Partner Attribution Bias in LLM-Assisted Export Control Screening

Salem Alotaibi[1,2], **Alexei Lisitsa**[1], **Antony McCabe**[1,3]

[1]Department of Computer Science, University of Liverpool, Liverpool, UK
[2]Department of Computer Science and Artificial Intelligence, University of Bisha, Bisha, KSA
[3]Computational Biology Facility, University of Liverpool, Liverpool, UK

s.alotaibi8@liverpool.ac.uk, salotaibi@ub.edu.sa, a.lisitsa@liverpool.ac.uk, tonymcc@liverpool.ac.uk

Abstract

Automating export control screening with large language models (LLMs) introduces risks of contextual sensitivity to non-technical signals, including partner country identity. This paper presents a reproducible counterfactual framework for diagnosing how partner country information influences LLM compliance decisions across five instruction-tuned models and 249 jurisdictions. Across four controlled experiments, partner identity systematically alters export control classifications even when project content is held constant or removed entirely, with BRICS and GCC partners attracting consistently higher control rates than EU and Five Eyes collaborations. Under causal partner controls, classification deltas become small and statistically non-significant, suggesting that observed disparities reflect contextual cue sensitivity rather than fixed geopolitical bias, though residual associative priors cannot be fully excluded. Model-level analysis reveals qualitatively distinct behavioural patterns, including directionally opposite responses to identical inputs across architectures, indicating that model selection carries consequences beyond accuracy in compliance-sensitive deployments. A structured two-stage prompting protocol reduces decision volatility by 60–80% while preserving interpretability. The proposed framework provides an auditable diagnostic method for evaluating contextual sensitivity in regulated AI systems, with implications for the design of accountable compliance automation.

1 Introduction

Large language models (LLMs) are increasingly explored in governance and compliance settings to support high-volume, text-based risk assessment (Bommasani et al., 2021). In the university context, they may assist export control screening by extracting technical indicators and generating rationales for regulatory review, though the quality of such outputs in operational settings remains an open question. Automation in this domain may also introduce legal and ethical risks (Weidinger et al., 2022): false negatives may permit unlawful transfers, while false positives may delay legitimate collaboration and amplify geopolitical inequities. Because export control determinations depend on technical content, end use, and jurisdictional context, potential sensitivity to non-technical signals warrants careful scrutiny.

Prior work demonstrates that LLMs can support regulatory interpretation when reasoning is transparent and auditable (Hassani et al., 2024), yet existing studies rarely examine whether geopolitical or contextual indicators independently influence model judgements in legally consequential domains (Meskó and Topol, 2023). This gap is particularly consequential for export control screening, where legal validity and procedural fairness must be maintained simultaneously.

We therefore investigate the following research question:

To what extent do large language models alter export control classifications based on partner country information when technical content is held constant?

We hypothesise that (1) partner country identity systematically affects classification outcomes through contextual sensitivity rather than fixed geopolitical bias, and (2) structured prompting protocols can attenuate these effects while preserving interpretability and auditability. Addressing this question extends fairness evaluation beyond demographic attributes to geopolitical context and provides a framework for analysing model accountability in legally consequential environments.

Export control regimes are intentionally asymmetric, reflecting codified distinctions between jurisdictions, technologies, and risk categories. Accordingly, this study adopts a diagnostic rather than normative perspective, examining whether partner-related contextual cues influence LLM behaviour

Proceedings of the Second International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2026)
11–12 June, 2026.

in measurable and procedurally controllable ways while prioritising transparency over outcome parity.

To address this objective, we introduce *partner attribution*, a counterfactual evaluation framework that isolates the effect of partner country information on automated export control classifications. By comparing model outputs across controlled prompt variants while holding project content constant, the framework reveals whether geopolitical context alone alters compliance decisions and provides an interpretable diagnostic for analysing contextual sensitivity in LLM-assisted decision support.

