Supplementary Material: 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

A Experimental Pipeline

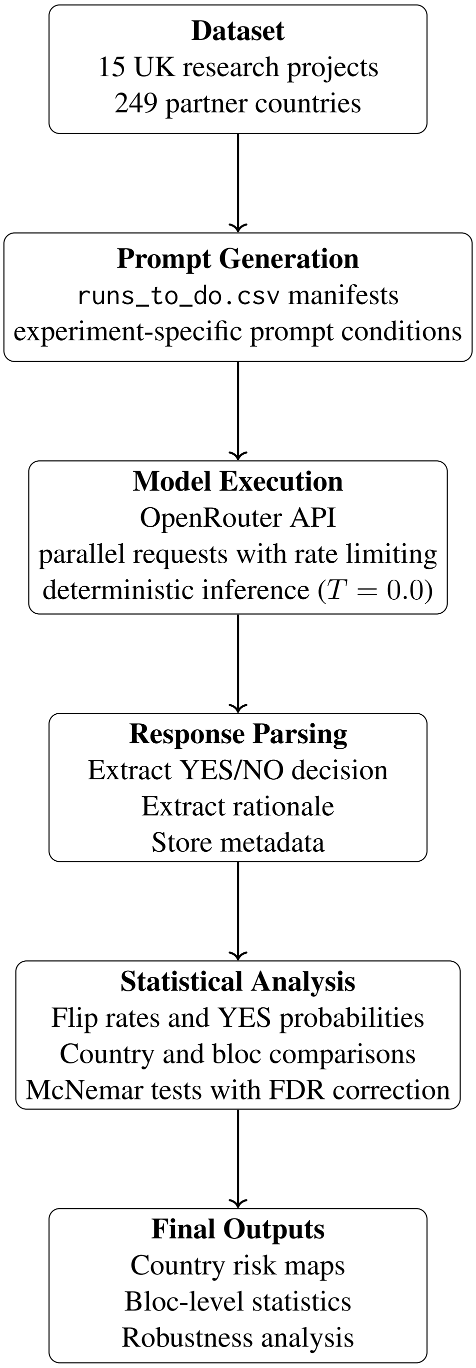


Figure 1: Overview of the experimental pipeline used across all four experiments.

Figure 1 illustrates the full experimental pipeline used in this study, including dataset construction, prompt generation, model execution, response parsing, and statistical analysis.

