Putting the Public Back in Public Debt: Citizen Narratives on Public Debt Burdens

Matthew DiGiuseppe
Leiden University*

Alessia Aspide
Leiden University

Bastián González-Bustamante
Universidad Diego Portales

Jaroslaw Kantorowicz
Leiden University

September 4, 2026

Abstract

Models of public debt accumulation and consolidation rest on assumptions about citizen engagement with fiscal policy and the cleavages that form around policy responses to growing debt. To assess competing models, it is necessary to build evidence for these underlying assumptions. Building on recent work that analyzes closed-ended survey responses, we explore how the public engages with public debt by analyzing original open-ended survey data from three highly indebted countries: Italy, Japan, and Brazil. We probe both the public's understanding of the consequences of rising debts and preferred policy solutions to the problem. We then examine if these responses align with assumptions of public preferences assumed in prominent models of debt politics. Our evidence suggests that the public is neither the well-informed voters assumed by Ricardian Equivalence and other models nor the uninformed voters assumed under Fiscal Illusion theories. The public in these countries largely expects that rising public debt has negative but unspecific consequences, as it does not readily connect debt to future policy outcomes (taxes and spending cuts). Next, we find little evidence of generational or political disagreement on policy solutions to high debt.

*Corresponding author: m.r.di.giuseppe@fsw.leidenuniv.nl

Scholars from several disciplines have proposed theories for why some countries can maintain sustainable debt levels while others fail to restrain deficits and pay down debt. These explanations largely focus on the decisions of elites within a government, whether they be heads of state, political parties, finance ministers, organized interests, or bureaucrats. Yet, until recently, scholars have spent little time trying to understand how the public reasons about public debt. This is not to say that others have ignored the public completely. Models theorizing about the behavior of elite decision-makers rest on explicit or implicit assumptions about how the public does or does not shape the incentives and behavior of politicians. Assumptions about the public range from passive bystanders, to myopic spendthrifts, to perfectly informed fiscal conservatives (Buchanan 1967; Buchanan and Wagner 1977; Alesina and Tabellini 1990; Alesina and Drazen 1991; Alesina et al. 2019; Yared 2019). Further, scholars assume that the public is divided on debt policy by either class lines (Alesina and Drazen 1991) or by age (Song et al. 2012; Cukierman and Meltzer 1989).

Surprisingly, researchers have only begun to sort out these assumptions that stand at the base of these prominent theoretical models. As with most studies of the public's economic preferences and attitudes, the nascent literature aimed at understanding citizens' preferences toward debt has utilized experimental or observational analysis of closed-ended questions in survey data and has focused on examining differences in opinion toward either deficits, debt financing, debt rules or, most commonly, fiscal adjustment to reduce debts (Aspide et al. 2023a,b; DiGiuseppe and Del Ponte 2026; Curtis et al. 2014; Curtis 2014; Ardanaz et al. 2025b, 2024; Kantorowicz 2023; Kantorowicz and Metelska-Szaniawska 2025; Roth et al. 2022; Barnes and Hicks 2018, 2022; Bremer and Bürgisser 2023). Other scholarship has also looked at how direct involvement of citizens through, e.g., referendum and initiatives on fiscal issues affects public debt and spending (Asatryan et al. 2017; Blume et al. 2009; Feld and Kirchgässner 2001; Funk and Gathmann 2011; Matsusaka 2018; Curtis et al. 2014).

In this paper, we diverge from these approaches and instead examine open-ended re-

sponses to questions about the consequences of and solutions to high public debt in three highly indebted countries: Italy, Brazil, and Japan. Open-ended responses allow us to examine preferences without directing respondents into prescribed and predefined answers that cue them to think about the topic in a specific way. As such, open-ended responses can give researchers more accurate, though more challenging to analyze, insights into citizens' mental models, reasoning, and depth of understanding of economic policy (Haaland et al. 2025; Andre et al. 2026). This is especially important in the area of public debt given the common assumptions that people heavily discount debt or have a limited understanding of its consequences (Buchanan and Wagner 1977; Shi and Svensson 2006; Bremer and Bürgisser 2023).

Our analysis first looks for evidence of two broad assumptions about the public found in the existing models. The first assumption is that high debt burdens ultimately stem from citizens' ignorance of fiscal matters that serve the short-term interests of politicians (Buchanan and Wagner 1977; Dollery and Worthington 1996; Lizzeri and Yariv 2017; Bisin et al. 2015). The second, in contrast, claims that citizens are sufficiently informed and fear the consequences of debt and deficits and that politicians have short-sighted preferences (Rogoff and Sibert 1988; Persson and Svensson 1989; Yared 2010; Brender and Drazen 2008; Alesina et al. 2019). This follows closely from the Ricardian equivalence assumption that citizens save in response to the expectation that government debt will mean higher future taxes (Barro 1974). However, more recent treatments do not strictly adhere to a tax-based response to high debts. To help make progress on this and other related debates, we examine if citizens expect, unprompted, future fiscal contraction (higher taxes, less government spending) or economic crises as debt grows. We find that citizens in all three countries, almost unanimously, expect negative economic consequences from a further increase in public debt. However, their expectations are not detailed. Instead of pointing to specific mechanisms, like higher taxes, they point to general effects like reduced quality of life and increased

poverty.

Next, we examine assumptions about cleavages over debt policy. Models of debt accumulation often center around distributive conflict along generational lines (Cukierman and Meltzer 1989; Song et al. 2012; Tabellini 1991; Aspide et al. 2023a) or ideological/class lines (Persson and Tabellini 2002; Alesina and Tabellini 1990; Alesina and Drazen 1991; Lizzeri 1999; Müller et al. 2016; Bansak et al. 2021; Barnes and Hicks 2018; Nelson and Steinberg 2018). We probe if these cleavages are present in citizens' own narratives by looking for a correlation between age and left-right placement and policies mentioned in citizens' responses. Our analysis reveals differences in proposed policies to address debt, showing slight variation with age and ideology. However, the effects are inconsistent across countries.

Importantly, we do not claim to provide decisive evidence for or against particular assumptions and the models that follow from them. Instead, our exercise is one of calibration. By looking at how descriptive evidence aligns with assumptions made by prominent models, scholars have additional information to weigh the value of competing models of sovereign debt in their application to explain politics in highly indebted countries like those we study here. Our evidence can help policymakers, on the other hand, have a better understanding of the constraints that policy proposals are likely to face in implementation. Our results suggest that assumptions of voter ignorance are probably overstated. The public is largely united in its distaste of increasing debt burdens. Further, we provide evidence that divisions over debt policy are not easy to classify across standard left-right or generational boundaries.

In addition to looking for evidence of existing assumptions about preferences for taxes or spending, our analysis also helps uncover latent themes in the public's public debt narratives. These insights can inform future research and models of debt accumulation and consolidation. For example, we find that citizens are keen to point to efficiency gains and fighting corruption as ways to reduce debt instead of austerity measures that are drawn on in most academic research. Further, a decent proportion of the public points to non-consequential and symbolic

policy solutions, like cutting politician salaries.

The Public in Sovereign Debt Models

Economists generally model decisions about budget deficits and debt consolidation as a top-level decision by a country, a benevolent social planner, or by competing political groups. Prominent models of sovereign debt are no exception. However, all these models rest on implicit or explicit assumptions about either the economic or political motivations or behaviors of citizens in responding to increasing debt burdens and the need for public debt reduction. In all regime types, citizens can constrain the options of politicians. Yet in democratic states especially, their influence can be hard to ignore. Historical and contemporary evidence points to many cases where efforts to rein in spending and repay debts have run up against public opposition (Walton and Ragin 1990; Rüdig and Karyotis 2014; Ponticelli and Voth 2020; Ketchley et al. 2024). As such, even models that do not consider the public are assuming the public is irrelevant to policy decisions or has consistent preferences aligned with decision makers.

There are several key assumptions that we will engage with in our analysis. The first two assumptions center on citizens' knowledge and expectations of public debt. Sorting between these assumptions allows us to better judge how we can incorporate public preferences in theoretical models that have varied assumptions.

On one extreme sits the assumption of Ricardian Equivalence that assumes economic actors, including voters, are informed and rationally anticipate the consequences of higher debt. Essentially, they equate debt with future taxation and must save and collect interest to repay higher government taxes in the future (Barro 1974, 1979). This idea of Ricardian Equivalence implies that citizens are a) paying attention to economic policy and b) have a nuanced understanding of the impact of public debt on both government policy and their

own future welfare. Although Ricardian Equivalence does not serve as the basis of canonical models of sovereign debt accumulation, related assumptions of fully rational, if imperfectly informed or temporarily uninformed voters, are commonplace (Rogoff and Sibert 1988; Persson and Svensson 1989; Yared 2010). For example, political business cycle theories assume that the public is informed and rational but it takes time for them to observe changes in government fiscal and monetary policy (Rogoff and Sibert 1988).

Models that assume voter rationality often see the short-sightedness of politicians as the cause of high public debt (e.g., Yared 2010). On the other end of the spectrum are scholars questioning the assumptions that voters are informed and rational in their expectations. They see the rising debt stemming from the public's inattention or ignorance of government debt and politicians' incentive to exploit this for short-term electoral gain. Suppose voters are unaware or ignorant of the costs of debt financing. In that case, politicians can engage in "fiscal illusion" to give them the impression that they are providing costless public or private goods while they are simply shifting the burden to future periods (Buchanan and Wagner 1977; Congleton 2001; Shi and Svensson 2006; Yared 2019). Such arguments can stem from a Downsian (1957) understanding of voters, which contends that they have an incentive to forgo the process of information acquisition. It can also stem from an assumption of time-preference in which voters have a bias that results in voters, not politicians, who discount the future (Bisin et al. 2015). The former functionally assumes ignorance, the latter assumes indifference.

Thus far, research on the rationality or attentiveness of voters has examined savings behavior in economic data or in lab experiments conducted on hypothetical economic circumstances (Ricciuti and Di Laurea 2003) or has tried to tease out cross-national differences. For example, several studies examine the correlation between country-level measures of informed voters and national budget deficits (Shi and Svensson 2006; Janků and Libich 2019). Data at such a high level of aggregation limits what can be said about citizens-level assump-

tions and lab behavior can treat 'ignorance,' but cannot tell us about how much of it exists in real-world scenarios. Other research consistent with the assumption of well-informed citizens examines how voters respond to deficits. For example, Brender and Drazen (2008) and Alesina et al. (2019) look at electoral outcomes and find a correlation between deficits and leader removal. However, the result might stem from the downstream effects of deficits – interest rates or high capital costs – rather than direct punishment of deficits themselves. As such, it is hard to directly infer what voters think or if their actions are explained by more readily observed confounders not picked up by macroeconomic statistics. Examining open-ended questions can provide insight into not only what citizens expect as debt increases but what is the depth of their understanding relative to macroeconomic theory (Haaland et al. 2025).

The next set of assumptions about the public's relationship with public debt concern the distributive consequences of debt reduction. Political economy models of debt often assume that differences in the importance and solutions to public debt fall along generational or left-right/class-based cleavages. For example, many have suggested that generational divides define preferences for debt reduction as the elderly have less interest in consolidation than the young who will be responsible for repaying future debts (Cukierman and Meltzer 1989; Tabellini 1991; Song et al. 2012) and recent evidence suggests a non-linear relationship between age and support for public debt reduction (Aspide et al. 2023a). This evidence largely relies on closed-ended questions from large, omnibus surveys and thus misses nuances in the differences in the ways young, middle-aged, and the old think about public debt.

Two strands of literature predict that traditional left-right divisions will generate cleavages in public debt consolidation. This stems from two separate arguments. The first is that material interest, often funneled through left-right distributional conflict, shapes cleavages. Class conflict then determines the shape of reform, either leading to spending cuts or progressive tax increases or the incentives to run further deficits (Tabellini and Alesina 1990;

Battaglini and Coate 2008; Dovis et al. 2016; Bierbrauer et al. 2021). Empirically, Curtis (2014) and Curtis et al. (2014) find that material interest variables, along with partisanship, are strong predictors of support for repaying Iceland's debt following the great recession.[1]

It is also possible that material interests do not generate political preferences, but instead, ideology or partisanship itself shapes attitudes. More recent research suggests that differences in attitudes towards debt accumulation and austerity result from partisan cues rather than material interests (Nelson and Steinberg 2018; Barnes and Hicks 2018; Bansak et al. 2021). Citizens' views on public debt, according to this argument, stem not from their material interest but rather follow from co-partisan political leaders shaping debates. This argument also stems from a view of uninformed voters who have been socialized to espouse particular policy positions with little reflection.

In the sections below, we use citizens' own narratives to find evidence consistent with these prominent assumptions about citizens' views on public debt. If assumptions of informed voters were to hold true, we would expect that voters are aware of the consequences of sovereign debt accumulation and potential default. Further, the models assume that respondents can tie specific policies responsible for an increase in debt and anticipate what policies may be required to deal with it. If voters are uninformed and ignorant, we would expect the opposite. They would have poor insight into the consequences of public debt – unable to tie rising public debt to specific future consequences.

Lastly, if there are salient cleavages, around age and partisanship, regarding the solutions to sovereign debt crises centered around who should shoulder the burden, we should observe that those solutions differ along class, partisanship or generational lines. Further those differences should reflect a preference to avoid contributing to the cost of adjustment.

[1]While class and partisanship are two distinct concepts, in practice, parties are often a vehicle to organize class interests and thus it is difficult to untangle the effects empirically.

Why Open-Ended Responses?

Closed-ended questions have been the standard in assessing citizen preferences over policy and politics. They have the advantage of standardization, making them easy to analyze in a systematic and structured way. The cost of this standardization is the bias generated by forcing respondents to choose among a set of responses introduced by the researcher that may not be initially considered by respondents (Connor Desai and Reimers 2019, 1427). As such, closed-ended questions likely overstate preferences for displayed options. Further, because the choice of options is limited, they inherently limit potential responses to those provided by the survey designers, even if the option "other" is provided. For example, a question about preferences for debt reduction may ask respondents to choose between a variety of taxes or spending cuts. This may exclude other preferences, such as those for increased investments (aimed at GDP growth) or greater tax enforcement, as we see below. The constrained choice might lead scholars to conclude, as Bremer and Bürgisser (2023) do, that citizens are unwilling to take action to address debt. This is especially a problem when it comes to citizen opinions on fiscal policy. Public debt's relevance to almost every aspect of the economy makes it difficult to distill it into a few neat options, irrespective of whether researchers are asking about the causes, consequences or solutions to high public debt. As such, the value of open-ended responses is likely to be particularly high in efforts to understand public debt attitudes and other issues with such a wide breadth.

Beyond providing insight into the diversity of preferences, open-ended responses may also allow for a greater understanding of the limits of public understanding. Given options, survey respondents may be reluctant to select "don't know" due to a variant of social desirability bias. This is a well-known issue with closed-ended public opinion surveys in which respondents report an opinion but really do not have strong views on the topic (Bishop et al. 1986; Converse 2006). Even if respondents select "don't know", they may be masking uncertainty

about the topic (Graham 2021). In an open-ended setting, they do not have an easy way out, and their ignorance or ambivalence is more easily revealed. Further, they may also reveal preferences for nonsensical solutions that are, for good reason, excluded from closed-ended options. For example, we find that a sizable proportion of Japanese citizens advocate for reducing politicians' salaries – a small solution to the large debt burden of 260% of GDP. As such, open-ended questions reveal the limits of citizen understanding to a greater extent than closed-ended questions. Further, open-ended responses can help the research discovery process by bringing attention to issues researchers have yet to (inductively) theorize about or consider in economic models (Haaland et al. 2025; Stantcheva 2021; Andre et al. 2026; Ferrario and Stantcheva 2022). As we see below, reducing tax-evasion and anti-corruption reforms play a non-trivial role in the public's preferred policy responses. Yet, survey research tends to ignore these options.

