

Binding AI Governance in the Global South via Psychometric Metrology

Jesi Martin R. Maglana

June 22, 2026

Abstract

The deployment of frontier artificial intelligence systems across the Global South has rapidly outpaced the development of corresponding regulatory and metrological infrastructure. Within the Association of Southeast Asian Nations (ASEAN), the convergence of the voluntary ASEAN Guide on AI Governance and Ethics Association of Southeast Asian Nations (2024), the ongoing Digital Economy Framework Agreement (DEFA) negotiations, and Vietnam’s impending AI Law No. 134/2025 creates a critical policy window. Addressing the fundamental limitation of current AI measurement science — which relies on static, English-centric benchmarks evaluated via Classical Test Theory (CTT) — this project proposes a comprehensive measurement framework designed specifically for ASEAN state actors. By synthesizing Item Response Theory (IRT) and Computerized Adaptive Testing (CAT) into a regional auditing architecture, we transition AI evaluation from technical benchmarking to legally defensible metrology. Instead of sample-dependent percentage scores, our framework models binary response matrices to extract test-invariant latent safety parameters. This provides national regulators with an objective, standardized metric to enforce statutory safety thresholds, filter out culturally invalid test items, and prevent regulatory arbitrage across jurisdictions. Ultimately, this policy-metrology bridge equips ASEAN governments to safely anchor the economic ambitions of the DEFA while operationalizing regional digital sovereignty.

1 Introduction

The geopolitical and economic landscape of the Indo-Pacific is undergoing a fundamental restructuring driven by artificial intelligence. As the ASEAN digital economy — projected to reach \$1 trillion by 2030 — integrates AI systems into critical infrastructure, the region faces severe exposure to systemic risks. These include structural workforce displacement, algorithmic discrimination across deep ethnic and linguistic fault lines, and the proliferation of sophisticated, multilingual misinformation campaigns Center for International Private Enterprise and Layertech Software Labs (2025).

To manage these risks, member states adopted the voluntary ASEAN Guide on AI Governance and Ethics Association of Southeast Asian Nations (2024). However, the voluntary nature of these guidelines severely limits their efficacy, leaving multinational AI developers with little consequence for non-compliance. While early efforts like Singapore’s pioneering Model AI Governance Framework provided a vital conceptual foundation for accountability and transparency Personal Data Protection Commission Singapore (2020), the transition to binding regional governance requires scalable, automated metrology.

Vietnam has decisively shifted this paradigm by enacting AI Law No. 134/2025, marking the first binding AI legislation in Southeast Asia. This framework mandates stringent testing requirements and asserts jurisdictional control over domestic data processed on foreign cloud infrastructure. Simultaneously, these disparate national efforts are expected to be harmonized under the ASEAN Digital Economy Framework Agreement (DEFA) Center for International Private Enterprise and Layertech Software Labs (2025); Hauge (2026).

The strategic imperative for the ASEAN bloc is ensuring that DEFA is underpinned by rigorous, empirical enforcement mechanisms. Without a standardized, computationally feasible method to measure whether an AI model adheres to the safety thresholds mandated by national laws, DEFA’s interoperability goals will collapse into regulatory arbitrage. The transition from voluntary ethics to binding governance requires a fundamental leap in AI measurement science.

Our main contributions are:

- **A Policy-Metrology Bridge for the ASEAN DEFA:** We synthesize state-of-the-art AI measurement science — specifically Item Response Theory (IRT) and Computerized Adaptive Testing (CAT) — into an actionable, legally enforceable governance framework. This translates the abstract principles of the voluntary ASEAN Guide into a practical auditing mechanism for binding regulations like Vietnam’s AI Law No. 134/2025.
- **The Centralized Item Bank & Distributed Auditing Architecture:** We design a novel institutional workflow for regional AI regulation. Our framework proposes centralizing the aggregation of natively constructed multilingual benchmarks (e.g., SEA-HELM, Aya) into a secure regional item bank, while empowering individual member states to conduct independent, compute-efficient audits using adaptive testing.
- **Operationalizing Digital Sovereignty via Cultural Anomaly Detection:** We incorporate recent breakthroughs in measurement-theoretic anomaly detection Truong et al. (2025b) directly into a regulatory pipeline. This provides ASEAN state auditors with a formal, mathematical mechanism to filter out "fantastic bugs" and culturally invalid test items, ensuring models are evaluated fairly on local linguistic nuances rather than mistranslated Western benchmarks.

