Mitigating Hallucinations in LLMs for International Trade: Introducing the TradeGov Evaluation Dataset and TradeGuard Hallucination Mitigation Framework for Trade Q&A

Kriti Mahajan
Amazon
kritimhj@amazon.com

Abstract

Given the constant flux in the world of geopolitics, staying up to date and compliant with international trade issues is challenging. But exploring if LLMs can aid this task is a frontier hither to unexplored in the LLM evaluation literature - primarily due to the lack of a dataset set for benchmarking the capabilities of LLMs on questions regarding international trade subjects. To address this gap, we introduce TradeGov - a novel, human audited dataset containing 5k international trade related question-answer pairs across 138 countries, created using ChatGPT based on the Country Commercial Guides on the International Trade Administration website. The dataset achieves 98% relevance and faithfulness and doesn't show any systematic biases along macroeconomic and geographical dimensions, lending itself to equal applicably for LLM assessment across countries. Testing the performance of ChatGPT-4o and Claude Sonnet 3.5 on this dataset - marking the first systematic evaluation of LLMs for answering questions about international trade - we find that ChatGPT-4o achieves 85% accuracy while Claude Sonnet 3.5 achieves 88% accuracy. Building on these insights, we develop TradeGuard - an ensemble trade regulation hallucination mitigation framework that leverages majority vote summarization and multi-agent debate to achieve 91% accuracy on the TradeGov dataset, outperforming vanilla versions of Claude and ChatGPT. TradeGuard's ensemble hallucination detection algorithm — combining entailment verification, cross-questioning, and Bayesian regression—achieves an F1 score of 91%, significantly enhancing reliability in legal contexts. Notably, we demonstrate that TradeGuard reduces "I don't know" responses while maintaining accuracy, particularly for low-income countries and demonstrates no systematic biases along key macroeconomic dimensions.

1 Introduction

In an increasingly globalized world, understanding and complying with international trade matters is crucial for both governments and businesses alike. For governments, it is essential to strike a balance between protecting domestic markets and integrating with the global economy. For businesses, staying abreast with international trade affairs is crucial for a. mitigating and minimizing losses due to fines on business operations and lost business opportunities, while b. also maximizing profits by taking advantage of legal opportunities for cross-border trade. However, navigating the complex legal landscape of international trade requires specialized legal expertise, which is not equitably available to all. Illustratively, larger businesses have the capital to leverage the expertise of lawyers specializing in the trade of a particular country (say India) while a small businesses are unlikely to have similar expertise, thus making them comparatively less competitive in the global economy. Large Language

Models (LLMs) have the potential to bridge this gap by offering reliable information regarding international trade. If LLMs can effectively interpret and provide information about international trade, they could assist both small and large businesses in understanding regulatory requirements and expanding into global markets. LLMs could also aid government entities in navigating complex policy negotiations and red tape associated with international trade regulations. Therefore, it is important to evaluate how well LLMs can handle questions related to international trade. However, the current LLM evaluation literature does not address the capabilities of LLMs for question-answering in this domain. A primary impediment is the lack of a dataset for benchmarking the performance of LLMs on Q&A tasks related to international trade.

We address this gap by introducing a novel dataset on international trade called TradeGov - constructed by leveraging Retrival Augmented Generation (RAG) with ChatGPT 4o to generate international trade question and answer pairs using the Country Commercial Guides on the International Trade Administration website. This paper describes the construction of this dataset and implements a framework for assessing the quality and biases (along global macroeconomic and geographical inequalities) of the Q&A pairs generated. It then carries out a novel LLM benchmarking exercise by evaluating the performance of ChatGPT 4o and Claude (V2, Sonnet and Sonnet 3.5) on TradeGov for answering questions related to international trade. Then, to improve the trade related Q&A capabilities of LLMs, we develop TradeGuard - an ensemble framework specifically designed to mitigate hallucinations in trade regulation responses which can be applied to any LLM of choice. By leveraging majority vote summarization and multi-agent debate, TradeGuard achieves 91% accuracy on the TradeGov dataset, significantly outperforming vanilla versions of ChatGPT and Claude. Moreover, TradeGuard's hallucination detection algorithm—combining entailment verification, cross-questioning, and Bayesian regression—achieves an F1 score of 91% on the TradeGov dataset, enhancing reliability in legal contexts. Notably, while maintaining high accuracy, TradeGuard reduces "I don't know" responses and eliminates the systematic biases along macroeconomic dimensions that we observed in single-shot responses generated by Claude and ChatGPT-4o.

Thus, this paper makes three main contributions: 1) we introduce TradeGov, the first comprehensive dataset for evaluating LLM performance on international trade questions, 2) we present the first systematic evaluation of LLM (ChatGPT and Claude) capability in this domain via Q&A evaluation on TradeGov, and 3) we propose TradeGuard, a novel, LLM agnoistic framework that significantly improves the reliability and fairness of LLM responses to international trade queries.

2 Literature Review

This paper situates itself at the intersection of four fields: applying LLMs to law, international trade law, hallucination mitigation, and creating novel datasets for LLM benchmarking.

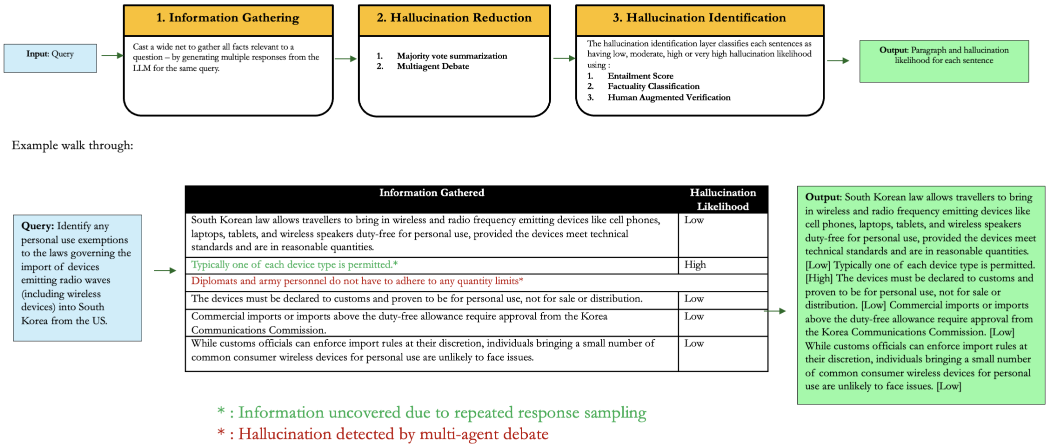
LLMs have been applied to various legal tasks such as summarization [19][10], Q&A [16][2], legal judgment prediction [7], text extraction, and reasoning. Numerous datasets support these tasks, including corpora for argument mining (Demosthenes, CDCP), legal case analysis (CaseHOLD, European Court of Human Rights Dataset), contract review (CUAD, ContractNLI), and regulatory analysis (EUR-Lex-Sum, Caselaw Access Project). While many datasets are multilingual and cover diverse legal domains, there is a notable gap in datasets focused on international trade law, which this paper aims to address.[1][2]. AI in international trade law is an emerging field, with most research focused on AI regulation from a trade perspective or the impact of generative AI on trade policy. However, no studies address the capability of LLMs to handle international trade law queries, which is the focus of this paper.

