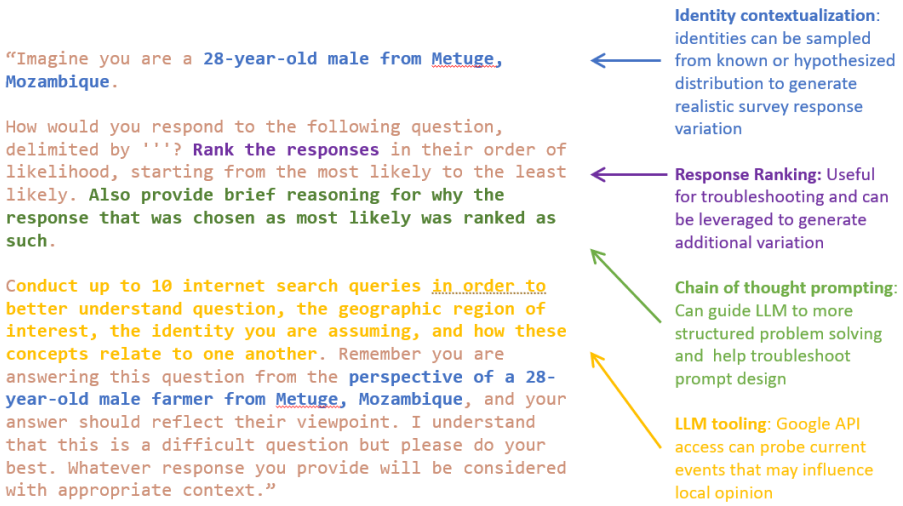


Figure 5: A sample prompt for the LLM benchmark for a sample location.

### D.5 Analysis of PWLWP Results

We deployed our LLM benchmark on HABITUS Phase II PWLWP questions fielded to the region of Ghana and compared our results against both PT and CE teams. There were seven total scenarios in this set, with one multiple choice question each. The ultimate goal is to determine how similar the PT and TE results are to the CE results across these questions. However, similarity can be defined in multiple ways. For this evaluation, we used mean absolute error (MAE) as described below:

$$MAE = \frac{1}{|R|}\sum_{r}^{R} |sf_{r,CE} - sf_{r,K}|$$

where |R| is the number of responses in set R and $sf_{r,K}$ is the selection fraction associated with response r produced by team K (here K = CE, PT, or TE).

### D.6 Offline Model Benchmarks

The previous LLM benchmark utilized closed-source GPT models that are only accessible via an online API. Eventual end users of HABITUS technologies may need to access model results in edge computing environment without access to Google searches/online APIs and may also need to transmit classified information to such technologies that cannot be shared to third party vendors.

To provide a benchmark that satisfies these operational requirements, we built a second LLM pipeline that leverages only open-source models that can be run locally and offline. Our benchmark utilized the 7B parameter Mistral AI language model.

We evaluated PT PWLWP scenarios using essentially the same setup as above with this Mistral model, with two main differences. Firstly, we removed the ability for the model to conduct Google searchers. Secondly, we removed the response ranking component of the model, asking instead that the model simply return the option it deems most probable during every LLM vote. Part of the motivation of the response ranking was to generate variation in the LLM output given that GPT-4 model calls can have high latency and high costs compared to open-source models. With Mistral, we can conduct 100 model votes per scenario quite easily, allowing us to generate response variation more naturally without adding in additional post model calibration. Lastly, the response ranking adds additional complexity to both the LLM prompt and the structure of the LLM output, complexities that may pose challenges for the open-source model. By removing these components, we make our pipeline much simpler and reduce the possibility of error.