B Experimental Logic Diagram

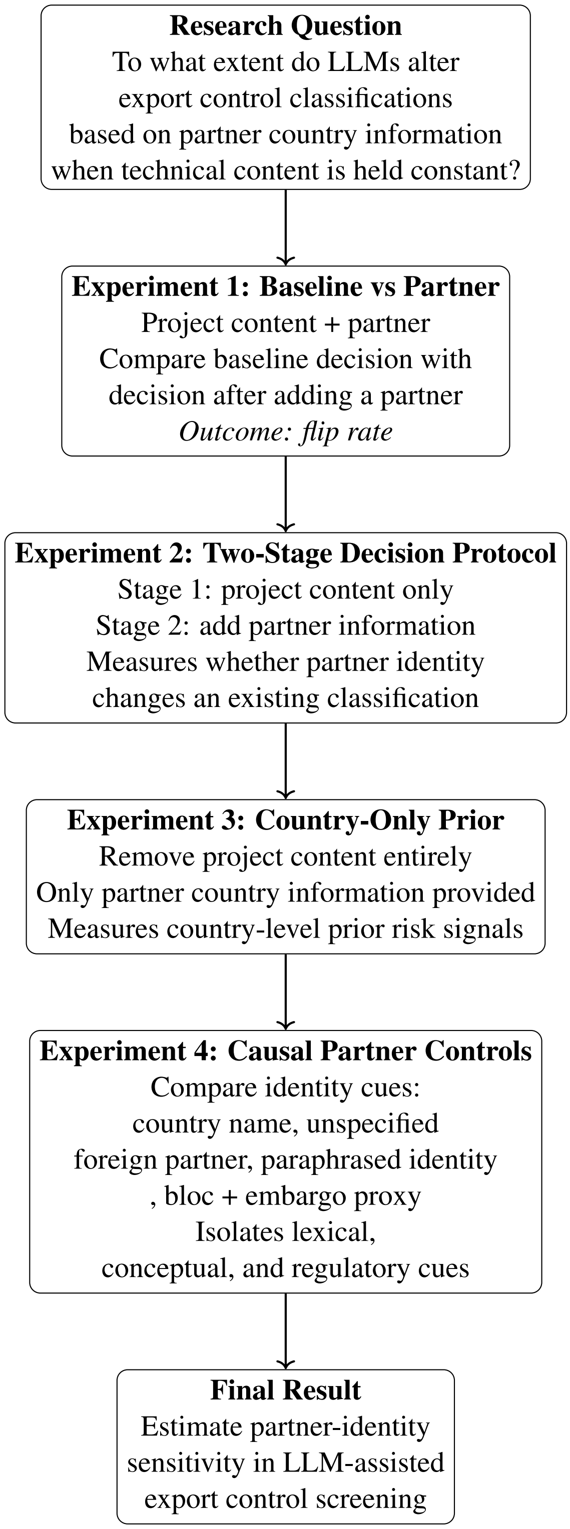


Figure 2: Experimental logic diagram showing how the four experiments progressively isolate the influence of partner country information on export control classification decisions.

Experiment 1 tests whether the presence of an international partner alters classifications relative to

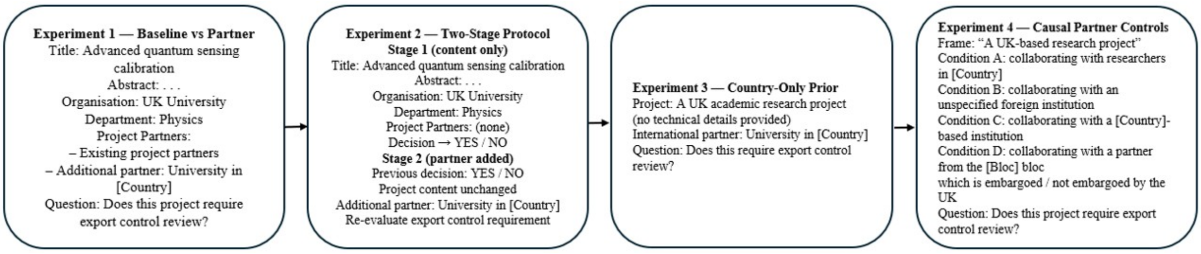


Figure 3: Prompt structures used across the four experiments. The experiments progressively manipulate the information provided about international partners while holding other prompt elements constant. Experiment 1 introduces partner countries into otherwise identical project descriptions. Experiment 2 separates technical evaluation from partner evaluation using a two-stage protocol. Experiment 3 removes project content entirely to measure country-level priors. Experiment 4 decomposes partner identity into lexical and regulatory cues to identify the mechanisms driving partner sensitivity.

a baseline project description. Experiment 2 examines whether partner identity changes decisions after an initial content-based evaluation using a two-stage protocol. Experiment 3 removes all project content to measure country-level prior risk signals associated with partner nationality alone. Experiment 4 then decomposes partner identity into lexical, conceptual, and regulatory proxy cues to identify which features drive the observed sensitivity.

C Prompt Structure Across Experiments

Figure 3 illustrates the prompt formats used across the four experiments. Each experiment modifies the information provided about international collaborators while keeping the underlying research context consistent.

Evaluation Metrics

Five metrics were used to assess contextual sensitivity and decision stability, combining measures of classification consistency with paired statistical comparison.

- **Mean control rate:** the proportion of prompts classified as requiring export control review (YES) within a condition.
- **Flip rate:** the proportion of paired prompts whose decisions differ between conditions.
- **Wilson confidence intervals:** binomial confidence intervals used to quantify uncertainty in estimated proportions.
- **McNemar paired test:** a non-parametric test for paired binary outcomes used to assess whether decision differences between prompt conditions are statistically significant.
- **Cross-temperature rank correlation:** Spearman's ρ computed between country-level rankings obtained at $T = 0.0$ and $T = 0.2$ to evaluate robustness to stochastic decoding.