[17], [14], [9] highlight significant hallucination rates in AI-driven legal research tools, underscoring the need for reliable evaluation methods and human oversight. Our work draws on hallucination mitigation and LLM uncertainty estimation literature, focusing on zero-resource, logit-free techniques due to constraints on model access and data as we limit the architecture to use one closed source, black box models like Claude[3]. Given the open-ended nature of legal Q&A, we prioritize

[1] Many of these datasets are multilingual, reflecting the global nature of legal practice

[2] Additionally, there are datasets focusing on specific legal domains such as patent litigation, tax law, refugee claims, datasets for responsible pre-training [15], legal reasoning

[3] Open source LLMs like Falcon, Flan-t, Adapt-LLM and SAULLM-7B were deemed unsuitable due to high inaccuracy in early experimentation


Input: Query
1. Information Gathering
2. Hallucination Reduction
1. Majority vote summarization
2. Multiagent Debate
3. Hallucination Identification
1. Entailment Score
2. Factuality Classification
3. Human Augmented Verification
Output: Paragraph and hallucination likelihood for each sentence
Example walk through:

Cast a wide net to gather all facts relevant to a question – by generating multiple responses from the LLM for the same query.

The hallucination identification layer classifies each sentences as having low, moderate, high or very high hallucination likelihood using :

Query: Identify any personal use exemptions to the laws governing the import of devices emitting radio waves (including wireless devices) into South Korea from the US.

Information Gathered	Hallucination Likelihood
South Korean law allows travellers to bring in wireless and radio frequency emitting devices like cell phones, laptops, tablets, and wireless speakers duty-free for personal use, provided the devices meet technical standards and are in reasonable quantities.	Low
Typically one of each device type is permitted.*	High
Diplomats and army personnel do not have to adhere to any quantity limits*	
The devices must be declared to customs and proven to be for personal use, not for sale or distribution.	Low
Commercial imports or imports above the duty-free allowance require approval from the Korea Communications Commission.	Low
While customs officials can enforce import rules at their discretion, individuals bringing a small number of common consumer wireless devices for personal use are unlikely to face issues.	Low

Output: South Korean law allows travellers to bring in wireless and radio frequency emitting devices like cell phones, laptops, tablets, and wireless speakers duty-free for personal use, provided the devices meet technical standards and are in reasonable quantities. [Low] Typically one of each device type is permitted. [High] The devices must be declared to customs and proven to be for personal use, not for sale or distribution. [Low] Commercial imports or imports above the duty-free allowance require approval from the Korea Communications Commission. [Low] While customs officials can enforce import rules at their discretion, individuals bringing a small number of common consumer wireless devices for personal use are unlikely to face issues. [Low]

* : Information uncovered due to repeated response sampling

* : Hallucination detected by multi-agent debate

Figure 1: TradeGuard - Model Architecture and Example Walkthrough

self-assessment methods for hallucination reduction and identification. [22] use Semantic-aware Cross-check Consistency for hallucination detection in Black-Box Language Models. [11] reduce hallucinations through Multiagent debate. [23] find that "ChatGPT and GPT-4 can identify 67% and 87% of their own mistakes, respectively". [25] introduce "SelfCheckGPT" - a sampling-based approach that can be used to fact-check the responses of black-box models in a zero-resource fashion. Additionally, (Ovadia et al., 2019; Dusenberry et al., 2020a; Band et al., 2021) find that ensembles and Bayesian neural nets are highly effective for uncertainty and robustness.

This paper address two critical gaps in the literature: TradeGov provides the first comprehensive benchmark for evaluating LLM capability on international trade Q&As, while TradeGuard advances the hallucination mitigation literature by demonstrating the first effective risk reduction strategies for open-ended international trade question answering. Within the existing legal LLM literature, the closest system to our TradeGov framework is Chatlaw [8] - a legal assistant using a Mixture-of-Experts (MoE) model and a multi-agent system to improve the reliability and accuracy of AI-driven legal services.

3 TradeGuard : Model Architecture

The TradeGuard model's architecture (Figure 1; see Appendix Figure 4 for an end to end example of a query traveling through the framework) involves sequential LLM queries as follows:

1. **Fact Generation**: Given that legal information for developing, under-developed and emerging markets is not present in the training data of foundational LLM services densely, it is important that we cast a wide net while generating responses to ensure greater coverage of facts (and avoiding "model unsure" / "doesn't know" responses). To enable this, the model generates one deterministic response (temperature and top p set to 0) and ten less deterministic responses i.e. more "creative" responses (temperature and top p > 0.2).[4].
2. **Hallucination Reduction**: Hallucination is a key concern for LLMs on factual questions, especially because we are generating less deterministic/ more "creative" responses. Thus for hallucination reduction we use: 1) **Majority vote summarization**: the 11 responses generated in (1.) are summarized into one paragraph, keeping only the statements which appear repeatedly in the 11 responses to reduce the inclusion of hallucinated statements and 2) **Multi-Agent Debate** [11]: Two instances of the same LLM family (i.e. agents) generate responses to the same query, and debate if the two responses agree with each other, regenerating the responses till they agree with each other (i.e. the model converges) or till the user specified number of debate iterations are done. This enables the model to align its response away from what it can see as erroneous in the summarized response and

[4] While outside the scope of the current paper, if documents are available for a country, Retrieval-Augmented Generation (RAG) with citations is used

thus reduce hallucinations. If the model doesn't converge after 5 iterations, it outputs a non-convergence string saying that the model couldn't agree on a response to the prompt.

3. **Hallucination Identification**: To minimize legal risks and maximise the number of hallucinations we catch, we employ a conservative approach using three combined zero to low data self-assessment to identify potential hallucinations in responses. [5]. The three methods are:

   3.1 **Entailment Score** (if <0.7, hallucination detected else not) : Inspired by SelfCheckGPT [25] and bi-directional entailment [26], for each sentence generated in the output response, the using single-shot prompting to Claude, the model checks how many generated sampled responses in the fact generate stage contain a sentence which "entails" the same meaning. For a given sentence, if more than 70% of the sampled responses contain a sentence which entails a similar meaning, then the sentence is classified as not being a hallucination because it is interpreted as having low uncertainty. We choose this method over measuring the semantic similarity between sentences because like [26] we found that semantic similarity was unable to capture the true meaning of sentences and thus could not correctly identify when two sentences entailed each other.

   3.2 **Factuality Classification** (if 1 , hallucination detected else not) : Our early experiments for international Q&A answering showed that, in addition to numerical quantities, Claude also hallucinated the names of laws, regulations, certifications, and labels. To identify the same, for each sentence generated in the in the summary paragraph generated by the the hallucination reduction stage, Claude is used to identify if the sentence contains laws, regulations, certifications and labels, and extract the same. If yes, then then Claude is asked to define the identified laws, regulations, certifications and label. If the Claude generated definition doesn't agree with the context in which those laws/regulations/certifications/label occurred in the sentence in the summary paragraph, then that sentence is classified as a hallucination.

   3.3 **Human Augmented Verification** (if 1, hallucination detected else not): Given that there are hallucinations that LLMs cannot self-report on i.e. which can only be identified by subject matter experts, we want to inject some human judgment into identifying hallucinations as well. We do this by leveraging a small dataset constructed from evaluations provided by legal experts, resulting in a labeled corpus of TradeGuard-generated sentences annotated as True, False, or Unsure - resulting in 400 sentence-hallucination assessment pairs. On this, we fit a Bayesian logistic regression which predicts the probability of a given sentence being False i.e a hallucination. The input features are reduced form embeddings (PCA with 8 dimensions) and whether the text contains a number or not. If the model is very sure / unsure about its classification (Bayesian uncertainty<0.25), the classifier outputs the predicted label , else it withholds decision making - this is done to ensure that if the input to the model doesn't resemble the training data (which is highly likely as sentence structures and language vary across countries), the model will avoid making a decision which has a high likelihood of being incorrect.