Contributions

- We introduce *partner attribution*, a counterfactual evaluation paradigm for isolating contextual variables in LLM decision pipelines, evaluated across five models and 249 jurisdictions.
- We propose a multi-stage causal prompting protocol that separates content-based reasoning from contextual sensitivity, enabling controlled analysis of partner-induced classification shifts.
- We provide empirical analysis of bloc- and country-level asymmetries using paired statistical testing and robustness checks across models and decoding temperatures.
- We show that nationality cues alone alter classifications in the absence of task content, and we use controlled conditions to isolate identity, surface, and proxy effects.
- We assess observed sensitivities against regulatory structures and release a reproducible auditing framework for evaluating contextual sensitivity in regulated AI systems.

2 Related Work

2.1 Fairness and Bias in Algorithmic and Language Model Systems

Research on algorithmic fairness highlights trade-offs between parity and accuracy in automated decision-making (Shrestha and Yang, 2019). These concerns extend to large language models, where behavioural evaluations and quantitative bias metrics demonstrate persistent disparities that are not resolved through prompt-level interventions alone (Bai et al., 2025; Huang et al., 2023).

Recent studies show that LLMs encode geopolitical, ideological, and normative associations across contexts (Salnikov et al., 2025; Motoki et al., 2023; Zhou and Zhang, 2024; Manvi et al., 2024). Additional work documents moral and value-laden biases embedded in pretrained language models (Schramowski et al., 2022). Benchmark resources such as *BOLD* support large-scale evaluation of open-ended bias (Dhamala et al., 2021), and empirical evidence indicates that biased model outputs can influence downstream human decision-making (Fisher et al., 2025).

2.2 Regulatory Oversight and Legal Compliance Automation

Regulatory frameworks increasingly emphasise lifecycle fairness and accountability in AI systems (Zhou et al., 2022). Oversight is particularly important when LLMs operate in legally consequential domains (Meskó and Topol, 2023). Prior work shows that foundation models can assist regulatory interpretation when supported by traceable reasoning interfaces (Hassani et al., 2024). Hybrid approaches have also been proposed for compliance verification and data protection tasks (Chen et al., 2024; Feretzakis et al., 2025).

2.3 Reliability and Interpretability in Regulated NLP

Reliability and factual consistency remain central challenges for LLM deployment in regulated environments. Inference variability can produce divergent reasoning under fixed prompts (Huang et al., 2025), and diagnostic prompting methods improve interpretability without guaranteeing stability across contexts (Savage et al., 2024). Similar domain-specific bias propagation has been observed in applied NLP settings such as medical radiology (Wiggins and Tejani, 2022). Surveys of legal NLP document rapid technical progress but limited attention to contextual sensitivity and jurisdictional reasoning (Siino et al., 2025).

3 Methods

3.1 Dataset

The dataset contains fifteen UK research project descriptions collected from the UKRI EPSRC funding portal[1]. Each entry includes the fields Title, Abstract, Project Partners, Organisation, and Department, spanning physical sciences, engineering, and digital technologies. The fifteen projects serve as a controlled proof-of-concept corpus rather than

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

a representative sample of all research disclosures; because all country-level estimates derive from repeated measures across the same descriptions, project-level idiosyncrasies will propagate across all 249 country pairings and findings should be interpreted accordingly.

Each project was paired with 249 ISO-3166 partner countries. Country metadata included geopolitical bloc membership, world region classifications, and UK embargo status for the embargo alignment analysis. Results are aggregated across five blocs: BRICS (Brazil, Russia, India, China, South Africa), the Gulf Cooperation Council (GCC), the European Union (EU), the Five Eyes alliance (Australia, Canada, New Zealand, the United Kingdom, and the United States), and a residual category labelled Other. These groupings represent pre-existing geopolitical frames with substantial presence in training corpora, making them meaningful categories for detecting learned associations. They are analytical rather than regulatory constructs, and internal heterogeneity within blocs means bloc-level results should be interpreted as directional tendencies rather than uniform country-level effects; within-bloc variation for BRICS is examined in Section D.4. Each record retained stable identifiers (`project_id`, `iso3`) to ensure traceability across prompt generation, model inference, and analysis.

