Reading the Fund with LLMs: An Analysis of IMF Fiscal Advice

Jantsankhorloo Amgalan, Jeremie Cohen-Setton, and Anton Korinek

This version: 01/31/2025

Abstract

This paper provides the first systematic analysis of IMF fiscal policy advice across its entire membership over 1998-2023. Using state-of-the-art large language models to analyze over 3,000 Article IV consultation reports, we classify the Fund's short- and medium-term fiscal policy recommendations and relate them to contemporary economic circumstances. We find that the Fund's advice has predominantly favored fiscal tightening, particularly in its medium-term recommendations where over 90% of advice to emerging markets and low-income countries called for consolidation. While fiscal policy advice has generally been counter-cyclical, with this relationship strengthening after the Global Financial Crisis, we document systematic differences in how economic factors influence advice across country groups. Our methodology demonstrates how large language models (LLMs) can enhance systematic policy evaluation while maintaining analytical rigor.

1. INTRODUCTION

The International Monetary Fund's (IMF) fiscal policy advice is a cornerstone of its surveillance activities, yet systematic evaluation of this advice across the Fund's membership has historically proven challenging. While individual Article IV (AIV) consultations are thoroughly documented, the sheer volume of reports has made comprehensive analysis of patterns in fiscal policy recommendations prohibitively time-consuming. This Background Paper, part of the Independent Evaluation Office's broader assessment of IMF's fiscal policy advice, leverages recent advances in large language models (LLMs) to overcome this constraint and provide the first systematic analysis of fiscal policy recommendations across the Fund's entire membership over a 26-year period.

The Fund has faced persistent questions about the consistency and evenhandedness of its fiscal policy advice. Civil society organizations and Executive Directors have regularly raised concerns about a lack of evenhandedness in the Fund's advice as well as about potential disconnects between the Fund's multilateral messaging and its country-specific recommendations. These latter concerns intensified following the Global Financial Crisis (GFC) and again during the COVID-19 pandemic, as the Fund's public communications emphasized the importance of fiscal support while some stakeholders perceived bilateral advice as remaining focused on consolidation. Similarly, questions about evenhandedness – whether similar circumstances receive similar advice – have been raised, particularly regarding differences in recommendations across country groups.

Until recently, addressing these questions comprehensively was not feasible. Traditional approaches relied either on case studies, which – although informative – could not establish broad patterns, or on manual coding of a limited sample of reports, which faced inevitable trade-offs between coverage and depth. The recent emergence of large language models has fundamentally changed this calculus (Korinek, 2023). These models can now process and interpret complex policy documents with a level of nuance approaching human analysis, while maintaining consistency across thousands of documents. This technological advance enables us to systematically examine the Fund's fiscal policy advice across its entire membership for the first time.

This paper makes three main contributions. First, it develops and validates a methodology for using LLMs to analyze IMF policy advice, establishing a framework that could be applied to other aspects of Fund surveillance. Second, it provides the first comprehensive characterization of how the Fund's fiscal policy advice varies with economic circumstances, documenting both systematic patterns and evolution over time. Third, we produce concrete evidence on the consistency and evenhandedness of Fund advice, helping move debates about these issues from anecdote to systematic analysis.

Our analysis delivers several important findings. First, we find that the Fund's fiscal policy advice has predominantly favored tightening, particularly in its medium-term recommendations where over 90% of advice to EMMIEs and LICs called for fiscal consolidation. Second, we find that the Fund's fiscal policy advice for the near-term has generally been counter-cyclical, with this relationship strengthening after the Global Financial Crisis. Third, we document systematic differences in how economic factors influence advice across country groups, though further analysis is needed to fully assess the implications for evenhandedness. Fourth, our ongoing work examines the alignment between bilateral advice and multilateral messages.

The paper proceeds as follows. Section 2 describes our methodology for extracting and classifying fiscal policy recommendations from Article IV reports, including our approach to validating the reliability of LLM-based analysis. It also discusses other macroeconomic data sources. Section 3 presents our empirical framework for relating fiscal advice to economic circumstances and our main results on the patterns and evolution of Fund fiscal advice. Section 4 explores several specific topics regarding the cyclicality of advice, the link with debt sustainability, and considerations for commodity exporters. Section 5 concludes.

2. DATA AND METHODOLOGY

2.1 Extracting advice on fiscal stance using LLMs

Large language models (LLMs) are well-suited to assist in gathering and classifying information from extensive amounts of unstructured data. In this project, we extract IMF staff advice on the fiscal stance in the short-term (ST) and medium-term (MT) from Article IV (AIV) consultation reports for all IMF member

countries, covering a 26-year period from 1998 to 2023.[1] Using internal sources, we compiled the entire set of AIV reports from this period, totaling 3,115 documents. Tables 1 and 2 detail the coverage by income group and IMF Area Departments, while Figure 1 illustrates the progression of this coverage over time.

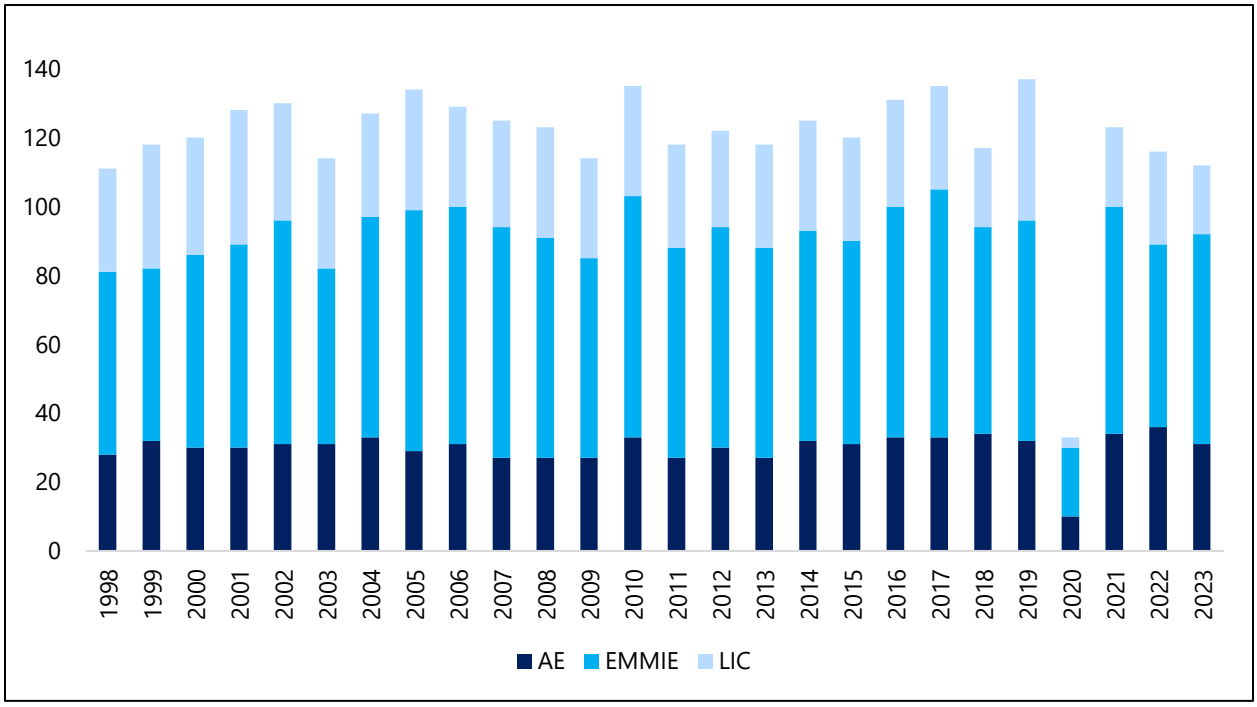
Table 1. AIV coverage by income group, 1998 - 2023

	AE	EMMIE	LIC	Total
Number countries per income group	39	96	58	193
AIVs per group	779	1565	771	3115
Average AIV coverage (% of years with AIV)	77%	63%	51%	62%

Table 2. AIV coverage by IMF Area Departments, 1998 - 2023

	EUR	WHD	APD	AFR	MCD
Number countries per region	45	35	36	45	31
Number of AIVs per region	826	579	633	603	469
Average AIV coverage (% of years with AIV)	71%	64%	68%	52%	58%

Figure 1. AIV publication by income grouping (1998 – 2023)



[1] We have collected the complete collection of AIVs since 1978 with the help of Strategy, Policy, and Review Department of the IMF but haven't yet processed those before 1998.

To categorize the fiscal policy advice from an AIV report, we use three different LLMs (OpenAI's GPT-4o, Anthropic's Claude 3.5 Sonnet and Google DeepMind's Gemini 1.5 Pro) that were at the cutting-edge of LLM capabilities at the time of writing. The process involves the following steps:

- Extract and Clean Data: We start by extracting the text of the staff appraisal section from the AIV report, as this is the section that contains the staff's fiscal policy recommendations, as described in further detail in the appendix.
- LLM Classification: We use the text of the staff appraisal as input for LLMs, alongside a carefully constructed prompt to guide the analysis, as detailed in Box 1. The models are tasked with interpreting the fiscal stance recommendations provided in the staff appraisal section.
- Categorizing Fiscal Stance: We feed the prompt and the staff appraisal into each of the three LLMs (OpenAI's GPT-4o, Anthropic's Claude 3.5 Sonnet, and Gemini 1.5 Pro) for each country-year pair for which an AIV report was compiled to classify the recommended fiscal stance for both the short and medium term into one of four categories: Loosen, Neutral, Tighten, or Unclear. The "Unclear" category is included because, in some cases, advice on the fiscal stance is ambiguous or not explicitly provided in the staff appraisal.
- Justification and Certainty: We instruct each LLM to follow a "chain of thought" approach, providing first the reasoning behind their classification and then the actual classification together with the corresponding level of certainty. This improves the model's classification and helps in understanding the justification behind the model's classification. We also use the provided level of certainty in some of our downstream analyses.
- Majority Voting System: To finalize the fiscal stance classification, the outputs of the three LLMs are compared. If two or more models agree on a classification (e.g., "Tighten"), that stance is selected. If there is no majority consensus, the output from OpenAI's GPT-4o, the most advanced model as of October 2024, is used for the final decision. To calculate the certainty weights, we average the certainty scores for the chosen stance across all three models.

The prompt used to extract the recommended fiscal stance from the staff appraisal section of the AIVs is reported in Box 1. In addition to the fiscal stance, we employ a similar procedure to extract the monetary policy advice contained in the staff appraisal section of the AIV reports.

Box 1. Prompt used to extract fiscal stance from Article IV reports

You will be given the staff appraisal section of an IMF Article IV report. The staff appraisal typically contains two types of information:

1) A factual description of the country's existing fiscal policies and their effects.
2) The IMF's advice for fiscal policy.

Disregard the factual description of existing policies. Your task is to identify and analyze only the IMF's advice on fiscal policies in the given text, for both the near-term and medium-term.

Near-term refers to the current economic cycle, for example, the recovery from an ongoing recession or crisis, or the immediate future (typically 1-2 years). It relates to addressing current economic challenges or imbalances. Near-term advice typically focuses on immediate policy responses to economic shocks or crises, measures to stabilize the economy in the short run, actions to support or moderate current economic growth, and steps to address urgent fiscal imbalances.

Medium-term generally refers to a longer time horizon, typically 3-5 years into the future. It often relates to structural reforms and long-term fiscal sustainability. Medium-term advice typically focuses on fiscal consolidation efforts to ensure debt sustainability, structural reforms to improve economic efficiency and potential growth, measures to address long-term demographic or economic challenges, policies to build fiscal buffers for future economic shocks.

Classify the fiscal advice into one of four categories: Tighten, Neutral, Loosen, or Unclear. These categories represent IMF advice on whether and how to change the fiscal policy stance compared to the current plans of the country's authorities.

Tighten: Refers to recommendations of fiscal restraint, deficit reduction, or fiscal consolidation. They may suggest reducing fiscal stimulus, a more ambitious fiscal strategy, public debt reduction, or continued fiscal prudence. Key terms: spending cuts, revenue increases, fiscal discipline, deficit reduction, higher cyclically adjusted primary balance, fiscal consolidation.

Neutral: Refers to recommendations of maintaining the current cyclically adjusted fiscal position. They may suggest that no change be made, that the current stance strikes the right balance or that only automatic stabilizers should operate. Key terms: neutral fiscal stance, maintain structural balance, limited role for discretionary fiscal policy.

Loosen: Refers to recommendations of a more supportive or accommodative fiscal policy. They may suggest fiscal expansion, higher cyclically adjusted primary deficit, stronger fiscal impulse, or implementing/expanding stimulus packages. Key terms: fiscal stimulus, fiscal easing, supportive fiscal policy, expansionary measures.

Unclear: Reflects that there are conflicting or ambiguous recommendations without a clear overall direction, or that the text offers no explicit fiscal advice, or that the advice is too vague to categorize definitively.

Here is the staff appraisal section from the IMF Article IV report on {country}, dated {date}:
<report_extract> {appraisal} </report_extract>

Please analyze this text carefully. Be concise and use only the information provided in the extract. Do not make assumptions or use outside knowledge.