In ensemble hallucination classification, a sentence is flagged as a hallucination if any of the three methods in the Hallucination Identification stage identifies it as such. If even one sentence in a response generated by TradeGuard is classified as a hallucination, the entire response paragraph is flagged as a hallucination, as hallucinations often occur in clusters [23]. This conservative approach is used because errors in the legal domain are costly, so suspect responses from TradeGuard should be verified with subject matter expert lawyers.

4 Model Evaluation Methodology

Given that there exists no open source dataset for international trade regulation Q&A pairs, we created a novel Q&A dataset containing 5k question & answer pairs about international trade regulation across 150 countries called TradeGov to benchmark the performance of TradeGuard in answering questions at scale and have made it publicly available. To determine the effectiveness of TradeGuard for every question we generate an answer using 1) Vanilla Claude (i.e. Claude accepting just the question as is as the prompt) 2) Vanilla ChatGPT-4o and 3) TradeGuard using Claude V2, Sonnet

[5] We also tried teaching models to express their uncertainty in words [20] but had no success and thus the same is excluded here from discussion

and Sonnet 3.5. We compare the responses generated by Vanilla Claude, Vanilla ChatGPT and TradeGuard along four dimensions to determine which is better: accuracy (Is the answer correct?), completeness (Does the answer contain all the necessary details?), specificity (Does the answer contain too many unnecessary details?) and null rate (Is the answer "I don't know"?).

4.1 TradeGov Dataset : Construction Methodology

To construct an open source benchmark dataset for measuring the performance of LLMs on international trade Q&A, four constraints were at the fore for the source data: 1) it must be non-proprietary, 2) it must be from a reliable, legally trusted source, 3) it should allow periodic updates to reflect changes in trade regulations across 150 countries, and 4) it must cover both high and low income countries. Ideally, this would involve extracting relevant information from each country's official government websites. However, this is an extremely difficult task because the degree to which the international trade regulation information is available for a country varies greatly. For instance in South Korea, the Korean Law Information website has all the required information in highly structured and searchable manner, but for Brazil, the information is neither available in a consolidated or well structured / searchable fashion. Thus, we forego this methodology to avoid bias in the quality and amount of information collected for each country due to a country's online government infrastructure. Using international trade books was ruled out due to copyright concerns. Thus, we determined the Country Commercial Guides on the International Trade Administration website maintained by the US government [27] to be the most suitable source. The website contains information on "market conditions, opportunities, regulations, and business customs prepared at the U.S. Embassies worldwide by Commerce Department,State Department and other U.S. agencies"[27] regarding all countries with any trade relation with the US. It is suitable because 1) it is not a proprietary domain and thus can be scraped and used for making a dataset(double checked with lawyers); 2) is considered to be a reliable source with up-to-date information for international trade regulation by lawyers; 3) updates information regularly and 4) it covers 150 countries. This data source also has the added advantage that is covers key World Trade Organization agreements / treaties as well. However, this website offers a trusted and comprehensive but limited high-level overview of the international trade landscape, with drawbacks including: 1) lack of information on the U.S. domestic trade policies, 2) potential omission of trade agreements to which the US is not a party, and 3) it being in English due to which nuances found in local language sources are lost. Despite these limitations, we argue that this provides a valuable starting point for evaluating LLM performance on international trade related questions at scale, given the current gap in the literature regarding the same.

To create the Q&A dataset, we scrape the information from the website for Customs, Regulations and Standards section for 150 countries. For each country, the website contains information about 11 categories : Trade Barriers, Import Tariffs, Import Requirements and Documentation, Labeling and Marking, Export Controls, Temporary Entry, Prohibited and Restricted Items, Customs Regulations, Standards for Trade, Trade Agreements and Licensing Requirements for Professional Services. To create Q&A pairs, we use ChatGPT-4o and follow these steps: 1) We provide ChatGPT-4o with text scraped from each category and country combination; 2) Using an optimized prompt (see Appendix Section 1), we instruct ChatGPT-4o to generate question-answer pairs based solely on the provided scraped text; to ensure that the generated Q&A pairs come only from the scraped text and not the model's internal world knowledge, we apply Retrieval-Augmented Generation (RAG) principles and ask the ChatGPT to provide exact quotes with citations for each answer it creates. To improve the quality and relevance of the generated Q&A pairs, we used in context learning (ICL) examples along with auto prompt tuning to create a dataset of 5,100 question-answer pairs regarding international trade (see Appendix Table 6 for a sample of generated Q&A pairs in the dataset) ([6]

4.2 Dataset Accuracy Evaluation

Having constructed the data, we determine the quality of the generated Q&A pairs using a human-in-the-loop audit with the following four criteria: 1) **Answer Relevance**: is the answer relevant to the question asked?; 2) **Faithfulness**: is the question-answer pair created only from the scraped text provided? ; 3) **Question Specificity**: is the created question very broad? ; 4) **Answer Specificity**: is the generated answer generic and lacking in details? Our dataset of 5,100 questions achieved 98%

[6] Due to a country name mapping error, the dataset currently has coverage for 138 out for 150 countries. These geographies will be included in forthcoming versions of the dataset.

Table 1: TradeGov Evaluation: Q&A Quality and Bias Assessment

Type	Mean	Correlation	Correlation	Correlation
Metric		Ease of Doing Business	GDP per capita	Trade % of GDP
Relevance	0.976657 (0.15)	0.089 (0.325)	-0.138 (0.156)	-0.040 (0.690)
Question Specificity	0.698419 (0.45)	0.374 (0.000)	-0.376 (0.000)	-0.174 (0.083)
Answer Specificity	0.981363 (0.13)	-0.045 (0.621)	0.046 (0.638)	0.092 (0.365)
Faithfulness	0.977786 (0.15)	-0.168 (0.062)	0.076 (0.435)	0.053 (0.597)
Scraped Text Length (characters)	3520 (4005.01)	-0.350 (0.000)	0.270 (0.005)	-0.020 (0.830)
# Questions per Country	36 (16.27)	-0.180 (0.045)	0.140 (0.141)	-0.190 (0.055)
# Categories per Country	7 (2.12)	-0.170 (0.056)	0.170 (0.087)	-0.150 (0.129)

Brackets in mean column/s contain standard deviation and for correlation columns contain p-values.

Faithfulness , Relevance, and Answer Specificity with 69% specific questions (see Table 1). If a Q&A pair lacks relevance, faithfulness and has a vague answer, it is removed from consideration, leaving us with 4992 Q&A pairs. This dataset consists of approximately 36 questions per country across 7 categories on average (see Table 1). The subject matter of majority of the Q&A pairs is import tariffs, trade standards, trade agreements, import requirements and documentation and trade barriers.