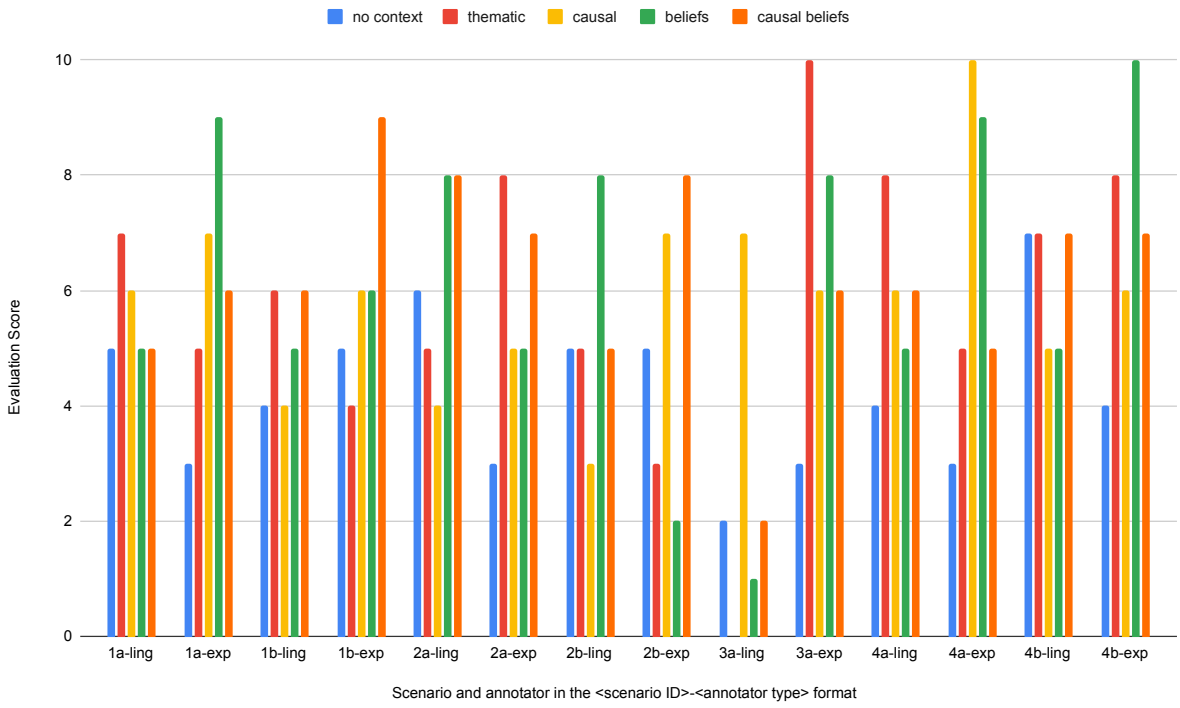


Figure 6: Breakdown of scores by two annotators per scenario. In the scenario and annotator code, "-exp" stands for domain expert and "ling" for linguist. In the corresponding scenarios (e.g., the two scenarios starting with 1a), the corresponding generated explanations are in the same color and order. For instance, in scenario 1a, the linguist gave the no context setting (blue color, first bar) the score 5 and the expert the score 3. We attribute the difference in scores to the two annotators prioritizing different types of criteria.

| | **Context Type** | **Scenario (paraphrase)** | **Generation Excerpt** |
|---|---|---|---|
| 1. | subjective view context | Illegal mining causing land degradation and shift of labor away from agriculture resulted in mass protests and tension. Explain how that impacts illegal mining. | ...Second, illegal mining activities can cause severe environmental damage, leading to the land's degradation and ruining the area's potential for future agricultural activities. According to the context, "it is acknowledged that while illegal mining supports livelihoods of many, it has equally caused severe damages to the environment in many parts of the district." Moreover, ... |
| 2. | causal context | Explain lack of decrease of illegal mining in a given location after a foreign country builds infrastructure for and promotes legal mining. | ...Second, there is widespread corruption and ongoing policy implementation challenges. The context reports "widespread corruption has hamstrung government efforts" in fighting against illegal mining, referring to illegal mining as a "menace" that may deter engagement in legal opportunities. ... |

Table 6: Examples for Section 5.3. Ex. 1: The quote is relevant for the topic sentence of the paragraph, but is only connected with a transition phrase and not elaborated on; additionally, the explanation does not match the scenario. Ex. 2: Use of short snippets of quotes from provided context, which, while successful, is harder to verify.