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

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 on this dataset - marking the first systematic evaluation of LLMs for answering questions about international trade - we find that it achieves 84% accuracy. However, we also show that ChatGPT-4o has bias, it performs better for countries with greater ease of business, higher GDP and higher trade shares. The TradeGov dataset thus fills a critical gap in the LLM evaluation literature and paves the way for greater understanding of how LLMs can assist in navigating the complex international trade landscape.

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

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: .

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[1] - constructed by leveraging Retrieval 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 for answering questions related to international trade.

2 Literature Review

This paper situates itself at the intersection of three fields: applying LLMs to law, international trade law, 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). However, there is a notable gap in datasets focused on international trade law, which this paper addresses. Relatedly, the use of LLMs in international trade law has limited literature, with most research focusing on AI regulation from a international trade perspective [3] or the impact of generative AI on international trade negotiations [1]. However, to the best of our knowledge no paper systematically addresses the ability of LLMs to answer international trade law related queries - a gap which this paper addresses through the creation of the TradeGov dataset and evaluating ChatGPT-4o on the same.

2.1 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.

[1]The TradeGov dataset can be found at: https://github.com/amazon-science/tradegov-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.

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 Figure 3), 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 5 for a sample of generated Q&A pairs in the dataset) ([2]

2.2 Dataset 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% 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 (see Appendix Table 3 for more details).

2.3 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[3]: 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 (see Figure 1 and Appendix Figure 4, 5, 6) in the number of Q&A pairs created for a country. The

[2]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.

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

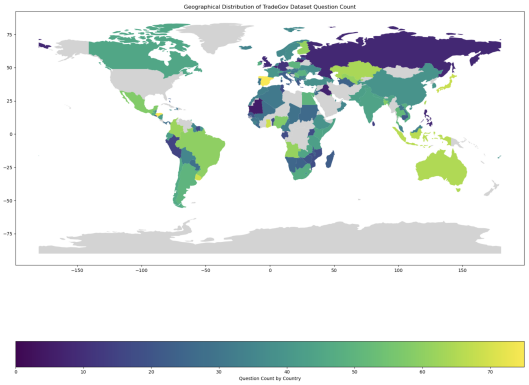

Figure 1: TradeGov Question Count Distribution

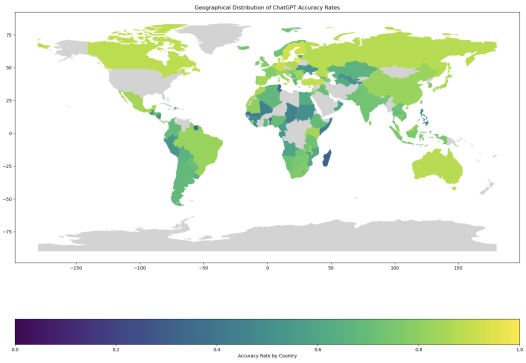

Figure 2: ChatGPT Accuracy Rate Distribution

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.

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 (see Appendix Table 4 for details). [4]. 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.

3 LLM Benchmarking Methodology

Equipped with the TradeGov dataset, we benchmark the performance of ChatGPT-4o for answering international trade policy related questions (see Appendix 1 for the prompt). We evaluate the responses generated across 4 dimensions: 1. **Accuracy**: Does the answer generated by the LLM contain the key facts in the benchmark TradeGov answer? ; 2. **Completeness**: Does the LLM answer contain all the details mentioned in the benchmark TradeGov answer? ; 3. **Specificity**: Does the LLM answer contain unnecessary details? ; 4. **Null response rate**: Is the answer "I don't know"? We use a human-in-the-loop audit to evaluate the LLM generated answer against the answers mentioned in the TradeGov dataset across all four criterion.

4 Results

Examining Table 2, we see that ChatGPT-4o has a Null Rate of 16%. Examining the relationship between null response rates by country and the macro economic indicators mentioned in section 2.3, we see that there is a strong positive correlation (0.5; statistically significant) between null rate and the Ease of Doing Business Index and 2) a weak but negative correlation between null rate and GPD PC (see Table 2). Thus, we conclude that ChatGPT-4o is more likely to respond with "I don't know" for countries with lower online policy documentation and lower economic development.