4.3 Dataset Bias Evaluation

Given that our dataset covers 150 countries, there is potential for representation biases. Particularly, it is possible that the dataset has a higher quantity and quality of Q&A pairs for nations that have 1) policies well documented on the internet, 2) are wealthier and 3) have trade as a big part of their economy. For each country in the dataset, we investigate these three potential biases using the correlation between country level average values for the dataset evaluation metrics mentioned in section 2.2 and three macro-economic indicators[7]: 1) **Ease of Doing Business Index**: A proxy for the level of digital documentation of a country's rules and regulations; 2) **GDP per capita (GDP PC)**: An indicator of economic development and 3) **Trade as %age of GDP**.

Referring to Table 1, we see that there is neither any statistically significant correlation between the dataset evaluation metrics and 3 macroeconomic indicators nor is there any discernible geographical bias in the number of Q&A pairs created for a country (see Appendix Figure 5). The only exception to this is Question Specificity - which has statistically significant but weak positive correlation with the Ease Of Doing Business Index and weak negative correlation with GDP PC. This finding holds true across all information categories. [8]. Notably, there is a statistically significant (weakly) positive correlation (0.3 at the 0.001 level) between the average length of the website text scraped and the number of Q&A pairs generated for a country. The average length of the text scraped is also statistically significantly: 1) negatively correlated with the Ease Of Doing Business Index and 2) positively correlated with GDP PC (see Table 1 and Appendix Figure 4). However, interestingly, the number of Q&A pairs generated for a country does not display a similar correlation - it is only weakly negatively correlated with the Ease Of Doing Business Index (at the 0.5 level of significance; see Table 1) ; we hypothesize this is due to the construction of our prompt which limits the number of Q& A pairs created for any country and topic to the range 5 to 10.

The above results are encouraging as they demonstrate that the TradeGov dataset does not have any obviously discernible biases in it, which helps the dataset have broad and credible applicability across all countries for international trade Q&A related tasks.

5 Model Performance

5.1 Model Accuracy, Completeness and Specificity

The evaluation on the TradeGov Dataset revealed that the TradeGuard framework always outperforms it's vanilla LLM counterparts in generation of correct answers (average accuracy uplift is 4%; see Table 2) and has a lower null response rate than it's vanilla LLM counterparts (average null rate

[7] Source : World Bank Open Data (https://data.worldbank.org/)

[8] Note: Topic modeling for each country using Latent Dirichlet Allocation didn't show any discernible differences in the content of the text scraped across countries and thus is omitted from discussion here.

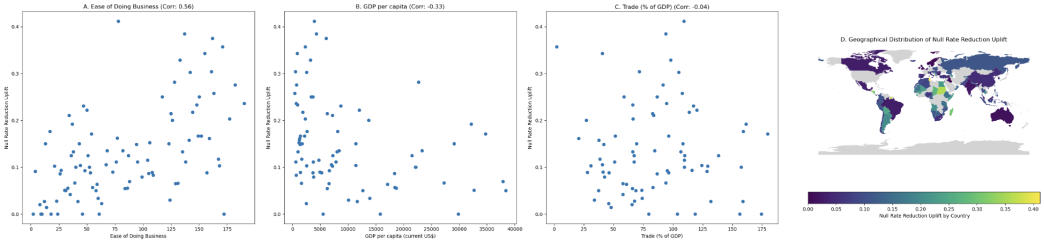


Figure 2: TradeGuard (Sonnet) - Null Rate Reduction Analysis (TradeGov Dataset)

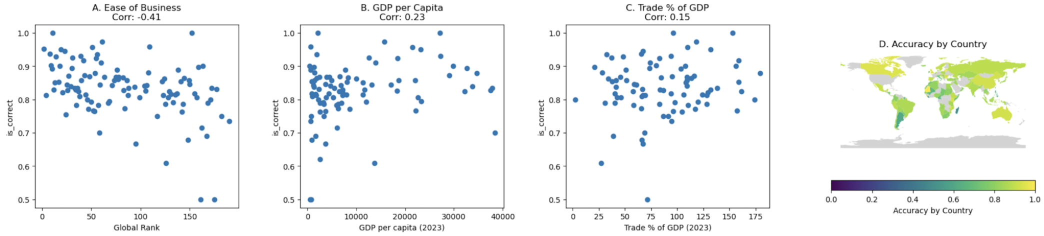


Figure 3: TradeGuard (Sonnet) - Accuracy Analysis (TradeGov Dataset)

reduction uplift is 18%; see Table 2). For both TradeGuard variants, for 60% of the answers where vanilla LLMs said "I don't know" TradeGuard generated the correct response. TradeGuard (Sonnet) has higher accuracy accuracy compared to vanilla Sonnet 3.5 and ChatGPT-40 (see Table 2 3)[9]. Thus, the remainder of the analysis focuses on TradeGuard (Sonnet). Examining the relationship between null response rates by country and the macro economic indicators, we see that using TradeGuard (Sonnet) 1) reduces the negative correlation between null rate and ease of doing business and 2) reduces the positive correlation between null rate and GPD PC (see Figure 2). The highest reduction in null rate is for lower income countries and countries with worse ease of doing business indexes. However, no such trend is observed between uplift in average correctness rates and these macroeconomic indicators.

Note that TradeGuard has lower completeness as compared to its vanilla ChatGPT and Sonnet 3.5. We hypothesize that this might be because of the use of majority vote summarization and multi-agent debate but leave further investigation into this for future versions of the paper. TradeGuard's Completeness is also negatively correlated with ease of doing business and weakly positively correlated with GDP PC, with less complete answers on average being concentrated in the Africa and South America (see Appendix Figure 6). TradeGuard (Sonnet) also has the lowest specificity (12%) as compared to other models.

5.2 Hallucination Identification

In the evaluation of TradeGuard's hallucination detection capabilities on the TradeGov dataset, the ensemble combining all three methods yielded the highest F1 score of 90% (see Ensemble (OR) in

[9] Due to throughput constrains, TradeGuard (Sonnet) was benchmarked against ChatGPT and Sonnet 3.5 on a smaller dataset of 1500 questions across 150 (it excluded the following categories: export-controls, temporary-entry, standards-trade, licensing-requirements-professional-services)

Table 2: Performance Metrics (TradeGov Dataset) : TradeGuard vs Vanilla Claude

Metrics	Claude V2	TradeGuard (Claude v2)	Sonnet	TradeGuard (Sonnet)
Accuracy	0.762560	0.814985	0.797203	0.827807
Null Rate	0.378331	0.124509	0.182854	0.058770
Completeness	0.512704	0.572863	0.727986	0.582252
Specificity	0.241694	0.097851	0.399578	0.144347

Table 3: Performance Metrics (TradeGov Dataset) : TradeGuard vs Vanilla ChatGPT vs Vanilla Sonnet 3.5

Metrics	ChatGPT-40	Sonnet 3.5	TradeGuard (Sonnet)
Accuracy	0.859661	0.882034	0.910508
Null Rate	0.030508	0.000000	0.000000
Completeness	0.769492	0.394576	0.719322
Specificity	0.432542	0.446780	0.124068

Table 4: Hallucination Identification Methods: TradeGov Dataset

Hallucination Identification Method	Recall	Precision	F1 Score	Corr(Ease of Doing Business)	Corr(GDP PC)	Corr(Trade %age of GDP)
Entailment Score	0.957763	0.832827	0.890936	0.00	0.11	0.03
Contains Numbers (Benchmark)	0.402855	0.832631	0.542992	-0.27	0.08	-0.03
Factuality Classification	0.903000	0.831099	0.865559	0.04	-0.07	0.06
Human Augmented Verification (Bayesian Regression)	0.512089	0.848865	0.638808	-0.35	0.21	0.19
Ensemble (OR)	0.988057	0.834646	0.904895	-0.01	0.70	-0.15

Table 4). Among the individual hallucination detection methods, the entailment score performed best with an F1 score of 89%. No significant trends were observed in hallucination detection related to ease of doing business, GDP per capita,trade share or geographical distribution (refer to Appendix Figure 7 for details). However, notably, when answering questions that stumped the vanilla version and resulted in null responses, TradeGuard was correct 60% of the time. However, TradeGuard's ensemble hallucination framework flagged all of these correct answers as potential hallucinations, regardless of which TradeGuard LLM variant was used.