Research Design and Original Survey Data

Despite the benefits of open-ended responses, they are infrequently used in research because imposing structure on the raw data had, until recently, required costly human coders. These coders would have to read respondent answers and categorize responses into predefined classes or identify other elements of the text, like sentiment or topics. When responses are sufficiently large in number this creates monetary costs beyond the reach of most social scientists. More recently, text-as-data innovations, like semantic networks, topic modeling or machine learning more generally, have allowed for the automation of this process. However, the techniques rely mainly on analyses of single words or 2-3 word strings that necessarily omit important context and relationships between concepts. Further, attempts to measure sophistication or knowledge with traditional text-as-data tools often diverge from domain knowledge. Instead, they assess proxies like language and discursive sophistication, which

evaluates the number of topics, dispersion of topics, and connections between topics divorced from the topic context (Kraft 2024). As such, they measure complexity rather than actual content. As described below, LLMs overcome cost and methodological limitations as they replicate the work of human coders and often do so at a higher level (Mellon et al. 2024; Heseltine and Clemm von Hohenberg 2024; Ornstein et al. 2025; Gilardi et al. 2023; Le Mens and Gallego 2025). This allows researchers to automate the classification process that then permits statistical description and analysis as we plan to do here.

Our analysis relies on original survey data collected in three countries with high debt burdens but which have not yet experienced a sovereign default or restructuring – Japan, Italy and Brazil. In each country, we recruited participants from quota samples of more than 1,500 that are representative of the domestic population on the dimensions of age, region and gender.[2] We demonstrate in the Electronic supplementary material how our sample compares to probability-samples in each country (Italy: European Social Survey, Japan: World Values Survey, Brazil: LAPOP AmericasBarometer). Compared with these benchmark surveys, our samples are non-trivially more educated. The share of respondents with tertiary or university-level education is about 35% in our Italian sample versus roughly 15% in the ESS, 32% in our Brazilian sample versus roughly 18% in LAPOP, and 44% in our Japanese sample versus roughly 28% in the WVS – gaps of 14–20 percentage points. On income, our Brazilian sample is somewhat lower-earning than the LAPOP benchmark, while the Italian and Japanese samples track their benchmarks closely. In all three samples, the distribution of left-right self-placement largely mirrors the benchmark surveys. We return to the educational skew in the Conclusion's limitations discussion.[3]

We chose three democracies that are at risk of a debt crisis to get a better understanding

[2]We recruited participants with Respondi in Italy in July 2022, Rakuten Research in Japan in October 2023, and Netquest in Brazil in September 2022.

[3]Given our writing task is cognitively intensive, we examined attrition. We find low attrition rates ($<$ 10%) and that attrition does not systematically correlate with key demographics like age, income, education, left-right orientation, or gender.

of narratives when debt is a salient political topic, but attitudes have not been crystallized or polarized by a recent costly default. Arguably, this positioning is most interesting from a theoretical perspective as it gives us insight into the politics on the path to a debt crisis and when the potential to change course remains an option. This is essentially the moment when we want to know why stabilizations are delayed. While informative, examining states where public debt is not salient is potentially less interesting as citizens have little incentive to develop narratives on public debt. As such, our findings cannot speak to citizen engagement in states with low but rising debt burdens. We leave such analyses for future research.

The countries we chose also differ on several dimensions. Each of the countries we selected differs in their level of economic development but also in their debt trajectory, monetary regime (monetary union vis-à-vis floating rates), and integration in external capital markets. Further, the countries differ in their political systems and underlying demographics. Japan and Italy have aging populations, while Brazil does not face strong demographic headwinds. Italy and Brazil have salient left-right political competition, while in Japan left-right structuring of fiscal policy is historically muted. This heterogeneity is by design. Our three-country selection is intended to support claims of external validity within the population of high-debt democracies. Importantly, the case selection does not to enable a formal cross-country comparison. Readers interested in cross-national comparison should treat our analysis as descriptive triangulation across a small, deliberately heterogeneous set of high-debt democracies rather than as a most-similar or most-different systems design.

In each country, we asked the same four questions[4] and asked respondents to answer in two to three sentences:

- In your opinion, what are the main causes of [Italy, Brazil, Japan]'s high public debt?

[4]In Italy and Brazil, we asked an additional question before these four questions: "When you think about [Country]'s public debt, in wondering whether it is too high or too low, what are the main considerations that come to mind?". We omitted this in Japan for space considerations after seeing it did not lead to useful results. In each case, the questions were preceded by basic demographic questions.

- **What do you think would happen to the [Italian, Japanese, Brazilian] economy if the national debt continues to increase?**
- **What do you think would happen to your economic situation if the national debt continues to increase?**
- **What policies should the government adopt to reduce [Italy, Brazil, Japan]'s public debt?**

We analyze the last three questions to probe both the anticipated consequences of a further increase in debt and cleavages in proposed policy solutions in the paper below.[5] Given the multilingual capabilities of frontier LLMs, we use the untranslated text. We remove nonsensical responses, such as a few instances of random typing to fill in the text box. We use the LLMs to identify these responses and omit them accordingly.

Annotation Workflow

We bring structure to our open-ended responses in several steps. For each question-response (e.g. national consequence) we first needed to identify topics/clusters in the raw data to classify downstream. Toward this end, we adopt a workflow akin to BERTopic (Grootendorst 2022) with several adaptations to accommodate the fact that we use multiple countries and each response can identify multiple consequences or policies.

We first turn each response into a respondent-topic summary. For each question, by country, we ask an LLM, `Qwen3-235B-A22B-Instruct-2507`, to identify and label (in English) each consequence, cause or policy mentioned in the respective question without providing prescribed categories. We then create a dataset that includes each respondent-item (e.g., respondent-policy) for each question in the event a respondent identified more than one.[6] Because the open-classification by the LLM can return different words for the same basic

[5] In the Electronic supplementary material, we present a classification of responses to the cause question.
[6] We present the prompt for each question in the Electronic supplementary material.

meaning, we then fed each phrase returned by the LLM into a text embedding model.[7] This converts each phrase into a numeric vector in a latent space. Like BERTopic, we reduce the dimensionality of these embeddings with Uniform Manifold Approximation and Projection (UMAP) to 15 dimensions and detected preliminary topics via Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN). We then condensed these preliminary topics by clustering via centroid cosine distances until k=10 categories were left for each country.

Lastly, to aid in human-labeling these categories, we generated human-readable labels and fuller descriptions by combining c-TF-IDF n-grams. At this point, the researchers used these labels, descriptions and examples of phrases in each category to determine 10 category names to be used in the final classification by an LLM. The research team created a standardized list of 10 categories (ten substantive categories, plus *Other* and *Don't know*, for twelve labels in total) to generalize across all countries by first including topics that were common in all 3 countries. Second, we prioritized topics that align with assumptions made in the public debt literature and, secondarily, topics that were heavily prominent in some countries. For example, we added a "no consequence" option to capture those who expressed there will be no change given an increase in high debt burdens as it aligns with fiscal illusion.

After we identify 10 categories for each question, we proceed to classify each response. To do so, we prompt an open-weights reasoning model, `Qwen3-235B-A22B-Thinking-2507`,[8] to engage in a zero-shot multi-label classification of the original text.[9] We set the model

[7]We use OpenAI's text-embedding-3-large. We use a proprietary embedding model here because, at the time of analysis, it offered the strongest multilingual performance for the short non-English phrases produced by the previous step. We note this as a deviation from the open-weights standard (Palmer et al. 2024) and discuss replication implications in the LLM Choice section of the Electronic supplementary material. Importantly, the embedding step provides scaffolding for the human-in-the-loop category-selection step described below. The published category prevalences and cleavage estimates are reproduced from the open-weights Qwen3 classification step against the human-decided 10-category schemes, not from the embeddings themselves.

[8]We explain our model choice in the Electronic supplementary material.

[9]In both sets of LLM calls we rely on an open-weights model so that our analysis could be replicated given the fear that proprietary models may be discontinued (Barrie et al. 2024). We called the model via

temperature to "0" allowing the LLM output to be deterministic. In the event that the response is unidentifiable or the response mentions a topic outside the categories, we instruct the LLM to code the response as "other" with a short description. We also instruct the LLM to identify if the respondent responds with "don't know". We use different prompts per question but retain the same prompt regardless of the country.[10]

LLM Prompt (Batch Chat Completions)

```
[System] You are a deterministic classifier. Output exactly one valid JSON object
    matching the user-provided format. No markdown/backticks/explanations.
[User] I asked survey respondents in Italy, Japan, and Brazil:
What do you think would happen to your economic situation if the national debt
    continues to increase? - Classify personal/household consequences the
    respondent expects from rising debt.
Output ONLY this exact JSON format (no other text):
{ "categories": [1], "other_note": null,"is_dk": false }
RULES:
- categories: array of 1--10 integers from list below (ordered by importance)
- other_note: string if using category 11 ("Other"), else null
- is_dk: true only if response is "don't know" (then categories=[12])
CATEGORIES:
1. Poverty or General Economic Hardship
2. Unemployment or Underemployment
3. Higher prices, Inflation, Currency Depreciation
4. Emigration (self or relatives)
5. Personal Sacrifice (e.g. sell assets, go hungry)
6. Increased Taxes
7. No Impact or change
8. Loss of government benefits or services (e.g. Pension, Healthcare)
9. Falling behind others (inequality)
10. Uncertain Future
11. Other
12. Don't know
STATEMENT: "<respondent free-text here>"
JSON OUTPUT:
```

We validate this classification in two ways, reported in full in the Electronic supplementary material. First, members of the author team independently hand-coded random samples of responses against the same category schemes. We find our expert-level agreement

batch processing on Fireworks.ai.

[10]We present prompts for the other questions in the Electronic supplementary material.

with the LLM (F1, excluding the residual "other" category) is consistent across countries and questions, ranging from roughly 0.60 to 0.73 and clustering near 0.70 – about the same level at which the human coders agree with one another. Second, we re-classified the same responses with two independently developed frontier models from other providers (OpenAI's GPT-5.5 and Anthropic's Claude Opus 4.6). These models agree with the Qwen model at comparable or higher levels than the expert coders (micro-F1 of roughly 0.74–0.82, excluding the residual "other" category). In both cases, agreement is strong on the substantive categories that carry our findings and is lowest only on a residual "other" bucket and a few rare categories. Taken together, the classification indicates that the Qwen model is sufficient for this multi-label classification task.

Evidence of Voter Sophistication

Recall that there are varying assumptions about the degree by which respondents are informed about the consequences of rising debt burdens. We first explore the degree by which respondents can identify the consequences of rising debt burdens at a national and individual level. Here, we examine the distribution of responses in light of expectations of responses that we infer from the assumptions we discuss above. In light of the fact that there is no statistical test we can run to either confirm or deny a particular assumption, we can only present descriptive data and then assess how they align with typical assumptions.

Figures 1 and 2 present the 12-categories for national and individual consequences and the percent of respondents that mentioned each category in their response to these two questions. Given multiple labels can be applied to one response, the percentages exceed 100%.

Again, we expect that rational and well-informed voters will anticipate that high debts today leads to either higher taxes or spending cuts tomorrow. The data presents a mixed

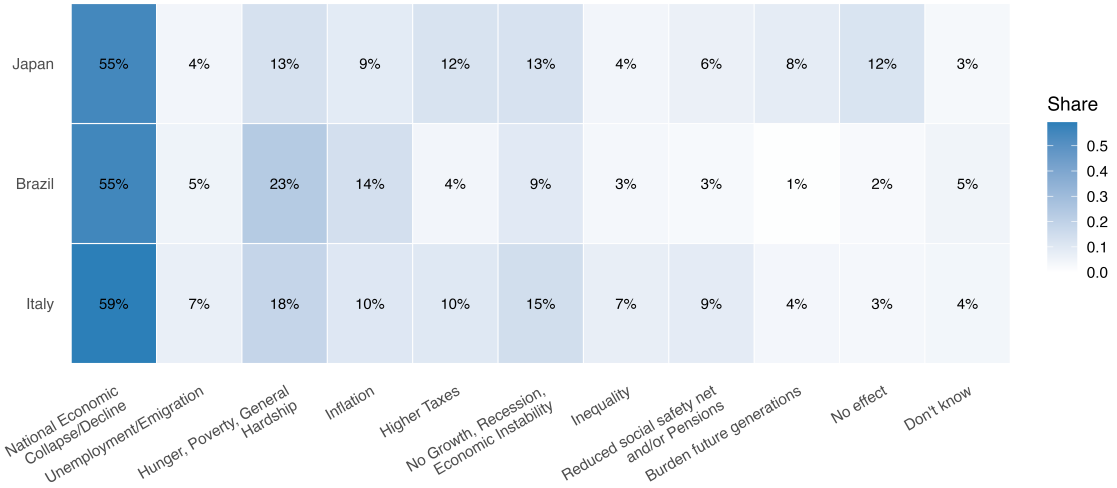

Figure 1: **National Consequences of Public Debt Increase:** Here we show the distribution of responses to the question "What do you think would happen to [country's] economy if the national debt continues to increase?" by country sample. Respondents can mention multiple consequences. As such, the categories do not sum to 100%.

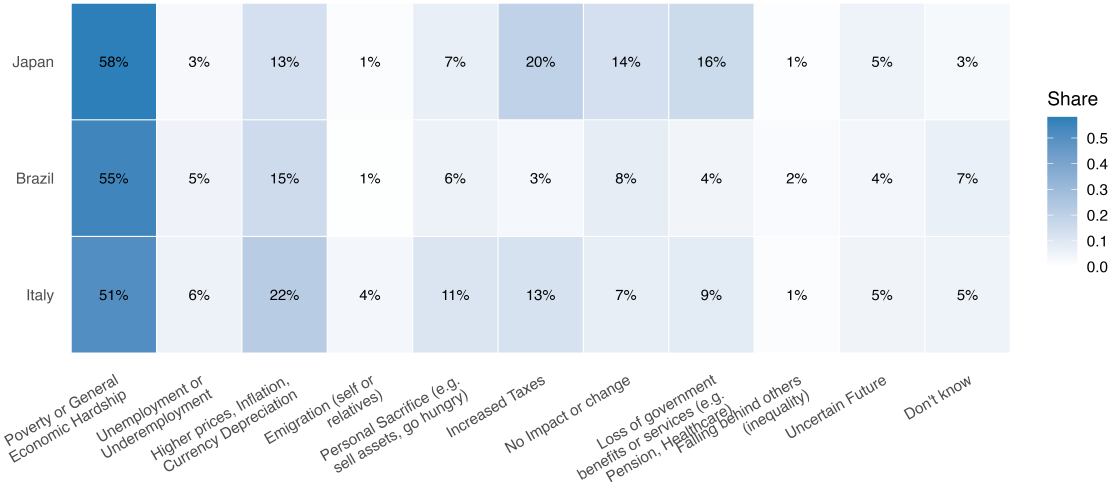

Figure 2: **Individual Consequences of Public Debt Increase:** Here we show the distribution of responses to the question "What do you think would happen to your economic situation if the national debt continues to increase?" by country sample. Respondents can mention multiple consequences. As such, the categories do not sum to 100%.

picture of citizens' responses to these questions. Economic orthodoxy would suggest that rising public debt can have numerous consequences. They would include higher taxes, in-

flation, reduction in public services, and higher costs of finance (Blanchard 2021). In the survey responses, few citizens point to specific policy outcomes that follow from high public debt when asked about the national economy or individual circumstances. As such, it is hard to conclude that a large percentage of the public naturally equate rising public debt with future taxation that would lend evidence to Ricardian-Equivalence like assumptions. On the other end of the spectrum, very few respondents claim ignorance on the matter by indicating they cannot name or do not know the consequences of rising debt burdens. Further, only in Japan do we see a meaningful amount of respondents claiming that debt burdens will have no impact on the economy or their situation.

We do see that a majority of responses in each country point to general negative economic consequences at both the national and individual level. At the national level, 55–59% of respondents point to national economic collapse or decline. Examples of phrases used in this category from our first labeling of the data indicate phrases such as: "economic collapse, economic crisis, recession, economic disaster, economic recession, loss of credibility, state failure" or "economic instability, citizen exhaustion, citizens suffer, citizens in distress, government instability, need for measures, riots, social instability." It is of note that, on average, the responses tend to lean more toward the dramatic rather than a mild recession.

