Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 441–449
DOI: 10.15439/2025F2314
ISSN 2300-5963 ACSIS, Vol. 43

# Leveraging Large Language Models for Automated Export Control Screening: Evaluating LLMs Framework

Salem Alotaibi*†, Alexei Lisitsa*, Antony McCabe*‡, Joanna MacSween§
*Department of Computer Science, University of Liverpool, Liverpool, UK
Emails: s.alotaibi8@liverpool.ac.uk, a.lisitsa@liverpool.ac.uk, tonymcc@liverpool.ac.uk
†Department of Computer Science and Artificial Intelligence, University of Bisha, Bisha, KSA
Email: salotaibi@ub.edu.sa
‡Computational Biology Facility, University of Liverpool, Liverpool, UK
§Export Control Officer, Legal and Governance, University of Liverpool, Liverpool, UK
Email: macsween@liverpool.ac.uk

***Abstract*—Export control (EC) compliance is a critical yet labour-intensive process within research institutions, where the classification of sensitive technologies and cross-border disclosures often depends on expert interpretation of complex legal frameworks. This paper investigates the potential of large language models (LLMs), specifically in this study ChatGPT-4o and LLaMA-3.3, to support EC screening through a multi-stage, expert-in-the-loop framework. The methodology includes prompt variation, regulatory conditioning, reflective reasoning, and expert-informed evaluation to simulate real-world compliance workflows. Using a curated dataset of UK research project descriptions and the UK Strategic Export Control List, we assess model performance across over 1,400 outputs. Results show that while both models benefit from domain-specific grounding, ChatGPT-4o consistently produces more stable and interpretable classifications. Prompt sensitivity, bias behaviour, and ambiguity handling are also examined to highlight model limitations. The findings suggest that LLMs can support early stage EC assessment but require structured prompting and human oversight to ensure regulatory alignment.**

***Index Terms*—Export control, large language models, regulatory compliance, prompt engineering, dual-use research, natural language processing, legal AI.**

## I. Introduction

EXPORT CONTROL (EC) regulations play a critical role in governing the dissemination of sensitive technologies, intellectual property, and research outputs. While this study focuses specifically on the UK's export control regime—including mechanisms such as the Open General Export Licence (OGEL)[1], Standard Individual Export Licence (SIEL)[2], and the SPIRE application system [3], we reference international frameworks such as the Export Administration Regulations (EAR)[4] and the International Traffic in Arms Regulations (ITAR)[5] to provide broader regulatory context. These frameworks aim to mitigate national security and foreign policy risks by restricting unauthorised access to dual-use or controlled items. Dual-use items include those with both civilian and military applications, such as advanced materials, sensors, and encryption technologies, which are particularly sensitive in academic research contexts. Relevant legal instruments include the UK Export Control Order 2008 [6], the EU Dual-Use Regulation (EU) 2021/821 [7], the US EAR [4], and the UK Strategic Export Control List [8].

Within the academic sector, ensuring compliance with these frameworks is increasingly complex due to international collaboration, evolving research domains (e.g., quantum computing, artificial intelligence), and growing scrutiny from government bodies.

Despite their significance, export control processes remain predominantly manual and highly dependent on expert interpretation. Compliance officers are often required to assess project descriptions against regulatory texts that are both voluminous and legally dense. This results in considerable administrative burden and interpretive inconsistency, especially in institutions managing large volumes of research proposals or operating across multiple jurisdictions.

Recent developments in natural language processing (NLP), particularly through the advent of LLMs, offer new avenues for supporting regulatory compliance. Models such as GPT-4 and LLaMA have demonstrated potential in legal classification, policy reasoning, and context-aware document analysis [9], [10], [11]. However, several limitations remain. Prior work highlights that existing LLM pipelines often fail to generalise across domains, lack grounding in domain-specific legal contexts, and exhibit limited capacity for justification or error correction [12], [13], [14]. As noted by Hussain et al. [9], automated compliance in sensitive areas such as export control demands a level of interpretability and consistency that many current approaches have yet to achieve.

This paper introduces a multi-stage, expert-in-the-loop framework for evaluating the capability of LLMs to assess research disclosures under export control criteria. The framework is designed to emulate key components of institutional compliance workflows, allowing for iterative refinement through prompt variation, domain-specific grounding, reflective reasoning, and expert-informed logic. Rather than

relying on retraining or symbolic transformation, the proposed approach evaluates whether LLMs can reason effectively over full regulatory documents and project metadata.

The contributions of this paper are threefold. First, we propose a structured methodology for assessing LLM performance across different stages of reasoning, incorporating domain knowledge and expert feedback. Second, we conduct a comprehensive evaluation of ChatGPT-4o and LLaMA-3.3 across 1400 outputs using varied prompts, conditions, and error correction procedures. Third, we analyse model behaviour under ambiguity and potential bias, offering practical insights into reliability and risk in compliance decision-making.

The remainder of the paper is structured as follows: Section II reviews related work on automated legal compliance and the application of LLMs in regulatory domains. Section III outlines our proposed five-stage framework and design rationale. Section IV details the experimental setup, data sources, and evaluation procedures. Section V presents empirical findings on model performance across prompts, stages, and error correction conditions. Section VI contextualises these findings with respect to model generalisability, domain adaptation, and deployment feasibility. Finally, Section VII concludes with key insights and directions for future work.

