Continuous Narrative Identification: Evidence from Tax Policy*

Tom Pesso†

January 28, 2026

Abstract

This paper develops a scalable and transparent framework for narrative identification that systematizes the use of qualitative evidence in empirical macroeconomics. The framework—Source-Augmented Narrative Identification (SANI)—combines semantic retrieval, structured summarization, and large language models to extract and classify policymakers' stated motivations from contemporaneous policy documents. SANI replaces binary narrative classifications with a continuous measure of exogenous support for each policy action and introduces a marginal narrative contribution procedure that traces model outputs to specific source passages and allows for placebo diagnostics to assess textual grounding. I validate the framework in the canonical setting of Romer and Romer (2010). First, when applied to the original narrative summaries, SANI reproduces the authors' classifications and their macroeconomic implications. Second, using a newly constructed dataset of U.S. federal tax measures drawn from annual budget documents, I show that SANI identifies exogenous tax shocks whose macroeconomic effects closely match those implied by the Romer–Romer series. As an application, I construct a new narrative dataset of French value-added tax (VAT) changes from 1960 to 2024 and show that exogenous VAT increases are moderately but persistently contractionary, with a short-lived rise in inflation reflecting tax pass-through.

*Preliminary draft. I am deeply grateful to Geert Mesters for his invaluable guidance and to Davide Debortoli for helpful comments and advice. I am grateful to Barbara Rossi, Luca Fornaro, and Hugo Reichardt for insightful discussions, and to the participants of the CREi internal seminar for their valuable feedback. This research was supported by the Scientific Computing Core Facility (MELIS-UPF), which generously provided access to computational resources. All remaining errors are my sole responsibility.

†Universitat Pompeu Fabra (UPF), Barcelona School of Economics (BSE). Contact: tom.pesso@upf.edu

Introduction

Evaluating macroeconomic policy is fundamental to understanding the sources of aggregate fluctuations, yet it remains empirically challenging because policy actions typically respond endogenously to economic conditions. The narrative approach, pioneered by Friedman and Schwartz (1963) and applied systematically to fiscal policy by Romer and Romer (2010), addresses this problem by using contemporaneous documentary evidence to distinguish policy actions driven by long-run objectives from those responding to short-run economic developments. This approach remains essential for when policymakers' intentions must be inferred from qualitative sources, and recent work emphasizes its continued relevance for empirical macroeconomics (Romer & Romer, 2023).

Yet the method faces a practical limitation: identifying and classifying policy motives requires extensive manual reading of archival material, making the procedure costly, difficult to replicate, and often limited to rare and large-scale policy interventions.[1] As a result, existing narrative series typically cover only a narrow subset of policy actions, leaving routine budgetary measures and granular tax instruments largely unexplored. Because qualitative evidence continues to expand, through digital archives, newly released historical records, and the increasing complexity of policy documents, the scalability of narrative identification has become a central bottleneck for its wider use.

This paper develops a scalable and transparent framework for narrative identification in large text corpora. The approach—*Source-Augmented Narrative Identification* (SANI)—integrates semantic retrieval, structured summarisation, and large language model (LLM) classification to automate the construction of narrative policy shocks while preserving the interpretive features that give narrative methods their credibility. SANI operates in three steps. First, a semantic retrieval operator identifies the text segments most relevant to establishing the government's stated motivations for each policy change, replacing manual document search with a systematic similarity measure. Second, an LLM synthesises the retrieved segments into concise summaries of the policy's stated motivations. Third, a classification model maps these summaries into narrative categories—such as exogenous or endogenous—according to decision rules tailored to the identification problem.[2]

A key challenge in scaling narrative identification is ensuring that automated outputs remain *grounded* in the underlying documentary evidence. LLMs can produce plausible classifications even when they underweight or ignore the text they are given, raising concerns about transparency and researcher oversight. To address this issue, I develop a diagnostic that quantifies the dependence of the model's classifications on the underlying documentary evidence, drawing on recent work in input-sensitivity analysis (Cohen-Wang et al., 2024; Feng et al., 2018; Prabhakaran et al., 2019). The resulting *randomized source-dependence* procedure measures how the model's output varies when individual retrieved passages are masked. By estimating each passage's *marginal narrative contribution*, the procedure identifies which pieces of evidence drive the model's classification and implements a placebo-based test to detect potential ungroundedness: cases in which the model assigns substantial weight to semantically irrelevant passages.[3] While not a guarantee against hallucination, the diag-

[1] As a result, narrative shock series typically exhibit extreme sparsity: roughly 80 percent of observations are zeros, with shocks concentrated in a few years; see Barnichon and Mesters (2025).

[2] The method was originally developed under the name RANI, emphasizing its link to retrieval-augmented architectures (RAG). I adopt the term SANI to reflect more clearly its reliance on contemporaneous documentary sources. For a recent survey of retrieval-augmented methods, see Brown et al. (2025).

[3] I refer to the overall framework as the *marginal narrative contribution* procedure, and to the per-passage weights $\hat{w}_\ell$ as *marginal source contributions*.

nostic ensures that the resulting classifications remain traceable to specific documentary sources and facilitates disciplined human validation.[4] Although implemented here within SANI, the framework is general and can be applied to any setting in which researchers rely on open-weight LLMs.

SANI produces a continuous measure of exogenous support for each policy change, constructed by aggregating the marginal narrative contributions that quantify how strongly each retrieved passage favors an exogenous versus endogenous interpretation. This extends the standard narrative design, in which shocks arise from binary researcher-coded classifications. By allowing the degree of exogeneity to vary smoothly with the underlying textual evidence, the framework avoids the discontinuities introduced by hard thresholds and yields continuous narrative shocks that preserve the conceptual structure of narrative identification while improving coverage, robustness, and econometric precision.

The paper provides two empirical applications. The first is a comprehensive external validation based on the canonical narrative tax series of Romer and Romer (2010). Using only the narrative material contained in their replication archive, I show that large language models can recover the original narrative classifications with high accuracy and a strongly conservative error structure, rarely misclassifying endogenous policy actions as exogenous. Crucially, when the resulting classifications are used to construct tax shock series, they preserve the macroeconomic implications of the original Romer–Romer shocks: estimated impulse responses closely track the benchmark both qualitatively and quantitatively. Extending the analysis, I construct a comprehensive dataset of routine federal tax changes drawn from the annual budget process covering more than 1,000 measures that fall outside the scope of the original Romer–Romer narrative sample, and show that augmenting the Romer–Romer shock series with SANI-identified exogenous measures increases effective sample size and tightens confidence intervals without materially altering estimated responses.

The second application constructs a new narrative dataset of French value-added tax (VAT) policy from 1960 to 2024. The VAT is the backbone of consumption taxation in advanced economies and accounts for more than 30 percent of global tax revenues. Yet despite its central role, the macroeconomic effects of VAT policy have received far less systematic attention than income or corporate taxes. Most existing studies rely on cross-country panel regressions or micro-level natural experiments, leaving open the question of how VAT changes affect aggregate dynamics within a single economy. Understanding these effects is increasingly important. VAT adjustments represent a key policy lever for many governments facing fiscal pressures, and although the United States is one of the few advanced economies without a federal VAT, a broad-based consumption tax features prominently in current debates on long-run fiscal sustainability.[5]

France offers a uniquely valuable setting for constructing a long-run narrative series of VAT policy. As the first country to introduce a modern VAT in 1954, it provides the longest continuous historical record for studying this instrument.[6] The dataset assembled here draws from the annual *Rapports généraux* of the Senate budget committee which provide detailed explanations of the motivations behind each proposed tax measure. Using SANI, I retrieve the relevant passages for every VAT change, generate structured summaries of the stated motivations, and classify each measure under the framework of Cloyne (2013). Exogenous changes are motivated by long-run structural, ideologi-

[4] Human validation on subsamples remains important as measurement error arising from systematic misclassification from the LLMs can bias downstream econometric estimates; see Ludwig et al. (2025) and Battaglia et al. (2024).

[5] As Mankiw emphasized in his 2025 Feldstein Lecture, debates over long-run debt sustainability naturally raise the question of whether the United States should adopt a federal VAT. At present, the United States relies instead on state and local retail sales taxes, which apply only to final consumption and do not operate as a credit-invoice VAT.

[6] Denmark adopted a VAT in 1967; Appendix E.1 lists adoption dates across countries.

cal, or institutional considerations; endogenous changes respond to contemporaneous macroeconomic conditions such as demand stabilization, debt management, or the financing of new expenditures.

Finally, using the identified exogenous VAT shocks, I estimate their dynamic macroeconomic effects. A one-percent-of-GDP VAT increase generates a moderate but persistent contraction: real GDP falls by about three percent at the trough and recovers only gradually, while inflation exhibits a brief spike reflecting the mechanical pass-through of the tax increase. The magnitude of the output response is similar to that found in Romer and Romer (2010) for broad-based tax hikes, despite VAT being conceptually less distortive. One interpretation is that the short-lived rise in inflation following a VAT increase may constrain monetary accommodation, amplifying its contractionary effect.

Taken together, the paper develops an identification framework that brings narrative methods to scale while preserving their interpretive foundations. By integrating semantic retrieval, structured summarisation, grounding diagnostics, and continuous shock construction, SANI enables systematic use of large documentary corpora for causal inference. The approach broadens the empirical reach of narrative identification and makes possible the analysis of granular fiscal policy domains, such as VAT, for which manual coding has long been infeasible.

Literature Review

Identification in Macroeconomics. The paper first situates itself within the literature on policy identification, which includes empirical strategies relying on reduced-form VARs identified through timing, sign, or other model-based restrictions (Blanchard & Perotti, 2002; Sims, 1980; Uhlig, 2005). A distinct line of work instead employs external instruments—variables correlated with the policy shock of interest but orthogonal to other structural innovations—to achieve identification (Ramey, 2016; Stock & Watson, 2018). Narrative approaches provide a primary methodology for constructing such instruments, using contemporaneous evidence on policymakers' stated motives. Foundational examples include Friedman and Schwartz (1963) for monetary policy and Hamilton (1985) for oil shocks, with subsequent applications (Ramey, 2011; Ramey & Shapiro, 1998; Ramey & Zubairy, 2018; Romer & Romer, 1989, 2004, 2010) establishing narrative identification as a leading empirical tool. A key limitation of these series is their extreme sparsity: as documented by Barnichon and Mesters (2025), typical narrative shock measures contain roughly 80 percent zeros, which substantially weakens statistical power. Their estimator proposes an econometric remedy for this power–credibility trade-off, but the underlying sparsity reflects, in part, the difficulty of constructing narrative shocks with limited coverage of smaller policy actions.

This paper contributes to this tradition by developing a scalable and transparent framework for extracting and classifying policymakers' stated motives from large document collections and for constructing narrative shocks that vary continuously with the underlying textual evidence. By grounding classifications in specific documentary passages and mapping them into a graded measure of exogenous support, the framework broadens coverage and enables more granular analysis than traditional binary narrative designs.

Text and AI in Economic Analysis. A growing body of work uses text as an empirical input in economics, enabling systematic measurement of beliefs, expectations, and policy intent. Recent surveys formalize core methodological tasks—document similarity, concept detection, relational inference, and text–metadata mapping—and emphasize the importance of transparent links between textual sources

and algorithmic outputs (Ash & Hansen, 2023; Gentzkow et al., 2019). The diffusion of large language models (LLMs) has broadened the scope of text-based analysis (Jamilov, 2025; Martell, 2025) but heightened concerns about validation and grounding and its impact on downstream estimated parameters (Battaglia et al., 2024; Carlson & Dell, 2025; Ludwig et al., 2025). This paper contributes to this literature by providing a structured mapping from policy documents to model-generated summaries through semantic retrieval and marginal source contribution, and by adapting a placebo-style validation test from applied econometrics to assess whether model outputs rely on the supplied context.

Recent research also applies text and AI methods to macroeconomic identification. In monetary policy, Aruoba and Drechsel (2025) extract shocks from Federal Reserve documents using natural-language representations, while Hansen and Kazinnik (2023) show that GPT 4 can extract Romer and Romer (2023) monetary policy shocks, reinforcing the feasibility of LLM-based structured classification as used in the final stage of this paper. In fiscal policy, LLMs have been used to detect policy announcements in unstructured text and relate them to high-frequency market reactions (Amodeo & Briganti, 2025; Bandeira et al., 2025). This paper contributes to this strand by developing a framework that, starting from known policy events, extracts and classifies policymakers' stated motives from historical sources, thereby extending narrative identification to large document collections.

Fiscal Policy Transmission and Tax Composition. A final strand of literature concerns the macroeconomic transmission of fiscal policy. Seminal VAR-based analyses (Auerbach & Gorodnichenko, 2012; Blanchard & Perotti, 2002; Mountford & Uhlig, 2009) show that exogenous fiscal expansions raise activity, with multipliers varying across instruments and business-cycle conditions. Narrative studies of government spending shocks (Ramey, 2011; Ramey & Zubairy, 2018) highlight the role of expectations, while analogous work on taxation (Cloyne, 2013; Mertens & Ravn, 2012, 2013; Romer & Romer, 2010) shows that tax multipliers depend on timing, composition, and economic context. Within this broader literature, value-added taxation (VAT) has received far less attention despite its centrality to modern tax systems and its distinctive transmission channel through consumer prices. Existing evidence comes primarily from event and panel studies (Dabla-Norris & Lima, 2023; Riera-Crichton et al., 2016) and micro-level quasi-experiments (Benzarti & Carloni, 2019; Benzarti et al., 2020), which shed light on incidence but provide limited evidence on aggregate dynamics. This paper contributes by constructing a long-run, single-country narrative series of VAT policy changes for France— the first nation to adopt the modern VAT—allowing for a systematic analysis of its macroeconomic effects comparable to the benchmark income-tax literature.

The remainder of the paper is structured as follows. Section 1 outlines the motivation for narrative identification and the conceptual framework for interpreting policy shocks. Section 2 presents the SANI methodology. Section 3 introduces the marginal narrative contribution procedure and the associated placebo-based grounding tests. Section 4.1 documents the construction of the French tax dataset and the underlying textual corpus. Section **??** applies SANI to value-added tax (VAT) policy, constructs the corresponding narrative shock, and estimates its macroeconomic effects. Section **??** validates the classification step by replicating Romer and Romer (2010). The final section concludes.

1 Identifying Policy Shocks: Motivation and Framework

A central objective of this paper is to recover the structural innovations that enter the government's policy rule. Any legislated policy action reflects both systematic responses to the government's information set and an unsystematic component orthogonal to contemporaneous macroeconomic conditions. Formally, for each policy measure j enacted in period t,

$$\Delta p_{j,t} \;=\; f(\Omega_t) \;+\; \varepsilon_{j,t}, \tag{1}$$

where $\Delta p_{j,t}$ denotes the legislated change in the policy instrument, $f(\cdot)$ captures systematic behaviour, and $\varepsilon_{j,t}$ is the structural shock of interest.

Following the narrative tradition, it is useful to decompose the policymaker's information set into two components:

$$\Omega_t \;=\; \Omega_t^{(1)} \cup \Omega_t^{(2)}. \tag{2}$$

The first component, $\Omega_t^{(1)}$, consists of observable macroeconomic indicators—output, inflation, unemployment, fiscal aggregates—that rationalize predictable policy responses. The second, $\Omega_t^{(2)}$, consists of qualitative statements of motivation found in contemporaneous documents. These sources record the government's own rationale: often long-run, institutional, or political considerations that do not map neatly into standard macro variables.

The traditional narrative approach

Conventional narrative identification isolates $\varepsilon_{j,t}$ by determining which elements of $\Delta p_{j,t}$ reflect $\Omega_t^{(2)}$ rather than $\Omega_t^{(1)}$. In practice, this involves two implicit thresholding steps. First, only policy actions exceeding a minimum importance threshold $\bar{s}$ are retained:

$$\mathbf{1}\{s_{j,t} \geq \bar{s}\}.$$

This truncation is driven by feasibility: the cost of manually processing each measure scales roughly linearly with the number of actions, making comprehensive coverage prohibitive in historical datasets.

Second, a policy action is classified as exogenous when its latent narrative motivation $\theta_{j,t}$ exceeds a conceptual boundary $\bar{\theta}$:

$$\mathbf{1}\{\theta_{j,t} \geq \bar{\theta}\}.$$

Formally, the latent variable $\theta_{j,t}$ represents a structural component of the policy decision: the share of the legislated action attributable to motivations orthogonal to the contemporaneous macroeconomic conditions summarized in $\Omega_t^{(1)}$. Traditional narrative studies implicitly rely on this structural object, but observe it only through a binary judgment based on the documentary record. The conceptual boundary $\bar{\theta}$ is not observable or formally defined; rather, it reflects an implicit cutoff applied when researchers deem the stated motives sufficiently exogenous.

The resulting narrative shock is

$$\varepsilon_{j,t}^{(\text{narrative})} \;=\; \mathbf{1}\{s_{j,t} \geq \bar{s}\} \cdot \mathbf{1}\{\theta_{j,t} \geq \bar{\theta}\} \cdot \Delta p_{j,t}. \tag{3}$$

While highly influential, this binary construction discards (i) all smaller measures, regardless of their motivations, and (ii) the information contained in the relative strength or balance of motives within each policy action. These limitations stem from the practical constraints of manually interpreting heterogeneous policy material, which make sharp but informally defined thresholds unavoidable in traditional applications.

A continuous narrative identification

The approach developed in this paper replaces these discrete classifications with a continuous measure of how strongly the documentary record supports an exogenous interpretation of each policy action. Instead of deciding whether a measure's stated motives exceed an implicit and unspecified boundary $\bar{\theta}$, the method quantifies the degree of exogenous versus endogenous support directly from the underlying documents.

For each policy measure (j, t), SANI extracts the stated motivations and assigns two marginal narrative contributions:

$$S_{j,t}^{\text{endo}}, \qquad S_{j,t}^{\text{exo}},$$

which summarise how strongly the documentary record aligns with endogenous or exogenous reasoning. These calibrated quantities lie on a common scale and are therefore directly comparable.

The exogenous share is defined as:

$$P_{j,t}^{\text{exo}} = \frac{S_{j,t}^{\text{exo}}}{S_{j,t}^{\text{endo}} + S_{j,t}^{\text{exo}}},$$

a continuous analogue of the latent structural variable $\theta_{j,t}$. By construction, $P_{j,t}^{\text{exo}}$ increases when the stated motivations more strongly support an exogenous interpretation of the policy action. Throughout the paper, $P_{j,t}^{\text{exo}}$ is interpreted as an empirical estimate of $\theta_{j,t}$—that is, as a continuous measure of the fraction of the policy action attributable to motivations orthogonal to contemporaneous macroeconomic conditions. In contrast to the traditional approach, which infers only whether $\theta_{j,t}$ exceeds an unspecified conceptual boundary, the measure constructed here offers a quantitative estimate of this structural component based on calibrated documentary evidence.

This continuous representation preserves the core logic of narrative identification—using stated motives to distinguish exogenous from endogenous actions—while avoiding sharp thresholds and making use of all available narrative evidence rather than discarding information through binary classification.

Narrative-weighted structural shocks

I construct the structural shock associated with policy measure (j, t) by scaling the legislated change by its exogenous narrative share. Given the interpretation above, a natural estimate of the structural shock is obtained by multiplying the observed policy action $\Delta p_{j,t}$ by the estimated structural motivation share $P_{j,t}^{\text{exo}}$. This mirrors the logic of traditional narrative identification: actions motivated entirely by exogenous considerations identify $\varepsilon_{j,t}$ directly, whereas actions reflecting mixed motives identify only the corresponding fraction of the shock. Importantly, this construction does not require specifying or estimating the systematic component $f(\Omega_t)$; as in classical narrative approaches, identification is achieved through variation in stated motives rather than through a parametric policy rule. However,

as described later in the paper, the narrative outputs generated by SANI can also be incorporated into the estimation of a text-augmented parametric policy rule, providing additional structure for understanding systematic policy behaviour.

The shock is therefore defined as:

$$\varepsilon_{j,t}^{(\text{SANI})} \;=\; P_{j,t}^{\text{exo}} \cdot \Delta p_{j,t}. \tag{4}$$

Equation (4) is a continuous generalisation of the traditional indicator-based narrative shock. When $P_{j,t}^{\text{exo}} \in \{0, 1\}$, the construction nests the classical case. More generally, the shock varies smoothly with the documentary evidence, allowing for partial exogeneity and mixed motivations within a single policy action.