3.2 Prompt Construction

Prompts were constructed by combining project descriptions with experiment-specific partner information. All prompts asked the model to determine whether a research project should require export control review under UK export control regulations. Responses were required in a binary format (YES or NO) accompanied by a brief explanatory rationale.

The overall prompt template remained consistent across experiments while specific elements were manipulated to isolate different aspects of partner attribution. These manipulations included the presence or absence of international collaborators, the stage at which partner information was introduced, and the lexical representation used to describe partner identity.

An illustrative prompt structure is shown below.

Project title: Advanced quantum sensing calibration

Abstract: This project develops calibration techniques for high-precision quantum sensing instruments used in laboratory environments.

Additional partner: A public university in Germany

Question: Should this project be classified as requiring export control review? Respond with YES or NO and briefly explain your reasoning.

The exact prompt templates used across the four experiments are provided in supplementary Section C.

3.3 Models

All models were accessed via the OpenRouter API, which routes requests to underlying model providers while preserving their default alignment and safety mechanisms. These provider-side controls were treated as part of the models' operational behaviour rather than as an experimental variable.

To assess cross-model robustness, we evaluated five widely used instruction-tuned models: `gpt-4o`, `claude-sonnet-4`, `gemini-2.5-pro`, `llama-3.3-70b-instruct`, and `qwen-2.5-72b-instruct`. Provider-reported model identifiers were logged at runtime and stored in the replication package for reproducibility.

Inference parameters were set to `temperature`=0.0, `max_tokens`=300, and `seed`=42.

3.4 Model Inference

Prompts were executed through parallelised API requests with rate-limited calls and automatic retry handling. Each model processed all project–country combinations, and for each query we recorded the prompt text, model identifier, decision output, rationale, token usage, and timestamp. Responses were parsed to extract the binary classification. Each model produced approximately 7,500 evaluations in total. Deterministic runs were replicated at `temperature`=0.2 to assess robustness to stochastic decoding. The full experimental workflow is provided in supplementary Section A.

3.5 Experimental Design

The experiments were designed as a controlled behavioural study of partner attribution effects in LLM-assisted export control screening. The independent variable is the representation of partner country identity within the prompt, while the technical content of the research project remains fixed across conditions. The outcome variable is the model's export control classification decision

(YES or NO). Four experiments were conducted under progressively simplified prompt conditions, each isolating a different aspect of partner attribution.

Experiment 1 measures whether adding an international partner changes baseline project classifications.

Experiment 2 introduces a two-stage decision protocol in which the model first evaluates project content alone and then re-evaluates the project after partner information is introduced.

Experiment 3 removes all technical project content to test whether partner nationality alone influences model decisions.

Experiment 4 introduces controlled variations of partner identity to isolate the contribution of lexical identity, foreignness, and regulatory attributes.

Model outputs were aggregated to estimate decision rates and behavioural changes across experimental conditions. Flip rates were computed for baseline versus partner-added prompts, and Stage 1–Stage 2 decision differences were analysed for the two-stage protocol. Country-level prior probabilities were estimated for partner-only prompts. Statistical significance was assessed using paired McNemar tests with Benjamini–Hochberg false discovery rate correction, and bloc-level estimates were computed as the mean of country-level effects. The overall experimental logic and evaluation metric definitions are provided in supplementary Sections B and C.

3.6 Experiment 1: Baseline vs Plus-One Partner

Partner sensitivity was first measured by comparing each project's classification with and without an international collaborator. For each UK project, a baseline prompt contained domestic information only, while a paired variant appended one partner country.

