Identifying Episodes of Fiscal Austerity: An LLM-Based Approach

Karan Bhasin and Prakash Loungani[1]

19th January 2026

Abstract

This paper introduces a hierarchical Large Language Model (LLM) framework for the automated identification of narrative fiscal shocks. We develop a multi-stage architecture to extract austerity episodes from IMF Article IV reports (2004–2022) for 17 OECD countries. Unlike manual coding, our approach ensures replicability and auditability through documented reasoning chains. Benchmarking against Adler et al. (2024), we find the LLM framework successfully filters endogenous policy shifts overlooked by human coders. Local projections demonstrate that LLM-identified shocks yield more precisely estimated fiscal multipliers whereas noisier traditional datasets show mixed results.

Keywords: Natural Language Processing; Large Language Models; Fiscal Shocks
JEL Classification:C55, C80, C83, C10, E62.

[1] Bhasin: University at Albany, State University of New York.
E: kbhasin@albany.edu
Loungani: Johns Hopkins University.
E: ploungani@jhu.edu

I. Introduction

As global public debt reaches historical levels, the need for urgent fiscal adjustments has intensified the debate over timing and impact of austerity. However, identification of austerity shocks remains a significant empirical challenge as traditional quantitative measures struggle to distinguish between proactive policy-driven adjustments and passive cyclical fluctuations. To address these limitations, this paper introduces a computational framework that leverages Large Language Models (LLMs) to systematically audit fiscal policy narratives and extract structural shocks from structured textual data.

The method for identifying fiscal shocks makes a big difference on the estimated effects of austerity on macroeconomic aggregates. For instance, Alesina et al. (2002) found fiscal consolidation to be associated with higher growth. This finding was in line with previous studies (Giavazzi and Pagano, (1990) Alesina and Perotti (1995) and Alesina and Perotti (1997)).[2] These studies relied on identification of austerity shocks based on the use of vector autoregression with exclusion restrictions. In contrast, Guajardo et al. (2014) relied on identification of austerity based on the narrative method pioneered by Romer and Romer (2010). The narrative approach relies on information obtained from policy documents to identify changes in fiscal policies motivated by the desire to reduce public debt and not by responding to economic conditions. Guajardo et al. (2014) found that fiscal consolidation had contractionary effects on private demand and GDP. The difference in these results was attributed to different choice of austerity shocks.

Despite the strength of the narrative approach in capturing policy intent it is inherently constrained by its high labor intensity and susceptibility to subjective interpretation. Coding vast archives of policy documents across multiple decades and jurisdictions is often prohibitively slow, limiting the scale and frequency of comparative fiscal analysis. By employing Natural Language Processing, this paper seeks to overcome these scalability and consistency hurdles, automating the identification of austerity shocks while maintaining the linguistic nuance required to separate debt-motivated actions from cyclical adjustments. Beyond labor intensity, the narrative approach lacks perfect replicability, as coding criteria can vary across researchers; our framework provides an algorithmic benchmark that ensures consistency across time and jurisdiction.

Recent work underscores these possibilities. For instance, Ash et al. (2025) discuss the advantages of transformer-based models like GPT, highlighting their ability to process large

[2] Also see Garcia (2025) for a recent paper that finds positive output effects of fiscal consolidations for the US.

corpora and analyze metadata such as tone or sentiment. Similarly, Korinek (2023, 2024) document how recent advances in Large Language Models (LLMs) have made it feasible for these tools to process and interpret complex policy documents with a level of consistency comparable to human judgment. These advances make it feasible to employ Natural Language Processing (NLP) and LLMs more systematically in macroeconomic analysis. The use of LLMs offers two principal advantages over traditional narrative methods. First, large language models can be restricted to contemporaneous textual sources, ensuring that classifications are not influenced by ex-post knowledge of macroeconomic developments. Critically, unlike human coding, these models provide 'Reasoning Chains' for every decision, allowing researchers to audit the specific logic used to identify a shock. Human classification, by contrast, inevitably possesses some degree of hindsight (or foresight), which risks contaminating the identification of policy shocks that are intended to be orthogonal to prevailing economic conditions. Second, NLP enhances scalability and ensures an objective criterion is consistently applied throughout the sample. These models can process vastly larger corpora at significantly lower cost and in a fraction of the time required by manual reading. Their application allows for a systematic classification of country-year observations as austerity episodes based on structured textual documents from Article IV reports.[3]

The use of data as text is not new to economics (Ash and Hansen, (2023)) and recent methodological advances have successfully applied NLP-based models and LLMs to identify monetary policy shocks from sources like the Federal Reserve's Beige Books (Aruoba & Drechsel, 2024; Fernández-Fuertes, 2025). Nevertheless, their application to fiscal policy remains nascent. These papers focus on generating a conditional expectation of a policy variable (such as interest rates) and use the deviation of the actual from this expectation as a measure of monetary policy shock. Expectations are generated either from an estimated equation that incorporate time-series data generated from textual data or from a multi-agent framework. Such frameworks work well for textual data that is available ex-ante, however, fiscal documents such as IMF's article IV reports are available ex-post. Consequently, we develop a hierarchical decision architecture tailored for ex-post policy archives. Our framework utilizes decompressed prompting, breaking complex identification tasks into a sequence of atomic logic gates, to identify episodes of austerity

[3] More recently, Amgalan et al. (2025) undertake a systematic analysis of IMF fiscal advice contained in Article IV consultations using a large language model. Their study focuses on identifying IMF advice to country authorities, which differs from the scope of this paper, which proposes a new method for identifying austerity shocks.

in a local, auditable setting. By developing this hierarchical architecture for ex-post policy documentation, we establish a replicable protocol for structural identification that supplements the growing 'Data as Text' literature and extends LLM applications beyond simple sentiment or expectations-based analysis.

We implement our framework for 17 OECD economies, improving upon the narrative austerity shocks compiled by Guajardo et al. (2014) and Adler et al. (2024). This comparison serves two purposes: first, to validate our approach in binary classification; and second, to demonstrate the superior precision of algorithmic auditing. We highlight specific cases, such as Denmark in 2010, where the LLM's hierarchical logic successfully identified exogenous adjustments that were previously obscured. Conversely, we identify numerous instances where human experts misidentified endogenous fiscal adjustments, driven by cyclical fluctuations, as exogenous shocks. Correcting for these systematic identification errors leads to a cleaner, more robust estimate of the fiscal contraction. Our results show that austerity is followed by statistically significant declines in output, consumption, and investment, with quantitative effects that provide a clearer resolution to the multiplier debate than previously possible.

The structure of the paper is as follows. Section II provides details of the data and our framework. In section III we provide our results and comparison with narrative shocks identified by Adler et al (2024). In Section IV, we use propensity score matching and local projections to measure the effects of austerity shocks and compare them with those in Guajardo et al. (2014). Section V concludes.

II. Methodology

A. Data

For our analysis, we use textual data from the IMF's article IV consultation. The article IV is an annual assessment of a member country's economy and is prepared in consultation with government officials. It summarizes the IMF's understanding of the country's economic situation and provides policy advice. Data on macroeconomic statistics are obtained from the World Bank and International Financial Statistics – International Monetary Fund.

B. Our Approach to Identifying Austerity

The literature on the effects of austerity policies has remained inconclusive not just in terms of the size but also in terms of the direction. Identifying the true effects of austerity is challenging because 'clean' data provided by natural experiments is rarely available in macroeconomics.

Nevertheless, numerous attempts have been made to better understand the effect of such policies resulting in a large and growing but inconclusive literature. The effect of austerity shocks on economic outcomes has been identified through exclusion restrictions (Alesina and Perotti, 1995), Perotti (1999) and Mountford and Uhlig (2009), instrument variables (Auerbach and Gorodnichenko (2012), Mertens and Ravn (2013) and Owyang et al (2013)) or through narrative-based measures (Guajardo et al. (2014) and Leigh et al. (2010)). A common finding in recent literature has been that austerity implemented through expenditure cuts are associated with smaller output costs and may, in some cases be expansionary (Alesina, Favero and Giavazzi, 2018).

