INTERNATIONAL MONETARY FUND

How Effectively Can Current LLMs Analyze Macrofinancial Issues?

Paola Ganum and Tohid Atashbar

WP/26/35

***IMF Working Papers* describe research in progress by the author(s) and are published to elicit comments and to encourage debate.**
The views expressed in IMF Working Papers are those of the author(s) and do not necessarily represent the views of the IMF, its Executive Board, or IMF management.

2026
FEB

WORKING PAPER

WP/26/35

IMF Working Paper
Strategy, Policy, and Review

How Effectively Can Current LLMs Analyze Macrofinancial Issues?
Prepared by Paola Ganum and Tohid Atashbar*

Authorized for distribution by Eugenio Cerutti
February 2026

***IMF Working Papers* describe research in progress by the author(s) and are published to elicit comments and to encourage debate.** The views expressed in IMF Working Papers are those of the author(s) and do not necessarily represent the views of the IMF, its Executive Board, or IMF management.

ABSTRACT: This paper empirically evaluates the ability of current Large Language Models (LLMs) to analyze macrofinancial coverage in IMF Article IV staff reports, using human economists' assessments as a benchmark. We test several GPT models on reports from 2016-2024, assessing their performance on both qualitative ratings and binary questions. Our findings indicate that the latest models can meaningfully assist economists, achieving an average accuracy of 71-75% on ratings and an average exact match rate of 76-81% on binary questions in 2024 across advanced GPT models. However, we find that LLMs tend to assign higher, less-dispersed ratings than human experts and struggle with open-ended questions that require deep contextual judgment. The paper provides quantitative evidence on current LLM accuracy in this domain, explores the drivers of its performance, and discusses key limitations such as optimistic bias.

RECOMMENDED CITATION: Ganum, Paola, and Tohid Atashbar. 2026. "How Effectively Can Current LLMs Analyze Macrofinancial Issues?" IMF Working Paper WP/26/35, International Monetary Fund, Washington, D.C.

JEL Classification Numbers:	E44, G01, C55, C81
Keywords:	AI, Large Language Model, Textual Analysis, Macrofinancial Surveillance, IMF Staff Reports, Human-AI Comparison
Author's E-Mail Address:	gganum@imf.org; tatashbar@imf.org

* The views expressed in this paper are those of the authors and should not be attributed to the IMF, its Executive Board, or IMF Management. We would like to thank Eugenio Cerutti, Nicolas Arregui, and Baoping Shang for invaluable guidance, Adeleke Adeyemi for outstanding programming support, and Wanning Shu for excellent research assistance. We also thank IMF seminar participants for helpful comments. All errors are ours.

WORKING PAPERS

How Effectively Can Current LLMs Analyze Macrofinancial Issues?

Prepared by Paola Ganum and Tohid Atashbar[1]

[1] “The views expressed in this paper are those of the authors and should not be attributed to the IMF, its Executive Board, or IMF Management. We would like to thank Eugenio Cerutti for invaluable guidance, Adeleke Adeyemi for outstanding programming support, and Wanning Shu for excellent research assistance. We also thank IMF seminar participants for helpful comments. All errors are ours.

Contents

Section I. Introduction

The rapid development of Generative Artificial Intelligence (gen-AI) tools, and in particular the public release of Large Language Models (LLMs) exemplified by ChatGPT, Claude, and Gemini, have marked an inflection point. This technology promises to be productivity-enhancing, particularly for occupations where it can be complementary to human work (IMF 2024). These LLM models offer unparalleled abilities to help economic research by conducting literature reviews, analyzing data, writing code, and preparing manuscripts (Korinek, 2023).

In this paper, we empirically evaluate the ability of current Generative Pre-trained Transformer (GPT) models to review the coverage of macrofinancial issues in Article IV staff reports, using human economists' assessment as a benchmark. Since the Global Financial Crisis, the IMF has taken substantial steps in strengthening macrofinancial surveillance and analysis (IMF 2017)[1]. In this paper, we would like to explore how well different LLMs could review macrofinancial coverage in Article IV staff reports if put to the task.