D Extended Results and Supplementary Analyses

This section provides detailed per-model statistics and supplementary analyses supporting the aggregate results reported in Section 4. These materials allow inspection of model-level variability and confirm that reported trends are consistent across architectures.

D.1 Cross-Experiment Bloc Comparison

Figure 4 provides a consolidated visual reference for the cross-experiment trends discussed in Section 4. This figure confirms that relative bloc ordering remains stable even as absolute magnitudes decrease under stronger controls.

D.2 Per-Model Embargo Overlap (TopN)

Table 1 disaggregates the embargo overlap results reported in Section 4.6 by model, revealing substantial inter-model variation in both mean YES rates and the proportion of embargoed countries among the top-50 most frequently controlled partners.

D.3 Experiment 4 Bloc-Level Summary

Table 2 reports bloc-level mean deltas for the four partner-identity conditions examined in Section 4.4. Bloc-level deltas are small and inconsistent across models and control conditions, with no stable directional pattern. This supports the main experimental finding that explicit partner naming has negligible causal impact once contextual framing is con-

Table 1: Per-model mean YES rate and share of UK-embargoed countries among top-50 partner countries across experiments.

Model	Experiment 1		Experiment 2		Experiment 3	
	Mean YES rate	Share embargoed	Mean YES rate	Share embargoed	Mean YES rate	Share embargoed
GPT-4o	0.801	0.44	0.513	0.32	0.500	0.40
Claude-Sonnet-4	0.831	0.64	0.908	0.46	0.300	0.36
Gemini 2.5	1.000	0.38	0.932	0.48	0.920	0.52
LLaMA 3.3	0.536	0.46	0.700	0.48	0.440	0.42
Qwen 2.5	0.712	0.36	0.380	0.32	0.140	0.32

Notes. Each row reports results for one model. Values denote the mean YES rate among the 50 most frequently controlled partner countries for that model and the proportion of those partners subject to current UK export restrictions.

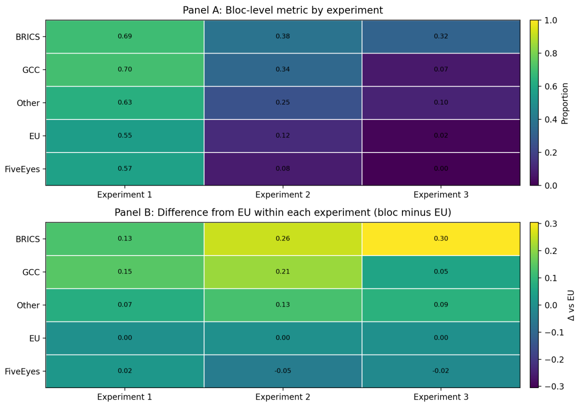


Figure 4: Aggregated bloc-level comparison across the three experiments. BRICS and GCC blocs show consistently higher control likelihoods than EU and Five Eyes across conditions, while overall magnitude decreases under stronger controls.

Table 2: Bloc-level mean deltas in YES rate for Experiment 4 (Causal Partner Controls).

Bloc	Metric	GPT-4o	Claude	Gemini	LLaMA	Qwen
BRICS	Δ_1	0.000	-0.267	0.000	0.000	0.000
	Δ_2	0.000	-0.200	0.000	0.000	0.000
	Δ_3	0.000	0.467	0.000	0.000	0.000
EU	Δ_1	-0.086	-1.000	0.000	-0.580	-0.333
	Δ_2	-0.086	0.000	0.000	0.012	-0.235
	Δ_3	-0.086	0.000	0.000	0.420	-0.333
Five Eyes	Δ_1	-0.200	-0.867	-0.133	-0.467	-0.267
	Δ_2	0.000	-0.067	-0.133	0.333	-0.067
	Δ_3	-0.200	0.133	-0.133	0.533	-0.200
GCC	Δ_1	0.000	-0.111	0.000	0.000	0.000
	Δ_2	0.000	-0.111	0.000	0.000	0.000
	Δ_3	0.000	0.222	0.000	0.333	0.000
Other	Δ_1	-0.086	-0.629	-0.032	-0.317	-0.031
	Δ_2	-0.044	-0.220	-0.011	-0.102	-0.026
	Δ_3	-0.086	-0.346	-0.032	-0.018	-0.031

