Dr Nikolaos Barkas

The Use of Artificial Intelligence in Trade Advisory Services

Brunel University of London
November 2024

This research was conducted as part of the Open Innovation Policy Fellowship in collaboration with Brunel University of London, under the supervision of Professor Tatiana Kalganova.

This research paper does not reflect official UK Govefrnment policy.

Note on the terminology used

The commonly used term for AI responses which are not justified by the training data is "hallucination(s)"; such terminology arguably perpetuates the stigma of mental illness (Østergaard and Nielbo, 2023). The suggestion to use the term "confabulation", (Berk 2024) must also be rejected as related to a neuropsychiatric disorder. Østergaard and Nielbo have suggested (with the assistance of AI) the terms "non sequitur" or "hasty generalisation"; both terms have their roots in Aristotelian logic and are subdivisions of the non-linguistic fallacy/apate (Athanassopoulos and Voscoglou 2020).

Fallacy/apate is associated with deception and indicates a degree of potential intent, which cannot be ascribed to AI. The same level of intentional deception also appears in the term "fabricating information" (Azamfirei 2023), and the problem is that neither makes room for answers that might be factually correct, but are not faithful to the source input.

Instead, a sense of (mental) wandering/going astray (from the truth or the query) might be more appropriate, and either "deviation(s)" or "misleading answer(s)" are better for conveying the unintentional departure from the training/input data. Nevertheless, the literature review forms a necessary part of this research and since the term "hallucination(s)" features prominently in the bibliography, the term was retained throughout this paper for clarity and cohesion but within quotation marks.

Disclaimer

This report includes an AI-generated image on the first and fifth pages. We acknowledge the use of artificial intelligence technology to create this visual content, which serves illustrative and aesthetic purposes only.

Cite the report

Barkas, N. (2024). The use of artificial intelligence in trade advisory services. Brunel University of London. Online resource: https://doi.org/10.17633/rd.brunel.28034627

Contents

Executive Summary

This report explores the integration of Large Language Models (LLMs) like GPT into trade advisory services, focusing on their potential benefits and challenges, particularly Artificial Intelligence(AI)-generated “hallucinations.” “Hallucinations” occur when AI models produce inaccurate or misleading information due to architectural limitations and incomplete training data. This is especially problematic in trade advisory, where errors can lead to significant reputational, financial and legal consequences.

While AI offers great potential for enhancing trade advisory services, success requires careful planning, continuous human oversight, and adherence to ethical standards. These principles are essential for maintaining trust and confidence in public service delivery.

Key Judgements:

- **Information Differentiation:** AI often struggles to distinguish between relevant and irrelevant data, leading to plausible but incorrect “hallucinations.”
- **Inevitability of “Hallucinations”:** These errors are currently statistically inevitable due to the probabilistic nature of transformer models and incomplete data.
- **High-Quality Data:** Diverse, domain-specific datasets are essential to minimise errors.
- **AI in Trade Advisory:** While AI can improve efficiency, “hallucinations” pose significant risks, requiring robust verification.
- **Verification:** A hybrid AI-human model is essential for accurate advice.
- **Data Management:** Effective data management involves continuous updates with high-quality information.
- **Ethical AI:** AI must operate transparently and align with public values and legal standards.

Recommendations:

- **Incremental Increase Strategy:** Start with a limited scope and clearly defined evaluation strategy, gradually expanding the AI’s knowledge base to ensure data quality and system efficiency.
- **Robust Data Management:** Use high-quality, diverse datasets and maintain up-to- date systems to reduce errors. It is necessary to establish a system to manage and verify the temporal relevance of data. This will ensure that the advice provided is based on the most current trade regulations and policies.
- **Using Smaller Language Models**: domain specific models trained on trade-related dataset. This can reduce the chances of irrelevant or incorrect responses.
- **Hybrid AI-Human Model:** Implement a system where AI handles routine queries and human experts oversee complex issues. Additionally, metrics to identify complex issues have to be determined.
- **Technical Solutions:** Use a combination of techniques like RAG and vectors to enhance AI accuracy. Alternatively, incorporate knowledge graphs to improve the AI’s contextual

understanding and interconnections between trade policies. This should reduce errors by providing structured data for decision-making.

- **Training and Oversight:** AI requires comprehensive training for human experts to understand both the technology and the relevant regulations, ensuring they can identify and correct AI-generated errors. Additionally, clear verification protocols should guide experts in consistently reviewing and validating outputs to maintain high standards of accuracy and reliability.
- **Provider Selection:** Set stringent criteria for AI providers, ensuring transparency and accountability with comprehensive records.
- **Ethical Considerations:** Ensure AI systems operate transparently, using Explainable AI (XAI) to align with public values and legal standards. There is a need for the creation of an independent office/ombudsman to oversee and audit AI advisory services.

1. Study aims and approach

This paper explores the integration of Large Language Models (LLMs), such as GPT, into trade advisory services, with a particular focus on AI-generated “hallucinations” -instances where models produce inaccurate or misleading information. These “hallucinations” present significant challenges in high-stakes environments like trade advisory services. The central hypothesis is that these “hallucinations” are currently an inevitable outcome of the architectural limitations of transformer-based models and the incomplete nature of their training data.

In complex and dynamic fields like trade advisory, where regulations are frequently updated, inaccurate AI-generated advice can lead to severe financial and legal repercussions. This paper argues that while AI holds great potential for enhancing efficiency and managing regulatory complexities, its effective deployment is constrained by the challenges posed by “hallucinations.”

To support this hypothesis, the paper will present evidence from experimental interactions with AI tools and a case study of the Finnish Government’s initially limited deployment of AI in customs services. These examples will highlight both the potential benefits of AI and the critical challenges related to “hallucinations.”

The paper first provides a detailed analysis of the nature and causes of AI “hallucinations”; next, it examines the practical implications of these errors in the context of trade advisory; finally, it concludes by advocating for a hybrid AI-human model, emphasising the need for careful planning, continuous human oversight, and adherence to ethical standards to ensure the accuracy and reliability of AI-generated advice.

Through this research, there’s an aspiration to contribute to broader goals by promoting AI diffusion across the economy and fostering capabilities related to the trustworthiness, adoptability, and transparency of AI technologies.

2. Research methodology

The approach proposed by Xiao and Watson (2019) for conducting a literature review involves the following eight steps, which were followed: (1) Problem formulation, (2) development and validation of the review protocol (3) literature search (4) screening for inclusion (5) quality assessment (6) data extraction (7) analysis and synthesis of data, and (8) reporting of results.

The review protocol was developed to focus on publications post-2022, following the introduction of GPT-3.5 and ChatGPT. The protocol prioritised English-language literature from reputable sources, with a specific emphasis on generative question answering while excluding works exclusively focused on abstractive summarization, data-to-text generation, machine translation, and visual-language generation.

The search strategy employed various strings related to:

- Generative AI, LLM, Natural Language Generation (NLG), Machine Learning, Dialogue Generation, Generative Question Answering

- Hallucinations and hallucination mitigation

- Reliability in AI-generated responses
- Fabricating or “making up” information
- Customer service applications of AI
- Personalised customer interactions
- Human-robot collaboration in customer service

Searches were conducted across multiple platforms, including academic databases, Google Scholar, ResearchGate, Academia.com, and AI-assisted research tools like Elicit and Microsoft Copilot.

Studies were screened based on their relevance to LLM “hallucinations”, with a particular focus on the “source of truth” for LLMs and the resulting errors and error tolerance levels. This screening process helped to narrow down the literature to the most pertinent sources for the research question. The quality of included studies was assessed based on their publication in reputable journals and books, as well as their citation in cross-referenced bibliographies. This ensured that the review incorporated high-quality research from authoritative sources in the field.

Data was extracted from the selected studies, focusing on key aspects such as the definition and nature of “hallucinations”, causes and mechanisms, challenges in mitigating “hallucinations” and implications for customer service and professional environments.

3. The scale of the challenge: why trade guidance needs AI

Trade guidance is a necessary service for both businesses and governments, but it is not without its challenges, making it an ideal area for AI support. For businesses, especially small and medium-sized enterprises (SMEs), keeping up with the ever-changing rules and regulations is a significant burden. The complexity and volume of information -from tariff classifications to customs procedures- are overwhelming and can lead to costly delays, compliance issues, and even penalties. These challenges are compounded by the fact that many companies lack the specialised staff needed to navigate these regulations effectively.