This formulation preserves the core logic of narrative identification, relying on stated motives to isolate the unsystematic component of policy, while avoiding the sharp and informally defined thresholds inherent in manual classification. It retains the sign and magnitude of the actual legislated change, incorporates all measures rather than only large ones, and weights each action according to the degree of exogenous support documented in contemporaneous sources.

The remainder of the paper describes how the exogenous shares $P_{j,t}^{\text{exo}}$ are recovered using the Source-Augmented Narrative Identification (SANI) framework and applies the resulting shock series to the empirical analysis that follows.

2 Source-Augmented Narrative Identification (SANI)

This section introduces the Source-Augmented Narrative Identification (SANI) framework, which applies narrative identification through a three-stage procedure: (i) semantic retrieval of the text segments most relevant to each policy measure, (ii) structured summarisation of their stated motivations, and (iii) classification of these motivations into researcher-defined narrative categories. Figure 1 summarises the procedure.

To fix ideas, consider the major VAT reform proposed in the 1977 French budget. This example illustrates the logic of the procedure and the nature of the outputs produced at each step and will be developed in detail in the following subsection. The next subsection first introduces the formal framework: it defines the retrieval ($\mathcal{R}$), summarisation ($\mathcal{S}$), and classification ($\mathcal{C}$) operators, introduces the relevant data objects and design choices, and establishes the end-to-end mapping from raw documents to narrative classifications.

2.1 Illustrative Example: VAT Reform in 1977

To illustrate the SANI workflow, this subsection examines a representative policy event: the major reform of the standard value-added tax (VAT) rate proposed in the 1977 French budget. The policy description is defined as reported in the policy documents: *"Government motivation for the tax change: reduction of the standard rate and unification of the standard rate and the intermediate rate"*. The measure's fiscal impact is reported as 8.5 billion francs (0.5% of GDP).

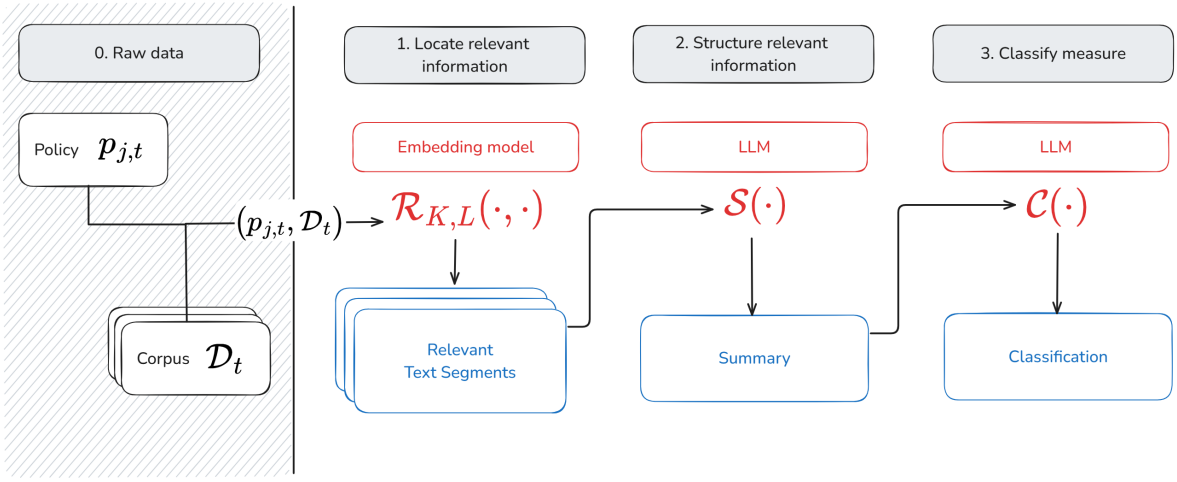


Figure 1: Four steps of the classification methodology

Step 1: Retrieval ($\mathcal{R}_{K,L}$). The retrieval operator identifies the passages in the contemporaneous Senate report that are most semantically related to the policy description p_j.[7] Among the top ten retrieved segments, the first three map directly to the article of law implementing the VAT reform and jointly cover more than 80% of its text. These passages contain the government's stated motivations for the measure. The first 5 retrieved text segments for this example are reported in Appendix A.2.

Step 2: Structured summarisation ($\mathcal{S}$). The summarisation operator processes the retrieved set $\hat{R}_j$ using a structured prompt designed to extract and organise the government's stated motives. Applied to the ten retrieved passages, this procedure produces the following structured narrative $\hat{S}_j$ (verbatim model output in italics):

- ***Mitigate the rise in prices after the end of the price freeze**: The measure aims to limit the inflationary rebound likely to occur at the end of the price-freeze period (December 31, 1976) by reducing the tax burden on consumption.*
- ***Complete a structural reform**: This change represents the final step in a reform undertaken since 1969, aimed at merging the standard VAT rate with the intermediate rate and reducing the number of VAT rates.*
- ***Fight inflation**: The reduction of the rate is part of the broader government programme to combat inflation by easing indirect taxation.*

These three motivations reflect the full substantive content of the official justification: two relate to short-run stabilisation concerns, while the second captures a longer-term structural objective.

[7] The corpus $\mathcal{D}_t$ contains the full Senate report on the 1977 budget, segmented into approximately 800 text blocks of $L = 2{,}000$ characters. The query embeds the policy description p_j in the same vector space as all document segments, and the operator selects the $K = 10$ closest segments under the embedding distance measure. The first three retrieved segments correspond to the article of law implementing the VAT adjustment and together account for more than 80% of its content; the remaining segments relate to other fiscal measures.

Step 3: Classification ($\mathcal{C}$). The classification operator evaluates each structured motivation against a set of narrative categories following Cloyne (2013). Categories are queried independently, with each prompt returning a binary label and a rationale tied explicitly to the structured narrative $\hat{S}_j$. For this example, the first and third motivations are classified as *Demand Management*, reflecting their explicit inflation-related objectives, while the second motivation is classified as *Long-Term Structural Reform*.

Discussion. The example highlights a common feature of fiscal policy: a single measure may combine short-run macroeconomic aims with longer-run structural goals. Whereas traditional narrative approaches typically resolve such cases by selecting a dominant motivation (Romer & Romer, 2010) or presenting multiple classifications as robustness checks (Cloyne, 2013), SANI records all stated motives and their classifications. This structure allows downstream analyses to incorporate mixed motivations directly, for instance through exogeneity shares.

2.2 Formal Framework

Let j index individual policy events within period t. Each event (j,t) is associated with a policy description $p_{j,t}$ and a contemporaneous document collection $\mathcal{D}_t$, which gathers texts relevant for identifying the motivations behind policy changes—for example, budget reports or legislative debates.[8] The collection is partitioned into N text segments $d_{t,k}$ of length L:

$$\mathcal{D}_t = \{d_{t,1}, \ldots, d_{t,N}\}.$$

For brevity, the time index is omitted in what follows. A maintained assumption is that the contemporaneous corpus $\mathcal{D}_t$ contains the relevant statements of motivation for each policy event.

Semantic retrieval. The retrieval operator identifies the text segments in $\mathcal{D}$ most relevant to the policy description p_j. An embedding model maps p_j and each segment d_k into a shared vector space, and relevance is measured by the distance

$$s_{j,k} = \text{dist}(p_j, d_k), \qquad d_k \in \mathcal{D}.$$

The operator returns the K closest segments:

$$\mathcal{R}_{K,L}(p_j; \mathcal{D}) = \underset{d_k \in \mathcal{D}}{\arg\min}^{K} \; s_{j,k} = \{d^{(1)}, \ldots, d^{(K)}\}, \tag{5}$$

where $d^{(k)}$ denotes the k-th smallest-distance segment.[9] Retrieval performance is governed by the text segment length L and the number of text segments retrieved K; Section 5.1.2 reports the calibration used in the application.

[8] In many applications, documents cover multiple measures within a period, making it natural to define a common corpus $\mathcal{D}_t$. When measure-specific documents are available, the framework can instead define a separate corpus $\mathcal{D}_{j,t}$ for each event.

[9] The threshold formulation $s_j^{(k)} \leq \max_{d \in \hat{R}_j} s_{j,d}$ is equivalent.

Structured summarisation. The second stage condenses the retrieved set $\hat{R}_j$ into a structured narrative $\hat{S}_j$ that records the policy change and its stated motivation:

$$\mathcal{S}(\hat{R}_j) = \hat{S}_j.$$

The operator is implemented using an LLM guided by a structured prompt that extracts and organises stated motives. The target is minimal transformation: the model filters and prioritises the information most relevant to the policy description while remaining faithful to the retrieved text.

Classification. The final stage instructs the LLM, through category-specific prompts, to map structured narratives into the categories defined for the application. Formally,

$$\mathcal{C}(\hat{S}_j) = \hat{c}_j \in \mathcal{G},$$

where $\mathcal{G}$ denotes the set of admissible categories—for example, $\mathcal{G} = \{\text{exogenous}, \text{endogenous}\}$ when studying tax shocks. Each category is queried independently, producing a binary label and a concise rationale linked to $\hat{S}_j$. This modular design isolates misclassifications within a single category and allows the category set $\mathcal{G}$ to be refined as needed. When mutually exclusive classifications are required, explicit resolution rules can be imposed.

Reproducibility. All prompting and decoding parameters are collected in a fixed vector θ, with stochastic elements of text generation disabled (temperature set to zero; see Appendix A.3). Under these settings, the retrieval, summarisation, and classification operators are deterministic, i.e., $\mathcal{S}(\cdot;\theta) = \mathcal{S}(\cdot)$. Open-source models ensure perfect reproducibility; proprietary models may be subject to silent updates, and researchers should therefore log metadata such as model version and timestamp.

End-to-end mapping. These steps define the mapping

$$(q_{j,t}, \mathcal{D}_t) \xrightarrow{\mathcal{R}_{K,L}} \hat{R}_{j,t} \xrightarrow{\mathcal{S}} \hat{S}_{j,t} \xrightarrow{\mathcal{C}} \hat{c}_{j,t},$$

which transforms each policy event into a structured narrative and its associated classifications.

3 Marginal Narrative Contributions and Grounding

The classification stage of SANI assigns each policy measure to economically meaningful categories (such as exogenous or endogenous) on the basis of the structured motivations extracted in the summarization step. Because these classifications play a central role in narrative identification, it is essential to determine whether they are actually supported by the underlying documentary evidence. A category assignment should reflect the content of the retrieved sources rather than patterns the model has absorbed during pretraining or generic associations loosely related to the policy context.

This concern motivates the need for a diagnostic that quantifies the *dependence* of the model's outputs on the textual evidence provided to it. In the classification stage, each structured motivation functions as one piece of evidence that may or may not support inclusion in a given category. Some motivations should strongly influence the model's decision, while others may be irrelevant or even point

toward a different category. A transparent identification strategy therefore requires a way to measure how much each motivation contributes to the model's final answer.

To fix ideas, consider a policy measure for which the structured narrative contains three motivations. When classifying this measure into a particular category, the model receives all three motivations as context. If one removes a motivation from the context and the model's answer is unchanged, then that motivation does not materially support the classification. If, by contrast, removing a motivation causes the model to reverse its decision or reduces its confidence in the original label, that motivation plays a substantive role in the inferred category. The key diagnostic object is therefore the *marginal narrative contribution*: the change in the model's output when a given motivation is withheld while all others are held fixed.

This approach connects to a growing literature examining how large language models respond to variations in their inputs (Cohen-Wang et al., 2024; Feng et al., 2018; Prabhakaran et al., 2019). In spirit, it is closest to Cohen-Wang et al. (2024), but applied at the level of entire retrieved passages rather than individual tokens or sentences.

Introducing these contributions first at the classification stage offers a direct window into the economic object of interest: how narrative evidence supports or contradicts the identifying assumptions behind each category. It also provides the simplest conceptual environment for understanding the broader framework. In the classification task, the number of motivations is small, allowing direct examination of the model's dependence without approximation. The same logic generalizes to the summarization stage, where the model processes many more retrieved passages, and where scalable estimation techniques are needed. The next subsection formalizes this idea and develops a unified framework for quantifying marginal dependence across both stages of SANI.

3.1 A Simple Framework: Marginal Contributions in Classification

This subsection formalizes the marginal narrative contribution in the simplest environment: the classification stage of SANI, where the model evaluates a small set of structured motivations for inclusion in a given category. This setting provides a clean introduction to the framework because each motivation enters as a single, well-defined unit of evidence. The dimensionality of the context is small, allowing dependence to be characterised without approximation. The same logic extends directly to the summarisation stage, where the context consists of many retrieved passages and scalable estimation becomes necessary.

Setup. Fix a policy measure j and a category c. Let the structured narrative $\hat{S}_j$ consist of m motivations,

$$\hat{S}_j = \{M_1, \ldots, M_m\}.$$

In the baseline classification query, the model receives all motivations and returns a binary label ("Yes" or "No") for category c. To analyse dependence on individual motivations, we introduce an inclusion vector $v \in \{0,1\}^m$, where $v_\ell = 1$ indicates that motivation M_ℓ is provided to the model and $v_\ell = 0$ indicates that it is withheld. Let

$$\text{CONTEXT}(\hat{S}_j, v)$$

denote the set of motivations shown to the model under inclusion pattern v.

Conditional on a fixed prompting and decoding specification $\theta_{\mathcal{C}}$ for the classification operator, define

$$f(v) = p_{\text{LM}}\big(\text{Yes} \mid \text{CONTEXT}(\hat{S}_j, v); \theta_{\mathcal{C}}\big),$$

the probability that the model assigns category c when presented with the subset of motivations specified by v. The function $f(\cdot)$ represents the model's *generation probability*: the likelihood of producing the label "Yes" for a given context.

Marginal narrative contributions. To quantify how much a motivation contributes to the classification, consider two contexts that differ only in the inclusion of motivation M_ℓ, holding all other motivations fixed at $v_{-\ell}$. The *marginal contribution* of M_ℓ at context $v_{-\ell}$ is

$$\Delta_\ell(v_{-\ell}) = f(v_\ell = 1, v_{-\ell}) - f(v_\ell = 0, v_{-\ell}).$$

This quantity measures the change in the probability of assigning category c that arises from showing the model motivation M_ℓ rather than withholding it. Large positive values indicate that M_ℓ provides strong supporting evidence for category c; values near zero imply irrelevance; and negative values indicate evidence against inclusion.

Because the number of motivations m is usually small, all the 2^m inclusion vectors v can be enumerated. This makes the marginal contributions $\Delta_\ell(v_{-\ell})$ directly computable without approximation. The resulting dependence pattern is therefore fully transparent: the analyst can trace how each motivation contributes to the classification decision and verify that the model's label reflects the intended documentary evidence.

VAT illustrative example. Table 1 reports attribution weights for the motivations underlying the 1977 VAT reform across the eight categories of Cloyne (2013). Categories such as *Spending-Driven* (SD), *Supply Stimulus* (SS), *Deficit Reduction* (DR), and *External* (EXT) receive values near zero, indicating little documentary support. By contrast, *Demand Management* (DM), *Ideological* (IL), and *Long-Run* (LR) exhibit higher narrative contribution, consistent with the motivations stated in the Senate report. The highlighted entry in each row marks the category most strongly supported by that motivation, making clear which elements of the narrative drive the classification.

Category-Specific Calibration and Relative Support

While raw marginal contributions reveal how each motivation affects the assignment decision within a category, the logits returned across categories are not directly comparable. Each category is queried using a distinct prompt, and prompt-specific framing differences induce arbitrary location and scale shifts in the corresponding logits. To place all categories on a common footing, I normalize logits using category-specific anchors.

This normalization procedure is related to recent work on calibrating language-model outputs to correct for prompt- and label-specific biases. Zhao et al. (2021) introduce *contextual calibration*, which estimates label priors from content-free inputs and adjusts logits so that predictions are not driven by the model's baseline preference for particular verbalizers. Fei et al. (2023) extend this approach through *domain-context calibration*, using random in-domain text as a placebo input to control for domain-specific shifts in label scores. While these methods anchor logits using neutral or randomly

Table 1: Estimated marginal narrative contribution across Cloyne (2013) classification categories

	Category	M1	M2	M3
N	Demand Management (DM)	0.55	0.31	0.38
	Spending-Driven (SD)	0.00	0.00	0.28
	Supply Stimulus (SS)	0.00	0.10	0.00
	Deficit Reduction (DR)	0.03	0.00	0.21
X	Deficit Consolidation (DC)	0.17	0.23	0.00
	External (EXT)	0.34	0.21	0.22
	Ideological (IL)	0.35	0.39	0.14
	Long-Run (LR)	0.21	0.67	0.30

Note: Entries report the *calibrated* marginal narrative contribution of each motivation to the corresponding classification category. Larger values indicate stronger documentary support for assigning the measure to that category. For each motivation, the shaded cell marks the category receiving the highest contribution, indicating the most strongly supported interpretation. "N" denotes endogenous subcategories and "X" denotes exogenous subcategories as defined in Cloyne (2013). Detailed category definitions appear in Table 2.

generated contexts, the procedure developed here instead employs semantically meaningful positive and negative reference prompts to identify the effective range of logits within each category. This anchor-based strategy parallels self-anchoring techniques in psychometrics and, to my knowledge, has not been applied to LLM calibration, thereby enabling interpretable comparisons of calibrated scores across categories.

Raw logits. For each category c and each motivation M_ℓ, let

$$Y_{\ell,c} = \log p_{\text{LM}}(\text{Yes} \mid X_{\ell,c}),$$

where $X_{\ell,c}$ denotes the prompt consisting of the category definition and motivation M_ℓ. Because prompts differ across categories, these raw logits are not comparable.

Anchor-based normalization. For each category c, I select two reference prompts: a placebo–No prototype yielding $Y_{\min,c}$ and a placebo–Yes prototype yielding $Y_{\max,c}$. These anchors span the relevant semantic range for category c. The logits are then rescaled to the unit interval:

$$S_{\ell,c} = \frac{Y_{\ell,c} - Y_{\min,c}}{Y_{\max,c} - Y_{\min,c}} \in [0,1]. \tag{6}$$

Values near zero (one) indicate that the motivation resembles the negative (positive) prototype for category c.

Interpretation. If the logits satisfy the affine form $Y_{\ell,c} = a_c + b_c\theta_{\ell,c}$ with $b_c > 0$, where $\theta_{\ell,c}$ is a latent semantic intensity, then (6) recovers

$$S_{\ell,c} = \frac{\theta_{\ell,c} - \theta_{\min}}{\theta_{\max} - \theta_{\min}},$$

removing all prompt-specific distortions (a_c, b_c). The normalised scores are therefore interpretable on a common latent scale across categories.

Relative support across categories. For each motivation, the set of calibrated scores $\{S_{\ell,c}\}_{c\in\mathcal{G}}$ summarises its semantic alignment across categories. Normalizing across categories yields a probability-like measure:

$$\tilde{P}_\ell(c) = \frac{S_{\ell,c}}{\sum_{c'\in\mathcal{G}} S_{\ell,c'}}.$$

This construction parallels normalizations in discrete-choice models and provides an interpretable distribution of support across categories.

Aggregation to Endogenous and Exogenous Supergroups

The empirical application groups the eight categories into two mutually exclusive supergroups: *endogenous* and *exogenous*. Because these supergroups represent the economically relevant identification criterion, I aggregate calibrated scores within each supergroup:

$$S_\ell^{\text{endo}} = \sum_{c\in\mathcal{G}_{\text{endo}}} S_{\ell,c}, \qquad S_\ell^{\text{exo}} = \sum_{c\in\mathcal{G}_{\text{exo}}} S_{\ell,c}.$$

Probability-like measures over the binary partition follow as:

$$\tilde{P}_\ell^{\text{endo}} = \frac{S_\ell^{\text{endo}}}{S_\ell^{\text{endo}} + S_\ell^{\text{exo}}}, \qquad \tilde{P}_\ell^{\text{exo}} = \frac{S_\ell^{\text{exo}}}{S_\ell^{\text{endo}} + S_\ell^{\text{exo}}}.$$

A motivation is assigned to the exogenous group whenever $\tilde{P}_\ell^{\text{exo}} > \tilde{P}_\ell^{\text{endo}}$, ensuring mutual exclusivity at the top level while preserving the underlying subcategory structure for diagnostic analysis.