Our multi-label classification does not allow us to infer the general distribution of negative expectations at a respondent level. As such, we ran an additional analysis, at both levels, to classify each response as identifying positive, negative, or neutral responses to the two questions. In this sentiment analysis we used a non-reasoning model – `Qwen3-235B-A22B-Instruct-2507` – given the relative simplicity of the task. Figure 3 presents the percentage of negative comments in each sample. We see that in Italy and Brazil, over 80% expect negative consequences for the national economy if debts continue to rise. In Japan, the number is slightly lower (77%). Expectations of individual consequences are somewhat lower (62.8–79%) likely given the heterogeneity in vulnerability to a crisis. However, there is still widespread agreement

that increasing the debt would yield negative personal consequences.

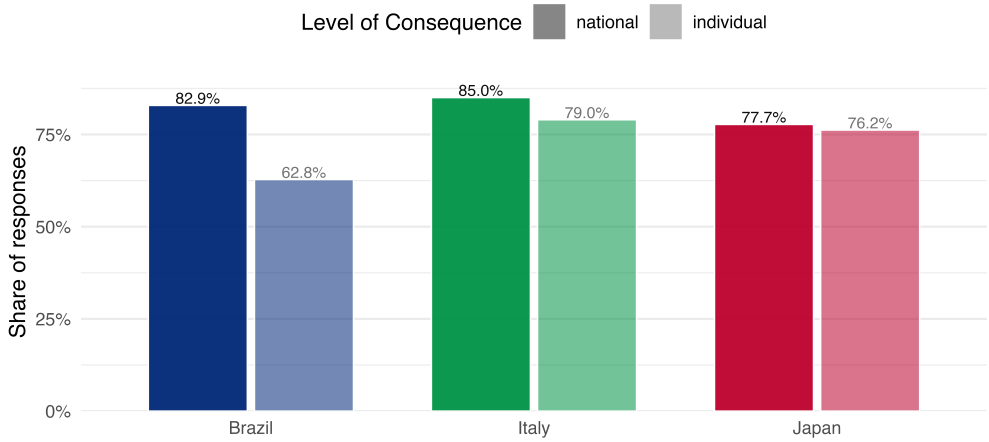

Figure 3: **Percentage of Respondents Expecting Negative Consequences:** Here we present the results of analysis in which we prompted an LLM to identify negative, neutral, or positive expectations for either the respondents or the respondent's country's economy if public debt were to increase further. Here we present the share of respondents that expected negative consequences.

Our analysis thus far reveals that, in these countries where public debt is salient, the public is neither perfectly ignorant nor perfectly informed about its consequences. Their understanding of how debt translates to policy outcomes is imperfect. However, we must also understand that such reasoning also requires an element of forecasting. Citizens are not expected to know which policies governments will take to address high debt or if they will wait until a crisis. This analysis should update the weight we place on models that assume fiscal illusion or that the public does not internalize the consequences of public debt. Our results here suggest that few respondents see increasing debt in an ambivalent manner. In fact, many anticipate dire consequences for themselves and their countries.

Evidence of Cleavages in Policy Solutions

The second aim of our study is to find evidence, in citizens' narratives, of cleavages over debt reduction policy in theoretical models. Recall that the two most prominent cleavages in the

literature center around age and left-right orientation. The latter is a proxy for conflict over the distributional burden of debt reduction. Toward this end, we run a series of independent linear probability models that regress the mention of a category in response to the question on policy solutions on age and left-right orientation.

Figure 4 presents the distribution of these responses for each country. We see considerable variation across the potential policies. Yet, some only garner a few mentions. Notably, we find that reducing government spending garners the most support – ranging from 41% in Brazil to a 60% majority of respondents in Japan. Across the categories, we only see double-digit support for the promotion of growth across all three countries, with Italian respondents mentioning pro-growth policies more often.

We now proceed to examine if our respondent's age and political orientation correlates with preferred policies for reducing their country's national debt. Given the skewed nature of the responses and to simplify the presentation, we display the findings for five categories here that most closely correspond with distributional conflict and have sufficient responses to analyze in a systematic manner. In addition to including age and left-right orientation, we include potentially confounding covariates – education, income, and gender. We include region covariates in Italy (North, Central, South) and Japan (Tokyo), to address confounding from regional differences not captured by income and education, while also serving an explanatory role in identifying other distributional cleavages.

Figure 5 presents the standardized coefficients from 15 linear probability models and the corresponding 95% confidence intervals. As such, the coefficients can be interpreted as a one standard deviation change in the independent variable corresponding to a percentage change in the dependent variable (scaled 0–1). The probability a category is mentioned is also determined by the length of the response and the total number of categories mentioned. We include the token length of each response and the number of total policies mentioned on the right-hand side of the equation although they are not shown here.

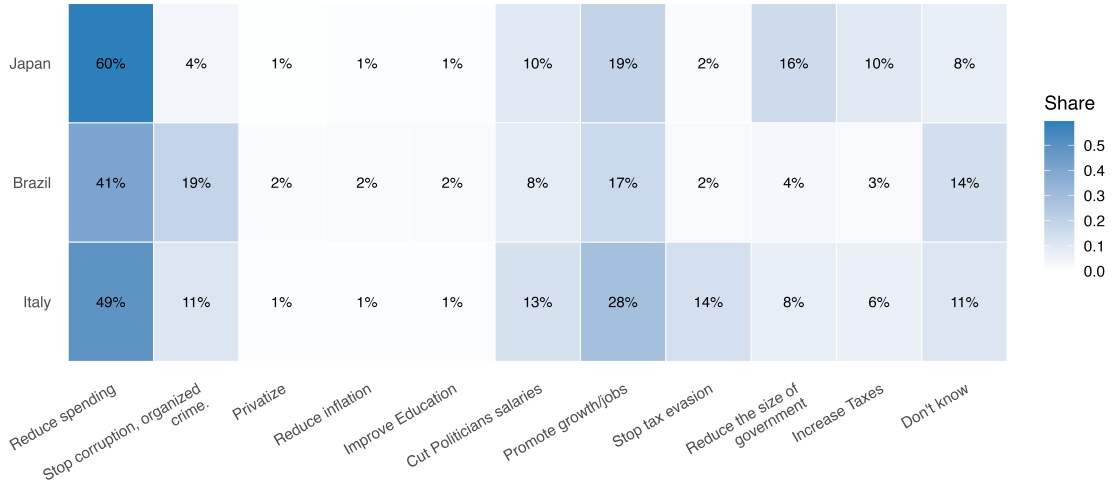

Figure 4: **Mentioned Policies to Reduce Public Debt:** Here we show the distribution of responses to the question "What policies should the government adopt to reduce [Italy, Brazil, Japan]'s public debt?" by country sample. Respondents can mention multiple policies. As such, the categories do not sum to 100%.

At the bottom of each plot are age and left-right coefficients. First, we see that left-right orientation appears to play a role only in Italy. Even so, it leads to slight differences in support for taxes. The scale from 0 (left) to 10 (right) shows that respondents who identify as more right are less likely to mention tax increases. Those on the left are more likely to mention efforts to reduce tax evasion. However, the differences are substantively small, ranging from 1.5% to 3% per standard deviation change. Surprisingly, we find no statistical difference in the mention of reducing government waste and cutting social spending. In Japan, we see only a significant difference in mentions to stimulate growth along the left-right dimension. In Brazil, we find that left-right orientation has no significant effect on the policies mentioned. Beyond left-right orientation, if we focus purely on income and the expected divisions between those dependent on government income and those that are creditors, we still see little evidence for cleavages across the tax and spending categories.

We observe more differences across age for all three countries.[11] However, the substantive

[11]Note that our analysis does not allow us to separate age and cohort effects since we only ask respondents at one point in time. Regardless, we refer to the correlation of age rather than age/cohort.

size of effects is still small and the effects are inconsistent across countries and inconsistent with theoretical expectations.

We would expect that older respondents prefer to preserve the size of government and would not want to reduce spending. However, we see that age has a positive correlation with shrinking the government in Japan and Brazil. In Italy, age has a positive relationship with support for a reduction in government spending. Aspide et al. (2023a) note that the relationship between age and support for debt reduction is non-linear. The college-aged would prefer to avoid reductions in spending until they earn higher incomes, and the elderly prefer to shift the burden to the next generation. However, further analysis including a quadratic function of age does not generate results in line with these expectations.

Our multi-step analysis thus far may obscure similarities in our respondents' underlying language. To address these concerns, we also analyze the clustering of the text embeddings of the entire raw responses to the policy question.[12] Figures 6 and 7 present two-dimensional UMAP projections of the full embedding space, colored by left-right self-placement and age tertiles, respectively. Across all countries, there is insufficient separation by left-right positions or age to indicate that these demographic divisions correspond to distinct ways of articulating debt policy solutions. In the Electronic supplementary material, we present a formal PERMANOVA test to examine separation in the multidimensional space without reduction to two-dimensions. Here we find that our variable groupings are statistically different from noise. However, they explain a very small amount of total variance in our responses (R^2 between 0.21% and 0.52%).[13] As such, it confirms that age and left-right political orientation have a trivial correlation with how people talk about policy solutions to high public debt.

[12]We embed each full-text response with the open-weights multilingual model `intfloat/multilingual-e5-large-instruct`. In the Electronic supplementary material, we replicate the analysis with OpenAI's proprietary `text-embedding-3-large`. The substantive conclusions are unchanged.

[13]We also find no substantial separation across income tertiles.

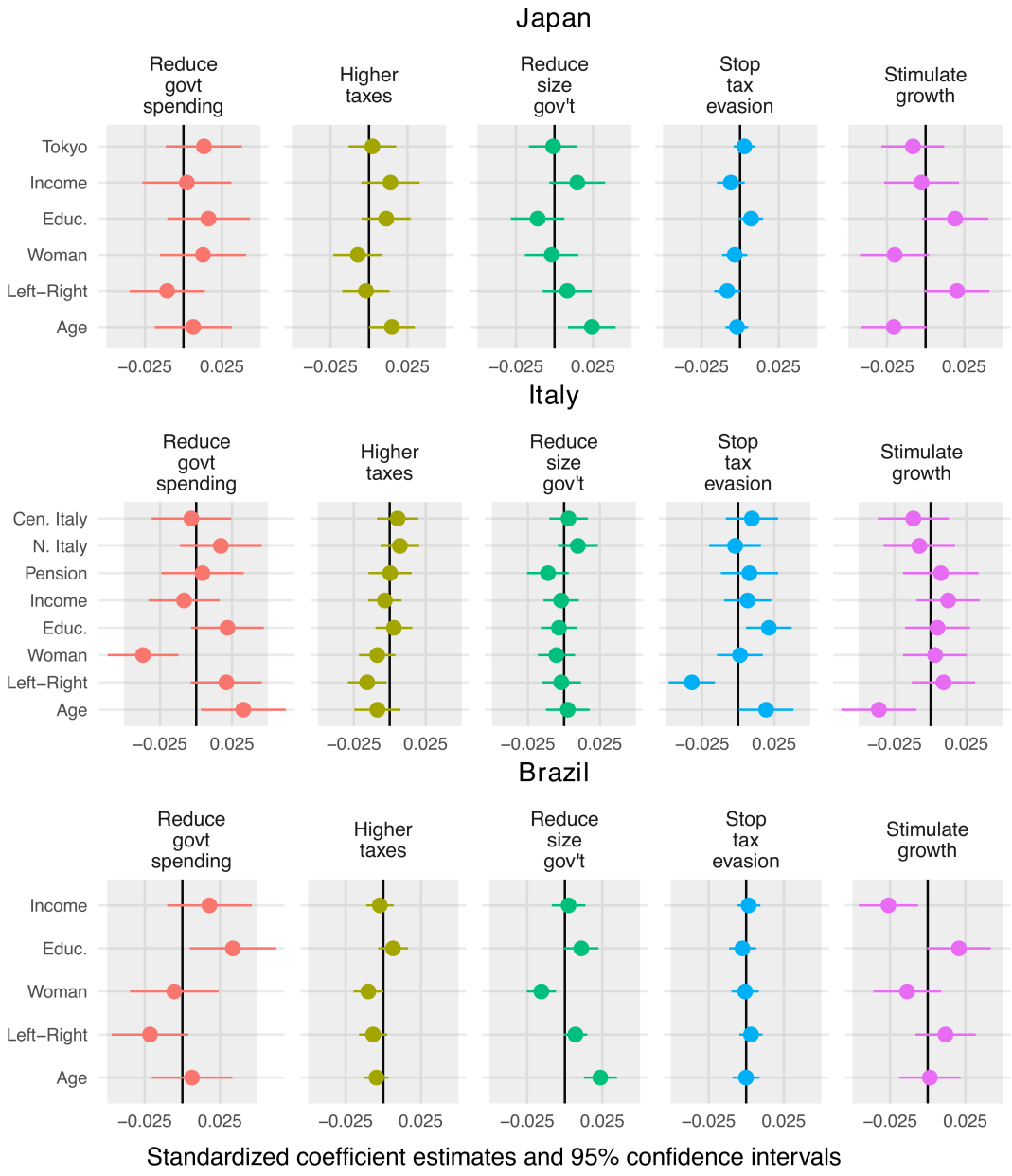

Figure 5: **Correlates of Mentioned Debt Reduction Policies:** Each coefficient plot presents the coefficient from separate linear probability models estimating whether a respondent mentioned the labeled policy in their response to an open-ended question for each country. Each model also includes the 'number of tokens' and the 'number of categories mentioned' on the right-hand side of the equation but they are not shown here. The dots indicate the standardized coefficient, and the bars indicate the 95% confidence intervals around the estimate.

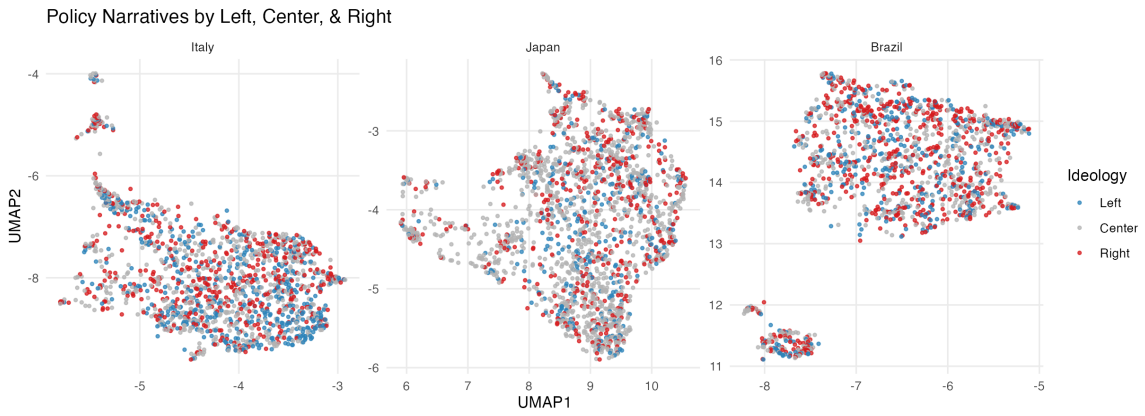

Figure 6: UMAP projection of open-ended policy responses, identified by political orientation. Embeddings are produced with the open-weights multilingual model `intfloat/multilingual-e5-large-instruct`. Points represent individual responses embedded in semantic space. Each point represents a single response. Axes correspond to the first and second dimensions of the reduced embedding space.