## II. Background and Related Work

Automating regulatory compliance has long posed a significant challenge for organisations operating within heavily regulated sectors such as finance, healthcare, and export control. The automation of regulatory compliance has traditionally relied on symbolic and rule-based approaches, including formal logic, ontologies, and information retrieval methods. These methods often apply sentence-level analysis and handcrafted rules to extract obligations or classify provisions. Traditional ML-based compliance checking in this area has relied on manually curated datasets and conceptual models, [15], [16]. While rule-based and symbolic methods offer transparency, many ML techniques particularly deep learning models require additional strategies to ensure interpretability and alignment with legal reasoning [9]. These methods typically focus on completeness and semantic coverage but struggle with ambiguity and cross-paragraph consistency [17]. This limitation becomes particularly pronounced in domains where legal language is dense and referential.

The emergence of large language models has prompted a shift in how regulatory artefacts are analysed. Unlike their symbolic predecessors, models such as ChatGPT-4o and LLaMA-3.3 are capable of interpreting language at the paragraph or document level, capturing contextual relationships and legal nuance in a way that more traditional techniques cannot [18]. Recent work by Hassani et al. [9] illustrates that LLMs outperform both keyword-based systems and fine-tuned BERT variants in legal classification tasks, particularly in domains like General Data Protection Regulation (GDPR) and food safety. An extension of this work [12] further examines paragraph-level grounding and contextual relevance. Similarly, [19] evaluated ChatGPT and LLaMA in the interpretation of clinical guidelines, highlighting their ability to follow structured protocols. Recent benchmarks by Yin et al. [20] also demonstrate LLM performance in semantic mapping of Sustainable Development Goals (SDGs), reinforcing their applicability in policy-oriented contexts.

LLMs have also shown potential for detecting inconsistencies in legal and technical requirements. In [21], they demonstrated that ChatGPT can be used to flag contradictions in natural language software specifications, highlighting broader utility for regulatory oversight. Similarly, Zhang et al. [11] showed that hallucination risks persist in extractive clinical NLP tasks, emphasising the need for cautious deployment in compliance settings.

Recent implementations have integrated LLMs into structured compliance systems. For example, the Gracenote platform [10] uses GPT-4 and prompt engineering to generate obligations registers, support consultation tools, and perform regulatory horizon scanning. LangChain pipelines have further enhanced LLM performance through document embeddings and retrieval-augmented generation (RAG). Al-Turki et al. [22] developed a human-in-the-loop LLM architecture for building regulation compliance, translating legal texts into structured YAML formats using GPT-4 with iterative expert feedback. Complementary work by Chen et al. [13] integrates deep learning and ontologies to check BIM (Building Information Modelling) compliance, combining LLMs with symbolic representation. These advances illustrate the growing role of LLMs as context-aware, domain-adaptable engines for compliance support.

This emphasis on context-aware interpretability and structured reasoning is echoed in recent work on legal NLP and compliance-oriented modelling. For example, a Bi-LSTM-based architecture has been used to detect explicit cause-and-effect relationships in court judgments, using labelled sentence structures to support traceable and explainable inference over legal arguments [23].

A risk-based quality control framework for legally regulated software further demonstrates how domain-specific structural representations can enhance traceability, regulatory robustness, and the auditability of automated compliance workflows [24].

The translation of natural language requirements into access control policies using LLMs has also been explored in the context of policy automation, offering insights relevant to export screening and classification workflows [25].

However, despite these developments, the use of LLMs in export control compliance remains largely underexplored. This domain involves a particularly challenging regulatory landscape governed by overlapping national and international regimes such as the UK Export Control Order, the EU Dual-Use Regulation, and the US EAR. In [14], the authors demonstrated a similar application in financial auditing, highlighting LLMs' potential to support layered and context-sensitive regulatory checks. The complexity of export classifications ranging from technical specification alignment to licensing decisions which requires a level of interpretive reasoning that

few systems currently offer.

Based on the literature, this paper highlights three key limitations:(1) the lack of annotated export control datasets for training or benchmarking; (2) minimal evaluation of LLMs across different legal jurisdictions; and (3) limited deployment of LLMs for core export tasks such as EC classification, red-flag screening, or licensing eligibility.

In this context, this paper explores a framework to evaluate whether ChatGPT-4o and LLaMA-3.3 can directly assess the applicability of export control regulations based on full-text legal documents and project descriptions. Unlike approaches that rely on intermediate representations or symbolic transformation, our framework ingests unstructured regulatory texts and infers, from context, whether export restrictions apply to a given scenario. This mirrors how compliance officers perform early-stage screening and allows us to evaluate the practical value of LLMs in complex classification tasks, particularly in domains like export control where project-specific variables and legal precision are critical.

## III. Methodology

This work proposes a five-stage, expert-in-the-loop framework to evaluate and iteratively improve the ability of LLMs to classify research outputs under export control regulations. The methodology simulates regulatory decision-making by combining baseline reasoning, regulatory conditioning, error analysis, reflective prompting, and expert-guided optimisation. Each stage builds upon the previous to promote contextual learning and alignment with expert judgement.

The framework is summarised in Fig. 1 and includes the following stages:

1) **Baseline Inference (Stage 1)**: In this stage, the language model is prompted to classify domain-specific research descriptions without prior instruction. This establishes a baseline for the model's default reasoning capabilities. The model also generates a natural language justification for each prediction
2) **Domain Conditioning (Stage 2)**: The UK Strategic Export Control List is introduced to the model as context. Models are then re-evaluated on the same dataset to assess domain knowledge integration.
3) **Disagreement Analysis (Stage 3)**: Model predictions are compared with expert annotations. Disagreements are categorised to identify systematic reasoning issues.
4) **Model Reflection (Stage 4)**: In this stage, disagreement examples are reprocessed using self-prompting techniques that guide the model to reflect on its earlier decision. Prompts are crafted to encourage regulatory reasoning, contextualization, and consistency with prior outputs. The goal is to assess whether reflective prompting improves alignment without additional training
5) **Expert-in-the-Loop Optimisation (Stage 5)**: Model outputs from the reflection stage are reviewed again by the expert, who provides feedback on both classification decisions and explanations. This feedback is used to refine prompts and adjust reasoning templates, forming a closed loop for improving the model's interpretability and regulatory conformity.