On the government side, the task of providing accurate and timely trade guidance is equally demanding. It requires significant time, resources, and expertise to stay on top of the frequent changes in trade laws and regulations. Any misstep in the advice given can result in serious consequences, both for the businesses involved and for the broader trade relationships between countries. This makes it essential for governments to ensure that the guidance they provide is not only accurate but also up-to-date, which is becoming increasingly difficult as global trade becomes more complex.

AI offers a promising solution to these problems. With its ability to process and analyse large datasets quickly, AI can help businesses navigate the complexities of trade regulations more efficiently, reducing the time and effort required to comply with the rules. For governments, AI can enhance the consistency and accuracy of the advice provided, freeing up human resources to focus on more nuanced and complex issues. By automating routine queries and flagging potentially

problematic areas for human review, AI can help ensure that businesses get the right information at the right time, while also helping governments manage their resources more effectively.

The potential benefits of applying AI to trade guidance are substantial. AI can make the process faster, more reliable, and more accessible, benefiting both businesses and government agencies. Given these advantages, it is clear why trade guidance is a strong candidate for AI support.

4. Hallucinations: definition, causes, frequency and solutions

Ji et al. (2023) provide a comprehensive overview of "hallucinations" in Natural Language Generation (NLG), defining them as instances where models generate unfaithful or nonsensical text. They categorise "hallucinations" into two types: intrinsic, where the model manipulates information present in the input, and extrinsic, where the model adds information that is not directly inferable from the input. However, this simple division can be problematic. While all intrinsic hallucinations are incorrect, the same does not apply for extrinsic hallucinations; following Maynez (2020), it is important to recognize that not all extrinsic "hallucinations" (understood as deviations from the training data) are incorrect or non-factual; some with information sourced from the internet may be factually accurate and enhance the reply provided, while others though factually accurate might still considered unfaithful because they are irrelevant or unnecessary in the given context. However, the focus will be on redundant, conflated, contradictory, and nonsensical replies -whether intrinsic or extrinsic- as well as their causes, frequency, and potential solutions. The next step is to examine the causes of "hallucinations" in AI.

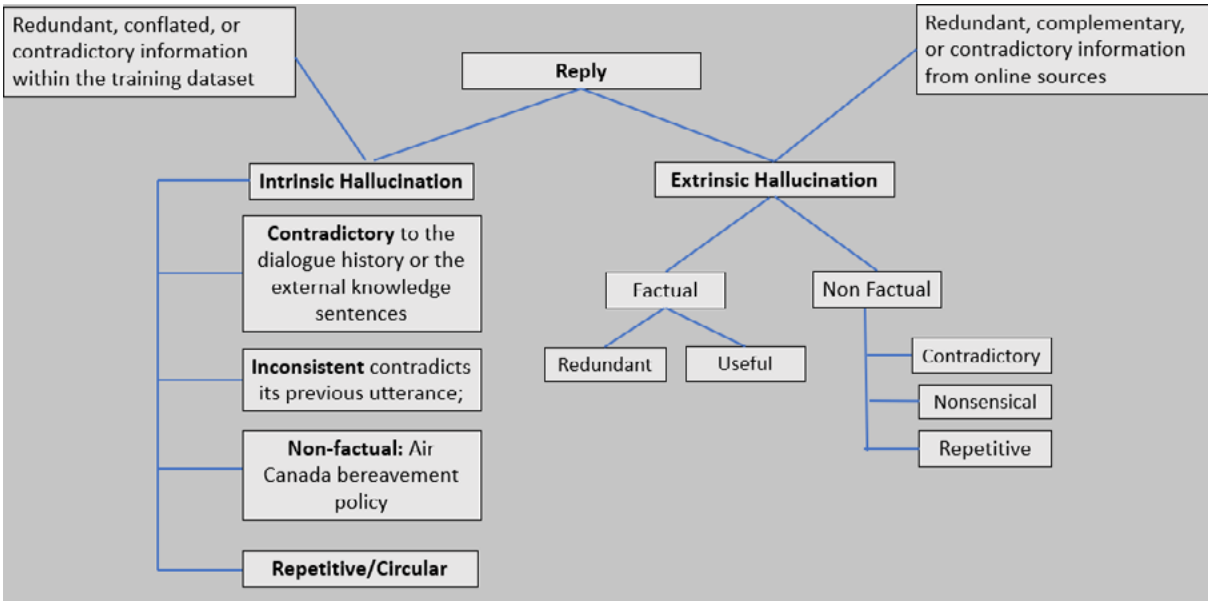

Figure 1. Types of Hallucinations

Causes of "hallucinations"

There are a variety of causes of AI hallucinations, which this section sets out. It is important to understand which of these are more likely to be relevant to an AI trade advisory service to identify potential challenges.

Training data quality: As AI models often do not take into account or cannot differentiate between useful information and irrelevant or incorrect data (noises or artefacts) data quality is paramount (Zhou et al., 2024). Noisy or biased data can lead to "hallucinations" where the AI generates

plausible but incorrect responses. One such “noise“ is the “reference divergence” which is a discrepancy between the training data and the expected output (Zhang et al 2023; Ji et al,2023). Ensuring the use of high-quality, diverse, and domain-specific datasets is paramount for minimising these intrinsic errors. For instance, Maynez et al. (2020) emphasise that factual consistency in training data is essential for improving the reliability of AI outputs.

Pretraining exposure: Pretraining exposure occurs because LLMs generate text by predicting word sequences based on patterns learned during pre-training. This process can introduce biases or assumptions into the model’s responses. For instance, if the training data includes a significant amount of text where goods like electronics are often linked to high import duties, the model might default to suggesting high import duties for electronics, even if the input document doesn’t explicitly state the duty rate. This happens because the model relies on statistical patterns learned during pre-training, which can lead it to ‘bake in’ certain details not present in the input data (Ji et al., 2023).

Contextual misunderstanding: LLMs occasionally struggle to grasp the context of a query, which can lead to the generation of irrelevant or incorrect information (Chan et al.,2022). This issue is particularly evident when LLMs lack the ability to understand and infer the broader context of a document. For example, if an LLM is asked about the export regulations for bolts and nuts but the broader context of the document is focused on mixed loads containing food, the model might generate information related to exporting items like walnuts or coconuts, which is irrelevant to the query. This occurs because the LLM fail to accurately interpret the specific context and nuances of the document, leading to errors in its response (Denning 2022).

Model architecture: The transformer architecture, which underpins models like GPT-3.5 and ChatGPT, is designed to process large volumes of text data and recognize patterns. However, when the model generates responses based on incomplete or ambiguous inputs, it can lead to errors (Huberman, 2023; McIntosh 2023). For example, a query about the import of a cream alcoholic beverage without the correct commodity code might lead the model to assume that “alcohol” refers to wine, resulting in incorrect calculations regarding the import duty. Because the model relies on pattern recognition, it can produce outputs that seem plausible but do not accurately address the specific query. Proof of this issue will be provided through a test conducted with an experimental GPT-type model available to the civil service

Input quality: The way a question is asked can significantly impact the accuracy of the AI’s response. LLMs are not able to abstain from answering when provided with no relevant information (Adlakha et al., 2023). Thus, to obtain a complete and unambiguous answer form GPT might require expert knowledge or a specialist prompt. In a later section, proof of this issue will be provided through a test conducted with an experimental GPT-type model available to the civil service. For example, if someone asks about the import duties for coffee imported from Kenya and provides the wrong commodity code, the AI might give an incomplete or incorrect answer. As with model architecture, a level of expert knowledge is required to obtain the correct answer (Huang et al., 2019). This is also demonstrated through a test with an experimental GPT-type model available to the civil service

It is important to understand the difference between input quality and model architecture; input quality refers to how clearly a question is asked. If the question is vague or unclear, even a well-designed AI might struggle to provide the correct answer. In contrast, model architecture relates

to how the AI is built -how it processes information and identifies patterns. Even with a well-phrased question, the AI might still make mistakes if its underlying architecture isn't designed to handle certain types of input well. For instance, if the AI relies too much on recognizing patterns rather than truly understanding the context, it might misinterpret details, such as mistaking the type of coffee or the relevant import regulations, especially if the input information is incomplete.

