Scaling Narrative Fiscal Shock Identification with LLMs

Concept Note

January 22, 2026

Task Team Lead: Agustín Samano

Country/Region: United States (training), Malaysia (pilot), Southeast Asia (extension)

Timeline:

- US training / benchmarking: January 2026
- Malaysia pilot: February 2026
- Core multi-country dataset: June 2026

1 Introduction and Motivation

Fiscal policy plays a central role in macroeconomic stabilization, long-run growth, and private-sector development. Yet we lack credible and comparable **exogenous fiscal shock series** for most emerging markets. Existing international evidence is heavily skewed toward the United States and a few OECD countries, due to the **high cost** of constructing narrative datasets like those of Romer and Romer (2010) or Mertens and Ravn (2013).

Romer and Romer's narrative analysis of postwar US tax changes provides act-by-act classifications of tax legislation, their motivations (spending-driven, countercyclical, deficit-driven, long-run), and the timing and size of revenue effects. These labels are exactly the kind of supervised signal that modern language models can learn from.

The rise of **Large Language Models (LLMs)** offers a practical pathway to extend the narrative approach systematically to developing countries. Governments now publish large volumes of digitized legislative documents, budget speeches, and parliamentary transcripts, often in multiple languages. Manually processing these sources is infeasible; LLMs can drastically reduce the bottleneck **if** they are trained and validated against high-quality human benchmarks first.

This project proposes to:

1. **Train and benchmark an LLM-assisted classifier on the US** using the Romer & Romer tax shock dataset and underlying narrative sources as labels and text.
2. **Adapt and validate this trained system to Malaysia** as a pilot country.
3. **Scale to Indonesia, Thailand, the Philippines, and Vietnam**, combining LLMs, multilingual translation, and human validation.

The outputs will be:

1. A **US-based benchmark module**: LLM classifiers and extraction routines that can reproduce US narrative tax shocks from raw documents.
2. A **narrative episode dataset** for each SEA country based on executive, legislative, and press documents.
3. A **shock-event dataset** representing actual changes in fiscal liabilities or expenditure paths, classified into the Romer and Romer (2010) motivation categories.
4. A set of **macro and firm-level impulse responses** to fiscal shocks using local projections and modern difference-in-differences methods (LP-DiD).

These will constitute the first systematic narrative fiscal shock series for Southeast Asia, backed by an explicit US benchmark.

2 Research Questions

The project focuses on four core questions:

1. **US Benchmarking / LLM Training:** Can an LLM trained on the **US narrative corpus** and Romer & Romer's tax-shock labels reliably:
   - detect fiscal acts,
   - classify their motivations, and
   - recover the timing and size of tax shocks closely enough to reproduce known US fiscal multipliers?
2. **Measurement in Emerging Markets:** Conditional on a successful US benchmark, can we construct reliable narrative-based fiscal shock series for Southeast Asian economies using the same LLM architecture, multilingual translation, and human validation?
3. **Validation and Transportability:**
   - How closely do LLM-based classifications align with (i) expert assessments and (ii) contemporaneous news coverage in each country?
   - Are the identified exogenous shocks orthogonal to short-run macro conditions?
   - How much degradation (if any) occurs when moving from US to Malaysia and from Malaysia to other SEA contexts?
4. **Impact:** What are the dynamic effects of tax and spending shocks on:
   - GDP, investment, FDI, employment, revenue, and expenditure?
   - Firm-level outcomes such as investment and employment?

3 Proposed Approach

We structure the work in **three phases**, with a deliberate **Phase 0 on the US** to de-risk the entire pipeline.

Phase 0: US Training and Benchmarking (2025)

Objective: Use the US as a sandbox to train and stress-test the LLM pipeline before touching emerging market data.

Data

- Romer & Romer's narrative dataset of postwar US federal tax actions (timing, size, motivation, present-value paths).
- Underlying narrative sources:
  - Economic Report of the President
  - Budget of the United States Government
  - Treasury Annual Reports
  - Presidential speeches and statements
  - Congressional reports and debates

Tasks

1. **Corpus reconstruction and alignment**
   - Reconstruct the narrative corpus used in Romer and Romer (2010).
   - Align each tax act in the Romer & Romer dataset with the source passages used to classify its motivation and revenue effects.
2. **LLM-based Act Detection (Model A)**
   - Train models to detect **candidate fiscal acts** and relevant passages in long documents (e.g., ERP chapters, budget speeches).
   - Use Romer & Romer's list of 50 significant postwar tax actions as positive labels and randomly sampled irrelevant passages as negatives.
3. **LLM-based Motivation Classification (Model B)**

   Using **Romer & Romer's four-way motivation**: spending-driven, countercyclical, deficit-driven, long-run/structural,
   - Fine-tune or instruction-tune a modest-size LLM (suitable for local hardware) on:
     - input: narrative passage(s) around the tax change,
     - output: motivation label + structured explanation.
   - Embed **best-practice rules for narrative work** (real-time sources, explicit criteria, documentation) from Romer & Romer's 2023 Presidential Address on the narrative approach.
4. **Information Extraction for Timing and Magnitude (Model C)**
   - Train models (or rule-based + LLM hybrid) to extract:
     - quarter of change in liabilities,
     - present-value quarter,
     - revenue impact (baseline and present value).
   - Compare extracted values with Romer & Romer's act-quarter series.