This iterative process promotes improved regulatory alignment and explanation quality without the need for model retraining.

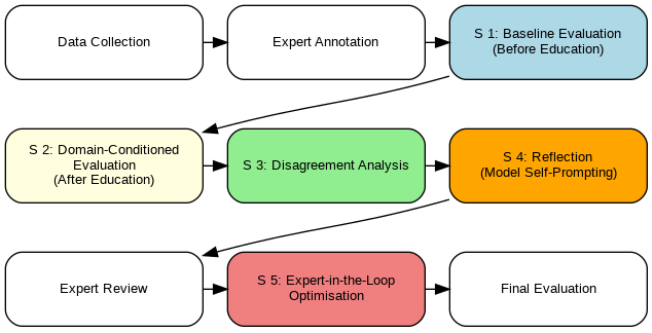


Fig. 1. Five-Stage Expert-Enhanced Framework for Export Control Classification.

## IV. Experimental Setup

To evaluate the proposed methodology, a series of experiments were conducted using ChatGPT-4o and LLaMA-3.3. The setup includes data sources, model prompting, evaluation metrics, and bias analysis.

### A. Data Collection

Two primary sources were used:

- **Regulatory Document:** The UK Strategic Export Control List was retrieved from official UK government publications.[1], [2], [8].
- **Project Dataset:** 50 research project descriptions were collected from the UKRI EPSRC funding portal[1]. Each entry included the fields: Title, Abstract, Project Partners, Organisation, and Department.

Each project description was independently annotated by a compliance officer from the University of Liverpool. Labels were binary, indicating whether the project should be subject to export control. These expert annotations were treated as ground truth across all experiments. Seven additional cases were marked as ambiguous for later bias and uncertainty evaluation.

### B. Prompt Design and Evolution

Prompt engineering was central to model evaluation. Each stage involved a different prompting strategy:

*1) Initial Prompt:*

> I will give you a description of many grant projects from EPSRC in the UK. The description includes Title, Abstract, Project Partners, Organisation, and Department. I need you to tell me whether the owner of this project should apply for export control items or not and give me an explanation for your answer.

[1]https://www.ukri.org/councils/epsrc/

*2) Prompt Diversity:* To evaluate model robustness and sensitivity to prompt phrasing, five semantically similar variants of the initial prompt were generated per model. These new prompts were then used to rerun both the Stage 1 and Stage 2 evaluations, producing six runs per model (original + 5 variants), resulting in 24 experimental conditions and 1200 model outputs in total.

*3) Reflective Prompting:* Incorrect predictions were re-evaluated using prompts designed to promote self-correction:

> You previously answered that export control applies. Re-assess this decision considering the regulation provided. Identify any mistakes in your previous reasoning and provide a revised classification and justification.

*4) Final Structured Prompt:* A compliance-driven decision flow was embedded into a structured Yes/No prompt:

> I will describe a research or teaching project. Follow this step-by-step Yes/No flow to assess whether export controls apply:
>
> 1) Is the project exporting materials, technology, software, or information outside the UK? (Consider collaborators, overseas travel, or postgraduate students overseas.)
> → If NO, export controls do not apply. If YES, continue.
> 2) Do you know or suspect that any participant (staff, student, collaborator, or end user) intends to use the materials for military or WMD purposes?
> → If YES, export controls apply — recommend contacting `exportcontrol@liverpool.ac.uk`. If NO, continue.
> 3) Are any items subject to US export controls?
> → If YES, flag as potentially restricted and recommend further review. If NO, continue.
> 4) Are any items on the UK Strategic Export Control List, or is the project working with sanctioned entities?
> → If NO, export controls do not apply. If YES, continue.
> 5) Does the project fall under any exemptions (public domain, basic scientific research, or patent filing)?
> → If YES, export controls may not apply. If NO, export controls apply — recommend contacting `exportcontrol@liverpool.ac.uk`.
>
> Always reference UK Government and University of Liverpool guidance in your reasoning [2], [26].

### C. Model Configurations

Two LLMs were tested: ChatGPT-4o, accessed via the OpenAI API, and LLaMA-3.3, accessed via the Hugging Face Inference API.

### D. Evaluation Metrics

The following metrics were used: Accuracy, Precision, Recall, F1-score, and confidence scores on incorrect predictions. Confidence scores were obtained by instructing the model to label each decision with one of three categories: Low (0–40%), Medium (41–70%), or High (71–100%), based on how confident it was in its prediction. Prompt variance was measured as the standard deviation in accuracy across prompt variants.

### E. Ambiguity and Bias Evaluation

Ambiguous cases were used to test confidence stability across prompts. Additionally, a bias test was conducted by modifying the institutional affiliation of partners (e.g., replacing EU with non-EU universities) to observe if geopolitical factors affected model judgement.