In this paper, we focus on developing a method for identification of austerity shocks using Large Language Models (LLMs). Our method is based on Romer and Romer (2010)[4] who used presidential speeches and congressional reports to identify principal motivation for post-war tax changes. The method was applied in the context of 17 OECD countries by Guajardo et al (2014) who used OECD and IMF country reports to construct their measure of austerity shocks. Their dataset was recently updated by Adler et al. (2024). A central idea behind the narrative approach pioneered by Romer and Romer (2010) is to isolate instances of austerity that were not driven by economic considerations. These shocks are subsequently used for identification of the macroeconomic effects of such policies.

Several recent advances have tried to revive the Romer and Romer identification approach using large language models. This rapidly growing literature has focused exclusively on identification of monetary policy shocks and have relied on textual data from Beige Books and Minutes released before the FOMC meetings (See Fernández-Fuertes (2025) and Aruoba and Drechsel (2024)). Our approach differs and supplements their methodological advances in several ways. A common theme between our method and those used by Fernández-Fuertes (2025) and Aruoba and Drechsel (2024) is the use of structured textual data for identification of policy shocks. Like the Beige Books and FOMC minutes, IMF's article IV provides structured textual data with relevant policy insights that help identify latent policy shifts.

A notable difference between our approach relative to theirs is textual data for monetary policy is often available ex-ante. That is, textual data in Beige Books and FOMC meetings is available before the decision on monetary policy is made. This allows them to construct conditional expectation of policy changes based on the textual data available prior to the FOMC

[4] Also see Romer and Romer (1989, 1997) for earlier work that focuses on monetary policy shocks.

meetings. Deviation of interest rate from this conditional expectation gives them a measure of unanticipated changes in interest rates. However, for fiscal policy, such data is seldom available. The most common and consistent cross-country source of textual information on fiscal policy is from the IMF's Article IV consultations. These consultations provide a structural textual dataset with some forward projections but more importantly, they provide important information on policy actions that have already been undertaken. In that sense, it serves as an ex-post measure or an account of policy changes that are already in effect. Our method accommodates the ex-post nature of our dataset and allows us to identify legislated changes that were taken due to exogenous reasons. This is in spirit of the Romer and Romer (2010) and Guajardo et al. (2014) approach for identifying instances of austerity.

C. General Framework

A notable advantage with our proposed method is that it relies on using textual data from Article IV and a series of prompts for identifying whether a country administered austere policies and if so, estimate the size of such policies relative to the GDP. We follow a relatively simple sequential prompt approach for identification of austerity shocks. The approach differs from more sophisticated analysis used by Aruoba and Drechsel (2024) who used textual data to construct time series data. They then use the time-series data along with macroeconomic forecasts to predict changes in interest rates. Fernández-Fuertes (2025) on the other hand relied on a multi-agent set-up where different LLMs interact with information in different stages.[5]

These approaches work well for identification of monetary policy shocks, which are much more frequent and display greater variation within a year. However, unlike monetary policy, fiscal policy decisions are often undertaken once a year and sometimes include multi-year plans. Deviating from these decisions within a year are often legislatively infeasible. Our simple-sequential prompt approach is therefore better suited for identification of austerity shocks. As a test of our approach, we rely on the datasets from Guajardo et al. (2014) and Adler et al. (2024) which serve as a validation set for our analysis. We later provide a detailed comparison between our identified instances and these established benchmarks.

For our analysis, we utilize Meta's Llama 3.3-70B as the core large language model (LLM). This model offers several distinct advantages, most notably its ability to deliver performance comparable to the significantly larger Llama 3.1-405B. This efficiency is critical, as it allows the

[5] Also see Li et al. (2024) and Tillmann (2025).

model to be deployed on local workstations rather than requiring enterprise-scale server clusters. Furthermore, the model's 128K context window is particularly advantageous for our research, as it provides the capacity to precisely define and identify instances of austerity that are truly exogenous. By leveraging these advanced reasoning capabilities, the model can effectively analyze structured textual data to isolate specific instances of austerity within the defined context.

A given textual data, in the form of a PDF file of IMF's article IV consultation is provided to the Llama 3.3-70B model. The model is provided with the following context: "*Identify episodes of austerity, defined as a reduction in government expenditures or an increase in tax rates. Additionally, please evaluate if the text indicates these measures were exogenous (motivated by long-term fiscal goals) rather than endogenous (a direct reaction to current economic fluctuations).*"

To determine if a tax hike is exogenous, the model needs to read the *entire* preamble of a budget or policy document. It needs to see if the government is raising taxes to "narrow a long-term deficit" (exogenous) or as a "panic response to a sudden market crash" (endogenous). The 128K context window allows it to hold all these supporting details in its "active memory" at once, ensuring the classification is based on the full narrative rather than an isolated sentence.[6] Since Llama 3.3-70B was trained on 15 trillion tokens, it understands the subtext of policy language. It can recognize that a "reduction in government expenditures" is occurring even when the document uses bureaucratic euphemisms. The model thus serves as a useful research assistant that processes large amount of textual data to arrive at binary classification.[7]

Task Decomposition and Logical Filtering: We begin our exercise by feeding the model with a PDF of the annual article IV report for a given country. We could directly ask the model if the country experienced austerity. However, a single-stage direct query increases the risk of unconstrained inference, where the model may rely on its pre-training data rather than the specific evidence contained within the provided document. We resolve these issues through a series of

[6] On technical fronts, Llama 3.3-70B has a top-tier 92.1% IFEval score. IFEval score measures LLM's ability to follow specific instructions using automated, verifiable tasks (such as word count) to give a precise score out of 100 on how well it adheres to rules. A high score such as 92.1% means it follows complex logical constraints better than other models.

[7] As a robustness check, we also applied OpenAI's GPT-4o and DeepSeek-R1 models. Somewhat unexpectedly, all three models produced highly similar classifications when evaluated on the rich textual information contained in the IMF's Article IV reports. We interpret this convergence as evidence that, when supplied with well-structured and detailed textual data, contemporary large language models tend to align closely in binary classification tasks.

prompts that serve as a logical filtering mechanism and create a chain of evidence. The first two prompts are as follows:

Prompt 1: If the {country name} is mentioned, does it discuss fiscal policies? Answer "Yes" or "No"

The first prompt serves as a geographical filter and tests the binary classification abilities of the model. It ensures that the model proceeds only if such policies were discussed in the Article IV with reference to a given country. LLM models are significantly more reliable with binary classification relative to open-ended questions. The first two prompts establish a logical gate such that if the model does not find discussion of tax policies in the PDF, then it does not identify an episode of tax-based austerity. These prompts also ensure that the methodology is auditable and if the model makes a mistake, then it could be traced to either failure to identify the country or to recognize fiscal policies.

Prompt 2: If the {country name} is mentioned, does it discuss tax policy? Answer "Yes" or "No"

We follow the first two prompts with a subsequent prompt that serves as a sophisticated classification instruction that transitions the model from general skimming of the document to a more technical identification of austerity. We used a technique called decompressed prompting for our third prompt by explicitly listing specific categories such as policy recommendations and fiscal targets that are of interest to our exercise. The intuition behind the prompt is the same as solving mathematical problems that are easier to solve when broken down into multiple parts. Similarly, decompressed prompts break the complex task of identification of austerity into small individual tasks. The advantage with this approach is that it forces the model to follow a multi-step reasoning which significantly reduces the risks associated with missing subtle measures such as references to broadening of tax base. The prompt further breaks down the term "austerity" into a technical definition used by multilateral institutions. For example, it asks the model to look for policy recommendations, fiscal targets and expenditure or revenue measures nudging it to search for policy intent. Moreover, providing specific "output indicators" such as debt-to-GDP ratio or primary surplus result in the model prioritizing such sections of the PDF and result in a much

higher recall rate for such textual information. The decompressed prompts used as the third prompt ensures that the Llama 3.3-70B model acts as a consistent economic analyst.[8]

Prompt 3 provides the necessary in-context grounding, overriding the model's general pre-trained definitions with a specific, four-pillar technical definition of austerity. This ensures that the binary classification is anchored strictly to the criteria used in the narrative fiscal literature.