While the LLMs are effective at textual tasks, analyzing macrofinancial coverage in staff reports requires technical knowledge, reasoning, and judgement, something the human mind with the appropriate technical expertise can do but it may be more challenging for LLMs to achieve.

Our research involves feeding staff reports one at a time in a standardized pdf format to an LLM and asking the model to answer a set of questions on it and producing an excel file for each report individually with its answers. We tried different GPT models, starting with the GPT-4o, we then moved on to more advanced models like GPT-4.1 (we also tried GPT-4.1-mini), GPT-o1, and GPT-5 (medium and high effort). Our primary interest is to understand the LLM's ability to answer questions and assign ratings while comparing it to the economists' answers in carrying out this task. But, more generally, we use this analysis to answer the following questions: (i) how accurate are LLMs relative to the human economist benchmark, (ii) what are the country/report characteristics that are associated with higher LLM and human ratings; (iii) how do different GPT models perform in this task; and (iv) what factors increase the likelihood of a match between human and LLM answers.

In pursuing this evaluation, this paper generates a novel dataset of LLM answers to explore how well the LLM can perform the review of macrofinancial coverage in staff reports. We will call "accuracy" (see section III for its definition) the LLM performance relative to the human. To our knowledge, this represents a novel systematic comparison in this specific domain, offering the first quantitative evidence on LLM accuracy for this granular assessment task, which constitutes a novel contribution to the related literature.

A key observation from this paper is that certain GPT models appear to achieve useful -but still limited- levels of accuracy in analyzing IMF staff reports, particularly on structured, fact-based questions. We found that GPT - o1[2], GPT-4.1, and GPT-5 offer an improvement over GPT-4o, with improved justifications and more agreement with economists' assessments.[3] Our findings indicate that the LLM can meaningfully assist economists in pursuing this assessment, with an average accuracy[4] across models of 71-75% on ratings in 2024 (44-74% in

[1] Approaches to Macrofinancial Surveillance in Article IV Reports; IMF Policy Paper

[2] This model is better at reasoning problems. See Introducing OpenAI o1 | OpenAI

[3] GPT- 4.1-mini was also tested for 2024 staff reports, but results resembled those of GPT-4o, comparing unfavorably to those of the more advanced models.

[4] As defined in Section III, accuracy measures the proportion of correct predictions (consisting of the set of true positives and true negatives) by the LLM out of all LLM predictions. A true positive, for instance, is a report that both the human and the LLM rated highly.

2022-23) and an average exact match rate of 76-81% on binary questions in 2024 (72-79% in 2022-23). This finding is consistent with OpenAI's description[5] of the improvement achieved in the GPT-o1 model from GPT-4o and suggests that model selection could be an important factor that may require periodic re-evaluation as new models become available. The newer models excel over their predecessor in performance, particularly on instruction following, reasoning ability, and long context aiding comprehension and ability to extract insights from large documents. These results could suggest that LLMs might be able to replicate a portion of the manual review process, possibly offering some efficiency gains.

We also found that LLM struggles mostly with open-ended questions and that the probability of a match between human and LLM answers is lower in complex and in ratings questions. We noticed a significant improvement in accuracy of LLM answers after introducing prompt refinements to our first prompt, with specific examples of economists' justifications and low/high ratings, as well as more advanced models (GPT-4.1, GPT-5, and GPT-o1). The distribution of ratings with the more advanced model GPT-4.1 and reasoning-capable models such as GPT-o1 and GPT-5 also appeared to align more closely with human assessments, although a tendency toward optimism remained. The distribution seems less compressed at the highest end of the scale compared to the GPT-4o results. This could suggest that the more advanced models, combined with a better prompt, might be slightly more discerning in their assessments. The source of this apparent optimistic tendency could be multifaceted stemming from the models' underlying training data or from the alignment process used by developers to make models helpful.

Beyond accuracy, we asked the LLM to produce a confidence score of its own answer to better distinguish between high-certainty and low-certainty outputs. LLMs show relatively high (self-assessed) confidence levels around 81%-88% in 2024, and confidence is the highest in questions asking about issues typically covered in staff reports such as banking sector issues, regulation and supervision policies, financial integrity issues, and identifying whether financial sector indicators (FSIs) were used, and only slightly lower around open-ended questions (where exact match rates with humans are lower).