Provide your reasoning for the near-term fiscal stance classification inside <near_term_reasoning> tags.
Provide your reasoning for the medium-term fiscal stance classification inside <medium_term_reasoning> tags.
Provide your response for the near-term fiscal stance classification inside <near_term_advice> tags.
Provide your response for the medium-term fiscal stance classification inside <medium_term_advice> tags.

Give your final categorizations in this format:
<near_term_advice> [Choose one of Loosen, Neutral, Tighten, or Unclear] </near_term_advice>
<medium_term_advice> [Choose one of Loosen, Neutral, Tighten, or Unclear] </medium_term_advice>

Provide your percentage certainty for each possible classification:
<near_term_certainty> Tighten: [0-100], Neutral: [0-100], Loosen: [0-100], Unclear: [0-100] </near_term_certainty>
<medium_term_certainty> Tighten: [0-100], Neutral: [0-100], Loosen: [0-100], Unclear: [0-100] </medium_term_certainty>

To validate the classification provided by the model, we selected a random sample of 105 country-year pairs of staff appraisal sections and classified them manually.

2.2 Summary of LLM Classification of IMF Staff Advice

In this section, we summarize the results of the LLM-based classification of fiscal and monetary advice provided in the staff appraisal sections of the AIV reports.

Table 3. Summary of LLM classification of recommended fiscal and monetary policy stances

	Unclear	Loose	Neutral	Tight

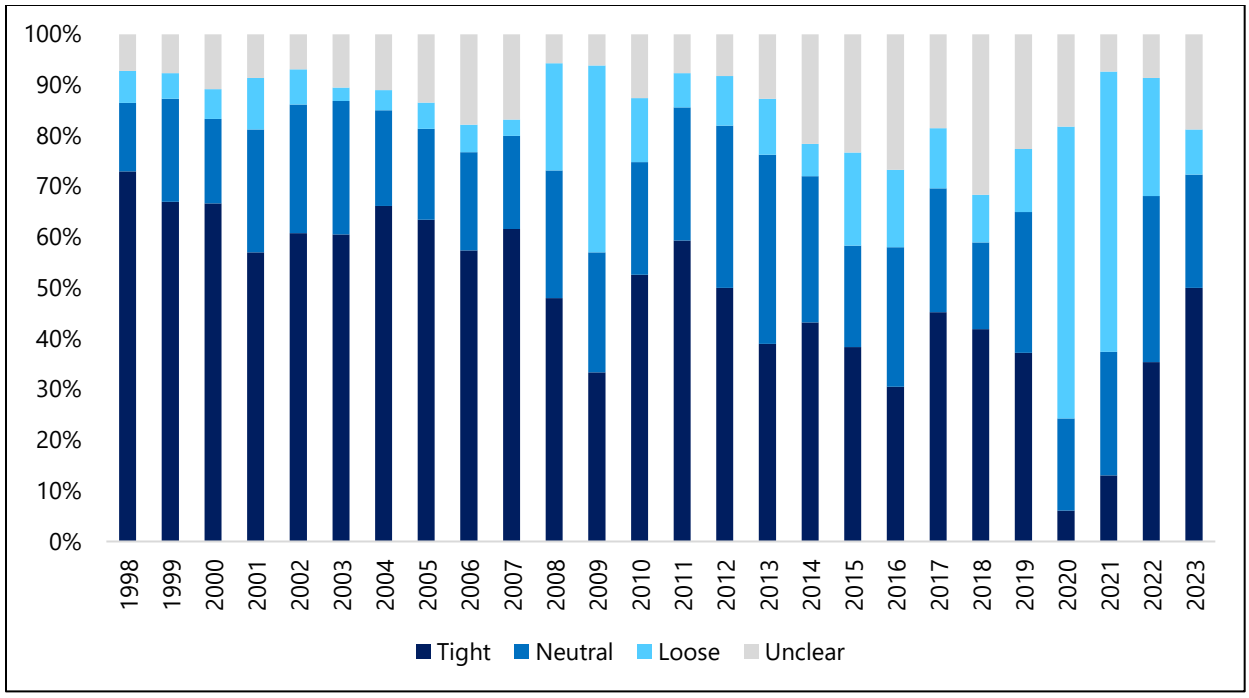
ST Overall	14%	13%	24%	50%
ST AE	11%	20%	31%	38%
ST EMMIE	16%	11%	22%	51%
ST LIC	13%	9%	20%	58%
MT Overall	2%	2%	3%	92%
MT AE	3%	6%	6%	85%
MT EMMIE	1%	1%	2%	96%
MT LIC	3%	1%	3%	93%
Monetary Overall	37%	8%	25%	31%
Monetary AE	50%	11%	22%	17%
Monetary EMMIE	30%	8%	28%	34%
Monetary LIC	36%	3%	23%	38%

Note: ST refers to short-term and MT refers to medium-term fiscal stances. Many of the monetary stance classifications are unclear. This is in part since many countries are in currency unions where no recommendation is provided for individual countries.

Regarding advice on the short-term fiscal stance, about 50% of the recommendations were for a tight fiscal stance across all years and countries. Recommendations for a loose fiscal stance generally have a lower frequency compared to neutral and tight. However, advice to loosen is more prominent around 2009 and 2020-2021, reflecting the IMF's advice during the global crises in those years to stimulate the economy.

For AEs, the distribution is more balanced, with 38% tight, 31% neutral, and 20% loose, suggesting a flexible policy approach compared to other country groups. EMMIEs had a tighter average stance, with 51% of recommendations being tight, reflecting more conservative advice compared to AEs. LICs had the most conservative advice on fiscal stance, with 58% tight advice, and relatively few recommendations for loose fiscal policies.

Figure 2. Short-term recommended fiscal stance, all countries

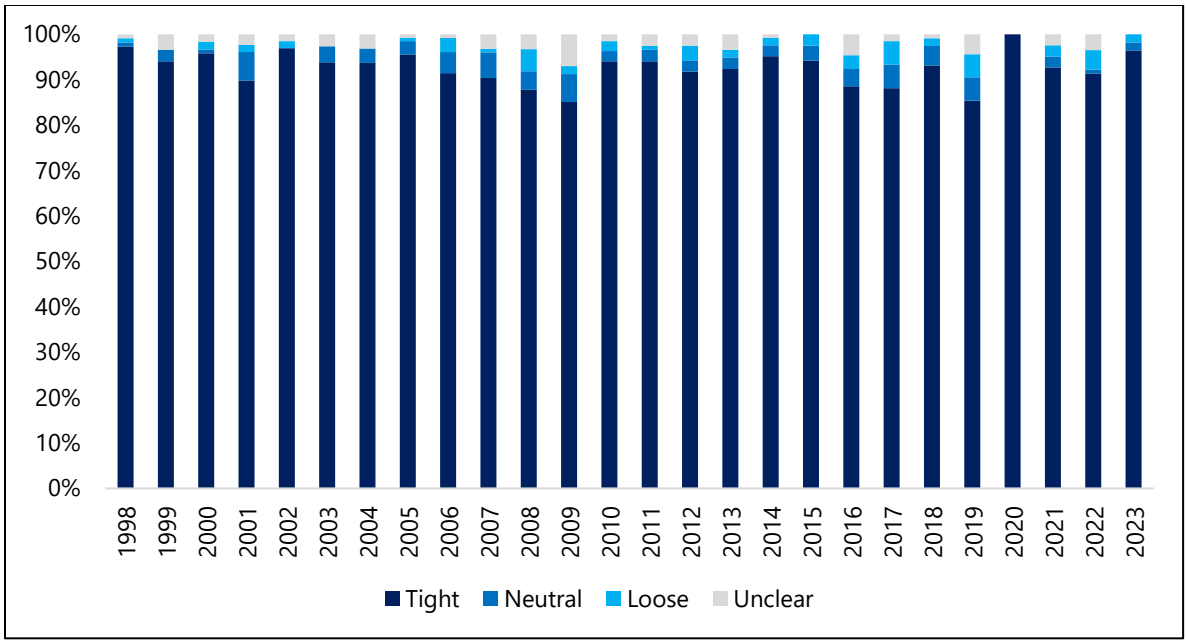


A tight fiscal stance overwhelmingly dominates the medium-term fiscal recommendations, making up the vast majority of each year's recommended fiscal stance. Even a neutral fiscal stance is advised rarely and only takes up a small portion in certain years but remains much less frequent compared to the short-term fiscal stance. Recommendations for a loose fiscal stance are rarely present, and only in a few years, reflecting very limited recommendations for expansionary policies in the medium-term. The category "unclear" fiscal stance barely appears throughout the period, indicating that the IMF's medium-term advice was generally quite clear, unlike in the short term where uncertainty might be more prevalent.

The overall medium-term stance shows that an overwhelming 92% of recommendations were for a tight fiscal stance, with neutral at 3%, loose at 2%, and unclear at 2%, indicating very little flexibility in the medium-term. For AEs, 85% of the recommendations were for a tight fiscal stance, while 6% were neutral and 6% were loose, showing a bit more diversity in recommendations compared to the overall advice but still a strong preference for fiscal tightening. EMMIEs were advised an even stricter 96% tight stance, with minimal recommendations for other stances, similar to the overall pattern. Advice to LICs followed the same trend, with 93% tight advice, 3% neutral, and 1% loose, again reflecting a strong emphasis on fiscal consolidation in the medium-term.

Further figures on the classification of short-and medium-term fiscal advice by income grouping are presented in the Appendix.

Figure 6. Medium-term recommended fiscal stance, all countries



In the table below, we analyze how frequently the IMF's advice changes in two consecutive AIV reports. The table shows that the short-term fiscal stance was unchanged in 59% of consecutive AIVs. Fund advice was more likely to change from one year to the next in AEs (46% probability of a change) than in EMMIEs and LICs (41% and 37% probability of a change, respectively). This is in part due to the higher probability of receiving tightening advice in EMMIEs and LICs.

Table 4. A shift in the policy advice between 2 periods

	All	AE	EMMIE	LIC
ST stance	41%	46%	41%	37%
MT stance	7%	13%	5%	7%
Monetary stance	41%	49%	40%	38%

Table 4. A shift in the policy advice between 2 periods

	All	AE	EMMIE	LIC
Short-term stance				
Loose to Neutral	5%	8%	4%	2%
Loose to Tight	5%	5%	5%	4%
Neutral to Tight	10%	11%	10%	10%
Neutral to Loose	5%	8%	4%	3%
Tight to Neutral	12%	10%	13%	12%
Tight to Loose	5%	5%	5%	5%
Total	41%	46%	41%	37%
Medium-term stance				
Loose to Neutral	4%	7%	3%	2%
Loose to Tight	14%	16%	14%	10%
Neutral to Tight	32%	24%	35%	41%
Neutral to Loose	4%	9%	1%	0%
Tight to Neutral	32%	28%	34%	35%

Tight to Loose	14%	16%	14%	12%
Total	7%	13%	5%	7%
Monetary stance				
Loose to Neutral	10%	18%	8%	6%
Loose to Tight	8%	11%	9%	4%
Neutral to Tight	32%	22%	32%	41%
Neutral to Loose	10%	17%	10%	3%
Tight to Neutral	32%	22%	32%	40%
Tight to Loose	9%	10%	9%	6%
Total	41%	49%	40%	38%

2.3 Data Sources for Macroeconomic Indicators

To relate the Fund's advice on fiscal stance to the economic circumstances at the time, we combine our LLM classification with macroeconomic data, as described in Table 5, which reports the resulting time series and data sources.

To the extent possible, we use data that are contemporaneous to the advice. For macroeconomic data released in vintages, such as those from the World Economic Outlook (WEO), we construct a time series by selecting the closest vintage produced before the publication date of the AIV consultation report, which we label specific vintage. This approach ensures that the selected data reflect the economic conditions as they were understood when the fiscal advice was provided. Further details on the construction of these time series are reported in the Appendix.

We select five indicators as the core variables in our baseline model: the output gap, fiscal balance, primary balance, current account, and public debt, all normalized by GDP. We use these indicators to estimate a Fiscal Advice Function (FAF) that captures how the IMF staff's short-term fiscal recommendations adjust to the prevailing macroeconomic and fiscal environment.

Our estimation procedure is closely related to the literature on fiscal rules and fiscal reaction functions. Eyraud et al. (2018) provide a guide on estimating fiscal rules based on the approach of Bohn (1998), which captures a country's historical fiscal behavior by examining the relationship between the primary balance, debt levels, and key macroeconomic variables. Mauro et al. (2015) document how a country's fiscal stance depends on changes in potential economic growth and on sovereign borrowing costs. Whereas traditional fiscal reaction functions focus on documenting observed fiscal behavior – how authorities have historically adjusted fiscal policy in response to factors such as debt levels and output gaps – our analysis evaluates the fiscal policy advice given by IMF staff.

However, our framework does not focus on observed (positive) fiscal policy decisions but instead on normative fiscal policy advice. This relates closely to the normative framework of Fournier and Lieberknecht (2020) who develop a nonlinear fiscal stance function that uses output gaps, public debt, and the last primary balance as inputs to produce a recommended fiscal stance for the following year. Their framework employs current forecasts of key macroeconomic variables and assesses whether fiscal advice should pay more attention to cyclical or debt-related considerations. This captures the importance

of weighing the two main competing arguments for how to adjust fiscal policy against each other—as illustrated by regular calls for expansionary fiscal policies in downturns (Delong and Summers, 2012; Krugman, 2015) and risk premiums and other debt sustainability considerations (Gruber and Kamin, 2012; Poghosyan, 2012; D'Agostino and Ehrmann, 2014; Fall and Fournier, 2015; Henao-Arbelaez and Sobrinho, 2017). For earlier papers providing practical guidance on a country's fiscal stance see Kanda (2011), Carnot (2014), and Bankowsky and Ferdinandusse (2017).