> Prompt 3: To identify fiscal austerity from an IMF Article IV document, look for the following indicators:
> 1. Policy Recommendations: Look for recommendations on reducing budget deficits and public debt. This may include suggestions for spending cuts, tax increases, or both.
> 2. Fiscal Targets: Check for specific fiscal targets such as deficit reduction goals, debt-to-GDP ratio targets, or primary surplus targets.
> 3. Expenditure Measures: Identify any proposed or implemented measures to cut public spending. This could include reductions in public sector wages, social benefits, subsidies, or public investment.
> 4. Revenue Measures: Look for measures aimed at increasing government revenue, such as tax hikes, broadening the tax base, or improving tax collection efficiency.
> By examining these elements, determine whether the country is experiencing an episode of fiscal austerity measures in the country. If yes, then answer "yes".

Prompt 4 serves as a structural filter, requiring the LLM to explicitly verify that the identified measures were newly introduced and primary deficit driven. By excluding rule-based adjustments and responses to contemporaneous economic fluctuations, this gate isolates the exogenous component of the fiscal policy, directly addressing the risk of endogeneity present in manual narrative datasets, consistent with Romer and Romer (2010). This refinement is essential for mitigating reverse causality, as it distinguishes policy "shocks" from endogenous responses to fluctuating economic conditions or pre-existing policy paths.

[8] Amgalan et al. (2025) used a single prompt for conducting their analyses of IMF article IV reports instead of following a series of sequential prompts. Their prompt also uses the decompressed prompting technique as they provide specific information to the LLM before providing it with instructions. As we show later, breaking down the instructions into small classification exercises through a series of prompts significantly improves the accuracy.

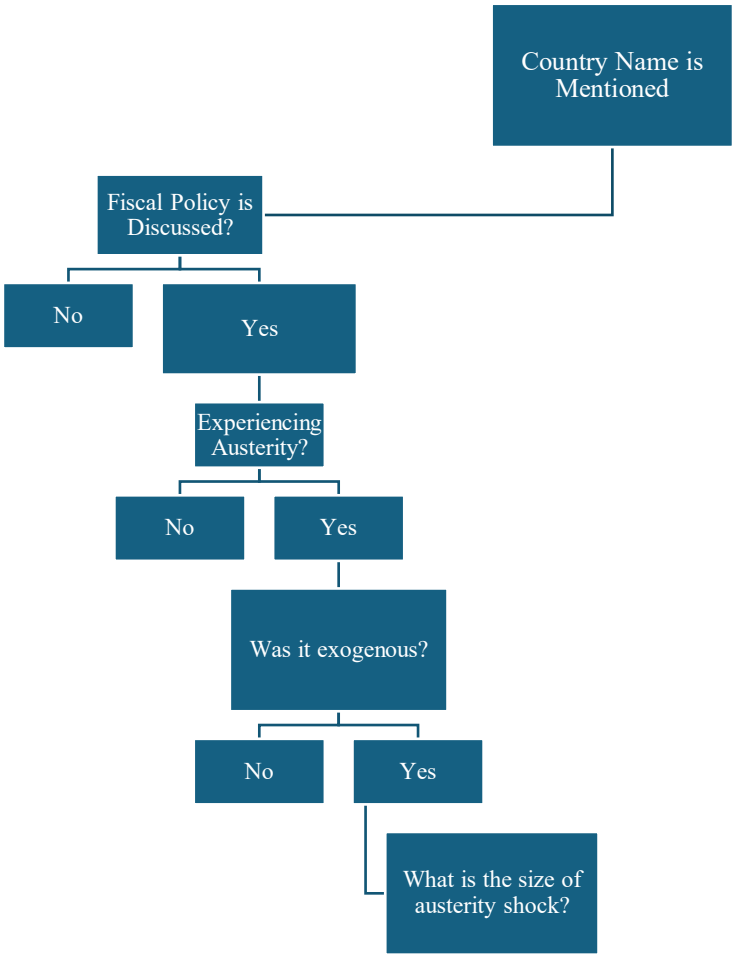
Prompt 4: Does the document explicitly state that the fiscal measures were newly introduced and motivated primarily by reducing the deficit, rather than by structural reforms, long-term goals, institutional rules, or previously announced plans?
Answer 'Yes' or 'No'.

The final step captures the intensity of these shocks using a fifth prompt that directs the model to focus on staff appraisals and other textual information in Article IV reports. The staff appraisal section typically includes data on fiscal adjustments, either embedded in the text or presented in tables. This prompt ensures that the model extracts these details and provides an estimate of the size of fiscal adjustments.

Prompt 5: If a country is experiencing an episode of fiscal austerity, then based on the document, what is the size of austere policies as a percentage of GDP?

As illustrated in Figure 1, our classification procedure follows a sequential decision pipeline designed to maximize precision. While the specific prompts are detailed above, the flowchart visualizes the cascading logic of the model. Prompts 1-2 serve as a geographic and thematic filter to eliminate false positives from non-fiscal sections of the IMF reports. Prompt 3 anchors the LLM's reasoning to a four-pillar definition of austerity, ensuring a consistent 'ground truth' across diverse country contexts. Prompt 4 then applies a stricter narrative-identification test by requiring the model to determine whether any identified fiscal measures were explicitly described as newly introduced and motivated primarily by deficit reduction. This step is crucial as it screens out multi-year reforms, rule-based adjustments, and structural measures, ensuring that only discretionary, deficit-driven actions are classified as exogenous shocks in the spirit of Romer and Romer (2010). Prompt 5 concludes the pipeline by extracting the magnitude of the policy shock, transforming the narrative text into a structured, quantitative metric for austerity shocks.

Figure 1: Hierarchical Structure for LLM-Assisted Austerity Classification



D. Implementation

For the implementation of our methodology, we provide the Llama 3.3-70B model with a PDF of the Article IV for a given country. This is followed by the five sequential prompts which gives the model careful instructions to classify whether the country-year pair experienced austerity, and if so, then quantify its intensity. After each country-year pair is classified and quantified, we re-initialize the model's state, clearing the context window and session history fully before proceeding to the next pair. This ensures the model does not carry over linguistic patterns or decision biases from previous years into subsequent classifications. This protocol ensures independent inference across the longitudinal sample, preventing the model from utilizing information from year $t+n$ to inform its classification of year t. By restricting the model's knowledge strictly to the provided PDF and preventing the accumulation of historical context, we ensure that each classification is based solely on the contemporary information available at the time the report was written.

To maximize the model's analytical precision and to comply with computational token limits, we specifically target the "Policy Discussions" and "Fiscal Policy" sections of the IMF Article IV reports. The explicit focus on these sections ensures that the Llama 3.3-70B model's 128K context window is populated with high-density technical discourse rather than peripheral descriptive data. By focusing the model's attention on these standardized technical chapters, we maximize the signal-to-noise ratio within the context window, allowing for an efficient classification process across our entire longitudinal dataset.

III. Identifying Austerity

Following the procedure outlined in Section II, we now examine instances of austerity across 17 OECD countries from 2004 to 2022. We treat the classification task as analogous to a conventional in-sample forecasting problem. In the same way that a forecast is evaluated against realized outcomes, we compare the binary austerity classifications produced by our procedure with those reported in Adler et al. (2024). Specifically, we validate our identification strategy by modeling austerity as a binary variable that equals 1 when a given country-year is classified as experiencing austerity. For each country-year pair, we obtain two independent classifications: one generated by our LLM-based approach and another drawn from the narrative episodes identified by Guajardo et al. (2014) and updated by Adler et al. (2024). Our validation procedure follows the spirit of Amgalan et al. (2025), who compare their LLM-based classifications with human judgments for a randomly selected set of country-year observations. The core idea is to assess the alignment between human and machine classifications. In our case, the availability of the Guajardo et al. (2014) and Adler et al. (2024) episodes allows us to conduct this comparison for 340 country-year pairs, substantially more than the 105 randomly selected observations used by Amgalan et al. (2025). We summarize the comparison in a 2×2 contingency table in Table 1.