From the simple case to the general framework. The classification environment thus provides a complete, exact, and interpretable application of marginal contributions and calibrated logits. The same conceptual structure applies when the evidence units are retrieved passages rather than structured motivations, but in that setting high dimensionality necessitates approximation. The next subsection extends the framework to this general setting and introduces a scalable method for estimating marginal dependence in the summarisation stage.

3.2 A General Framework: Marginal Contributions in Summarisation

The classification case formalizes marginal narrative contributions in the simplest environment, where the number of motivations is small and all 2^m inclusion patterns can be evaluated directly. The summarisation stage poses the same conceptual problem—quantifying how each unit of evidence affects the model's output—but in a high-dimensional setting. For each policy measure j, the retrieval operator returns a set of passages $\hat{R}_j = \{d_1, \ldots, d_K\}$, typically with K between ten and several dozen. The summariser processes these passages jointly and outputs $\hat{S}_j$. Assessing grounding therefore requires measuring how strongly the regeneration probability $f(v)$ responds to the inclusion or exclusion of each passage, where $v \in \{0,1\}^K$ is the inclusion vector defined in the previous subsection.

When K is large, exhaustive evaluation of all 2^K contexts is infeasible. To approximate the marginal log-odds effects $\Delta_\ell^g(v_{-\ell})$ introduced earlier, I approximate the log-odds function $g(v)$ with a linear surrogate,

$$g(v) \approx b + \sum_{\ell=1}^{K} w_\ell v_\ell.$$

The coefficients w_ℓ are estimated by querying the model on a random sample of inclusion patterns $v^{(1)}, \ldots, v^{(n)}$ drawn uniformly from $\{0,1\}^K$. In the empirical implementation, I use $n = 64$ ablations per measure and at most $K \approx 30$ retrieved passages, ensuring that the linear approximation remains well behaved. For each sampled pattern, the regenerated log-odds $g\big(v^{(i)}\big)$ provide an observation that identifies the local dependence of the summary on each passage.[10] The resulting coefficients $\hat{w}_\ell$ estimate the average marginal log-odds contribution of passage d_ℓ, $\hat{w}_\ell \approx \mathbb{E}_{v_{-\ell}}[\Delta_\ell^g(v_{-\ell})]$. In practice, only a subset of passages receives substantively large or statistically meaningful coefficients, so empirical dependence is effectively sparse even without explicit penalisation. Positive coefficients indicate passages that support the generated summary, near-zero coefficients indicate irrelevance, and negative coefficientsPositive coefficients indicate passages that support the generated summary, near-zero coefficients indicate irrelevance, and negative coefficients[11]signal passages whose inclusion would push the model away from the observed output. signal passages whose inclusion would push the model away from the observed output. Combined with the placebo test described below, this surrogate yields a faithful and interpretable diagnostic of grounding in the summarisation stage.

VAT example for dependence. In practice, only a few passages receive substantively large positive weights, indicating that the summariser anchors each motivation in a small subset of the retrieved record. Table 2 illustrates this pattern for the 1977 French VAT reform. Across the three generated motivations, positive contributions are highly concentrated, with the largest coefficients assigned to passages that describe the government's stated goals for the measure.

Table 2: Attribution weights and placebo-test results across top ten retrieved passages

	Retrieved segments (1–10)									
	1	2	3	4	5	6	7	8	9	10
M_1	41.56	20.92	7.33	-0.24	-1.76	8.36	2.78	-0.69	-4.93	4.47
M_2	18.82	-4.13	15.29	-4.56	0.25	1.20	1.93	-3.47	-0.51	-3.26
M_3	2.88	8.99	1.19	0.76	-1.26	1.04	1.05	-1.40	-0.46	2.67

Note: Each entry reports the estimated marginal log-odds contribution of the corresponding passage to regenerating the motivation. Shaded cells indicate passages that pass the placebo-based grounding test at the 5% level. These entries exhibit positive marginal contributions that are statistically larger than would arise under random assignment of placebo labels. Further implementation details are provided in Section 3.2.

[10] Appendix B.1 reports stability of the estimated dependence patterns across OLS, ridge, and LASSO surrogates.

[11] Negative coefficients are more difficult to interpret than positive ones. While a positive weight has a clear substantive meaning—support for the generated motivation in the direction specified by the prompt—a negative weight may reflect heterogeneous mechanisms: contradiction, a competing frame, or linguistic dilution of the signal. Because there is no unified notion of "negative support," the placebo test focuses on positive coefficients.

A Placebo-Based Test for Grounding

The marginal contributions estimated in the previous subsection identify which retrieved passages are influential for each motivation. However, a large coefficient does not, by itself, guarantee that the influential passage is substantively relevant to the policy event. A passage may receive a high positive weight simply because its linguistic features resemble generic patterns the model has learned, rather than because it contains meaningful evidence. To diagnose this, I implement a placebo-based test that evaluates, for each retrieved passage individually, whether its marginal contribution is unusually large relative to contributions assigned to semantically irrelevant passages.

Relevant and placebo sets. For each policy event, the retrieved passages are partitioned into:

- *Real passages*: the first J retrieved segments, which contain the material most likely to include policy-relevant motivations;
- *Placebo passages*: the remaining segments, drawn from the same document but selected to be semantically unrelated to the specific measure.

This setup yields a vector of marginal source contributions $\{w_0, \ldots, w_{J-1}, w_J, \ldots, w_{K-1}\}$, where the first J coefficients correspond to real passages and the remaining $K - J$ to placebo passages.

Positive-only comparison. Since only positive coefficients can provide support for a motivation, I restrict attention to real passages with $w_j > 0$. For each such passage j, the test compares w_j to the empirical distribution of *positive* placebo coefficients:

$$\mathcal{P} = \{w_k : k \geq J,\ w_k > 0\}.$$

This ensures that the benchmark reflects placebo passages that the surrogate model treats as supportive in expectation.

Monte Carlo p-values. The null hypothesis is that the marginal contribution w_j is not systematically larger than the positive contributions assigned to placebo passages. Under the null, w_j should behave like a random draw from $\mathcal{P}$. I approximate the corresponding p-value via Monte Carlo sampling:

$$p_j = \Pr(\tilde{w} \geq w_j), \qquad \tilde{w} \sim \mathcal{P},$$

which is estimated by repeatedly sampling one element from $\mathcal{P}$ (with replacement) and counting the fraction of iterations in which the sampled value exceeds w_j. A small p_j indicates that w_j is unusually large relative to the placebo distribution and therefore represents meaningful dependence on passage j. For real passages with non-positive w_j, the test is skipped and p_j is set to 1.

Interpretation. This coefficient-level test identifies which specific passages genuinely support the summarised motivations and which appear influential only due to noise or stylistic correlations. A real passage is deemed substantively important when (i) its marginal contribution is positive and (ii) it exceeds the benchmark provided by positive placebo passages at conventional significance levels. Because each passage is evaluated independently, the diagnostic provides a fine-grained map of grounding: it

reveals exactly which sources drive the model's output and which can be dismissed as incidental. In practice, this greatly reduces the burden of manual verification. The researcher may focus attention on a small subset of statistically supported passages when checking the validity of a motivation, rather than inspecting the full retrieved context. In this way, the test complements the marginal contribution analysis by adding a transparent statistical filter that both sharpens interpretability and improves the efficiency of human review.

VAT example for placebo grounding. A second reading of Table 2 incorporates the coefficient-level placebo test. Highlighted cells mark real passages whose positive marginal contributions exceed the empirical distribution of positive placebo coefficients at the 5% level. These statistically supported passages constitute the evidentiary basis on which the summariser relies. A generated motivation is therefore deemed grounded whenever its row contains at least one highlighted entry, indicating that the model draws substantively on the retrieved record rather than on incidental patterns.[12]

4 Data Construction

4.1 French Tax Data

After presenting the methodology, I now turn to its application to a substantive case: value-added tax (VAT) policy in France. Implementing this analysis requires the construction of a comprehensive dataset of French tax policy changes between 1960 and 2024. The dataset combines detailed fiscal information from official budget documents with contemporaneous legislative texts that record policy-makers' stated motivations. This dual structure allows both the quantitative identification of policy measures and the qualitative classification of their underlying rationales. The remainder of this section describes the construction of the dataset in two steps: first, the identification of tax policy changes in the *Voies et Moyens Tome I* (VMT), and second, the compilation of the textual corpus used to recover their motivations.

4.1.1 Identifying Tax Policy Changes

Primary Source: *Voies et Moyens Tome I* (VMT). To construct a comprehensive dataset of tax policy changes in France between 1960 and 2024, I systematically processed the information contained in each year's *Voies et Moyens Tome I* (VMT), an annex to the annual draft budget law (*Projet de Loi de Finances*). While most recent volumes of the VMT are accessible online, earlier editions were not digitized. I therefore retrieved and scanned the missing volumes[13] from archival collections held by public institutions.

The VMT provides a detailed breakdown of the government's revenue projections and is published annualy. It covers both fiscal and non-fiscal revenues, showing how these are expected to evolve in the coming fiscal year. The VMT distinguishes between the mechanical evolution of tax bases, the effects of newly adopted legislation, and the delayed consequences of past reforms. Importantly, the section on

[12] Because multiple passages are tested within each measure, some significant cells are expected under the null. The placebo test is designed as a diagnostic rather than as a vehicle for formal inference, so the main text focuses on the pattern of supported passages rather than on multiplicity-adjusted thresholds. Appendix B.2 reports a simple global test that aggregates evidence across passages and confirms that the main results are not driven by chance significance.

[13] The online VMT documents were unavailable for the years 1960–1983, as well as for 1988 and 1989.

the impact of past reforms also includes certain tax changes legislated between two VMTs, capturing some of the measures enacted outside the scope of the annual budget planning process. As some tax changes remain unaddressed, I will need to review relevant decrees ('décrets') and amended finance laws ('lois de finances rectificatives'). This will be incorporated into a next version of the paper.

Decomposition of Annual Revenue Changes. The change in net tax revenues from one year to the next, as presented in the VMT, is broken down into four main components[14]:

$$\begin{aligned}\Delta\text{Revenues} = \text{Spontaneous change} + \text{Scope and transfer measures} \\ + \text{Previous measures} + \text{New measures}\end{aligned} \tag{7}$$

1. **Spontaneous change**: This represents the natural (or endogenous) evolution of tax revenues at constant legislation. It captures changes due to economic dynamics, such as GDP growth, household income, consumption, inflation, and tax base expansion. It also includes the effects of tax progressivity, e.g., an increase in wages moving taxpayers into higher brackets.

2. **Previous measures**[15]: These are the effects of measures adopted in previous years, which continue to affect revenues through delayed application, phased implementation, or full-year effects. For example, the extension of a tax credit that starts mid-year in $t-1$ and applies fully in year t.

3. **New measures**[16]: These are new tax measures proposed in the current year's draft budget law (PLF). They include newly introduced taxes, rate changes, deductions, or credits that introduce a change in legislation.

4. **Scope and transfer measures**[17]: These reflect technical adjustments in how revenues are accounted for, such as reclassification of taxes from one budget entity to another, or rebudgetisation of previously off-budget revenues. A change in this category should not have any direct macroeconomic impact and will be excluded of the main dataset.

This structure allows for the identification of discretionary policy actions with potential macroeconomic implications—specifically, those classified as new measures (3), introduced in the current draft budget law, and previous measures (2), carried over from earlier reforms whose effects extend into the current fiscal year. In this paper, I will restrict the analysis of the dataset to new measures; integrating previous measures awaits reliable motivation mapping and is left to future work at this stage.

Tax categories. The VMT systematically disaggregates annual revenue changes across a standardized set of 6 major tax categories[18]. This structured breakdown reflects the legal and economic distinctions between different revenue sources. The consistency of this classification across years ensures comparability and aids in constructing a coherent dataset of tax reforms[19].

[14]The explanations related to this breakdown is present in any year of the VMT documents. The years before 2000 do not include the component "Scope and transfer measures".

[15]Before 2009, this section is named "Other factors of variation".

[16]Before 2009, this section is named "Adjustment of rights".

[17]Before 2009, there is no dedicated section for this. Scope and transfer measures are present in the two former section (2) and (3).

[18]I merge the two categories related to "Other direct taxes" in one single categories for simplicity.

[19]Categories are broadly consistent over the years appart from few example that will be documented in the Appendix.

I report the decomposition of revenues[20] for each of the major tax groups presented in the VMT in table 3 as well as some statistics about the data. For example, in 2024, VAT accounts for 38% of total gross revenues and 6% of GDP. Between 1980 and 2024, the VMT records a total of 444 changes in VAT[21].

Table 3: Gross Tax Revenue Breakdown by Source (France, 2024)

Tax Category	**Share of Revenue (%)**	**As % of GDP**	**#New Measures**	**#Old Measures**
Personal Income Tax	24.3%	3.89%	359	1940
Other Direct Taxes	7.2%	1.15%	185	284
Corporate Income Tax	18.5%	2.96%	137	1459
Energy Consumption Tax	3.8%	0.60%	65	114
Value-Added Tax (VAT)	37.8%	6.06%	140	324
Other Indirect Taxes	8.5%	1.36%	243	295
Gross Tax Revenues	**100.0%**	**16.02%**	1145	4444

Dates. The data collected is at the yearly frequency. To convert the dataset to the quarterly frequency, the following procedure is applied.

Measures that have been enacted during the annual budget draft are enacted the 1st of January of the coming year. This represent the majority of the measures. However, there are exception: some measures are applied by anticipation, some others are planned to be applied in the coming years. This information is partly taken from the data provided by the *Institut des politiques publiques* that collects the date of implementation of the different taxes[22].

Measures coming from Law finances that are legislated in between the annual budget procedure are allocated to the quarter where it has been legislated. The same applies to other type of laws that do not correspond to LF or LFR but still affect the ressources of the state.

4.1.2 Corpus of Text to Identify the Motivations behind the Tax Changes

A final step in building the dataset is to curate a corpus of contemporaneous texts which likely include the motivations behind the different granular tax changes. As highlighted by Romer and Romer, the sources should be carefully selected to ensure they reflect the political, economic, and social context of the time when the tax changes were proposed and enacted. In addition, the source should be reliable and consistent over time. To achieve this, I collect the Senate reports as the primary source for analysis from 1960 to 2024.

[20]The higher figures for the tax-to-GDP ratio reported by institutions like the OECD and IMF arises from differences in definitions and the scope of what is considered "tax revenue." They usually include compulsory social security contributions in their tax revenue calculations, which is not the case in the figures presented in the annual draft budget law. In France, these contributions are substantial and significantly impact the tax-to-GDP ratio reported.

[21]This number excludes the new measures related to "Scope and transfer" when identifiable.

[22]We timestamp measures using the *Institut des politiques publiques* (IPP) statute-level database, which records for each change the legal basis (law/decree/arrêté or social-partner agreement), the publication date in the Official Journal (JORF), the effective date, and links to the source on legifrance.gouv.fr. For tax schedules, coverage is historical (income tax back to 1914; social contributions to 1945; most items from the 1980s) and maintained with the OpenFisca community. In my quarterly mapping, I use IPP effective dates; when unspecified, I fall back to the JORF publication quarter.

In the analysis of finance bills, Senate reports provide a particularly useful entry point. They are often more critical and technical than those of the National Assembly, making them well suited for identifying the government's underlying motivations behind changes in taxation. By highlighting inconsistencies, emphasizing the fiscal consequences of certain choices, or questioning the official justifications advanced by the government, the Senate offers insights that go beyond the surface of the legislative text. In addition, because the Senate examines the text after the Assembly's first reading, its reports also indicate which measures were adopted, amended, or deleted upstream. My current analysis therefore relies on Senate reports both to trace the motivations behind tax reforms and to verify which provisions survived the Assembly's first examination.

That said, the analysis of reports from the National Assembly would have their own advantages. They tend to align more closely with the government majority, and thus provide the most direct view of the legislative choices of the ruling coalition. While they are less critical in tone, they are essential for confirming the government's political priorities and for tracking in real time the measures approved at each stage. For now, my focus remains on the Senate's perspective, which combines a critical reading of motivations with information on legislative outcomes, but adding the Assembly reports could be a natural extension of the corpus.

4.2 US Tax Data

4.2.1 Identifying Tax Policy Changes

Existing narrative coverage. The narrative tax dataset of Romer and Romer (2010) identifies major federal tax actions in the United States over the postwar period, focusing on large and salient legislative episodes with substantial macroeconomic relevance. The series includes comprehensive tax reforms and major standalone acts, such as income tax reforms, Social Security legislation, and federal highway funding acts, that generate sizable and clearly identifiable changes in aggregate tax liabilities. By construction, the dataset abstracts from routine fiscal adjustments and from smaller tax provisions embedded in annual budgetary processes. In particular, it excludes most year-to-year budget measures and minor tax changes implemented outside of major legislative acts, even when these provisions affect statutory tax parameters or tax bases. As a result, while the Romer and Romer series provides a high-quality measure of large, infrequent tax shocks, it offers only partial coverage of legislated tax changes at an annual frequency and at a disaggregated tax-instrument level.

Extension. To recover the tax changes omitted by existing narrative datasets, I construct a comprehensive annual series of legislated federal tax measures by systematically processing each year's *Budget of the United States Government* from 1950 to 2024. For each fiscal year, I extract all reported changes in federal taxation and record them at the highest level of detail available, including a disaggregation by tax instrument whenever possible. The same extraction procedure is applied to the major tax acts identified by Romer and Romer (2010), yielding a finer tax-level decomposition than that available in the extension of Mertens and Ravn (2013).

In the present paper, this disaggregation is not exploited to estimate differential macroeconomic effects across tax instruments. Instead, it is used to recover policy motivations at the instrument level and to relate these motivations to the overall narrative classification of each legislative package. This structure allows the analysis to distinguish between act-level motivations and the heterogeneous policy

rationales underlying their constituent tax measures, while leaving the study of instrument-specific transmission mechanisms to future work.

To provide an overview of the composition of the resulting dataset, Table 4 reports the distribution of new tax measures introduced within the annual budget framework. The table aggregates all tax changes explicitly proposed in the yearly *Budget of the United States Government* over the sample period and classifies them into six major federal tax categories (individual income, corporate income, payroll/social insurance, consumption, wealth transfer, and other receipts).

Importantly, the table does not include tax changes enacted through stand-alone tax legislation outside the annual budget process. It therefore reflects the structure of tax policy adjustments as they appear in the executive budget proposals, rather than the universe of all federal tax legislation.

Table 4: Distribution of New Tax Measures by Category (United States)

Tax Category	**# New Measures**
Individual Income Tax	371
Corporate Income Tax	361
Consumption Tax	238
Payroll / Social Insurance Tax	114
Other Receipts	89
Wealth Transfer Tax	17
Total	**1190**

4.2.2 Corpus of Text to Identify the Motivations behind the Tax Changes

A final step in constructing the U.S. tax dataset is to assemble a corpus of contemporaneous documents that record policymakers' stated motivations for legislated tax changes. Sources are selected to reflect the policy rationale articulated at the time of enactment and to remain consistent and comparable over long horizons.[23] The analysis relies on two primary sources: the *Budget of the United States Government* and the *Annual Report of the Secretary of the Treasury*.

The Budget documents serve as the main source of narrative evidence for tax changes embedded in the annual budget process, as they provide detailed descriptions of proposed measures and explicitly articulate the administration's rationale within the broader fiscal strategy. For tax changes implemented outside the budget cycle, including one-off legislative actions of the type identified by Romer and Romer (2010), the Annual Treasury Reports are used instead, as they systematically document all tax changes enacted in a given year and discuss their intended objectives. Together, these sources allow motivations to be recovered for both recurring budgetary adjustments and standalone tax legislation, ensuring that narrative classifications are grounded in contemporaneous policy documents rather than ex post interpretation.

[23]Romer and Romer (2010) draw on a broad set of contemporaneous sources, including legislative debates and discussions in congressional chambers. Some of these materials are not systematically used here, primarily because their access and coverage are uneven over time. The analysis therefore focuses on recurring executive and Treasury documents that are consistently available throughout the sample period.

5 VAT in France

5.1 Applying SANI

5.1.1 Embedding model and LLM choice

To implement the SANI framework, I first select a pretrained embedding model to represent the textual content of policy documents in vector form. I use the `text-embedding-large` model developed by OpenAI, which provides a one-to-one mapping between each text segment and a high-dimensional embedding vector.