Our analysis aims to capture a comprehensive set of factors reflecting these considerations—namely the output gap, public debt, current account, and the primary balance—in the fiscal advice contained in the IMF staff recommendations.

Table 5. Data description

Indicator	N	Description	Source	Vintage	Transformation
Macro policy advice					
ST_stance	3115	Majority vote (OpenAI if no agreement) for short-term fiscal stance	AIV		None
MT_stance	3115	Majority vote (OpenAI if no agreement) for medium-term fiscal stance	AIV		None
monetary_stance	3115	Majority vote (OpenAI if no agreement) for monetary policy stance	AIV		None
Cyclical conditions					
ngap	2513	Output gap, % (combination of WEO and Hamilton filter estimates)	WEO	Yes	Winsorized
ngap_r	1432	Output gap, % (only WEO estimates)	WEO	Yes	Winsorized
ngap_r_ha	2513	Output gap, % (only Hamilton filter estimates)	WEO	Yes	Winsorized
5_yr_dev	1859	Real GDP growth deviation from 5y average, %	SPR	No	Winsorized
VE_Real_percentile	1863	Percentile of a country's risk of real sector vulnerability	SPR	No	None
pcip_d_pch	2655	Inflation, %	DSF + DSA, SPR	Yes	Winsorized
Fiscal sustainability					
fb_gdp	3075	Fiscal balance, % of GDP	WBG[2]	No	Winsorized
pb_gdp	2729	Primary balance, % of GDP	DSF + DSA, SPR	Yes	Winsorized
du_gdp	2683	Gross public sector debt, % of GDP	DSF + DSA, SPR	Yes	Winsorized
dugfn_gdp	2567	Gross financing needs, % of GDP	DSF + DSA, SPR	Yes	Winsorized
ds_gdp	2115	Gross public sector debt service, % of GDP	DSF + DSA, SPR	Yes	Winsorized
ds_ggr	2110	Gross public sector debt service, % of Revenue	DSF + DSA, SPR	Yes	Winsorized
External balance					
bca_gdp	3038	Balance of Payments, Current Account, % of GDP	WEO	No	Winsorized
tt	2832	Terms of Trade	WEO	No	Log-demean+ Winsorized

[2] Kose et al. (2022)

Note: The vintage column indicates whether specific vintages are selected for the data series. bca_gdp and tt do not have specific vintages when constructed to increase data coverage. DSF and DSA[3] refer to the IMF debt sustainability framework datasets for LIC and Market-Access Countries, respectively.

Complementary to this approach, we also examine two variables that capture medium-term adjustment needs, such as how much the primary balance needs to change to achieve a sustainable debt level within a defined timeline. Saxegaard et al. (2012) develop indicators that help gauge the medium and long-term adjustment needs. The first indicator focuses on reaching a predetermined debt level within a specified timeframe, while the second indicator identifies the necessary primary balance to maintain a stable debt-to-GDP ratio. This type of reasoning underpins, for example, the European Commission's standard procedure for calculating sustainability indicators in its regular Sustainability Report for EU member states.

3. DRIVERS OF ADVICE ON FISCAL STANCE

3.1 Econometric methodology

Since the LLMs classification is ordinal (Tight > Neutral > Loose) but not cardinal, we use ordinal logistic regressions to analyze the drivers of the staff advice on fiscal stance. We exclude the "unclear" category in our main analysis. We denote by $\mathbf{y} \in \{\text{Loosen } (1), \text{Neutral } (2), \text{Tighten } (3)\}$ the observed realization of the dependent variable, and assume that this observed variable depends on a latent continuous variable $\mathbf{y}^*$ that in turn depends on our explanatory variables according to the following equation:

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\mathbf{y}^*$ is an $\mathrm{N} \times 1$ vector of latent dependent variables for N observations,
- $\mathbf{X}$ is an $\mathrm{N} \times \mathrm{K}$ matrix of independent variables, where K is the number of independent variables,
- $\boldsymbol{\beta}$ is a $\mathrm{K} \times 1$ vector of coefficients,
- $\boldsymbol{\epsilon}$ is an $\mathrm{N} \times 1$ vector of error terms which follows a logistic distribution.

The observed ordinal outcome $\mathbf{y}$ is related to the latent $\boldsymbol{y}^*$ through two threshold points τ_1 and τ_2, which divide the range of $\mathbf{y}^*$ into the corresponding ordinal categories:

$$y = \begin{cases} \text{Loosen} & \text{if } \mathbf{y}^* \leq \tau_1 \\ \text{Neutral} & \text{if } \tau_1 < \mathbf{y}^* \leq \tau_2 \\ \text{Tighten} & \text{if } \mathbf{y}^* > \tau_2 \end{cases}$$

The probability of being below a certain category j (fiscal stance) is modeled using the cumulative logistic distribution:

[3] See Appendix on DSA and DSF.

$$P(y \leq j \mid \mathbf{X}) = \frac{1}{1 + \exp\left(-(\tau_j - \mathbf{X}\boldsymbol{\beta})\right)}$$

The probability of being in category j is the difference between the cumulative probabilities of being in category j and the category just below it:

$$P(y = j \mid \mathbf{X}) = P(y \leq j \mid \mathbf{X}) - P(y \leq j - 1 \mid \mathbf{X})$$

The estimation for n observations and $J = 3$ categories is performed using maximum likelihood estimation (MLE) with the following log-likelihood function:

$$L(\boldsymbol{\beta}, \boldsymbol{\tau}) = \sum_{i=1}^{n} w_i \sum_{j=1}^{J} d_{ij} \log P(y_i = j \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau})$$

where the d_{ij} is an indicator equal to 1 if $y_i = j$, and 0 otherwise, w_i is the weight of observation i, and $\log P(y_i = j \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau})$ is the probability of observing category j for observation i as described in the equation above.

When we employ individual country fixed effects, we simply extend the model with country dummy variables, which allows us to effectively account for unobserved heterogeneity across countries. The latent model then becomes:

$$\mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

- $\mathbf{Z}$ is $\mathrm{N} \times (\mathrm{M} - 1)$ matrix of country dummy variables, where $\mathrm{M} - 1$ is the number of countries minus one, which serves as the reference category to avoid perfect multicollinearity,
- γ is $(\mathrm{M} - 1) \times 1$ vector of country fixed effects (coefficients for each country dummy).

To limit the influence of extreme values in the covariates, we apply winsorization by capping the data at the 1st and 99th percentiles. Further details on this are provided in the Appendix.

We use two separate approaches to determining the weights of different observations. First, we report results where each country-year observation is equally weighted. Additionally, we compute weights based on the certainty values provided by the LLMs when classifying fiscal stance advice.

3.2 Baseline regression results

Our analysis explores the role of three categories of variables—cyclical conditions, fiscal sustainability concerns, and the extent of external imbalances—as illustrated in Table 6.

Our baseline specification includes only five explanatory variables from these categories: the output gap, fiscal balance, primary fiscal balance, public debt-to-GDP ratio, and current account balance. Each variable is expressed as a percentage of GDP (or potential GDP for the output gap). This selection strikes a balance

between including key macroeconomic indicators and data availability, minimizing reductions in sample size.

Table 6 reports four different columns: (1) reports results with no fixed effects and weights, (2) reports results with country fixed effects without weighted observations, (3) reports results with weights but not country fixed effects, and (4) reports results with weighted observations and country fixed effects. With country fixed effects we aim to control for unvarying country-specific effects. We introduce weights, based on LLM certainty of fiscal stance classification, to potentially increase efficiency in our estimation. We find that the results from weighted and unweighted estimation are similar.

The output gap (ngap) consistently shows a positive association with recommendations for a tighter fiscal stance, with highly significant coefficients across all models, ranging from 0.062 to 0.100. This finding indicates that IMF staff advice is, on average, counter-cyclical: they recommend fiscal tightening when output exceeds potential (a positive output gap) and loosening when output is below potential. This cyclical pattern aligns well with standard textbook recommendations for fiscal policy. The relationship remains robust, even after controlling for country-specific effects and applying different weighting schemes.

In contrast, the association between a country's fiscal indicators and IMF staff advice is less clear-cut. Although the simple correlation between a country's overall fiscal balance (fb_gdp) and short-term fiscal advice is significant and negative, the regression coefficients in our baseline regression reported in Table 6 are largely insignificant, and moderately negative in only one specification. A negative relationship reflects that a larger deficit is associated with more tightening advice, as would be suggested by textbook models. Table 6 shows, that the negative relationship is significantly present only for advanced economies. In further analysis, we found that the fiscal balance for our full sample only becomes significant if we drop the country's current account/GDP ratio from our list of controls, which is correlated with the fiscal balance.

After controlling for the overall fiscal balance, a country's primary balance (pb_gdp) shows a positive and significant association with a tighter fiscal stance across all models. This likely reflects that countries with higher debt levels need to maintain a higher primary balance for a given level of the fiscal balance. The gross public sector debt (du_gdp) has a weak association with the fiscal stance recommendations that is insignificant in most specifications. This indicates that, when controlling for the other variables, variations in public debt levels do not appear to significantly drive the Fund's fiscal stance recommendations.

The current account balance (bca_gdp) exhibits a negative and significant relationship with recommendations for fiscal tightening, indicating that countries with larger current account surpluses are more likely to receive advice to loosen their fiscal stance. This effect is consistently strong across models, even after incorporating country fixed effects and different data vintages. This association aligns well with textbook Mundell-Fleming model predictions.

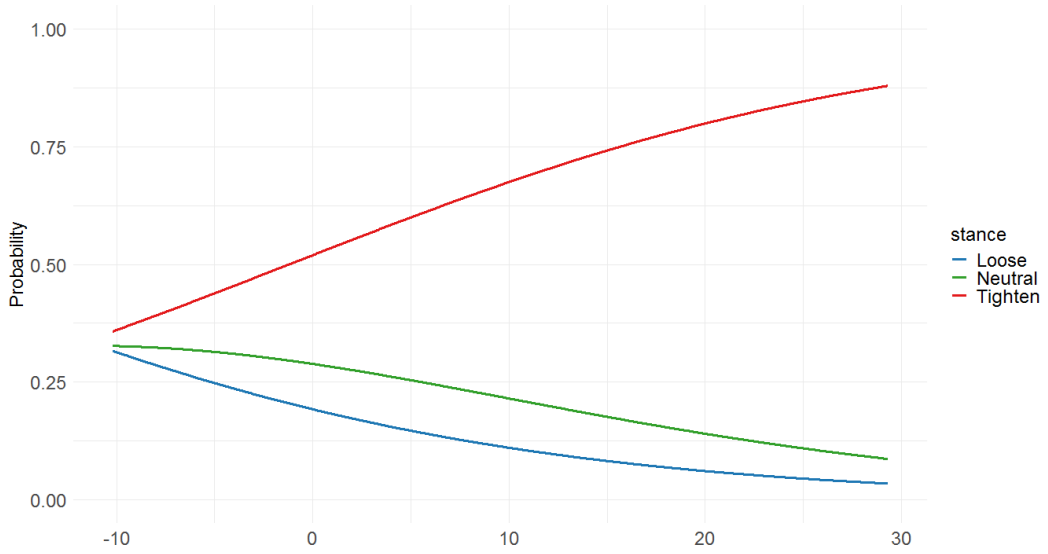
Table 6. Short-term stance, AIV specific vintages[4]

	(1)	(2) Country FE	(3) Weights	(4) Weights + Country FE
ngap	0.062***	0.096***	0.065***	0.100***
	(0.007)	(0.010)	(0.008)	(0.012)
fb_gdp	-0.013	0.005	-0.016	0.001
	(0.010)	(0.012)	(0.012)	(0.014)
pb_gdp	0.029***	0.030***	0.033***	0.035**
	(0.010)	(0.011)	(0.012)	(0.013)
du_gdp	0.002*	0.003*	0.002	0.003
	(0.001)	(0.002)	(0.001)	(0.002)
bca_gdp	-0.034***	-0.031***	-0.037***	-0.034***
	(0.005)	(0.008)	(0.006)	(0.009)
N	1904	1904	1896	1896

Note: Standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01.

Figure 7. Probability plot, short-term stance, output gap, specific vintage

[4] Specific vintages refer to selecting a variable value from the vintage published immediately before the AIV publication date. If no such vintage exists, the most recent available vintage—currently the July 2024 WEO—is selected to maximize coverage, and the value is applied retrospectively.



3.3 Results by Income Grouping

Our next analysis evaluates the extent to which IMF staff on the ST fiscal stance varies based on a country's income level, as captured by its membership in one of the three income group classifications in the WEO: Advanced Economies (AEs), Emerging Markets and Middle-Income Economies (EMMIEs), and Low-Income Countries (LICs). To explore this, we include interaction terms between these income groupings and our key baseline variables from the ordered logit regression model. The results are presented in the following table.

Table 7. Short-term stance, AIV specific vintages and income grouping interaction.