Behavioural change was quantified using the **flip rate**, defined as the proportion of projects whose decisions differed between paired prompts. Uncertainty was estimated using Wilson 95% confidence intervals, and statistical significance was assessed using the procedures described above. Results were aggregated across the five geopolitical blocs defined in Section 3.1. A flip rate of zero would indicate complete classification stability, while higher rates indicate greater sensitivity to partner identity.

3.7 Experiment 2: Two-Stage Decision Protocol

To improve causal interpretability, a two-stage prompting protocol was implemented.

Stage 1: the model produced an initial export control decision and rationale based solely on project content.

Stage 2: the model reconsidered the same project with an added partner country after being shown its Stage 1 decision and reasoning.

This protocol reduces regeneration effects and isolates behavioural change attributable to partner identity. Stage 1–Stage 2 differences were analysed using the same statistical procedures as Experiment 1. Robustness was assessed via cross-temperature correlations ($T = 0.0$ vs $T = 0.2$) using Spearman's ρ over country-level flip-rate rankings.

3.8 Experiment 3: Country-Only Partner Evaluation

To test whether nationality alone influences decisions, all project content was removed from the prompts. The prompt described only *"a research project based in the United Kingdom"* followed by a partner country. This configuration measures partner-driven effects independent of technical information. Outputs were aggregated by country and bloc and analysed using the same statistical procedures described above.

3.9 Experiment 4: Causal Partner Controls

The final experiment isolates the specific contribution of partner identity. Building on Experiment 3, we compared four partner conditions across three semantically equivalent UK project frames:

1. **Country-identified:** explicit country reference (e.g., "collaborating with researchers in China"), capturing the full partner identity signal.

2. **Unspecified partner:** collaboration with an unspecified foreign institution, removing the country identity while preserving the concept of international collaboration.

3. **Paraphrased identity:** reformulated country reference (e.g., "a China-based institution"), testing whether effects arise from the specific token sequence or the underlying country concept.

4. **Proxy identity:** replacement of the country name with bloc membership and embargo status, testing whether regulatory attributes explain the observed differences.

The dataset comprised 249 countries grouped into five blocs. For each country, three semantically equivalent UK project frames were combined with four partner-identity conditions, producing approximately 3,000 queries per model.

For each country we computed YES proportions and contrasts Δ_1, Δ_2, and Δ_3 between identity conditions. Paired differences were tested using McNemar statistics with FDR correction (0.05) across countries, and bloc estimates were computed as mean country-level deltas.

4 Results

The following sections report classification outcomes across the four experiments, with results aggregated by geopolitical bloc and analysed using the procedures described in Section 3.5.

4.1 Experiment 1: Baseline vs Plus-One Partner Design

We evaluate whether adding an international partner alters export control classifications when project content remains constant.

Across models, consistent bloc-linked disparities emerge (Figure 1), indicating a shared behavioural pattern. McNemar tests confirm significant partner effects primarily for BRICS and GCC, with negligible change in EU and Five Eyes (Table 1). Replication at `temperature`=0.2 yields 93.5% agreement with deterministic runs, with moderate rank stability for GPT-4o ($\rho = 0.73$), Gemini 2.5 ($\rho = 0.68$), and Claude ($\rho = 0.48$), but near-zero correlations for LLaMA and Qwen, indicating weaker rank stability across temperature settings.

4.2 Experiment 2: Two-Stage Decision Protocol

We assess whether requiring models to revisit their prior reasoning stabilises classifications when partner information is introduced.