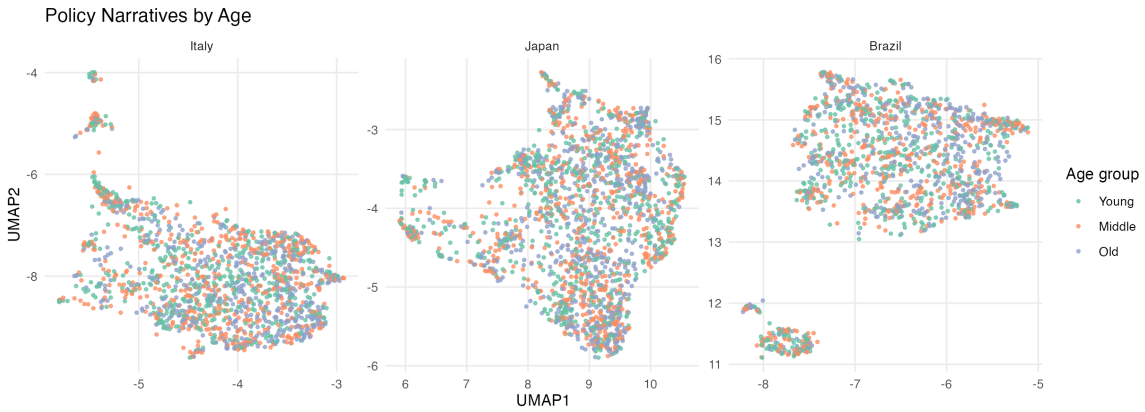

Figure 7: UMAP projection of open-ended policy responses, identified by tertile age groups. Embeddings are produced with the open-weights multilingual model `intfloat/multilingual-e5-large-instruct`. Points represent individual responses embedded in semantic space. Each point represents a single response embedded in semantic space; axes correspond to the first and second dimensions of the reduced embedding space.

Taken together, these findings suggest that citizens across ideological and generational divides articulate debt reduction solutions using remarkably similar language, contrary to expectations of sharp rhetorical cleavages along partisan or demographic lines.

Additional Insights

Beyond finding evidence consistent with theoretical assumptions, open-ended questions are helpful for exploratory research as they bring to light issues that researchers may have ignored. Our study reveals several things that may inform future research on the public's relationship with sovereign debt.

Academic research tends to focus on the austerity aspects of debt consolidation (taxes and spending cuts). Yet, the public often holds narratives that less costly policies will reduce public debt. Many voters see pro-growth policies as preferred policies, and others point to vague conceptions of "waste" in public spending. Both can have a reasonable impact on the debt/GDP. However, reducing inefficiencies could have various meanings, and respondents might have different policies in mind, or no policies in mind, when mentioning reducing government waste. Other policies like reducing corruption or fighting tax evasion are rarely considered by political economists but play heavily in domestic narratives of debt burdens and are also legitimate avenues of debt reduction. Yet, corruption plays a small role in theoretical work on public debt despite a strong correlation in observational data (Cooray et al. 2017).

Other policies mentioned, like reducing politicians' salaries, are clearly not sufficient to impact government debt. Along with the prominence of the "don't know" category, it exposes the limits of some voters' ability to identify policies that can sufficiently reduce debt. Our findings point to a need for more research to explain this phenomenon.

Conclusion

The public remains central to our understanding of public debt accumulation and consolidation. This is why most models of public debt accumulation and consolidation rest on assumptions of citizen preferences. In this paper, we join a nascent literature that is at-

tempting to shed light on how the public thinks (or does not think) about public debt (Bremer and Bürgisser 2023; Aspide et al. 2023a,b; Ardanaz et al. 2025b; Barnes and Hicks 2018, 2021; Ardanaz et al. 2024; Aspide and DiGiuseppe 2026) and a broader literature that seeks to understand how the public engages with markets and economic policy in economics and political science (Stantcheva 2021; Andre et al. 2026; Curtis 2014; Aspide et al. 2023a,b; Nelson and Steinberg 2018). Our contribution is to analyze what the public thinks when given space to provide their own narratives in open-ended responses. We think this is informative, especially when assessing assumptions about voter sophistication. Traditional surveys tend to skew responses due to the fact that they require forced, and often polarized, choices. By looking at open-ended responses, we provide a new evidence on citizen's engagement with public debt that complements work on closed-ended questions.

In this regard, our results are useful in a Bayesian sense, for those who hold on to voter rationality and those who assume the public heavily discounts public debt. Our analysis requires each to give pause and potentially update their assumptions about the public toward a middle ground. The public, on one hand, did in fact anticipate negative economic consequences from rising debt burdens. This lends support to other empirical findings that show the public may be a constraint on myopic politicians (Brender and Drazen 2008; Alesina et al. 2020). On the other hand, the public did not mention specific consequences, indicating a limited understanding of the specific mechanisms by which public debt would impact the national economy and themselves leaving the door open for fiscal illusion.

Our analysis of policy solutions complements what we have learned from closed-ended survey questions. Our analysis shows that, when unprompted, voters can name policies that they would like to enact to reduce debt. However, the divisions over policy do not appear as robust in the absence of menu-cuing. We fail to find strong evidence of generational or left-right conflict in both the preferred policies or how respondents talk about policy solutions. In conjunction with our analysis of citizen consequences, our findings suggest several things

may drive policy responses to high debt. First, politicians may be prospective in trying to protect their constituents from harmful consequences those constituents are unaware of. Second, it may suggest that voters are more amenable to policy solutions that have sizable costs as suggested by Bansak et al. (2021) and Ardanaz et al. (2025b). As we see in our analysis, close to a majority of voters mention spending cuts as a solution to debt. It may also suggest that the key distributional conflicts are non-material or shielded by information deficits (DiGiuseppe and Del Ponte 2026; Ardanaz et al. 2025a). This would imply existing models are misspecified. Unfortunately, we don't have the space to directly address specific models in this article. Here, we only provide new evidence that questions existing practice and points in several directions.

Our analysis provides an important counterweight to the non-testing of assumptions and the use of closed-ended questions to probe citizen engagements and preferences. It is not, however, without limitations. Our analysis captures one snapshot in time within three countries. We hope that the relative consistency of findings across three samples speaks to external validity within the scope of countries that have high debt and are at risk of crisis or default. However, that can't be guaranteed. A second limitation concerns the educational composition of our samples. As discussed in the Research Design section, our quota-based samples leans toward the highly educated. This poses a small threat to our central conclusions. The first finding, of the widespread expectation of negative consequences from rising debt, runs in the opposite direction of what an education-induced bias would predict. To the extent that better-educated respondents are more likely to articulate any consequence at all (rather than respond "don't know"), our over-educated sample should, if anything, understate rates of unspecific or absent reasoning relative to the population. The second finding, of weak generational and left-right cleavages, is similarly insulated from reweighting as we directly control for education in the models. Next, survey research helps us see factors outside of market and political outcomes. Yet, the low stakes involved question

its validity to capture true preferences or mental processes. Further, this lack of effort or insincerity might be heterogeneous. This is a fair criticism and any interpretation must weigh the possibility that respondents are answering in a shallow or insincere manner. Still if respondents were simply drawing on readily available information, we should expect that they draw from elite cues which would be reflected in our analysis of cleavages. We fail to see such sorting. Yet, we also cannot rule out heterogeneity in attention as impacting our findings.

Despite the limitations, survey research and open-ended responses are important because they are the only way to probe thinking on economic policy among a diverse pool of respondents. There are simply no other ways to probe how citizens think about and engage with economic policy at the scale necessary to say something about the public more broadly.

Acknowledgments

We thank Joachim Wehner for comments on an initial draft.

Funding Statement

This research was funded by European Research Council Horizon 2020 Grant #852334.

Statement on the use of AI

Large language models (LLMs) are a central methodological instrument of this study: as described in the Research Design and Electronic supplementary material, we use them to extract, cluster, and classify open-ended survey responses, and to conduct the cross-model validation of that classification. All such uses, including the specific models, prompts, parameters, and validation procedures, are documented in the paper and the accompanying replication materials.

Beyond this instrumental use, generative AI tools (including code assistants and conversational models) were used by the authors to assist with writing and editing code for data processing and analysis, to audit the replication package, and to copyedit prose. The audit of our analysis code in the course of preparing this revision identified an error in how a validation statistic had been computed in an earlier draft, which we have corrected and describe transparently in the Electronic supplementary material. The authors directed all analyses, verified all outputs, and take full responsibility for the content of the manuscript, including any errors.

References

Alesina, Alberto, and Allan Drazen. 1991. Why are stabilizations delayed? *American Economic Review* 81: 1170–1188.

Alesina, Alberto, Carlo Favero, and Francesco Giavazzi. 2019. *Austerity: When it works and when it doesn't*. Princeton: Princeton University Press. https://doi.org/10.1515/9780691185019

Alesina, Alberto, and Guido Tabellini. 1990. A positive theory of fiscal deficits and government debt. *Review of Economic Studies* 57: 403–414. https://doi.org/10.2307/2298021

Alesina, Alberto F., Davide Furceri, Jonathan D. Ostry, Chris Papageorgiou, and Dennis P. Quinn. 2020. Structural reforms and elections: Evidence from a world-wide new dataset. NBER Working Paper 26720. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w26720

Andre, Peter, Ingar Haaland, Christopher Roth, Mirko Wiederholt, and Johannes Wohlfart. 2026. Narratives about the macroeconomy. *Review of Economic Studies*. https://doi.org/10.1093/restud/rdag014

Ardanaz, Martín, Evelyne Hübscher, Philip Keefer, and Thomas Sattler. 2024. Voter responses to fiscal crisis: New evidence on preferences for fiscal adjustment in emerging markets. IDB Working Paper IDB-WP-1545. Washington, DC: Inter-American Development Bank. https://doi.org/10.18235/0012884

Ardanaz, Martín, Evelyne Hübscher, Philip Keefer, and Thomas Sattler. 2025a. Policy misperceptions, information, and the demand for redistributive tax reform: Experimental evidence from Latin America. *Fiscal Studies* 46: 373–397. https://doi.org/10.1111/1475-5890.70001

Ardanaz, Martín, Evelyne Hübscher, Philip Keefer, and Thomas Sattler. 2025b. Voters' preferences over the composition of fiscal adjustment. Unpublished manuscript.

Asatryan, Zareh, Thushyanthan Baskaran, Theocharis Grigoriadis, and Friedrich Heinemann. 2017. Direct democracy and local public finances under cooperative federalism. *Scandinavian Journal of Economics* 119: 801–820. https://doi.org/10.1111/sjoe.12169

Aspide, Alessia, Kathleen J. Brown, Matthew DiGiuseppe, and Alexander Slaski. 2023a. Age and support for public debt reduction. *European Journal of Political Research* 62: 1191–1211. https://doi.org/10.1111/1475-6765.12577

Aspide, Alessia, Kathleen J. Brown, Matthew DiGiuseppe, and Alexander Slaski. 2023b. Culture & European attitudes on public debt. *New Political Economy* 28: 509–525. https://doi.org/10.1080/13563467.2022.2143490

Aspide, Alessia, and Matthew DiGiuseppe. 2026. The mass politics of public debt, immigration, and austerity. *Journal of European Public Policy* 33: 2010–2038. https://doi.org/10.1080/13501763.2025.2510523

Bansak, Kirk, Michael M. Bechtel, and Yotam Margalit. 2021. Why austerity? The mass politics of a contested policy. *American Political Science Review* 115: 486–505. https://doi.org/10.1017/S0003055420001136

Barnes, Lucy, and Timothy Hicks. 2018. Making austerity popular: The media and mass attitudes toward fiscal policy. *American Journal of Political Science* 62: 340–354. https://doi.org/10.1111/ajps.12346

Barnes, Lucy, and Timothy Hicks. 2021. All Keynesians now? Public support for countercyclical government borrowing. *Political Science Research and Methods* 9: 180–188. https://doi.org/10.1017/psrm.2019.48

Barnes, Lucy, and Timothy Hicks. 2022. Are policy analogies persuasive? The household budget analogy and public support for austerity. *British Journal of Political Science* 52: 1296–1314. https://doi.org/10.1017/S0007123421000119

Barrie, Christopher, Alexis Palmer, and Arthur Spirling. 2024. Replication for language models: Problems, principles, and best practice for political science. Working paper. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf. Accessed 3 September 2026.

Barro, Robert J. 1974. Are government bonds net wealth? *Journal of Political Economy* 82: 1095–1117. https://doi.org/10.1086/260266

Barro, Robert J. 1979. On the determination of the public debt. *Journal of Political Economy* 87: 940–971. https://doi.org/10.1086/260807

Battaglini, Marco, and Stephen Coate. 2008. A dynamic theory of public spending, taxation, and debt. *American Economic Review* 98: 201–236. https://doi.org/10.1257/aer.98.1.201

Bierbrauer, Felix J., Pierre C. Boyer, and Andreas Peichl. 2021. Politically feasible reforms of nonlinear tax systems. *American Economic Review* 111: 153–191. https://doi.org/10.1257/aer.20190021

Bishop, George F., Alfred J. Tuchfarber, and Robert W. Oldendick. 1986. Opinions on fictitious issues: The pressure to answer survey questions. *Public Opinion Quarterly* 50: 240–250. https://doi.org/10.1086/268978

Bisin, Alberto, Alessandro Lizzeri, and Leeat Yariv. 2015. Government policy with time inconsistent voters. *American Economic Review* 105: 1711–1737. https://doi.org/10.1257/aer.20131306

Blanchard, Olivier. 2021. *Macroeconomics*, 8th ed. Harlow: Pearson. Global edition.

Blume, Lorenz, Jens Müller, and Stefan Voigt. 2009. The economic effects of direct democracy—a first global assessment. *Public Choice* 140: 431–461. https://doi.org/10.1007/s11127-009-9429-8

Bremer, Björn, and Reto Bürgisser. 2023. Do citizens care about government debt? Evidence from survey experiments on budgetary priorities. *European Journal of Political Research* 62: 239–263. https://doi.org/10.1111/1475-6765.12505

Brender, Adi, and Allan Drazen. 2008. How do budget deficits and economic growth affect reelection prospects? Evidence from a large panel of countries. *American Economic Review* 98: 2203–2220. https://doi.org/10.1257/aer.98.5.2203

Buchanan, James M. 1967. *Public finance in democratic process: Fiscal institutions and individual choice*. Chapel Hill: University of North Carolina Press.

Buchanan, James M., and Richard E. Wagner. 1977. *Democracy in deficit: The political legacy of Lord Keynes*. New York: Academic Press.

Congleton, Roger D. 2001. Rational ignorance, rational voter expectations, and public policy: A discrete informational foundation for fiscal illusion. *Public Choice* 107: 35–64. https://doi.org/10.1023/A:1010337412291

Connor Desai, Saoirse, and Stian Reimers. 2019. Comparing the use of open and closed questions for Web-based measures of the continued-influence effect. *Behavior Research Methods* 51: 1426–1440. https://doi.org/10.3758/s13428-018-1066-z

Converse, Philip E. 2006. The nature of belief systems in mass publics (1964). *Critical Review* 18: 1–74. https://doi.org/10.1080/08913810608443650

Cooray, Arusha, Ratbek Dzhumashev, and Friedrich Schneider. 2017. How does corruption affect public debt? An empirical analysis. *World Development* 90: 115–127. https://doi.org/10.1016/j.worlddev.2016.08.020

Cukierman, Alex, and Allan H. Meltzer. 1989. A political theory of government debt and deficits in a neo-Ricardian framework. *American Economic Review* 79: 713–732.

Curtis, K. Amber. 2014. In times of crisis: The conditions of pocketbook effects. *International Interactions* 40: 402–430. https://doi.org/10.1080/03050629.2014.902816

Curtis, K. Amber, Joseph Jupille, and David Leblang. 2014. Iceland on the rocks: The mass political economy of sovereign debt resettlement. *International Organization* 68: 721–740. https://doi.org/10.1017/S0020818314000034

DiGiuseppe, Matthew, and Alessandro Del Ponte. 2026. Moral judgment links private and public debt attitudes. Unpublished manuscript.

Dollery, Brian E., and Andrew C. Worthington. 1996. The empirical analysis of fiscal illusion. *Journal of Economic Surveys* 10: 261–297. https://doi.org/10.1111/j.1467-6419.1996.tb00014.x

Dovis, Alessandro, Mikhail Golosov, and Ali Shourideh. 2016. Political economy of sovereign debt: A theory of cycles of populism and austerity. NBER Working Paper 21948. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w21948

Downs, Anthony. 1957. *An economic theory of democracy.* New York: Harper and Brothers.