After filtering out null responses, we are left with 4200 questions. On this subset, ChatGPT-4o achieves an accuracy of 84%, completeness of 74% and specificity of 40%. To determine if these results are on account of ChatGPT-4o parroting text it has memorized from the International Trade Administration website, we split the answer for each query into half and ask ChatGPT-4o to complete the sentence. Then, we use longest sub-string match and sub-string overlap to determine if it outputs text exactly matching the one found on the International Trade Administration website or not. Table 2 and Appendix Fig. 9 show this to not be the case - for majority of the dataset, the longest sub-string match is less than 20 characters.

Given that ChatGPT-4o is likely to have representation biases of the nature mentioned in section 2.3 - we apply the same bias evaluation frame work to analyze the answers generated by ChatGPT-4o. We compute the per country mean values of Null Rate, Accuracy, Completeness, Specificity and Longest sub-string overlap length and measure the correlation of the same with the Ease of Doing Business Index, GDP PC and Trade share % of GDP. We find that 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 (see Figure 2; Table 2; Appendix Figure 8). Particularly, the null rate, accuracy and completeness are statistically significant, strongly negatively correlated with the Ease of Doing Business Index, signaling that the lower the digital documentation for a country, the worse ChatGPT-4o performs. They are also statistically significantly (but weakly) positively correlated with GDP PC, implying higher a country's per capita income, the better ChatGPT-4o performs. Accuracy and completeness are also statistically significantly (but weakly) positively correlated with Trade %age of GDP - indicating that ChatGPT-4o knows more about the trade regulation of countries that trade more. Lastly, answer specificity being weakly positively and negatively correlated with Ease of Doing Business and Trade % of GDP (see Table 2; Appendix Figure 8) respectively potentially indicates that ChatGPT-4o generates answers with more details than needed for countries with lower online documentation and smaller trade shares as it is more uncertain of its knowledge and thus wants to cast a wider net while answering. We leave investigations into these claims to forthcoming versions of the paper.

5 Conclusion

We introduced the TradeGov dataset - the first human-audited, open source dataset for evaluating the performance of LLMs within the domain of on international trade related Q&A. Using this, we were able to show that while current state of the art LLMs can achieve high performance (84% accuracy) in answering factual questions about international trade, this performance is not equitable and is biased in the favour of countries with greater ease of business, higher GDP and higher trade share. 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

[4]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.

questions. To improve the grading of the questions, we will engage lawyers next as opposed the current non-expert auditors. This is especially important because given the low specificity, the subject matter expertise of a lawyer is required to understand if the additional generated facts generated by LLMs - not contained in the TradeGov dataset - are correct or not. This will also aid in establishing a robust human base line for answering international trade related questions, against which the performance of LLMs can be better contextualized.

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: 14 September 2024).

[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: 14 September 2024).

[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. $doi : 10.1007/978 - 3 - 319 - 10061 - 6_14$.

[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

Table 3: Number of Questions per Category

Category	Num. Questions
import-tariffs	768
standards-trade	763
trade-agreements	627
import-requirements-and-documentation	626
trade-barriers	606
customs-regulations	471
temporary-entry	379
licensing-requirements-professional-services	329
labelingmarking-requirements	301
prohibited-restricted-imports	122

.1 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:"""
```

Figure 3: Prompt Template for TradeGov Dataset Q&A Generation

