

**Task Team Lead:** Agustín Samano

**Region:** Malaysia (pilot) → Southeast Asia (extension)

**Timeline:** US Benchmark January 2026; Malaysia Pilot February 2026; Multi-Country Dataset June 2026

## 1 Motivation

Fiscal policy is one of governments' most powerful levers for macroeconomic stability and private-sector development. Yet for most emerging markets we lack the cornerstone input required for credible analysis: **consistent and comparable measures of exogenous fiscal shocks<sup>1</sup>**.

The United States is the only country where this has been done systematically, through the “narrative approach” pioneered by Romer and Romer (2010). Their method uses historical documents—budget speeches, economic reports, legislative records—to identify why taxes or spending changed, and to distinguish exogenous shocks from policy actions responding to the business cycle.

Replicating this approach manually is costly. For low- and middle-income countries, it has simply never been feasible.

**Recent advances in Large Language Models (LLMs)** make this possible for the first time. These models can read long historical documents, interpret motivations, and extract structured information at scale—opening the door to producing robust fiscal shock series for developing countries.

This project proposes to build a **validated, scalable LLM pipeline** for fiscal shock identification, starting with Malaysia and extending across Southeast Asia.

## 2 Novelty and Innovation

### A. A US-Trained LLM Core

The project begins by training and benchmarking an LLM on the US narrative corpus with Romer & Romer's original tax-shock labels. This ensures:

- we anchor the method in a gold-standard dataset,
- we quantify model accuracy against known results,
- we avoid region-specific overfitting and ensure robustness.

This step is essential for **credibility**: the model must reproduce known US results before we trust it elsewhere.

### B. A Scalable Pipeline for Emerging Markets

We adapt the US-trained model to Malaysia, where English-language archives are extensive. The pipeline automatically:

1. extracts narrative episodes where fiscal policy is discussed,

---

<sup>1</sup>See for example Mertens and Ravn (2013) and Romer and Romer (2010)

2. classifies motivations using Romer & Romer's categories,
3. identifies actual tax or spending changes, their timing, and magnitude,
4. flags potential political or cyclical influences, and
5. integrates human expert review at critical points.

Once validated on Malaysia, the pipeline extends to Indonesia, Thailand, the Philippines, and Vietnam, with built-in multilingual translation capability.

### **3 Project Outputs (June 2026)**

#### **A. New Data Assets**

- Narrative fiscal episode datasets for 5 Southeast Asian economies.
- Shock-event datasets identifying exogenous tax and spending changes: timing, size, and motivation.
- Harmonized multi-country panel suitable for macro and micro analysis.

#### **B. Analytical Products**

- Macro impulse responses estimated with local projections.
- Firm-level effects (investment, employment) using modern LP-DiD methods (Dube et al., n.d.).
- A methodological paper on LLM-assisted narrative identification, similar to (Romer and Romer, n.d.).

#### **C. Tooling and Capacity**

- A fully documented, open-source pipeline (R-based).
- Country-specific LLM prompts, rules, and evaluation protocols.
- A tested architecture that can be extended to Africa, South Asia, and Latin America.

### **4 Implementation Plan**

#### **Phase 0 – US Benchmark (Jan 2026)**

Reconstruct the US narrative corpus used by Romer and Romer (2010); train LLMs to identify fiscal acts, classify motivations, and extract timing/magnitude of shocks. Validate by testing against known labels.

#### **Phase 1 – Malaysia Pilot (Feb 2026)**

Deploy the pipeline, adapt models to local documents, and complete the first full narrative and shock-event datasets.

#### **Phase 2 – Southeast Asia Scaling (June 2026)**

Extend to Indonesia, Thailand, the Philippines, and Vietnam, incorporating expert review and news-based cross-checks.

## 5 Risk Management

Risk	Mitigation
Weak or uneven archives	Combine parliamentary, budget, and press sources; clearly document gaps.
LLM misclassification	US benchmarking + targeted human review; conservative thresholds for exogeneity.
Translation quality for non-English sources	Dedicated translation stage; multilingual adaptation; manual review of samples.
Political bias in source documents	Independent cross-validation with contemporaneous news archives.

## 6 Strategic Value for the World Bank

1. Transforms a previously infeasible task into a reproducible method for potentially dozens of client countries.
2. Provides a missing input for macroeconomic modeling enabling better fiscal forecasting and policy advice.
3. Offers the first systematic estimates of firm-level impacts of fiscal policy in Southeast Asia using modern causal methods.
4. Establishes the World Bank as a pioneer in responsible, auditable LLM use for economic analysis.
5. Creates a transferable framework that can serve as a global public good for future fiscal reform programs.

## Conclusion

This project brings together frontier methodology, practical policy relevance, and global scalability. The US-benchmarked LLM pipeline ensures methodological rigor; the Malaysia pilot provides a clear proof of concept; and the multi-country rollout fills a large and longstanding data gap.

## References

- Dube, Arindrajit, Daniele Girardi, Òscar Jordà, and Alan M. Taylor. n.d. *A Local Projections Approach to Difference-in-Differences Event Studies*.
- Mertens, Karel, and Morten O. Ravn. 2013. “The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States.” *American Economic Review* 103 (4): 1212–47.
- Romer, Christina D, and David H Romer. 2010. “The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks.” *American Economic Review* 100 (3): 763–801.
- . n.d. *A Narrative Analysis of Postwar Tax Changes*.