Table 1: Concordance Between Machine-Generated and Narrative Austerity Measures

	Adler: Austerity	Adler: No Austerity
LLM: Austerity	54	1
LLM: No Austerity	32	253

Notes: 1) LLM refers to the classification by our LLM based approach.
2) The number represents the count of instances where country–year observations falling into each classification category.
3) Adler austerity classification is from Adler et al. (2024).

Table 1 shows that, on average, the LLM performs well in identifying austerity: it matches the classification in Adler et al. (2024) in 90 percent of cases. Precision is 0.98, indicating that

when the LLM identifies an austerity episode, it is correct 98% of the time.[9] Specificity is high at 1.00, reflecting a strong ability to correctly classify non-austerity years. The F1 score, which balances precision and recall, is 0.77, and the balanced accuracy, which averages sensitivity and specificity to adjust for class imbalance, is 0.81. The Matthews correlation coefficient (MCC) = 0.74 reflects a well-balanced correlation between predicted and benchmark classifications. From a diagnostic perspective, the model exhibits a positive likelihood ratio of 159.49, meaning that an LLM-identified austerity episode is over 150 times more likely when austerity is truly present, and a negative likelihood ratio of 0.37, indicating a moderate ability to rule out austerity. This yields a high diagnostic odds ratio of 426.9, a summary measure of separability between the two classes. These results show that the LLM provides substantial signal in identifying austerity episodes. The Kuipers score (also known as the True Skill Statistic) for the comparison between our LLM-based classification of fiscal austerity episodes and the benchmark Adler classification is 0.624, indicating a strong level of discriminatory power. This value suggests that the LLM approach is effective at identifying true austerity cases while maintaining a very low false-positive rate. A Kuipers score close to 0.62 implies that the LLM method provides meaningful signal beyond chance, making it a reliable tool for detecting austerity episodes in narrative fiscal datasets.

As shown in the table, the LLM method generally provides classifications consistent with those in the literature. We first examine some select cases where the model aligns with existing episodes of austerity, followed by a discussion of the single instance uniquely identified by our approach, and finally review some of the 32 episodes classified as austerity shocks in the literature that our method rejects. We begin by examining the five instances of austerity that are jointly identified by both the LLM-based procedure and the manual narrative approach for the United Kingdom. Consistent with established historical accounts, both methods classify the years 2010–2015 as episodes of austerity. The 2010 Article IV highlighted the front-loaded fiscal consolidation package introduced in the June 2010 budget, while the 2011 report documented further discretionary tightening, including wage freezes and deep cuts to capital spending. A new round of discretionary consolidation followed in 2013 as part of the ongoing effort to restore fiscal balance, and such measures continued through 2015–16 as the UK pursued long-term fiscal

[9] This compares with 86 percent match rate in Amgalan et al. (2025). Although the two approaches are not directly comparable due to differences in classification criteria and objectives, the contrast in match rates is consistent with sequential prompts improving classification accuracy.

sustainability. This period is widely recognized as one of the most prominent and sustained fiscal consolidation phases in the UK, initiated by the 2010 welfare reforms introduced by the newly formed coalition government and followed by successive rounds of expenditure cuts and revenue-enhancing measures. A substantial literature has since evaluated the social, economic, and political consequences of these policies.[10] For instance, Fetzer (2019) documents a significant relationship between local exposure to austerity measures and increased support for the Brexit referendum.

Our focus now turns to the remaining 10 percent of cases where the LLM's classification diverges from Adler et al. (2024). The sole instance in which the LLM identifies an austerity episode that Adler et al. do not is Denmark in 2010. In this case, the LLM detects an exogenous fiscal consolidation associated with the "Fiscal Consolidation Agreement" reached in May 2010, which was a decisive policy shift responding to the fiscal deterioration following the global financial crisis. The Article IV report highlights that policymakers redirected their efforts toward "bringing fiscal policy back to the pre-crisis trajectory," safeguarding the credibility of the exchange rate peg, and ensuring compliance with the EU's 3 percent deficit limit. The agreement included several newly introduced discretionary measures such as suspending the indexation of income-tax thresholds, reducing the duration of unemployment benefits, and capping child benefits, amounting to a front-loaded consolidation designed to deliver a structural improvement of 1.5 percent of GDP. Because these measures were explicitly new, discretionary, and primarily motivated by immediate deficit reduction rather than long-term structural reforms or automatic stabilizers, the LLM classifies Denmark's 2010 fiscal package as an episode of exogenous austerity shock.

We next turn to the 32 cases classified as austerity by Adler et al. (2024) but rejected by the LLM. Rather than reflecting a lack of sensitivity, these 'rejections' represent a critical refinement of the narrative series. In these instances, Prompt 4 identified that the fiscal consolidations, while present, were either responses to pre-existing fiscal rules (e.g., Belgium 2011) or pre-announced reversals of previous stimulus (e.g., Canada 2010). Under the strict Romer and Romer (2010) criteria, these are endogenous adjustments. By systematically filtering these episodes, our method produces a 'cleaner' structural shock series that is less susceptible to the endogeneity bias present in traditional narrative datasets.

[10] Also see Sawyer (2012), Amann and Middleditch (2017) and Stuckler et al. (2017).

Table 2: Adler-Only Austerity Classifications

Year	Country	Year	Country
2011	Belgium	2013	Ireland
2010	Canada	2014	Ireland
2006	Germany	2007	Italy
2007	Germany	2010	Italy
2013	Germany	2004	Japan
2011	Denmark	2005	Japan
2012	Denmark	2014	Japan
2013	Denmark	2015	Japan
2017	Spain	2019	Japan
2011	Finland	2020	Japan
2012	Finland	2012	Netherlands
2013	Finland	2005	Portugal
2014	Finland	2010	Portugal
2014	France	2011	Portugal
2018	France	2013	Portugal
2011	Ireland	2014	Portugal

Notes: Here we provide instances of austerity shocks as classified by Adler et al (2024) that were either not classified as instances of austerity by the LLM or failed to comply with the exogeneity definition outlined in the fourth prompt.

We proceed by analyzing specific examples from these 32 cases to clarify why they were excluded from our classification of austerity episodes. For instance, Belgium's 2011 consolidation does not qualify as exogenous because its Article IV report indicates that the measures were undertaken to comply with pre-existing fiscal rules, rather than introduced as a newly motivated discretionary deficit-reduction package. For Canada, although the LLM initially classified the 2010 episode as austerity, it was ultimately ruled out as an exogenous change since the IMF (2010, Canada Article IV consultation) notes that the fiscal actions had been pre-announced and that the authorities' plans remained largely unchanged. In Germany, fiscal measures in 2006, 2007, and 2013 were primarily driven by ongoing structural reforms or adherence to established fiscal rules and thus do not qualify as fiscal shocks under our criteria. As shown in Table 2 and illustrated by the specific cases discussed above, applying a somewhat stricter interpretation of exogeneity in Prompt 4, consistent with the Romer and Romer (2010) approach, naturally excludes several interventions that Adler et al. (2024) categorize as austerity shocks.

