What Do LLMs Know About International Trade? Introducing TradeGov Dataset for International Trade Q&A Evaluation

Kriti Mahajan , Amazon, kritimhj@amazon.com

TradeGov Dataset is Forthcoming at:

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

Introduction & Contribution Summary

Problem: Keeping Up with International Trade Regulations in a rapidly evolving global geopolitical landscape is a key challenge for governments and businesses alike . Understanding and complying with international trade 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.

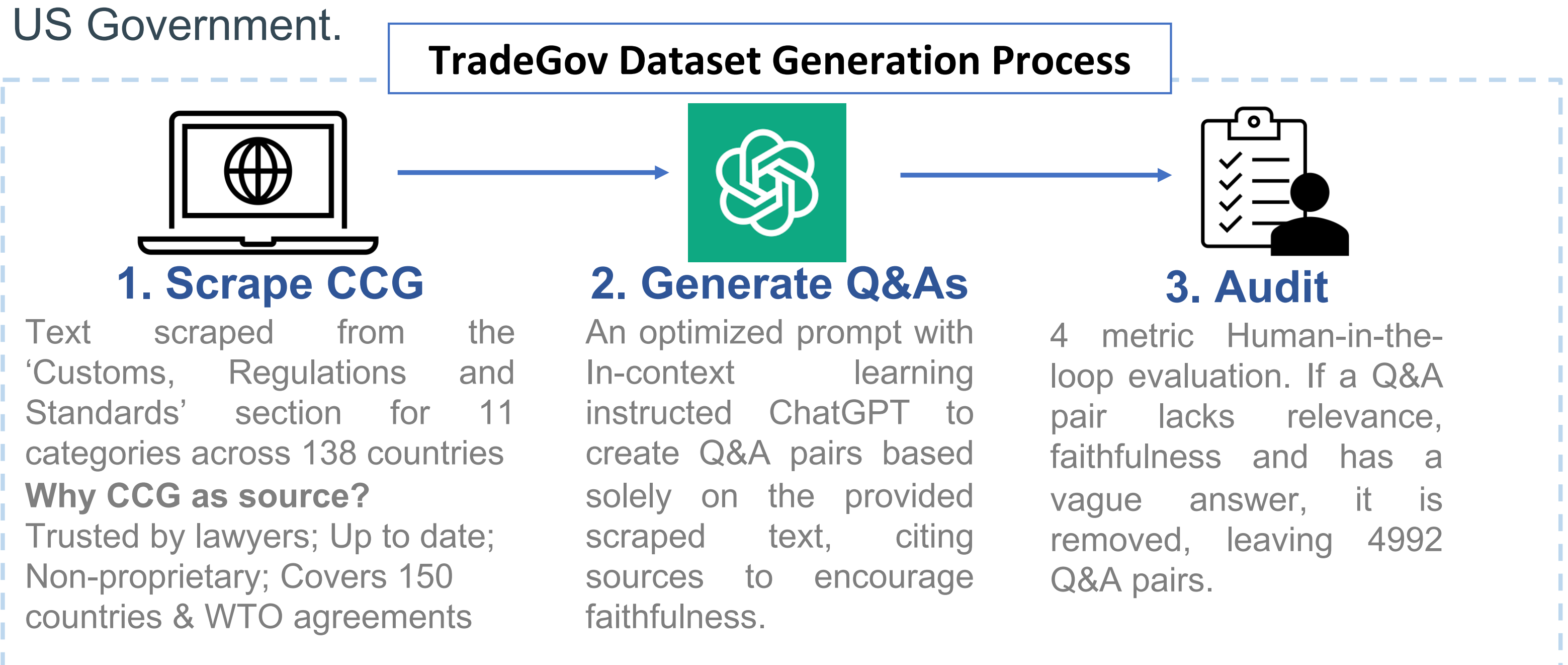
Main Contribution: To address the gap in the literature, we :

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 on the TradeGov dataset but it performs better for countries with greater ease of business, higher GDP and higher trade shares.

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?