```
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}"""
```

Table 4: Correlation of Metrics with Economic Indicators, with Statistical Significance

Info Type	Metric	Ease of Doing Business	GDP per capita	Trade % of GDP
-customs-regulations	is_correct	-0.277 (0.032)	0.143 (0.322)	0.210 (0.151)
-customs-regulations	completeness_bool	-0.412 (0.001)	0.266 (0.062)	0.287 (0.048)
-customs-regulations	specificity_bool	0.023 (0.863)	0.235 (0.100)	-0.157 (0.285)
-import-requirements-and-documentation	is_correct	-0.348 (0.002)	0.315 (0.011)	0.149 (0.277)
-import-requirements-and-documentation	completeness_bool	-0.306 (0.008)	0.419 (0.001)	0.120 (0.382)
-import-requirements-and-documentation	specificity_bool	0.029 (0.806)	-0.051 (0.689)	0.072 (0.603)
-import-tariffs	is_correct	-0.427 (0.000)	0.241 (0.036)	0.139 (0.243)
-import-tariffs	completeness_bool	-0.434 (0.000)	0.268 (0.019)	0.176 (0.138)
-import-tariffs	specificity_bool	0.186 (0.081)	-0.072 (0.538)	-0.157 (0.285)
-prohibited-restricted-imports	is_correct	-0.286 (0.235)	0.113 (0.701)	0.378 (0.183)
-prohibited-restricted-imports	completeness_bool	-0.283 (0.241)	0.080 (0.785)	0.478 (0.084)
-prohibited-restricted-imports	specificity_bool	-0.063 (0.799)	0.413 (0.142)	-0.602 (0.023)
-standards-trade	is_correct	-0.324 (0.002)	0.139 (0.233)	0.044 (0.709)