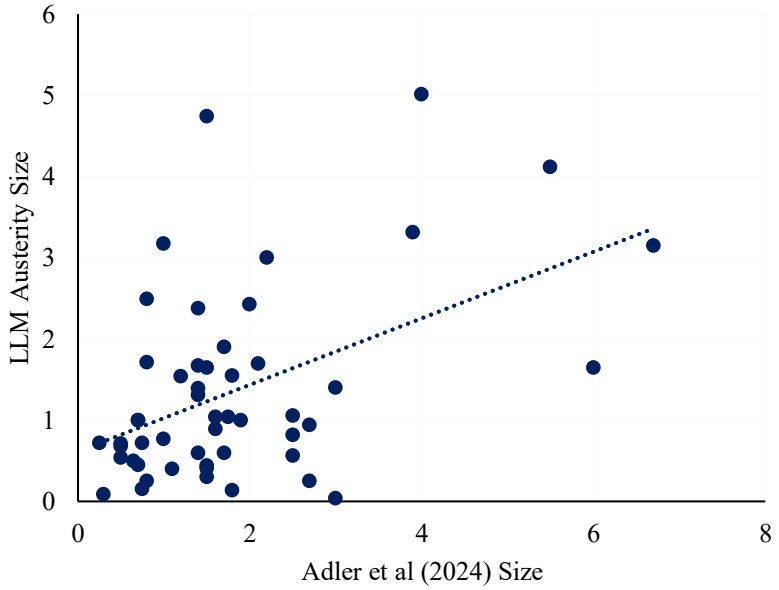
These examples underscore a broader methodological point. Such classification requires careful, consistent application of a given criteria and LLMs may achieve that consistency better than manual reading due to two reasons. For starters, LLMs can be constrained to rely solely on

the textual content, ensuring that classification is based only on what is explicitly stated in the documents. Second, LLMs allow for consistent application of the criteria, something that is difficult to achieve in manual classification, where fatigue, cognitive biases, and subjective judgment may introduce inconsistencies.

Having established that our LLM-based approach yields a consistent and text-grounded measure of austerity shocks, we now turn to comparing the estimated size of these shocks. Accurate measurement of the magnitude of austerity episodes serves as a secondary validation test for the procedure discussed in Section II. In particular, we ask: when both approaches identify an austerity episode, do they also agree on its relative size?

Figure 2 plots the size of austerity shocks estimated by our LLM procedure on the vertical axis against the corresponding sizes reported in Adler et al. (2024) on the horizontal axis, restricting the sample to episodes that both methods classify as austerity. Each point represents an austerity episode jointly identified by the two approaches. The upward-sloping fitted line indicates a systematic positive relationship: as Adler et al. assign larger fiscal consolidations, the LLM method also tends to assign larger sizes. The correlation coefficient of 0.48 reflects moderate concordance in relative magnitudes, which is notable given that our approach applies a stricter exogeneity standard and relies solely on text-based signals from Article IV reports. In other words, even though our method is designed to capture a narrower and more tightly defined concept of "exogenous" austerity, the two approaches nonetheless agree meaningfully on the ranking of the size of shocks when they overlap. This concordance provides secondary validation of the LLM-based measurement of shock magnitude.

Figure 2: Comparison of Austerity Shock Sizes for Jointly Identified Episodes



Lastly, we turn to whether our identified shocks are truly endogenous and compare them with the existing narrative shocks. To evaluate the structural integrity of our shocks, we regress the AI-generated measure and the narrative benchmark on lagged macroeconomic fundamentals, with standard errors two-way clustered by country and year. As shown in Table 3, unlike the conventional narrative measure, the AI-based fiscal shock satisfies the requirements for an exogenous innovation: it is orthogonal to lagged debt levels, past output growth, and its own lagged values.

Table 3: Predictability of Fiscal Shocks: AI-Based vs. Narrative Measures

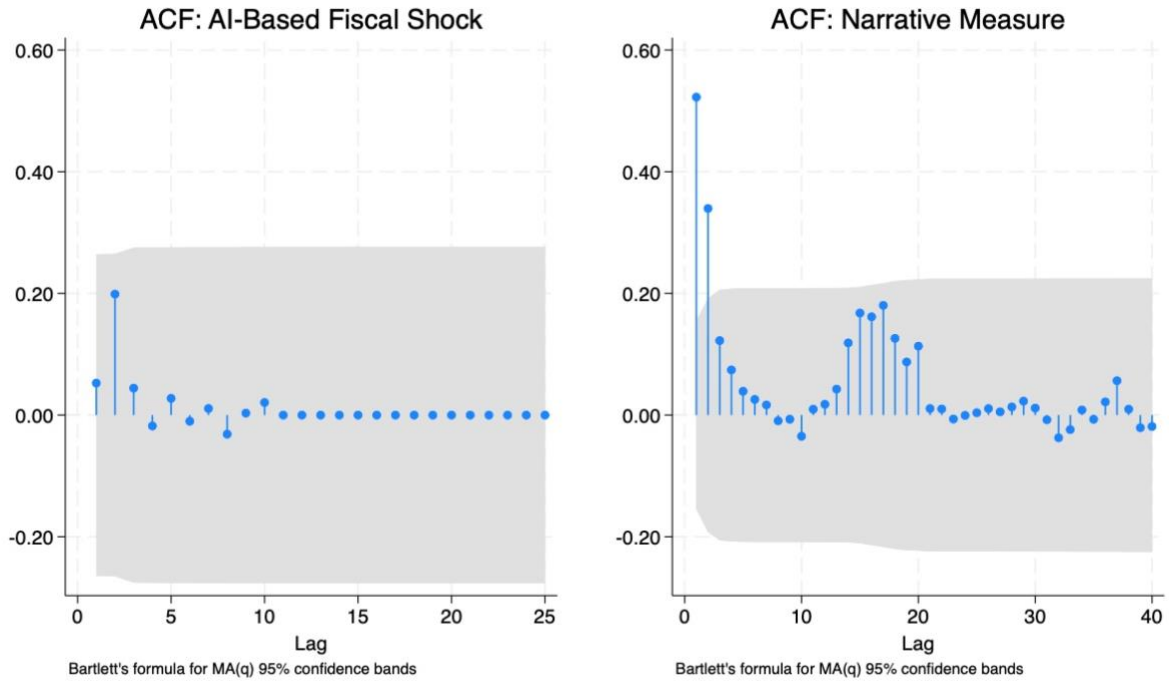
	AI Austerity Shock at *t+1*	Narrative Measure Shock *at t+1*
Government Debt to GDP Ratio *at t*	0.00	0.00
	(0.00)	(0.00)
GDP Growth *at t*	-5.47	-7.91**
	(3.11)	(2.41)
AI Austerity Shock *at t*	0.06	
	(0.13)	
Narrative Measure Shock *at t*		0.50***
		(0.05)
Constant	0.19	0.35***
	(0.11)	(0.07)
$Adjusted\ R^2$	0.27	0.56
N	502	502

Notes: 1. Dependent Variable are the size of fiscal shocks as identified by both the approaches.
2. Standard errors in parentheses.
*3. * p<0.05, ** p<0.01, *** p<0.001*

The results in Table 3 highlight two critical advantages of our computational framework. First, while the narrative measure (Column 2) remains strongly correlated with lagged GDP growth *($p < 0.01$)*, our AI-based shock is orthogonal to the state of the business cycle. This suggests that the hierarchical 'logic gates' successfully filter out endogenous adjustments that human coders may inadvertently include. Second, the AI shock exhibits virtually no serial correlation (0.06), whereas the traditional narrative measure displays substantial persistence (0.50). This persistence in the narrative measure indicates that it likely captures ongoing fiscal regimes rather than isolated innovations. By contrast, our framework identifies idiosyncratic shocks that are significantly less predictable, as evidenced by the much lower adjusted $Adjusted\ R^2$ (0.27 vs. 0.56).

This difference in shock structure is further visualized in the Autocorrelation Function (ACF) plots in Figure 3. The AI-based fiscal shock (Left Panel) exhibits the properties of a white-noise process, with correlations quickly falling within the 95% confidence bands and showing no significant persistence. In contrast, the narrative measure (Right Panel) displays significant spikes at multiple lags and a slower decay. This visual evidence confirms the regression results: the algorithmic audit provides a higher-fidelity identification of fiscal innovations by preventing the 'narrative overlap' often found in manual datasets.

Figure 3: Comparison of Austerity Shock Sizes for Jointly Identified Episodes



Ultimately, our results show that both the binary and quantitative classifications generated by the LLM are highly robust. The model reliably identifies austerity episodes consistent with established historical narratives, but its intentionally stricter definition of exogeneity filters out the policy persistence and cyclical endogeneity identified in Table 3. In line with Romer and Romer (2010), we require fiscal measures to be both newly introduced and primarily deficit motivated. By enforcing this sharper conceptual and statistical boundary, our approach produces a more disciplined and internally consistent set of fiscal shock measures, providing a cleaner foundation for the estimation of fiscal multipliers.