Notes. Values are mean differences between the *Country-identified* condition and each control frame: Δ_1 = Country–Unspecified, Δ_2 = Country–Paraphrase, Δ_3 = Country–Proxy. Negative values indicate fewer YES classifications when a partner country is explicitly named.

trolled. Occasional larger deviations, particularly for individual models, do not generalise across architectures or conditions, indicating model-specific sensitivity rather than systematic geopolitical bias.

Overall, the table corroborates the conclusion that earlier partner effects arise primarily from contextual cues rather than fixed representational preferences.

D.4 Within-Bloc Country-Level Variation (BRICS)

Table 3 reports country-level YES rates for the five BRICS members under Experiment 1, disaggregating the bloc-level summary to address within-bloc heterogeneity. The Russian Federation consistently attracts the highest control rates across models (mean: 0.87), while South Africa produces the lowest (mean: 0.59), a gap of 0.28 within the same analytical bloc. Brazil and India occupy intermediate positions similar to South Africa, while China sits closer to Russia. This pattern is consistent with the caveat in Section **??** that bloc-level summaries reflect directional tendencies rather than uniform country-level effects, and confirms that the framework is sensitive enough to detect within-bloc variation driven by country-specific regulatory and geopolitical associations.

Table 3: Country-level mean YES rates for BRICS members in Experiment 1 (Baseline vs Plus-One Partner). Values represent the proportion of projects classified as requiring export control review when the named country is added as a partner.

Country	GPT-4o	Claude	Gemini 2.5	LLaMA 3.3	Qwen 2.5	Mean
Russian Federation	1.00	1.00	1.00	0.60	0.73	0.87
China	0.73	0.73	1.00	0.53	0.67	0.73
India	0.67	0.53	1.00	0.47	0.47	0.63
Brazil	0.60	0.53	1.00	0.47	0.53	0.63
South Africa	0.53	0.53	0.93	0.47	0.47	0.59

E Illustrative Rationale Comparison

To complement the quantitative findings, this appendix presents illustrative rationale extracts drawn from Experiment 1 using a single fixed project: *TECAN – Telemetry-Enabled Carbon Aware Networking* (University of Oxford, Department of Engineering Science). The project description remains constant across all conditions. The only variable manipulated is the identity of the international partner, drawn from different geopolitical blocs. The resulting binary classifications and rationale texts were recorded for each condition. Five extracts are presented below: two from GPT-4o, two from Qwen 2.5, and one from Claude Sonnet 4, collectively covering all five geopolitical blocs.

GPT-4o

Saudi Arabia (GCC) – YES:

> The project involves international collaboration with a university in Saudi Arabia, which may require export control review due to the transfer of potentially sensitive technology or information. Additionally, the focus on telemetry solutions and routing algorithms could involve dual-use technology, which is subject to export controls.

Poland (EU) – NO:

> The project focuses on reducing carbon emissions in computer networks, which is a civilian application. The involvement of a public university in Poland, an EU member state, does not inherently trigger export control concerns. There is no indication of military or dual-use technology development that would require export control review.

Qwen 2.5

Brazil (BRICS) – YES:

> The project involves telemetry solutions and routing algorithms for carbon-aware networking, which may include the development and deployment of technology that could be subject to export control regulations. The international collaboration with a partner in Brazil adds an additional layer of complexity, potentially requiring export control review to ensure compliance with UK export control laws.

Australia (Five Eyes) – NO:

> The project focuses on reducing carbon emissions in computer networks through telemetry and routing algorithms, which are not typically subject to export controls. The involvement of international partners does not inherently trigger export control requirements unless the technology or information shared is specifically controlled under UK or international regulations.