Given the probabilistic nature of LLMs, a concern might exist that the same prompt and model could yield different results on subsequent runs. While perfect one-to-one identity is not always expected, our experience in this study indicated a high degree of practical consistency. In repeated tests using the same model, prompt, and temperature settings[6], we observed that the outputs were very similar, with 93% consistency in binary answers and 83% accuracy in ratings answers. This level of stability could suggest that for the purpose of a large-scale review exercise, the results may be sufficiently reproducible to be useful. The next section will delve into the related literature. Section III will describe our data sources and methodology. Section IV will present our main findings. Section V discusses results, LLM issues, and possible avenues for future research. Section VI concludes.

Section II. Related Literature

This paper contributes to several related literatures. The first and most direct is the rapidly growing field of studies evaluating the capabilities of AI, and specifically LLMs, to perform complex economic and financial analysis. Within this, our work is situated among studies that compare LLM-produced outputs against human benchmarks to gauge their accuracy, reliability, and limitations. The second related literature consists of the extensive macrofinancial literature. Our work bridges these two areas by applying AI techniques to the

[5] Introducing GPT-4.1 in the API | OpenAI

[6] Temperature is a parameter that controls the degree of randomness in a model's responses: lower temperature settings produce more deterministic and repeatable outputs, while higher settings allow for greater variability and creativity.

established practice of macrofinancial surveillance, demonstrating how LLMs can be used to systematically analyze the qualitative information embedded in official policy documents like the IMF's Article IV reports.

The application of textual analysis in economics is not new. Early methods relied on dictionary-based approaches, creating word lists to measure sentiment, policy uncertainty, or other concepts (Baker, Bloom, and Davis, 2016). While foundational, these methods can be rigid and miss the nuance and context inherent in human language. The development of Natural Language Processing (NLP) introduced more sophisticated models like BERT (Devlin et al., 2019), which could understand context. However, the recent advent of generative LLMs like the GPT (Generative Pre-trained Transformer) family represents a paradigm shift, moving from simple classification to complex generation, summarization, and reasoning tasks. Our paper leverages these state-of-the-art models to go beyond what was previously possible with older NLP techniques. For a comprehensive review of generative AI applications in economic research, see Korinek (2023).

I. LLMs in Analyzing Central Bank Communications

A prominent application of LLMs in macrofinancial contexts is analyzing central bank communications. Central bank statements are carefully crafted, dense with meaning, and have significant market-moving potential, making them an ideal test case for an LLM's ability to "read between the lines." One of the most cited studies in this area, Hansen and Kazinnik (2023), examines whether GPT models can "decipher Fedspeak"—the technical language used by the U.S. Federal Reserve. They evaluate GPT-based models on classifying the monetary policy stance (hawkish vs. dovish) in Federal Open Market Committee (FOMC) announcements, using human expert assessments as a benchmark. Their findings are striking: GPT models achieved substantially higher classification accuracy than prior NLP methods (e.g., BERT or dictionary-based approaches). Notably, the latest GPT-4 model could also generate text-based explanations for its classifications that were judged to be on par with the reasoning of human economists. Furthermore, the same study showed that GPT-4 can help identify monetary policy shocks in historical transcripts using the narrative approach pioneered by Romer and Romer (1989), a task that traditionally requires intensive human reading and judgment. Recent work has extended this to broader central bank analysis. For instance, Christiano Silva et al. (2025) use a fine-tuned LLM to classify sentences from over 75,000 central bank documents across 169 institutions, extracting insights on topics, stance, sentiment, and audience. This approach reveals how LLMs can quantify policy shifts and uncertainty at scale, aligning with our use of LLMs for structured analysis of IMF reports. Our paper adopts a similar philosophy by providing the LLM with detailed definitions of macrofinancial coverage and a structured questionnaire, guiding its analysis of IMF reports.