5. **US Back-Testing**
   - Re-estimate US tax multipliers using only **LLM-generated** episodes and shocks:
     - aggregate, exogenous tax changes à la Romer and Romer (2010),
     - dynamic effects via local projections.
   - Benchmark whether:
     - sign and timing match the original estimates,
     - magnitudes are within acceptable bands (e.g., ±20–30%).
   - This gives a **quantitative yardstick** for "good enough" LLM performance.

Deliverable of Phase 0: an **audited, documented LLM toolkit** that can (i) reconstruct the US series reasonably well and (ii) be ported to other countries with known limitations.

Phase 1: Malaysia Pilot (2025)

Malaysia is the optimal starting point due to extensive English-language archives and relatively stable political institutions.

Data Sources

- Parliamentary transcripts, Hansards
- Budget speeches and fiscal policy statements
- Bills and explanatory memoranda
- Prime Minister / Minister of Finance speeches
- Official press releases
- Secondary validation archives (international/local newspapers)

Narrative Identification Strategy

The project distinguishes two levels of analysis:

1. **Narrative Episodes (Talk)** All instances where policymakers discuss tax or spending actions. Output: a comprehensive episode-level dataset with motivation labels.
2. **Shock Events (Action)** Subset of episodes that result in measurable changes to fiscal liabilities/expenditures (legislated or executive). Output: a shock-event dataset for econometric analysis (timing, size, tax base, motivation).

LLM Pipeline (Human-in-the-Loop)

We reuse the **US-trained models** as the backbone and adapt them to Malaysia.

Step 1 — Ingestion & Parsing Automatic collection and segmentation of documents into candidate narrative episodes, using the Act Detection model trained in Phase 0 and adjusted to Malaysian document formats.

Step 2 — Translation LLM-based translation for non-English texts into English, with:

- consistent terminology for fiscal concepts,
- retention of uncertainty / qualifications in the original text.

Step 3 — Motivation Classification (US-initialized)

Each episode is classified into Romer & Romer's four motivation categories, using the US-trained classifier as a prior:

- Spending-driven
- Countercyclical
- Deficit-driven
- Long-run/structural

The LLM produces:

- a proposed label,
- a short justification using explicit criteria (e.g., "responding to current macro weakness" vs "paying for a new program"),
- a confidence score,
- an "inconsistency flag" comparing **stated motives** (in-text) to **inferred motives** (based on macro context).

Step 4 — Domain Adaptation & Human Validation

- Draw stratified samples by:
  - label,
  - model confidence,
  - time period,
  - type of document.
- Local experts review and re-label:
  - obvious misclassifications (e.g., misreading of local institutional language),
  - ambiguous cases (e.g., mixed motivations).
- Use these corrections to:
  - update prompts and decision rules,
  - optionally fine-tune the model on Malaysia-specific examples (few-shot adaptation).

Step 5 — Shock Identification

- Cross-walk narrative episodes with:
  - actual legislation,
  - implementation regulations,
  - budget tables.
- Code whether and when **tax/spending liabilities** actually change:
  - change in liabilities (quarter, size, tax base),
  - present value of announced changes,
  - exogenous vs endogenous classification (based on motivation and timing, as in Romer and Romer (2010)).

Phase 2: Extension to Indonesia, Thailand, the Philippines, Vietnam (2025–2026)

- Apply the same pipeline, with:
  - country-specific corpus discovery and scraping,
  - language adaptation (translation and terminology),
  - targeted human validation in each country.
- Leverage **transfer learning**:
  - US ⟶ Malaysia serves as the main calibration jump,
  - Malaysia ⟶ other SEA countries leverages experience with mixed English/local language environments and similar fiscal institutions.

4 Validation

The project embeds a rigorous, multi-method validation strategy, now explicitly in **two layers**: US and SEA.

1. US Benchmark Validation

- Compare LLM-derived US shocks to Romer & Romer's original series:
  - confusion matrix for motivation labels,
  - distribution of timing errors (quarter misalignment),
  - revenue size errors (absolute and relative).
- Re-run baseline US macro regressions using:
  - local projections for GDP, investment, and unemployment,
  - exogenous tax changes only.
- Require that impulse response functions are **qualitatively identical** and **quantitatively close** to original published estimates before deploying the model to SEA.