Frequency of "hallucinations"

Based on the above, the frequency of "hallucinations" can vary depending on the complexity of the query, the specificity of the training data, and the model's inherent limitations in understanding nuanced or domain-specific information. Studies have shown that even well-trained models like GPT-3.5 can frequently produce "hallucinations", especially in specialised fields. Welleck et al. (2021) note that neural text generation models often encounter challenges in maintaining factual accuracy, leading to frequent "hallucinations". Overall, "hallucinations" occur in 15% to 20% of responses from models like ChatGPT, with higher rates in complex fields (Woodie, 2023). A recent study has confirmed these, and even higher "hallucinatory" scores for most ChatGPT and similar models, however it showed that ChatGPT-4 has a 94% resilience to "hallucinations" (McIntosh, 2023).

Inevitability of "hallucinations"

Due to the probabilistic nature of word prediction in transformer architectures, and the incomplete nature of training data, "hallucinations" are statistically inevitable under the current technology. They occur because the model is optimised to maximise the likelihood of generating text that resembles the training data (Wang and Sennrich, 2020) rather than ensuring factual accuracy or logical coherence. When faced with uncertain or unfamiliar inputs, the model is more prone to generate outputs that are incorrect or nonsensical.

Thus, the reasons that render «hallucinations» currently inevitable are:

- Architectural limitations: current transformer-based architectures inherently produce some "hallucinations" as they predict the next word in a sequence (Zhao et al 2020; Ji et al., 2023).
- Data limitations: comprehensive training data cannot cover every possible query, leading to "hallucinations" with unfamiliar inputs(Zhao et al 2020; Ji et al., 2023).

This does not mean that -in the near future- advances in the architecture of LLMs or combination of solutions could not eliminate hallucinations completely. Nevertheless, at this junction no model can promise 100% accuracy, and this should be a consideration in employing AI in trade advisory services where erroneous responses can cause serious reputational and financial damage.

The complexity of queries as a catalyst for the presence of "hallucinations"

The common denominator regarding the frequency and inevitability of "hallucinations" is the complexity of the queries (McIntosh, 2023); this is important for an AI-powered trade advisory environment with frequent changes in laws and regulations, as well as interdependencies and trade-offs. The table below provides the potential challenges AI would face in a complex and changing trade regulations environment:

Problem	AI Challenge	Hallucination
Frequent Changes in Laws	Constantly changing regulations require frequent updates.	AI might provide outdated compliance advice if unaware of recent amendments.
Different Departments and Agencies	Multiple departments with different rules can confuse the AI.	Offering advice for the import of goods based on HMRC rules only overlooking the need for certificates which are administered by DEFRA and enforced by the Home Office
International Trade Agreements and Sanctions	Tracking and applying specific rules of numerous trade agreements.	Misinterpreting rules of origin and declaring goods duty-free or declaring banned imports as permitted
Customs Procedures and Tariffs	Detailed customs procedures and tariff classifications can be misinterpreted.	Misclassifying a product under the wrong tariff code, leading to incorrect duty calculations
Interdependencies and Trade-Offs	Balancing various policies and understanding trade-offs is complex.	Misadvising on environmental regulations without considering economic benefits or international agreements,
Documentation Requirements	Extensive documentation requirements increase the risk of “hallucinations”.	Advising the use of outdated or incorrect formats for permits

Table 1. Potential hallucinated responses to trade queries

Having explained the nature, causes, and frequency of “hallucinations”, it is necessary to turn to the potential solutions and try to understand what solution(s) might provide a reliable and if possibly uncomplicated fix to the problem. Alaswad and Kalganova (2023) discuss the use of ChatGPT and other LLMs in professional environments, highlighting the need for understanding the context of questions and providing appropriate, relevant answers. They emphasise the importance of sophisticated information retrieval techniques and knowledge representation methods. Nevertheless, apart from the technical/internal solutions there is also need for external intervention by engaging service providers and users alike.

Technical/internal solutions

Improved training data: utilising high-quality, diverse, and domain-specific training datasets can significantly reduce the incidence of “hallucinations”. This involves curating data that is accurate, relevant, and free from biases that could distort the model’s output. See et al. (2019) highlight the importance of clean and well-annotated data in minimising “hallucinations” in generative AI models. Nevertheless, gathering, cleaning, and annotating large datasets is time-consuming and labour-intensive and therefore expensive . Ensuring the data is free from biases and covers the necessary breadth of topics requires significant effort and expertise. A potential solution is to use smaller, domain-specific language models that are trained on targeted datasets. This approach can reduce the size of the models while minimising the chances of generating irrelevant or incorrect responses.

Fact-Checking models: fact-checking models are designed to detect factual inconsistencies in AI-generated content. Kryscinski et al. (2019) introduced models that can identify when a claim made by an AI does not align with known facts. These models work by comparing the AI’s output against a database of verified information. For instance, a fact-checking model might break down a generated summary into individual claims and then assess the probability of each as being correct based on available data. This approach helps identify and correct “hallucinations” before the content is used or published (Kryscinski et al., 2019). Developing and maintaining

a comprehensive database of verified information can be resource-intensive. Fact-checking models also need regular updates to stay current with new information, and integrating these systems with AI models can be technically complex.

Reinforcement learning: it can be employed to improve the informativeness of AI outputs and reduce contradictory information. Some researchers (Li et al 2020; Mesgar et al. 2021) proposed using textual entailment-based rewards, where the AI is rewarded for generating text that logically follows from the given input. This method encourages the model to produce more coherent and factually accurate responses by reinforcing correct information and penalising “hallucinations”. Reinforcement learning involves complex reward structures that require careful design and extensive computational resources. Training these models can be time-consuming, and ensuring the reward system aligns with the desired outcomes demands significant expertise. It is indicative of the efficiency of the method that ChatGPT-4, the least hallucinated of the GPT types (6% hallucinated responses), used it to reduce open and closed domain hallucinations (Ji et al., 2023).

Enhanced model architecture (EMA): developing EMA architectures that incorporate better contextual understanding and fact-checking mechanisms can help mitigate “hallucinations”. Techniques such as reinforcement/supervised learning and fine-tuning have shown promise in improving the accuracy and reliability of AI-generated responses (Alaswad and Kalganova 2023). Holtzman et al. (2020) discuss using unlikelihood training to reduce neural text degeneration and improve output quality. However, such methods are labour-intensive, and for such an enormous, evolving, and disparate body of information as that required to run an import/export advisory service, the resource and time required for such solutions is prohibitive.

Contextual reasoning: enhancing models’ document-level understanding and inference capabilities can help in mitigating “hallucinations”. By improving the AI’s ability to understand and reason about the context of the information it processes, the model can generate more accurate and contextually appropriate responses (Ji et al., 2023). The need for significant computational resources required for this could mean high energy demands which renders this technology as non-environmentally friendly. Ensuring these models generalise well across various contexts is also challenging.

Retrieval-Augmented Generation (RAG): RAG works by augmenting generative models with retrieval mechanisms in order to provide factual grounding. The relevant information is retrieved from a predetermined knowledge base before generating a response. Thus, RAG might pull information from a company’s HR policies database to answer specific employee queries accurately. This approach reduces “hallucinations” by anchoring the AI’s output in verified data (Zhao et al., 2020; Gao et al.2023). However, maintaining an up-to-date and comprehensive knowledge base for retrieval is resource-intensive. A solution to this might be starting with a small area of knowledge i.e. exports of manufactured goods and then expanding the body of knowledge by stages.

Ji et al. (2023) have provided a further list of RAG shortcomings, which RAG shares in common with LLM systems: at the retrieval stage RAG must identify and filter out noise, irrelevant, or fake information; at the augmentation stage, it faces difficulties in integrating diverse, independent, and sometimes conflicting information. Additionally, RAG must avoid generating outputs when there is insufficient information. However, Ji et al. (2023) state that new methods might improve efficiency.