IV. Macroeconomic Effect of Austerity Shocks

Having shown in the previous section that our methodology performs well in both identifying austerity episodes and estimating their magnitudes, we now provide a final layer of external validation by examining the macroeconomic effects implied by our LLM-identified shocks. Beyond providing a final layer of external validation, this section highlights the importance of our identification strategy as we examine whether the economic responses generated by austerity shock identified by LLM align with other narrative shocks identified in the literature. If our shocks indeed capture the same underlying contractionary fiscal events with a stricter definition of exogeneity, then their estimated effects should display patterns broadly consistent with prior work.[11]

To evaluate this, we compare macroeconomic responses using our shocks with those obtained from Adler et al. (2024). We implement two complementary approaches. First, we use propensity score matching to evaluate the quality of our binary classification by treating austerity episodes as the "treated" group and non-austerity episodes as the "control," allowing us to compare economic outcomes across comparable units that differ only in whether an austerity shock occurred. Second, employing a panel time-series framework, we use local projections to trace out the dynamic responses of output, consumption, and investment. Local projections help validate our shock measures by showing whether the dynamic responses associated with LLM-identified austerity shocks resemble those obtained from benchmark series, thereby offering a direct comparison of the economic impact implied by each measure. Together, these exercises allow us

[11] See Alesina and Perotti (1995, 1997), Perotti (1999), Ramey and Shapiro (1998), Alesina and Ardagna (2010), Broadbent and Daly (2010), and Guajardo, Leigh and Pescatori (2011), Perotti (2013), Broadbent and Daily (2010), Auerbach and Gorodnichenko (2012), Jorda and Taylor (2015) and Garcia (2025).

to assess whether the LLM-based shocks generate macroeconomic effects that align with accepted empirical patterns, thereby offering a robust behavioral validation of our identification approach.

Propensity Score Matching: Having outlined our empirical strategy, we now turn to the evidence. Table 4 presents the first piece of evidence, the Average Treatment Effect on the Treated (ATET) from an augmented inverse probability weighting estimator. Here, austerity episodes are treated as the "treated" group and matched to comparable non-austerity observations using public debt and austerity shocks in preceding period as predictors of treatment assignment. This design holds constant observable factors that influence a country's propensity to consolidate, thereby isolating the average contemporaneous effect of an austerity shock on macroeconomic outcomes.

The results reveal a clear and internally consistent pattern. For the LLM-identified shocks, both GDP and consumption fall significantly in austerity episodes: GDP declines by 0.75 percentage points ($p = 0.01$), and consumption by just over 1 percentage point ($p < 0.01$). Investment also falls by a sizable amount; however, the effect is not statistically significant. These effects are strongly aligned with expectations from the narrative fiscal literature, where consumption often exhibits the clearest and most immediate contraction.

When the same exercise is applied to the Adler et al. (2024) shock series, the results look broadly similar: consumption again displays a significant decline of about 0.9 percentage points, and GDP and investment again fall. The key difference is that the LLM-based shocks produce a statistically significant GDP effect, whereas the Adler shocks do not.

The propensity-score evidence establishes two key findings: (i) both shock series produce contractionary effects with the expected signs, demonstrating broad substantive agreement, and (ii) the LLM shocks exhibit stronger statistical precision for GDP. This discrepancy in statistical significance despite similar point estimates is highly revealing. It suggests that our stricter exogeneity filter successfully reduces the residual variance and "noise" typically introduced by endogenous or anticipated policy actions. While traditional narrative measures may suffer from a "dilution" of the shock effect due to the inclusion of rule-based or predictable consolidations (as demonstrated by the high autocorrelation in Section III), the LLM-based identification yields a cleaner estimate of the contraction associated with truly discretionary episodes. Consequently, the gain in precision observed in Table 4 provides empirical validation that the hierarchical logic gates produce a higher-fidelity measure of structural fiscal innovations.

Table 4: Average Treatment Effects on Treated (Augmented Inverse Probability Weighting)

	ATET	**SE**	**P-value**
LLM Shocks			
GDP	-0.75	0.30	0.01
Consumption	-1.02	0.29	0.00
Investment	-1.55	1.09	0.15
Adler et al Shocks			
GDP	-0.69	0.45	0.12
Consumption	-0.90	0.35	0.01
Investment	-1.23	1.15	0.28

Notes: 1) Estimates represent the Average Treatment Effect on the Treated (ATET) calculated via Augmented Inverse Probability Weighting.
2) Each treated unit is matched to its three nearest neighbors using a logistic model for treatment assignment.
3) Reported standard errors are Abadie and Imbens (2016) robust standard errors, which account for the fact that the propensity scores are estimated rather than known.

From a theoretical perspective, these findings shed light on the nature of the fiscal transmission mechanism in OECD economies. The observation that consumption remains the primary and most immediate channel of contraction across both shock measures reinforces the role of liquidity constraints in the propagation of fiscal policy. However, the higher statistical precision of the GDP response under our LLM-based identification suggests that the 'Fiscal Multiplier' is often obscured in traditional narrative datasets by policy anticipation and endogenous rule-following. By isolating truly idiosyncratic innovations, we find that the economic process is highly sensitive to unexpected discretionary shifts, even when those shifts are relatively small in magnitude. This supports the view that the LLM method is not merely reclassifying episodes but is instead capturing a sharper, more genuinely exogenous subset of fiscal consolidations. In the next subsection, we examine whether these conclusions continue to hold in a dynamic setting using local projections, which trace out the medium-term responses of output, consumption, and investment to the two-shock series.

Local Projections: To complement the propensity-score analysis and to examine the dynamic effects of austerity shocks over time, we estimate impulse responses using the local projections (LP) framework of Jordà (2005). Local projections offer a flexible and transparent way to recover the multi-period path of macroeconomic variables following an identified fiscal shock without imposing strong parametric restrictions. In the present context, LPs allow us to ask whether the time-profile of responses generated by our LLM-identified shocks aligns with the

dynamic patterns associated with established narrative shock series. We estimate the following equation:

$$y_{i,t+k} - y_{i,t} = \alpha_i^k + \theta_t^k + \Sigma_{j=0}^{2} \beta_j^k S_{i,t-j} + \Sigma_{m=0}^{2} \gamma_m^k V_{i,t-m} + \epsilon_{i,t}^k \quad (1)$$

in which y is the dependent variable alternatively the log of output, consumption, investment as share of GDP; α_i^k are country fixed effects, θ_t^k are time fixed effects; $S_{i,t}$ is the size of austerity shocks; and $V_{i,t-m}$ are controls such as lagged values of dependent variable. We separately evaluate the macroeconomic impact of austerity shocks identified via the narrative approach of Adler et al. (2024) and our LLM-based identification strategy. In all specifications, statistical inference is based on robust standard errors, clustered by country to provide robustness against arbitrary patterns of intra-group correlation.

Figure 4 presents the impulse responses of output, investment, and consumption to an austerity shock, comparing our LLM-identified shocks with the conventional narrative-based measure. Output falls in the first year after the shock and remains below baseline, while investment and consumption exhibit immediate and persistent declines.