	(1)	(2)
		Country FE
ngap x AE	0.203***	0.025***
	(0.047)	(0.056)
ngap x EMMIE	0.052***	0.071***
	(0.012)	(0.019)
ngap x LIC	0.038**	0.090***
	(0.017)	(0.026)
fb_gdp x AE	-0.198***	-0.187***
	(0.050)	(0.067)
fb_gdp x EMMIE	-0.022	-0.024
	(0.020)	(0.028)
fb_gdp x LIC	-0.022	-0.015
	(0.030)	(0.036)
pb_gdp x AE	0.227***	0.175***
	(0.047)	(0.061)
pb_gdp x EMMIE	0.047**	0.076***
	(0.019)	(0.026)
pb_gdp x LIC	-0.054	-0.066

	(0.034)	(0.045)
du_gdp x AE	-0.001	-0.005
	(0.002)	(0.005)
du_gdp x EMMIE	0.003	0.001
	(0.003)	(0.005)
du_gdp x LIC	0.011***	0.012***
	(0.003)	(0.004)
bca_gdp x AE	-0.061***	-0.041
	(0.019)	(0.032)
bca_gdp x EMMIE	-0.039***	-0.041**
	(0.009)	(0.0161)
bca_gdp x LIC	-0.001	-0.018
	(0.011)	(0.018)
AE	-0.081	-0.328
	(0.321)	(0.995)
EMMIE	0.05	0.87
	(0.307)	(0.544)
N	1896	1896

Note: Standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01. Total effect of the interaction term is reported.

The output gap (ngap) exhibits a positive and significant association with the likelihood of receiving advice for fiscal tightening across all income groups, suggesting that, on average, staff recommendations for the short-term (ST) fiscal stance are counter-cyclical. This effect is, however, most pronounced for AEs than for EMMIEs, implying that other factors may hold greater weight in fiscal stance decisions for these groups. For 1 percentage point increase in the output gap, the odds[5] of receiving tightening advice for AEs increases by 22.5%, while it is 5.3% for EMMIEs and 3.8% for LICs.

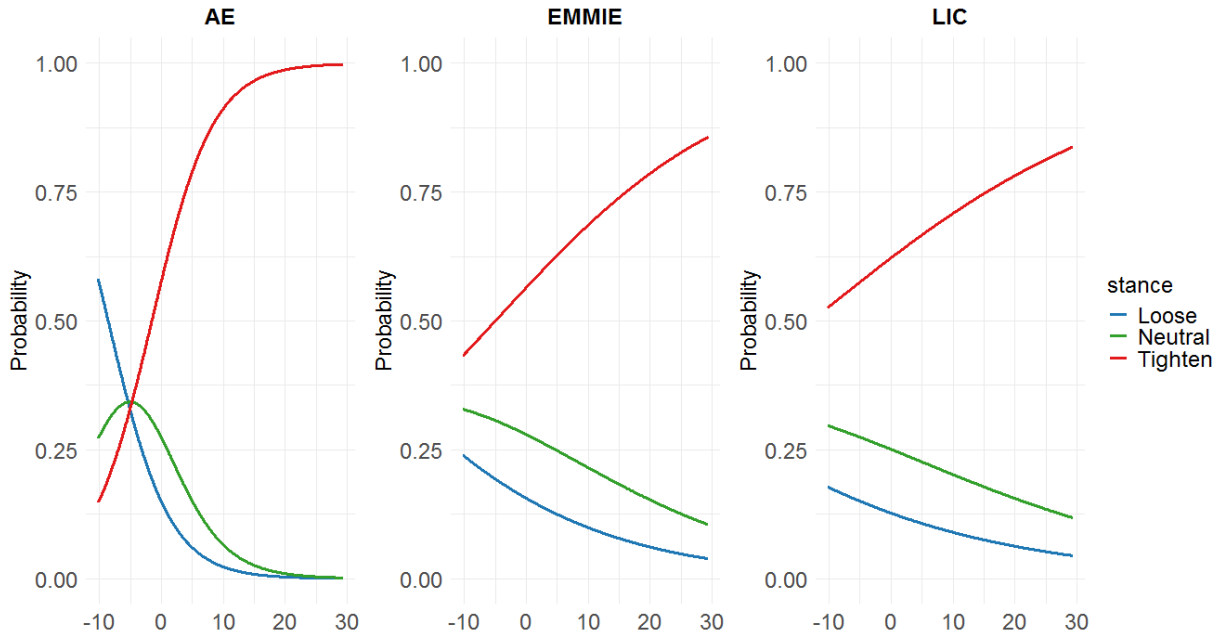
For fiscal balance (fb_gdp), there is a negative and significant effect only for AEs, as already observed, indicating that a deterioration in the fiscal balance increases the likelihood of staff recommending a tightening of the fiscal stance. This relationship does not hold significantly for EMMIEs or LICs, indicating that fiscal balance plays a less influential role in shaping fiscal stance recommendations for these economies. Consistent with earlier findings, the primary balance (pb_gdp) has a positive and significant effect for both AEs and EMMIEs, where higher primary balances (surpluses) increase the likelihood of a fiscal tightening recommendation. In contrast, this effect is insignificant for LICs.

Public sector debt (du_gdp) does not have a significant effect for AEs or EMMIEs. However, for LICs, the relationship between public sector debt and fiscal tightening advice is positive and significant, indicating that higher debt levels in LICs increase the likelihood of receiving a recommendation for fiscal tightening. This suggests that debt sustainability concerns are more prominent in fiscal advice for LICs compared to other groups.

[5] By exponentiating the coefficients from ordered logit model, we can get the odds ratio, which tells us about the multiplicative change in the odds of recommending a tightening stance, for example, for one unit increase in an explanatory variable, ceteris paribus. For the case of output gap for AEs, it is $e^{0.203} \approx 1.225$, which is 22.5% increased odds of receiving tightening advice for 1 percentage point increase in output gap.

Current account balance (bca_gdp) has a negative and significant effect for AEs, where higher current account surpluses are associated with a higher likelihood of fiscal loosening advice. For EMMIEs, the effect remains significant, though somewhat smaller, while for LICs, current account balances do not appear to significantly influence fiscal stance recommendations.

Figure 8. Probability plots, output gap



Sources: See text; IEO calculations.

Note:

3.4 Inertia in Advice

We now investigate to what extent AIV advice depends on the advice provided in prior years. If our macroeconomic variables could perfectly control for the macroeconomic situation of a country and there was no inertia, then the advice given to a country in a given year should be independent of the previous year's advice. As indicated in Table 8, this is far from the case.

The table measures the dependence of current advice on past advice in two complementary ways: in columns 1 and 2 of Table 8, we include the prev_stance variable as a cardinal variable and find that it has a highly significant coefficient – past tightening advice is followed by more tightening advice, even controlling for all other variables. Then, we account for the ordinal nature of fiscal advice by including separate dummies for previous loosening and tightening advice (columns 3 and 4 of Table 8). Previous loosening advice is only modestly associated with current advice, yet previous tightening advice significantly increases the probability of further tightening advice.

Table 8. Short-term stance, previous advice on stance

	(1)	(2) Country FE	(3)	(4) Country FE
ngap	0.054***	0.096***	0.053***	0.096***
	(0.009)	(0.015)	(0.009)	(0.015)
fb_gdp	-0.013	-0.002	-0.013	-0.002
	(0.013)	(0.019)	(0.013)	(0.019)
pb_gdp	0.029**	0.037**	0.030**	0.038**
	(0.013)	(0.018)	(0.013)	(0.018)
du_gdp	0.002	0.002	0.001	0.002
	(0.001)	(0.003)	(0.001)	(0.003)
bca_gdp	-0.028***	-0.033***	-0.028***	-0.033***
	(0.006)	(0.011)	(0.006)	(0.011)
prev_stance	0.032***	0.083**		
	(0.035)	(0.040)		
prev_stance_1			-0.274*	0.083
			(0.152)	(0.168)
prev_stance_5			0.923***	0.356**
			(0.123)	(0.138)
N	1874	1874	1874	1874

Note: Standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01. Weights included.

4. ROBUSTNESS ANALYSIS

4.1 On the Counter-Cyclicality of the Advice

Next, we assess the robustness of our findings on the counter-cyclicality of fiscal advice by testing alternative indicators for a country's cyclical position beyond our baseline measure, ngap. The baseline measure ngap is constructed by combining current output gap estimates from the World Economic Outlook (WEO), when available, with calculated output gaps derived from a Hamilton filter applied to contemporary WEO real GDP series.

We investigate four alternative measures for a country's cyclical position. The first, ngap_r, includes only the WEO output gap estimates (column 2). The second, ngap_r_ha, applies a Hamilton filter to calculate output gaps for all countries (column 3). The third measure, 5_yr_dev, captures deviations in underlying growth from a 5-year average (column 4). The fourth measure, VE_Real_percentile, measures a country's ranking the real crisis risk index calculated by the IMF's Strategy and Policy Review (SPR) Department 's as part of the Vulnerability Exercise (column 5). A higher value of the index reflects a higher probability of a sharp slowdown in economic activity or a more moderate but prolonged slowdown episodes. All WEO variables are derived from vintages closest to when the advice was provided. We find that no matter which of the indicators of cyclical conditions we employ, the association with IMF staff advice continues to be countercyclical and significant.

Table 9. Short-term stance, different indicators for cyclical conditions

	(1) Baseline	(2)	(3)	(4)	(5)
ngap	0.065***				

	(0.008)				
ngap_r		0.121***			
		(0.024)			
ngap_r_ha			0.059***		
			(0.009)		
5_yr_dev				0.049**	
				(0.022)	
VE_Real_percentile					0.004**
					(0.002)
fb_gdp	-0.016	-0.092***	-0.020	-0.031**	-0.033**
	(0.012)	(0.026)	(0.012)	(0.014)	(0.014)
pb_gdp	0.033***	0.136***	0.039***	0.059***	0.063***
	(0.012)	(0.026)	(0.012)	(0.014)	(0.014)
du_gdp	0.002000	-0.0004	0.002*	-0.0005	-0.002
	(0.001)	(0.002)	(0.001)	(0.001)	(0.002)
bca_gdp	-0.037***	-0.064***	-0.035***	-0.037***	-0.038***
	(0.006)	(0.011)	(0.006)	(0.006)	(0.007)
N	1896	1107	1896	1511	1513

Note: Standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01. Weights and controls included.

Evolution of Counter-Cyclicality of Fiscal Advice Over Time

We revert to using our baseline measure of cyclical conditions to examine whether the counter-cyclicality of IMF fiscal advice has evolved over time. We begin by incorporating a linear time trend into the baseline specification to assess whether the likelihood of recommending a tighter fiscal stance has decreased over time. Following this, we explore the evolution of counter-cyclicality in the advice by comparing pre- and post-Global Financial Crisis (GFC) periods. We further segment the analysis by examining five-year intervals starting from 1998 and by measuring the degree of counter-cyclicality in the advice on an annual basis.

Column 1 of Table 10 uses time indicators for five buckets of five years each, starting in 1998, 2003, 2008, 2013, and 2018 (with the last covering six years). From looking at the constant terms in the table, we can see that the overall advice of the fund has become significantly looser over time – the period from 2008 onwards is associated with significantly lower odds of the tighter stance compared to the pre-2003 period. Moreover, the coefficient on the output gap has become larger in more recent periods (especially 2013-2023) so the Fund's advice has become more responsive to the output gap after the crisis and in the post-recovery periods.

Columns 2 and 3 (with FE) of Table 10 shows the results of adding a linear time trend ("year") to the baseline. The coefficient is negative and highly significant in the models with and without country fixed effects, indicating that, over time, there is a decreasing likelihood of recommending tighter fiscal stance, possibly reflecting a shift toward more counter-cyclicality of the IMF's fiscal advice.

Table 10. Short-term stance, with various time trends

	(1)	(2)	(3) Country FE	(4)
ngap	0.017	0.059***	0.090***	0.044**

	(0.027)	(0.009)	(0.015)	(0.009)
ngap x 2003 - 2007	0.076***			
	(0.026)			
ngap x 2008 - 2012	0.032**			
	(0.013)			
ngap x 2013 - 2017	0.080***			
	(0.020)			
ngap x 2018 - 2023	0.091***			
	(0.019)			
2003-2007	0.164			
	(0.240)			
2008-2012	-0.577***			
	(0.212)			
2013-2017	-0.694***			
	(0.211)			
2018-2023	-1.140***			
	(0.209)			
Year		-0.054***	-0.063***	
		(0.008)	(0.012)	
post-gfc				-0.354***
				(0.093)
ngap x post-gfc				0.073***
				(0.011)
fb_gdp	-0.023*	-0.02	-0.013	-0.017
	(0.013)	(0.012)	(0.018)	(0.012)
pb_gdp	0.025**	0.028**	0.032*	0.030**
	(0.013)	(0.012)	(0.017)	(0.012)
du_gdp	0.002	0.002*	-0.004	0.002
	(0.001)	(0.001)	(0.003)	(0.001)
bca_gdp	-0.037***	-0.036***	-0.032***	-0.037***
	(0.006)	(0.006)	(0.011)	(0.006)
N	1896	1896	1896	1896

Note: Standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01. Weights and controls included. Total effect of the interaction term is reported.