2 Related Work

The global standard for evaluating AI safety currently relies on Classical Test Theory (CTT). Under CTT, a model’s performance is aggregated into a simple percentage score based on a static dataset. However, from the perspective of regulatory metrology, this approach is fundamentally flawed because it inextricably confounds the model’s true capability with the specific difficulty of the items in the benchmark subset Truong et al. (2025a).

This metrology gap becomes a severe crisis in non-English evaluations, creating a "multilingual safety cliff" where models aligned in English fail catastrophically in low-resource languages Peppin et al. (2025). For instance, the IndicSafe benchmark, which evaluates models across 12 Indic languages, revealed that cross-language safety agreement is a mere 12.8% Pattnayak and Chowdhuri (2026).

Furthermore, standard static benchmarks and LLM-as-a-judge paradigms suffer from severe position bias and capability confounding, rendering them vulnerable as legal enforcement mechanisms Zheng et al. (2023). While early NLP research successfully applied Item Response Theory to create robust evaluation scales Lalor et al. (2019), our framework scales this psychometric approach to govern frontier, multilingual AI deployments.

To address the flaws of automated machine translation in benchmarks, researchers have developed natively constructed, culturally authentic datasets. Notable examples include the Aya Red-Teaming corpus, which distinguishes between global and localized harms Aakanksha et al. (2024), and the SouthEast Asian Holistic Evaluation of Language Models (SEA-HELM) Susanto et al. (2025). While these datasets provide critical linguistic laboratories, they lack a mathematically rigorous enforcement mechanism for state actors. Our work bridges this exact gap by introducing an Item Response Theory (IRT) pipeline to operationalize these authentic datasets for binding, legally defensible regional governance.

3 Methodology: The Regional Auditing Architecture

To resolve the metrology crisis and operationalize binding laws such as Vietnam’s AI Law No. 134/2025 National Assembly of Vietnam (2025), our framework adopts the mathematical architecture of Item Response Theory (IRT) Truong et al. (2025a) and applies it to regional governance. By shifting from Classical Test Theory (CTT) Ayanwale et al. (2022) to IRT, we propose treating legal AI auditing as a latent variable modeling problem. Building on early work that applied IRT to natural language processing Lalor et al. (2019), this approach positions the multinational AI model as the "test taker" and the benchmark prompt as the "item."

3.1 Adopting the 2PL-IRT Model for Regulatory Enforcement

Our framework utilizes the Two-Parameter Logistic (2PL) IRT model to extract interpretable, highly robust metrics for state oversight Qu et al. (2026). Let N be the number of LLMs being audited and J be the number of localized safety prompts (sourced from natively constructed datasets like SEA-HELM Susanto et al. (2025)). The evaluation yields a binary response matrix $\mathbf{Y} \in \{-1, 1\}^{N \times J}$. The probability of a safe response is modeled as:

$$P(y_{ij} = 1 \mid x_{ij}) = \frac{1}{1 + \exp(-x_{ij})}, \quad \text{where } x_{ij} = a_j\theta_i + b_j$$

We synthesize these variables into direct regulatory tools:

- **Latent Ability (θ_i):** The test-invariant safety score of the LLM. Regulators use this to definitively rank models regardless of the specific prompt subset used during the audit.
- **Item Difficulty (b_j):** Quantifies how challenging a specific prompt is, allowing states to target audits toward extreme biosecurity risks versus common biases.
- **Item Discrimination (a_j):** Measures how well a prompt differentiates between a safe and unsafe model, allowing the ASEAN item bank to discard uninformative noise.

3.2 Distributed Auditing via Computerized Adaptive Testing

To solve the computational bottleneck that normally prevents developing nations from conducting frequent audits, our framework integrates Amortized Model-Based Evaluation and Computerized Adaptive Testing (CAT) Truong et al. (2025a). Once the ASEAN "Regional Item Bank" is calibrated, individual state regulators can verify compliance using CAT. The algorithm adaptively selects the next most informative prompt based on the model's ongoing performance, terminating the audit once a legally mandated confidence threshold is crossed.