Knowledge graphs: graphs act like smart maps, connecting critical data points such as regulations, tariffs, and products. These graphs could enable the visualisation of relationships between various elements, making it easier to spot opportunities or compliance risks. With AI powering these graphs, systems can autonomously discover new connections, improving the precision of data-driven advice (Kejriwal, 2022; Zhou et al., 2022). However, maintaining these systems can be costly, requiring consistent investments in data management and updates (Jones, 2023). Additionally, there is a risk of data inconsistencies, which may lead to incorrect insights if the input data isn't regularly verified (Kejriwal, 2022; Jones, 2023). Despite these challenges, knowledge graphs greatly enhance decision-making, making complex international trade processes more manageable while ensuring compliance (Zhou et al., 2022; Kejriwal, 2022).

Prompt engineering: clear and specific prompts can guide AI models to generate more accurate responses. By providing detailed instructions and specifying the desired context and details, prompt engineering helps narrow the AI's focus and prevents it from making unwarranted assumptions or fabrications (Ji et al., 2023; Bozkurt, 2024). Effective prompt engineering requires a deep understanding of both the AI model's capabilities and the domain-specific knowledge relevant to the task. Crafting precise prompts can be time-consuming and requires continual refinement to achieve the desired outcomes.

Fine-grained AI feedback: this involves using detailed feedback mechanisms to detect and mitigate "hallucinations". This method allows for more precise identification of inaccuracies and provides targeted corrections, improving the overall accuracy of the AI's responses (Ji et al., 2023). The implementation of fine-grained feedback mechanisms requires significant expertise in error analysis and system design. Collecting and integrating detailed feedback can be labour-intensive and demands sophisticated data management practices.

Conformal prediction techniques: the techniques enable AI models to abstain from answering (e.g., by saying "I don't know") when the responses are likely to be nonsensical or incorrect. This method uses self-consistency and similarity measures to evaluate the confidence of the model's responses. If the confidence is low, the model can choose to abstain from answering, thereby reducing the risk of "hallucinations" (Yadkori et al., 2024) Ensuring the model accurately evaluates its confidence in a wide range of scenarios can be challenging.

Explainable AI (XAI): integrating XAI models enhances transparency and trust. XAI techniques provide insights into how AI models make decisions, allowing users to understand and verify the reasoning behind the AI's outputs. This transparency helps in identifying and correcting potential "hallucinations", ensuring the reliability of the AI system (Zodage et al., 2024). Developing XAI systems that provide meaningful and accurate explanations without oversimplifying complex models is difficult. Ensuring these explanations are accessible and useful to non-expert users adds an additional layer of complexity.

External solutions

User training: Educating users on how to phrase questions clearly and specifically can reduce the likelihood of misunderstandings and incorrect answers. This is particularly important in professional and specialised environments. Huang et al. (2019) emphasise the role of user input quality in achieving consistent and accurate AI responses. Users will not only need ongoing support to adapt to best practices for interacting with AI systems, but also have a good grounding on trade rules and regulations.

Hybrid AI-Human systems: Maynez et al. (2020) advocate for human-in-the-loop systems to enhance the factual accuracy of AI-generated content. Human reviewers can identify and correct inaccuracies that AI models might miss. This process involves regular review and correction of AI-generated content, which improves the AI's performance over time. For example, a hybrid human-in-the-loop framework for fact-checking combines automatic AI methods, crowdsourcing, and expert review to verify the veracity of information at scale (Ji et al., 2023; Aditya,2024).

Disclaimers: not a remedy in itself, but in the rapidly evolving field of AI-powered trade advisory services, the inclusion of a clear and comprehensive disclaimer serves to remind users that the AI's outputs are not guaranteed to be accurate and that human judgement should prevail in decision-making (Metzger et al., 2024)

Providing trade advice carries significant potential liability, especially when decisions based on this advice can result in financial loss or regulatory breaches. A disclaimer helps manage this risk by clearly stating that the service is for informational purposes only and that the responsibility for verifying and acting on the information rests with the user. This can protect the service provider from legal claims and ensure that users are aware of the need to exercise caution.

Overall, successfully mitigating AI "hallucinations" involves several strategies. The best approach often involves a combination of intrinsic and extrinsic methods. For instance, combining RAG with human oversight (human-in-the-loop) and improved training data can significantly reduce "hallucinations" by grounding AI responses in verified information and allowing human reviewers to correct any remaining inaccuracies. Additionally, using conformal prediction techniques can further enhance the reliability of the AI by enabling it to abstain from uncertain responses. Finally, a disclaimer that the service is only advisory and that ultimate responsibility lies with the user is a necessary safeguard. This multi-faceted approach leverages the strengths of each method to provide a robust solution to the problem of AI "hallucinations".

5. Employing AI in Trade Advisory Services

The integration of AI into trade advisory services offers substantial potential to improve efficiency and manage the complexities of constantly evolving trade regulations. However, this potential is tempered by the significant challenge posed by AI "hallucinations". information. Such errors can have severe financial and legal consequences for businesses, underscoring the need for robust verification mechanisms and a hybrid AI-human model to ensure the accuracy and reliability of advice (Ji et al., 2023).

As noted above, the first step for reducing "hallucinations" is the training data: utilising high-quality, diverse, and domain-specific training datasets can significantly reduce the incidence of "hallucinations". Data management is not limited to the initial collection and curation of data; it requires maintaining up-to-date systems by ensuring they are fed with high-quality and current data. The whole data management process is labour intensive, even more so in a complex and changing environment as trade.

Arguably starting small, with a limited area of trade and then expanding is one of the best strategies to ensure the quality of the data, the efficiency of the technology and the effectiveness of any fine tuning. This is the approach Finnish Customs has adopted. Finnish Customs implemented an off-

the-shelf chatbot to manage customs and duty fees introduced on January 1, 2021, for parcels across Europe. This required creating 60 interaction scenarios, and the chatbot served 32,000 customers in its first year with a 75% success rate. Approximately 8% of the conversations had to be transferred to their chat service to interact with Finnish Customs agents.

The Finnish experience provides a measure of the limitations of such an endeavour, the work required to make it operational, and the potential level of success. Scaling this to more complex and larger datasets, such as those required for comprehensive trade advisory services, would demand substantial human and computational resources even if the system used was off-the-shelf. Currently, the Finish Customs tool handles 160 scenarios and has around 90% rate of success, which demonstrates the incremental increase in data and is an excellent approach to countering “hallucinations”.

Private companies might offer an end-to-end service and promise to undertake the collection and curation of data for the public sector; this is a temptation that should be avoided. Firstly, it might lead to vendor lock-in, making future changes costly and difficult. Secondly, it is unlikely that external providers would have a high level of domain expertise, and to ensure high-quality, accurate information, the organisation might have to provide and curate the initial data, adding to the overall cost. Thirdly, since it is such a novel area, caution is required around setting KPIs, accuracy/success forecasting, the requirements in human support, and issues such as liability for errors and the traceability and evaluation of the outputs. More critically, there are serious concerns about unethical practices in data input and labelling processes involving exploitative practices in low-wage regions (Muldonn et al. 2023, Nwachukwu et al, 2023). Therefore, to ensure ethical standards and specialised service quality, the public sector should maintain control over data collection and curation

A recent case highlights the potential pitfalls of AI-generated ‘hallucinations”: Air Canada was ordered to pay compensation after their chatbot gave a customer inaccurate and fabricated information regarding their bereavement policy, misleading him into buying a full-price ticket. Air Canada attempted to distance itself from the error by claiming that the bot was responsible for its own actions.

Identifying “hallucinations” to customs procedures queries in a GPT- type model

The Borders Innovation Team conducted experimental queries about customs procedures on a generic ChatGPT-based model and identified serious issues with “hallucinations”. The version used was up-to-date until November 2021 and had no access to the internet. However, this lack of internet access did not hinder the AI’s ability to find the correct information, indicating that “hallucinations” occur even when the model is not dependent on outdated sources.