Feld, Lars P., and Gebhard Kirchgässner. 2001. Does direct democracy reduce public debt? Evidence from Swiss municipalities. *Public Choice* 109: 347–370. https://doi.org/10.1023/A:1013077121942

Ferrario, Beatrice, and Stefanie Stantcheva. 2022. Eliciting people's first-order concerns: Text analysis of open-ended survey questions. *AEA Papers and Proceedings* 112: 163–169. https://doi.org/10.1257/pandp.20221071

Funk, Patricia, and Christina Gathmann. 2011. Does direct democracy reduce the size of government? New evidence from historical data, 1890–2000. *Economic Journal* 121: 1252–1280. https://doi.org/10.1111/j.1468-0297.2011.02451.x

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* 120: e2305016120. https://doi.org/10.1073/pnas.2305016120

Graham, Matthew H. 2021. "We don't know" means "they're not sure". *Public Opinion Quarterly* 85: 571–593. https://doi.org/10.1093/poq/nfab028

Grootendorst, Maarten. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794

Haaland, Ingar, Christopher Roth, Stefanie Stantcheva, and Johannes Wohlfart. 2025. Understanding economic behavior using open-ended survey data. *Journal of Economic Literature* 63: 1244–1280. https://doi.org/10.1257/jel.20251780

Heseltine, Michael, and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics* 11: 20531680241236239. https://doi.org/10.1177/20531680241236239

Janků, Jan, and Jan Libich. 2019. Ignorance isn't bliss: Uninformed voters drive budget cycles. *Journal of Public Economics* 173: 21–43. https://doi.org/10.1016/j.jpubeco.2019.01.003

Kantorowicz, Jaroslaw. 2023. Testing public reaction to constitutional fiscal rules violations. *Constitutional Political Economy* 34: 483–509. https://doi.org/10.1007/s10602-022-09387-5

Kantorowicz, Jaroslaw, and Katarzyna Metelska-Szaniawska. 2025. Debt beliefs and public support for restrictive fiscal rules. *Economics Letters* 247: 112104. https://doi.org/10.1016/j.econlet.2024.112104

Ketchley, Neil, Ferdinand Eibl, and Jeroen Gunning. 2024. Anti-austerity riots in late developing states: Evidence from the 1977 Egyptian Bread Intifada. *Journal of Peace Research* 61: 952–966. https://doi.org/10.1177/00223433231168188

Kraft, Patrick W. 2024. Women also know stuff: Challenging the gender gap in political sophistication. *American Political Science Review* 118: 903–921. https://doi.org/10.1017/S0003055423000539

Le Mens, Gaël, and Aina Gallego. 2025. Positioning political texts with large language models by asking and averaging. *Political Analysis* 33: 274–282. https://doi.org/10.1017/pan.2024.29

Lizzeri, Alessandro. 1999. Budget deficits and redistributive politics. *Review of Economic Studies* 66: 909–928. https://doi.org/10.1111/1467-937X.00113

Lizzeri, Alessandro, and Leeat Yariv. 2017. Collective self-control. *American Economic Journal: Microeconomics* 9: 213–244. https://doi.org/10.1257/mic.20150325

Matsusaka, John G. 2018. Public policy and the initiative and referendum: A survey with some new evidence. *Public Choice* 174: 107–143. https://doi.org/10.1007/s11127-017-0486-0

Mellon, Jonathan, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman. 2024. Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale. *Research & Politics* 11: 20531680241231468. https://doi.org/10.1177/20531680241231468

Müller, Andreas, Kjetil Storesletten, and Fabrizio Zilibotti. 2016. The political color of fiscal responsibility. *Journal of the European Economic Association* 14: 252–302. https://doi.org/10.1111/jeea.12154

Nelson, Stephen C., and David A. Steinberg. 2018. Default positions: What shapes public attitudes about international debt disputes? *International Studies Quarterly* 62: 520–533. https://doi.org/10.1093/isq/sqy020

Ornstein, Joseph T., Elise N. Blasingame, and Jake S. Truscott. 2025. How to train your stochastic parrot: Large language models for political texts. *Political Science Research and Methods* 13: 264–281. https://doi.org/10.1017/psrm.2024.64

Palmer, Alexis, Noah A. Smith, and Arthur Spirling. 2024. Using proprietary language models in academic research requires explicit justification. *Nature Computational Science* 4: 2–3. https://doi.org/10.1038/s43588-023-00585-1

Persson, Torsten, and Lars E. O. Svensson. 1989. Why a stubborn conservative would run a deficit: Policy with time-inconsistent preferences. *Quarterly Journal of Economics* 104: 325–345. https://doi.org/10.2307/2937850

Persson, Torsten, and Guido Tabellini. 2002. Political economics and public finance. In *Handbook of public economics*, ed. Alan J. Auerbach and Martin Feldstein, vol. 3, 1549–1659. Amsterdam: Elsevier. https://doi.org/10.1016/S1573-4420(02)80028-3

Ponticelli, Jacopo, and Hans-Joachim Voth. 2020. Austerity and anarchy: Budget cuts and social unrest in Europe, 1919–2008. *Journal of Comparative Economics* 48: 1–19. https://doi.org/10.1016/j.jce.2019.09.007

Ricciuti, Roberto, and Davide Di Laurea. 2003. An experimental analysis of two departures from Ricardian equivalence. *Economics Bulletin* 8: 1–11. http://www.accessecon.com/pubs/EB/2003/Volume8/EB-03H30001A.pdf

Rogoff, Kenneth, and Anne Sibert. 1988. Elections and macroeconomic policy cycles. *Review of Economic Studies* 55: 1–16. https://doi.org/10.2307/2297526

Roth, Christopher, Sonja Settele, and Johannes Wohlfart. 2022. Beliefs about public debt and the demand for government spending. *Journal of Econometrics* 231: 165–187. https://doi.org/10.1016/j.jeconom.2020.09.011

Rüdig, Wolfgang, and Georgios Karyotis. 2014. Who protests in Greece? Mass opposition to austerity. *British Journal of Political Science* 44: 487–513. https://doi.org/10.1017/S0007123413000112

Shi, Min, and Jakob Svensson. 2006. Political budget cycles: Do they differ across countries and why? *Journal of Public Economics* 90: 1367–1389. https://doi.org/10.1016/j.jpubeco.2005.09.009

Song, Zheng, Kjetil Storesletten, and Fabrizio Zilibotti. 2012. Rotten parents and disciplined children: A politico-economic theory of public expenditure and debt. *Econometrica* 80: 2785–2803. https://doi.org/10.3982/ECTA8910

Stantcheva, Stefanie. 2021. Understanding tax policy: How do people reason? *Quarterly Journal of Economics* 136: 2309–2369. https://doi.org/10.1093/qje/qjab033

Tabellini, Guido. 1991. The politics of intergenerational redistribution. *Journal of Political Economy* 99: 335–357. https://doi.org/10.1086/261753

Tabellini, Guido, and Alberto Alesina. 1990. Voting on the budget deficit. *American Economic Review* 80: 37–49.

Walton, John, and Charles Ragin. 1990. Global and national sources of political protest: Third World responses to the debt crisis. *American Sociological Review* 55: 876–890. https://doi.org/10.2307/2095752

Yared, Pierre. 2010. Politicians, taxes and debt. *Review of Economic Studies* 77: 806–840. https://doi.org/10.1111/j.1467-937X.2009.00584.x

Yared, Pierre. 2019. Rising government debt: Causes and solutions for a decades-old trend. *Journal of Economic Perspectives* 33: 115–140. https://doi.org/10.1257/jep.33.2.115

A Electronic supplementary material

A.1 Sample Comparison: Education, Income, Left-Right

Our sample was collected with quotas for age, region and sex. Here we show how our sample differs by comparing our measures of education, income and left-right orientation to benchmark surveys known for their rigorous probability-based samples.

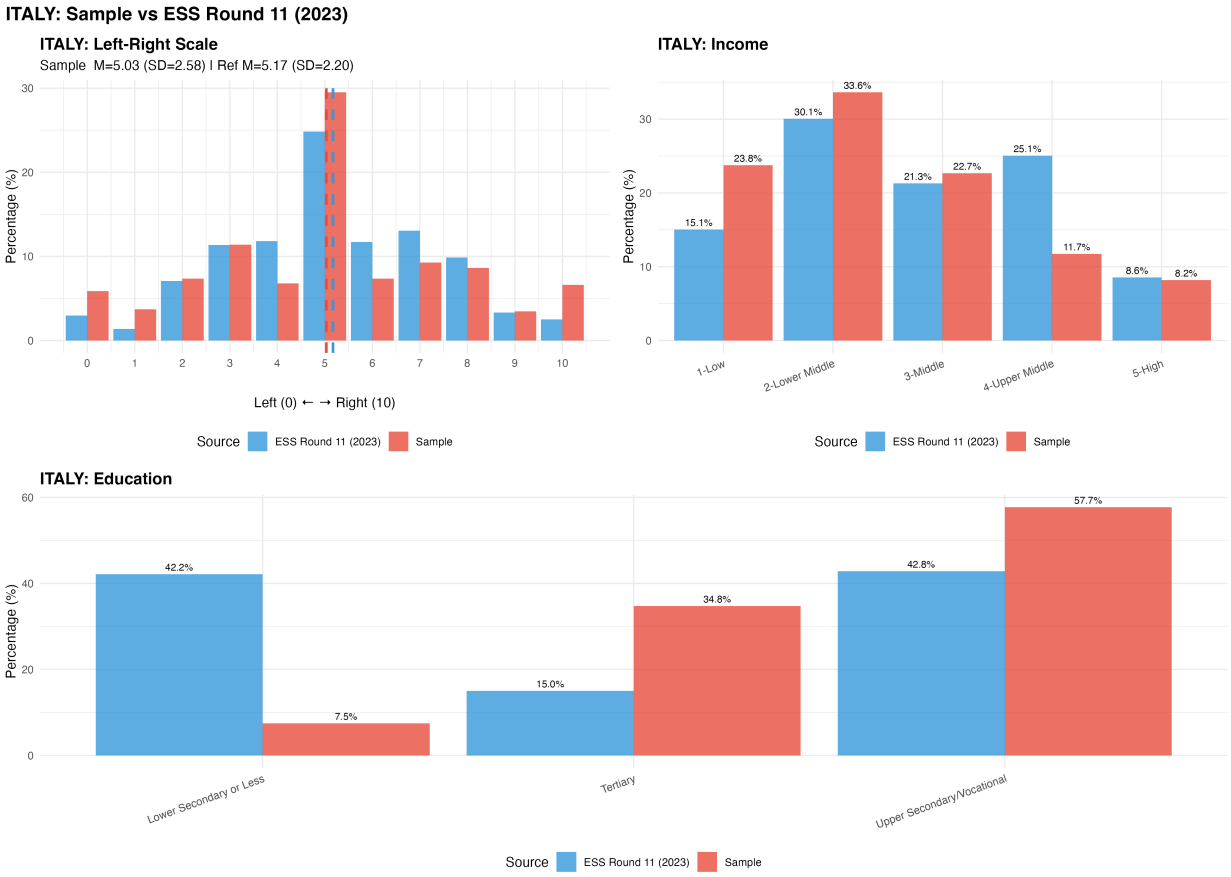

Figure A1: Comparison of Italian Survey Sample with ESS Round 11 (2023). **Top left panel:** Left-right political scale distributions. Both surveys use 0–10 scales where 0=left and 10=right. Dashed vertical lines indicate means (red=sample, blue=ESS). Sample: $n = 1{,}709$, $M = 5.27$, $SD = 2.31$; ESS: $n = 2{,}407$, $M = 5.15$, $SD = 2.18$. **Top right panel:** Income distributions collapsed into five ordinal categories based on monthly household income in euros. Sample income categories were mapped to ESS deciles (hinctnta) to create comparable quintile-like groups. Sample: $n = 1{,}611$; ESS: $n = 2{,}013$. **Bottom panel:** Education distributions harmonized into three categories reflecting comparable levels of educational attainment within the Italian system: (1) Lower Secondary or Less includes no formal education through lower secondary completion; (2) Upper Secondary/Vocational includes 2–3 year vocational qualifications and 5-year upper secondary diplomas (maturità); (3) Tertiary includes university degrees and postgraduate qualifications. ESS education (eisced) follows ISCED 2011 classification. Sample: $n = 1{,}815$; ESS: $n = 2{,}862$.

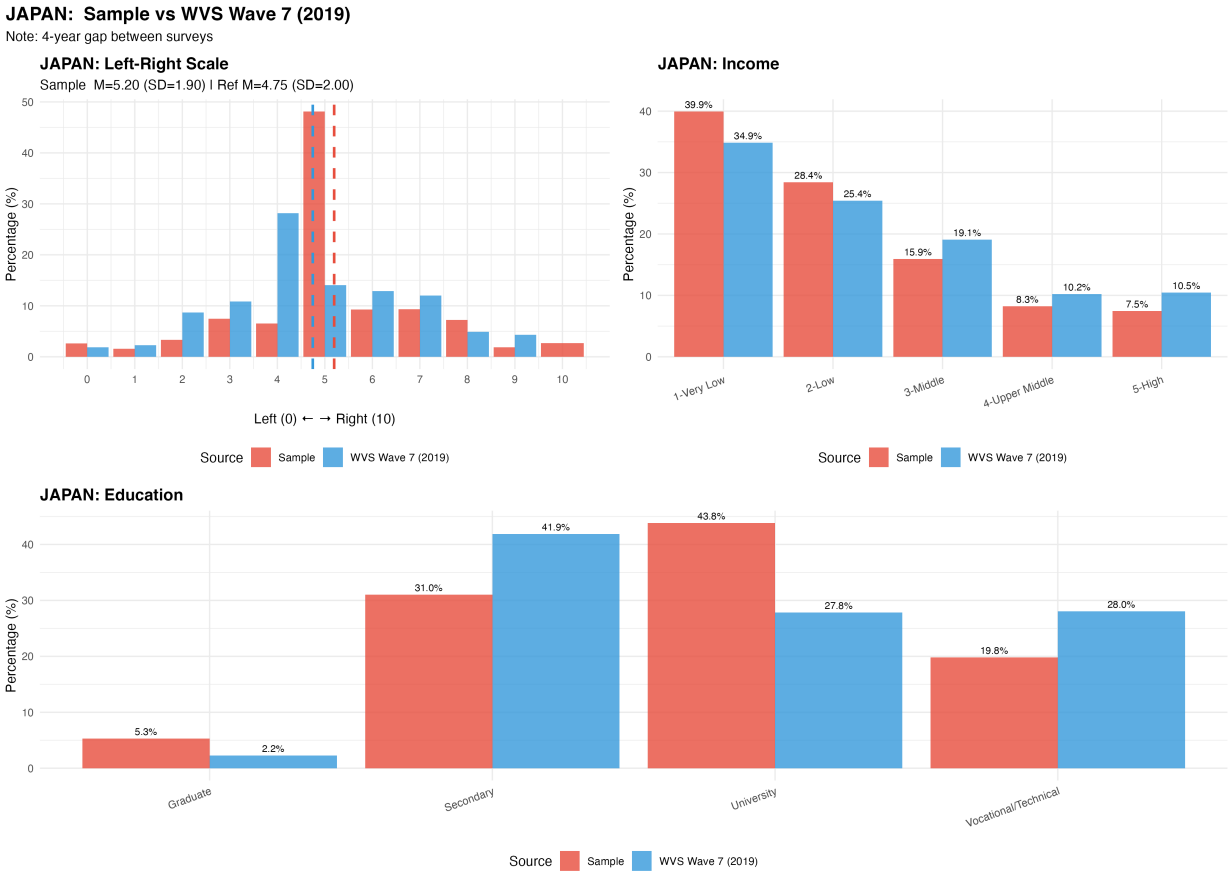

Figure A2: Comparison of Japanese Survey Sample with World Values Survey Wave 7 (2019). **Note:** Four-year temporal gap between surveys (2023 vs. 2019). **Top left panel:** Left-right political scale distributions. Sample uses 0–10 scale; WVS uses 1–10 scale (converted to 0–10 by subtracting 1 for comparability). Dashed vertical lines indicate means (red=sample, blue=WVS). Sample: $n = 2{,}218$, $M = 5.16$, $SD = 1.68$; WVS: $n = 1{,}188$, $M = 5.44$, $SD = 1.89$. **Top right panel:** Income distributions collapsed into five ordinal categories based on annual household income in Japanese yen. Sample categories (ranging from no income/under ¥1M to over ¥10M) were mapped to WVS income scale (Q288: 1–10 subjective income scale) to create comparable groups. Sample: $n = 1{,}839$; WVS: $n = 1{,}012$. **Bottom panel:** Education distributions harmonized into four categories: (1) Secondary includes junior high, high school, and specialized training schools (kōtō senshū gakkō); (2) Vocational/Technical includes professional training schools (senmon gakkō), junior colleges (tanki daigaku), and technical colleges (kōtō senmon gakkō); (3) University (daigaku); (4) Graduate school (daigakuin). WVS education (Q275) follows ISCED 2011 classification, grouped to align with Japanese educational structure. Sample: $n = 2{,}256$; WVS: $n = 1{,}330$.