For large language model (LLM) operations, I employ a two-model strategy that balances performance and reproducibility. The primary outputs of the SANI procedure—summarization of retrieved segments and classification of policy motivations—are generated using `gemini-3.0-pro`. The use of a very large model at these stages offers high performance in natural language understanding and reasoning, ensuring high-quality generation of structured summaries and consistent application of categorization rules.

For robustness and interpretability analyses described in Section 3, I rely on the open-weight model `Qwen-8B`. Using an openly available model allows full reproducibility of the marginal narrative contribution analysis. The combination of these two models thus leverages the superior generative capabilities of frontier proprietary systems for the main classification outputs, while preserving the replicability and auditability afforded by open-weight alternatives for robustness checks. That said, the entire SANI procedure can, if needed, be implemented exclusively with open-weight models without altering its logic or structure.

This separation between generation and marginal source contribution stages is deliberate and methodologically valid. The computation of source contributions operates conditional on a fixed output: its goal is to quantify how individual retrieved segments influence that output, regardless of how it was produced. Consequently, the interpretability analysis can be conducted with an alternative or open-weight model without affecting the validity of the results. In principle, the same procedure could be applied to outputs generated by any model or even by human analysts, provided the classification or summary is held constant.

5.1.2 Retrieval

The first step in applying SANI to French VAT policy is to retrieve, for each identified tax change, the passages in contemporaneous Senate reports that contain the government's stated motivations. Because retrieval quality directly affects all subsequent stages, this subsection evaluates the performance of the semantic retriever and documents the trade-offs involved in choosing segmentation length L and the number of text segments retrieved K. Using a manually constructed benchmark, I assess both the optimal parameter configuration and the value of exploiting document structure relative to a fully unsupervised segmentation approach.

Benchmark design: golden articles. Each VAT measure in the *Voies et Moyens Tome I* (VMT) is manually linked to the specific article of the Senate report that discusses the corresponding provision of the budget law. I refer to this as the *gold article*. Because French budget laws are highly

structured—each article of the draft law introducing a specific tax measure, followed by a dedicated Senate discussion—the mapping between policy events and relevant text is clear. However, in the retrieval stage, documents are divided into fixed-length text segments of length L, and a single gold article spans several segments. The goal is therefore not to retrieve one exact chunk, but to ensure that among the top-K retrieved segments there is sufficient overlap with the true gold article.

Ensuring coverage. Performance is assessed by verifying, for each event, whether at least one of the top-K retrieved segments belongs to the corresponding gold article. This criterion captures whether the model successfully localizes the relevant discussion. Across segment lengths $L \in \{512, 1024, 2048, 4096\}$, results are highly robust: in nearly all cases, at least one segment from the gold article appears among the top ten retrieved passages, and on average 5.5 of the ten retrieved segments overlap with the gold article.[24] Increasing K or L beyond these ranges yields only marginal improvements, indicating that retrieval quality is high even with modest parameter values. The configuration $L = 1024$ is adopted as the baseline, but given the low computational cost of the full procedure, alternative values of L are also implemented and will be reported as robustness checks in a future version of the paper.

Supervised versus unsupervised segmentation. A key practical question concerns whether retrieval requires explicit alignment with the legal structure of the documents. In the *supervised* approach, the Senate reports are preprocessed so that segment boundaries coincide with article boundaries. In the *unsupervised* approach, the text is divided mechanically into equal-length segments without any prior formatting or alignment. Both methods deliver nearly identical recall among the top ten retrieved passages: the supervised variant attains full coverage with only a slightly smaller K for segment lengths $L \in \{512, 1024\}$. Given that the supervised approach entails substantial additional work, parsing, aligning, and annotating article boundariesk the improvement in accuracy is minor relative to its cost. This result is central to the contribution of SANI: it suggests that reliable narrative retrieval can be conducted directly on raw legislative text, without the need for costly manual preparation of sources. By removing this preprocessing requirement, the procedure markedly lowers the barriers to scaling the narrative approach across countries and policy domains.

Overall, the retrieval stage consistently identifies the correct Senate passages explaining each VAT measure. The method achieves near-perfect recall at low computational cost and without relying on document-specific alignment. This ensures that the subsequent summarization step operates on the relevant textual evidence for each policy change, while preserving the scalability that motivates the use of AI-assisted narrative identification.

5.1.3 Summarization

Having established that retrieval reliably identifies the relevant passages for each VAT measure, the next step condenses these passages into structured narratives that summarize the government's stated motivations. This stage applies the summarization operator $\mathcal{S}$, which transforms the retrieved text segments $\hat{R}_{j,t}$ into a compact and standardized representation $\hat{S}_{j,t}$. The goal is to translate lengthy legislative discussions into concise, factual statements that preserve the logic and evidential grounding of the original text.

[24] For $L = 2048$, this is true for 97% of the measures.

Prompt design. Summarization is implemented through a LLM instructed to extract the government's stated motivations from the retrieved passages and to present them in a structured format. The model receives a fixed prompt of the following form:[25]

English translation of the prompt:

Based on the [EXCERPTS] from the Senate report on the draft budget law for {year}, what are the main motivations of the government related to the tax change {tax type}, description: {description}? Summarize these motivations and limit yourself to a maximum of three main ones.

- Rely only on the content of the excerpts, without adding interpretation.
- Consider only motivations directly linked to the described measure.

[EXCERPTS]
[Excerpt no. 1] ... [End Excerpt no. 1]
[Excerpt no. 2] ... [End Excerpt no. 2]
[...]

Prompt 1: Summarization Prompt

Dynamic fields. The placeholders {`year`}, {`tax type`}, and {`description`} in the prompt are automatically populated with information specific to each VAT measure. The variable `year` indicates the fiscal year of the corresponding budget law, `section` identifies the legal article or provision introducing the tax change, and `description` provides a short textual description of the measure directly drawn from the French budgetary documents *Voies et Moyens Tome I*.

Fidelity and grounding. The credibility of the summarisation stage hinges on the extent to which generated motivations remain faithful to the retrieved passages. I assess this using two complementary approaches. First, grounding is evaluated through the perturbation-based attribution test introduced in Section 3, which measures how the probability of regenerating a given motivation responds to ablating individual retrieved segments. For 98.5 percent of generated motivations, at least one retrieved passage exhibits a significant positive attribution weight at the 2.5 percent level, indicating that the model's output depends on the supplied evidence rather than on prior knowledge. Together, these results show that $\mathcal{S}$ functions as a faithful compression operator that extracts motivations without distortion or hallucination. Second, manual inspection of a benchmark sample of VAT measures is underway to provide additional qualitative validation.

The structured narratives produced by this step serve as direct inputs for the classification operator $\mathcal{C}$, which maps each summarized motivation into a policy category according to codified decision rules. The next subsection applies this operator to distinguish exogenous from endogenous VAT changes based on their stated motivations.

[25] The prompt specifies as well a uniform output structure across all measures: motivations must appear as bullet points beginning with "–", and listed directly without introductory text.

Group	Sub category	Explanation and examples
Endogenous (N)	1. Demand management (DM)	• Targeting the aggregate level of demand e.g. to boost investment, consumption, growth, or curb inflation. • Specific help to households and individuals by stimulating disposable income. • Dealing with a balance of payments crisis via demand.
	2. Supply stimulus (SS)	• Certain help for businesses during a downturn (e.g. NIC cut). • Short term sector support (e.g. targeted tax cuts for a sector).
	3. Deficit reduction/BoP crisis (DR)	• Direct measures to deal with a budget or external deficit contemporaneously caused.
	4. Spending-driven (SD)	• Taxes which fund specific spending commitments.
Exogenous (X)	1. LR performance (LR)	• Measures to improve competitiveness, productivity, efficiency and long-run growth (but not taken to offset a shock). • Simplification and deregulation measures. • Long-term support for business or sectors of the economy.
	2. Ideological (IL)	• Long-term social or political goals, independent of their effect on performance and not to offset current shocks. • Some anti-avoidance measures (where no other motive is given).
	3. External (ET)	• Court rulings and enforcement of directives.
	4. Deficit consolidation (DC)	• Measures to lower inherited deficit for reasons of economic philosophy or to offset current actions in the future. • Does not include actions forced on the government, or decisions contemporaneous motivated by a current shock.

Figure 2: Classification of Tax Measures Following Cloyne (2013)

5.1.4 Design of Classification Categories

Whereas the retrieval and summarization stages of SANI are general, the design of classification categories is inherently application-specific. It defines the conceptual boundary between types of policy actions and therefore determines the interpretation of the resulting narrative shocks. In this application, the objective is to classify value-added tax (VAT) measures as *endogenous* or *exogenous*, following the narrative taxonomy developed by Cloyne (2013) for the United Kingdom. This framework provides economic criteria for distinguishing tax changes responding to contemporaneous macroeconomic conditions from those motivated by long-run structural, ideological, or institutional considerations.

Table 2 recall the categories defined by Cloyne (2013), along with representative examples. In this taxonomy, *endogenous* measures are those explicitly linked to current economic or fiscal conditions—for instance, to stimulate demand, stabilize activity, or address short-term budget pressures. By contrast, *exogenous* measures are justified by objectives unrelated to the current state of the economy, such as improving efficiency, ensuring long-run fiscal sustainability, or advancing ideological or institutional reforms. This distinction is conceptually consistent with Romer and Romer (2010), who classify U.S. tax changes according to their stated motives, but adapted by Cloyne to more granular categories.

Operationalization in SANI. In practice, the classification operator $\mathcal{C}$ applies codified decision rules to the structured motivations $\hat{S}_{j,t}$ generated by the summarization stage. Each subcategory is assessed independently using a dedicated prompt, yielding binary outcomes that indicate whether a given motivation supports that category. A single prompt example, corresponding to the *external* category, is reported below; the remaining category prompts are presented in Appendix C.2.

English translation of the classification prompt (External / EXT):

Task: Classify a fiscal measure based on the stated motivations. Does this measure correspond to the label "External (EXT)"?

Definition of External (EXT):
A passage is classified as External if the measure is motivated by external economic pressures, trade obligations, or international agreements.

This includes:

- Measures adopted in response to balance-of-payments problems, exchange rate pressures, or the need to protect reserves.
- Tax modifications explicitly linked to supporting exports, foreign trade operations, or competitiveness in international markets.
- Adjustments mandated by commitments to international organizations or treaties (e.g., WTO, EU, EFTA).
- Policy changes justified by concerns about investor confidence or market reactions to the government's financing needs.

*Do **not** classify as External if:*

- The motivation is purely domestic (e.g., deficit reduction, consolidation, demand management, social equity, environment).
- The change is a routine adjustment unrelated to trade, finance, or international agreements.
- The measure concerns only domestic consumption, revenue, or distribution without an external framing.

Output format:

- Respond with one option only for the entire set of motivations (do not provide multiple answers):
  - Yes — [short rationale]
  - No — [short rationale]

Prompt 2: Classification Prompt for the "External (EXT)" Category

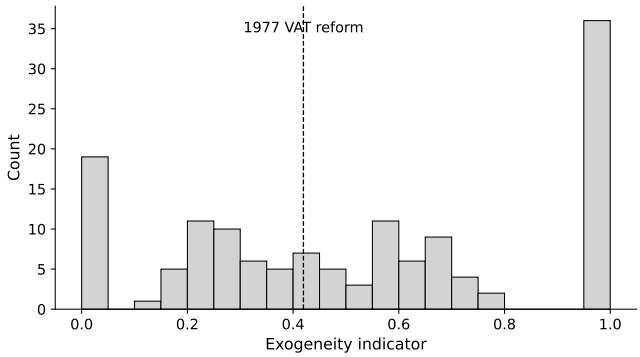


Figure 3: Distribution of Exogeneity Shares Across VAT Measures

Note: The histogram reports the distribution of exogeneity shares constructed from singleton marginal narrative contributions. Mass at zero corresponds to measures whose motivations are entirely endogenous, while the spike at one reflects measures classified as fully exogenous. Intermediate values indicate mixed motives and partial exogeneity. Shares are based on the max-assignment rule described in the text.

5.2 Constructing the VAT shock

5.2.1 Source Augmented Narrative Shock

To construct the exogenous share for each VAT measure, I apply the marginal narrative contribution framework of Section 3 to the calibrated classification scores obtained for every summarized motivation. In principle, the exogenous share would combine not only the singleton effects of individual motivations but also their interaction effects when presented jointly. However, preliminary exploration indicates that higher-order interactions tend to blur the motivation-specific signal in ways that are not yet well understood and are the subject of ongoing investigation. I therefore rely solely on singleton marginal effects: for each motivation, the category with the highest calibrated score is selected via a simple max rule, and measure-level exogenous support is computed by aggregating these assignments across motivations. The resulting exogenous shares display substantial heterogeneity across VAT measures, ranging from fully endogenous to fully exogenous cases, with a broad distribution in between. Figure 3 documents this empirical distribution for the universe of VAT changes.

Figure 4 plots the resulting narrative-weighted exogenous VAT shock series over the full sample, alongside the corresponding series of all legislated VAT changes. The exogenous series still displays pronounced spikes, while remaining comparatively muted during periods in which VAT changes were motivated primarily by short-run fiscal or stabilisation concerns. The 1968 shock is muted because its stated motives of revenue compensation, financing deduction expansions, and sectoral alignment were fully endogenous.

5.2.2 A text-augmented VAT policy rule

Beyond delivering narrative shocks, the SANI outputs also permit a parametric characterisation of systematic VAT behaviour. By treating the calibrated motivation intensities as observable components of the policymaker's information set, I estimate a text-augmented policy rule that complements the narrative identification. Following the policy-rule framework introduced in Section 1, and in the spirit of the text-augmented approach of Aruoba and Drechsel (2025), the unobserved information set is

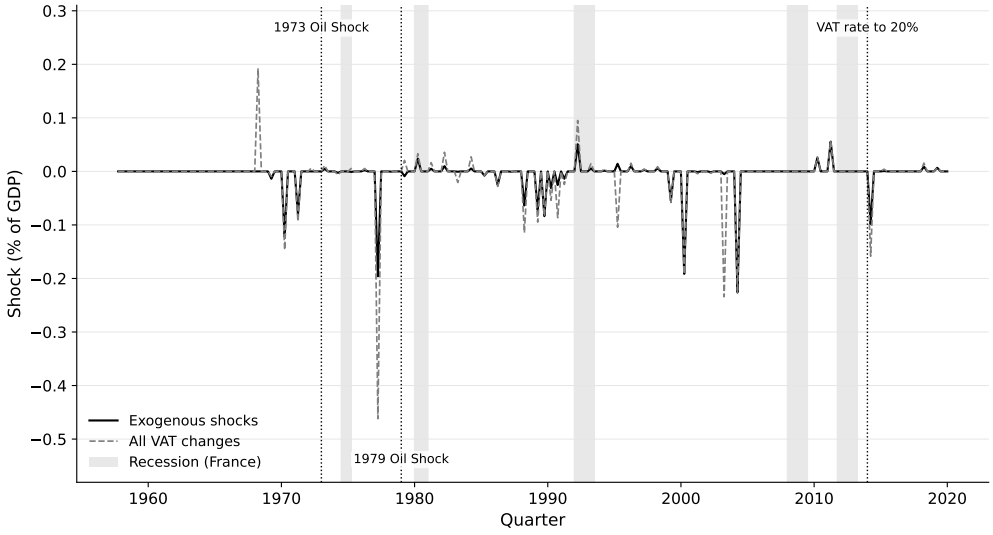


Figure 4: Time Series of Narrative-Weighted Exogenous VAT Shocks and All VAT Changes

Note: The figure plots the time series of narrative-weighted exogenous VAT shocks together with the corresponding series of all VAT changes. Grey shaded areas denote recession periods in France, while vertical dotted lines mark major macroeconomic and fiscal events. Shocks are expressed as a percentage of GDP.

approximated using both macro-fiscal variables and the qualitative motives extracted by SANI. This allows systematic VAT behaviour to be modelled jointly from quantitative fundamentals and stated motivations, while SANI continues to isolate the unsystematic component through narrative evidence.

To operationalise this policy rule, I first introduce the relevant observable covariates. Let X_t denote a vector of macro-fiscal controls (output gap, inflation, deficit, and lagged VAT variables). For each VAT measure (j,t) and each classification category $k \in \{1,\dots,K\}$, the randomised source-dependence procedure produces marginal source contributions $\{\hat{w}_{j,t,\ell}^{(k)}\}_{\ell=1}^{L_{j,t}}$, where $\hat{w}_{j,t,\ell}^{(k)}$ measures how strongly motivation $M_{j,t,\ell}$ shifts the log-odds of assignment to category k.[26]

Direction-adjusted narrative intensities. The marginal source contributions lie in $[0,1]$ and measure the strength of narrative support for category k, independently of whether the VAT measure raises or reduces the tax burden. In the policy-rule regression, however, the effect of a motivation must reflect the economic direction of the measure. I therefore construct the category–specific intensity as

$$m_{j,t}^{(k)} = \operatorname{sign}(\Delta\tau_{j,t}) \sum_{\ell=1}^{L_{j,t}} \hat{w}_{j,t,\ell}^{(k)}, \qquad k = 1,\dots,K. \tag{8}$$

Thus, the magnitude of $m_{j,t}^{(k)}$ increases with the cumulative narrative support for category k, while its sign reflects whether the measure raises or lowers the VAT burden.

Collecting these into $m_{j,t} = (m_{j,t}^{(1)},\dots,m_{j,t}^{(K)})'$, I estimate the following text-augmented VAT policy rule:

$$\Delta\tau_{j,t} = \alpha + \beta' X_{t-1} + \gamma' m_{j,t} + \delta' Z_t + u_{j,t}, \tag{9}$$

[26]See the subsection on the randomised source-dependence method for details of the log-odds surrogate model and the definition of marginal source contributions.

where β captures systematic responses to lagged macro-fiscal conditions, γ loads the direction-adjusted narrative intensities, and $u_{j,t}$ is the innovation. The vector Z_t collects contemporaneous changes in other tax instruments, such as personal and corporate income taxes, constructed from the same dataset. Including Z_t filters out systematic comovement across fiscal instruments and isolates the component of VAT policy uniquely associated with its own narratives. The fitted values represent the systematic component of VAT policy conditional on $(X_{t-1}, m_{j,t}, Z_t)$, and the residual $u_{j,t}$ is interpreted as the VAT policy shock orthogonal to both quantitative and narrative information.

In the current version of the paper, I do not yet report estimates of (9); the empirical analysis instead relies on the baseline source-augmented narrative shocks.

5.3 Macroeconomic effects of exogenous VAT shocks

Econometric Specification. To estimate the dynamic effects of fiscal shocks, we rely on a local projection framework in the spirit of Jordà (2005). Specifically, for each horizon h, we estimate the following regression:

$$\Delta^h y_{t+h} = \alpha_h + \beta_h \Delta\tau_t + \Gamma_h X_t + u_{t+h}, \tag{10}$$

where $\Delta^h y_{t+h}$ denotes the h-period-ahead change in the variable of interest, $\Delta\tau_t$ is the identified tax shock, and X_t is a vector of controls including six lags of real GDP growth and inflation. The coefficients β_h trace the impulse response of the outcome variable to the tax schock. Standard errors are computed using Newey–West corrections to account for serial correlation across horizons.

The shock as an instrumental variable. A potential concern with narrative or constructed fiscal shocks is that they are only imperfect proxies for the true structural innovations. As emphasized by Stock and Watson (2018), such measures typically contain noise or capture only part of the intended disturbance, so treating them as exact shocks can induce bias. In our case, the shocks are derived from ex-ante estimations by the French Treasury made at the time of policy design. These estimates are plausibly well correlated with the underlying fiscal innovations but remain imperfect reflections of their true ex-post realizations. Accordingly, the appropriate interpretation of these series is as external instruments that provide exogenous variation correlated with the structural fiscal shock but uncorrelated with other macroeconomic disturbances. To address this issue, an instrumental-variable version of the local projection (LP-IV) is planned.