Claude Sonnet 4

Afghanistan (Other) – YES:

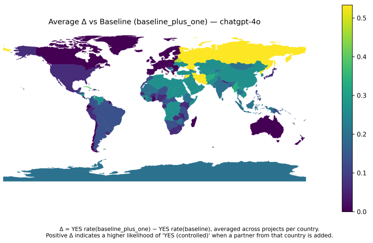
> The project involves networking technologies, telemetry systems, and routing algorithms that could have dual-use applications in telecommunications infrastructure. The inclusion of a public university in Afghanistan as a project partner raises export control concerns, as Afghanistan is subject to comprehensive sanctions and export restrictions. The combination of potentially controlled networking technologies with collaboration involving an Afghan institution requires export control review under UK regulations.

These extracts illustrate two patterns across conditions. First, identical technical content is characterised differently depending on partner identity: telemetry and routing algorithms are framed as potentially dual-use when a GCC or BRICS partner is present, but as civilian applications when paired with EU or Five Eyes collaborators. Second, divergence between models is visible in the rationale text itself. GPT-4o foregrounds partner-linked risk signals, Qwen 2.5 emphasises jurisdictional complexity, and Claude Sonnet 4 integrates both technical and regulatory considerations. These qualitative differences suggest that partner identity influences not only the classification outcomes but also the reasoning pathways constructed to justify them, a distinction relevant to auditability in compliance-sensitive deployments.

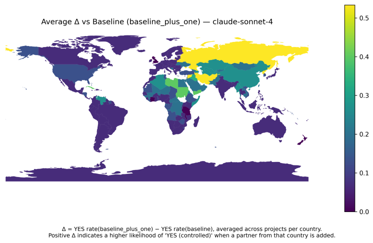
F Geographic Sensitivity Visualisations

F.1 Experiment 1 Maps

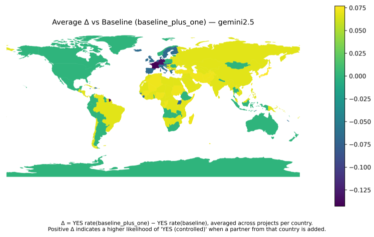
These maps show country-level changes in YES classification probability when a partner is introduced. Values represent differences between baseline and partner conditions for each country. Figure 5



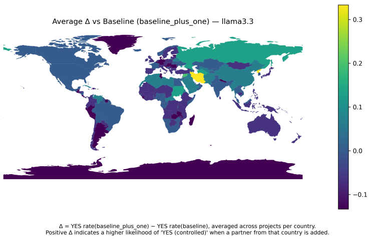
(a) GPT-4o



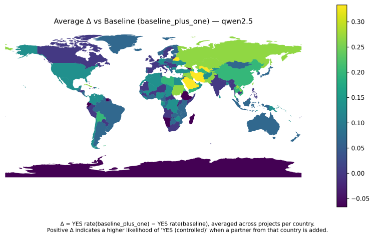
(b) Claude-Sonnet-4



(c) Gemini 2.5



(d) Llama 3.3



(e) Qwen 2.5

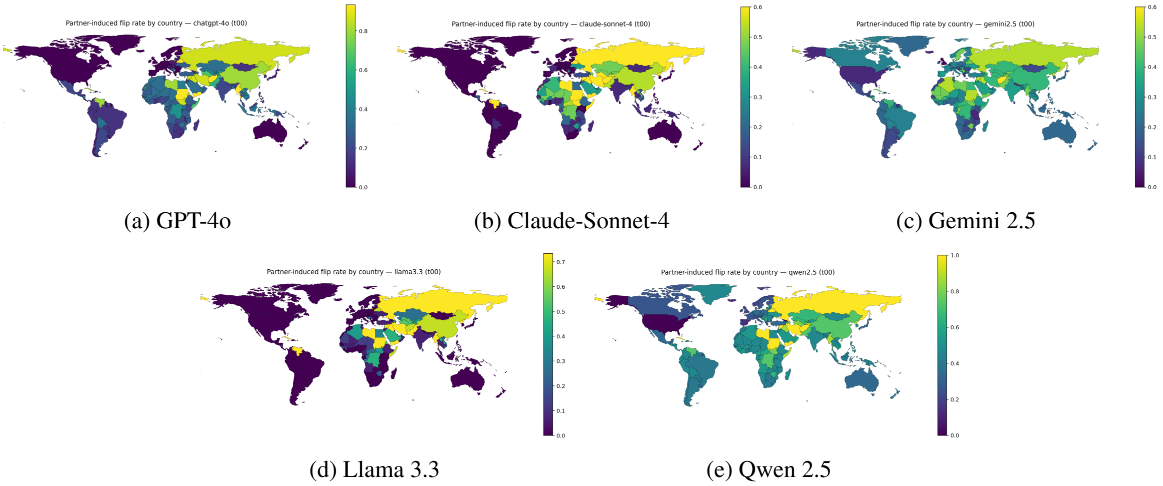
Figure 5: Geographic distribution of partner-induced increases in YES classifications in the baseline experiment. High-sensitivity clusters appear in Eastern Europe, Central Asia, and the Middle East, indicating contextual effects of partner identity.