While this multi-stage design may appear complex, it is intended to emulate the layered decision-making practices of institutional export control review. Compliance officers often rely on contextual interpretation, iterative clarification, and expert oversight—requirements that cannot be captured by one-shot classification alone. The staged approach allows the model to integrate domain knowledge, reflect on prior outputs, and adapt to expert feedback, thereby aligning more closely with real-world compliance workflows.

## V. Results

### A. Initial Prompt Performance

Initial classification performance was evaluated using the original prompt, which produced three output labels: `YES`, `NO`, and `MAYBE`. In Stage 1, ChatGPT-4o tended to produce confident binary classifications (40% YES, 60% NO, 0% MAYBE), whereas LLaMA-3.3 showed greater uncertainty, assigning 30% YES, 15% NO, and 55% MAYBE. Following domain-specific conditioning in Stage 2, both models showed improved calibration. ChatGPT-4o's MAYBE rate increased slightly to 2%, while LLaMA-3.3 reduced its MAYBE usage to 8%, achieving better balance across output categories. These shifts indicate a positive effect of domain exposure in aligning outputs with regulatory expectations.

### B. Prompt Sensitivity and Variance

To complement the overall classification performance analysis, we examined the extent to which prompt phrasing influences model reliability. Each model was evaluated using five semantically equivalent prompt variants differing in structure and wording. Although this introduced some fluctuation, the overall impact on classification metrics was relatively minor, particularly for ChatGPT-4o.

As illustrated in Fig. 3 and Fig. 2, ChatGPT-4o maintained stable F1-scores across all five prompts in Stage 1, ranging from 60.61% to 70.59%, with a standard deviation of 4.33 percentage points. In Stage 2, the F1-score range expanded to 64.52%–75.29%, with a standard deviation of 4.24 percentage points, reflecting slightly greater variability under conditioning. Similarly, its accuracy fluctuated by only 6.97 percentage points between the best- and worst-performing prompt.

In contrast, LLaMA-3.3 exhibited greater prompt sensitivity in Stage 1, with F1-scores ranging from 63.41% to 72.22% and an accuracy spread of 11.62 percentage points. In Stage 2, the accuracy range narrowed to 4.65 percentage points, though F1-score variability remained comparable (62.86%–70.27%, standard deviation of 3.36 percentage points). This suggests modest gains in output consistency, but not a clear improvement in classification performance.

Prompt paraphrasing thus introduced measurable variation in classification performance across both models. Structured prompts that incorporated regulatory keywords or decision templates tended to elicit more consistent and context-aware outputs, whereas open-ended variants were more sensitive to ambiguity. ChatGPT-4o also exhibited lower overall variance in prediction confidence, suggesting more stable calibration under prompt variation.

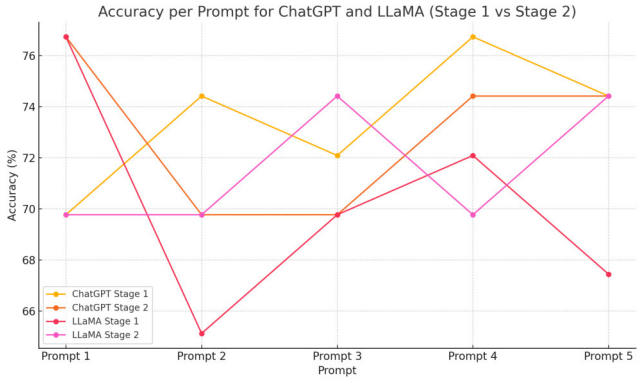


Fig. 2. Accuracy per prompt across ChatGPT-4o and LLaMA-3.3 in Stage 1 and Stage 2.

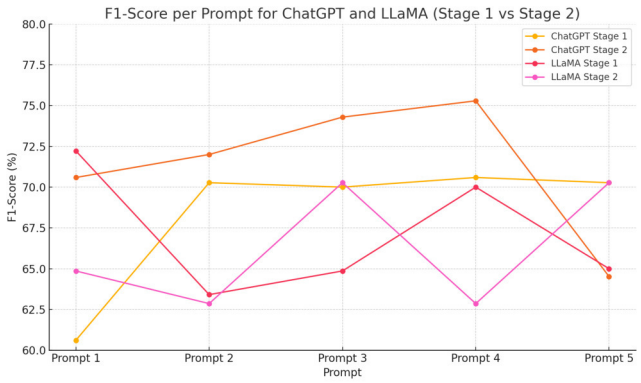


Fig. 3. F1-score per prompt for ChatGPT-4o and LLaMA-3.3 in Stage 1 and Stage 2.

### C. Effect of Domain Conditioning

To evaluate the effect of domain-specific education, we compared model performance across all five prompt variants in Stage 1 (pre-conditioning) and Stage 2 (post-conditioning). This setup reflects how each model generalises its export control reasoning before and after exposure to the UK Strategic Export Control List.

Across all five prompt variants, ChatGPT-4o maintained a strong baseline with an average accuracy of 73.49% in Stage 1, and a comparable 73.02% in Stage 2. While the difference in accuracy was negligible, the model exhibited slightly improved calibration through more consistent outputs and reduced variance in confidence scores. LLaMA, by contrast, demonstrated more visible changes in response to domain conditioning. Its average accuracy improved from 70.23% to 71.63%. More notably, LLaMA-3.3's confidence in incorrect predictions decreased by approximately 8.8 percentage points which indicating better uncertainty management after incorporating regulatory context.

### D. Reflective Inference Impact

Stage 4 introduced reflective prompting to evaluate whether large language models could revise incorrect classifications when explicitly asked to reassess their prior outputs. ChatGPT-4o corrected 4 out of 11 previous errors, resulting in a 36.36% resolution rate and an F1-score increase from 64.52% to 78.79%. In contrast, LLaMA-3.3 corrected none of its 11 prior misclassifications, and its F1-score declined due to additional false negatives.