The two-stage protocol substantially reduces decision variation while retaining the same bloc hierarchy (Figure 2). Median flip rates decline by 60–80% across models, with BRICS and GCC remaining most sensitive and EU/Five Eyes largely stable. Cross-model agreement exceeds 80%, and

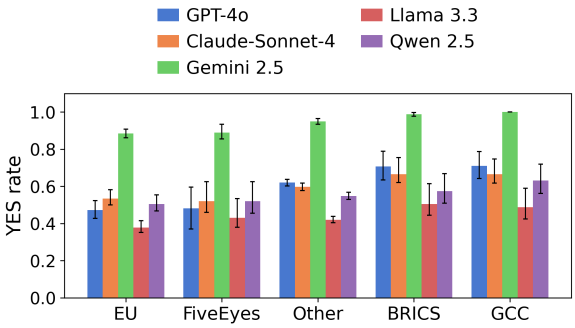

Figure 1: Bloc-level YES classification rates in the baseline experiment with constant project content. BRICS and GCC partners show the highest control rates, while EU and Five Eyes are lowest.

under mild perturbation ($t = 0.2$) correlations remain low (0.04–0.30), indicating residual model-specific priors rather than random noise.

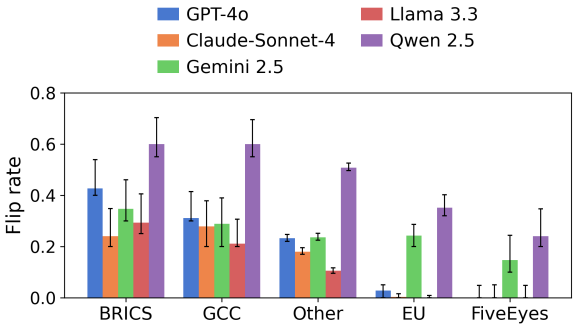

Figure 2: Bloc-level flip rates between Stage 1 and Stage 2 of the two-stage protocol. Flip rates decrease by 60–80%, while BRICS and GCC remain more sensitive than EU and Five Eyes.

4.3 Experiment 3: Country-Only Partner Evaluation

We test whether partner nationality alone elicits a controlled classification when technical content is removed. The same bloc ordering persists (BRICS > GCC > Other > EU/Five Eyes), though with lower overall YES rates (Figure 3). Persistence of these patterns without project content indicates that partner identity alone appears capable of driving model differentiation under the conditions tested, consistent with the presence of learned associative tendencies that persist independently of task content. Robustness checks show high rank consistency for Claude ($\rho = 1.00$) and LLaMA ($\rho = 0.97$), moderate consistency for GPT-4o ($\rho = 0.84$), and lower consistency for Gemini ($\rho = 0.63$) and Qwen ($\rho = 0.77$).

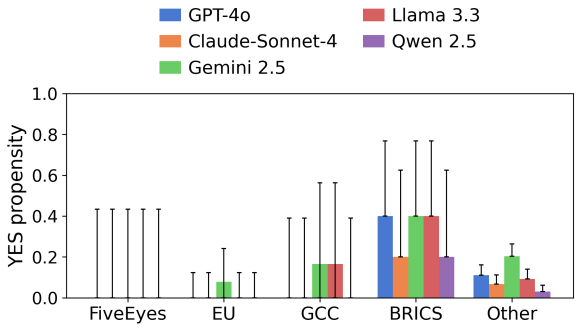

Figure 3: Bloc-level YES classification rates in the country-only experiment. The same bloc hierarchy (BRICS > GCC > Other > EU/Five Eyes) persists, indicating partner effects independent of project content.

4.4 Experiment 4: Causal Partner Controls

We reintroduced the Country-Only prompts from Experiment 3 under controlled modifications to test whether removing, paraphrasing, or proxying partner references alters model classifications. The design compares each prompt across four matched conditions (*country*, *unspecified*, *paraphrase*, and *proxy*) while holding all other content constant. Across all five models, affirmative classifications remained largely stable across conditions. Mean deltas were small (average $\Delta_1 = -0.23$, $\Delta_2 = -0.07$, $\Delta_3 = -0.08$) and none reached statistical significance after FDR correction. This indicates that explicitly naming a partner country produces no measurable additional effect beyond what is already present in the broader contextual framing, though this finding is consistent with both the absence of stable associations and the robustness of those associations to lexical variation.