Query	Specialist Prompt (Commodity Code)	Hallucinations	Type
Import duty for 17% Abv. Alc. Cream Beverage from Spain to the UK	Y	N	N/A
Import duty for 17% Abv. Alc. Cream Beverage from Spain to the UK	N	Y	Intrinsic
Formalities for the transport to the EU of personal property from the UK	Y	Y	Repetition/ Conflicting info
Import duty for coffee from Kenya to the UK	N	Y	Intrinsic

Table 2. AI responses to trade queries from a generic GPT-type system

The first query regarding the amount of UK excise duty for a 70 cl bottle of creme alcoholic beverage 17% Abv imported from Spain was answered correctly, and the steps for confirming the answer were provided as well. However, when the question was resubmitted and the commodity code was omitted, AI could not discern between the rate for wine and the rate for "spirits other than UK-produced whisky," resulting in a £1.00 per litre miscalculation. If the second reply was followed, a business importing a 10,000 bottles consignment would need an additional £7,000 than what was quoted by AI.

The next question was about transporting to the EU cars, bicycles, motorbikes, boats, etc., within, upon, behind, or adjacent to a main vehicle. Since the vehicle carried over the border is not the main means of transport, it is classified as "goods" or "personal property" and therefore the question is whether to use an ATA Carnet. ATA Carnets allow users the temporary export of commercial samples, trade fair or exhibition goods and professional equipment (including music and sport equipment) to countries that are part of the ATA Carnet system. Another possibility was using a CPD Carnet, (Carnet de Passages en Douane) which allows for the temporary export of a UK registered vehicle into certain countries. CPDs are not used for vehicles visiting the EU.

AI was correct that since the destination was the EU and the intended use was private, neither an ATA nor a CPD carnet were necessary. AI also provided the standard guidance the government would give to business: "It is generally recommended to contact the customs authorities of the country you will be entering to inquire about the specific requirements for temporarily importing a vehicle for personal use."

On the other hand, AI could not get around the issue of an oral declaration being the solution to customs procedures provided it is accompanied by an inventory. Moreover, AI needed some prompting regarding the carnets and did not arrive at this answer by itself. Finally, AI replies became confused, stating that you needed some form of declaration, and then ultimately saying that this might be a type of carnet in a subsequent response. So, for those without a relatively good knowledge of customs, this ended up being quite a confusing and ultimately self-contradictory response.

The final question concerned the import of coffee beans from Kenya. An erroneous commodity code was inputted to the AI. The AI did not recognize the error and proceeded to fabricate the whole answer, assigning a 7.5% import duty to the coffee beans from Kenya. If AI had processed the information contextually, then it would have concluded that no coffee imports pay duty, but it didn't. Hence, in the absence of any concrete info, it made the answer up. This was not an issue of outdated information, as the UK has signed a duty-free agreement with Kenya since March 2021; apart from that, all coffee beans imports (unless from Russia or Belarus) have no import duty, which predates 2021 (last update of the ChatGPT used).

These examples highlight the critical challenges and limitations associated with the use of AI in complex and regulated environments such as customs and trade advisory services. While AI has the potential to streamline processes and improve efficiency, the risk of errors and "hallucinations" underscores the necessity of robust verification systems and human oversight. Thus, the second step for minimising "hallucinations" is to employ a technical solution or better a combination of solutions both internal and external as exemplified earlier in this essay. Any AI systems not continuously updated with real-time data risk providing outdated or incorrect advice, which could lead to compliance failures or financial losses (Kryscinski et al., 2019). When tariff classifications undergo changes,

an AI model lacking the capability to dynamically update in response to new information would neglect these changes, leading to mistakes such as incorrect duty calculations or misclassification of goods. RAG, which grounds responses in the most up-to-date and verified information, seems a good solution, despite being resource-intensive (Welleck et al., 2021).

Due to the complexity and specificity inherent in trade-related inquiries, human intervention is indispensable to ensure the precision of generated responses. A hybrid AI-human model capitalises on the strengths of both, with AI efficiently handling routine queries and human experts focusing on the more intricate and nuanced issues that require deep regulatory understanding (Huang et al., 2019). Misleading responses are thoroughly vetted and corrected by knowledgeable professionals before they reach the end-users (See et al., 2019). This division of labour not only streamlines service efficiency but also elevates the accuracy of the advisory outputs.

The Human-AI hybrid approach has been used in the development by i.AI and the Citizens Advice Bureau of Caddy, an AI-powered assistant which acts as a copilot for customer service agents, empowering them to provide high-quality, actionable advice quickly and securely. Caddy employs a "human in the loop" validation system to mitigate risk, ensuring advice accuracy and reliability.

Caddy employs embeddings, which are numerical representations of words, to understand the context of a query better. When a question is asked, Caddy converts it into this numerical form and conducts a vector search within its knowledge base to find the most relevant information. This method requires a very rigorous collection and curation of data. Moreover, if this is the only grounding technique, there is a risk that vector search may not always provide the contextual depth needed for complex queries without additional processing. Thus, a hybrid RAG (vector+full search) would benefit accuracy.

Caddy could be effectively adapted for use in a trade advisory service by leveraging its capabilities in information retrieval, customer support, document management, and compliance monitoring.

Comparing solutions for AI accuracy in trade advisory

When applying AI to trade advisory services, choosing the right technology is essential due to the complexity and constant changes in trade regulations. Two prominent approaches are Retrieval-Augmented Generation (RAG) combined with vector embeddings and knowledge graphs. Both offer unique benefits, but they also come with limitations that need to be carefully weighed, especially in a domain where accuracy, context, and transparency are critical for making informed decisions.

When evaluating AI solutions for trade advisory services, two leading approaches are Retrieval-Augmented Generation (RAG) with vector embeddings and knowledge graphs. Each of these technologies has distinct strengths and weaknesses that become particularly significant when applied to the complex, dynamic world of trade regulation. The choice between these two depends not only on their technical merits—such as accuracy and relational understanding—but also on practical considerations like cost, time to implement, and the level of expertise required.

Vectors: the foundation with limitations

Vectors form the backbone of many language models, such as GPT and BERT. These models map words or phrases as numbers in a high-dimensional space, enabling the AI to capture the

relationships between terms based on context. In practice, vectors act as a map, grouping related concepts together and allowing AI to make sense of large amounts of data quickly. This has revolutionised natural language processing and has been particularly useful in sectors like trade advisory, where the AI needs to manage vast datasets of regulations and tariffs (Mikolov et al., 2013).

However, vectors alone have several critical limitations, especially when it comes to accuracy and explainability. A key issue is that vectors identify relationships based on statistical patterns without providing a clear, human-readable explanation of why certain terms are connected. For example, a trade advisory system might recognize that two regulations are related, but it won't explain why or how these rules interact. This lack of transparency is particularly problematic in the regulatory environment, where stakeholders need to trust and understand the reasoning behind the AI's advice (Bender et al., 2021). Additionally, vectors can struggle with contextual understanding—misinterpreting ambiguous terms or failing to account for nuances in language. This is risky in trade advisory, where such misunderstandings could lead to incorrect regulatory advice (Adadi & Berrada, 2018).

Another challenge arises from the scaling complexity inherent in vector-based systems. As the dataset grows, with more regulations and rules constantly being added, the vector space becomes more complex and harder to manage. This can undermine both the accuracy and efficiency of the system over time, especially in fields like trade advisory where regulations are frequently updated (Mikolov et al., 2013). Furthermore, vectors are limited in their ability to represent structured relationships, which are necessary for understanding the hierarchy and interaction between trade rules and agreements.

How RAG can enhance vectors

Retrieval-Augmented Generation (RAG) improves upon these limitations by combining vector-based language models with external knowledge retrieval. In trade advisory, RAG can enhance vectors by pulling up-to-date, context-specific information from external databases, improving the AI's ability to deliver more accurate and context-aware responses. For instance, when asked about recent changes in export tariffs, RAG retrieves the most relevant and current data from its knowledge base to inform the AI's response, reducing the likelihood of errors related to outdated or incomplete information (Lewis et al., 2020).

RAG also addresses the issue of hallucinations, where AI generates incorrect or irrelevant responses. By grounding responses in factual data, RAG improves reliability. However, it does not fully resolve the explainability and relational understanding gaps of vector-based systems. While RAG helps pull accurate data, it still struggles to model complex relationships between regulations, which is a significant drawback in trade advisory services that often require multi-layered reasoning (Ji et al., 2023).