Data and Detrending. All data series are obtained from the French national accounts published by INSEE. We use quarterly data on GDP, its expenditure components, and the GDP deflator, ensuring internal consistency across the usual national accounting identity. All nominal aggregates—including GDP and its subcomponents—are deflated by the GDP deflator to obtain real series. We then remove long-term trends following the procedure of Gordon and Krenn (2010): (i) we regress real GDP on a quadratic time trend, and (ii) divide all real variables by the fitted quadratic trend. This transformation expresses each component as a percentage of trend GDP, thereby preserving the exact decomposition of GDP into its constituent parts. Compared to logarithmic detrending, the Gordon–Krenn approach has the advantage of maintaining the accounting identity across the disaggregated components of GDP. In

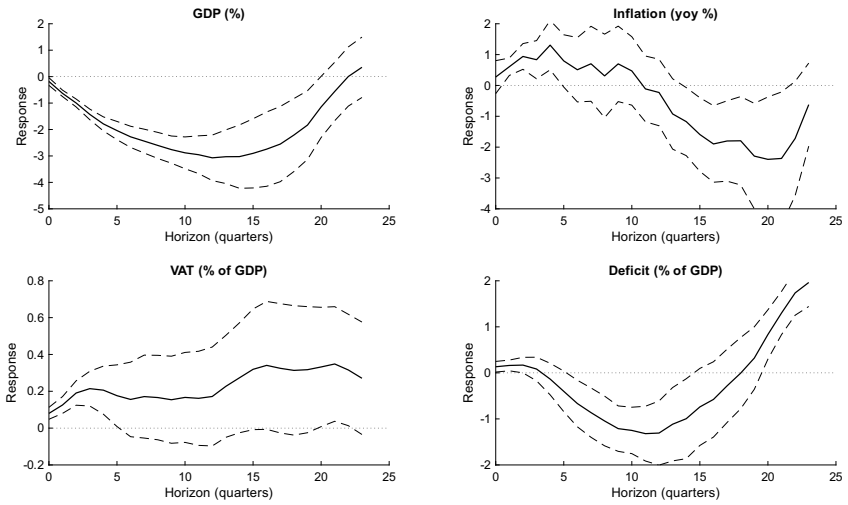


Figure 5: Responses to a +1%-of-GDP shock in VAT (68% Confidence Bands)

5.3.1 Empirical results

Baseline specification. Figure 5 presents the estimated responses of real GDP, inflation, VAT revenues, and the government deficit to a VAT increase of one percent of GDP. Real GDP declines persistently, reaching a trough of about three percent below trend around the twelfth quarter, before recovering gradually and returning to zero by roughly the twenty-second quarter. Inflation exhibits a short-lived spike: it rises sharply on impact, peaking around the fourth quarter as the higher indirect tax is passed through to consumer prices, and then declines steadily toward zero.

VAT revenues increase mechanically on impact, and the deficit improves for the same reason, before deteriorating once the fall in GDP depresses the broader tax base. As activity normalizes, the deficit gradually returns toward balance.

Overall, the results indicate that a VAT increase generates a moderate but prolonged contraction in output and a brief inflationary surge. The magnitude of the GDP response is strikingly similar to that reported by Romer and Romer (2010) for broad-based tax increases, which also produce output losses of around three percent.However, VAT is generally regarded as less distortive than the income and profit taxes that dominate the Romer–Romer measures, so one might have expected a weaker output response.

The fact that both tax changes produce effects of comparable size might be driven by differences in the monetary-policy environment surrounding the two shocks. Broad-based tax increases in Romer and Romer (2010) could plausibly trigger monetary accommodation that mitigates the short-run contraction, whereas a VAT hike generates a short-lived but pronounced spike in inflation, which may limit the scope for such accommodation. While this interpretation is only suggestive, it offers a consistent explanation for why a conceptually less distortive tax change produces output effects similar to those observed following broader tax hikes.

6 External Validation: The Romer and Romer (2010) Benchmark

This section externally validates the proposed framework using the canonical narrative tax series of Romer and Romer (2010). The Romer–Romer classification provides a natural benchmark: it is widely used in empirical macroeconomics, is based on careful reading of contemporaneous policy documents, and has well-established macroeconomic implications.

The validation proceeds in two steps. I first assess whether large language models can recover the Romer–Romer narrative identification when applied to the same narrative evidence used by the original authors. This benchmark exercise evaluates whether the decision rules underlying narrative identification can be performed in a transparent and replicable way, and whether doing so preserves the empirical content of the original shock series. I then extend the analysis by applying the full SANI procedure to U.S. policy documents, thereby relaxing the assumption that curated narratives are available ex ante.

6.1 Validation using the Romer–Romer narratives

This subsection evaluates whether large language models can reproduce the narrative identification of Romer and Romer (2010) when applied to the same narrative summaries used in the original study. The goal is not to reinterpret the historical record, but to assess whether an LLM-based procedure can recover expert judgments on policy exogeneity and, crucially, preserve their macroeconomic implications.

By conditioning on the Romer–Romer narratives, this exercise focuses on the implementation of the narrative decision rules themselves, abstracting from issues related to document retrieval or summarization. This benchmark therefore provides a stringent and transparent test of whether narrative identification can be automated without altering its economic content.

6.1.1 Setup

The Romer–Romer replication material documents 50 major federal tax changes in the United States between 1945 and 2007. For each episode, the authors provide a detailed narrative discussion of the motivations underlying the reform, drawing on contemporaneous sources such as presidential speeches, budget documents, and Congressional records. Each narrative concludes with a classification of the tax change as either exogenous or endogenous.

To construct the benchmark validation task, I extract the narrative text describing each policy episode and remove the portion in which the authors explicitly state and justify their classification. In particular, I delete the sentence reporting the final classification along with the two immediately preceding sentences, which often contain direct cues revealing the intended label (see Appendix D.1.1 for examples). The remaining narrative text is used as input to the language model.

The model is prompted to classify each tax change as exogenous or endogenous based solely on the narrative evidence provided. This design ensures that the LLM is not exposed to the authors' conclusions, allowing for a clean comparison between human-coded and model-generated classifications.

6.1.2 Main classification results

Table 5 reports the classification performance of several state-of-the-art large language models, including both closed-source frontier models and an open-source alternative. The closed-source models correctly recover approximately 82% of the Romer–Romer classifications, while the open-source model achieves an accuracy of 70%. These results indicate that LLMs can recover a substantial share of expert narrative judgments using transparent and replicable prompts.

Table 5: LLM Classification Performance Summary

LLM Model	**False Endog. (%)**	**False Exog. (%)**	**Overall Accuracy (%)**
GPT-4o-mini (Closed)	31%	4%	82%
Gemini-2.0-flash (Closed)	35%	0%	82%
Llama3.2:4B (Open)	54%	4%	70%

Beyond overall accuracy, the composition of classification errors is particularly informative. Most errors take the form of false negatives, in which tax changes classified as exogenous by Romer and Romer are instead labeled endogenous by the model. By contrast, false positives—endogenous policies misclassified as exogenous—are rare. This asymmetric error structure implies that the LLM-based procedure is conservative, tending to exclude borderline episodes rather than incorrectly admitting endogenous policy changes as exogenous.

From the perspective of empirical identification, this pattern is economically meaningful. The primary threat to narrative identification is the inclusion of policy changes that respond endogenously to contemporaneous macroeconomic conditions, which would invalidate the resulting shock series as an instrument. The scarcity of false positives therefore indicates that the LLM-based procedure preserves the core identifying restriction emphasized by Romer and Romer (2010). By contrast, false negatives mainly reduce the number of usable shocks, potentially weakening first-stage power but not introducing bias into estimated responses. Overall, this validation exercise suggests that automation primarily trades off statistical power for robustness, without materially altering the economic content of the narrative identification.

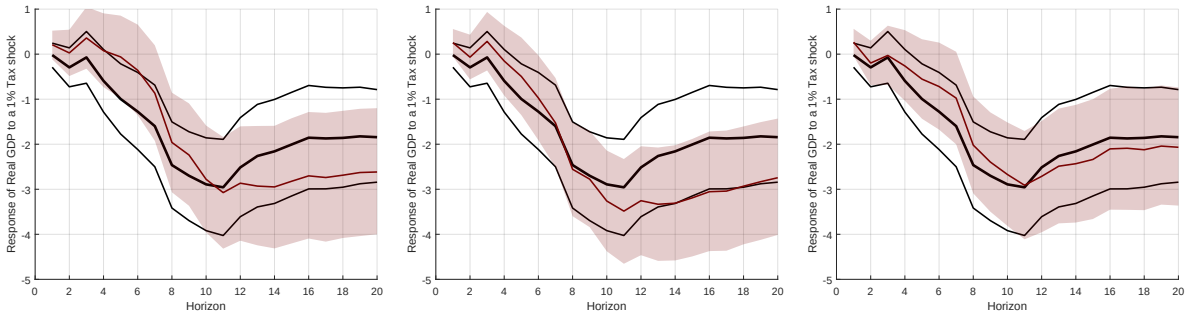
Appendix D.2 reports a set of robustness checks, including sensitivity to prompt design and tests for potential training-data leakage, which confirm that the classification results are not driven by prompt overfitting or memorization. Appendix D.3 further analyzes the confidence of the LLM classifications using diagnostics derived from token-level output probabilities, shedding light on the sources of remaining classification errors.

6.1.3 Empirical implications

I next assess whether the LLM-based classifications preserve the empirical implications of the Romer–Romer shock series. Following Romer and Romer (2010), I construct a quarterly series of exogenous tax changes expressed as a percentage of GDP at the time of implementation and estimate a two-variable VAR with log real GDP and the tax shock series, ordering tax shocks first.

Figure 6 compares the impulse responses obtained using the original Romer–Romer series with those derived from the LLM-based classifications. Across model variants, the estimated responses

of output to a tax increase are close to the benchmark results. While the LLM-based responses are slightly attenuated at short horizons, they converge to—and in some cases exceed—the contractionary effects estimated by Romer and Romer at longer horizons.



(a) GPT-4o-mini (b) Gemini-2.0-flash (c) llama3.2

Figure 6: Impact of Tax Changes on Real GDP across Classifications

Note: Black lines reproduce the original results from Romer & Romer (2010), Section IV.B. Red lines represent the empirical estimates obtained using the LLM-based classifications. Following the original paper, 68% confidence bands are reported.

These results contrast sharply with the attenuated responses obtained when broader measures of tax changes are used, such as cyclically adjusted revenues or all legislated tax changes, which Romer and Romer argue are contaminated by endogenous policy responses. Consistent with the conservative error pattern documented above, the LLM-based instruments appear to preserve the identifying variation emphasized by Romer and Romer, yielding macroeconomic responses that are close to the canonical benchmark.

Taken together, these results show that large language models can recover the Romer–Romer narrative identification and its macroeconomic implications when applied to curated narrative summaries. However, this benchmark exercise abstracts from the earlier stages of the SANI procedure, in particular the retrieval and summarization of policy documents. The next subsection therefore extends the validation by applying the full procedure to U.S. policy documents, assessing whether similar conclusions obtain under more realistic data conditions.

6.2 End-to-end validation using U.S. policy documents

The previous subsection shows that large language models can reproduce the narrative identification of Romer and Romer (2010) when applied to the curated narrative summaries assembled by the original authors. I now extend the validation by applying the full SANI procedure to U.S. federal tax policy using contemporaneous policy documents, thereby relaxing the assumption that narrative summaries are available ex ante. This exercise provides an end-to-end assessment of the framework in a canonical benchmark environment.

Importantly, the purpose of this exercise is not to construct a competing narrative tax series for the United States, nor to replicate the Romer and Romer classifications using automated methods. The objective is instead to provide an initial validation of the SANI framework by applying it to federal tax changes that fall outside the scope of existing narrative datasets and assessing whether the resulting identifying variation yields macroeconomic implications comparable to established benchmarks.

Specifically, I identify a set of federal tax measures that are not documented by Romer and Romer (2010), assemble contemporaneous policy documents associated with these measures, and use SANI to classify their stated motivations and construct exogenous tax shocks. The macroeconomic effects implied by these shocks are then compared to those obtained using the Romer–Romer narrative series. This comparison provides an indirect validation of the framework, while extending the full SANI procedure to the Romer and Romer tax episodes themselves remains an important direction for future work.

Data and SANI implementation. I start from the extended annual dataset of federal tax measures constructed from the *Budget of the United States Government*, as described in Section 4.2. This dataset captures tax changes introduced through the regular budgetary process and therefore excludes most of the large, stand-alone legislative acts studied by Romer and Romer (2010).

For each selected measure, I use contemporaneous budget documentation to extract stated policy motivations and apply the SANI framework to classify these motivations as exogenous or endogenous. The implementation follows the same settings adopted in the analysis of French VAT policy; details are provided in Section 5.1. At this stage, the classification relies exclusively on information contained in the budget documents and does not incorporate additional contemporaneous sources such as the *Economic Report of the President* or Treasury annual reports.

Figure 7 summarizes the distribution of exogeneity shares across the budget-based tax measures classified using SANI. A substantial mass is concentrated at zero, with more than 500 tax changes classified as fully endogenous. These measures are therefore excluded from the construction of the exogenous tax shock series. The remaining tax changes are distributed over the interior of the unit interval, reflecting varying degrees of mixed motivations and partial exogeneity. This dispersion highlights the prevalence of ambiguous or multi-purpose policy rationales in routine budgetary tax adjustments and underscores the conservative nature of the classification procedure.

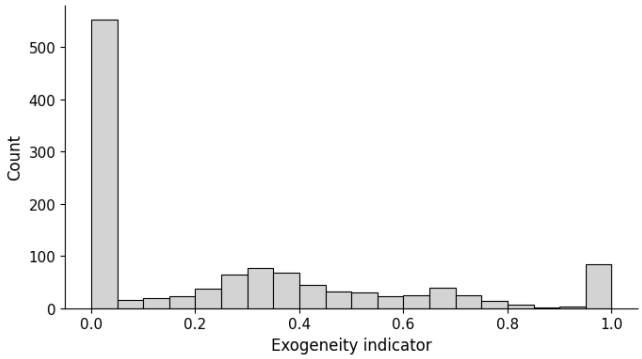


Figure 7: Distribution of Exogeneity Shares Across VAT Measures

Note: The histogram reports the distribution of exogeneity shares constructed from singleton marginal narrative contributions. Mass at zero corresponds to measures whose motivations are entirely endogenous, while the spike at one reflects measures classified as fully exogenous. Intermediate values indicate mixed motives and partial exogeneity. Shares are based on the max-assignment rule described in the text.

Figure 8 plots the quarterly time series of budget-based tax changes together with the narrative-weighted shocks implied by the SANI classifications. The dashed black line reports the aggregate series that treats all legislated tax changes recorded in the annual budget symmetrically. The solid black

line rescales each observation by its estimated exogeneity weight, which both attenuates mixed-motive measures and sets to zero those tax changes classified as fully endogenous. As a result, several sizeable movements in the raw budget series disappear entirely once narrative information is incorporated. For example, large tax changes observed in the mid-1950s, the early 1960s, and the mid-1970s are fully eliminated in the narrative-weighted series, reflecting their classification as endogenously motivated responses to prevailing macroeconomic conditions. The red line overlays the Romer–Romer narrative tax shock series for comparison.

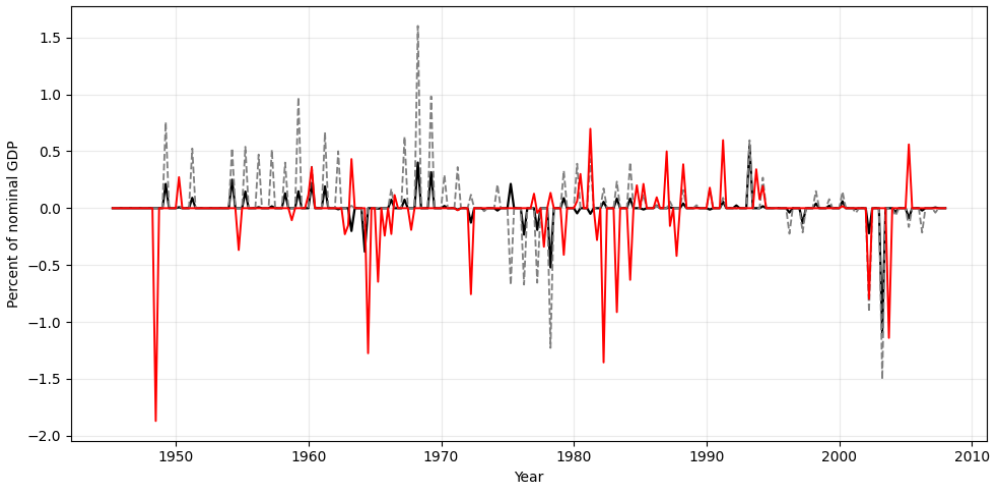


Figure 8: Budget-Based Tax Changes, Narrative-Weighted Shocks, and Romer–Romer Tax Shocks

Note: The dashed black line reports the aggregate series of all legislated federal tax changes appearing in the annual budget, expressed as a percentage of GDP. The solid black line reports the corresponding series reweighted by the estimated exogenous share produced by the SANI procedure, so that each tax change is scaled by the degree to which its stated motivations are classified as exogenous. The red line plots the Romer–Romer (2010) exogenous tax shock series for comparison. All series are dated at the time of implementation and expressed as a percentage of contemporaneous GDP.

Empirical implications. To assess whether the shocks generated by SANI carry economically meaningful identifying variation, I estimate their macroeconomic effects using the same empirical framework employed in Romer and Romer (2010). The resulting impulse responses are then compared to the benchmark responses obtained using the Romer–Romer narrative shocks.

Despite differences in policy composition and in the types of tax measures considered, the estimated responses display similar qualitative features. Output declines persistently following an exogenous tax increase in both cases, with effects that build up over time and remain highly persistent. At short to early medium horizons, the magnitude of the contraction is comparable to that implied by the Romer–Romer narrative shocks, while at longer horizons the responses based on budget-derived exogenous tax measures are more pronounced. Consistent with this interpretation, Appendix D.4 shows that, within each framework, including all tax changes leads to substantially attenuated responses relative to specifications that isolate exogenous measures.

Augmenting Romer–Romer shocks with SANI-identified measures. As a complementary exercise, I construct a combined tax shock series that augments the original Romer–Romer (RR)

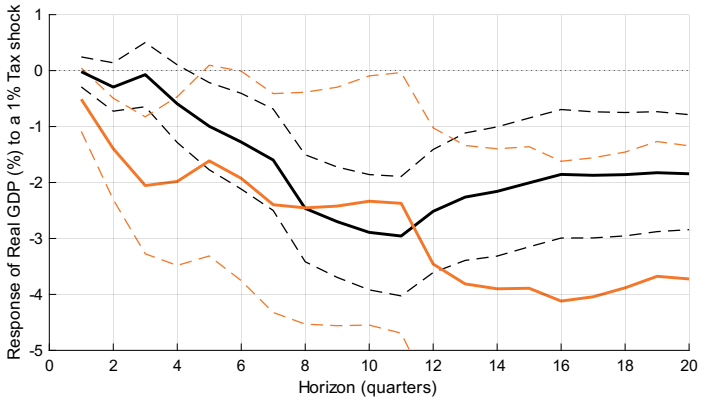


Figure 9: Comparison of Romer–Romer and Budget-Based Tax Shocks

Note: The black curve reports impulse responses obtained using the Romer–Romer (RR) exogenous tax shock series. The red curve corresponds to impulse responses estimated using exogenous tax measures constructed from U.S. federal budget reforms and classified using the SANI procedure. Shaded areas denote 68% confidence intervals.

exogenous tax shocks with the exogenous tax measures identified by SANI outside the RR sample. The RR shocks are left unchanged, and the SANI-based shocks enter only in periods not covered by the original narrative series. The purpose of this exercise is not to alter the identifying assumptions underlying the RR shocks, but to assess whether SANI adds identifying variation that is economically consistent with the benchmark while improving effective sample size.

Figure 10 reports the impulse responses obtained using the combined shock series. The estimated responses closely track those implied by the original Romer–Romer shocks. Output declines persistently following a tax increase, with a similar timing and degree of persistence across specifications. Differences in point estimates are small relative to sampling uncertainty at all horizons.

At the same time, the inclusion of SANI-identified exogenous tax measures leads to a modest but systematic reduction in standard errors, particularly at short and medium horizons. This tightening of confidence bands reflects the additional identifying variation contributed by routine budgetary measures, while leaving the qualitative and quantitative features of the responses largely unchanged.

Overall, these results indicate that SANI extends the support of the Romer–Romer shock series without materially shifting the implied macroeconomic effects. The framework primarily improves precision by increasing coverage, rather than altering the estimand identified by the canonical narrative approach.