A representative example illustrates this dynamic: ChatGPT-4o initially classified a cybersecurity research project with UK defence partners (e.g., RAF, Rolls-Royce) as controlled due to perceived dual-use concerns. After reflection, it revised the label to "not controlled," noting that the project lacked direct transfers or controlled components, and instead recommended internal review. This revision aligned with the expert's nuanced judgment and indicates improved reasoning alignment through structured metacognitive intervention.

These findings suggest that reflective prompting can enhance decision quality in models with sufficient contextual comprehension, such as ChatGPT-4o, but may be ineffective for models like LLaMA-3.3, which lack adequate task adaptation capacity or grounding.

### E. Final Prompt Effectiveness

Stage 5 evaluated an expert-informed prompt based on institutional decision logic. As shown in Table I, this prompt led to the highest F1-scores for both models: 82.40% for ChatGPT-4o and 81.08% for LLaMA-3.3. ChatGPT-4o achieved balanced improvements in both precision and recall, while LLaMA-3.3's performance was driven by a substantial increase in recall 93.75%

TABLE I
PERFORMANCE USING EXPERT-INFORMED FINAL PROMPT (STAGE 5)

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|
| ChatGPT-4o (Stage 1) | 73.49 | 58.58 | 82.67 | 68.35 |
| ChatGPT-4o (Stage 5) | 86.00 | 77.80 | 87.50 | 82.40 |
| LLaMA-3.3 (Stage 1) | 70.23 | 54.89 | 86.67 | 67.10 |
| LLaMA-3.3 (Stage 5) | 83.72 | 71.43 | 93.75 | 81.08 |

The table and figure demonstrate that the expert-informed prompt (Stage 5) led to the best overall performance. ChatGPT-4o reached 82.40% with balanced precision and recall, while LLaMA-3.3 attained its highest F1-score 81.08% primarily through a notable increase in recall 93.75%.

### F. Ambiguity Handling and Bias Evaluation

Model behaviour on ambiguous export control cases was assessed using five prompts with uncertain control status. ChatGPT-4o maintained high and stable average confidence

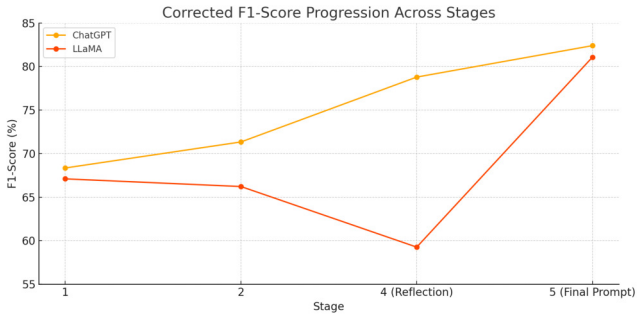


Fig. 4. F1-score progression across stages for both models. Stage 5 shows peak performance, with ChatGPT-4o reaching 82.40% and LLaMA-3.3 achieving 81.08%, driven by improved precision and recall.

levels across all prompts in Stage 1 85–92% and Stage 2 82–91%, as shown in figures 5 and 6. In contrast, LLaMA-3.3 exhibited broader fluctuations, with confidence levels ranging from 65% to 80%.

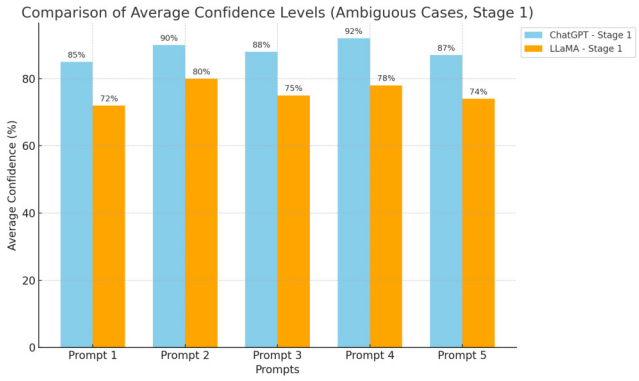


Fig. 5. Average confidence levels for ambiguous cases Stage 1. ChatGPT-4o shows higher and more consistent confidence than LLaMA-3.3 across all prompts.

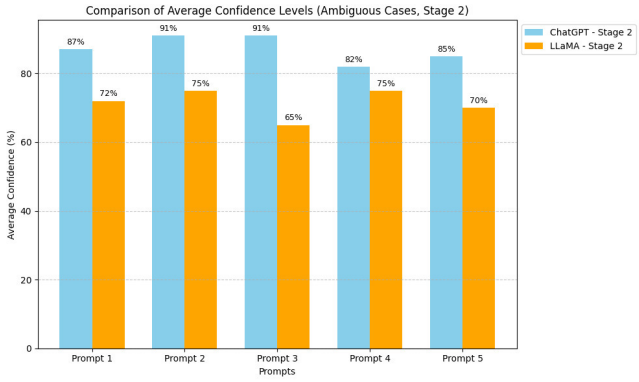


Fig. 6. Average confidence levels for ambiguous cases Stage 2. ChatGPT-4o maintains stable high confidence; LLaMA-3.3 remains more variable and generally lower.

Figure 7 illustrates the variance in confidence across stages. ChatGPT-4o's variance remained low and stable, at 43.7 in Stage 1 and 44.6 in Stage 2, whereas LLaMA-3.3's was substantially higher, at 180.3 in Stage 1 and 173.4 in Stage 2, indicating persistent volatility.