6 Conclusion

In this paper we introduced two novel artifacts: 1) the TradeGuard framework - for mitigating hallucinations in trade related queries to LLMs and 2) the TradeGov dataset - the first open source dataset for measuring the performance of LLMs on international trade regulation related questions. We show that 1) TradeGuard has an accuracy of 82% and outperforms vanilla Claude (V2, Sonnet and Sonnet 3.5) and ChatGPT in answering trade law related questions and 2) TradeGuard has an F1 score of 91% for hallucination detection to highlight incorrect claims - capabilities that apply equally across all 150 countries. We also demonstrated that TradeGuard reduces the null answers, particularly for low income countries. As future improvement, we aim to dive deeper into the biases and errors of both TradeGuard and the TradeGov dataset to improve them iteratively. To provide continued support for such analysis, improving the generation of Q&A pairs for the TradeGov dataset iteratively is key. More context needs to be added to the questions to reduce ambiguity and improve Question Specificity. The adherence of the Q&A generation to instructions regarding no duplication needs to be addressed as well - despite asking the model to not generate duplicate question, we get questions which are very similar in meaning (Ex: "What is the role of INMETRO in Brazil's regulatory regime?" ; "What is INMETRO responsible for in Brazil according to international trade law?" are the same question). Furthermore, most questions are factual (96% are "what" questions) and focus on recalling information rather than understanding the international trade landscape. The TradeGov dataset also lacks information regarding agriculture - only 2% of the queries include agriculture or food. This is a critical gap for emerging markets where majority of trade policies deals with agriculture. We shall use few-shot ICL and iterative prompt tuning to improve question specificity, reduce duplication and encourage generation of more cause and effect related questions to make both TradeGuard and TradeGov more robust in the future.

References

[1] Abad, A.A. (2024) Artificial Intelligence and the future of International Trade Law and Dispute Settlement, SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract$_id = 4849453(Accessed : 14September2024)$.

[2] Abdallah, A., Piryani, B. and Jatowt, A. (2023) Exploring the state of the art in Legal QA Systems - Journal of Big Data, SpringerOpen. Available at: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-023-00802-8 (Accessed: 14 September 2024).

[3] CHATWTO: An analysis of generative artificial intelligence and international trade 2024 (no date) World Economic Forum. Available at: https://www.weforum.org/publications/chatwto-an-analysis-of-generative-artificial-intelligence-and-international-trade/ (Accessed: 14 September 2024).

[4] Choi, J.H. et al. (2023) Chatgpt goes to law school, SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract$_id = 4335905(Accessed : 14September2024)$.

[5] Cohen, R. et al. (2023) LM vs LM: Detecting factual errors via cross examination, arXiv.org. Available at: https://arxiv.org/abs/2305.13281 (Accessed: 14 September 2024).

[6] Colombo, P. et al. (2024) SAULLM-7B: A pioneering large language model for law, arXiv.org. Available at: https://arxiv.org/abs/2403.03883 (Accessed: 14 September 2024).

[7] Cui, J. et al. (2022) A survey on legal judgment prediction: Datasets, Metrics, models and challenges, arXiv.org. Available at: https://arxiv.org/abs/2204.04859 (Accessed: 14 September 2024).

[8] Cui, J. et al. (2024) Chatlaw: A multi-agent collaborative legal assistant with knowledge graph enhanced mixture-of-experts large language model, arXiv.org. Available at: https://arxiv.org/abs/2306.16092 (Accessed: 14 September 2024).

[9] Dahl, M. et al. (2024) Large legal fictions: Profiling legal hallucinations in large language models, arXiv.org. Available at: https://arxiv.org/abs/2401.01301 (Accessed: 14 September 2024).

[10] Deroy, A., Ghosh, K. and Ghosh, S. (2023) How ready are pre-trained abstractive models and llms for legal case judgement summarization?, arXiv.org. Available at: https://arxiv.org/abs/2306.01248 (Accessed: 14 September 2024).

[11] Du, Y. et al. (2023) Improving factuality and reasoning in language models through Multiagent Debate, arXiv.org. Available at: https://arxiv.org/abs/2305.14325 (Accessed: 14 September 2024).

[12] eClear (2023) Leveraging large language models in customs, eClear AG. Available at: https://eclear.com/article/leveraging-large-language-models-in-customs/ (Accessed: 14 September 2024).

[13] Guha, N. et al. (2023) LegalBench: A collaboratively built benchmark for measuring legal..., OpenReview. Available at: https://openreview.net/forum?id=WqSPQFxFRC (Accessed: 14 September 2024).

[14] Hallucinating law: Legal mistakes with large language models are pervasive (no date) Stanford HAI. Available at: https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive (Accessed: 14 September 2024).

[15] Henderson, P. et al. (2022) Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset, arXiv.org. Available at: https://arxiv.org/abs/2207.00220 (Accessed: 14 September 2024).

[16] Kim, M.-Y. et al. (2014) 'Answering yes/no questions in legal bar exams', Lecture Notes in Computer Science, pp. 199–213. doi:10.1007/978-3-319-10061-6$_1$4.

[17] Magesh, V. et al. (2024) Hallucination-free? assessing the reliability of leading AI Legal Research Tools, arXiv.org. Available at: https://arxiv.org/abs/2405.20362 (Accessed: 14 September 2024).

[18] Mündler, N. et al. (2024) Self-contradictory hallucinations of large language models: Evaluation, Detection and Mitigation, arXiv.org. Available at: https://arxiv.org/abs/2305.15852 (Accessed: 14 September 2024).

[19] Polsley, S., Jhunjhunwala, P. and Huang, R. (no date) Casesummarizer: A system for automated summarization of legal texts, ACL Anthology. Available at: https://aclanthology.org/C16-2054/ (Accessed: 14 September 2024).

[20] Tian, K. et al. (2023) Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback, arXiv.org. Available at: https://arxiv.org/abs/2305.14975 (Accessed: 14 September 2024).

[21] Turpin, M. et al. (2023) Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, arXiv.org. Available at: https://arxiv.org/abs/2305.04388 (Accessed: 14 September 2024).

[22] Zhang, J. (no date) SAC3: Reliable hallucination detection in black-box language models via Semantic-aware cross-check consistency. Available at: https://arxiv.org/html/2311.01740v2 (Accessed: 14 September 2024).

[23] Zhang, M. et al. (2023) How language model hallucinations can snowball, arXiv.org. Available at: https://arxiv.org/abs/2305.13534 (Accessed: 14 September 2024).