Bloc-level patterns showed modest, model-specific shifts, with small reductions for EU, Five Eyes, and Other blocs and inconsistent responses for BRICS and GCC. Bloc-level mean differences are reported in supplementary Section D, showing that directional effects were small and statistically unstable across models.

4.5 Cross-Experiment Comparison

Table 1 summarises key quantitative results across the first three experiments. Progressive procedural control consistently reduces volatility, while bloc-level asymmetries remain stable, suggesting that structured prompting mitigates behavioural instability without eliminating contextual sensitivity.

The range of significant effects after FDR correction (11–70% across models) reflects substantial model-level variation. While Claude, Gemini, and LLaMA show net increases in YES classifications when a partner is introduced, Qwen 2.5 exhibits a systematic decrease across 245 of 249 jurisdictions (mean effect rate: -0.36), whereas GPT-4o remains broadly neutral. Qwen 2.5 responds to the presence of an international partner as a moderating signal rather than an escalating one, a qualitatively distinct behavioural pattern from that observed in other models. This directional divergence indicates that model-specific priors may produce not only different magnitudes of partner sensitivity but also opposite decision shifts on identical inputs, suggesting that multi-model evaluation is advisable rather than optional in compliance-sensitive deployments.

Illustrative rationale extracts showing how partner identity shapes model reasoning pathways across blocs and architectures are provided in supplementary Section E.

4.6 Alignment with UK Embargo Structure

To evaluate whether high YES-rate partners correspond to legitimate regulatory concerns, we compared the top 50 most frequently controlled partner countries per model with the UK export embargo list. As shown in Table 2, roughly half of these partners fall under formal restrictions (45.6%, 41.2%, and 40.4% for Experiments 1–3). Mean TopN YES rates decrease from 0.78 to 0.46 as contextual information is removed, indicating that structured reasoning moderates but does not eliminate alignment with formal embargo structures. Model-level variation is reported in supplementary Section D. This mixture of lawful correspondence and residual sensitivity suggests that model outputs partially align with formal regulatory structures without evidence that legal criteria are explicitly represented or applied.

5 Discussion

5.1 Interpretation

The four experiments together reveal a consistent pattern: partner country identity influences export control classifications across all conditions tested, including when project content is held constant, progressively simplified, and removed entirely. Bloc-level asymmetries persist across models and experimental conditions, with BRICS and GCC partners attracting systematically higher control rates than EU and Five Eyes collaborations.

Metric	Experiment 1	Experiment 2	Experiment 3
Mean flip rate (BRICS/GCC)	0.45–0.60	0.10–0.18	0.22–0.30
Mean flip rate (EU/Five Eyes)	0.05–0.12	0.02–0.05	0.03–0.06
% significant effects (FDR < 0.05)	15–30%	5–69%	9–14%
Spearman ρ (0.0 vs 0.2)	0.04–0.30	0.04–0.30	0.63–1.00
High-sensitivity countries	Belarus, Afghanistan, UAE	Afghanistan, Belarus, Eritrea, Laos	Afghanistan, Belarus, Eritrea, Laos
Dominant pattern	BRICS/GCC highest; EU/Five Eyes lowest	Same pattern, smaller magnitude	Pattern persists without content

Table 1: Summary of quantitative results across the first three experiments. Values represent aggregate trends across five models.

Experiment	TopN per model	Mean TopN YES rate	Mean share embargoed
Experiment 1	50	0.776	0.456
Experiment 2	50	0.687	0.412
Experiment 3	50	0.460	0.404

Table 2: Proportion of UK-embargoed countries among the top-50 YES-rate partners (averaged across models) for each experiment.