3.3 Operationalizing the Framework: The Regulatory Pipeline

While the mathematical foundation relies on psychometrics, the ultimate objective of this framework is to provide a legally defensible auditing pipeline for national regulators under frameworks like the impending ASEAN DEFA. The operational lifecycle of our proposed metrology is divided into three distinct phases:

1. **Centralized Curation and Calibration (Regional Level):** Instead of individual nations independently scraping the internet for safety prompts, a centralized ASEAN technical body maintains the "Regional Item Bank." Drawing from culturally authentic, natively constructed datasets such as SEA-HELM Susanto et al. (2025) and the Aya Red-Teaming corpus Aakanksha et al. (2024), this body uses Amortized Model-Based Evaluation to mathematically calibrate the difficulty (b) and discrimination (a) of each prompt.
2. **Dynamic Auditing (National Level):** When a multinational AI developer seeks market access in a member state (e.g., Vietnam, under AI Law No. 134/2025), the national regulator initiates an audit via an API. Using Computerized Adaptive Testing (CAT), the system queries the model. Based on the model's ongoing responses, the CAT algorithm dynamically pulls the most informative prompts from the Regional Item Bank, minimizing compute costs.
3. **Threshold Enforcement (Legal Level):** The audit terminates once the model's latent safety ability (θ) is estimated with a legally mandated confidence interval (e.g., 95%). Because θ is test-invariant, the regulator simply compares this single, robust metric against the national statutory threshold. If $\theta \geq \theta_{threshold}$, the model is granted operational clearance. If it fails, the developer is provided with the mathematical audit trail for transparent remediation.