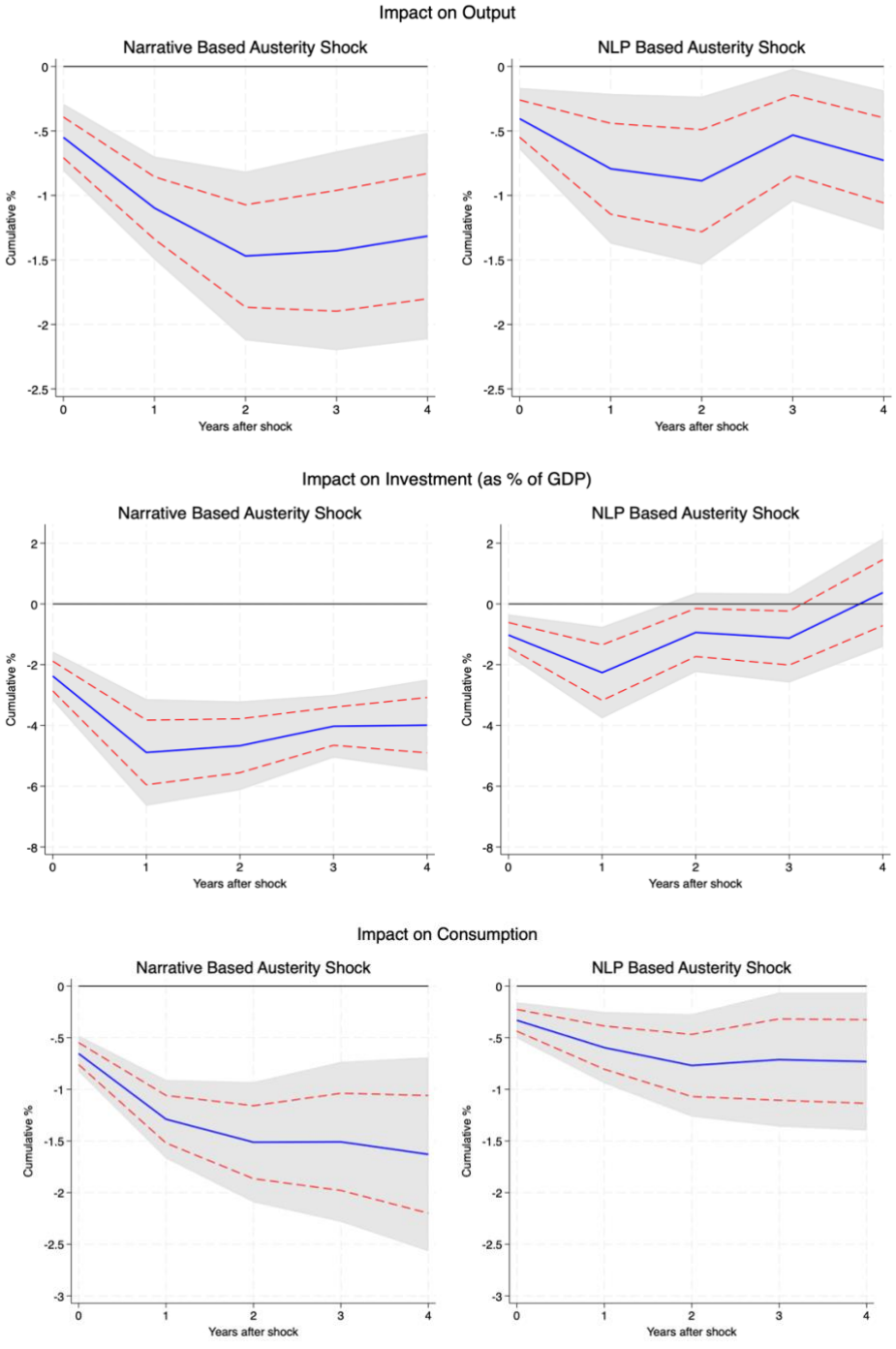
However, a formal comparison of the coefficients reveals a systematic quantitative divergence. The AI-based shocks yield results that are consistently and significantly smaller in magnitude than those of the benchmark narrative series. For example, while the narrative measure implies a GDP contraction of -1.47% by the second year, the AI-based innovation suggests a more moderate response of -0.89% ($p = 0.055$). This divergence is even more pronounced for consumption and investment, where the null hypothesis of coefficient equality is rejected across nearly all horizons ($p < 0.05$).[12]

This quantitative difference provides a strong form of validation for our methodological refinement. Rather than "understating" the impact of austerity, the LLM approach likely provides a more accurate estimation by filtering out the policy persistence documented in Section III. Because traditional narrative shocks often exhibit high serial correlation, they tend to capture the cumulative "pain" of multi-year policy regimes and the associated economic downturns. By

[12]The p-values are derived from a Wald test of coefficient equality using Seemingly Unrelated Estimation. This approach allows for a joint test of the coefficients across models by accounting for the cross-equation correlation of error terms that arises from using the same sample and dependent variables. We then perform Wald tests of equality at each horizon h to determine if the AI-based identification yields statistically distinct macroeconomic multipliers. The p-values for the equality of the AI-based and Leigh/Adler coefficients are: For GDP Growth (H0: 0.262, H1: 0.316, H2: 0.055, H3: 0.023, H4: 0.107); for Consumption (H0: 0.008, H1: 0.005, H2: 0.019, H3: 0.037, H4: 0.045); and for Investment (H0: 0.005, H1: 0.005, H2: 0.001, H3: 0.014, H4: 0.010).

isolation truly idiosyncratic, white-noise innovations, our framework successfully unbundles the shock from the regime.

Figure 4: Impulse Response Functions for Following Austerity Shocks



The formal comparison of the dynamic multipliers reveals a systematic quantitative divergence. Across nearly all horizons for consumption and investment, and at the three-year horizon for GDP, we statistically reject the hypothesis that the two measures yield identical responses. Notably, the AI-based coefficients are consistently smaller in magnitude. For example, at H=3, the narrative benchmark suggests a cumulative GDP contraction of -1.43%, while the AI-based measure implies a significantly milder contraction of -0.53% (p=0.023).

This finding suggests that traditional narrative identification may inadvertently capture "policy regimes" rather than isolated "policy innovations." If a fiscal consolidation is part of a multi-year, predictable program, the narrative shock, which Table 3 demonstrated to be highly persistent, will pick up the compounded effects of that entire regime. Because our LLM framework enforces a strict white-noise criterion, it isolates the impact of the idiosyncratic innovation alone. Consequently, the smaller multipliers found here likely represent a "Pure Discretionary Multiplier": the effect of a single, unexpected austerity shock, unfiltered by the endogenous persistence of long-term fiscal cycles. This suggests that the true cost of an isolated austerity measure is more moderate than prior narrative estimates would imply.

Taken together, the propensity scores matching results and the local projection estimates demonstrate that the LLM-identified shocks are qualitatively consistent with the benchmark narrative series, matching in sign and temporal propagation. At the same time, the improved statistical precision and the more moderate multipliers produced by the LLM suggest that the approach offers a meaningful refinement in fiscal shock identification. By more effectively excluding rule-driven or anticipatory adjustments, the algorithmic audit provides a higher-fidelity measure of structural innovations. The consistency of these results across multiple empirical frameworks validates the LLM approach as a credible, and perhaps more disciplined, tool for identifying austerity shocks.

V. Conclusion

This paper develops a scalable and conceptually grounded framework for identifying fiscal austerity shocks using large language models. By operationalizing a hierarchical decision structure, the "Logic Gates", applied to structured policy narratives, the approach reduces the scope for subjective interpretation that has long constrained manual narrative methods. Implemented on a panel of 17 OECD economies, the framework demonstrates that modern NLP can recover

credible fiscal shocks at scale while preserving the core strength of capturing policy intent rather than cyclical reactions.

A central contribution of this study is the demonstration that LLM-based identification can operate effectively on ex-post policy documents, such as IMF Article IV reports. This extends text-based shock identification beyond the domain of monetary policy, which typically relies on rich, ex-ante real-time communications. Our hierarchical structure enables the LLM to infer discretionary intent from ex-post narratives without importing hindsight bias, providing a bridge between qualitative policy history and quantitative structural analysis.

The empirical evidence is mutually reinforcing. In propensity-score analyses, austerity episodes classified by the LLM produce statistically significant contractions in GDP and consumption. However, the most striking finding emerges from the formal comparison of dynamic multipliers. While the LLM-identified shocks are qualitatively consistent with established benchmarks, they are quantitatively distinct. The AI-based shocks result in multipliers that are systematically smaller, often by more than half, compared to traditional narrative measures.