The relationship between Experiment 3 and Experiment 4 requires careful interpretation. Experiment 3 shows that models differentiate between national identities with high rank stability even in the complete absence of task content, with rank correlations reaching $\rho = 1.00$ for Claude and $\rho = 0.97$ for LLaMA, while Experiment 4 shows that removing, paraphrasing, or proxying the country name produces small, statistically non-significant deltas. Two interpretations are consistent: earlier disparities may be primarily prompt framing artefacts that dissolve once surface cues are controlled, or all four conditions may be sufficient to activate the same underlying associative structure such that lexical reformulation does not reduce the signal. What the experiments establish is that structured prompting reduces decision volatility and residual asymmetries persist across all conditions, consistent with behaviour one would expect from probabilistic associative systems rather than explicit legal reasoners, though the experiments cannot establish the underlying mechanism.

The directional divergence identified in Section 4.5, where most models increase `YES` classifications when a partner is introduced while Qwen 2.5 systematically decreases them, indicates that contextual sensitivity is not uniform across architectures, and that model selection in deployment carries consequences beyond accuracy or calibration.

5.2 Limits of Causal Attribution

Counterfactual prompting provides a useful diagnostic lens but does not fully isolate underlying causal mechanisms. Observed disparities may arise from surface framing effects, from statistical correlations between national identifiers and risk signals in training data, or from partial pattern matching to embargo-related signals in alignment procedures; the present experiments constrain these influences but cannot conclusively disentangle them. Qualitative analysis of rationale text in supplementary Section E suggests that partner identity influences not only classification outcomes but also the reasoning pathway constructed to justify them, indicating contextual cue sensitivity operating at the level of model continuation rather than post-hoc rationalisation.

Two further limitations bear on interpretation. This study measures behavioural consistency across controlled prompt conditions rather than classification accuracy against a legal standard; accuracy evaluation would require annotated project–country pairs from qualified practitioners, which do not currently exist at the scale required. Holding prompt organisation constant is necessary for controlled comparison but means that sensitivity to information arrangement cannot be fully separated from sensitivity to partner identity itself; results should therefore be interpreted as evidence of context-sensitive behaviour under structured prompting rather than proof of the absence of deeper representational bias.

5.3 Provider-Side Alignment Constraints

The reported results characterise models as deployed through their public APIs rather than as isolated foundation checkpoints. Providers may update weights, alignment procedures, or safety filters without external visibility, and moderation mechanisms cannot be experimentally separated from base model behaviour. For open-weight models such as Llama and Qwen, provider-side modera-

tion layers commonly applied to closed models are less likely to be present, and safety-influenced responses that superficially conform to the required format cannot be fully excluded. These constraints apply to any behavioural evaluation conducted through public APIs and should be considered when interpreting results across all five models examined here.

5.4 Implications for Export Control Screening

These results reveal both the potential utility and limits of LLM-assisted screening. Under constrained prompting conditions, models produce classifications and traceable rationales that may support triage functions, though operational performance in real screening environments remains untested. Partner-linked asymmetries raise considerations for procedural fairness and legal defensibility, and the two-stage protocol represents a candidate deployment approach by preserving intermediate reasoning and supporting auditability. Comparable asymmetric partner classifications appear in the US Export Administration Regulations (EAR) and EU Regulation 2021/821, suggesting the diagnostic framework extends beyond the UK context.

5.5 Broader Implications and Ethical Considerations

The partner attribution framework is domain-agnostic and may extend to analysing contextual sensitivity in LLMs deployed in high-stakes settings, offering a candidate approach for operationalising contextual robustness as a measurable behavioural property prior to deployment. LLM outputs in legally consequential environments should function as advisory signals subject to human review; embedding auditability within LLM-assisted screening remains an important safeguard against automation that displaces rather than supports expert judgement.

6 Reproducibility and Data Availability

All code, prompts, model outputs, analysis scripts, project descriptions, and supplementary material are publicly available.[2]

7 Conclusion

This study introduced and evaluated a methodological framework for assessing the reliability of large language models in regulated decision contexts, using export control screening as a case study. Across four controlled experiments, partner country identity systematically influenced export control classifications even when technical project content was held constant or removed. These disparities diminished under counterfactual partner controls, indicating behaviour consistent with sensitivity to contextual cues under structured prompting, although residual effects suggest that associative tendencies cannot be fully eliminated through procedural control alone.