F.2 Experiment 2 Maps

These maps show country-level flip rates between Stage 1 and Stage 2 under the two-stage protocol, indicating spatial patterns of classification volatility. Figure 6

F.3 Experiment 3 Maps

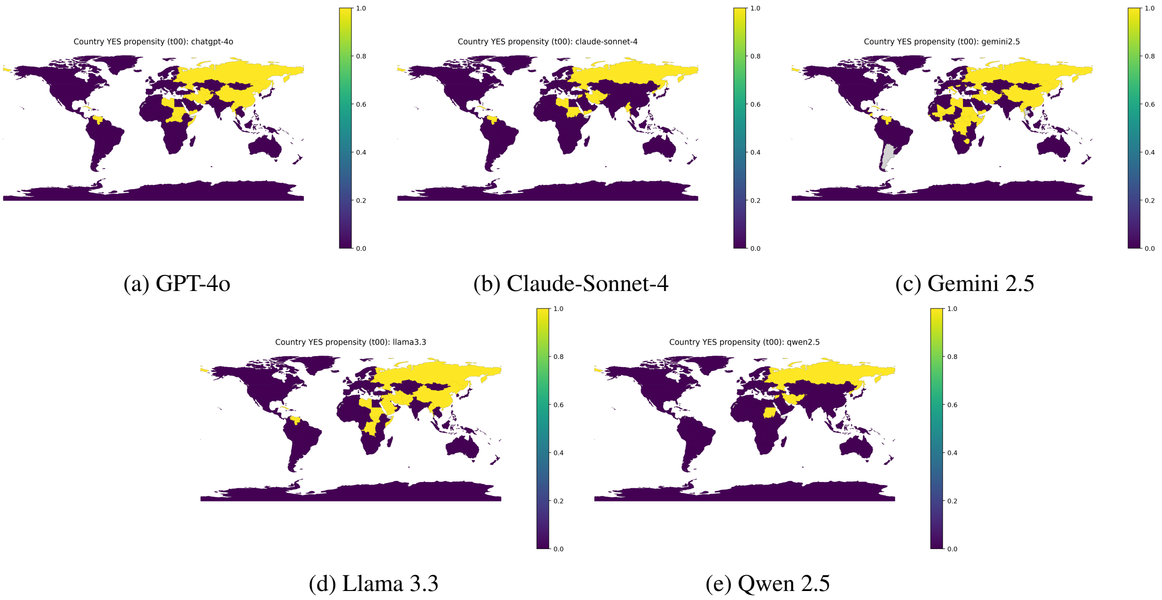
These maps show country-level YES propensities when only partner nationality is provided, enabling comparison of spatial consistency across models. Figure 7



(a) GPT-4o (b) Claude-Sonnet-4 (c) Gemini 2.5

(d) Llama 3.3 (e) Qwen 2.5

Figure 6: Geographic distribution of decision reversals under the two-stage protocol. High-sensitivity regions persist in Eastern Europe, the Horn of Africa, and Central Asia, indicating residual contextual variability.



(a) GPT-4o (b) Claude-Sonnet-4 (c) Gemini 2.5

(d) Llama 3.3 (e) Qwen 2.5

Figure 7: Geographic distribution of YES propensities in the country-only experiment. Clusters of heightened sensitivity appear in Central Asia, the Middle East, and Eastern Europe, revealing partner-linked representational structure.