Column 4 of Table 10 examines whether the counter-cyclicality of IMF fiscal advice shifted post-GFC. To do this, we introduce a dummy variable set to 1 for years beginning in 2009 and interact it with the output gap. As the coefficient on the dummy itself indicates, the IMF's fiscal advice was less focused on tightening in the post-GFC period. Moreover, the output gap's positive effect on the advice is amplified after the GFC, indicating that the relationship between the output gap and the outcome is more counter-cyclical in the post-crisis period.

4.2 On the link with Debt Sustainability

We assess the robustness of our findings on the relationship between fiscal advice and debt sustainability concerns by expanding the set of relevant indicators beyond the fiscal balance, primary balance, and debt level. Specifically, we include measures of a country's debt service capacity (ds_ggr, defined as debt service as a percentage of general government revenue) and financing needs (dugfn_gdp, representing

gross financing needs as a percentage of GDP). This expansion reduces our sample size, as data is missing for mostly in 2000-2010 periods for some countries.

Table 13. Short-term stance, debt service and financing needs

	(1)	(2)
		country FE
ngap	0.058***	0.102***
	(0.010)	(0.017)
fb_gdp	-0.034**	-0.016
	(0.017)	(0.025)
pb_gdp	0.058***	0.048**
	(0.015)	(0.022)
du_gdp	-0.002	-0.009**
	(0.002)	(0.004)
bca_gdp	-0.040***	-0.041***
	(0.007)	(0.012)
ds_ggr	0.008***	0.013**
	(0.003)	(0.005)
dugfn_gdp	-0.007	-0.010
	(0.011)	(0.020)
N	1619	1619

Note: Standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01. Weights included.

Table 13 shows the results of adding a country's debt service capacity and gross financing needs as additional explanatory variables. The coefficients for a country's debt service capacity are positive and significant in both models, indicating that higher debt service burdens increase the likelihood of recommending a tighter fiscal stance. On the other hand, the coefficient for a country's gross financing needs is insignificant in both models.

Table 14 does the same exercise but allows for different effects for AEs, EMMIEs, and LICs. Debt service plays a significant role in influencing IMF fiscal recommendations for EMMIEs, with higher debt service obligations leading to tighter fiscal stance recommendations. In contrast, debt service does not appear to significantly influence recommendations for LICs, and its influence on AEs diminishes when controlling for country-specific factors. Gross financing needs significantly influence fiscal recommendations in AEs in the model without fixed effects, suggesting that larger financing requirements lead to fiscal tightening recommendations. Gross financing needs do not play a significant role for EMMIEs or LICs.

Table 14. Short-term stance, debt service and financing needs income group interaction

		Country FE
ngap x AE	0.209***	0.239***
	(0.051)	(0.059)
ngap x EMMIE	0.051***	0.081**
	(0.014)	(0.021)
ngap x LIC	0.050**	0.112***
	(0.020)	(0.033)
fb_gdp x AE	-0.183***	-0.167**
	(0.052)	(0.069)
fb_gdp x EMMIE	-0.021	-0.030
	(0.027)	(0.037)

fb_gdp x LIC	-0.047	-0.014
	(0.040)	(0.055)
pb_gdp x AE	0.307***	0.220**
	(0.057)	(0.088)
pb_gdp x EMMIE	0.068***	0.075**
	(0.025)	(0.036)
pb_gdp x LIC	-0.061	-3.435
	(0.052)	(0.077)
du_gdp x AE	-0.006	-0.017*
	(0.004)	(0.009)
du_gdp x EMMIE	-0.003	-0.01
	(0.004)	(0.007)
du_gdp x LIC	0.011**	0.011
	(0.005)	(0.010)
bca_gdp x AE	-0.051**	-0.037
	(0.021)	(0.034)
bca_gdp x EMMIE	-0.044***	-0.051***
	(0.010)	(0.018)
bca_gdp x LIC	0.002	-0.009
	(0.014)	(0.021)
ds_ggr x AE	-0.019*	0.011
	(0.010)	(0.019)
ds_ggr x EMMIE	0.006	0.018*
	(0.004)	(0.010)
ds_ggr x LIC	-0.007	-0.013
	(0.007)	(0.010)
dugfn_gdp x AE	0.080**	0.042
	(0.035)	(0.060)
dugfn_gdp x EMMIE	0.014	-0.004
	(0.019)	(0.038)
dugfn_gdp x LIC	0.008	0.0001
	(0.030)	(0.037)
AE	0.129	-0.641
	(0.388)	(1.257)
EMMIE	0.177	1.113
	(0.374)	(0.690)
N	1619	1619

Note: Standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01. Weights included. Total effect of the interaction term is reported.

4.3 Commodity Exporters

This section examines if the drivers of fiscal advice are different for commodity exporters (defined in Appendix 7.1.) Table 15 shows that the advice for commodity exporters is more counter-cyclical than for non-commodity exporters. Interestingly, the somewhat puzzling result that a higher primary balance is associated with a tighter fiscal stance appears to be driven by commodity exporters. For non-commodity exporters, the coefficient is statistically insignificant. Other variables such as public debt, current account balance, and terms of trade do not seem to have a differential effect on the advice to commodity exporters.

Table 15. Short-term stance, commodity exporters and income grouping intercept

ngap	0.042***
	(0.013)
ngap x cex	0.059***
	(0.003)

fb_gdp	-0.021
	(0.019)
fb_gdp x cex	-0.023
	(0.026)
pb_gdp	0.017
	(0.018)
pb_gdp x cex	0.057**
	(0.024)
du_gdp	0.002
	(0.001)
du_gdp x cex	0.005
	(0.003)
bca_gdp	-0.049***
	(0.009)
bca_gdp x cex	-0.011
	(0.013)
tt	0.563
	(0.364)
tt x cex	-0.167
	(0.352)
AE	-0.500***
	(0.185)
EMMIE	-0.070
	(0.158)
cex	-0.107
	(0.241)
N	1798

Note: Standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01. Weights included. Total effect of the interaction term is reported.

5. CONCLUSION

This paper has developed and implemented a novel methodology for analyzing the IMF's fiscal policy advice across its membership over a 26-year period (1998-2023). The analysis leverages state-of-the-art large language models to systematically extract and classify fiscal policy recommendations from Article IV consultation reports, while carefully relating this advice to contemporary economic circumstances.

We introduce a robust approach for using LLMs to analyze IMF policy advice. By employing three different cutting-edge models (GPT-4, Claude 3.5 Sonnet, and Gemini 1.5 Pro) with carefully designed prompts and a majority voting system, the paper established a reliable framework for classifying fiscal stance recommendations. This multi-model approach, combined with certainty weights and chain-of-thought reasoning, helps ensure the robustness of the classification results while maintaining transparency about potential ambiguities in the advice.

Then we develop a comprehensive framework for relating fiscal advice to economic circumstances. By constructing contemporary measures of cyclical conditions, fiscal sustainability concerns, and external imbalances, the analysis ensures that the relationship between economic conditions and policy advice is evaluated using information available at the time the advice was given.

The results reveal several important patterns in the IMF's fiscal policy advice. The analysis found that fiscal advice has generally been counter-cyclical, with this relationship strengthening after the Global Financial Crisis. The paper also documented systematic differences in how various economic factors influence advice across different country groups. For instance, debt sustainability concerns appear more influential for LICs, while current account considerations play a larger role in shaping advice to advanced economies and to commodity exporters.

This analysis makes several important contributions to the evaluation objectives outlined for BP5.

First, the paper establishes a rigorous framework for employing LLMs in policy evaluation, demonstrating how these tools can be used to systematically analyze large volumes of institutional documents while maintaining high standards of accuracy and transparency. This framework may also be useful in future IEO evaluations.

Second, our methodology successfully extracted and classified fiscal stance recommendations from over 3,000 Article IV reports, providing a comprehensive picture of the Fund's bilateral advice across time and country groups. The analysis goes beyond simple classification to capture nuances in the certainty and reasoning behind different policy recommendations.

Third, by relating bilateral advice to economic circumstances, the paper provides a systematic way to evaluate whether similar circumstances received similar advice. Through detailed analysis of how different factors influence advice across country groups, the paper provides concrete evidence on the evenhandedness of Fund advice.

While finding some systematic differences across country groups, the analysis helps distinguish when these differences reflect legitimate tailoring to country circumstances versus instances where advice may have diverged from typical patterns. This is part of the focus of our ongoing work. We are also still examining how bilateral advice evolved in relation to major shifts in the Fund's overall policy framework, particularly around the Global Financial Crisis. This will help evaluate the consistency between multilateral and bilateral messaging on fiscal policy.

Several overall findings emerge from our analysis:

1. The Fund's fiscal advice has become more sophisticated and nuanced over time, with increasing attention to country-specific circumstances and constraints.

2. The strengthening of counter-cyclical advice post-GFC suggests the Fund has successfully incorporated lessons from major crises into its surveillance practices.

3. There are systematic differences in how advice responds to economic conditions across country groups, and we are still examining the extent to which these differences can be justified as reflecting legitimate tailoring of advice rather than lack of evenhandedness.

4. The methodology developed here could be employed to analyze other aspects of Fund advice and could be valuable for future evaluations of surveillance activities.

REFERENCES

Saxegaard, E. C. A., Alper, E., Jonas, J., Weber, A., Marc, G., Kinda, T., Callegari, G., Shabunina, A., Schaechter, A., & Caceres, C. 2012. A toolkit for Assessing Fiscal Vulnerabilities and Risks in Advanced Economics. *IMF Working Paper* No. 2012/011

Bankowsky, K. & Marien F. 2017. "Euro Area Fiscal Stance."*European Central Bank, Occasional Paper Series*, 182.

Bailey, M. A., Strezhnev, A., & Voeten, E. 2015. Estimating Dynamic State Preferences from United Nations Voting Data. *Journal of Conflict Resolution*, 61(2), 430–456. https://doi.org/10.1177/0022002715595700

Blanchard, O. J. & Perotti, R. 2002. "An Empirical Characterization of the Dynamic Effects of Changes in Government Spending and Taxes on Output," *Quarterly Journal of Economics*, 117: 1329–68.

Bohn, Henning, 1998, The Behavior of U.S. Public Debt and Deficits*, Quarterly Journal of Economics*, August, Vol. 113, No. 3, pp. 949–963.

Carnot, N. 2014. "Evaluating Fiscal Policy. A Rule of Thumb." *European Commission, Economic Papers* 526.

Celasun, O. & Kang, J. S. 2006. On the Properties of Various Estimators for Fiscal Reaction Functions. *IMF Working Paper* 06/182

Delong, J. B. & Summers, L. H.. 2012. "Fiscal Policy in a Depressed Economy," *Brookings Papers on Economic Activity*.

Dreher, A., & Sturm, J. 2010. Do the IMF and the World Bank influence voting in the UN General Assembly? *Public Choice*, 151(1–2), 363–397. https://doi.org/10.1007/s11127-010-9750-2

Dreher, A., Lang, V., Rosendorff, B. P., & Vreeland, J. R. 2021. Bilateral or multilateral? International financial flows and the Dirty-Work hypothesis. *The Journal of Politics*, 84(4), 1932–1946. https://doi.org/10.1086/718356

Eyraud, L., Baum, A., Hodge, A., Jarmuzek, M., Ture, H. E., Mbaye, S., & Kim, Y. 2018. How to Calibrate Fiscal Rules: A Primer. *IMF How-To Notes* No. 2018/002

Fournier, J. M. & Lieberknecht, P. 2020. A Model-based Fiscal Taylor Rule and a Toolkit to Assess the Fiscal Space. *IMF Working Paper* No. 2020/033

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

International Monetary Fund. 2017, Fiscal Monitor 2017/1

International Monetary Fund. 2022. Guidance Note for Surveillance Under Article IV Consultations. Policy Paper No. 2022/029.

Jha, M., Qian, J., Weber, M., & Yang, B. 2024. ChatGPT and Corporate Policies. *National Bureau of Economic Research*. https://doi.org/10.3386/w32161

Kanda, D. 2011. "Modeling Optimal Fiscal Consolidation Paths in a Selection of European Countries," *IMF Working Paper* No. 11/164.

Korinek, A. 2023. "Generative AI for Economic Research: Use Cases and Implications for Economists." *Journal of Economic Literature* 51(4):1281-1317.

Korinek, A. 2024. "LLMs Level Up—Better, Faster, Cheaper. June 2024 Update to Section 3 of "Generative AI for Economic Research: Use Cases and Implications for Economists." *Journal of Economic Literature,* June 2024.

Kose, M. A., Kurlat, S., Ohnsorge, F., & Sugawara, N. 2022. A cross-country database of fiscal space. *Journal of International Money and Finance*, 128, 102682. https://doi.org/10.1016/j.jimonfin.2022.102682

Krugman, P. R. 2015. St. Augustine and Secular Stagnation. *Paul Krugman's blog, The New York Times*.

Mauro, P., R. Romeu, A. Binder, and A. Zaman. 2015. "A Modern History of Fiscal Prudence and Profligacy." *Journal of Monetary Economics* 76: 55-70

6. APPENDIX

6.1 Staff Appraisal section in AIVs

The Article of Agreement, Board decisions, Board-approved policy papers, frameworks, and guidance from IMF management establish clear requirements for the Article IV consultation report, including the staff appraisal section. According to IMF (2022), the staff appraisal must summarize the staff's analysis, policy views, and recommendations, and may only address issues discussed both with the authorities and in the main report. It is also encouraged to integrate risk assessments, spillovers, policy mix, and trade-off analyses. Since the Article IV report is published following the Board's meeting, the staff appraisal—along with the rest of the report—becomes part of the public record, and the Board issues a press release noting the Executive Directors' endorsement of the appraisal.