More broadly, the results demonstrate how structured prompting and counterfactual experimentation can function as diagnostic tools for interpretable and auditable AI evaluation. By linking behavioural analysis with procedural safeguards, the framework offers a candidate approach for examining model reliability in legally consequential domains. Although grounded in the UK export control regime, the methodology may be adaptable to other jurisdictions, languages, and regulatory environments, though cross-jurisdictional validation remains a direction for future work.

Future work should extend this framework to expert-labelled datasets, retrieval-augmented systems, and cross-jurisdictional evaluations to better understand how contextual effects interact with domain knowledge and institutional constraints.

Acknowledgements

The authors thank Joanna MacSween, Export Control Officer at the University of Liverpool, for guidance on export control procedures.

References

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2025. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences of the United States of America*, 122(8):e2416228122.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Yifan Chen, Jian Wang, and Zhihong Li. 2024. Hybrid large language model and ontology-driven approach for automated bim compliance checking. *Buildings*, 14(9):1983.

[2] https://zenodo.org/records/20532919

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 862–872. ACM.

Georgios Feretzakis, Panagiotis Panagiotakopoulos, Christos Prasinos, Georgios Kouroupetroglou, and Ilias Maglogiannis. 2025. Gdpr and large language models: Technical and legal obstacles. *Future Internet*, 17(4):151.

Jon Fisher, Sherry Feng, Rachel Aron, Trevor Richardson, Yejin Choi, Daniel W. Fisher, and Katharina Reinecke. 2025. Biased llms can influence political decision-making. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6559–6607, Vienna, Austria. Association for Computational Linguistics.

Sina Hassani, Mehrdad Sabetzadeh, Daniel Amyot, and Jian Liao. 2024. Rethinking legal compliance automation: Opportunities with large language models. *arXiv preprint arXiv:2404.14356*.

Dong Huang, Qingwen Bu, Jie Zhang, Xiaofei Xie, Junjie Chen, and Heming Cui. 2023. Bias testing and mitigation in llm-based code generation. *arXiv preprint arXiv:2309.14345*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, ..., and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Rohin Manvi, Samar Khanna, Marshall Burke, David Lobell, and Stefano Ermon. 2024. Large language models are geographically biased. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*. PMLR.

Bertalan Meskó and Eric J. Topol. 2023. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *npj Digital Medicine*, 6(120):1–3.

Fabio Motoki, Daniel R. Vieira, Hugo Rodrigues, and Felipe Lacerda. 2023. More human than human: measuring chatgpt political bias. *Scientific Reports*, 13(15541).

Roman Salnikov, Tobias Weber, and Andrei Kurenkov. 2025. Geopolitical biases in llms: what are the "good" and the "bad" countries according to contemporary language models. *arXiv preprint arXiv:2506.06751*.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H. Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20.

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3):258–268.

Y. R. Shrestha and Y. Yang. 2019. Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems. *Algorithms*, 12(9):199.

Marco Siino, Mariana Falco, Daniele Croce, and Paolo Rosso. 2025. Exploring llms applications in law: A literature review on current legal nlp approaches. *IEEE Access*, 13:18253–18276.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Amelia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

Walter F. Wiggins and Ali S. Tejani. 2022. On the opportunities and risks of foundation models for natural language processing in radiology. *Radiology: Artificial Intelligence*, 4(4):e220119.

Di Zhou and Wei Zhang. 2024. Political biases and inconsistencies in bilingual gpt models: The cases of the u.s. and china. *Computational Communication Research*, 6(2):123–147.

Nengfeng Zhou, Zach Zhang, Vijayan N. Nair, Harsh Singhal, and Jie Chen. 2022. Bias, fairness and accountability with artificial intelligence and machine learning algorithms. *International Statistical Review*, 90(3):468–480.