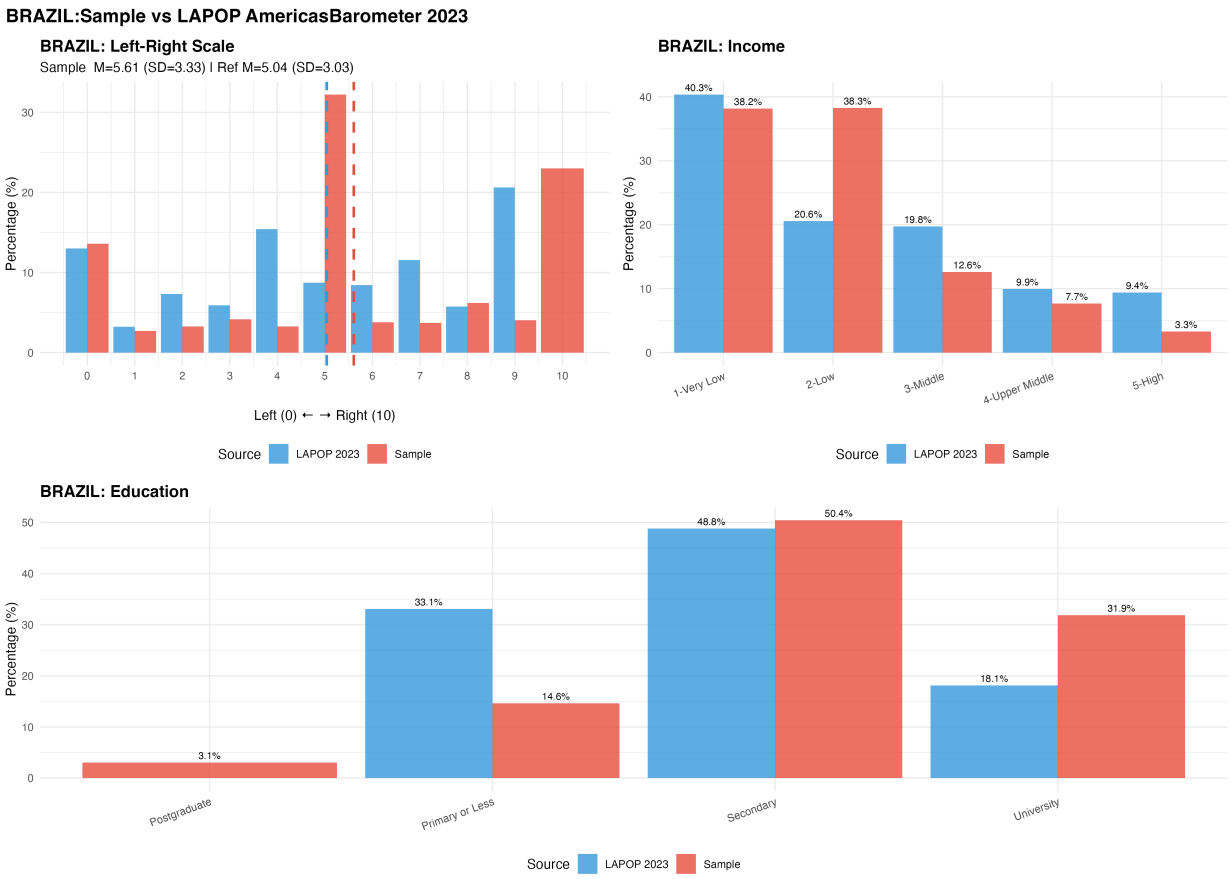

Figure A3: Comparison of Brazilian Survey Sample with LAPOP AmericasBarometer (2023). **Top left panel:** Left-right political scale distributions. Sample uses 0–10 scale; LAPOP uses 1–10 scale (l1n; converted to 0–10 by subtracting 1 for comparability). Dashed vertical lines indicate means (red=sample, blue=LAPOP). Sample: n = 1,773, M = 5.42, SD = 3.28; LAPOP: n = 1,421, M = 5.02, SD = 2.82. **Top right panel:** Income distributions collapsed into five ordinal categories based on monthly household income in Brazilian reais. Sample categories were aligned with LAPOP income ranges (q10inc: 15 categories from R$0–500 to over R$9,600, excluding missing responses coded as 888888, 988888, or 999999). Categories represent roughly comparable positions in the income distribution within Brazil. Sample: n = 1,787; LAPOP: n = 1,448. **Bottom panel:** Education distributions harmonized into three categories (four in sample) based on the Brazilian education system: (1) Primary or Less includes no formal schooling and incomplete/complete Ensino Fundamental; (2) Secondary includes incomplete/-complete Ensino Médio; (3) University includes incomplete/complete Graduação; (4) Postgraduate includes Pós-graduação, Mestrado, and Doutorado. LAPOP education (edre: 0–6 scale) was harmonized accordingly, where 0–2=Primary or Less, 3–4=Secondary, 5–6=University. Sample: n = 2,249; LAPOP: n = 1,521.

A.2 Attrition Analysis

Table A1: Attrition Rates by Country

country	N	Attrition Rate (%)	Number Attrited
Brazil	1773	9.70	172
Italy	1610	6.15	99
Japan	1784	6.78	121

Note:
Attrition defined as missing response to the final open-ended question: the recommended policies question.

Table A2: Demographic Predictors of Attrition (Logistic Regression)

	Italy	Brazil	Japan
Age	0.000	0.001	0.000
	(0.000)	(0.001)	(0.000)
Woman	0.017	−0.022	0.006
	(0.012)	(0.014)	(0.013)
Education	−0.021**	−0.004	−0.003
	(0.007)	(0.011)	(0.003)
Income	0.000	0.001	0.003
	(0.002)	(0.004)	(0.003)
Left-Right	0.001	0.003	0.001
	(0.002)	(0.002)	(0.003)
Num.Obs.	1610	1773	1784
R2	0.007	0.004	0.002

+ p <0.1, * p <0.05, ** p <0.01, *** p <0.001
Robust standard errors in parentheses. Attrition defined as missing response to the policies question (Q4).

A.3 Word count per question

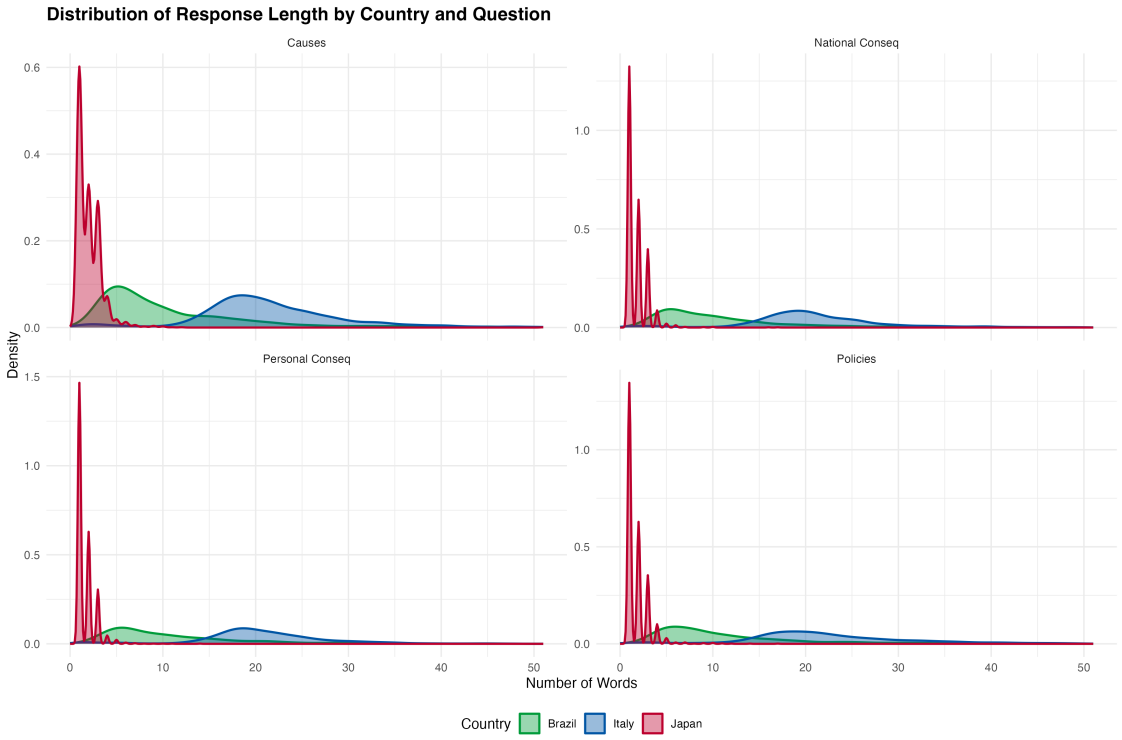

Figure A4: Here we present the word count by country-question for each of the four open-ended responses used in our analyses. We use the original Japanese characters in our analysis. This explains why Japan has a lower "word" count.

Table A3: Word Count Progression Across Questions by Country

Country	N	Q1: Causes	Q2: National Conseq	Q3: Self Conseq	Q4: Policies	Drop (Q1 to Q4)
Brazil	1773	11.06	10.13	11.01	11.06	0.01
Italy	1610	21.75	20.63	20.27	22.61	-0.86
Japan	1784	1.95	1.68	1.54	1.67	0.28

Note:
Mean word count for each question in sequence. Drop shows difference between Q1 (causes) and Q4 (policies).

A.4 LLM Choice

Recent advances in generative Large Language Models (LLMs) have dramatically reduced the costs of using open-ended responses in survey research and analyzing other text data. Prior to the recent advances in LLMs, the annotation of text data was either coded by humans (research assistants or crowd workers) at great expense or limited to the analysis of specific words using structural topic models that are limited in their ability to pull nuanced context from the data.

While LLMs have demonstrated an impressive ability to annotate tasks that exceeds STMs and is on par with humans, the performance is conditional on the model selected. Beyond performance, other considerations, like replicability, play into choosing an LLM for research purposes. Proprietary models, like those offered by OpenAI (GPT-4o, GPT-5, o3), hold no guarantee of being available in the future as newer models are rolled out. Consequently, their use is closer to using human coders in that the exact coding decisions are created by a black box and not easily replicated. As such, it is recommended to use open-weights models (Palmer et al. 2024).

In this paper, we are using data from three different non-English samples. Given the added complexity of non-English languages, we opt for a large but open-weights reasoning model for multi-label classification: `Qwen3-235B-A22B-Thinking-2507`.[14] This model, according to several benchmarks, was the best open-weights model available at the time of research. By choosing a thinking model, we are taking advantage of chain of thought processing which allows the model to spend additional tokens to sort through the multiple potential responses presented in each response. While the additional reasoning tokens come at an additional monetary expense, they are likely to lead to a more accurate classification than non-reasoning models.

One element of our pipeline does depart from a strict open-weights standard: the BERTopic-style category-discovery step described in the Annotation Workflow section uses OpenAI's `text-embedding-3-large` to embed the short multilingual phrases returned by the upstream Qwen3 extraction call. We considered open-weights multilingual embedding alternatives (e.g., `BGE-M3`, `multilingual-E5-large`) but chose OpenAI's model on the basis of its multilingual benchmark performance on short text at the time of analysis.

[14] We accessed the model via the fireworks.ai API.

Human Validation

We validate the LLM annotation against human coders in two ways. Both use F1, the harmonic mean of precision and recall computed over the categories a coder assigns to a response, aggregated across labels and excluding the residual "other" category. $\text{F1} = 2\,(\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$, with $\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$ and $\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$. F1 is informative in this context because this is a multi-label task: respondents may mention more than one policy or consequence in the same open-ended answer. Therefore, our validation exercise should be read as evidence of convergence across trained coders and independent model annotators, rather than as a claim that any single coder or model provides error-free ground truth.

Expert coding by the author team. Four members of the author team independently coded random samples of responses against the same category schemes used in the manuscript, in the responses' original languages with English translations provided.[15]

We coded the policies question and the individual-consequences question for all three countries. Table A4 reports per-country agreement between these human codings and the published `Qwen3-Thinking-2507` pipeline. Agreement is consistent across countries and questions, with micro-F1 between 0.60 and 0.73 and most cells near 0.70. These values are slightly lower than the cross-model agreement reported below, but broadly in the same empirical range. They are also consistent with the level of ambiguity observed among human coders on overlapping items. Given the multi-label structure of the task and the possibility of overlap across categories, we interpret these results as evidence that the LLM pipeline is behaving comparably to trained human coders, especially for categories with clear substantive anchors such as tax cuts, emigration, inflation, or anti-corruption.

Table A4: Agreement between human (author) coding and the published `Qwen3-Thinking-2507` pipeline, by country and question. Micro F1, excluding the residual "Other" category. n is the number of coded response-items.

Question	Country	n	Precision	Recall	Micro F1
Policies	Italy (prior expert)	100	0.79	0.69	**0.73**
Policies	Brazil	123	0.78	0.62	**0.69**
Policies	Japan	117	0.79	0.61	**0.69**
Individual consequences	Italy	77	0.72	0.70	**0.71**
Individual consequences	Brazil	70	0.67	0.54	**0.60**
Individual consequences	Japan	73	0.67	0.66	**0.66**

Inter-coder agreement. Because coders drew random subsets, a subset of consequence responses ($n = 44$) was coded by two or more authors, allowing us to benchmark human–LLM agreement against human–human agreement on the same kind of task. On these

[15] With the exception of the Italian Policies question, which was coded in the original Italian by a fluent Italian speaker (Aspide).

co-coded items, the human coders agreed with one another at a mean pairwise micro-F1 of 0.61 (excluding the residual "other" category), and selected an identical label set on 48% of items. Human–LLM agreement on the consequences question (F1 $\approx$ 0.66, computed the same way) is therefore at least as high as the agreement among the trained human coders themselves. Although this overlap sample is small, the comparison is informative: trained human coders also disagree on a considerable share of responses. The LLM does not behave as an obvious outlier relative to trained human coders; rather, many disagreements appear to arise from the same ambiguous responses on which human coders also diverge.

Cross-Model Validation: Qwen3-Thinking vs. GPT-5.5 and Claude Opus 4.6

To complement the human-coder validation described above, we conducted a cross-model agreement exercise between the Qwen model used in the analysis and two frontier reasoning models from different providers. Specifically, we re-classified a stratified random sample of 300 policy responses per country (n=900 total; `set.seed(42)`) using OpenAI's `gpt-5.5` and Anthropic's `claude-opus-4-6`, and compared the resulting labels to those produced by `Qwen3-235B-A22B-Thinking-2507`.

Table A5 reports the exact inference parameters and access mode used for each model. All three were given the published system and user prompts verbatim, with the response text inserted into the `STATEMENT:` field. We did not enable any model-specific features (no tool use, no JSON mode, no extended-thinking budget tuning); the prompt itself constrains the output format to a single JSON object.