Within this framework, the staff appraisal stands as the pivotal section for distilling key findings and guidance for the authorities, especially on policies affecting the member's balance of payments and domestic conditions. Although IMF (2022) does not explicitly mention the need to provide advice on short-term fiscal stance, given the centrality of fiscal policy in sustaining stability, advice on the short-term fiscal stance is essential. By evaluating the appropriateness of fiscal measures over the economic cycle, the appraisal would ensure policymakers can respond swiftly to emerging risks and seize opportunities for economic improvement.

6.2 Construction of macroeconomic variables

This section describes in detail how we constructed the macroeconomic variables to ensure that our analysis is reproducible.

We derived categorical data on the staff's recommended policy stance by inputting textual data into large language models (LLMs) and prompting them to classify the advice. To ensure that our analysis utilizes data contemporaneous with these recommendations, we collected fiscal sustainability series from the Market Access Countries' Debt Sustainability Analysis (MAC-DSA) and the Low-Income Countries' Debt Sustainability Framework (LIC-DSF) databases, both maintained by the IMF's Strategy, Policy, and Review Department. These databases, similar in nature to the World Economic Outlook (WEO) dataset, provide country-specific debt sustainability metrics drawn from various scenarios and projections, but lack a fixed release schedule and are published in vintages.

We then compiled macroeconomic variables from all available WEO vintages (January, April, July, and October) dating back to 1998. After retrieving these data internally, we matched each country-year pair for our AIV reports with the WEO vintage released immediately prior to the AIV's publication date. In case when there is no immediate prior vintage is available to the publication of an AIV, we use the latest available vintage retrospectively.We also constructed data series based on the latest available vintage of a

data source. This was done to test if a cyclical macroeconomic series constructed based on different vintages would yield vastly different results. As seen in section 7.7, the difference was minimal.

The dummy variable cex is constructed using the World Commodity Exporters Database, maintained by the IMF's Fiscal Affairs Department (FAD). This database compiles key macro-fiscal indicators for 52 countries that are significant exporters of oil, gas, and metals (such as copper, gold, iron, and silver), where these commodities comprise 20 percent or more of total exports or fiscal revenues. The FAD list differs from the one used in the April 2024 WEO Statistical Appendix Table D (employed in BP2 to classify commodity exporters). Although we have not conducted a detailed comparison of the two lists, we selected the FAD list as it includes additional countries, such as Australia, Canada, Colombia, Indonesia, Norway, and Russia, which are absent from the WEO list. On the other hand, the WEO list used in BP2 contains countries such as Afghanistan, Burkina Faso, Central African Republic, Eritrea, Liberia, Marshall Islands, Turkmenistan, which are absent from the FAD list.

6.3 Proportional odds assumption

The proportional odds assumption is a key assumption for ordinal logistic regression. The model is designed to handle ordinal dependent variables, where the outcome has an order but the distances between categories are not known. It assumes that the odds of being in a higher category (as opposed to all lower categories combined) are the same across all thresholds. In other words, the proportional odds assumption implies that the effect of each predictor is consistent across the different levels of the ordinal outcome. The effect of the independent variables on the odds of moving from "Loosen" to "Neutral or Tighten" is the same as the effect of the independent variables on the odds of moving from "Loosen or Neutral" to "Tighten."

$$\log\left(\frac{P(y \leq j)}{1 - P(y \leq j)}\right) = \tau_j - \mathbf{X}\boldsymbol{\beta}$$

Where $P(y \leq j)$ is the probability that the ordinal outcome is less than or equal to category j, $\mathbf{X}\boldsymbol{\beta}$ is a linear combination of the predictors and coefficients, and τ_j are the threshold values that separate the ordinal categories.

Table A1. Results from Brant test

Test for	χ^2	df	p-value
Omnibus	9.45	5	0.09
fb_gdp	2.19	1	0.14
pb_gdp	0.01	1	0.91
du_gdp	1.6	1	0.21
ngap	1.05	1	0.31
bca_gdp	0.17	1	0.68

Table A1 shows the results from Brant test for proportional odds assumption. The omnibus test is an overall test that checks whether the proportional odds assumption holds for the model as a whole, rather

than for each individual variable. The result suggests the proportional odds model because the assumption holds for both the overall model and for individual variables.

6.4 Winsorization

To limit the influence of extreme values to our empirical analysis results, we use winsorization to reduce their effect by replacing those extreme values (both in the upper and lower tails) with the nearest values within 1st and 99th percentile thresholds. In other words, for a data series, the values below the 1st percentile are set equal to the value at the 1st percentile, and those above the 99th percentile are set equal to the 99th percentile value.

Table A.2 Summary Statistics of Winsorized Variables

Variables	Raw			Winsorized			# of obs winsorized
	min	mean	max	min	mean	max	n
ngap	-26.63	3.58	212.28	-10.19	3.25	29.34	44
ngap_r	-18.83	0.14	212.28	-9.94	-0.27	26.67	26
ngap_r_ha	-27.65	6.90	145.76	-12.09	6.70	25.10	44
fb_gdp	-59.74	-1.79	44.63	-16.28	-1.85	18.32	54
pb_gdp	-37.75	0.65	46.92	-12.67	0.63	19.03	48
du_gdp	-249.37	57.99	457.61	1.52	57.78	227.85	46
bca_gdp	-68.78	-1.20	314.91	-29.78	-1.55	33.66	54
ds_ggr	-2.55	29.75	332.54	0.17	29.32	177.42	36
dugfn_gdp	-35.18	9.31	411.68	-18.06	8.98	53.38	44
tt	-2.10	0.02	1.15	-0.73	0.02	0.84	50
5_yr_dev	-7.39	3.50	27.12	-2.78	3.48	12.42	32

6.5 A Primer on Random Forest Algorithm

[Here we would provide a short technical overview of Random Forest Algorithm]