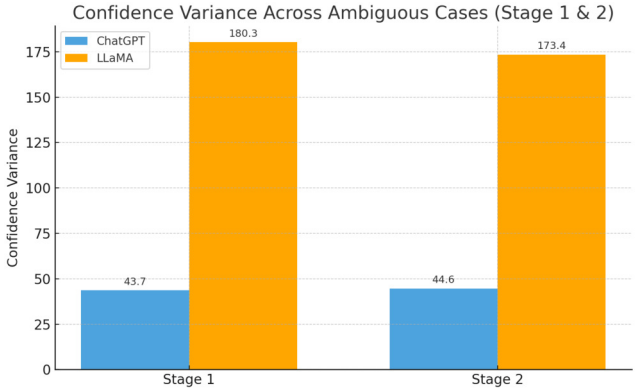


Fig. 7. Confidence variance on ambiguous cases. ChatGPT-4o shows stable variance across stages, while LLaMA-3.3's variance remains significantly higher, despite domain conditioning.

These findings suggest that ChatGPT-4o is better calibrated under uncertainty, whereas LLaMA-3.3 continues to exhibit erratic confidence distributions, reducing reliability in ambiguous classifications.

### G. Bias in Partner Attribution

To evaluate the influence of institutional affiliation on compliance classification, we tested models on paired inputs with identical technical descriptions but different international partners. As shown in Table II, the inclusion of a non-EU collaborator (King Saud University) resulted in a classification change from "not controlled" to "controlled" for both ChatGPT-4o and LLaMA-3.3, despite no changes in project content, scope, or end-use.

The shift in decision was reflected in the model's generated justifications, which included references to U.S. export control frameworks (e.g., ITAR/EAR), reclassification of the technology as dual-use, and heightened concern over sensitive or military end-use. By contrast, substitution with a European partner (University of Amsterdam) produced no change in classification or in the model's accompanying rationale.

These findings indicate a model-level inclination to infer elevated risk based on geopolitical affiliation. These results suggest that models may be over-weighting partner nationality or geopolitical affiliation, even when the technical content of the project remains unchanged. This raises concerns about fairness and legal validity in automated screening, highlighting the need for explicit bias mitigation strategies in export control applications.

## VI. Discussion

This study aimed to critically evaluate the capacity of large language models to reason about export control classifications using a staged framework. The experimental objectives focused on four dimensions of model behaviour: (1) prompt variation sensitivity, (2) the effect of domain-specific conditioning

TABLE II
IMPACT OF ADDING FOREIGN PARTNERS ON EXPORT CONTROL COMPLIANCE DECISIONS

| Title | Project Partners | Organisation | ChatGPT-4o | LLaMA-3.3 |
|---|---|---|---|---|
| XXXXXX | BT, Intel, Open Networking Foundation, Ori Industries 1 Ltd, Yale University, King Saud University (Saudi Arabia) | Univ. of Oxford | YES | YES |
| XXXXXX | BT, Intel, Open Networking Foundation, Ori Industries 1 Ltd, Yale University, University of Amsterdam | Univ. of Oxford | NO | NO |

*Note: For both examples shown above, the LLMs' original classification decisions (before the addition of new partners) were "NO."*

("education"), (3) the impact of reflective prompting for error correction, and (4) the value of expert-informed structured prompts. In addition, model responses to ambiguous cases and bias-inducing input modifications were assessed to explore practical risks in regulatory deployment.

Prompt variation tests revealed that both models were somewhat sensitive to surface-level differences in input phrasing, though this effect was less pronounced than expected. ChatGPT-4o demonstrated stronger robustness to linguistic variation across all stages, aligning with findings by Bogireddy and Dasari [18], who observed consistent performance from ChatGPT-4o across diverse tasks. LLaMA-3.3, on the other hand, showed substantial variance in Stage 1, with up to 11.6 percentage points difference in accuracy across prompts. This instability improved notably post-conditioning. These results suggest that while prompt phrasing can influence model responses, its effect is comparatively modest in classification contexts, particularly for better calibrated models like ChatGPT-4o. Nevertheless, this sensitivity may still pose risks in high-stakes regulatory settings if users craft prompts inconsistently.

Domain-specific conditioning through the provision of the UK Strategic Export Control List was introduced in Stage 2 to simulate regulatory education. ChatGPT-4o's accuracy rose by just 0.1%, while LLaMA-3.3 improved by 0.6%. The more significant impact was observed in confidence calibration: LLaMA's confidence on incorrect predictions dropped by nearly 9 percentage points, indicating improved uncertainty management. More sophisticated grounding techniques, such as retrieval-augmented generation (RAG), may be required to meaningfully shift classification outcomes, as RAG models have been shown to improve factual accuracy, specificity, and task performance by combining pre-trained parametric memory with non-parametric knowledge sources [27].

Stage 4 explored whether models could improve their decisions through structured reflection. ChatGPT-4o successfully revised 36.4% of its prior errors, demonstrating an ability to align with regulatory expectations after self-assessment. This supports research by Lawal et al. [25], who found that prompting LLMs to reflect on prior outputs can lead to more policy aligned decisions. In contrast, LLaMA-3.3 failed to correct any outputs, reinforcing concerns around the model's adaptability and interpretive reasoning under pressure. The divergent performance highlights the need for models to not only generate initial outputs reliably but also to re-engage critically when prompted, particularly in domains that require explainability and self-consistency.