Knowledge graphs: structured, relational understanding

Knowledge graphs, by contrast, offer a far more structured and explicit way of representing entities and their relationships. This is particularly advantageous in trade advisory, where regulations, tariffs, goods, and trade agreements often have intricate, hierarchical relationships. Knowledge graphs enable the AI to understand not only what the entities are, but how they interact—essential

for delivering precise, context-aware advice in complex regulatory environments (Hogan et al., 2021).

For example, a knowledge graph can represent how a free trade agreement impacts specific tariff rules and how those changes influence other regulations downstream. This relational mapping allows the AI to provide deeper, more nuanced advice compared to vector-based systems. Moreover, knowledge graphs excel in explainability: they offer clear reasoning paths that users can follow to understand how decisions were made, which is invaluable in environments where trust in AI's output is critical (Paulheim, 2017).

The downside of knowledge graphs, however, lies in their high implementation and maintenance costs. Building a knowledge graph requires significant time, resources, and expertise. Domain experts need to manually map out entities and relationships, and as trade regulations evolve, maintaining the graph requires continuous updates, making it resource-intensive (Ehrlinger & Wöß, 2016). Furthermore, the upfront time and costs involved in deploying a knowledge graph system are much higher compared to RAG + vectors, which can be implemented more quickly and at a lower cost.

Cost, time, and expertise: key practical considerations

From a cost and time perspective, RAG + vectors is the more cost-effective and quicker-to-deploy solution. Since it leverages existing models and databases, RAG + vectors requires less manual input and can be implemented relatively quickly. This makes it ideal for situations where trade advisory services need a flexible system that can rapidly adapt to new regulations or handle large-scale queries efficiently (Lewis et al., 2020). Furthermore, less technical expertise is required to implement RAG, making it suitable for environments with limited access to AI experts.

On the other hand, while knowledge graphs are more resource-intensive in terms of both cost and time, they provide a superior long-term solution when accuracy, explainability, and relational understanding are the top priorities. Knowledge graphs require significant expertise to design and maintain, but once established, they offer far better performance for complex, high-stakes trade advisory tasks. They are particularly valuable for systems that need to handle detailed regulatory relationships and provide transparent, accountable advice (Paulheim, 2017).

Balancing technical merits and practical constraints

In the context of trade advisory services, the choice between RAG + vectors and knowledge graphs ultimately depends on the specific needs of the system. RAG + vectors offers a scalable, cost-effective solution for handling routine inquiries and rapidly changing information, making it suitable for environments with tight budgets and limited time for deployment. However, it lacks the relational depth and transparency required for more complex advisory tasks.

On the other hand, knowledge graphs offer a more robust and explainable framework for handling the intricate, hierarchical relationships that are often involved in trade regulations. Though they require more time, resources, and expertise to build and maintain, they provide the long-term precision and contextual understanding that is critical for high-stakes advisory environments. In situations where accuracy, trust, and explainability are paramount, knowledge graphs are the superior choice, despite their higher initial cost and longer implementation timeline.

Both RAG with vectors and knowledge graphs help tackle issues like limited context windows, query dependence, and clarification. RAG extends the AI's reach by retrieving relevant information from external sources, ensuring more accurate responses even when the query goes beyond what the model can immediately handle. This makes the AI's answers more context-dependent and tailored to the query. Knowledge graphs, on the other hand, provide structured, interconnected data, helping the AI understand complex relationships, like those between trade policies, and reducing errors. Either of these techniques enhance the AI's accuracy by clarifying and expanding context dynamically.

In addition to choosing between RAG + vectors and knowledge graphs, integrating Explainable AI (XAI) techniques can significantly enhance both systems, particularly in trade advisory, where trust and transparency are critical. XAI methods can be applied to both approaches to provide clearer reasoning paths and improve user understanding of AI decisions. In RAG + vectors, XAI could help explain how retrieved data was selected, mitigating concerns about black-box decision-making. For knowledge graphs, XAI can further clarify the relationships between entities, enhancing transparency in complex regulatory environments (Arrieta et al., 2020). By incorporating XAI, both systems can offer greater accountability and trust, making them more reliable in high-stakes trade advisory tasks.

Implementation strategy

For the hybrid model to achieve its full potential, it is imperative that human experts receive thorough training in the use and oversight of AI tools. This training should cover not only the AI's technical capabilities and limitations, but also equip experts with the skills necessary to identify and correct errors in AI outputs. Comprehensive training programs are vital to maximising the effectiveness of hybrid systems (Maynez et al., 2020). These programs should extend beyond technical proficiency to include a deep understanding of trade regulations, enabling experts to provide contextually relevant oversight.

Establishing clear, structured verification protocols is critical for maintaining high standards of accuracy and reliability in AI-generated advice, especially when adapting AI systems to the intricacies of trade advisory services, where the stakes for accuracy are particularly high. These protocols should meticulously detail the procedures that human experts must follow when reviewing AI responses, ensuring consistency and thoroughness in the verification process (Huang et al., 2019). By adhering to these protocols, human experts can systematically identify and rectify any errors in AI outputs, thereby mitigating the risk of disseminating inaccurate advice.

In addition, implementing a comprehensive data governance framework is essential to manage and verify the temporal relevance of the datasets that AI relies on. Tools like Delta Lake or DVC can track version histories, ensuring that historical and current data are easily accessible and verifiable (Atlan, 2023; Ji, 2023). Regular automated data refresh processes using Apache Airflow help ensure that datasets remain up-to-date, particularly in the fast-changing world of trade regulations (IBM, 2022; KPMG, 2022). Moreover, incorporating data validation checks and audit trails will enable timely detection and correction of errors, allowing the system to remain accurate and reliable. This approach reduces the risk of outdated information influencing AI outputs, thereby improving decision-making in trade advisory (IBM, 2022; KPMG, 2022).

A robust feedback mechanism is essential for the continuous refinement of the AI model, which involves regularly updating training data based on human input and fine-tuning the model's parameters to enhance its performance over time. The iterative feedback loop is particularly relevant in this context, as it enables the AI model to adapt to new regulatory changes and evolving industry practices (See et al., 2019). By integrating insights from human experts, the AI system becomes increasingly responsive to the specific needs of trade advisory services, ensuring that its outputs remain both accurate and pertinent.

The challenges for policy

Navigating the complexities of trade regulations and policies requires an advisory service that can handle a significant volume of inquiries and stay adaptive to ongoing changes. Integrating AI into this process has the potential to greatly enhance efficiency, particularly when dealing with the intricate trade-offs inherent in different policies. However, the complexity and ever-changing nature of trade regulations also heighten the risk of AI producing errors or '"hallucinations".' Examples like incorrect tariff classifications or miscalculated duties highlight the urgent need for robust verification systems to mitigate such mistakes (Maynez et al., 2020).

To effectively meet these challenges, it's essential for the government to establish a stringent selection process for both internal and external AI providers, prioritising transparency, accuracy, and accountability. Providers must have a demonstrated history of delivering reliable AI solutions, especially in areas requiring precise regulatory compliance, such as trade advisory services. This can be supported by implementing a certification process for AI providers, ensuring they adhere to specific technical and ethical standards before being contracted. Furthermore, providers should be required to keep comprehensive, auditable records of their AI systems' training data and decision-making processes. Strict penalties -within the existing framework of government procurement- including financial sanctions and potential contract termination, should be enforced for providers who fail to meet the defined accuracy and reliability standards.

Maintaining robust oversight is paramount to continuously monitor AI systems' performance in trade advisory services. This oversight should include regular audits, real-time error monitoring, and a transparent reporting process for issues arising from AI-generated advice. Establishing an independent oversight body empowered to conduct these audits and reviews, is essential. This body should have the authority to intervene and make necessary adjustments or suspend AI operations when needed. Providers found non-compliant during audits should face immediate penalties, such as mandatory retraining of AI models at their own cost or suspension of services until issues are resolved. The existence of an independent body supervising AI advisory operations would also resolve the problem of any potential conflict of interest where the same civil service entity is both providing guidance and assessing compliance with regulations

The government should invest in training programs designed to enhance AI literacy among civil servants, particularly those in regulatory and oversight roles. A tailored curriculum that ranges from basic AI literacy courses to advanced training for AI management can effectively address the varying levels of expertise within the workforce. By offering practical workshops, online courses, and certification programs, the government can provide accessible learning opportunities that are both flexible and comprehensive.