We interpret this divergence as a significant methodological refinement. As demonstrated by the white-noise properties of our shocks, the LLM successfully isolates idiosyncratic innovations from the persistent "policy regimes" that often contaminate manual datasets. Traditional narrative shocks, which exhibit high serial correlation, likely capture the compounded pain of multi-year austerity paths and their associated economic downturns. By isolating the "Pure Discretionary Multiplier," our results suggest that the true cost of an isolated, unexpected austerity measure is more moderate and more precisely identified than prior estimates would imply.

Beyond validation, the broader implication is that NLP-driven methods can serve as a viable alternative capable of applying the strict Romer and Romer (2010) definition of a policy shock in a more consistent and transparent manner. As textual data becomes more abundant and standardized, the approach offers a path toward high-frequency, cross-country indicators of fiscal actions that were previously infeasible at scale. This, in turn, can enable richer comparisons of consolidation strategies and a closer integration of policy communication and quantitative analysis.

In sum, this paper demonstrates that large language models, equipped with a hierarchical logic, can deliver a higher-fidelity identification of fiscal shocks. The discovery of a milder, purer fiscal multiplier underscores the promise of modern NLP for revitalizing empirical

macroeconomics. As public debt pressures intensify globally, the need for reliable, reproducible, and "clean" identification strategies will be essential for informing both academic inquiry and the design of sustainable fiscal policy.

References

Abadie, A., & Imbens, G. W. (2016). Matching on the estimated propensity score. *Econometrica*, *84*(2), 781-807.

Adler, G., Allen, M. C., Ganelli, M. G., & Leigh, M. D. (2024). *An updated action-based dataset of fiscal consolidation*. International Monetary Fund.

Ash, E., & Hansen, S. (2023). Text algorithms in economics. *Annual Review of Economics*, *15*, 659-688.

Ash, E., Hansen, S., Marangon, C., & Muvdi, Y. (2025). Large language models in economics. *Handbook of Economics and Language, Palgrave Handbooks*.

Amgalan, J., Cohen-Setton, J., & Korinek, A. (2025). *Reading the Fund: A systematic analysis of IMF fiscal advice using large language models* (APAIEO Background Paper No. BP/25-01/05). Independent Evaluation Office of the International Monetary Fund.

Alesina, A., & Perotti, R. (1995). Fiscal expansions and adjustments in OECD countries. *Economic Policy*, 10(21), 205-248.

Alesina, A., & Perotti, R. (1997). Fiscal adjustments in OECD countries: composition and macroeconomic effects. *IMF Staff Papers*, *44*(2), 210-248.

Alesina, A., Ardagna, S., Perotti, R., & Schiantarelli, F. (2002). Fiscal policy, profits, and investment. *American Economic Review*, *92*(3), 571-589.

Alesina, A., Favero, C. A., & Giavazzi, F. (2018, May). What do we know about the effects of austerity?. In *AEA papers and proceedings* (Vol. 108, pp. 524-530). 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association.

Aruoba, S. B., & Drechsel, T. (2024). *Identifying monetary policy shocks: A natural language approach* (No. w32417). National Bureau of Economic Research.

Auerbach, A. J., & Gorodnichenko, Y. (2012). Measuring the output responses to fiscal policy. *American Economic Journal: Economic Policy*, *4*(2), 1-27.

Broadbent, B., & Daly, K. (2010). *Limiting the fall-out from fiscal adjustment* (Global Economics Paper No. 195). Goldman Sachs Global Economics, Commodities and Strategy Research.

Fernández-Fuertes, R. (2025). Monetary Policy Schocks: A New Hope Large Language Models and Central Bank Communication. *BAFFI CAREFIN Centre Research Paper*, (257).

Fetzer, T. (2019). Did austerity cause Brexit?. *American Economic Review*, *109*(11), 3849-3886.

García, C. G. (2025). Fiscal consolidation in heavily indebted economies. *Journal of Economic Dynamics and Control*, 173, 105046.

Giavazzi, F., & Pagano, M. (1990). Can severe fiscal contractions be expansionary? Tales of two small European countries. *NBER macroeconomics annual*, *5*, 75-111.

Guajardo, J., Leigh, D., & Pescatori, A. (2014). Expansionary austerity? International evidence. *Journal of the European Economic Association*, *12*(4), 949-968.

International Monetary Fund. (2006). Germany: 2006 Article IV consultation. Staff report. International Monetary Fund.

International Monetary Fund. (2007). Germany: 2007 Article IV consultation. Staff report. International Monetary Fund.

International Monetary Fund. (2010). Canada: 2010 Article IV consultation. Staff report. International Monetary Fund.

International Monetary Fund. (2010). Denmark: 2010 Article IV consultation. Staff report. International Monetary Fund.

International Monetary Fund. (2010). United Kingdom: 2010 Article IV consultation. Staff

report. International Monetary Fund.
International Monetary Fund. (2011). Belgium: 2011 Article IV consultation. Staff report. International Monetary Fund.
International Monetary Fund. (2011). United Kingdom: 2011 Article IV consultation. Staff report. International Monetary Fund.
International Monetary Fund. (2013). Germany: 2013 Article IV consultation. Staff report. International Monetary Fund.
International Monetary Fund. (2013). United Kingdom: 2013 Article IV consultation. Staff report. International Monetary Fund.
International Monetary Fund. (2015). United Kingdom: 2015 Article IV consultation. Staff report. International Monetary Fund.
Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review*, *95*(1).
Jordà, Ò., & Taylor, A. M. (2016). The time for austerity: estimating the average treatment effect of fiscal policy. *The Economic Journal*, *126*(590), 219-255.
Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, *61*(4), 1281-1317.
Korinek, A. (2024). LLMs level up—Better, faster, cheaper: June 2024 update to section 3 of "Generative AI for economic research: Use cases and implications for economists". *Journal of Economic Literature*, *61*(4), 1-38.
Leigh, D., Devries, P., Freedman, C., Guajardo, J., Laxton, D., Pescatori, A., ... & Yang, J. (2010). Will it hurt? Macroeconomic effects of fiscal consolidation. *World Economic Outlook*, 93-125.
Li, J., Zhang, Q., Yu, Y., Fu, Q., & Ye, D. (2024). *More agents is all you need*. OpenReview.
Mountford, A., & Uhlig, H. (2009). What are the effects of fiscal policy shocks?. *Journal of Applied Econometrics*, *24*(6), 960-992.
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.
Owyang, M. T., Ramey, V. A., & Zubairy, S. (2013). Are government spending multipliers greater during periods of slack? Evidence from twentieth-century historical data. *American Economic Review*, *103*(3), 129-134.
Perotti, R. (1999). Fiscal policy in good times and bad. *The Quarterly Journal of Economics*, *114*(4), 1399-1436.
Perotti, R. (2013). The "Austerity Myth": Gain without pain? In A. Alesina & F. Giavazzi (Eds.), *Fiscal Policy after the Financial Crisis* (pp. 307–354). University of Chicago Press.
Ramey, V. A., & Shapiro, M. D. (1998, June). Costly capital reallocation and the effects of government spending. In *Carnegie-Rochester conference series on public policy* (Vol. 48, pp. 145-194). North-Holland.
Romer, C. D., & Romer, D. H. (2010). The macroeconomic effects of tax changes: estimates based on a new measure of fiscal shocks. *American Economic Review*, *100*(3), 763-801.
Romer, C. D., & Romer, D. H. (1989). Does monetary policy matter? A new test in the spirit of Friedman and Schwartz. *NBER Macroeconomics Annual*, *4*, 121-170.
Romer, C. D., & Romer, D. H. (1997). Identification and the narrative approach: A reply to Leeper. *Journal of Monetary Economics*, *40*(3), 659-665.
Sawyer, M. (2012). The tragedy of UK fiscal policy in the aftermath of the financial crisis. *Cambridge Journal of Economics*, *36*(1), 205-221.

Stuckler, D., Reeves, A., Loopstra, R., Karanikolos, M., & McKee, M. (2017). Austerity and health: the impact in the UK and Europe. *The European Journal of Public Health*, *27*(Supplement 4), 18-21.

Tillmann, A. (2025). Literature Review Of Multi-Agent Debate For Problem-Solving. *arXiv preprint arXiv:2506.00066*.