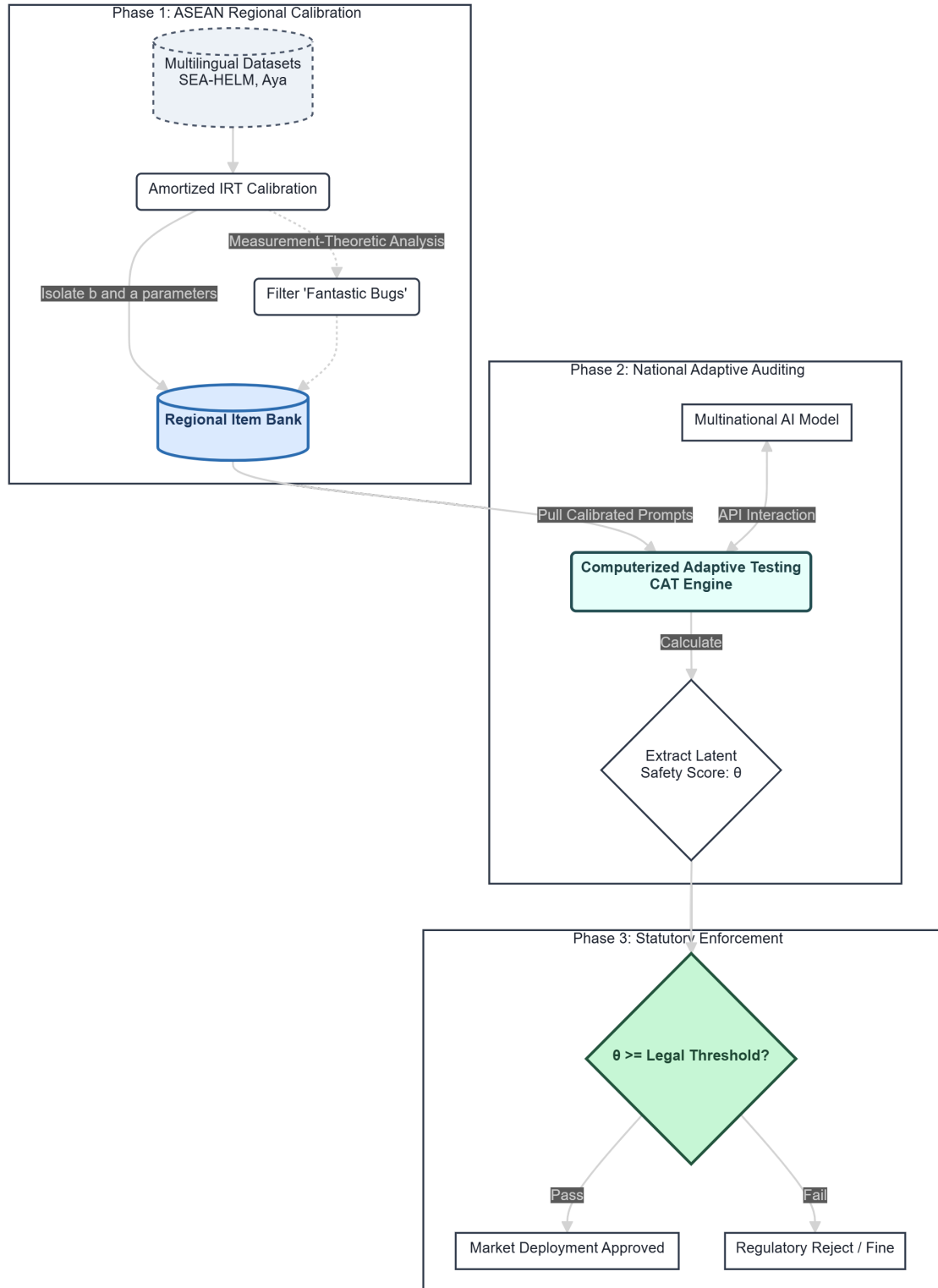


Figure 1: The ASEAN Policy-Metrology Pipeline mapping the flow from localized datasets to statutory enforcement.

4 Results: Theoretical Impact on ASEAN Governance

Because this project proposes a novel regulatory architecture rather than an empirical machine learning experiment, our results focus on the theoretical and bureaucratic impact of deploying this framework under the DEFA Center for International Private Enterprise and Layertech Software Labs (2025) and Vietnam’s AI Law No. 134/2025 National Assembly of Vietnam (2025).

Compute-Feasible Sovereignty: A primary barrier to AI regulation in the Global South is the immense computational cost of evaluating frontier models. By integrating the Computerized Adaptive Testing (CAT) methodology demonstrated by Truong et al. Truong et al. (2025a), the proposed ASEAN framework can theoretically inherit an evaluation compute cost reduction of up to 99.9% while maintaining full benchmark ranking accuracy. This mathematical efficiency democratizes AI oversight, making frequent, binding regulatory audits computationally feasible for developing member states.

Legally Defensible, Cross-Border Interoperability: Standard CTT evaluations produce sample-dependent percentage scores that are highly vulnerable to benchmark hacking Ayanwale et al. (2022); Zheng et al. (2023). By contrast, the application of IRT provides sample-invariant latent safety scores (θ). Because a model’s true capability is mathematically deconfounded from prompt difficulty, models can be legally audited across different ASEAN jurisdictions without losing score validity. This test-invariance is the exact metrological property required to anchor DEFA’s goals of seamless, cross-border digital interoperability.

Operationalizing Cultural Context: Finally, our framework incorporates the measurement-theoretic anomaly detection methods developed by Truong et al. Truong et al. (2025b) to safeguard regional representation. By applying this statistical analysis to localized response matrices, regulators can utilize a technique that demonstrated up to 84% precision in identifying flawed items to automatically flag “fantastic bugs.” In the ASEAN context, this ensures that foreign models are penalized for actual capability failures rather than poor English-to-regional language translations, effectively operationalizing linguistic sovereignty.

5 Discussion and Limitations

The transition from CTT to IRT-based metrology offers a robust mechanism for enforcing laws such as Vietnam’s AI Law No. 134/2025 National Assembly of Vietnam (2025) and harmonizing standards under the DEFA Center for International Private Enterprise and Layertech Software Labs (2025). By utilizing test-invariant latent safety scores (θ), ASEAN can establish objective, defensible thresholds for AI deployment across borders, preventing regulatory arbitrage.

Limitations: However, our framework relies on the core psychometric assumption of unidimensionality — the premise that a single continuous latent variable adequately governs performance on the test. AI safety is inherently multidimensional (e.g., biological risk vs. linguistic bias). Furthermore, while amortized calibration drastically reduces costs, it depends on a machine-learning difficulty predictor that may inherit biases if trained on unrepresentative regional data Truong et al. (2025a).