Additionally, expanding these educational efforts to include public awareness campaigns about AI's role in government can promote transparency and build public trust. To further encourage participation, the government could introduce incentives such as certification bonuses or career advancement opportunities tied to AI competencies.

Democratising AI training is another essential step that can help mitigate errors and reduce biases in decision-making processes (Dessimoz and Thomas, 2024). By equipping a broader range of individuals with AI knowledge, the government can enhance the ethical use and effectiveness of AI tools, leading to more informed procurement decisions and improved public service delivery. Research indicates that incorporating fairness and bias analysis in AI applications significantly improves outcomes by addressing unintended biases (Chen, Wu, & Wang,2023). By implementing these strategies, the government can cultivate a well-prepared and inclusive workforce, ready to meet the challenges of AI integration in public services.

Placing a strong emphasis on the ethical use of AI in all advisory services is crucial for the government, ensuring that AI systems operate transparently and are aligned with public values and legal standards. Mandating the use of Explainable AI (XAI) technologies will allow users and overseers to understand the processes behind AI-generated advice. Developing and enforcing ethical guidelines for AI use in public services -including data privacy, bias prevention, and public accountability-is essential. Establishing channels for public and internal feedback on AI performance, along with processes to address concerns and promptly implement necessary adjustments, will further enhance trust and reliability. While fostering AI innovation is important, the government must ensure that any new technologies introduced into public services, particularly in advisory roles, meet stringent standards of reliability and fairness.

Conclusion

Bringing AI into trade advisory services offers a great way to boost efficiency and handle the ever-changing complexities of trade regulations. But with this potential comes the challenge of "hallucinations" -when AI gets things wrong- which could have serious financial or legal consequences. To keep these risks in check, it is crucial to pair AI with human expertise, creating a system that leverages the strengths of both.

Key Recommendations

1. **Start small and manage data wisely:** Governments should begin by using AI on a smaller scale, focusing on specific areas, and then gradually expand as the technology proves its worth. Smaller language models with more manageable datasets can greatly assist in this direction.

2. **Keeping the data accurate and up-to-date is essential to minimise mistakes and ensure that AI provides reliable advice.** Implementing a comprehensive data governance framework is essential to manage and verify the temporal relevance of the datasets that AI relies on. Tools like Delta Lake or DVC can track version histories,

ensuring that historical and current data are easily accessible and verifiable (Atlan, 2023; Ji, 2023).

3. **Train the policy experts/AI users and set clear guidelines:** it is important to train human experts not just on the technical side of AI but also on the trade regulations themselves. Clear guidelines for how to verify AI-generated advice will help ensure that the information provided is both accurate and dependable.

4. **Prioritise ethics and transparency:** AI in public services needs to operate under strict ethical standards. Using technology like Explainable AI (XAI) can help make AI's decision-making process clearer and easier to trust. Also, having an independent body to oversee and audit these AI systems is vital to maintaining legal and ethical integrity.

Forward look

As AI technology continues to improve, we can expect more sophisticated systems that are better at avoiding mistakes and providing contextually accurate advice. But this won't happen on its own—it will require ongoing investment in research, collaboration between public and private sectors, and a commitment to ethical practices.

In the bigger picture, the way we integrate AI into trade advisory could set an example for other areas of public service. If done right, it could show how AI can improve how governments serve the public while still keeping everything transparent and trustworthy. By sticking to these principles and staying alert to the challenges, governments can lead the charge in using AI responsibly, ultimately benefiting everyone.

Bibliography

- Adlakha, V., BehnamGhader, P., Lu, X. H., Meade, N., & Reddy, S. (2023). *Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering*. arXiv preprint arXiv:2307.16877. https://doi.org/10.48550/arXiv.2307.16877

- Alaswad, M., & Kalganova, T. (2023). Using ChatGPT and other LLMs in Professional Environments. *Information Sciences Letters*, 12(9), 2097–2108. https://doi.org/10.18576/isl/120916

- Alkhaqani, A. L. (2023). Potential Benefits and Challenges of ChatGPT in Future Nursing Education. *Maaen Journal for Medical Sciences*, 2(2). https://doi.org/10.55810/2789-9128.1020

- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*. arXiv preprint arXiv:1606.06565. https://arxiv.org/abs/1606.06565

- Arivazhagan, N., Bapna, A., Firat, O., Lepikhin, D., Johnson, M., Krikun, M., Chen, M. X., Cao, Y., Foster, G., Cherry, C., Macherey, W., Chen, Z., & Wu, Y. (2019). Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges. arXiv preprint arXiv:1907.05019. https://doi.org/10.48550/arXiv.1907.05019

- Arquilla, J., & Denning, P. J. (2022). The Context Problem in Artificial Intelligence: The Artificial Intelligence Design Challenge of Teaming Humans and Machines Is Difficult Because Machines Cannot Read the Context of Use. *Communications of the ACM, 65*(12), 18–21. https://doi.org/10.1145/3567605

- Athaluri, S. A., Manthena, S. V., Kesapragada, V. S. R. K. M., Yarlagadda, V., Dave, T., & Duddumpudi, R. T. S. (2023). Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References. *Cureus*, 15(4). https://doi.org/10.7759/cureus.37432

- Athanassopoulos, E., & Voskoglou, M. Gr. (2020). Quantifying Aristotle's Fallacies. *Mathematics*, 8(9), 1399. https://doi.org/10.3390/math8091399

- Azamfirei, R., Kudchadkar, S. R., & Fackler, J. (2023). Large Language Models and the Perils of Their Hallucinations. *Critical Care*, 27(1). https://doi.org/10.1186/s13054-023-04393-x

- Bansal, M., Pasunuru, R., & Krysciński, W. (2018). Multi-Reward Reinforced Summarization with Saliency and Entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 646–653). https://doi.org/10.18653/v1/N18-2102

- Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, 104(3), 671–732. https://doi.org/10.15779/Z38BG3

- Berk, H. (2024). Beware of Artificial Intelligence Hallucinations or Should We Call It Confabulation? *Acta Orthopaedica et Traumatologica Turcica*, 58(1), 1–3. https://doi.org/10.5152/j.aott.2024.130224

- Bozkurt, A., & Sharma, R. C. (2023). Generative AI and Prompt Engineering: The Art of Whispering to Let the Genie Out of the Algorithmic World. *International Journal of Interactive Multimedia and Artificial Intelligence*, 18(2). https://doi.org/10.5281/zenodo.8174941

- Bradley, A. P. (1997). The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. *Pattern Recognition*, 30(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

- Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., & Dafoe, A. (2020). *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*. arXiv preprint arXiv:2004.07213. https://arxiv.org/abs/2004.07213