Table A6 reports per-category precision, recall, and F1 against the published `Qwen3`. Pooled micro F1 (excluding the residual *Other* category) is 0.81 for GPT-5.5 and 0.75 for Claude Opus; macro F1 excluding *Other* is 0.80 and 0.73 respectively. Mean per-response Jaccard similarity is 0.68 (GPT-5.5) and 0.63 (Opus); exact-match label-set agreement is 56% and 46%. Both models converge closely on the substantive categories (per-category F1 typically 0.62–0.88 across both) and diverge primarily on the residual *Other* category (F1 = 0.28 for GPT-5.5, 0.22 for Opus), where the two cross-check models use the catch-all far more liberally than `Qwen3` (GPT-5.5: 281 vs. 47 mentions; Opus: 225 vs. 47). Opus also over-uses *Cut politicians' salaries* (203 vs. 105 mentions), which depresses precision on that category but recall remains perfect. We read the broad three-way convergence as evidence that the published Qwen3 classification is not an artefact of one model pipeline. Two frontier reasoning models from different providers, prompted with the same instructions and category list, recover substantively similar labels for the same responses. This is best interpreted as evidence of convergent validity rather than as an external ground truth or gold standard accuracy estimate, since the published Qwen3 labels serve as the reference in these comparisons.

To locate the disagreement, Table A7 reports F1 by category for both cross-check models alongside each category's share of all cross-model disagreement (the share of total false-positive-plus-false-negative cells attributable to that category). Disagreement is highly con-

Table A5: Inference settings used for each model in the cross-model validation.

	Paper pipeline (reference)	**GPT-5.5 (cross-check)**	**Claude Opus 4.6 (cross-check)**
Model ID	`qwen3-235b-a22b-thinking-2507`	`gpt-5.5`	`claude-opus-4-6`
Provider / access	Fireworks.ai Batch Inference	OpenAI `/v1/chat/completions` (synchronous)	Anthropic `/v1/messages/batches` (asynchronous, 50% discount)
Temperature	0	default (omitted – reasoning model)	default (omitted)
Top-p	1	default (omitted)	default (omitted)
Max output tokens	32,768	8,000 (`max_completion_tokens`)	1,024 (`max_tokens`)
Reasoning mode	native thinking model	native reasoning model (reasoning tokens billed as completion)	native; no extended-thinking budget specified
Structured output	no (prompt-constrained)	no (prompt-constrained)	no (prompt-constrained)
Mean prompt tokens (observed)	n/a (cached output)	506	320
Mean output tokens (observed)	n/a (cached output)	105	25
Cost basis	published main pipeline; cached outputs reused	\$1.25/M input, \$7.50/M output	\$5.00/M input, \$25.00/M output (50% batch discount applied)

Table A6: Cross-model agreement between the paper's published `Qwen3-Thinking-2507` pipeline and two independent frontier reasoning models (OpenAI `gpt-5.5`; Anthropic `claude-opus-4-6`) on the same 900 randomly sampled policy responses (300 per country, `set.seed(42)`). All three models received the published Set-C system and user prompts verbatim. Per-category precision and recall treat `Qwen3` as the reference. The pooled micro and macro F1 in the table footer exclude the residual *Other* category.

			GPT-5.5			Claude Opus 4.6		
#	Category	N_{Qwen}	N	P	R/F1	N	P	R/F1
1	Reduce spending	449	333	0.93	0.69/**0.79**	358	0.89	0.71/**0.79**
2	Stop corruption	92	97	0.76	0.80/**0.78**	92	0.76	0.76/**0.76**
3	Privatize	16	21	0.71	0.94/**0.81**	20	0.65	0.81/**0.72**
4	Reduce inflation	10	18	0.56	1.00/**0.71**	16	0.50	0.80/**0.62**
5	Improve education	13	21	0.62	1.00/**0.76**	22	0.59	1.00/**0.74**
6	Cut politicians' salaries	105	156	0.67	0.99/**0.80**	203	0.52	1.00/**0.68**
7	Promote growth / jobs	210	189	0.87	0.79/**0.83**	172	0.85	0.70/**0.77**
8	Stop tax evasion	59	52	0.85	0.75/**0.79**	62	0.76	0.80/**0.78**
9	Reduce the size of government	92	108	0.77	0.90/**0.83**	114	0.62	0.77/**0.69**
10	Increase taxes	47	61	0.74	0.96/**0.83**	86	0.51	0.94/**0.66**
11	Other	47	281	0.16	0.98/**0.28**	225	0.13	0.64/**0.22**
12	Don't know	92	88	0.90	0.86/**0.88**	128	0.69	0.96/**0.80**
Micro F1 (pooled, excl. *Other*)				0.82/0.79	**0.81**		0.73/0.78	**0.75**
Macro F1 (excl. *Other*)					**0.80**			**0.73**
Mean Jaccard similarity per response					0.68			0.63
Percentage of responses with exact-match label set					56%			46%

centrated: for both models the residual *Other* category and the two most prevalent substantive categories (*Reduce spending*, *Promote growth/jobs*) together account for roughly 60–69% of all disagreement. These patterns point to a difference in classification style at the boundary between substantive categories and the residual bucket: `Qwen3` more often commits ambiguous responses to a substantive category, whereas the cross-check models more often route them to *Other*. Across the remaining categories, per-category F1 is generally in the 0.62–0.88 range, with no clear category-specific breakdown other than Opus's tendency to over-attribute *Cut politicians' salaries*. Importantly, the categories that carry the substantive findings of the paper – the prevalence of spending cuts, growth, anti-corruption, and tax measures – show strong three-way agreement. The remaining disagreement should, therefore, be treated as annotation uncertainty concentrated in residual and borderline classifications, rather than as evidence that the substantive category structure breaks down.

Table A7: F1 by category for each cross-check model against the published `Qwen3-Thinking-2507` pipeline (n=900; 300 per country). "Disagr. share" is the percentage of all disagreeing cells (false positives + false negatives relative to `Qwen3`) attributable to that category. Categories are ordered as in the published prompt.

			GPT-5.5		Claude Opus 4.6	
#	Category	N_{Qwen}	F1	Disagr. share	F1	Disagr. share
1	Reduce spending	449	0.79	23.9%	0.79	21.0%
2	Stop corruption	92	0.78	6.0%	0.76	5.3%
3	Privatize	16	0.81	1.0%	0.72	1.2%
4	Reduce inflation	10	0.71	1.2%	0.62	1.2%
5	Improve education	13	0.76	1.2%	0.74	1.1%
6	Cut politicians' salaries	105	0.80	7.8%	0.68	11.9%
7	Promote growth / jobs	210	0.83	10.1%	0.77	10.7%
8	Stop tax evasion	59	0.79	3.4%	0.78	3.3%
9	Reduce the size of government	92	0.83	5.0%	0.69	7.8%
10	Increase taxes	47	0.83	2.6%	0.66	5.5%
11	*Other (residual)*	47	**0.28**	**34.6%**	**0.22**	**25.7%**
12	Don't know	92	0.88	3.2%	0.80	5.3%
Macro F1 (mean over 12 categories)			**0.76**		**0.69**	
Macro F1 excluding *Other*			**0.80**		**0.73**	

Cross-Model Validation: The Consequence Questions

Using the same design – 300 randomly sampled responses per country (n=900 per question; `set.seed(42)`), we carried out the same cross-model variation for the national and individual consequences questions.

Table A8 reports the pooled agreement statistics for all three questions side-by-side, and Table A9 gives the category-level F1 for the two consequence questions. The results closely track the policy cross-validation: agreement is consistently high for most substantive categories, while the residual category again displays low F1 because it is used differently across models.

Table A8: Pooled cross-model agreement with the published `Qwen3-Thinking-2507` pipeline for all three classification questions (n=900 each; 300 per country). All models received the published prompt and category list verbatim with the inference settings in Table A5. Micro and macro F1 exclude the residual *Other* category.

Question	Model	Micro F1	Macro F1	Mean Jaccard	Exact-match
Policies	GPT-5.5	0.81	0.80	0.68	56%
Policies	Opus 4.6	0.75	0.73	0.63	46%
Nat. consequences	GPT-5.5	0.82	0.81	0.72	59%
Nat. consequences	Opus 4.6	0.80	0.76	0.70	53%
Indv. consequences	GPT-5.5	0.78	0.76	0.70	55%
Indv. consequences	Opus 4.6	0.74	0.71	0.65	45%

Robustness: Propagating Classification Uncertainty into the Cleavage Estimates

The cleavage models in Figure 5 regress a binary, LLM-classified outcome – whether a respondent mentioned a given policy – on age, left-right orientation, and controls. Because that outcome is classified rather than directly observed, it is measured with error, and a natural concern is whether the weak age and left-right cleavages we report could be an artifact of classification error. We probe this in two complementary ways.

Consensus labels across models. On the cross-model subset (300 policy responses per country; $n = 900$), we rebuild each of the five Figure 5 outcomes as a two-of-three majority vote across the three models (Qwen3, GPT-5.5, and Claude Opus) and re-estimate the cleavage regressions on the consensus outcome, comparing against the single-model (Qwen3) fit on the same subset so that any difference is attributable to the labeling rather than to the smaller sample. The two sets of estimates are nearly identical (correlation of point estimates $= 0.75$), and the same 2 of 30 age and left-right coefficients are statistically significant under either labeling (Figure A5). Using a three-model consensus does not surface cleavages that the published single-model classification misses.

Misclassification-aware Monte Carlo. On the full estimation sample, we propagate classification error into the coefficients directly. For each of 200 draws, we redraw every binary outcome from its posterior probability of being a true positive given the observed label – using the per-category precision and recall measured in the cross-model validation above –

Table A9: Category-level F1 for the two consequence questions, comparing `gpt-5.5` and `claude-opus-4-6` to the published `Qwen3` pipeline (n=900 per question). N_{Qwen} is the number of responses the Qwen3 pipeline assigned to each category. Categories ordered as in the published prompts.

Category	N_{Qwen}	GPT-5.5 F1	Opus F1
National economy			
National Economic Collapse/Decline	526	0.86	0.87
Unemployment/Emigration	49	0.87	0.75
Hunger, Poverty, General Hardship	154	0.77	0.78
Inflation	90	0.86	0.81
Higher Taxes	78	0.85	0.79
No Growth, Recession, Instability	101	0.59	0.60
Inequality	33	0.81	0.75
Reduced safety net / Pensions	56	0.83	0.76
Burden future generations	39	0.82	0.68
No effect	55	0.81	0.82
Other (residual)	12	**0.13**	**0.08**
Don't know	47	0.83	0.78
National: micro F1 (excl. *Other*)		**0.82**	**0.80**
National: macro F1 excl. *Other*		**0.81**	**0.76**
Individual economic situation			
Poverty or General Economic Hardship	505	0.81	0.80
Unemployment or Underemployment	35	0.72	0.70
Higher prices, Inflation, Depreciation	131	0.83	0.78
Emigration (self or relatives)	17	0.87	0.87
Personal Sacrifice	70	0.66	0.53
Increased Taxes	86	0.88	0.85
No Impact or change	99	0.80	0.78
Loss of government benefits / services	88	0.78	0.70
Falling behind others (inequality)	14	0.69	0.63
Uncertain Future	38	0.50	0.43
Other (residual)	7	**0.12**	**0.07**
Don't know	48	0.80	0.69
Individual: micro F1 (excl. *Other*)		**0.78**	**0.74**
Individual: macro F1 excl. *Other*		**0.76**	**0.71**

re-estimate all 15 models, and combine the draws with Rubin's rules so that the reported standard errors absorb both sampling and classification uncertainty. As expected when a weak signal is diluted by noise, the coefficients attenuate (the mean absolute age/left-right coefficient falls by roughly a third) and standard errors widen (by about 19% on average); the number of significant age/left-right coefficients falls from 9 of 30 under the naive fit to 2 of 30 (Figure A6). The substantive conclusion – that demographic and ideological cleavages in proposed debt solutions are weak – is unchanged, and if anything reinforced: injecting realistic classification error shrinks already-small coefficients rather than revealing hidden structure.

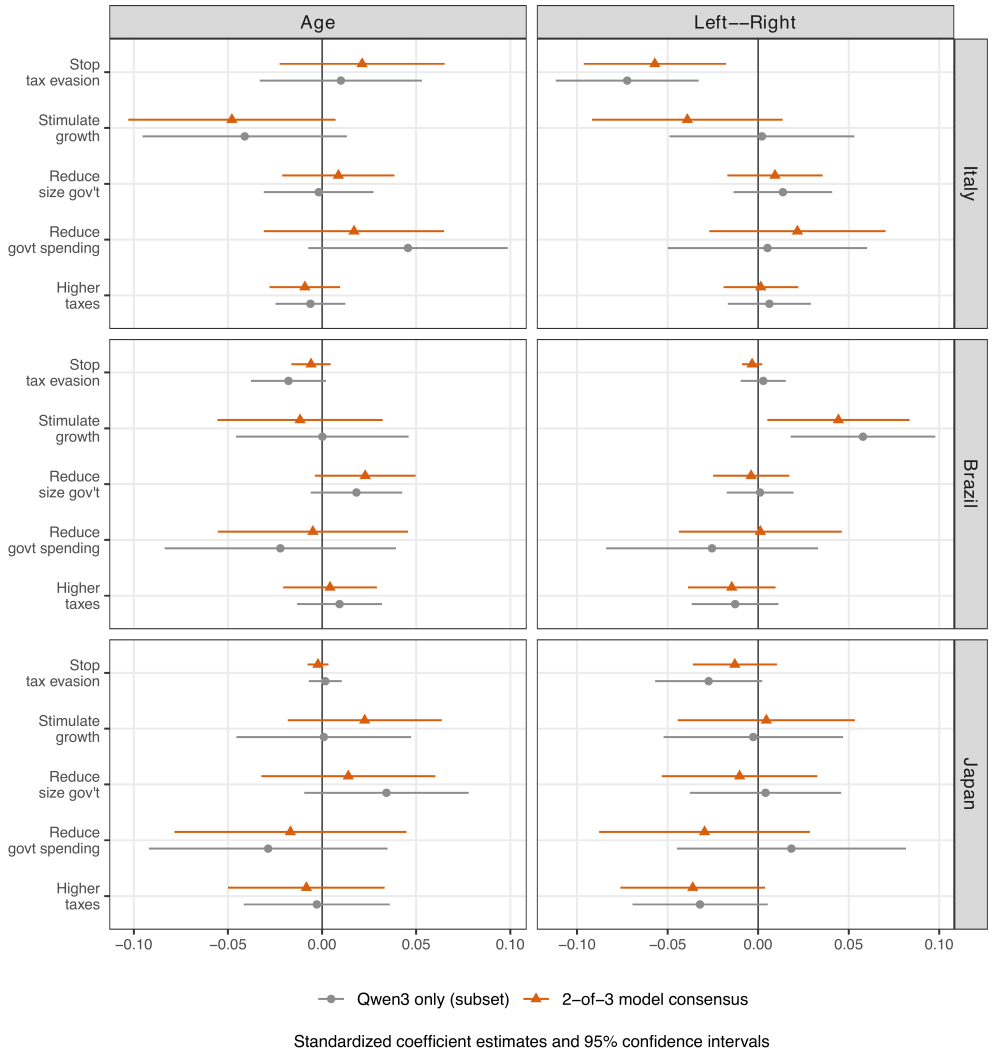

Figure A5: Cleavage estimates under single-model (Qwen3) versus two-of-three cross-model consensus labeling, on the $n = 900$ cross-model subset. Points are standardized age and left-right coefficients with 95% confidence intervals; the two labelings yield nearly identical estimates. Compare with main-text Figure 5.