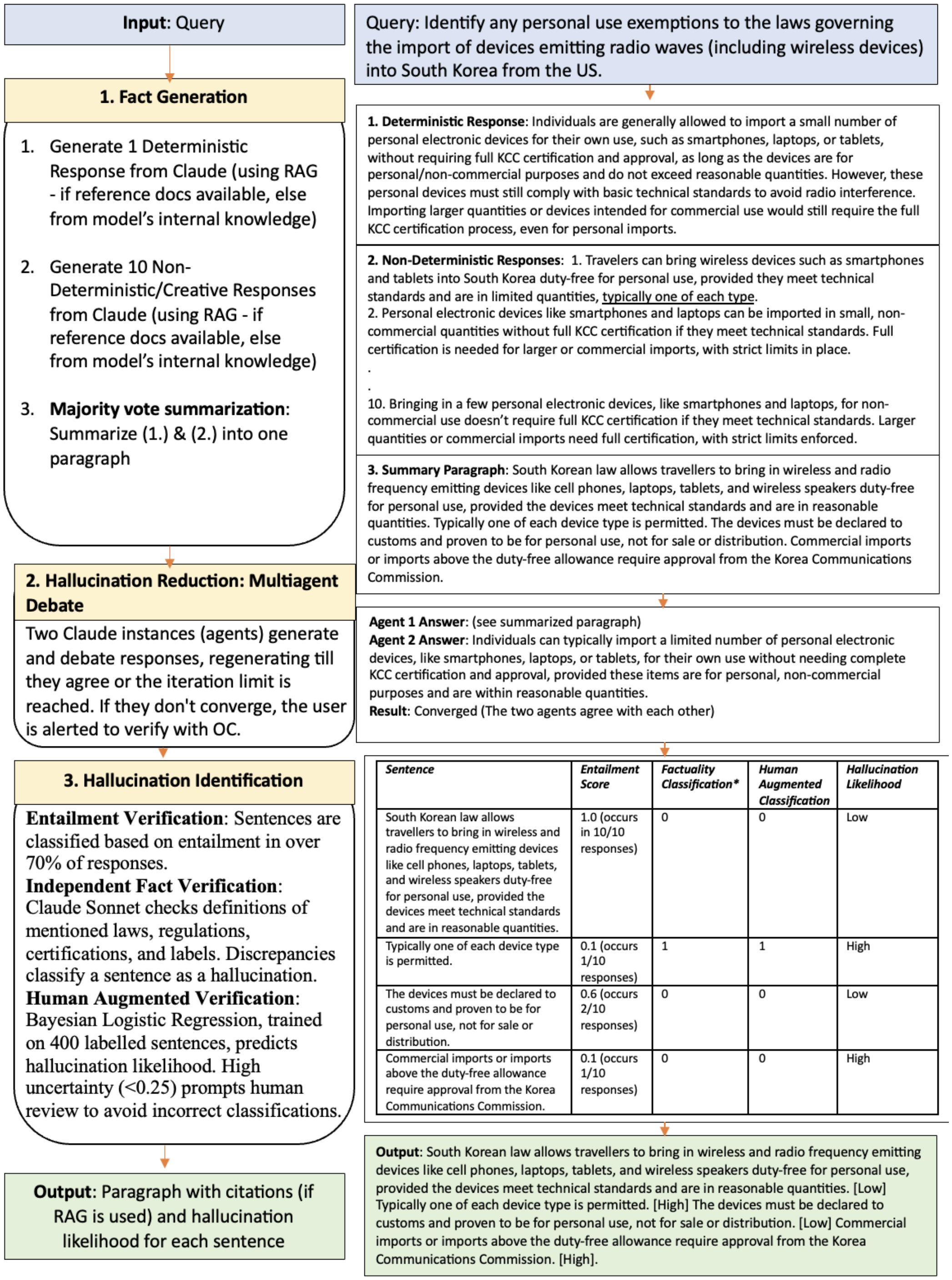
[24] Yin, Z. et al. (2023) 'Do large language models know what they don't know?', Findings of the Association for Computational Linguistics: ACL 2023, pp. 8653–8665. doi:10.18653/v1/2023.findings-acl.551.

[25] Manakul, P., Liusie, A. and Gales, M.J.F. (2023) SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models, arXiv.org. Available at: https://arxiv.org/abs/2303.08896 (Accessed: 14 September 2024).

[26] Kuhn, L., Gal, Y. and Farquhar, S. (2023) Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation, arXiv.org. Available at: https://arxiv.org/abs/2302.09664 (Accessed: 14 September 2024).

[27] Country commercial guides, International Trade Administration | Trade.gov. Available at: https://www.trade.gov/country-commercial-guides (Accessed: 14 September 2024).

Appendix



Input: Query

1. Fact Generation

1. Generate 1 Deterministic Response from Claude (using RAG - if reference docs available, else from model's internal knowledge)
2. Generate 10 Non-Deterministic/Creative Responses from Claude (using RAG - if reference docs available, else from model's internal knowledge)
3. **Majority vote summarization**: Summarize (1.) & (2.) into one paragraph

2. Hallucination Reduction: Multiagent Debate

Two Claude instances (agents) generate and debate responses, regenerating till they agree or the iteration limit is reached. If they don't converge, the user is alerted to verify with OC.

3. Hallucination Identification

Entailment Verification: Sentences are classified based on entailment in over 70% of responses.
Independent Fact Verification: Claude Sonnet checks definitions of mentioned laws, regulations, certifications, and labels. Discrepancies classify a sentence as a hallucination.
Human Augmented Verification: Bayesian Logistic Regression, trained on 400 labelled sentences, predicts hallucination likelihood. High uncertainty (<0.25) prompts human review to avoid incorrect classifications.

Output: Paragraph with citations (if RAG is used) and hallucination likelihood for each sentence

Query: Identify any personal use exemptions to the laws governing the import of devices emitting radio waves (including wireless devices) into South Korea from the US.

1. Deterministic Response: Individuals are generally allowed to import a small number of personal electronic devices for their own use, such as smartphones, laptops, or tablets, without requiring full KCC certification and approval, as long as the devices are for personal/non-commercial purposes and do not exceed reasonable quantities. However, these personal devices must still comply with basic technical standards to avoid radio interference. Importing larger quantities or devices intended for commercial use would still require the full KCC certification process, even for personal imports.

2. Non-Deterministic Responses: 1. Travelers can bring wireless devices such as smartphones and tablets into South Korea duty-free for personal use, provided they meet technical standards and are in limited quantities, typically one of each type.
2. Personal electronic devices like smartphones and laptops can be imported in small, non-commercial quantities without full KCC certification if they meet technical standards. Full certification is needed for larger or commercial imports, with strict limits in place.
.
.
10. Bringing in a few personal electronic devices, like smartphones and laptops, for non-commercial use doesn't require full KCC certification if they meet technical standards. Larger quantities or commercial imports need full certification, with strict limits enforced.

3. Summary Paragraph: South Korean law allows travellers to bring in wireless and radio frequency emitting devices like cell phones, laptops, tablets, and wireless speakers duty-free for personal use, provided the devices meet technical standards and are in reasonable quantities. Typically one of each device type is permitted. The devices must be declared to customs and proven to be for personal use, not for sale or distribution. Commercial imports or imports above the duty-free allowance require approval from the Korea Communications Commission.