2. Expert Review in SEA

- Engage World Bank economists (including KL office) and local experts.
- Evaluate a stratified sample of episodes and shocks in each country.
- Document disagreements systematically to refine classification rules and prompts.

3. News-Based Cross-Validation

- Use institutional access (e.g. ITAM, WB) to global and regional newspaper archives.
- Assess whether contemporaneous reporting aligns with classified motivations, especially for:
  - sharp tax hikes or cuts,
  - politically sensitive episodes,
  - "borderline" exogenous vs endogenous cases.

4. Internal Consistency Checks

- Test correlation of "exogenous" shocks with real-time macro indicators (GDP, unemployment, inflation).
- Exogenous shocks should show weak contemporaneous correlation with short-run macro conditions; if not, re-examine classification.

This two-tier validation (US + SEA) turns the US into a **truth-benchmark** and SEA into the **test of portability**.

5 Empirical Analysis

1. Macro-Level Effects

Use **local projections** to estimate dynamic responses of:

- GDP growth
- Private investment
- FDI
- Employment
- Government revenue and expenditure

Separate estimates will be produced for:

- tax vs spending shocks,
- personal vs corporate income tax shocks (where identifiable),
- exogenous vs endogenous shocks.

2. Firm-Level Effects

Using firm-level panels (e.g. Orbis, national administrative data where available), estimate how fiscal shocks affect:

- investment,
- employment,
- capital formation.

Identification:

- **LP-DiD (Local Projections DiD)** following Dube et al. (2022), which unifies local projections with clean-control difference-in-differences in staggered treatment settings.
- Treatment assignment based on exposure (e.g., sector-level tax incidence, import/export orientation, foreign ownership).

This will generate the first micro-based fiscal multipliers for these countries.

6 Outputs and Deliverables

By mid-2025 (US Benchmark)

- Reconstructed US narrative corpus aligned with Romer & Romer tax acts.
- Trained and documented LLM modules:

  - act detection,
  - motivation classification,
  - timing and magnitude extraction.
- Quantitative comparison of LLM-based vs original US tax multipliers.

By December 2025 (Malaysia Pilot Completion)

- Fully operational R-based and LLM-assisted pipeline for Malaysia.
- Narrative episode dataset (Malaysia).
- First version of the shock-event dataset (Malaysia).

By early February 2026 (Pilot Validation)

- Expert- and news-based validation for a subset of Malaysian episodes and shocks.
- Technical documentation and reproducible code for the Malaysia pipeline.

By June 2026 (Core Dataset Completion)

- Narrative and shock-event datasets for: Malaysia, Indonesia, Thailand, the Philippines, Vietnam.
- Harmonized multi-country episode and shock datasets.
- Fully validated motivation classification framework and LLM toolkit.
- Public dissemination-ready code and documentation (subject to licensing/permissions).

Post-2026 (Analytical Outputs)

- Macro-level impact study of tax and spending shocks in SEA.
- Firm-level impact study using LP-DiD.
- Methodological paper on **LLM-assisted narrative identification**, emphasizing:
  - US benchmark reproduction,
  - cross-country transportability,
  - human-in-the-loop design.
- Data package for public release (subject to permissions).

7 Risks and Mitigation

Risk	Mitigation
Incomplete/uneven archives	Combine multiple sources; transparently document coverage; adjust sample windows.
Translation inaccuracies	Dedicated translation stage; manual checks; standardized prompts and glossaries.
LLM misclassification	US-based benchmarking; human-in-the-loop refinement; conservative thresholds for exogeneity.
Domain shift US → SEA	Explicit diagnostics on model confidence, label drift, and validation error; country-specific fine-tuning where needed.

Risk	Mitigation
Political bias in narrative documents	Cross-check with news; inconsistency flags; robust alternative classifications.
Weak identification for some countries	Use Malaysia as a strong pilot; downweight uncertain periods; exploit LP-DiD with clean controls.

8 Policy Relevance and Value Added to the World Bank

- Fills critical data gaps for fiscal-policy modeling in major middle-income economies.
- Provides **scalable infrastructure** for narrative shock identification, with a **US-validated LLM core** exportable to other regions (Latin America, South Asia, Africa).
- Enhances understanding of **how fiscal policy affects private investment and firm behavior**, central to the Bank's structural reform and growth agenda.
- Supports better macroeconomic forecasting and policy advice in client countries by plugging a key missing input: credible exogenous fiscal shocks.

The project also showcases responsible, auditable use of LLMs for applied economic research—grounded in best-practice narrative methods —rather than opaque black-box automation.

References

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.