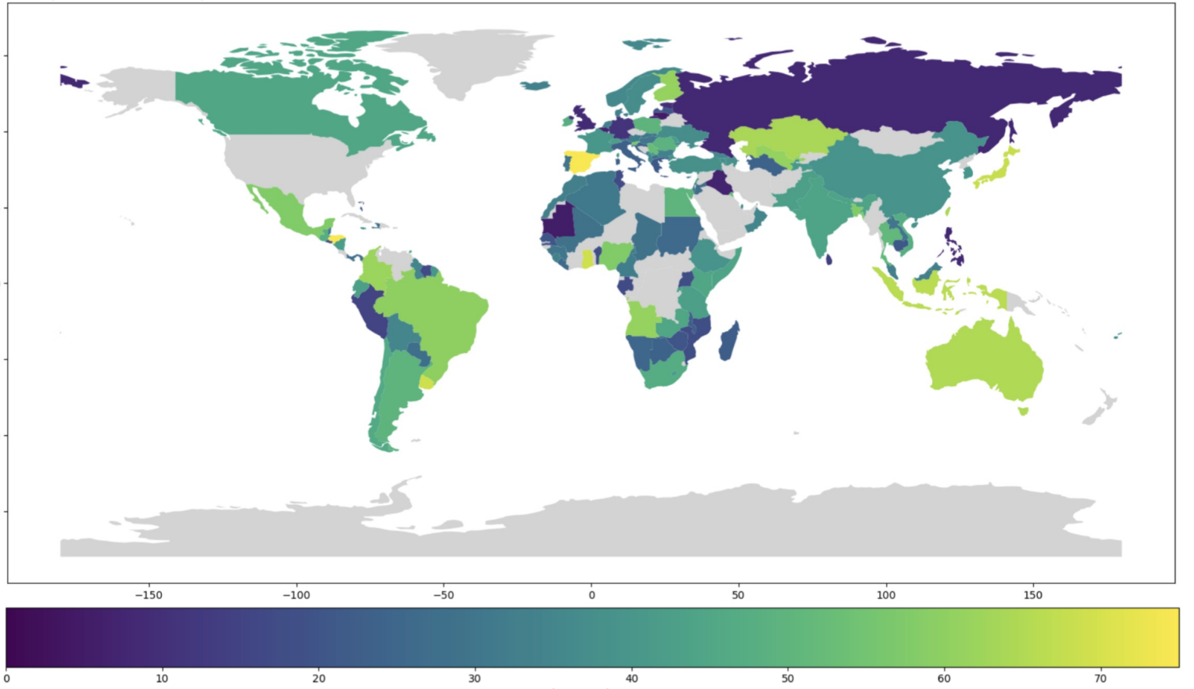
Answer: No, Certificates of Origin are not required for U.S. goods. (Paragraph 4, Sentence 10)

TradeGov Dataset Evaluation

A. Quality Assessment: Human-in-the-loop evaluates 4 criterions:

- **Answer Relevance:** is the answer relevant to the question asked?
- **Faithfulness:** is the Q&A pair created only from the scraped text provided?
- **Question Specificity:** is the created question very broad?
- **Answer Specificity:** is the generated answer generic, lacking in detail?

Fig. 1: Geographical Distribution of TradeGov Dataset Question Count



B. Macroeconomic and Geographical Bias Assessment: 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 and three macro-economic indicators:

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 (GDPPC):** An indicator of economic development

3) **Trade as %age of GDP**.

Result:

- **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.**

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.

ChatGPT-4o Benchmarking on TradeGov

Responses generated by ChatGPT on the questions in the TradeGov dataset are evaluated across 4 dimensions:

- **Accuracy**: Does the answer generated by the LLM contain the key facts in the benchmark TradeGov answer?
- **Completeness**: Does the LLM answer contain all the details mentioned in the benchmark TradeGov answer?
- **Specificity**: Does the LLM answer contain unnecessary details?
- **Null response rate**: Is the answer "I don't know"?