Agent 1 Answer: (see summarized paragraph)
Agent 2 Answer: Individuals can typically import a limited number of personal electronic devices, like smartphones, laptops, or tablets, for their own use without needing complete KCC certification and approval, provided these items are for personal, non-commercial purposes and are within reasonable quantities.
Result: Converged (The two agents agree with each other)

Sentence	*Entailment Score*	*Factuality Classification**	*Human Augmented Classification*	*Hallucination Likelihood*
South Korean law allows travellers to bring in wireless and radio frequency emitting devices like cell phones, laptops, tablets, and wireless speakers duty-free for personal use, provided the devices meet technical standards and are in reasonable quantities.	1.0 (occurs in 10/10 responses)	0	0	Low
Typically one of each device type is permitted.	0.1 (occurs 1/10 responses)	1	1	High
The devices must be declared to customs and proven to be for personal use, not for sale or distribution.	0.6 (occurs 2/10 responses)	0	0	Low
Commercial imports or imports above the duty-free allowance require approval from the Korea Communications Commission.	0.1 (occurs 1/10 responses)	0	0	High

Output: South Korean law allows travellers to bring in wireless and radio frequency emitting devices like cell phones, laptops, tablets, and wireless speakers duty-free for personal use, provided the devices meet technical standards and are in reasonable quantities. [Low] Typically one of each device type is permitted. [High] The devices must be declared to customs and proven to be for personal use, not for sale or distribution. [Low] Commercial imports or imports above the duty-free allowance require approval from the Korea Communications Commission. [High].

Figure 4: TradeGuard: Model Architecture and Example Walkthrough

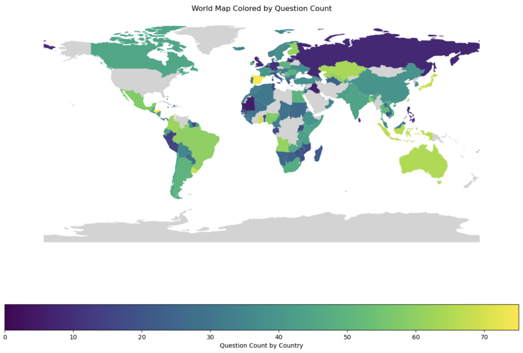

World Map Colored by Question Count
0
10
20
30
40
50
60
70
Question Count by Country

Figure 5: TradeGov Dataset: Question Count Geographical Distribution

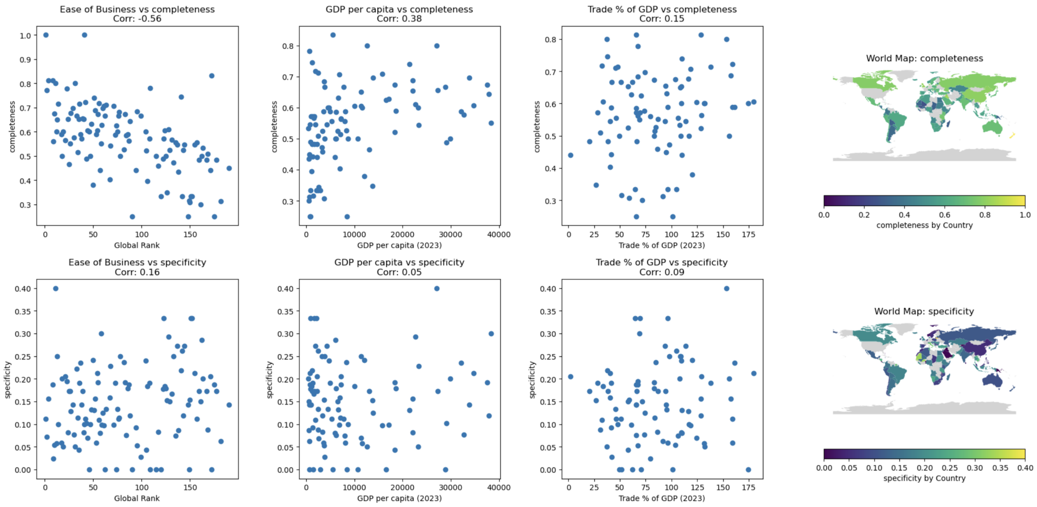


Figure 6: TradeGuard (Sonnet): Completeness and Specificity Analysis

Table 5: Accuracy After Removing Null Responses

	Claude V2	TradeGuard (Claude v2)	Sonnet	TradeGuard (Sonnet)
Accuracy	0.894238	0.807385	0.858003	0.834062
# Samples	2846	4008	3155	4116