6.6 Further Figures on LLM Classification of IMF Fiscal Advice

Figure A1. Short-term recommended fiscal stance, AEs

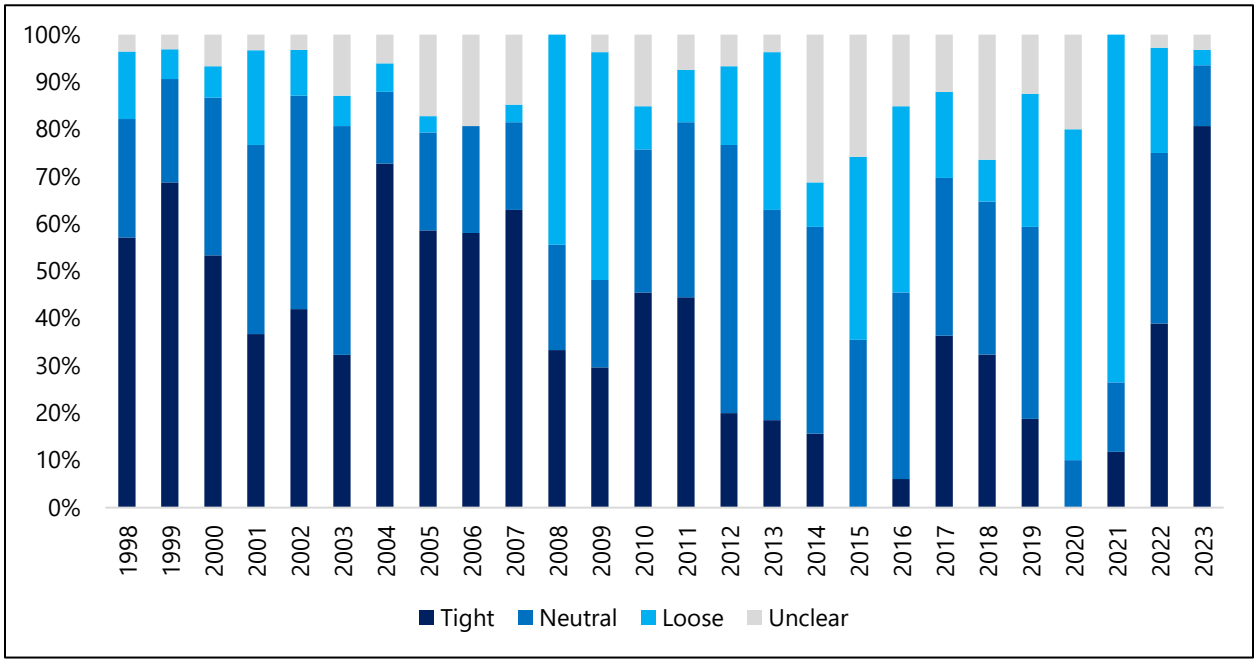


Figure A2. Short-term recommended fiscal stance, EMMIEs

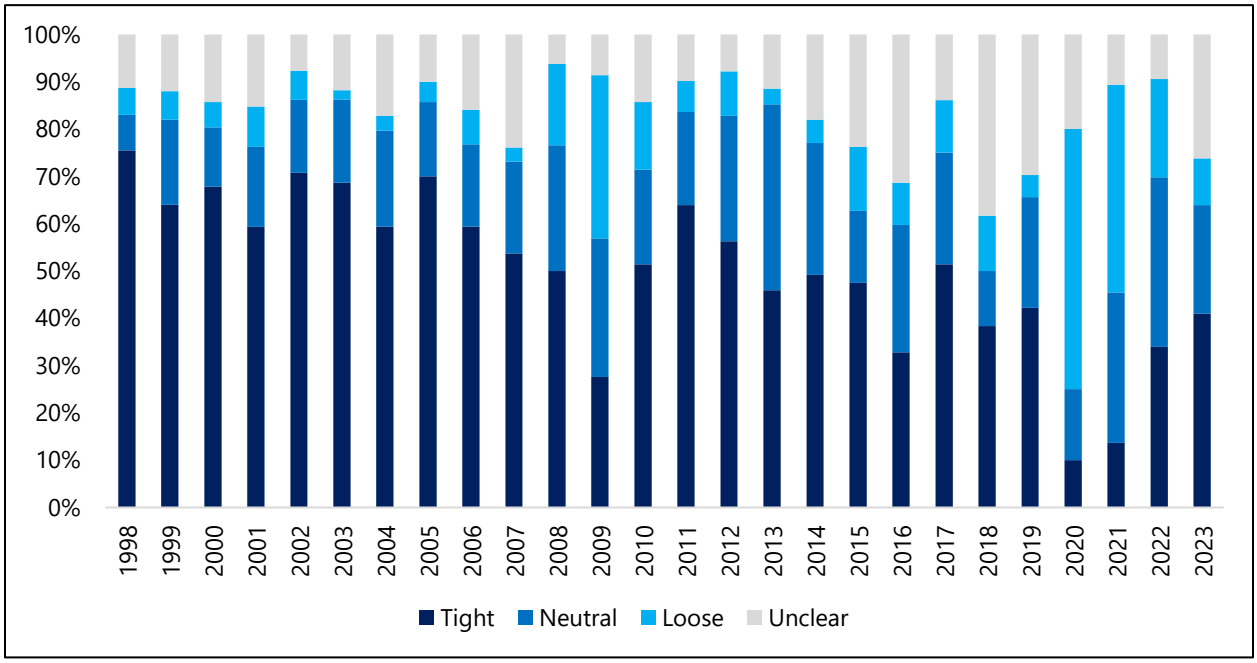


Figure A3. Short-term recommended fiscal stance, LICs

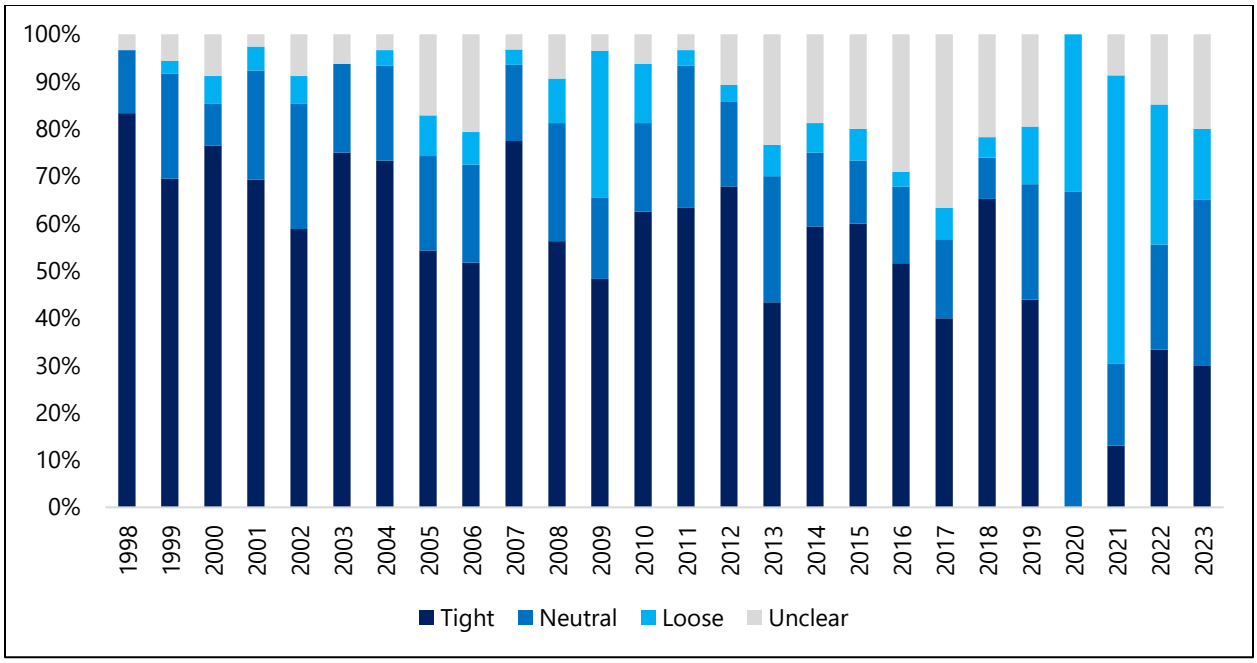


Figure A4. Medium-term recommended fiscal stance, AEs

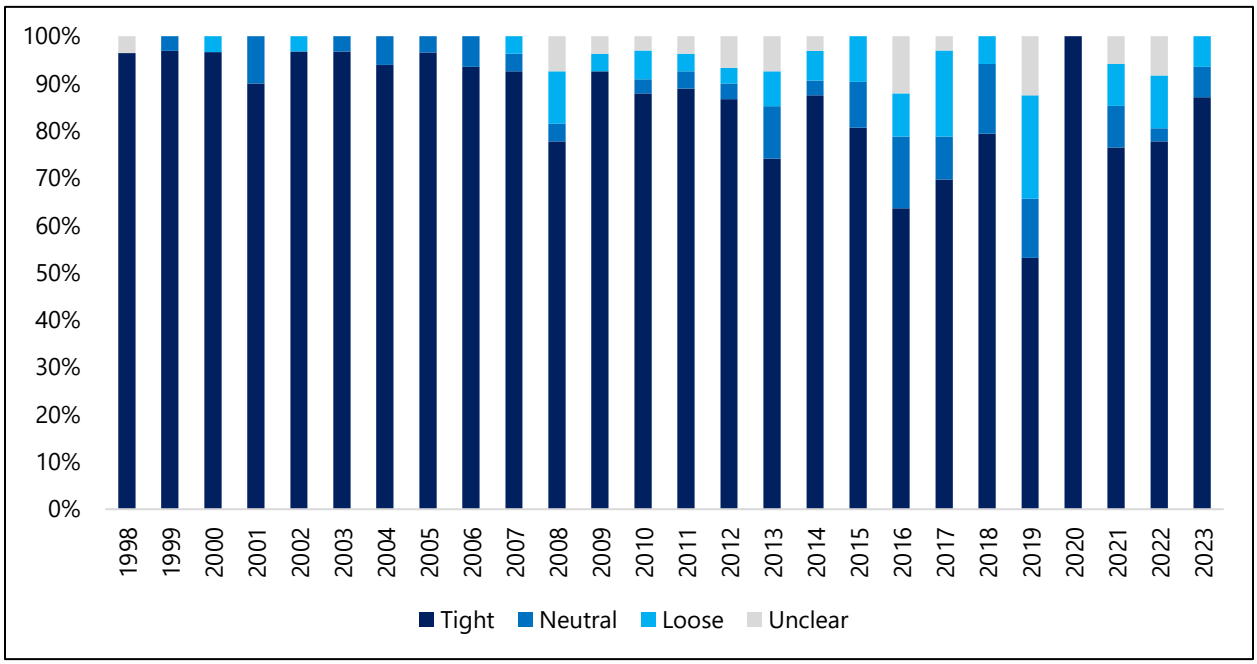


Figure A5. Medium-term recommended fiscal stance, EMMIEs

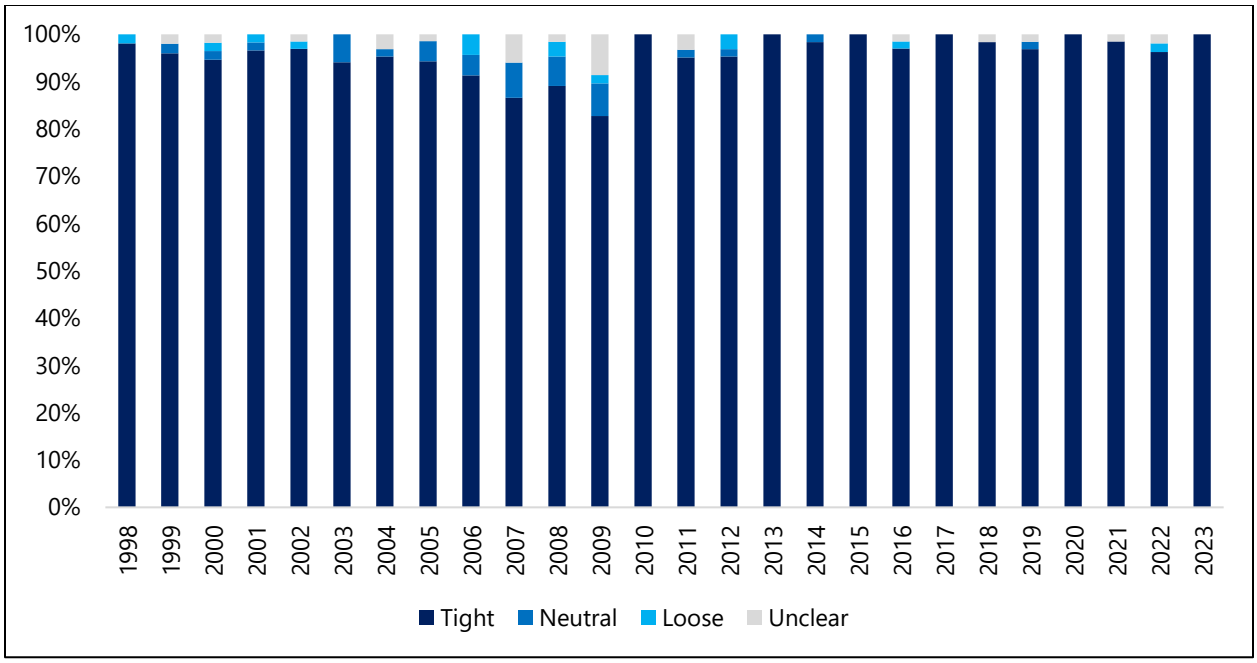
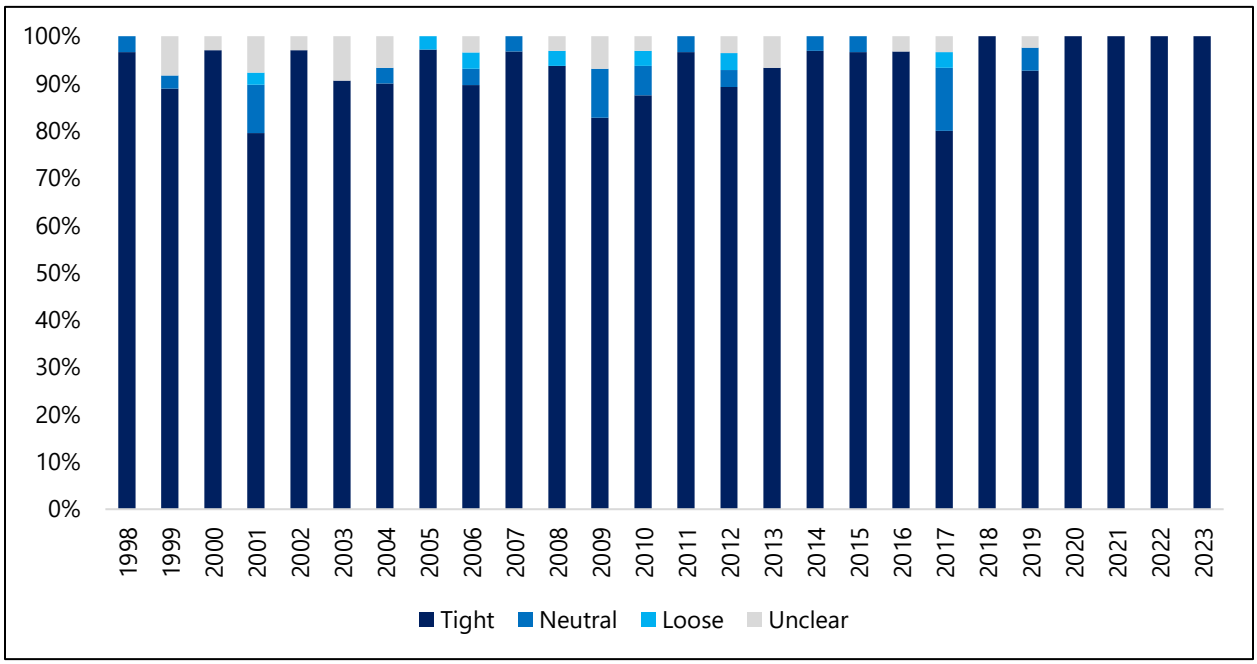


Figure A6. Medium-term recommended fiscal stance, LICs



6.7 Additional Regression Results

CORRELATION ANALYSIS

We also report the results of simple bilateral ordinal regressions of the variables contained in our main specification on fiscal advice in the Table below. As can be seen, the association of all the employed with the IMF staff's advice on fiscal policy is significant.

Table A.3 Short-term stance, specific vintages, individual predictors

	(1)	(2)	(3)	(4)	(5)
ngap	0.056***				
	(0.005)				
pb_gdp		0.038***			
		(0.007)			
du_gdp			0.004***		
			(0.001)		
bca_gdp				-0.041***	
				(0.002)	
fb_gdp					-0.028***
					(0.006)
N	2635	2326	2286	2607	2129

Note: Standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01. Weights included.

BASELINE REGRESSION WITH FIXED VINTAGES

Table A4. Short-term stance, fixed vintages.

	(1)	(2) Country FE	(3) Weights	(4) Weights + Country FE
ngap	0.052***	0.078***	0.055***	0.084***
	(0.005)	(0.010)	(0.006)	(0.012)
pb_gdp	0.043***	0.037***	0.052***	0.046***
	(0.008)	(0.013)	(0.009)	(0.013)
fb_gdp	-0.023***	0.002	-0.030****	-0.006
	(0.008)	(0.013)	(0.009)	(0.015)
du_gdp	0.001	-0.008	0.001	-0.001
	(0.001)	(0.002)	(0.001)	(0.002)
bca_gdp	-0.041***	-0.038***	-0.044***	-0.039***
	(0.004)	(0.007)	(0.004)	(0.009)
N	2035	2035	2024	2024

Note: Standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01.

The results in the table are based on data series constructed using fixed vintages. Fixed vintage means using the variable value from the April 2024 WEO vintage consistently across all years for all WEO series included in the regression. For DSA-DSF variables, the values are selected from the latest available issue date specific to that country.

INSTITUTIONAL QUALITY

Table A5. Short-term stance, institutional quality

ngap	0.039***
	(0.012)
polity_score	0.172
	(0.126)
bureacracy	-0.482***
	(0.090)
fb_gdp	-0.042**
	(0.021)
pb_gdp	0.045**
	(0.020)
du_gdp	-0.001
	(0.002)
bca_gdp	-0.030***
	(0.010)
N	1139

Note: Standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01. Weights included.

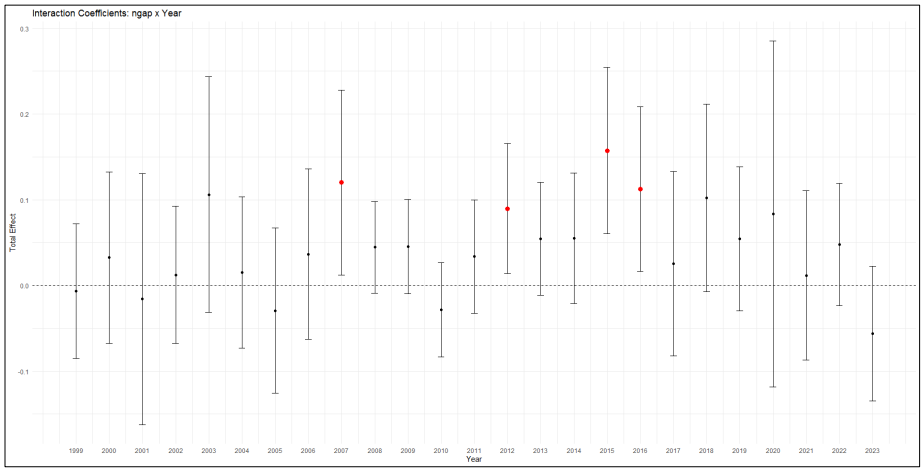
Polity Score[6] (democracy vs. autocracy) does not have a significant effect on fiscal stance recommendations, indicating that the IMF does not weigh regime type heavily in its fiscal advice. The quality of bureaucracy, on the other hand, has a strong and significant effect, with better bureaucratic quality reducing the likelihood of fiscal tightening recommendations.