Discussion. Taken together, these results provide an initial validation of the SANI framework by showing that narrative shocks constructed from budget-based tax measures yield macroeconomic implications consistent with those obtained from canonical narrative series. While the current analysis does not apply the full SANI procedure to the Romer and Romer tax episodes themselves, the similarity in estimated responses suggests that the framework is able to recover economically relevant identifying variation even in settings characterized by routine, lower-salience policy changes. Extending the end-to-end application of SANI to stand-alone major tax acts remains an important direction for future work.

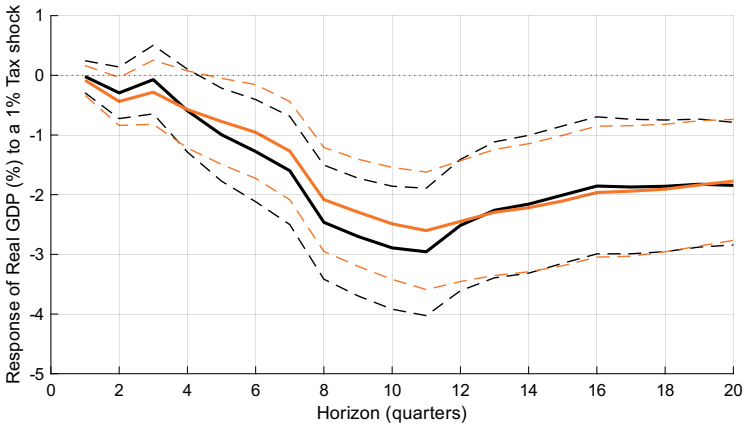


Figure 10: Impulse Responses to Tax Shocks: Romer–Romer vs. Romer–Romer Augmented with SANI

Note: The black line reports impulse responses obtained using the original Romer–Romer exogenous tax shock series. The red line corresponds to a combined series that augments the Romer–Romer shocks with exogenous tax measures identified by SANI outside the original narrative sample. Shaded areas denote 68% confidence intervals.

7 Conclusion

This paper develops a scalable and transparent framework for narrative identification, SANI (Source-Augmented Narrative Identification), which combines semantic retrieval and large language models to recover policymakers' stated motives from historical text. By formalizing the qualitative steps underlying traditional narrative analysis, the framework extends narrative identification to large document collections while maintaining the source grounding required for credible causal inference.

A comprehensive validation using the canonical narrative tax series of Romer and Romer (2010) shows that the proposed framework preserves the identifying content of established narrative designs. When applied to the original narrative evidence and decision rules, large language models recover the Romer–Romer classifications with high accuracy and a strongly conservative error structure. Crucially, the resulting tax shock series reproduce the macroeconomic responses implied by the original shocks, and augmenting the Romer–Romer series with SANI-identified measures outside the original sample improves precision without materially altering estimated effects. Together, these results demonstrate that automation can scale narrative identification while preserving its economic content.

Applying SANI to the full history of French VAT policy since 1960, I construct a new narrative series of exogenous VAT changes and estimate their macroeconomic effects. The results indicate that exogenous VAT increases generate a sizeable and persistent contraction in activity, with output falling sharply, investment responding strongly, and inflation rising temporarily due to the mechanical pass-through to consumer prices. These dynamics highlight the distinctive transmission channel of indirect taxes relative to income-based instruments.

More fundamentally, the framework recasts narrative identification as a continuous rather than binary exercise. By quantifying the degree of exogenous support attached to each policy action, SANI eliminates the hard thresholds that traditionally restrict narrative series to a small number of large measures. This continuous design mitigates sparsity and yields narrative-weighted shocks that

more closely reflect the underlying structural objects motivating policy. These properties improve econometric precision while preserving the interpretive foundations of narrative identification.

Beyond the empirical findings, the paper shows that large language models can serve as disciplined empirical tools when embedded in a document-grounded workflow. The marginal narrative contribution diagnostics ensure that classifications remain anchored in identifiable sources, strengthening transparency and researcher oversight. By reducing the cost of incorporating qualitative evidence and standardizing the analysis of large textual corpora, SANI enhances replicability and expands the feasibility of narrative identification in settings where manual methods are prohibitively costly.

More broadly, the framework demonstrates how the expanding body of historical and contemporary policy documents can be integrated systematically into empirical macroeconomics. By combining computational scale with interpretability, SANI enables more granular narrative designs in fiscal, monetary, and structural policy, and provides a foundation for future work on text-augmented identification strategies.

References

Amodeo, F., & Briganti, E. (2025). High-Frequency Cross-Sectional Identification of Military News Shocks.

Aruoba, S. B., & Drechsel, T. (2025). Identifying Monetary Policy Shocks: A Natural Language Approach. *American Economic Journal: Macroeconomics*. https://doi.org/10.1257/mac.20240184

Ash, E., & Hansen, S. (2023). Text Algorithms in Economics. *Annual Review of Economics*, *15*(Volume 15, 2023), 659–688. https://doi.org/10.1146/annurev-economics-082222-074352

Auerbach, A. J., & Gorodnichenko, Y. (2012). Measuring the Output Responses to Fiscal Policy. *American Economic Journal: Economic Policy*, *4*(2), 1–27. https://doi.org/10.1257/pol.4.2.1

Bandeira, M., Carvalho, C., Bonomo, M., & Mendes, M. (2025). Contractionary Fiscal Stimulus (and Expansionary Austerity).

Barnichon, R., & Mesters, G. (2025). Innovations Meet Narratives -Improving the Power-Credibility Trade-off in Macro. *Working Papers*, (1475).

Battaglia, L., Christensen, T., Hansen, S., & Sacher, S. (2024, December). Inference for Regression with Variables Generated by AI or Machine Learning. https://doi.org/10.48550/arXiv.2402.15585

Benzarti, Y., & Carloni, D. (2019). Who Really Benefits from Consumption Tax Cuts? Evidence from a Large VAT Reform in France. *American Economic Journal: Economic Policy*, *11*(1), 38–63. https://doi.org/10.1257/pol.20170504

Benzarti, Y., Carloni, D., Harju, J., & Kosonen, T. (2020). What Goes Up May Not Come Down: Asymmetric Incidence of Value-Added Taxes. *Journal of Political Economy*, *128*(12), 4438–4474. https://doi.org/10.1086/710558

Blanchard, O., & Perotti, R. (2002). An Empirical Characterization of the Dynamic Effects of Changes in Government Spending and Taxes on Output. *The Quarterly Journal of Economics*, *117*(4), 1329–1368.

Brown, A., Roman, M., & Devereux, B. (2025, September). A Systematic Literature Review of Retrieval-Augmented Generation: Techniques, Metrics, and Challenges. https://doi.org/10.48550/arXiv.2508.06401

Carlson, J., & Dell, M. (2025, July). A Unifying Framework for Robust and Efficient Inference with Unstructured Data. https://doi.org/10.48550/arXiv.2505.00282

Cloyne, J. (2013). Discretionary Tax Changes and the Macroeconomy: New Narrative Evidence from the United Kingdom. *American Economic Review*, *103*(4), 1507–1528. https://doi.org/10.1257/aer.103.4.1507

Cohen-Wang, B., Shah, H., Georgiev, K., & Madry, A. (2024, September). ContextCite: Attributing Model Generation to Context. https://doi.org/10.48550/arXiv.2409.00729

Dabla-Norris, E., & Lima, F. (2023). Macroeconomic effects of tax rate and base changes: Evidence from fiscal consolidations. *European Economic Review*, *153*, 104399. https://doi.org/10.1016/j.euroecorev.2023.104399

Fei, Y., Hou, Y., Chen, Z., & Bosselut, A. (2023, August). Mitigating Label Biases for In-context Learning. https://doi.org/10.48550/arXiv.2305.19148

Feng, S., Wallace, E., Grissom II, A., Iyyer, M., Rodriguez, P., & Boyd-Graber, J. (2018). Pathologies of Neural Models Make Interpretations Difficult. In E. Riloff, D. Chiang, J. Hockenmaier, & J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

Processing (pp. 3719–3728). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1407

Friedman, M., & Schwartz, A. J. (1963). *A monetary history of the United States 1867-1960*. Princeton University Press.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature*, *57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Hamilton, J. D. (1985). Historical Causes of Postwar Oil Shocks and Recessions. *The Energy Journal*, *6*(1), 97–116.

Hansen, A. L., & Kazinnik, S. (2023). Can ChatGPT Decipher Fedspeak? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4399406

Jamilov, R. (2025). Measuring Conflict Inflation.

Jordà, Ò. (2005). Estimation and Inference of Impulse Responses by Local Projections. *American Economic Review*, *95*(1), 161–182. https://doi.org/10.1257/0002828053828518

Ludwig, J., Mullainathan, S., & Rambachan, A. (2025, January). *Large Language Models: An Applied Econometric Framework* (tech. rep. No. w33344). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w33344

Martell, E. (2025). Official Arm-Twisting? Measuring the Federal Reserve's Use of Moral Suasion.

Mertens, K., & Ravn, M. O. (2012). Empirical Evidence on the Aggregate Effects of Anticipated and Unanticipated US Tax Policy Shocks. *American Economic Journal: Economic Policy*, *4*(2), 145–181. https://doi.org/10.1257/pol.4.2.145

Mertens, K., & Ravn, M. O. (2013). The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States. *American Economic Review*, *103*(4), 1212–1247. https://doi.org/10.1257/aer.103.4.1212

Mountford, A., & Uhlig, H. (2009). What are the effects of fiscal policy shocks? *Journal of Applied Econometrics*, *24*(6), 960–992. https://doi.org/10.1002/jae.1079

Prabhakaran, V., Hutchinson, B., & Mitchell, M. (2019). Perturbation Sensitivity Analysis to Detect Unintended Model Biases. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 5740–5745). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1578

Ramey. (2011). Can Government Purchases Stimulate the Economy? *Journal of Economic Literature*, *49*(3), 673–685. https://doi.org/10.1257/jel.49.3.673

Ramey. (2016). Macroeconomic Shocks and Their Propagation. In *Handbook of Macroeconomics* (pp. 71–162, Vol. 2). Elsevier. https://doi.org/10.1016/bs.hesmac.2016.03.003

Ramey & Shapiro, M. D. (1998). Costly capital reallocation and the effects of government spending. *Carnegie-Rochester Conference Series on Public Policy*, *48*, 145–194. https://doi.org/10.1016/S0167-2231(98)00020-7

Ramey & Zubairy, S. (2018). Government Spending Multipliers in Good Times and in Bad: Evidence from US Historical Data. *Journal of Political Economy*, *126*(2), 850–901. https://doi.org/10.1086/696277

Riera-Crichton, D., Vegh, C. A., & Vuletin, G. (2016). Tax multipliers: Pitfalls in measurement and identification. *Journal of Monetary Economics*, *79*, 30–48. https://doi.org/10.1016/j.jmoneco.2016.03.003

Romer & Romer. (1989). Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz. *NBER Macroeconomics Annual*, *4*, 121–170. https://doi.org/10.1086/654103

Romer & Romer. (2004). A New Measure of Monetary Shocks: Derivation and Implications. *American Economic Review*, *94*(4), 1055–1084. https://doi.org/10.1257/0002828042002651

Romer & Romer. (2010). The Macroeconomic Effects of Tax Changes: Estimates Based on a New Measure of Fiscal Shocks. *American Economic Review*, *100*(3), 763–801. https://doi.org/10.1257/aer.100.3.763

Romer & Romer. (2023). Presidential Address: Does Monetary Policy Matter? The Narrative Approach after 35 Years. *American Economic Review*, *113*(6), 1395–1423. https://doi.org/10.1257/aer.113.6.1395

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, *48*(1), 1–48. https://doi.org/10.2307/1912017

Stock, J. H., & Watson, M. W. (2018). Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments. *The Economic Journal*, *128*(610), 917–948. https://doi.org/10.1111/ecoj.12593

Uhlig, H. (2005). What are the effects of monetary policy on output? Results from an agnostic identification procedure. *Journal of Monetary Economics*, *52*(2), 381–419. https://doi.org/10.1016/j.jmoneco.2004.05.007

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021, June). Calibrate Before Use: Improving Few-Shot Performance of Language Models. https://doi.org/10.48550/arXiv.2102.09690

8 Appendix

A SANI

A.1 Embedding Model and Distance Metric

Semantic retrieval relies on vector representations of text segments and policy descriptions generated by a transformer-based embedding model. In the baseline implementation, embeddings are obtained using OpenAI's `text-embedding-3-large` model, which maps each text input to a 3,072-dimensional dense vector in $\mathbb{R}^{3072}$. The model is trained to capture semantic relationships between texts such that semantically similar passages are represented by vectors with small angular separation.

For each policy description p_j and document segment $d_k \in \mathcal{D}$, the corresponding embedding vectors q_j and v_k are normalized to unit length. Semantic proximity is then evaluated using the cosine distance,

$$s_{j,k} = 1 - \frac{q_j \cdot v_k}{\|q_j\| \, \|v_k\|},$$

so that smaller values of $s_{j,k}$ indicate greater semantic similarity. Semantic proximity between text embeddings is measured using the cosine distance. Because all embeddings are normalized to unit length, cosine and Euclidean distances are mathematically equivalent in this setting, yielding identical retrieval rankings under either metric. Distances are computed through standard nearest-neighbor search routines provided by the underlying vector store, which efficiently identifies the text segments closest to each policy query in the embedding space. This normalization ensures that retrieval reflects semantic similarity rather than differences in vector magnitude. The retrieval operator defined in Equation (5) in the main text therefore selects, for each policy event j, the K document segments with the smallest cosine distances to the corresponding policy query q_j.

A.2 Example details

Applying the retrieval operator $\mathcal{R}_{K,L}$ to the query $q_{j,t}$ and corpus $\mathcal{D}_t$, I obtain the top-K most relevant passages. Here are the top-5 passages:

- **Chunk 1.** On December 31, 1976, to mitigate, by a reduction of the weight of taxes bearing on consumption, the upward movement likely to occur at the end of this phase of price freeze; — it constitutes the last stage of a reform undertaken since 1969 — the date on which the standard rate of VAT was 23.456% — which results, ultimately, in a reduction, in less than seven years, of one quarter of the standard rate and in the merging of the latter with the intermediate rate.
- **Chunk 2.** An exceptional contribution would be requested in 1977 from the wealthiest French; it would affect all of them and only them, contrary to what had been provided for in the deleted article 5. This taxation would have as its base the total of the bases corresponding to the elements of lifestyle enumerated in article 168 of the General Tax Code, with the exception of main residences and cars of a power equal to or less than 16 horsepower. The income-equivalent threshold, calculated according to the norms of the aforementioned article 168, above which the exceptional contribution applies, is set at 60,000 F and the number of accumulated elements provoking taxation is three. Your Finance Committee proposes that you adopt the present article. VAT (Value Added Tax) – Article 6. Reduction of the standard rate and unification of

the standard rate and the intermediate rate. Text. — I. — The standard rate of value-added tax is set at 17.60%. In the departments of Guadeloupe, Martinique and Réunion, it is set at 7.50II. — The upper limit of the special rebate for artisans fixed in paragraph 3 of article 282 of the General Tax Code is raised to 20,000 F. III. — The provisions of the present article enter into force on January 1, 1977. Comments. — It is proposed in the present article to lower by 2.4 points, starting from January 1, 1977, the standard rate of value-added tax which is currently 20% and, thereby, to bring it down to the level of the intermediate rate, namely 17.60%. This measure fits within a double perspective: — it is part of the program of combating inflation and is of a nature to permit, after the price freeze applied during the period from September 15 to December 31, 1976, to mitigate, by a reduction of the weight of taxes bearing on consumption, the upward movement likely to occur at the end of this phase of price freeze;

- **Chunk 3.** The number of rates applied in the matter of value-added tax, starting from next January 1, will be reduced to three instead of four; there would remain only: — the increased rate (33 %: cars, motorcycles, photography and cinematography equipment, sound and image recording devices, precious stones and precious metals, etc.); — the standard rate (17.6%: energy products, beverages, services and all transactions hitherto subject to the 20% rate, that is to say all those which were not taxed at the increased rate, nor at the intermediate rate, nor at the reduced rate); — the reduced rate (7%: essentially food products, pharmaceutical products, water and fertilizers). At the same time, it is planned to carry out a reduction of 2.5 points of the standard VAT rate and to align it with the intermediate rate in the departments of Guadeloupe, Martinique and Réunion, French Guiana not being concerned since the value-added tax regime is not applied there. In the three departments concerned, where, due to constraints inherent in the economies of developing countries, VAT rates are significantly lower than those of Metropolitan France, starting from January 1, 1977, one will record: — the increased rate: 14%; — the standard rate: 7.5%; — the reduced rate: 3.5%. In order to support the artisanal professions, it is proposed in paragraph II of the present article to raise, starting from January 1, 1977, the upper limit of the special rebate from which they benefit, raising it from 13,500 F to 20,000 F. Let us recall that, in order to lighten the burdens of small taxpayers, and in particular artisans, article 282 of the General Tax Code has provided, in their favor, special provisions resulting in a waiver or a reduction of the VAT normally incumbent upon them.

- **Chunk 4.** Standard of 6.5% per year. b) The elimination of the levy. As had been provided by the Finance Law for 1976, in the initial text presented by the Government, the levy was to cease to apply when the progression of prices was below a set norm: this one, in the governmental project, was fixed at 2.50% for the consumer price index of the group of manufactured products during a period of six consecutive months.

- **Chunk 5.** Article: — in paragraph I, the objective is to encourage business leaders to convert into capital the advances they are led to grant in the form of deposits in current accounts to the companies they manage. Up to now, in such cases, the ordinary contribution right was a proportional duty of 1%; it is planned to substitute for it a fixed duty of 220 F on the condition, however, that the sums thus incorporated into the capital have been placed at the constant disposal of the company, for a minimum period of twelve months.

A.3 Technical Implementation Details

The deterministic implementation relies on several technical specifications to ensure reproducibility. Both $\mathcal{S}$ and $\mathcal{C}$ operators use fixed prompt templates with the following decoding parameters: temperature = 0, top-p = 1.0, no nucleus sampling, and fixed maximum token limits. For transformer models, this eliminates stochastic sampling during generation.

For open-source models with frozen weights (e.g., LLaMA, Mistral family), deterministic decoding combined with identical computational environments guarantees bitwise-identical outputs across runs. This provides the strongest form of reproducibility, as the complete model architecture, weights, and inference code are publicly auditable.

For proprietary APIs (e.g., OpenAI GPT models, Google Gemini), additional logging is implemented to track potential sources of non-reproducibility. Each API call logs: (i) the specific model identifier and version, (ii) SHA-256 hashes of the prompt template and input content, (iii) all decoding parameters, (iv) timestamp and API response metadata, and (v) the complete generated output. This comprehensive logging enables detection of silent model updates or API changes that could affect results.

Even with deterministic settings, proprietary models may exhibit minor non-determinism due to distributed inference infrastructure, floating-point precision differences, or undocumented system changes. Additionally, model providers may silently update their systems, potentially altering outputs for identical inputs. While such changes are typically small, they can accumulate across large-scale classification tasks. The metadata logging system provides an audit trail that enables researchers to verify whether results were generated under identical conditions and to detect potential sources of variation in replication attempts.

B Marginal Narrative Contributions

B.1 Robustness of the Linear Surrogate Estimator

Because the surrogate model is estimated with $n = 64$ ablations and up to $K \approx 30$ retrieved passages, I assess whether the resulting coefficients are sensitive to the choice of regularisation. Ridge regression yields estimates that are nearly identical to the OLS baseline and improves numerical stability when inclusion patterns exhibit collinearity. A LASSO with $\lambda = 0.01$ selects the same high-weight passages as OLS and produces coefficients of similar magnitude. These results indicate that the dependence patterns reported in the main text are not driven by estimator choice or overfitting. For settings with substantially larger K, penalised estimators would provide a natural extension.

B.2 Global Placebo Test for Summarisation Grounding

The coefficient-level placebo test in the main text evaluates, for each passage individually, whether its positive marginal contribution is unusually large relative to the distribution of positive placebo contributions. Because this diagnostic is applied to multiple passages within each event, some significant cells may be mechanically expected under the null of no grounding. This appendix reports an omnibus test that aggregates evidence across passages and verifies that the overall pattern of significant contributions cannot be attributed to chance.