Stage 5 incorporated an expert-designed prompt that mirrored institutional export control procedures. This structured, decision-flow-based format led to the highest performance for both models, especially in recall. LLaMA-3.3's recall increased to 93.75%, a substantial leap from earlier stages. These results mirror findings from Al-Turki et al. [22], who showed that expert-guided input schemas can compensate for model instability. Unlike earlier stages, which tested the models' inherent flexibility, this final configuration emphasised alignment with operational logic. The results underscore the importance of designing prompts that do not merely inform the model but actively shape its reasoning observable in the generated outputs in a way that mirrors human expert processes. This strategy loosely aligns with principles of chain-of-thought prompting, where structured input sequences encourage more consistent and interpretable decision-making.

The final experimental tasks evaluated model behaviour in ambiguous or politically sensitive cases. ChatGPT-4o exhibited stable confidence across all uncertainty cases, while LLaMA-3.3 continued to show high variance. Bias testing further revealed problematic behaviours: both models flagged higher regulatory risk when international collaborators were from non-EU regions, despite identical technical content. This highlights the urgent need for robust bias mitigation strategies if LLMs are to be deployed in real-world compliance workflows.

In summary, the results suggest that while domain-specific input and expert-informed prompts can improve regulatory alignment, gains from prompt variation and naive domain conditioning remain modest. Models like ChatGPT-4o show stronger baseline reliability, while LLaMA-3.3 benefits more from external structure. Ultimately, consistent improvements were achieved only when expert logic was embedded directly into the input format, pointing toward the value of hybrid frameworks that combine human expertise with adaptable LLM reasoning.

Although Stage 5 achieved the highest performance metrics, it also required the most expert intervention and prompt engineering effort. In contrast, Stage 2 produced modest

but reliable gains with minimal human oversight, offering a more scalable compromise between accuracy and automation. Future deployment strategies may adopt a hybrid model, where expert input is reserved for edge cases flagged by lower-confidence outputs.

## VII. Conclusion and Future Work

This study investigated the use of LLMs for export control classification of research disclosures, proposing a multi-stage, expert-in-the-loop framework. Through prompt variation, domain-specific conditioning, reflective prompting, and expert-informed refinement, we evaluated the classification performance and behavioural consistency of ChatGPT-4o and LLaMA-3.3. Our findings suggest that structured prompting and regulatory context can meaningfully enhance model reliability, with ChatGPT-4o demonstrating more consistent and interpretable outputs. However, challenges remain, particularly in handling ambiguous phrasing, edge cases, and politically sensitive affiliations.

Future work will address several areas. First, we plan to significantly expand the dataset, incorporating a broader range of research descriptions and licensing scenarios. This will enable more rigorous statistical analysis and improve generalisability. Second, we aim to evaluate additional LLM architectures, including newer open-source models and multilingual variants to assess comparative strengths across regulatory regimes. Third, we intend to explore the integration of automated outputs into institutional review processes, to assess whether human-machine collaboration improves both decision accuracy and accountability. Finally, we will extend this approach to adjacent compliance domains, such as sanctions screening and dual-use research oversight, where similar regulatory reasoning is required.

We also intend to experiment with retrieval-augmented generation (RAG) architectures to assess whether real-time document retrieval can enhance legal grounding and factual consistency, while monitoring for risks such as retrieval drift or hallucinated justifications.

## Acknowledgments

Salem is a PhD student funded by the University of Bisha, Saudi Arabia.

## References

[1] Export Control Joint Unit, Department for International Trade, and Department for Business and Trade, "Open general export licences (ogels)," 2025, last updated: 6 March 2025. [Online]. Available: https://www.gov.uk/government/collections/open-general-export-licences-ogels

[2] UK Department for International Trade and Export Control Joint Unit, "Do i need an export licence?" 2019, last updated: 13 August 2019. [Online]. Available: https://www.gov.uk/guidance/beginners-guide-to-export-controls

[3] UK Department for Business and Trade, "Spire export control system," 2024. [Online]. Available: https://www.spire.trade.gov.uk

[4] U.S. Department of Commerce, Bureau of Industry and Security, "Export administration regulations (ear)," 2024, accessed: May 2024. [Online]. Available: https://www.bis.gov/regulations/ear

[5] U.S. Department of State, Directorate of Defense Trade Controls, "International traffic in arms regulations (itar)," 2024. [Online]. Available: https://www.pmddtc.state.gov/ddtc_public?id=ddtc_public_portal_itar_landing

[6] UK Government, "The export control order 2008," 2008, statutory Instrument No. 3231. [Online]. Available: https://www.legislation.gov.uk/uksi/2008/3231/contents/made

[7] European Parliament and Council, "Regulation (eu) 2021/821 setting up a union regime for the control of exports, transfer, brokering and transit of dual-use items," 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32021R0821

[8] Export Control Joint Unit, Department for International Trade, and Department for Business and Trade, "Export controls: Military goods, software and technology," 2024, last updated: 10 April 2024. [Online]. Available: https://www.gov.uk/guidance/export-controls-military-goods-software-and-technology

[9] S. Hassani, M. Sabetzadeh, D. Amyot, and J. Liao, "Rethinking legal compliance automation: Opportunities with large language models," in *2024 IEEE International Requirements Engineering Conference (RE)*, 2024, pp. 432–440. [Online]. Available: https://doi.org/10.1109/RE59067.2024.00051

[10] J. Ioannidis, J. Harper, M. S. Quah, and D. Hunter, "Gracenote.ai: Legal generative ai for regulatory compliance," in *Proceedings of the 3rd International Workshop on AI for Legal Professionals (LegalAIIA)*, 2023. [Online]. Available: https://doi.org/10.2139/ssrn.4494272