ChatGPT-4o Benchmarking Results

Table 2: ChatGPT Evaluation: Answer Quality and Bias Assessment

Type	Mean	Correlation	Correlation	Correlation
Metric		Ease of Doing Business	GDP per capita	Trade % of GDP
Null Rate	0.163 (0.369)	0.51 (0.0)	-0.28 (0.004)	-0.18 (0.07)
Accuracy	0.845 (0.361)	-0.586 (0.0)	0.345 (0.0)	0.236 (0.018)
Completeness	0.740 (0.438)	-0.539 (0.0)	0.377 (0.0)	0.22 (0.028)
Specificity	0.400 (0.4900)	0.229 (0.01)	-0.111 (0.255)	-0.218 (0.029)
Longest Substring Overlap Length (Memorization proxy)	14.000 (16)	-0.36 (0.0)	0.19 (0.04)	0.1 (0.331)

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

- **ChatGPT-4o has a 16% Null Respone Rate and is more likely to respond with "I don't know" for countries with lower online policy documentation and lower economic development.**

Fig. 2: Geographical Distribution of ChatGPT Accuracy

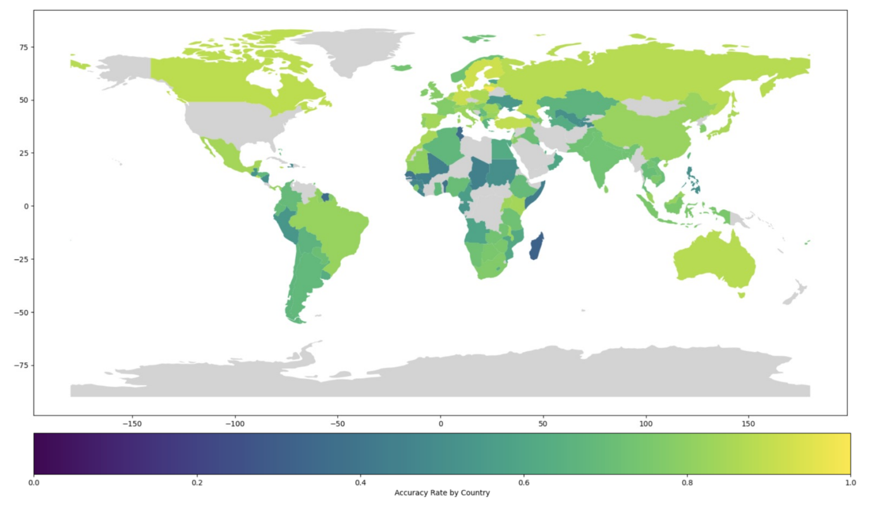
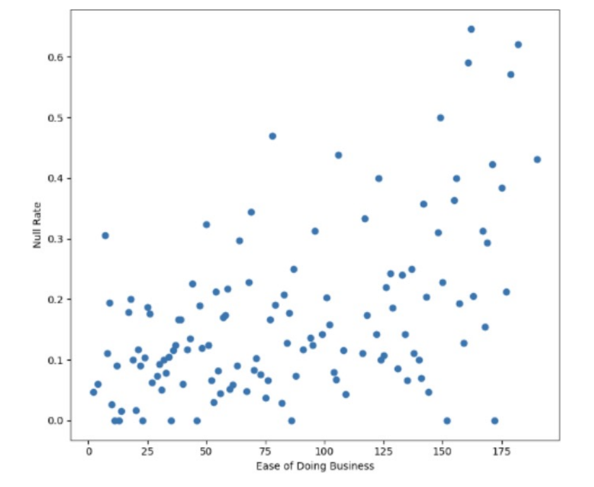
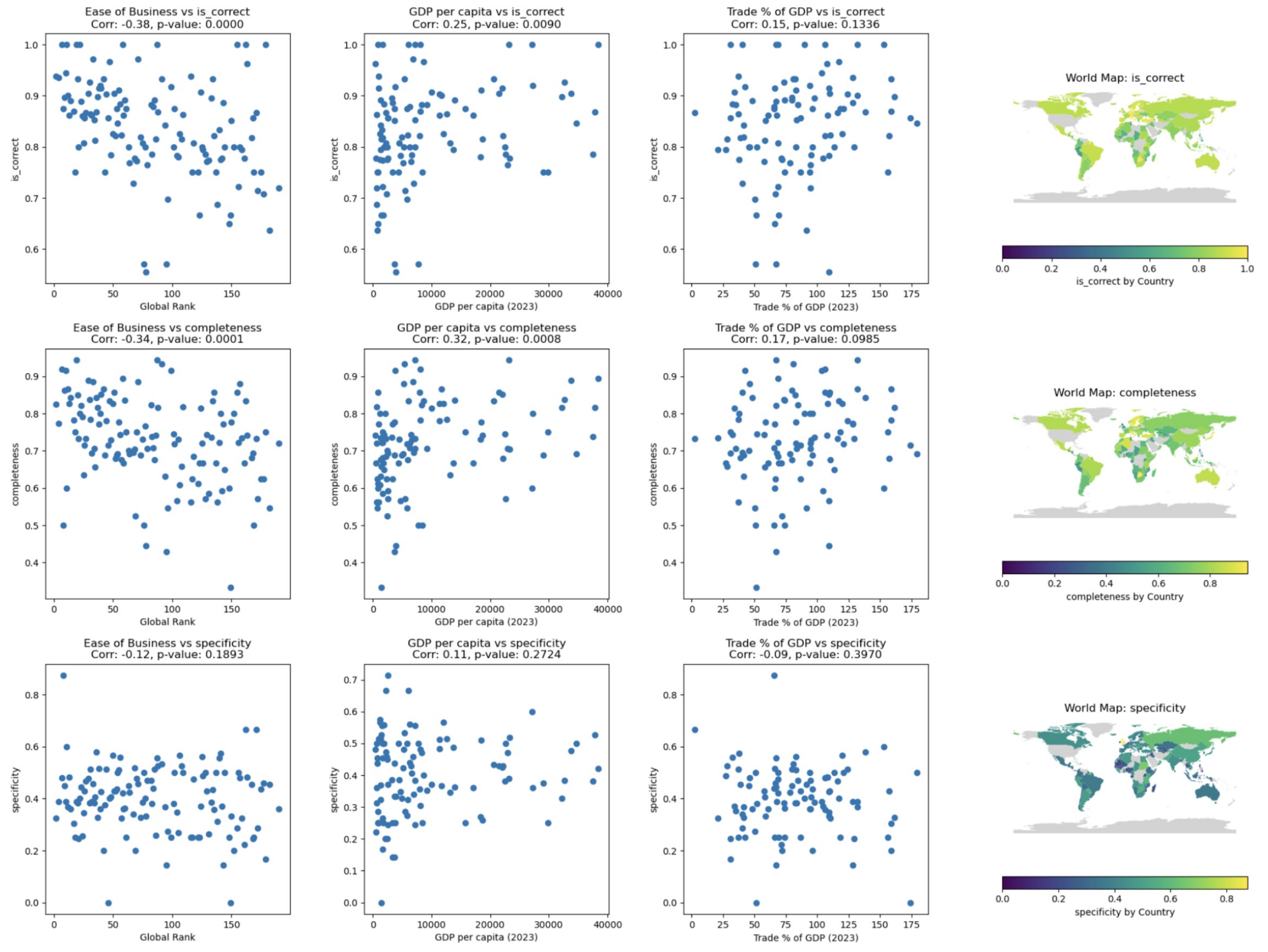


Fig. 3: Cross-Country Null Rate vs. Ease of Doing Business



- Filtering out null responses, leaves 4200 Q&A pairs on which: **ChatGPT-4o achieves 84% accuracy , 74% completeness, 40% specificity**
- There is statistically significant evidence of **ChatGPT-4o performing better for countries with greater ease of business, higher GDP PC and a larger share of trade in their GDP**, with worse performing countries being concentrated in Africa.

Fig. 4: ChatGPT-4o Evaluation - Answer Quality & Bias Assessment



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).