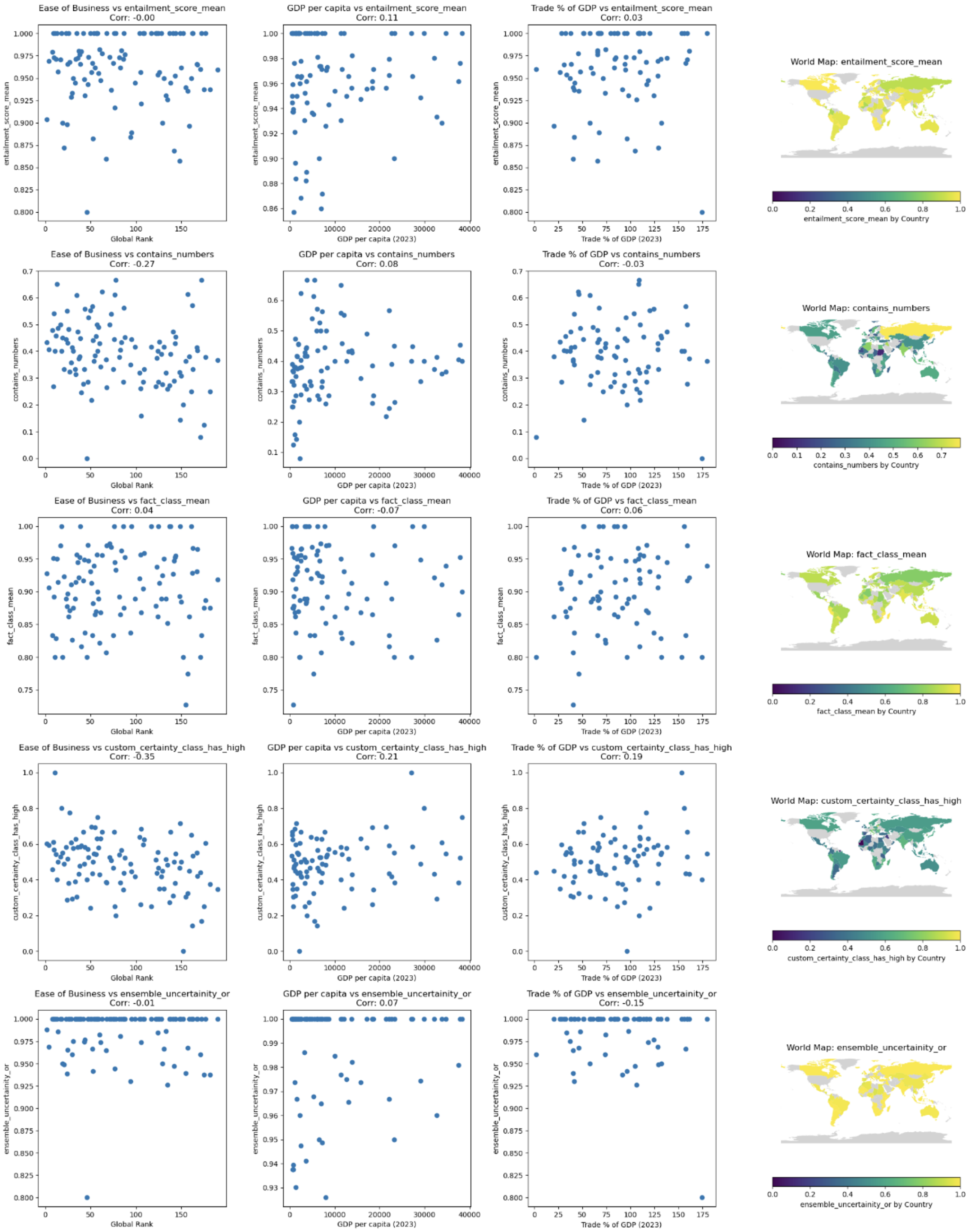


Figure 7: TradeGuard (Sonnet): Hallucination Analysis

Table 6: TradeGov Dataset : Sample Q&As

Questions	Answers
With which agency must products that affect the human body be registered in Brazil?	Such products must be registered with Brazil's Health Regulatory Agency, ANVISA. (Paragraph 1, Sentence 5)
What is the VAT rate on all imports and domestically manufactured goods in Korea?	Korea has a flat 10 percent Value Added Tax (VAT) on all imports and domestically manufactured goods. (Paragraph 3, Sentence 1)
What is the purpose of the CE Mark in Turkey's international trade law?	The CE Mark was established by the EU to ensure products circulating within Europe met certain health, safety, consumer, and environmental protection standards. (Paragraph 2, Sentence 2)
Which organizations certify the quality of most non-medical goods in Zimbabwe?	The Standards Association of Zimbabwe and Bureau Veritas certify the quality of most non-medical goods produced or imported into the country. (Paragraph 1, Sentence 3)
Which ministry in Vietnam publishes a list of goods with HS codes in the Import and Export Tariffs?	The Ministry of National Defense publishes a list of goods with HS codes in the Import and Export Tariffs. (Paragraph 2, Sentence 1)
Are Certificates of Origin required for U.S. goods imported into Ireland?	No, Certificates of Origin are not required for U.S. goods. (Paragraph 4, Sentence 10)
What is the role of the Uzbek Agency for Technical Regulation in Uzbekistan?	The Uzbek Agency for Technical Regulation is responsible for certification and standardization policy. (Paragraph 3, Sentence 1)
How long is an import license valid for motor vehicles in Uruguay?	An import license is valid for 60 days (90 days for motor vehicles) after approval. (Paragraph 1, Sentence 8)
How is VAT charged on imported goods in the UK?	VAT is charged as though it is a customs duty. (Paragraph 2, Sentence 3)
What document details the commodity codes for VAT in the UK?	VAT liability is ascertained using 'commodity codes,' detailed in the 'UK Trade Tariff: Volume 1' from HMRC. (Paragraph 3, Sentence 1)
What are the three rates of import duties in Ukraine's tariff schedule?	Ukraine's import tariff schedule includes Full, Most Favored Nation (MFN), and Preferential rates. (Paragraph 2, Sentence 1)
What does Brazil's conformity assessment system follow?	Brazil's conformity assessment system follows ISO guidelines. (Paragraph 3, Sentence 2)
How does Tunisia calculate VAT on imported goods?	VAT is calculated on the base price plus import duties, surcharges, and consumption taxes. (Paragraph 1, Sentence 12)
What system does Thailand use for import classification?	Thailand classifies imports using the Harmonized System (HS). (Paragraph 2, Sentence 2)
How many Free Trade Zone (FTZ) authorities exist in Singapore?	Singapore has three FTZ authorities: PSA Corporation Ltd, Jurong Port Pte Ltd, and Changi Airport Group. (Paragraph 3, Sentence 1)
Are tariffs on U.S. imports the same as those on EU imports in Serbia?	No, tariffs/duties on U.S. imports differ from those on EU imports. (Paragraph 2, Sentence 6)
What labeling regulations apply to food in Serbia?	The Rulebook on Declaration, Labeling, and Advertising of Food (RS OG No. 19/17 and 16/18) defines food labeling regulations. (Paragraph 3, Sentence 1)
How can low-value commercial samples be imported into Poland?	Zero or low-value samples can be imported duty-free with a written statement confirming their value. (Paragraph 1, Sentence 4)
What documents are needed for customs clearance in Nigeria?	Required documents include a bill of lading, commercial invoice, exit note, Form 'M' entry declaration, packing list, single goods declaration, and a product certificate. (Paragraph 3, Sentence 1)
When were import quotas on yellow corn and pork phased out in Nicaragua?	Import quotas on yellow corn and pork meat were phased out in 2020. (Paragraph 1, Sentence 10)

(Continued on next page)

(Continued from previous page)

Questions	Answers
Where can a list of prohibited items and HS codes for Mexico be found?	The list is available on the Prohibited Items List at the Mexican Customs website. (Paragraph 1, Sentence 9)
What does the Mauritius-Turkey free trade agreement cover?	The agreement allows duty-free access for industrial products and specific agricultural products, including chilled fish and tropical fruits. (Paragraph 1, Sentence 16)
What duty is assessed on tobacco products in Kuwait?	Tobacco products are subject to a 100% duty. (Paragraph 2, Sentence 5)
At what stage is labeling not required for imports in Japan?	Labeling is not required at customs clearance but at the point of sale. (Paragraph 1, Sentence 2)

.1 Prompt for creating TradeGov Dataset Q&A Pairs using ChatGPT 4o

```
Example Text Extract : The following labeling information must be in Croatian on the
     original package of products subject to quality control: name of the product;
    full address of the producer or full address of the importer; net quantity,
    weight, or volume; ingredients; usage and storage particulars; and any
    important warnings about the product for the consumer. Technically complicated
    products must include instructions for use, the manufacturers specifications, a
     list of authorized maintenance offices, warranty, and other applicable data.
    Every certified product must carry a CE mark indicating that the product has
    undergone appropriate testing and that it conforms to the provisions of the
    relevant regulations. Foreign labels, including the U.S. standard label, are
    not acceptable; stick-on labels that meet local requirements are allowed for
    products that contain a foreign label.

Prompt :

f"""Read the following text and create 5 to 10 question-answer pairs related to
    international trade law for {country_name}. Each question must include the name
     of the country. Answers should be exact quotes from the text with citations in
     the format (paragraph number, sentence number). Avoid non-trade related
    questions and duplicates.

Examples:

Question: What registration process must Brazilian importers follow according to
    Brazilian international trade law?

Answer: "Brazilian importers must register with the Foreign Trade Secretariat (SECEX
    ), a branch of the Ministry of Development, Industry, Trade and Services (MDIC)
     via its Integrated System for Foreign Trade (Siscomex)." (Paragraph 1,
    Sentence 1)

Question: What determines if additional documentation is required for imported
    products in Brazil?

Answer: "Depending on the product, Brazilian authorities may require additional
    documentation." (Paragraph 1, Sentence 2)

Question: Which ministry controls products that may affect the human body in Brazil?

Answer: "For instance, the Ministry of Health controls all products that may affect
    the human body, including pharmaceuticals, vitamins, cosmetics and medical
    equipment/devices." (Paragraph 1, Sentence 3)

Text: {text_extract}"""
```

.2 Prompt for evaluating the performance of ChatGPT on TradeGov Dataset

```
Question : What is required for all vehicles, both new and used, that are imported
    into Russia according to technical regulation TR TS 018-2011?

Prompt :

f"""Answer the following question. If you don’t know the answer to a particular
    question, answer with ’I dont know’.\nQuestion: {question}\nAnswer:"""
```