Placebo mass. For each event, let $w_k^+ = \max\{\hat{w}_k, 0\}$ denote the positive marginal contribution of passage k, and let R and P denote the sets of relevant and placebo passages, respectively. The share of positive weight falling on placebo passages is

$$\phi = \frac{\sum_{k \in P} w_k^+}{\sum_{k \in R \cup P} w_k^+} \in [0, 1],$$

which summarises, at the event level, how much of the model's total positive dependence is placed on semantically irrelevant text. Grounded behaviour is characterised by low ϕ, while ungrounded dependence yields large values.

Randomisation inference. To assess whether the observed value ϕ_{obs} is unusually large, I test the sharp null that placebo labels are irrelevant to the pattern of marginal contributions. Let K^+ denote the number of passages with positive weights. Under the null, any assignment of K^+ placebo labels across passages is equally likely. I therefore compute the randomisation distribution by permuting placebo labels across the K^+ positions, recomputing placebo mass ϕ_b for each permutation, and forming the right-tailed p-value:

$$p = \frac{1 + \sum_{b=1}^{B} \mathbb{1}\{\phi_b \geq \phi_{\text{obs}}\}}{1 + B}.$$

A small p-value indicates that the model places significantly more positive weight on placebo passages than expected under random assignment, implying ungrounded behaviour. Conversely, when ϕ_{obs} falls well within the reference distribution, the observed coefficient-level significance pattern is unlikely to arise from multiple testing alone.

Interpretation. This global test complements the passage-level diagnostic in the main text. While the main procedure identifies *which* passages provide statistical support for each motivation, the global placebo mass test assesses *whether the overall pattern of support is itself meaningful.* Across events, the observed placebo masses cluster well below their randomisation-based critical values, confirming that the significant passages identified in the main text do not arise from chance alone.

C Application of SANI

C.1 Retrieval Performance Metrics and Parameter Selection [TO BE ADDED]

This Appendix provides additional detail on the evaluation of the retrieval operator $\mathcal{R}_{K,L}$, the calibration of its parameters (K, L), and the performance of alternative segmentation strategies. The objective is to document the robustness of retrieval accuracy and to clarify how the parameter settings used in the main text were chosen.

Benchmark Setup Each VAT measure (j, t) is manually linked to its corresponding Senate report article, denoted the *gold article*. For every candidate pair of parameters (K, L), I perform retrieval and record all text segments d_k whose span overlaps the gold article. These are counted as *hits*. Because each article can extend across multiple segments, a single event may have several true positives.

C.2 Classification Prompts and Category Scheme

This appendix reports the prompt templates used to classify value-added tax (VAT) measures into categories following the narrative taxonomy of Cloyne (2013). Each category corresponds to one of the subtypes summarized in Table 2. The classification operator $\mathcal{C}$ applies these prompts independently to the structured motivations $\hat{S}_{j,t}$ obtained from the summarization stage, returning binary (*Yes/No*) assessments and a textual rationale grounded in the retrieved evidence.

Prompt templates

Each prompt is presented in a standardized format and corresponds to one of the classification categories listed in Table 2. The prompts are organized by group—*endogenous* (Demand Management 3, Supply Stimulus 4, Deficit Reduction 5, Spending-driven 6) and *exogenous* (Long-term 7, Ideological 8, External 9, Deficit Consolidation 10). Each prompt provides the task definition, inclusion and exclusion criteria, and output format used in the classification procedure. Together, they operationalize the decision rules summarized in Table 2, ensuring full transparency and reproducibility of the classification process.

English translation of the classification prompt (Demand Management / DM):

Task: Classify a fiscal measure based on the stated motivations. Does this measure correspond to the label "Demand Management (DM)"?

Definition of Demand Management (DM):
A passage is classified as Demand Management if the measure is motivated by the need to influence, in the short term, the level or composition of aggregate demand in the economy—either to curb inflationary pressures and cool overheated sectors, or to stimulate demand during a slowdown or recession.

This includes:

- Tax increases or restrictions explicitly justified as reducing consumption, restraining demand, or mitigating inflationary pressures.
- Tax cuts or reliefs explicitly presented as stimulating demand, consumption, investment, or economic activity.
- Measures related to concerns over the balance of payments, inflation control, or unemployment, where the mechanism operates through aggregate demand.
- Temporary or reversible measures (e.g., surcharges, freezes, deferrals) justified by immediate economic circumstances, particularly short-term demand conditions.

*Do **not** classify as Demand Management if:*

- The motivation is solely deficit reduction or fiscal consolidation.
- The measure is driven exclusively by environmental, health, or equity objectives without reference to aggregate demand.
- The change is a routine indexation or neutral adjustment without intent to influence demand.
- The measure is presented as a long-term structural reform (improving incentives or competitiveness) rather than targeting short-term demand effects.

Output format:

- Respond with one option only for the entire set of motivations (do not provide multiple answers):
  - Yes — [short rationale]
  - No — [short rationale]

Prompt 3: Classification Prompt for the "Demand Management (DM)" Category

English translation of the classification prompt (Supply stimulus / SS):

Task: Classify a fiscal measure based on the stated motivations. Does this measure correspond to the label "Supply stimulus (SS)"?

Definition of Supply stimulus (SS):
A passage is classified as Supply stimulus if the measure is motivated by the need to stimulate, in the short term, production, investment, competitiveness, or employment on the supply side of the economy. Such measures typically involve tax cuts, reliefs, or structural changes aimed at improving incentives, profitability, liquidity, or business conditions to increase output and growth potential.

This includes:

- Reductions in income tax rates, allowances, or thresholds explicitly presented as improving work incentives or effort.
- Corporate or business tax cuts justified as encouraging investment, profitability, or competitiveness.
- Investment allowances or reductions in social contributions intended to lower hiring costs or promote employment.
- Measures described as supporting recovery or short-term growth through improvements on the supply side.
- Cyclical adjustments designed to maintain the immediate competitiveness of a sector.

*Do **not** classify as Supply stimulus if:*

- The justification emphasizes long-term structural, productivity, or competitiveness effects.
- The measure is presented as deficit reduction or fiscal consolidation.
- The measure targets aggregate demand management (short-term demand stimulation or restraint).
- The justification is primarily environmental, health-related, or social without a supply-side argument.
- The change is a neutral indexation, a temporary demand stimulus, or a routine adjustment without an incentive-based rationale.

Output format:

- Respond with one option only for the entire set of motivations (do not provide multiple answers):
  - Yes — [short rationale]
  - No — [short rationale]

Prompt 4: Classification Prompt for the "Supply stimulus (SS)" Category

English translation of the classification prompt (Deficit Reduction / DR):

Task: Classify a fiscal measure based on the stated motivations. Does this measure correspond to the label "Deficit Reduction (DR)"?

Definition of Deficit Reduction (DR):
A measure is classified as Deficit Reduction if its objective is to generate immediate revenue or prevent a loss of revenue in order to contain the short-term budget deficit.

This includes:

- The freezing or restriction of tax allowances or reliefs to raise funds for the national treasury "this year" or "next year."
- Increases in taxes or duties explicitly presented as necessary to raise additional revenue (e.g., "I need to raise X in additional receipts," "to finance this," "given the need to generate extra revenue").
- The elimination or reduction of costly reliefs whose expense is deemed unsustainable or unjustified.
- Adjustments to consumption taxes—such as fuel, alcohol, tobacco, or VAT—to "protect real yields" or prevent erosion of the tax base.
- Temporary or exceptional levies introduced to cover revenue needs within the current fiscal year.

*Do **not** classify as Deficit Reduction if:*

- The measure is primarily motivated by medium- or long-term fiscal consolidation (this belongs to another category).
- The measure is justified mainly by non-fiscal motives (environmental, health, or social equity) without clear reference to deficit control.
- The measure reflects routine indexation or neutral adjustments that do not raise revenue.

Output format:

- Respond with one option only for the entire set of motivations (do not provide multiple answers):
  - Yes — [short rationale]
  - No — [short rationale]

Prompt 5: Classification Prompt for the "Deficit Reduction (DR)" Category

English translation of the classification prompt (Spending-driven / SD):

Task: Classify a fiscal measure based on the stated motivations. Does this measure correspond to the label "Spending-driven (SD)"?

Definition of Spending-driven (SD):
A passage is classified as Spending-driven if the measure is explicitly motivated by the need to finance, offset, or enable increases in public spending or the introduction of new expenditure programs.

This includes:

- Tax increases or restrictions on reliefs introduced to cover the cost of new spending or the expansion of existing programs (e.g., food subsidies, NHS funding, social reforms).
- Statements indicating that additional revenue is needed to "finance," "help finance," or "enable" proposed spending measures.
- Exceptional or one-off taxes whose purpose is to fund a specific expenditure program (e.g., employment initiatives, child poverty measures).
- Reallocation or withdrawal of tax reliefs to free resources for targeted spending programs.

*Do **not** classify as Spending-driven if:*

- The motivation is deficit reduction or fiscal consolidation for long-term sustainability.
- The measure is justified by demand management (inflation control, balance-of-payments adjustment, or demand stimulus).
- The change is primarily motivated by environmental, health, or equity considerations without explicit reference to spending needs.
- The measure is a routine indexation or neutral adjustment with no link to expenditure financing.

Output format:

- Respond with one option only for the entire set of motivations (do not provide multiple answers):
  - Yes — [short rationale]
  - No — [short rationale]

Prompt 6: Classification Prompt for the "Spending-driven (SD)" Category

English translation of the classification prompt (Long-term / LR):

Task: Classify a fiscal measure based on the stated motivations. Does this measure correspond to the label "Long-term (LR)"?

Definition of Long-term (LR):
A passage is classified as Long-term if the measure is motivated by structural, supply-side, or long-term incentive effects on the economy—aimed at improving productivity, competitiveness, investment, saving, or economic efficiency over time.

This includes:

- Tax cuts or reforms justified by improving work effort, incentives, or morale.
- Measures designed to stimulate business investment, capital formation, or technological modernization.
- Reforms targeting international competitiveness, efficiency, or simplification of the tax system.
- Policies encouraging entrepreneurship, risk-taking, or saving to support future growth.
- Changes presented as necessary for the transition to a new economic regime (e.g., from wartime to peacetime).

*Do **not** classify as Long-term if:*

- The main motivation is short-term revenue needs or deficit consolidation.
- The measure concerns only short-term demand management (stimulus or demand restraint).
- The justification is purely environmental, health-related, or equity-driven without reference to long-term incentives or growth.
- The change is a routine indexation or minor administrative adjustment with no structural intent.

Output format:

- Respond with one option only for the entire set of motivations (do not provide multiple answers):
  - Yes — [short rationale]
  - No — [short rationale]

Prompt 7: Classification Prompt for the "Long-term (LR)" Category

English translation of the classification prompt (Ideological / IL):

Task: Classify a fiscal measure based on the stated motivations. Does this measure correspond to the label "Ideological (IL)"?

Definition of Ideological (IL):
A passage is classified as Ideological if the measure is primarily motivated by political values, social philosophy, principles of fairness, or a broader vision of society and the tax system—rather than by short-term macroeconomic management (demand control) or fiscal balance (deficit reduction or consolidation).

This includes:

- References to fairness, justice, tradition, or "social values," including environmental or ethical considerations.
- Tax reliefs or exemptions targeted at "deserving" groups (e.g., retirees, widows, children, families, vulnerable populations).
- Explicitly moral, social, or political justifications (e.g., "people hate income tax more than any other tax," "to reward work," "to encourage saving," "to support family unity").
- Long-term reforms aimed at making the tax system simpler, fairer, or more logical, without references to deficit or demand management.
- Measures presented as fulfilling electoral promises or party commitments (e.g., VAT cuts on essential goods, new redistributive taxes).
- Reforms introduced as part of a broader vision of taxation reflecting social or political values (e.g., taxing capital gains "as income," inheritance thresholds for family fairness).

*Do **not** classify as Ideological if:*

- The measure is primarily motivated by deficit reduction, fiscal sustainability, or budget consolidation.
- The change is framed in terms of short-term demand management (stimulus, inflation control, balance of payments).
- The measure is a routine indexation or a neutral technical adjustment.

Output format:

- Respond with one option only for the entire set of motivations (do not provide multiple answers):
  - Yes — [short rationale]
  - No — [short rationale]

Prompt 8: Classification Prompt for the "Ideological (IL)" Category

English translation of the classification prompt (External / EXT):

Task: Classify a fiscal measure based on the stated motivations. Does this measure correspond to the label "External (EXT)"?

Definition of External (EXT):
A passage is classified as External if the measure is motivated by external economic pressures, trade obligations, or international agreements.

This includes:

- Measures adopted in response to balance-of-payments problems, exchange rate pressures, or the need to protect reserves.
- Tax modifications explicitly linked to supporting exports, foreign trade operations, or competitiveness in international markets.
- Adjustments mandated by commitments to international organizations or treaties (e.g., WTO, EU, EFTA).
- Policy changes justified by concerns about investor confidence or market reactions to the government's financing needs.

*Do **not** classify as External if:*

- The motivation is purely domestic (e.g., deficit reduction, consolidation, demand management, social equity, environment).
- The change is a routine adjustment unrelated to trade, finance, or international agreements.
- The measure concerns only domestic consumption, revenue, or distribution without an external framing.

Output format:

- Respond with one option only for the entire set of motivations (do not provide multiple answers):
  - Yes — [short rationale]
  - No — [short rationale]

Prompt 9: Classification Prompt for the "External (EXT)" Category

English translation of the classification prompt (Deficit Consolidation / DC):

Task: Classify a fiscal measure based on the stated motivations. Does this measure correspond to the label "Deficit Consolidation (DC)"?

Definition of Deficit Consolidation (DC):
A passage is classified as Deficit Consolidation if the measure is motivated by a structural or medium- to long-term strategy aimed at strengthening public finances or reducing deficits, borrowing, or public debt.

This includes:

- Measures presented as part of a sustained fiscal plan to ensure the soundness of public finances.
- Structural revenue generators (e.g., base broadening, duty escalators, restrictions on reliefs) intended to improve the budget balance over several years.
- Policies explicitly linked to fiscal consolidation, fiscal sustainability, or medium-term deficit reduction, even when other justifications are also mentioned.
- Measures whose initial intent was deficit consolidation, even if the rhetoric later evolved.

*Do **not** classify as Deficit Consolidation if:*

- The motivation concerns short-term revenue needs ("to raise additional revenue this year," "to finance support for the economy")—this falls under Deficit Reduction (DR), not Consolidation.
- The change is justified mainly by environmental, health, or social equity objectives without a fiscal framing.
- The measure reflects routine indexation or neutral adjustments without a consolidation goal.
- The policy reduces taxes or increases spending without compensating measures.

Output format:

- Respond with one option only for the entire set of motivations (do not provide multiple answers):
  - Yes — [short rationale]
  - No — [short rationale]

Prompt 10: Classification Prompt for the "Deficit Consolidation (DC)" Category

D Replication of Romer and Romer

D.1 Classification Settings

D.1.1 Narrative extraction

This appendix provides two illustrative examples of the narrative editing procedure described in Section **??**. Each example corresponds to a tax change classified by Romer and Romer (2010) as either exogenous or endogenous. The full original narrative is displayed, with the portion removed prior to LLM classification highlighted in blue. This includes the sentence in which the classification appears, as well as the two preceding sentences. These examples serve to clarify the rationale behind the procedure: removing these sentences prevents the LLM from directly accessing the authors' stated judgment or any surrounding language that might strongly signal the intended label. As shown below, the remaining narrative retains substantial contextual information, allowing the model to infer the likely classification while ensuring the comparison with human-coded labels remains fair.

Example 1: an endogenous measure

The motivation for the Revenue Act of 1951 was again the increase in spending related to the Korean War. The *1951 Economic Report* stated: "These new taxes are required to finance the defense effort; and to help keep total spending within the capacity of current production, so that inflation does not reduce the purchasing power" (p. 17). This same sentiment was echoed in a number of presidential speeches. On February 2, 1951, Truman said, "If we do not tax ourselves enough to pay for defense expenditures, the Government will ... add to total purchasing power and inflationary pressures" (*Special Message to the Congress Recommending a "Pay as We Go" Tax Program*, p. 2). These quotations make it clear that the tax increase was designed to keep growth normal.

It stated: "The military action in Korea, coupled with the general threat to world peace, has made it necessary to provide extraordinary increases in revenues to meet essential national defense expenditures" (82d Congress, 1st Session, House of Representatives Report No. 586, 6/18/51, p. 1). It also stressed the temporary nature of the tax, saying: "The 12$\frac{1}{2}$-percent flat across-the-board increase [in individual income tax rates] was selected as the form of increase, since it is not intended that this increase will be permanent, and therefore, it was desired to provide an increase which would not become an integral part of the rate structure" (p. 8). Because the tax increase was designed to counteract the effects of higher military spending and thereby keep growth normal, I classify it as an endogenous, spending-driven action.

Truman's speech upon signing the bill stated that it would raise about $5.5 billion of additional revenues at an annual rate (*Statement by the President Upon Signing the Revenue Act of 1951*, 10/20/51, p. 1). Both the *1952 Economic Report* (p. 135) and the *1951 Treasury Annual Report* (p. 44) gave the figure as $5.4 billion. I follow my usual practice and use the number from the *Economic Report*.

The increase in corporate taxes was retroactive to April 1, 1951, while the remainder took effect on November 1, 1951 (*1951 Treasury Annual Report*, p. 51). Because the action took place before the middle of the quarter, I date the tax increase in 1951Q4. Approximately $2.3 billion of the tax increase was from corporate taxes (p. 501). Therefore, the retroactive portion of the bill increased taxes at an annual rate of $2.3 billion · 2 extra quarters, or $4.6 billion. Adding this to the steady-state effect of $5.4 billion yields a tax increase of $10 billion in 1951Q4. The return to the steady-state level of tax increase implies a tax cut of $4.6 billion in 1952Q1. If one ignored the retroactive part of the legislation, there would be just a tax increase of $5.4 billion in 1951Q4.

The tax increase largely took the form of an increase in marginal rates. The act also raised the capital gains tax, the tax on corporate profits, and some excise taxes (*1951 Treasury Annual Report*, pp. 50–52). Most of the changes were explicitly temporary. The individual income tax increases were legislated to expire on January 1, 1954; the corporate and excise tax increases were to continue until March 31, 1954 (*1951 Treasury Annual Report*, pp. 51–52).

Example 2: an exogenous measure

This bill was a significant change of the income tax system. It was the first comprehensive revision of the Internal Revenue Code in 75 years (*Statement by the President Upon Signing Bill Revising the Internal Revenue Code*, 8/16/54, p. 1).

This bill was passed primarily to increase the fairness and efficiency of the tax system. It had been under discussion since 1952 (Holmans, 1961, p. 219), well before the 1953–54 recession. At the signing ceremony, Eisenhower stressed: "It is a law which will help millions of Americans by giving them fairer tax treatment than they now receive." He also said, "economic growth will be fostered by such provisions as more flexible depreciation and better tax treatment of research and development costs, thus encouraging all business—large and small—to modernize and expand" (*Statement by the President Upon Signing Bill Revising the Internal Revenue Code*, 8/16/54, p. 1). The Report of the Committee on Ways and Means to Accompany H.R. 8300 (Internal Revenue Code of 1954) repeated this motivation. It stated: "In general, the purpose of these changes has been to remove inequities, to end harassment of the taxpayer and to reduce tax barriers to future expansion of production and employment" (83d Congress, 2d Session, House of Representatives Report No. 1337, 3/9/54, p. 1).

From this perspective, the reduction in revenues at a given level of income was the result of a belief that the bill would stimulate growth and of a desire to make the reform roughly revenue neutral after accounting for the growth effects. The *1954 Economic Report* stated, "A number of additional tax measures should now be taken in order to strengthen the forces of growth in employment and production. These measures, which involve some immediate sacrifices of revenue, contain the seeds of important future revenue gains to be reaped from the economic growth they will stimulate" (p. 79). The *1955 Economic Report* described the motivations for the bill as "to stimulate competitive enterprise, to strengthen the floor of security for the individual, and to curb tendencies toward either depression or inflation," and referred to "the loss of 1.4 billion dollars as a result of enacting structural tax changes" as almost a side effect of the reform (pp. 19–20).

The possible countercyclical benefits of the bill were certainly mentioned frequently. As discussed above, the administration favored some additional tax relief in mid-1954 because of further declines in defense spending. And, much of the Congressional debate focused on the need for additional stimulus. Indeed, several amendments were offered that involved increasing the personal exemption and cutting tax rates (Holmans, 1961, pp. 229–233). However, once the Excise Tax Reduction Act of 1954 was passed, the Administration argued strenuously against further tax cuts, especially the increase in the personal exemption.