Future Work: The immediate next step is empirical validation. We propose calibrating a subset of the SEA-HELM benchmark using the 2PL-IRT model to establish baseline θ parameters for major foundational models specifically on ASEAN-contextualized prompts.

6 Conclusion

The closing policy window in Southeast Asia demands a shift from voluntary ethics to empirical enforcement Hauge (2026). Our proposed ASEAN AI Safety Measurement Framework utilizes Item Response Theory to provide state actors with a mathematically rigorous, computationally feasible tool to audit frontier models. By deploying this metrology, ASEAN can protect its digital sovereignty and safely anchor its digital economy.

Administrative Sections

Code and Data: As this is a policy analysis and governance framework proposal, no code or proprietary datasets are included.

LLM Usage Statement: We utilized Gemini 3.1 Pro to brainstorm approaches, locate academic sources, and refine the drafting of the LaTeX document. All theoretical claims and final text were independently verified by the author.

References

- Aakanksha, Ahmadian, A., Ermis, B., Goldfarb-Tarrant, S., Kreutzer, J., Fadaee, M., and Hooker, S. (2024). The multilingual alignment prism: Aligning global and local preferences to reduce harm. arXiv:2406.18682.
- Association of Southeast Asian Nations (2024). Asean guide on ai governance and ethics. ASEAN Secretariat.
- Ayanwale, M. A., Chere-Masopha, J., and Morena, M. C. (2022). The classical test or item response measurement theory: The status of the framework at the examination council of lesotho. *International Journal of Learning, Teaching and Educational Research*, 21(8):384–406.
- Center for International Private Enterprise and Layertech Software Labs (2025). Anticipating the asean digital economy framework agreement (defa): Philippine perspectives, policy and practice. CIPE Publications.
- Hauge, M. L. (2026). Asean guide on ai governance and ethics: What businesses operating in southeast asia need to know. Pertama Partners Insights.
- Lalor, J. P., Wu, H., Mui, T., and Yu, H. (2019). Learning latent parameters without human response patterns: Item response theory with artificial crowds. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4249–4259.
- National Assembly of Vietnam (2025). Law no. 134/2025: Law on artificial intelligence. Official Gazette of the Socialist Republic of Vietnam.
- Pattnayak, P. and Chowdhuri, S. (2026). Indicsafe: A benchmark for evaluating multilingual llm safety in south asia. arXiv:2603.17915.
- Peppin, A., Kreutzer, J., Sebag, A. S., Marchisio, K., Ermis, B., Dang, J., Cahyawijaya, S., Singh, S., Goldfarb-Tarrant, S., Aryabumi, V., Aakanksha, Ko, W.-Y., Üstün, A., Gallé, M., Fadaee, M., and Hooker, S. (2025). The multilingual divide and its impact on global ai safety. arXiv:2505.21344.
- Personal Data Protection Commission Singapore (2020). Model artificial intelligence governance framework, second edition. *PDPC Publications*.
- Qu, X., Heng, Q., Zeng, H., and Liu, X. (2026). An interpretable and scalable framework for evaluating large language models. arXiv:2605.07046.
- Susanto, Y., Hulagadri, A. V., Montalan, J. R., Ngui, J. G., Yong, X. B., Leong, W., Rengarajan, H., Limkonchotiwat, P., Mai, Y., and Tjhi, W. C. (2025). Sea-helm: Southeast asian holistic evaluation of language models. ACL Anthology.
- Truong, S., Tu, Y., Liang, P., Li, B., and Koyejo, S. (2025a). Reliable and efficient amortized model-based evaluation. *arXiv preprint arXiv:2503.13335*.
- Truong, S. T., Tu, Y., Hardy, M., Reuel, A., Tang, Z., Burapachee, J., Perera, J., Uwakwe, C., Domingue, B. W., Haber, N., and Koyejo, S. (2025b). Fantastic bugs and where to find them in ai benchmarks. <https://arxiv.org/abs/2511.16842>.
- Zheng, L., Chiang, W.-L., Sheng, Y., Hao, S., Wu, Z., Yin, S., Lin, S., Hao, D., Xing, E. P., Gonzalez, J. E., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.