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

TradeGov Dataset is Forthcoming at:

https://github.com/amazon-science/tradegov-dataset

Introduction & Contribution Summary

Problem: Understanding and complying with the rapidly evolving international trade landscape is crucial to minimize losses and maximize revenue generation. However, **navigating global trade requires specialized, expensive legal expertise which is not equitably available leading to competitiveness gaps** - large entities can afford this expertise while smaller entities cannot, hindering their ability of compete globally.

Solution : LLM Assisted International Trade Information Generation & Retrieval can bridge the resource and knowledge gap by generating reliable on-demand information about international trade regulations.

- **But can LLMs provide reliable, accurate and fair information regarding international trade regulation? We don't know** because the LLM evaluation literature does not address the capabilities of LLMs for international trade related tasks. A primary impediment is the lack of a dataset for benchmarking the performance of LLMs on Q&A tasks related to international trade.

Main Contribution: To address the gap in the literature, we :

1) Introduce the TradeGov Dataset - the first benchmark for international trade Q&A: TradeGov is a human audited dataset containing 5k international trade related question-answer pairs across 138 countries.

2) First Systematic Evaluation of LLMs on International Trade Related Questions: ChatGPT-4o achieves 84% accuracy while Claude Sonnet 3.5 achieves 88% accuracy on the TradeGov dataset

3) Introduce TradeGuard – the first trade regulation hallucination mitigation framework that leverages majority vote summarization and multi-agent debate to achieve 91% accuracy on the TradeGov dataset

TradeGov Dataset Construction Methodology

TradeGov is created using human audited, ChatGPT backed Retrieval Augmented Generation (RAG) based on the Country Commercial Guides (CCG) on the International Trade Administration website maintained by the US Government.

TradeGov Dataset Generation Process

1. Scrape CCG

Text scraped from the 'Customs, Regulations and Standards' section for 11 categories across 138 countries

Why CCG as source?

Trusted by lawyers; Up to date; Non-proprietary; Covers 150 countries & WTO agreements

2. Generate Q&As

An optimized prompt with In-context learning instructed ChatGPT to create Q&A pairs based solely on the provided scraped text, citing sources to encourage faithfulness.

3. Audit

4 metric Human-in-the-loop evaluation. If a Q&A pair lacks relevance, faithfulness and has a vague answer, it is removed, leaving 4992 Q&A pairs.

Example Output : Generated Q&A Pair In TradeGov Dataset

Question: Are Certificates of Origin required for U.S. goods imported into Ireland?

Answer: No, Certificates of Origin are not required for U.S. goods. (Paragraph 4, Sentence 10)

TradeGov Dataset Evaluation

- **TradeGov achieves 98% relevance and faithfulness**
- **Doesn't show any systematic biases along macroeconomic and geographical dimensions, lending itself to equal applicably for LLM assessment across countries.**

Fig. 1: Geographical Distribution of TradeGov Dataset Question Count

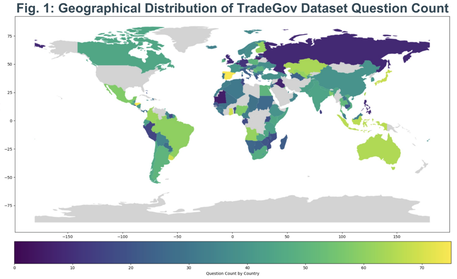
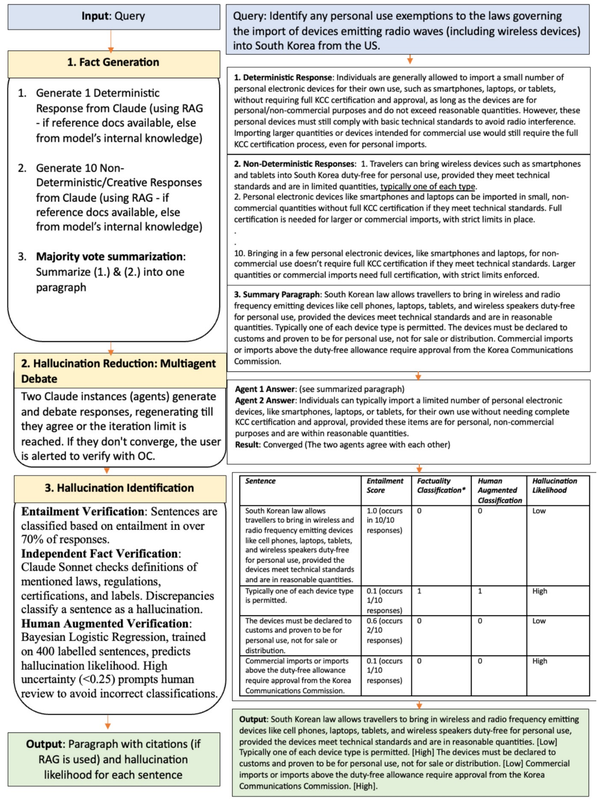


Table 1: TradeGov Evaluation: Q&A Quality and Bias Assessment

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.

TradeGuard Architecture



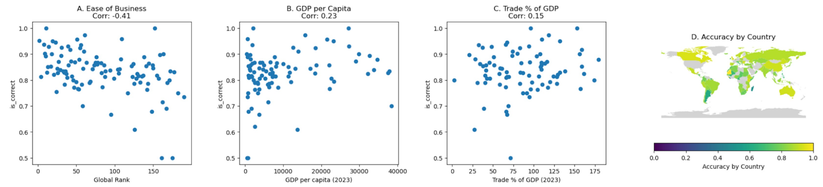
LLM Benchmarking Results

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

- **TradeGuard always outperforms it's vanilla LLM counterparts in generation of correct answers (average accuracy uplift is 4%) and has a lower null response rate than it's vanilla LLM counterparts (average null rate reduction uplift is 18%).**

Fig. 3: TradeGuard Accuracy Analysis



- **TradeGuard 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.**

Fig. 4: TradeGuard Null Rate Reduction Analysis

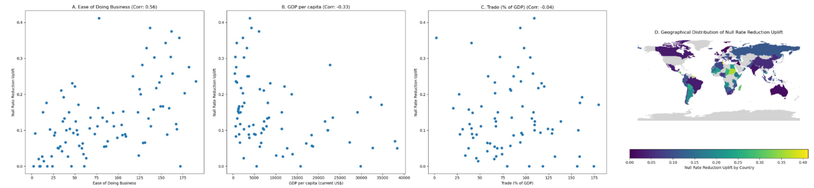


Table 3: Hallucination Identification Methods

Hallucination Identification Method	Recall	Precision	F1 Score
Entailment Score	0.957763	0.832827	0.890936
Contains Numbers (Benchmark)	0.402855	0.832631	0.542992
Factuality Classification	0.903000	0.831099	0.865559
Human Augmented Verification (Bayesian Regression)	0.512089	0.848865	0.638808
Ensemble (OR)	0.988057	0.834646	0.904895

- TradeGuard's ensemble hallucination detection algorithm — combining entailment verification, cross-questioning, and Bayesian regression— achieves an F1 score of 91%

Limitations and Future Work

- **Engage legal experts** for subject matter expert driven dataset and LLM evaluation
- **Improve question diversity** beyond fact recall (currently 96% are 'what' questions) to include cause and effect questions
- **Expand topical coverage** particularly by including more agriculture related questions (currently only 2% of the queries).