-standards-trade	completeness_bool	-0.321 (0.002)	0.145 (0.210)	-0.037 (0.753)
-standards-trade	specificity_bool	0.181 (0.093)	-0.054 (0.644)	-0.084 (0.480)
-temporary-entry	is_correct	-0.368 (0.003)	0.198 (0.151)	0.035 (0.803)
-temporary-entry	completeness_bool	-0.384 (0.002)	0.267 (0.051)	0.133 (0.347)
-temporary-entry	specificity_bool	0.288 (0.022)	-0.200 (0.148)	-0.011 (0.941)
-trade-agreements	is_correct	-0.176 (0.121)	0.105 (0.396)	-0.078 (0.569)
-trade-agreements	completeness_bool	-0.203 (0.073)	0.185 (0.134)	-0.020 (0.883)
-trade-agreements	specificity_bool	-0.004 (0.974)	-0.111 (0.370)	0.027 (0.846)
-trade-barriers	is_correct	-0.455 (0.000)	0.282 (0.019)	0.267 (0.033)
-trade-barriers	completeness_bool	-0.292 (0.009)	0.283 (0.019)	0.109 (0.393)
-trade-barriers	specificity_bool	0.245 (0.029)	-0.220 (0.069)	-0.224 (0.076)

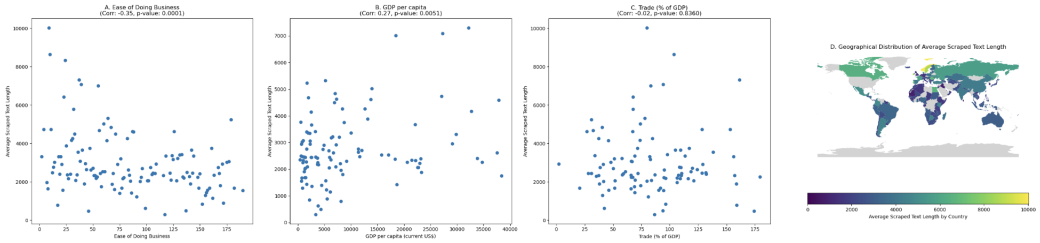

Figure 4: TradeGov Dataset Evaluation: Avg. Country Scraped Text Length

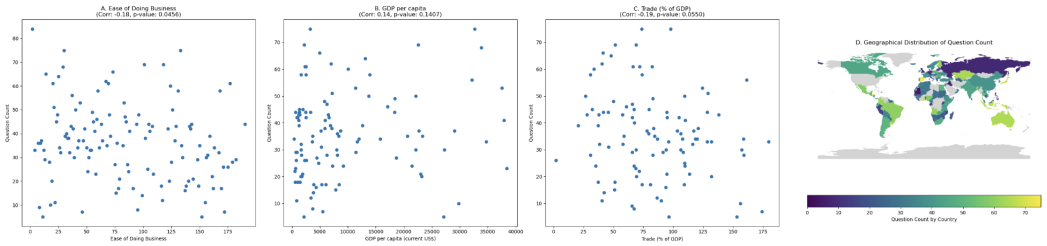

Figure 5: TradeGov Dataset Evaluation: Country Question Count

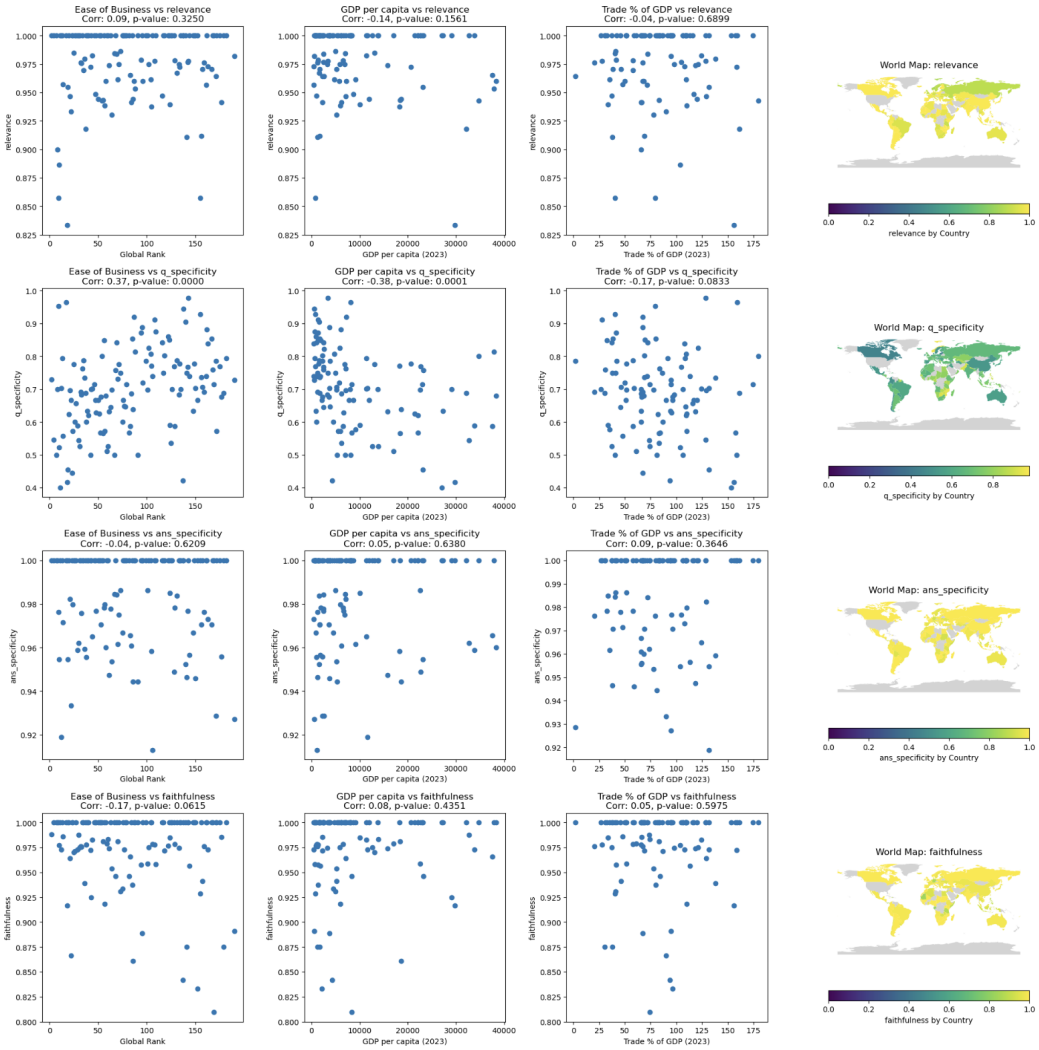

Figure 6: TradeGov Dataset Evaluation: Macroeconomic Analysis

Table 5: 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)

(Continued on next page)