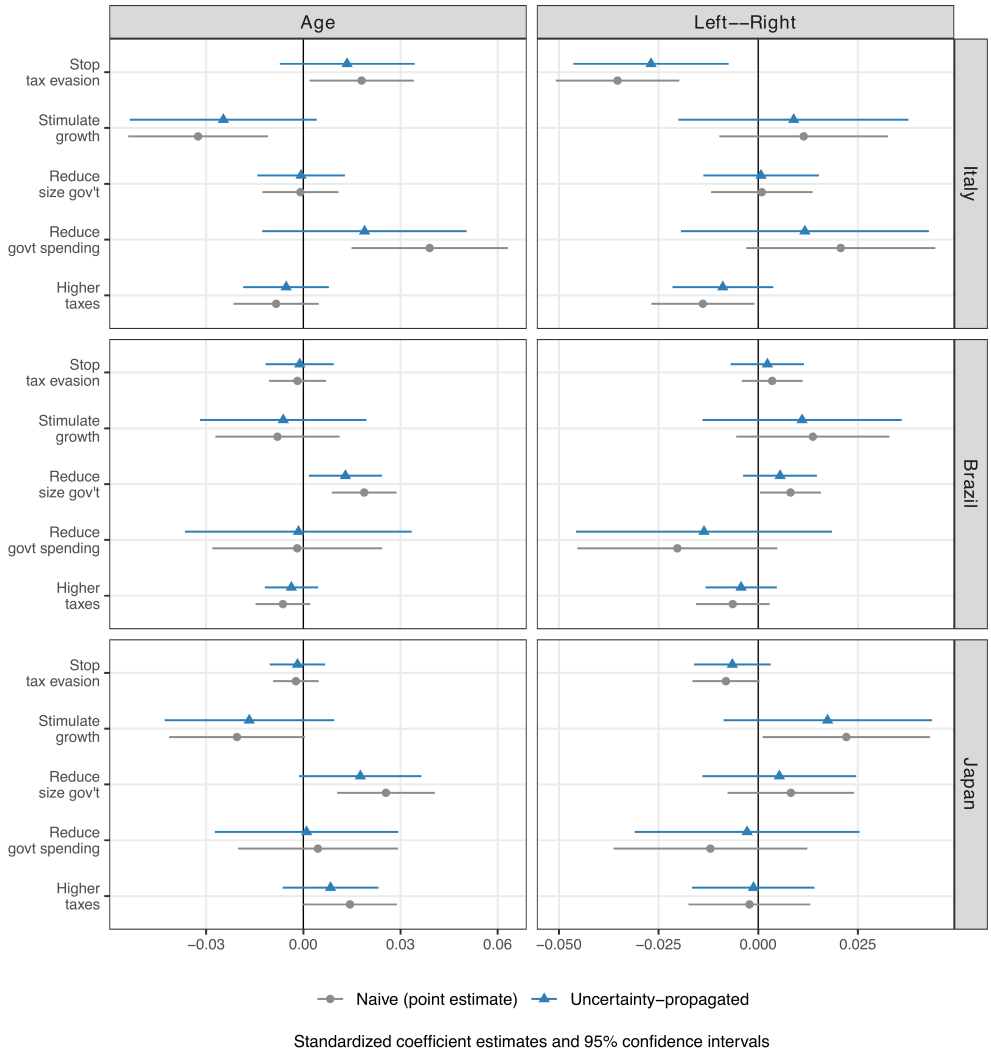

Figure A6: Cleavage estimates before (naive) and after propagating classification uncertainty through a misclassification-aware Monte Carlo (200 draws, combined via Rubin's rules). Propagated intervals are wider and coefficients attenuated, but the weak-cleavage conclusion is unchanged. Compare with main-text Figure 5.

A.5 Semantic Separation: PERMANOVA

In our final analysis we plotted two dimensions of the embedding space of our respondents' recommended policies to address high debt. We saw no visible separation by age and partisanship. Here we present an analysis (PERMANOVA) in which we use the entire multidimensional embedding space to examine if the tertiles of age, ideology, and income are significantly different from a random placement in the text embedding space and the amount of variance explained by the groupings.

While we find statistical significance, the groupings explain a small amount of variance

(R^2 between 0.21% and 0.52% under the open-weights multilingual-E5 embeddings used in the main text, and between 0.18% and 0.63% under the OpenAI `text-embedding-3-large` embeddings used in the prior version of the manuscript). This corresponds with our linear probability model analysis that these factors play a small role in explaining how respondents talk about solutions to public debt.

Table A10 reports both embedding models side-by-side. The substantive conclusion – a statistically detectable but substantively trivial association between demographic tertiles and the policy-narrative embedding space – is unchanged across embedding models. The pattern also holds for income tertiles (reported only under the OpenAI embeddings, since the main-text analysis follows the editor's request to use open-weights models for the central claim).

Table A10: PERMANOVA results for demographic tertiles in the policy embedding space, under both the open-weights `multilingual-E5-large-instruct` embeddings (used in the main-text figures) and OpenAI's `text-embedding-3-large` embeddings (retained as a robustness check). 999 permutations; cosine distance on the full-dimensional embedding matrix. Sample sizes (N) vary across rows because observations missing the relevant tertile variable are dropped listwise. Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Embedding	Country	Variable	N	R^2	F	p	
E5	Italy	Left-Right	1691	0.00447	3.79	0.001	***
OpenAI	Italy	Left-Right	1691	0.00371	3.15	0.001	***
E5	Japan	Left-Right	2027	0.00309	3.14	0.001	***
OpenAI	Japan	Left-Right	2027	0.00246	2.50	0.001	***
E5	Brazil	Left-Right	1601	0.00212	1.70	0.010	**
OpenAI	Brazil	Left-Right	1601	0.00185	1.48	0.030	*
E5	Italy	Age	1691	0.00516	4.38	0.001	***
OpenAI	Italy	Age	1691	0.00482	4.09	0.001	***
E5	Japan	Age	2056	0.00483	4.98	0.001	***
OpenAI	Japan	Age	2056	0.00393	4.05	0.001	***
E5	Brazil	Age	1601	0.00488	3.92	0.001	***
OpenAI	Brazil	Age	1601	0.00522	4.19	0.001	***
OpenAI	Italy	Income	1691	0.00280	2.37	0.001	***
OpenAI	Japan	Income	1685	0.00300	2.54	0.001	***
OpenAI	Brazil	Income	1601	0.00630	5.08	0.001	***

Robustness: OpenAI `text-embedding-3-large`

Figures A7 and A8 re-plot the main-text UMAP projections using the OpenAI embeddings the manuscript used in its prior round. Both the visual pattern (no separation by ideology or

age within any country) and the formal PERMANOVA results in Table A10 are substantively unchanged.

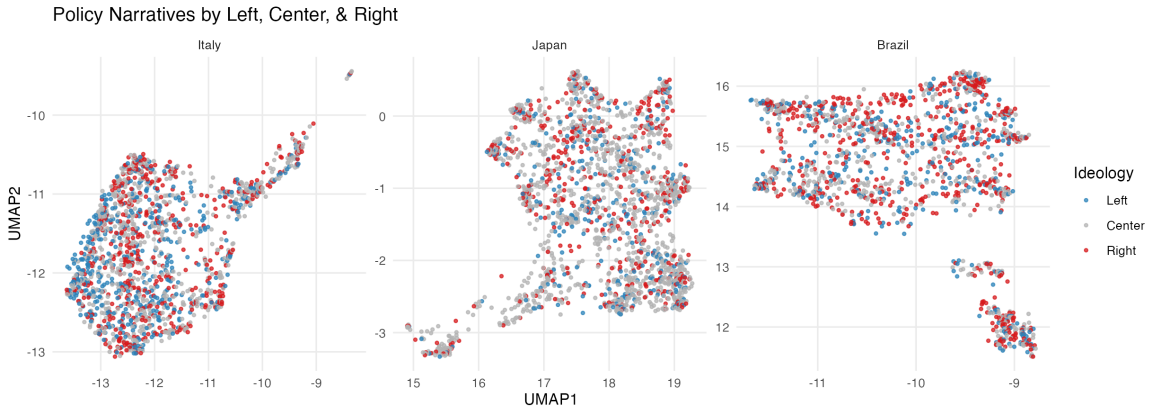

Figure A7: Robustness UMAP projection by political orientation using OpenAI's `text-embedding-3-large` embeddings. Compare with main-text Figure 6.

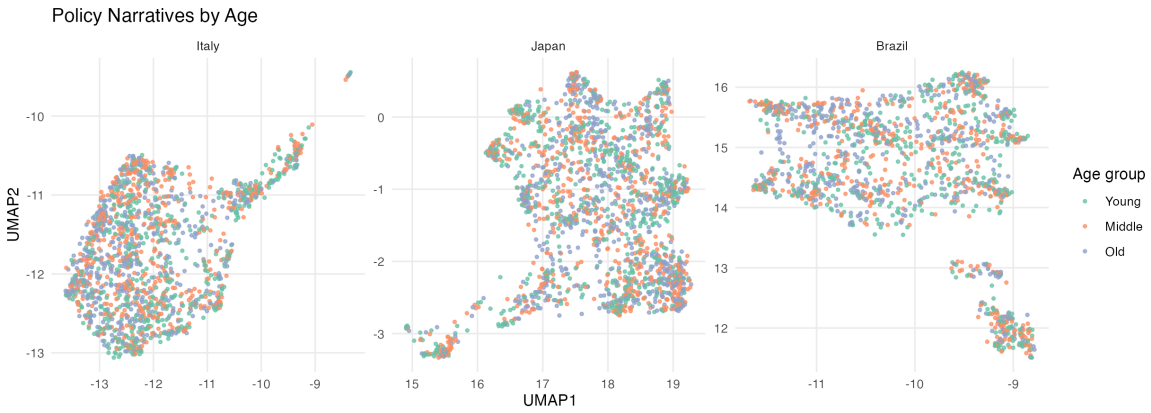

Figure A8: Robustness UMAP projection by tertile age groups using OpenAI's `text-embedding-3-large` embeddings. Compare with main-text Figure 7.

A.6 Analysis of 'Cause of Debt' Responses

As we note in the manuscript, we ask four open-ended questions to our respondents. We omitted the analysis of the "causes" of a country's high debt as it did not map neatly onto debates over base assumptions. Here we present those findings for full transparency.

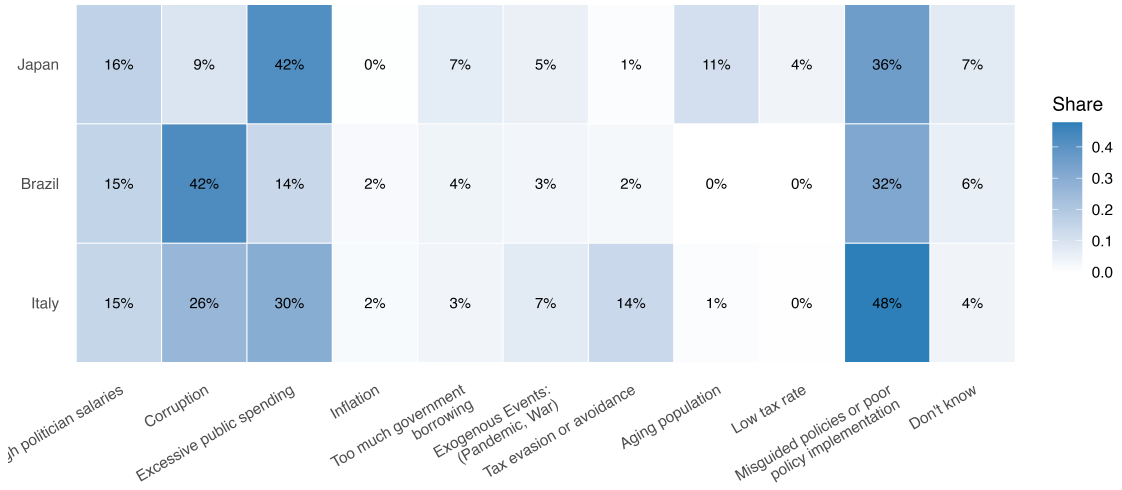

Figure A9: Here we present an analysis of the respondents' response to our question about the causes of high debt in their respective countries. We use the same annotation workflow we used for the other question-responses.

A.7 Prompts For Classification

LLM Prompt (Batch Chat Completions)

```
[System] You are a deterministic classifier. Output exactly one valid JSON object
    matching the user-provided format. No markdown/backticks/explanations.
[User] I asked survey respondents in Italy, Japan, and Brazil: What policies should
     the government adopt to reduce your country's public debt? - Classify what the
     respondent says the government should do to reduce public debt.
Output ONLY this exact JSON format (no other text):
{ "categories": [1], "other_note": null,"is_dk": false }
RULES:
       - categories: array of 1--10 integers from list below (ordered by importance
           )
       - other_note: string if using category 11 ("Other"), else null
       - is_dk: true only if response is "don't know" (then categories=[12])
CATEGORIES:

       1. Reduce spending
       2. Stop corruption, organized crime.
       3. Privatize
       4. Reduce inflation
       5. Improve Education
       6. Cut Politicians salaries
       7. Promote growth/jobs
       8. Stop tax evasion
       9. Reduce the size of government
       10. Increase Taxes
       11. Other
       12. Don't know
STATEMENT: ""
JSON OUTPUT:
```

LLM Prompt (Batch Chat Completions)

```
[System] You are a deterministic classifier. Output exactly one valid JSON object
    matching the user-provided format. No markdown/backticks/explanations.
[User] I asked survey respondents in Italy, Japan, and Brazil: In your opinion,
    what are the main causes of your country's high public debt? - Classify the
    reasons the respondent gives for high public debt.
Output ONLY this exact JSON format (no other text):
{ "categories": [1], "other_note": null,"is_dk": false }
RULES:
       - categories: array of 1--10 integers from list below (ordered by importance
           )
       - other_note: string if using category 11 ("Other"), else null
       - is_dk: true only if response is "don't know" (then categories=[12])
CATEGORIES:
```

```
       1. High Politician Salaries
       2. Corruption
       3. Excessive Public spending (benefits, healthcare)
       4. Inflation
       5. Too much government borrowing
       6. Exogenous Events: (e.g. Pandemic, Ukraine-Russia War)
       7. Tax Evasion or tax avoidance
       8. Aging population
       9. Low tax rate
       10. Misguided policies or poor policy implementation
       11. Other
       12. Don't know
STATEMENT: ""
JSON OUTPUT:
```

LLM Prompt (Batch Chat Completions)

```
[System] You are a deterministic classifier. Output exactly one valid JSON object
    matching the user-provided format. No markdown/backticks/explanations.
[User] I asked survey respondents in Italy, Japan, and Brazil: What do you think
    would happen to your economic situation if the national debt continues to
    increase? - Classify personal/household consequences the respondent expects
    from rising debt.
Output ONLY this exact JSON format (no other text):
{ "categories": [1], "other_note": null,"is_dk": false }
RULES:
       - categories: array of 1--10 integers from list below (ordered by importance
           )
       - other_note: string if using category 11 ("Other"), else null
       - is_dk: true only if response is "don't know" (then categories=[12])
CATEGORIES:

       1. Poverty or General Economic Hardship
       2. Unemployment or Underemployment
       3. Higher prices, Inflation, Currency Depreciation
       4. Emigration (self or relatives)
       5. Personal Sacrifice (e.g. sell assets, go hungry)
       6. Increased Taxes
       7. No Impact or change
       8. Loss of government benefits or services (e.g. Pension, Healthcare)
       9. Falling behind others (inequality)
       10. Uncertain Future
       11. Other
       12. Don't know
STATEMENT: ""
JSON OUTPUT:
```

LLM Prompt (Batch Chat Completions)

```
[System] You are a deterministic classifier. Output exactly one valid JSON object
    matching the user-provided format. No markdown/backticks/explanations.
[User] I asked survey respondents in Italy, Japan, and Brazil: What do you think
    would happen to your country's national economy if the national debt continues
    to increase? - Classify economy-wide consequences the respondent expects from
    rising debt.
Output ONLY this exact JSON format (no other text):
{ "categories": [1], "other_note": null,"is_dk": false }
RULES:
       - categories: array of 1--10 integers from list below (ordered by importance
           )
       - other_note: string if using category 11 ("Other"), else null
       - is_dk: true only if response is "don't know" (then categories=[12])
CATEGORIES:

       1. National Economic Collapse/Decline
       2. Unemployment/Emigration
       3. Hunger, Poverty, General Hardship
       4. Inflation
       5. Higher Taxes
       6. No Growth, Recession, Economic Instability
       7. Inequality
       8. Reduced social safety net and/or Pensions
       9. Burden future generations
       10. No effect
       11. Other
       12. Don't know
STATEMENT: ""
JSON OUTPUT:
```