[11] H. Zhang, N. Jethani, S. Jones, N. Genes, V. J. Major, I. S. Jaffe, and et al., "Evaluating large language models in extracting cognitive exam dates and scores," *PLOS Digit Health*, vol. 3, no. 12, p. e0000685, 2024. doi: 10.1371/journal.pdig.0000685. [Online]. Available: https://doi.org/10.1371/journal.pdig.0000685

[12] S. Hassani, "Enhancing legal compliance and regulation analysis with large language models," in *2024 IEEE International Requirements Engineering Conference (RE)*, 2024. doi: 10.1109/RE59067.2024.00065 pp. 507–511. [Online]. Available: https://doi.org/10.1109/RE59067.2024.00065

[13] N. Chen, X. Lin, H. Jiang, and Y. An, "Automated building information modeling compliance check through a large language model combined with deep learning and ontology," *Buildings*, vol. 14, no. 7, p. 1983, 2024. [Online]. Available: https://doi.org/10.3390/buildings14071983

[14] A. Berger, L. Hillebrand, D. Leonhard, T. Deußer, T. B. F. D. Oliveira, T. Dilmaghani, and R. Sifa, "Towards automated regulatory compliance verification in financial auditing with large language models," in *2023 IEEE International Conference on Big Data (BigData)*, 2023, pp. 4626–4635. [Online]. Available: https://doi.org/10.1109/BigData59044.2023.10386518

[15] S. Wilson, F. Schaub, Y. Agarwal, A. Acquisti, L. Cranor, and N. Sadeh, "The creation and analysis of a website privacy policy corpus," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, vol. 1, 2016, pp. 1330–1340. [Online]. Available: https://doi.org/10.18653/v1/P16-1126

[16] O. Amaral, S. Abualhaija, D. Torre, M. Sabetzadeh, and L. C. Briand, "AI-enabled automation for completeness checking of privacy policies," *IEEE Transactions on Software Engineering*, vol. 48, no. 11, pp. 4647–4674, 2022. doi: 10.1109/TSE.2021.3124332. [Online]. Available: https://doi.org/10.1109/TSE.2021.3124332

[17] K. M. Sathyendra, S. Wilson, F. Schaub, S. Zimmeck, and N. Sadeh, "Identifying the provision of choices in privacy policy text," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1294 pp. 2774–2779. [Online]. Available: https://doi.org/10.18653/v1/D17-1294

[18] S. R. Bogireddy and N. Dasari, "Comparative analysis of chatgpt-4 and llama: Performance evaluation on text summarization, data analysis, and question answering," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024. doi: 10.1109/ICCCNT61001.2024.10725662 pp. 1–7. [Online]. Available: https://doi.org/10.1109/ICCCNT61001.2024.10725662

[19] S. Pandya, T. E. Bresler, T. Wilson, Z. Htway, and M. Fujita, "Decoding the nccn guidelines with ai: A comparative evaluation of chatgpt-4.0 and llama 2 in the management of thyroid carcinoma," *The American Surgeon*, vol. 91, no. 1, pp. 94–98, 2025. [Online]. Available: https://doi.org/10.1177/00031348241269430

[20] H. Yin, A. Aryani, and N. Nambiar, "Evaluating the performance of large

language models for sdg mapping," arXiv, techreport arXiv:2408.02201, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2408.02201

[21] S. Fantechi, L. Gnesi, L. Passaro, and L. Semini, "Inconsistency detection in natural language requirements using chatgpt: A preliminary evaluation," in *Proceedings of the 2023 IEEE 31st International Requirements Engineering Conference (RE)*, 2023. [Online]. Available: https://doi.org/10.1109/RE57278.2023.00045

[22] D. Al-Turki, H. Hettiarachchi, M. M. Gaber, M. M. Abdelsamea, S. Basurra, S. Iranmanesh, H. Saadany, and E. Vakaj, "Human-in-the-loop learning with llms for efficient rase tagging in building compliance regulations," *IEEE Access*, 2024. doi: 10.1109/ACCESS.2024.3512434 Early Access. [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3512434

[23] Łukasz Kurant, "Mechanism for detecting cause-and-effect relationships in court judgments," in *Annals of Computer Science and Information Systems*, vol. 35, 2023. doi: 10.15439/2023F4827 pp. 1041–1046. [Online]. Available: https://doi.org/10.15439/2023F4827

[24] M. Esche, L. Ho, M. Nischwitz, and R. Meyer, "Risk-based continuous quality control for software in legal metrology," in *Annals of Computer Science and Information Systems*, vol. 35, 2023. doi: 10.15439/2023F6171 pp. 451–461. [Online]. Available: https://doi.org/10.15439/2023F6171

[25] S. Lawal, X. Zhao, A. Rios, R. Krishnan, and D. Ferraiolo, "Translating natural language specifications into access control policies by leveraging large language models," in *2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*, 2024. doi: 10.1109/TPS-ISA62245.2024.00048 pp. 361–370. [Online]. Available: https://doi.org/10.1109/TPS-ISA62245.2024.00048

[26] University of Liverpool Legal and Compliance, "Export controls: How export control legislation applies to collaborating internationally," 2024, accessed: May 2024. [Online]. Available: https://www.liverpool.ac.uk/legal/exportcontrols/

[27] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Y. Chang, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474. [Online]. Available: https://doi.org/10.48550/arXiv.2005.11401