[6] Polity Score: single regime score, runs from 1 (full democracy) to -1 (full autocracy)

OUTPUT GAP INTERACTION WITH YEAR DUMMIES

Figure A7 presents an interaction plot of ngap × Year, illustrating the generally positive relationship between the output gap and IMF fiscal stance recommendations over time. This relationship, however, exhibits fluctuations across years. In the first bucket, from 1998 to 2002, the IMF's advice did not appear to be counter-cyclical. Thereafter it has generally been significantly counter-cyclical. This may reflect an evolution in the IMF staff's reaction function to macroeconomic conditions.

Figure A7. Short-term stance, output gap interacted with year dummies



Sources: IEO calculations.

Note: Significant terms at 5% level are market with red dots. Weights included, but no controls. The plot shows total effect with 95% CI bands.

PROBABILITY PLOTS

Figure A7. Probability plot, short-term stance, primary balance, income grouping interaction

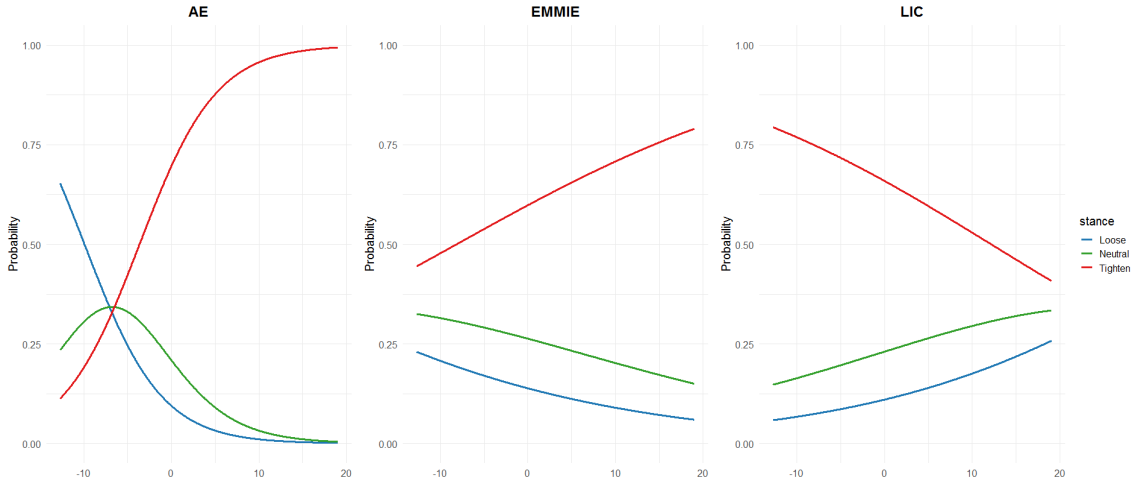


Figure A8. Probability plot, short-term stance, fiscal balance, income grouping interaction

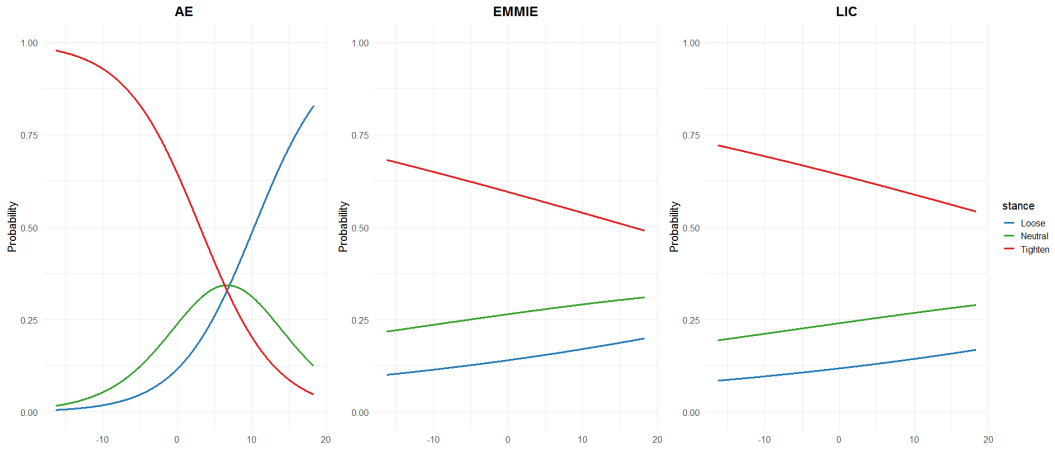


Figure A9. Probability plot, short-term stance, debt-to-GDP, income grouping interaction

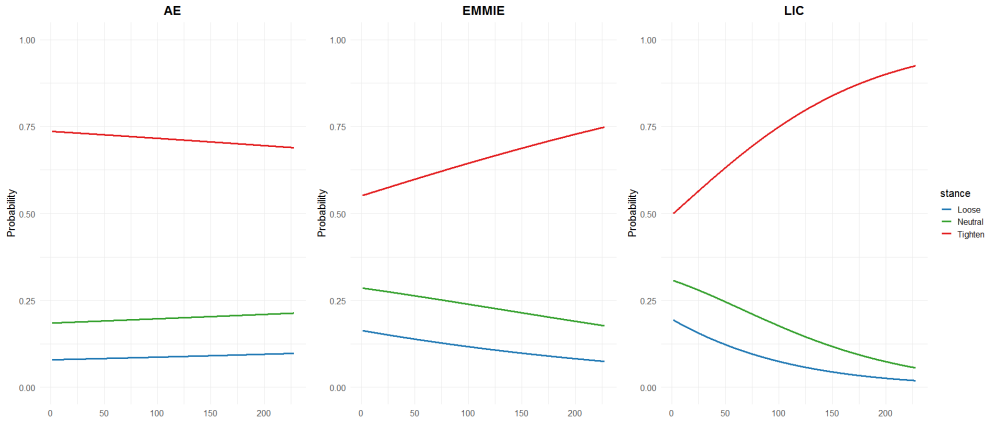


Figure A10. Probability plot, short-term stance, current account as % of GDP, income grouping interaction

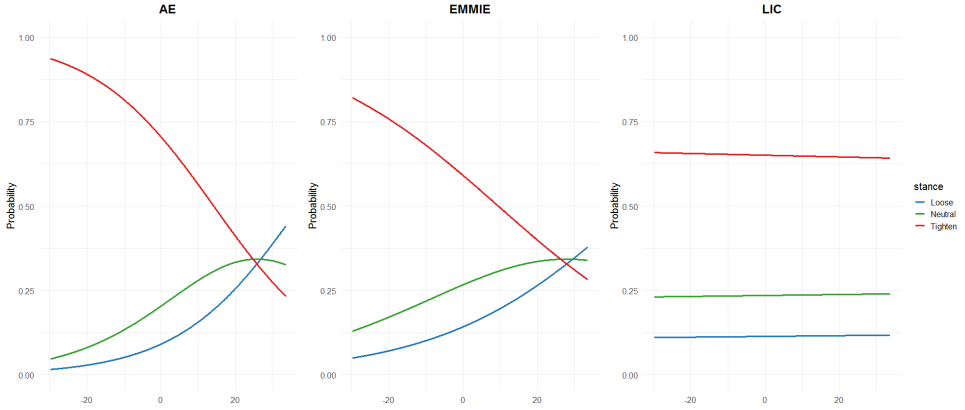


Figure A11. Probability plot, short-term stance, fiscal balance

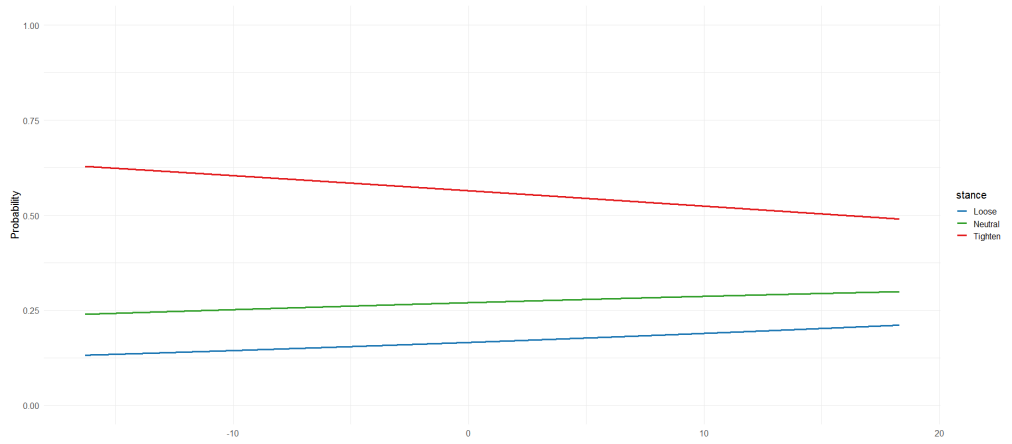


Figure A12. Probability plot, short-term stance, primary balance

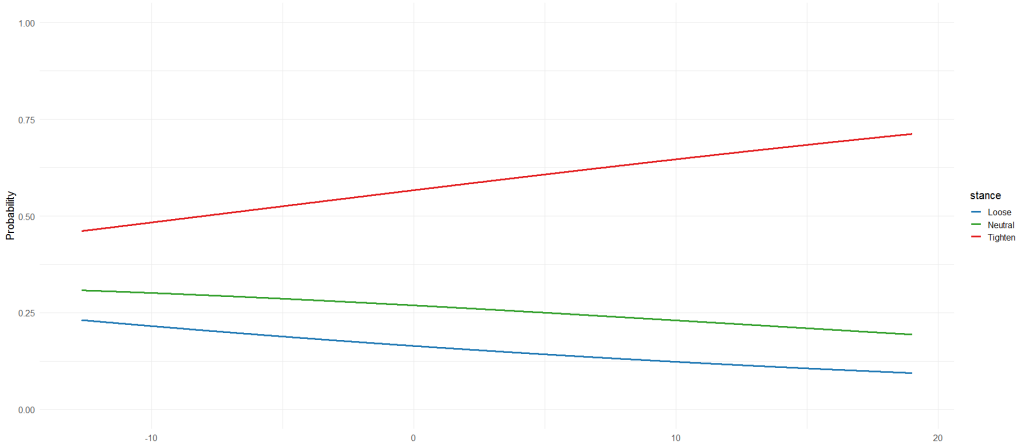


Figure A13. Probability plot, short-term stance, debt to GDP

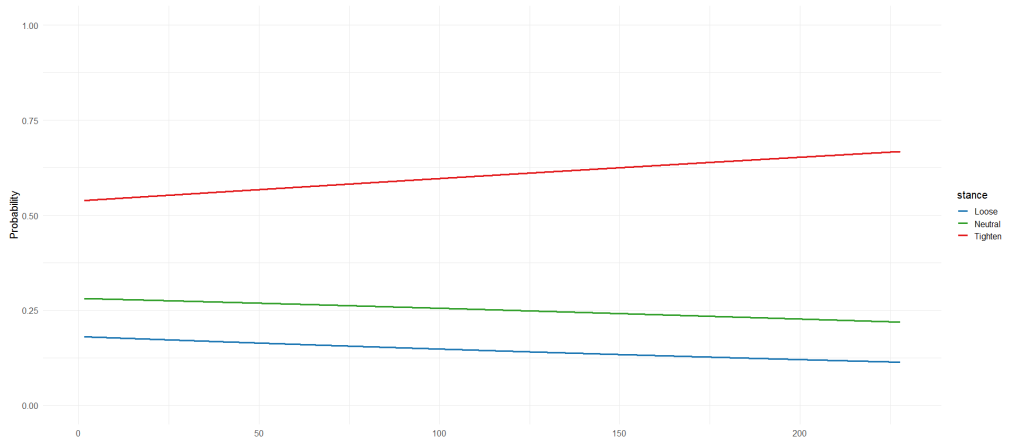
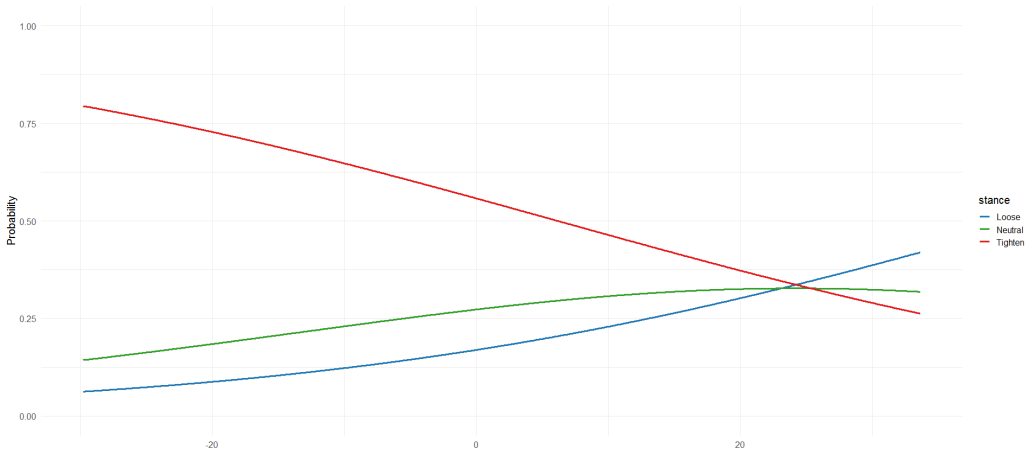


Figure A14. Probability plot, short-term stance, current account


1.00
0.75
0.50
0.25
0.00
Probability
-20
0
20
stance
Loose
Neutral
Tighten