- Capita. (2023). *Can AI Reshape Trade Compliance and Border Procedures?* Retrieved from https://www.capita.com/our-thinking/can-ai-reshape-trade-compliance-and-border-procedures
- Chinchor, N. (1992). MUC-4 Evaluation Metrics. In *Proceedings of the Fourth Message Understanding Conference (MUC-4)* (pp. 22–29). https://aclanthology.org/M92-1002
- Courville, A., Goodfellow, I., & Bengio, Y. (2016). *Deep Learning.* MIT Press. https://doi.org/10.1007/978-3-319-94463-0
- Dafoe, A., Brundage, M., & Garfinkel, B. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *arXiv preprint*, arXiv:2004.07213. https://doi.org/10.48550/arXiv.2004.07213
- Das, D., Gehrmann, S., & Wang, X. (2020). ToTTo: A Controlled Table-To-Text Generation Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 1173–1186). https://doi.org/10.18653/v1/2020.emnlp-main.89
- Dean, J., & Ghemawat, S. (2008). MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, 51(1), 107–113. https://doi.org/10.1145/1327452.1327492
- DeKay, C. (2023). 'Source?' 'I Made It Up': The Ethics of Citing ChatGPT in Academia. *Final Projects Summer 2023*. https://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=1009&context=fims_evolvingtech_finalproj_summer2023
- Denning, P. J., & Arquilla, J. (2022). The Context Problem in Artificial Intelligence: The Artificial Intelligence Design Challenge of Teaming Humans and Machines Is Difficult Because Machines Cannot Read the Context of Use. *Communications of the ACM*, 65(12), 18–21. https://doi.org/10.1145/3567605
- Dessimoz, C., & Thomas, P. D. (2024). AI and the Democratization of Knowledge. *Scientific Data*, 11(1), 268. https://doi.org/10.1038/s41597-024-03099-1
- Dziri, N., Madotto, A., Zaïane, O. R., & Bose, A. J. (2021). Neural Path Hunter: Reducing Hallucination in Dialogue Systems via Path Grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 2192–2207). https://doi.org/10.18653/v1/2021.emnlp-main.168
- Filippova, K. (2020). Controlled Hallucinations: Learning to Generate Faithfully from Noisy Data. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 864–870). https://doi.org/10.18653/v1/2020.findings-emnlp.76
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv preprint arXiv:2312.10997. https://doi.org/10.48550/arXiv.2312.10997
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org
- Hu, X., Tian, Y., Nagato, K., Nakao, M., & Liu, A. (2023). Opportunities and Challenges of ChatGPT for Design Knowledge Management. *Procedia CIRP*, 119, 21–28. https://doi.org/10.1016/j.procir.2023.05.001
- Huang, J., & Sciuchetti, M. (2023). ChatGPT Unveiled: Unleashing AI Magic in Online Shopping and Digital Marketing. *Atlantic Marketing Journal*, 12(2). https://digitalcommons.kennesaw.edu/cgi/viewcontent.cgi?article=1362&context=amj
- Huberman, B. A., & Mukherjee, S. (2024). Hallucinations and Emergence in Large Language Models. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4676180
- Hutson, J., & Plate, D. (2023). Human-AI Collaboration for Smart Education: Reframing Applied Learning to Support Metacognition. In *Applications of Artificial Intelligence in Engineering*. IntechOpen. https://doi.org/10.5772/intechopen.1001832

- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), 1–38. https://doi.org/10.1145/3571730

- Kryściński, W., McCann, B., Xiong, C., & Socher, R. (2019). Evaluating the Factual Consistency of Abstractive Text Summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 707–717). https://doi.org/10.18653/v1/D19-1075

- Lappalainen, Y., & Narayanan, N. (2023). Aisha: A Custom AI Library Chatbot Using the ChatGPT API. *Journal of Web Librarianship*, 17(3), 37–58. https://doi.org/10.1080/19322909.2023.2221477

- Lipton, Z. C. (2018). The Mythos of Model Interpretability. *Communications of the ACM*, 61(10), 36–43. https://doi.org/10.1145/3233231

- Maleki, N., Padmanabhan, B., & Dutta, K. (2024). AI Hallucinations: A Misnomer Worth Clarifying. arXiv preprint arXiv:2401.06796. https://doi.org/10.48550/arXiv.2401.06796

- Marcus, G., & Davis, E. (2019). *Rebooting AI: Building Artificial Intelligence We Can Trust*. Pantheon Books.

- McIntosh, T. R., Liu, T., Susnjak, T., Watters, P., Ng, A., & Halgamuge, M. N. (2023). A Culturally Sensitive Test to Evaluate Nuanced GPT Hallucination. *IEEE Transactions on Artificial Intelligence*, 1–13. https://doi.org/10.1109/TAI.2023.3332837

- Mesgar, M., Simpson, E., & Gurevych, I. (2021). Improving Factual Consistency Between a Response and Persona Facts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 521–535). https://doi.org/10.18653/v1/2021.eacl-main.44

- Metzger, L., Miller, L., Baumann, M., & Kraus, J. (2024). Empowering Calibrated (Dis-)Trust in Conversational Agents: A User Study on the Persuasive Power of Limitation Disclaimers vs. Authoritative Style. In *Proceedings of the 2024 ACM Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613904.3642122

- Panagopoulou, F., Parpoula, C., & Karpouzis, K. (2023). Legal and Ethical Considerations Regarding the Use of ChatGPT in Education. arXiv preprint arXiv:2306.10037. https://doi.org/10.48550/arXiv.2306.10037

- Parikh, A., Wang, X., Gehrmann, S., Faruqui, M., Dhingra, B., Yang, D., & Das, D. (2020). ToTTo: A Controlled Table-To-Text Generation Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 1173–1186). https://doi.org/10.18653/v1/2020.emnlp-main.89

- Parker, M. J., Anderson, C., Stone, C., & Oh, Y. (2024). A Large Language Model Approach to Educational Survey Feedback Analysis. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00414-0

- Pasunuru, R., & Bansal, M. (2018). Multi-Reward Reinforced Summarization with Saliency and Entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 646–653). https://doi.org/10.18653/v1/N18-2102

- Sanders, N. R., & Ganeshan, R. (2018). Big Data in Supply Chain Management. *Production and Operations Management, 27*(10), 1745–1748. https://doi.org/10.1111/poms.12845

- See, A., Pappu, A. S., Saxena, R., Yerukola, A., & Manning, C. D. (2019). Do Massively Pretrained Language Models Make Better Storytellers? In *Proceedings of the 23rd Conference on Computational Natural Language Learning* (pp. 843–861). https://doi.org/10.18653/v1/K19-1079

- Sokolova, M., & Lapalme, G. (2009). A Systematic Analysis of Performance Measures for Classification Tasks. *Information Processing & Management*, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971. https://arxiv.org/abs/2302.13971

- Wang, C., & Sennrich, R. (2020). On Exposure Bias, Hallucination and Domain Shift in Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3544–3552). ttps://doi.org/10.18653/v1/2020.acl-main.326

- Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper No. 601). https://doi.org/10.1145/3290605.3300831

- Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., & Weston, J. (2019). Neural Text Generation with Unlikelihood Training. In *Proceedings of the 8th International Conference on Learning Representations*. https://arxiv.org/abs/1908.04319

- Williams, A., Miceli, M., & Gebru, T. (2022). The Exploited Labor Behind Artificial Intelligence. *Noema Magazine*. https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/

- Woodie, A. (2023). Hallucinations, Plagiarism, and ChatGPT. *Datanami*. https://www.datanami.com/2023/01/17/hallucinations-plagiarism-and-chatgpt/

- Yadkori, Y. A., Kuzborskij, I., Stutz, D., György, A., Fisch, A., Doucet, A., Beloshapka, I., Weng, W.-H., Yang, Y.-Y., Szepesvári, C., Cemgil, A. T., & Tomasev, N. (2024). Mitigating LLM Hallucinations via Conformal Abstention. arXiv preprint arXiv:2405.01563. https://doi.org/10.48550/arXiv.2405.01563

- Ye, H., Liu, T., Zhang, A., Hua, W., & Jia, W. (2023). Cognitive Mirage: A Review of Hallucinations in Large Language Models. arXiv preprint arXiv:2309.06794. https://doi.org/10.48550/arXiv.2309.06794

- Yun, G. (2022). Ideasquares: Utilizing Generative Text as a Source of Design Inspiration. In *Proceedings of DRS2022 International Conference on Design Research Society*. https://doi.org/10.21606/drs.2022.484

- Zhang, S., Pan, L., Zhao, J., & Wang, Y. (2023). The Knowledge Alignment Problem: Bridging Human and External Knowledge for Large Language Models. arXiv preprint arXiv:2305.13669. https://arxiv.org/abs/2305.13669

- Zhao, Z., Cohen, S. B., & Webber, B. (2020). Reducing Quantity Hallucinations in Abstractive Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2237–2249). https://doi.org/10.18653/v1/2020.findings-emnlp.203

- Zodage, P., Harianawala, H., Shaikh, H., & Kharodia, A. (2024). Explainable AI (XAI): History, Basic Ideas, and Methods. *International Journal of Advanced Research in Science Communication and Technology*, 560–568. https://doi.org/10.48175/ijarsct-16988

- Østergaard, S. D., & Nielbo, K. L. (2023). False Responses from Artificial Intelligence Models Are Not Hallucinations. *Schizophrenia Bulletin*, 49(5), 1105–1107. https://doi.org/10.1093/schbul/sbad068