(Continued from previous page)

Questions	Answers
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)
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)

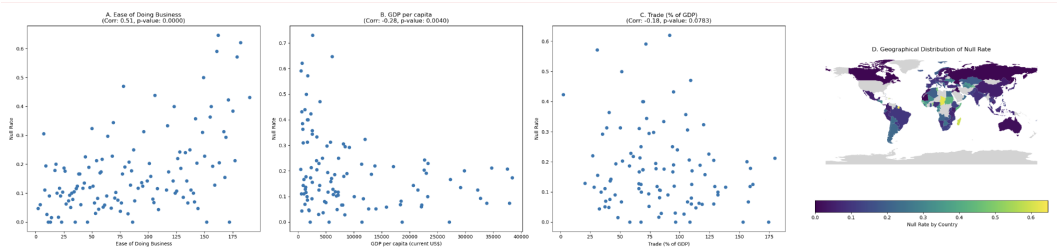

Figure 7: Null Rate Analysis - ChatGPT 4o

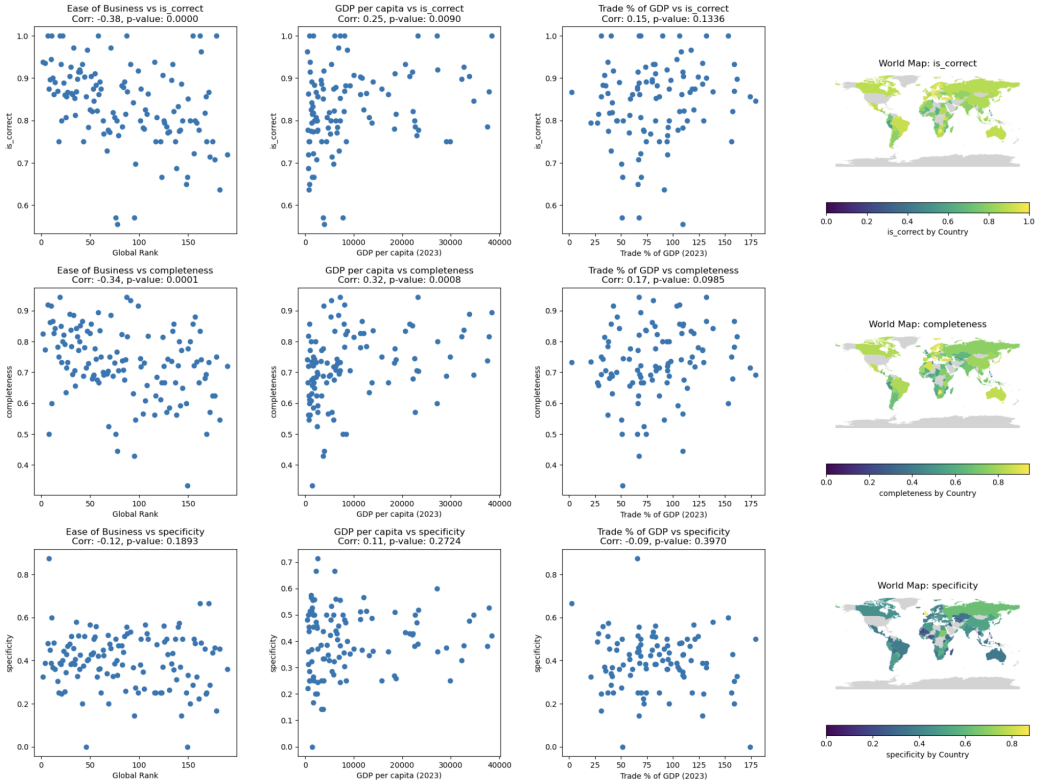

Figure 8: Response Quality Analysis - ChatGPT 4o

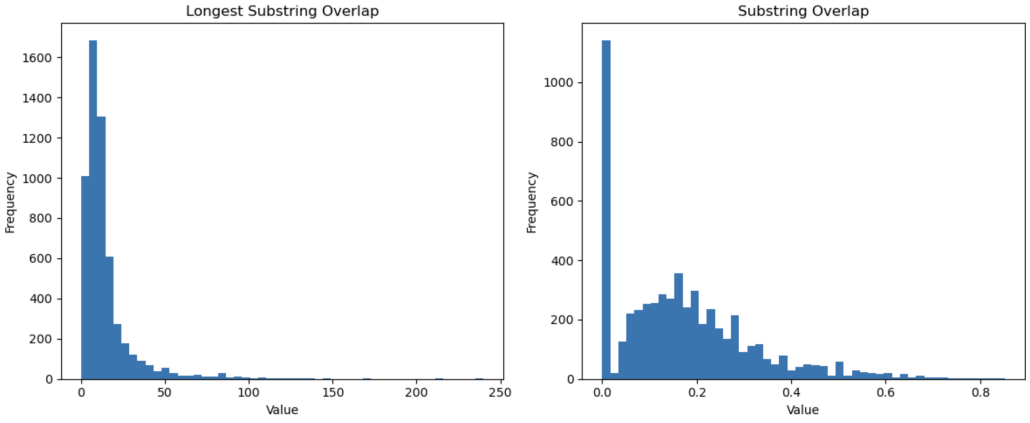

Figure 9: Memorization Quantification - ChatGPT 4o

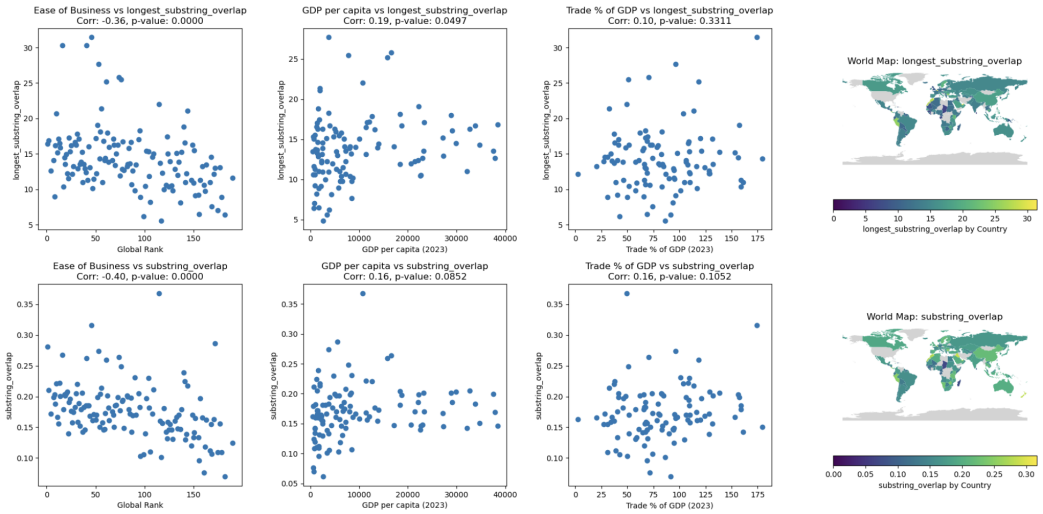

Figure 10: Memorization Quantification Analysis - ChatGPT 4o

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract states that a novel dataset TradeGov has been introduced and used for evaluating ChatGPT on international trade related Q&A and the paper elaborates on that.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: see section 2.1

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Paper has no theoretical results

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the algorithms and datasets required to reproduce the same.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The dataset was cleared by IP review very close to the submission date. If accepted, the paper will be updated with a link to the public dataset repo and relevant replication scripts for the paper based on the same. paper.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Sections 2 and 3 elaborate on this

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Standard errors are reported where applicable

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: sections 2 and 3 elaborate on this

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Reviewed and complied with NeurIPS Code of Ethics

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: see sections 2 and 3

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Artifacts with high risk for misuse are not part of this publication.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All authors have been notified and mentioned in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The dataset was cleared by IP review very close to the submission date. If accepted, the paper will be updated with a link to the public dataset repo and relevant replication scripts for the paper based on the same. paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not have crowdsourcing experiments and research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: The paper does not have study participants

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper evaluates the ChatGPT on a new dataset.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.