[1] On March 15, 1954, Eisenhower said: "at this time economic conditions do not call for an emergency program that would justify larger Federal deficits and further inflation through large additional tax reductions" (*Radio and Television Address to the American People on the Tax Program*, 3/15/54, p. 4; see also *Annual Budget Message to the Congress: Fiscal Year 1955*, 1/21/54, p. 8). Also, the House report on the bill explicitly downplayed the countercyclical aspects, saying: "This bill is a long overdue reform

measure which is vitally necessary regardless of the momentary economic conditions and should not be confused with other measures which may be, or might become, appropriate in the light of a particular short-run situation" (House Report No. 1337, pp. 1–2). Thus, it seems clear that at least some of the tax cuts contained in the Excise Tax Reduction Act of 1954 and the Internal Revenue Code of 1954 should be classified as exogenous.

[2] Thus, it seems clear that at least some of the tax cuts contained in the Excise Tax Reduction Act of 1954 and the Internal Revenue Code of 1954 should be classified as exogenous. There was a cut beyond what the administration thought was justified by the decline in spending and the state of the economy. Because the excise tax reduction came first and had no obvious other motivation, I choose to classify it as endogenous and the Internal Revenue Code of 1954 as an exogenous, long-run action.

[3] There was a cut beyond what the administration thought was justified by the decline in spending and the state of the economy. Because the excise tax reduction came first and had no obvious other motivation, we choose to classify it as endogenous and the Internal Revenue Code of 1954 as an exogenous, long-run action. We feel this makes sense given that the Internal Revenue Code of 1954 clearly had a substantial exogenous component and indeed was passed after it was widely understood that the trough of the recession (May 30, 1954) had been turned (Holmans, 1961, p. 233).

[4] Because the excise tax reduction came first and had no obvious other motivation, we choose to classify it as endogenous and the Internal Revenue Code of 1954 as an exogenous, long-run action. We feel this makes sense given that the Internal Revenue Code of 1954 clearly had a substantial exogenous component and indeed was passed after it was widely understood that the trough of the recession (May 30, 1954) had been turned (Holmans, 1961, p. 233). But, at some level, which of the two tax cuts is classified as endogenous and which as exogenous is somewhat arbitrary.

The act was signed in mid-August 1954, so I date the initial tax change in 1954Q3. My sources give a figure of –$1.4 billion for the revenue effects at an annual rate (for example, *1955 Economic Report*, p. 20; *1954 Treasury Annual Report*, p. 44). The bill was retroactive to January 1, 1954 (*1954 Treasury Annual Report*, p. 286). It thus reduced taxes by 3 · $1.4 billion, or $4.2 billion at an annual rate, in 1954Q3. The return to the steady-state reduction of $1.4 billion in 1954Q4 implies a tax increase of $2.8 billion in 1954Q4. If one chose to neglect the retroactive element, there would be just a tax cut of $1.4 billion in 1954Q3.

The changes consisted largely of the removal of obvious inequities, complications, ambiguities, and opportunities for tax avoidance, and of provisions designed to stimulate investment. All the changes were legislated to be permanent.

D.1.2 Prompt

The classification task is implemented using a structured prompt composed of three components: an explicit persona instruction (Prompt 11), a set of base classification instructions mirroring the criteria developed in Romer and Romer (2010) and Cloyne (2013) (Prompt 12), and a small number of illustrative examples implementing a few-shot learning approach (Prompt 13). Drawing on recent work in computer science on prompt engineering to improve performance (White et al., 2023; Chen et al., 2024), this structure is designed to elicit consistent and transparent classifications while remaining fully replicable.

You are a professional economist specializing in Fiscal Policy and Causal Inference. You are highly skilled at identifying whether tax policy changes are endogenous or exogenous, based on economic intent and context.

Prompt 11: Persona Prompt

You will be provided descriptions of legislated tax changes in the U.S. from 1945 to 2007. Please classify each tax change as either endogenous (N) or exogenous (X) based on the following definitions:

- **Endogenous (N)**: If the primary motivation for the tax change was related to addressing current or expected economic conditions — e.g., reacting to a recession, inflation, unemployment, or funding new programs.
- **Exogenous (X)**: If the tax change was mainly introduced to deal with structural or long-term issues — e.g., reducing an existing deficit, encouraging growth, improving fairness or efficiency, or philosophical shifts in policy. These actions should not be tied to short-term economic fluctuations.

Guidance: If there are multiple motivations, use the predominant one cited at the time of passage.

Your response must ALWAYS be a JSON object containing two fields:

- `classification` (string, either N or X)
- `justification` (a brief explanation)

Prompt 12: Base Instruction Prompt

Endogenous

- Responds to contemporaneous or forecasted macro conditions (recession, inflation, unemployment)? Yes→N
- Primarily to finance newly enacted or announced spending? Yes→N
- Focus on short-term stabilization rather than long-term reform? Yes→N

Exogenous

- Primarily to reduce a non-program deficit, not used as a stabilization tool? Yes→X
- Chiefly to promote long-run growth, equity, efficiency, or philosophical changes? Yes→X
- Intended to raise long-run potential output or reduce structural unemployment without reacting to a current shock? Yes→X

Prompt 13: Few Shots Prompt

D.2 Classifications Robustness checks

D.2.1 Sensitivity to Prompting variations

To evaluate the sensitivity of the classification results to the specific structure of the prompt, I conduct a systematic robustness check in which I vary each of the main components of the prompt used for the main classification. In particular, I test prompts both with and without the in persona framing (Prompt 11), with and without explicit instructions to articulate chain-of-thought reasoning (adding "Think carefully."), and with or without the inclusion of few-shot examples (Prompt 13). This design follows the framework proposed in Ludwig et al. (2025), who emphasize that prompt formulation can significantly influence LLM behavior. The results of this exercise are presented in Table 6. Despite the deliberate variations, the overall classification patterns remain stable across specifications, suggesting that the LLM's performance is not overly reliant on the precise prompt architecture chosen in the main analysis. This lends additional credibility to the classification outcomes, reinforcing the conclusion that the LLM's judgments are robust to reasonable changes in instruction.

Table 6: Robustness of Miss-classification to Prompt Variation

Persona	**CoT**	**Few-shot**	**GPT-4o-mini (Closed)**		**Gemini-2.0-flash (Closed)**	
			N (%)	**X (%)**	**N (%)**	**X (%)**
			35	8	35	13
✓			35	4	39	4
	✓		35	4	35	8
		✓	35	8	35	4
✓	✓		35	4	39	0
✓		✓	31	4	35	0
	✓	✓	35	4	35	4
✓	✓	✓	31	4	35	4

Note: Entries report the share of misclassified tax changes to the categories endogenous (N) and exogenous (X) under each prompt variant. Columns indicate whether the base prompt included a professional economist persona, chain-of-thought reasoning (CoT), or few-shot examples. The different variation blocks are specified in Prompts 11, 12, 13, except for CoT which is performed by adding "Think carefully." to the prompt.

D.2.2 Training-data leakage robustness check

To assess whether the language model has memorized specific source material, I implement a sentence-reconstruction test taken from Golchin and Surdeanu's (2024) contamination detection framework. The intuition is the following: if a model has encountered a document during training, it may be able to reproduce missing content from that document with unusual precision—especially when provided with cues pointing to the original source. To implement this test, I systematically remove sentences from the Romer and Romer narratives and asks the model to reconstruct them under two conditions: one relying solely on the surrounding context, and the other explicitly guiding the model by referencing the source paper and section. If the model performs significantly better under guided prompting, this may suggest prior exposure to the original text. To quantify and evaluate the reconstruction quality, I use both semantic (embedding cosine similarity) and lexical (ROUGE-L) metrics, followed by a bootstrap-based significance test to flag cases of potential memorization.

Using both metrics, and evaluating with `gpt-4o-mini` and `gemini-2.0-flash`, no narrative is flagged as a potential case of memorization. This result strengthens confidence that the model was not trained on Romer and Romer (2010), and supports the validity of the replication exercise by reducing concerns about training-data leakage as a source of bias in the model's classifications.

Procedure.

1. **Multiple draws & prompts.** For each narrative d I perform $T = 10$ draws. In draw t, one sentence $s_{d,t}$ is chosen uniformly at random, removed, and replaced by the token `[MISSING SENTENCE]`. I then issue two prompts a general prompt, and a guided prompt (Prompts 14).

> **General prompt:**
> You are provided with a passage in which one sentence has been removed and replaced with `[MISSING SENTENCE]`. Based ONLY on the context provided, generate a plausible sentence that could fill the gap. You do not need to match any specific source text — just ensure the result is coherent and contextually appropriate.
>
> {`narrative text from Romer and Romer`}
>
> Reconstructed sentence:

> **Guided prompt:**
> You are provided with a passage from the scientific paper *A Narrative Analysis of Postwar Tax Changes* by Christina D. Romer and David H. Romer, published in 2010. The passage comes from the section titled '{`title`}" in the main body of the paper.
> One sentence has been removed and replaced with `[MISSING SENTENCE]`. Based **only** on the context provided and your knowledge of this specific paper, reconstruct the missing sentence exactly as it appeared.
>
> {`narrative text from Romer and Romer`}
>
> Reconstructed sentence:

Prompt 14: General and Guided Prompt

   The model returns two reconstructions $\hat{s}^{\text{gen}}_{d,t}$ and $\hat{s}^{\text{guide}}_{d,t}$.

2. **Similarity metrics.** For each original $s_{d,t}$ and each reconstruction, compute:
   - *Embedding cosine similarity* $\rho^{\text{gen}}_{d,t} = \cos\big(s_{d,t}, \hat{s}^{\text{gen}}_{d,t}\big)$, $\rho^{\text{guide}}_{d,t} = \cos\big(s_{d,t}, \hat{s}^{\text{guide}}_{d,t}\big)$, using `text-embedding-3-large`.
   - *ROUGE-L F1 score* $r^{\text{gen}}_{d,t}$ and $r^{\text{guide}}_{d,t}$ between the original and each reconstruction.

3. **Paired bootstrap test.** For each narrative d, define the mean differences

$$\Delta\rho_d \;=\; \frac{1}{T}\sum_{t=1}^{T}\big(\rho^{\text{guide}}_{d,t} - \rho^{\text{gen}}_{d,t}\big), \quad \Delta r_d \;=\; \frac{1}{T}\sum_{t=1}^{T}\big(r^{\text{guide}}_{d,t} - r^{\text{gen}}_{d,t}\big).$$

   We perform a nonparametric paired bootstrap with $B = 1{,}000$ resamples of the T pairs $\{(\rho^{\text{guide}}_{d,t}, \rho^{\text{gen}}_{d,t})\}$ (and similarly for $\{(r^{\text{guide}}_{d,t}, r^{\text{gen}}_{d,t})\}$), yielding bootstrap distributions of $\Delta\rho_d$ and Δr_d. Each indicator $\mathbf{1}\{\Delta\rho^{*(b)}_d \le 0\}$ (or $\mathbf{1}\{\Delta r^{*(b)}_d \le 0\}$) counts a bootstrap sample in which the similarity under

the guided prompt is no better than—or worse than—that under the general prompt; that is, the model failed to improve reconstruction quality despite being explicitly cued about the source. The empirical p-values are

$$p_d^{\rho} \;=\; \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{\Delta\rho_d^{*(b)} \le 0\}, \quad p_d^{r} \;=\; \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{\Delta r_d^{*(b)} \le 0\}.$$

Decision rule. A narrative d is flagged as potentially memorized if its bootstrap p-value for either similarity metric falls below 0.05, i.e. if guided prompting produces significantly closer reconstructions than general prompting.

D.2.3 Manually Investigating Misclassifications

The misclassification pattern, mostly exogenous policies misclassified as endogenous, can possibly be attributed to an asymmetry inherent in macroeconomic narrative analysis: while it is relatively straightforward to construct narratives supporting the endogeneity of a policy intervention, establishing its exogeneity typically requires more stringent evidence. In the absence of sufficiently strong signals of such independence LLMs may default to a conservative interpretation, erring on the side of endogeneity.

Steering growth above normal levels. A recurring source of systematic misclassification involves tax cuts enacted with the aim of boosting growth above normal levels. In at least three instances, the LLM interprets these as countercyclical, and thus endogenous, due to the presence of growth-oriented motivation. However, according to Romer and Romer's decision rule, such motivations qualify as exogenous, as they are not designed to return output to trend but to exceed it. This kind of mislabeling can be mitigated by explicitly including this rule in the prompt design or few-shot examples, highlighting that aspirational growth policies, unless tied to negative output gaps, are to be treated as exogenous.

Conflicting motivations. Another set of misclassifications arises from cases with mixed or conflicting motivations. For example, one narrative discusses a tax increase intended primarily for deficit reduction and solvency of the Highway Trust Fund, both exogenous motives, but also references inflation concerns and contemporaneous spending increases. While the latter suggest endogenous intent, the original classification deems the measure exogenous with noted ambiguity. Such nuanced judgment is difficult for LLMs to replicate without additional guidance on prioritizing primary over secondary motivations.

"Package" related measure. Finally, there are highly specific edge cases that hinge on bespoke decision rules. A salient example is the Tax Adjustment Act of 1966. Despite being implemented to counteract high growth following an exogenous tax cut, Romer and Romer classify it as exogenous, arguing that treating it as endogenous would yield nonsensical results. This type of reasoning, context-specific and precedent-based, poses a significant challenge for LLMs, particularly in the absence of similar examples in the prompt.

D.3 Confidence Assessment of LLM Classifications

It is crucial to assess the confidence with which the LLMs make their classifications. A model that consistently produces accurate classifications but is highly uncertain about its judgments might be less reliable in practice, especially in ambiguous cases. To gauge this, I examine two key metrics: the log probabilities of generated tokens and the perplexity of the overall output.

Log probabilities / ambiguity. The log probabilities of generated tokens provide a direct measure of the model's certainty at each step of its output. When the LLM outputs a classification (e.g., "X" for exogenous or "N" for endogenous), each token in that word has an associated log probability. A higher (less negative) log probability indicates greater confidence in that specific token's prediction. By aggregating these values, I can assess the model's conviction in its final classification. For instance, a significantly lower log probability for the chosen classification in a misclassified case, compared to a correctly classified one, would suggest that the model was less certain in its incorrect judgment.

To quantify this uncertainty more systematically, I define an *ambiguity* score based on the model-assigned probability for the predicted class label. Specifically, for a classification such as "X", I extract the associated probability p from the model output and compute ambiguity as $\min(p, 1-p)$. This transformation maps probabilities near 0 or 1 (high confidence) to values near 0 (low ambiguity), and probabilities near 0.5 (low confidence) to values near 0.5 (high ambiguity). This score allows for a consistent interpretation of how uncertain the model is about its decision, regardless of whether the predicted label is "X" or "N".

Perplexity. In addition to token-level confidence, I evaluate the perplexity of the overall output, including any explanatory text or justification provided by the LLM alongside its classification. Perplexity is a measure of how well a probability model predicts a sample. A lower perplexity score indicates that the model is "less surprised" by its own output, suggesting a more coherent and confident generation. Conversely, a high perplexity implies greater uncertainty or hesitation, potentially indicating that the model struggled to formulate a clear rationale or was less sure about its classification.

Relation between model confidence and misclassification. To empirically assess whether these uncertainty measures predict classification errors, I fit three logistic regression models where the binary outcome variable is whether the classification was incorrect. The first model includes ambiguity (based on the model's token-level output probabilities), the second includes perplexity, and the third includes both predictors. The results, shown in Table 7, indicate that ambiguity is a strong and statistically significant predictor of error: higher ambiguity is associated with a significantly increased likelihood of misclassification. Specifically, in the full model, the coefficient on ambiguity is 4.26, which implies that a one-unit increase in ambiguity multiplies the odds of misclassification by more than 70 times (since $e^{4.26} \approx 70.7$). This is a large effect, underscoring the critical role of ambiguity in driving model errors.

In contrast, the effect of perplexity—while statistically significant when ambiguity is excluded—disappears once ambiguity is included. In the joint model, the coefficient on perplexity becomes small and non-significant, suggesting that ambiguity subsumes most of the predictive value of perplexity. In

other words, while perplexity correlates with misclassification on its own, it does not offer additional explanatory power beyond what is already captured by token-level ambiguity.

Table 7: Logistic Regression of Error on Ambiguity and Perplexity

	Ambiguity	**Perplexity**	**Ambig+Perp**
const	-1.358***	-3.464***	-1.068
	(0.085)	(0.454)	(0.788)
ambig	3.893***	—	4.258***
	(0.606)		(1.159)
perp.	—	2.187***	-0.292
		(0.417)	(0.787)
N	1000	1000	1000
Pseudo R^2	0.037	0.025	0.037

Note: Coefficients are log-odds estimates; standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$. A dash (—) indicates the variable was omitted from that model.

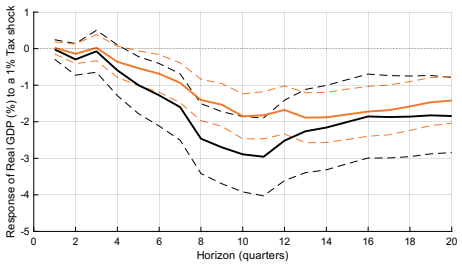
D.4 The Role of Narrative Exogeneity in Tax Shock Identification

This subsection jointly examines the role of narrative exogeneity in the identification of tax shocks under the Romer–Romer (2010) framework and the SANI procedure. Figure 11 reports impulse responses obtained using, for each approach, (i) all legislated tax changes without conditioning on policy motivations, and (ii) tax shocks constructed using narrative-based exogeneity restrictions.

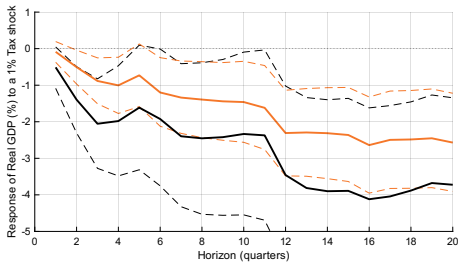
Across both frameworks, the results point to a common pattern. Including all tax changes without narrative filtering leads to a substantial attenuation of the estimated output response to a tax increase. In the Romer–Romer case, peak effects are reduced by roughly one third relative to the baseline specification that isolates exogenous shocks, while preserving a broadly similar hump-shaped dynamic profile. In the SANI case, the attenuation is even more pronounced: treating all routine budgetary tax measures as exogenous yields responses that are markedly smaller in magnitude than those implied by narrative-weighted shocks.

In both panels, confidence bands are mechanically tighter when all tax changes are included. This increase in precision reflects the inclusion of endogenous policy variation rather than improved identification. Conditioning on stated policy motivations, whether through discrete narrative classification or continuous exogeneity weighting, is therefore essential to recover economically meaningful tax shocks.

Taken together, the side-by-side comparison highlights the central role of narrative information in fiscal shock identification. While the institutional settings and policy universes differ across the two approaches, the qualitative implications are strikingly similar: absent narrative exogeneity restrictions, estimated tax multipliers are severely attenuated, despite appearing more precisely estimated. This pattern underscores the importance of narrative methods for isolating exogenous fiscal variation in the presence of policy endogeneity.



(a) Romer–Romer: Exogenous vs. All Tax Shocks



(b) SANI: Narrative-Weighted vs. All Tax Shocks

Figure 11: The Role of Narrative Exogeneity in Tax Shock Identification

Note: Each panel reports impulse responses of real GDP to a one-percent tax increase estimated using the same empirical specification. The black curve corresponds to shocks constructed using narrative exogeneity restrictions (Romer–Romer exogenous shocks in the left panel and SANI narrative-weighted shocks in the right panel). The red curve reports responses obtained when all legislated tax changes (endogenous included), without narrative filtering. Shaded areas denote 68% confidence intervals.

E Other

E.1 VAT table adoption

Table 8: VAT (or GST-equivalent) Adoption by Year in Developed Economies

Year	Countries
1954	France
1967	Denmark
1968	Germany
1969	Netherlands; Sweden
1970	Luxembourg; Norway
1971	Belgium
1972	Ireland
1973	Austria; Italy; United Kingdom
1977	South Korea
1980	Mexico
1985	Turkey
1986	New Zealand
1989	Japan
1990	Iceland
1991	Canada
1994	Finland
1995	Switzerland
2000	Australia