II. LLMs in Macroeconomic Forecasting

Another major research area is the use of LLMs for macroeconomic forecasting. Traditional forecasting models are typically econometric, relying on structured numerical data. LLMs, however, offer the potential to extract forward-looking information from the vast universe of unstructured text, such as news articles, policy statements, and corporate filings. A key question is whether this textual information contains predictive power beyond what is already captured in standard economic indicators. A 2025 comparative analysis by Carriero, Pettenuzzo, and Shekhar used the FRED-MD database to test LLMs against traditional vector autoregression (VAR) models. They found that LLMs can capture intricate patterns in data but have limitations in certain forecasting scenarios compared to traditional methods. This suggests that while LLMs can identify patterns, they have inherent constraints in replicating economists' deep causal reasoning processes, a limitation we also explore in our analysis of IMF reports. One key reason for this is that LLMs are pre-trained by developers on vast but general datasets, which may not contain the specific domain knowledge needed for specialized economic analysis. In a different approach, Chen et al. (2025) investigate whether models like ChatGPT and

DeepSeek can predict the stock market and macroeconomy. They use these models on news articles and find that they can generate useful signals for predicting stock returns. Their work suggests that LLMs are more effective tools for textual analysis compared to both traditional word-based methods and number-based technical analysis. Complementing this, Chen et al. (2025) evaluate LLMs' macroeconomic knowledge by testing their ability to recall historical variables and data release dates, finding that while LLMs excel in pattern recognition, they struggle with precise factual recall in specialized domains.

III. LLMs in Financial Statement and Corporate Policy Analysis

Beyond macroeconomic applications, LLMs are increasingly being used to analyze firm-level data, particularly financial statements and the transcripts of earnings calls. This literature is closely related to our work, as it involves asking an LLM to "read" a detailed report and extract specific, structured information—a task very similar to our review of Article IV reports. In a study close to our work, Jha et al. (2025) prompted ChatGPT to generate an "investment score" based on the introductory remarks of corporate earnings calls. They found that the LLM-generated score was a significant predictor of future returns, though it often failed to capture critical forward-looking nuances discussed in the Q&A portion of the calls—a limitation that human analysts were better able to navigate. Moving from analysis to generation, Fang, Jia, Li, and Lu (2025) use LLMs to analyze Chinese government documents, identifying and structuring industrial policies. They create a detailed, structured dataset capturing policy objectives, tools, and targets. By combining this with firm-level data, they document how policy choices evolve and affect firm productivity. This study showcases the power of LLMs for conducting large-scale policy analysis that would be impossible for humans to perform manually and provides a compelling use case for their application in institutional settings like the IMF.

IV. LLMs for Measuring Sentiment, Risk, Financial Stability, and Cross-border Flows

Finally, a growing body of work uses LLMs to construct novel measures of sentiment and risk from textual sources, which are key components of macrofinancial analysis. Chen, Kelly, and Xiu (2022) use LLMs to extract contextualized representations from news articles to predict stock returns, finding that these advanced models are more effective than traditional methods. More directly related to financial stability, studies like those by the Bank for International Settlements (Kwon et al., 2024) provide a primer for economists on how LLMs can be used to monitor financial markets in real-time by processing news flows and analyst reports to detect emerging risks. Bergant et al. (2026) construct a new high-frequency multi-dimensional dataset of *de jure* cross-border flow restrictions since the 1950s by applying LLMs to IMF AREAER official documents, thus utilizing LLMs to classify and organize large volumes of text in a consistent way. This aligns with our paper's objective of using LLMs to perform a granular, risk-focused assessment of IMF reports.

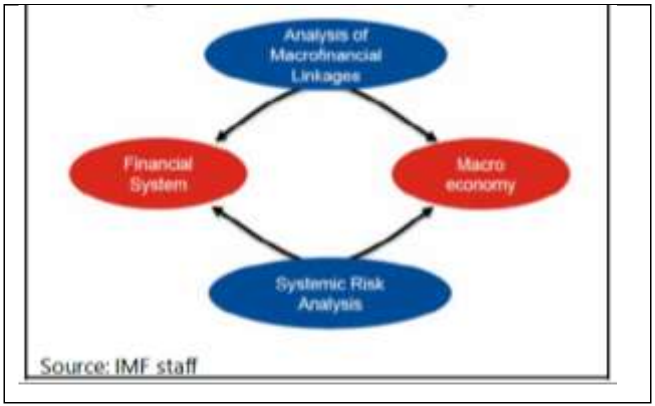
Our paper builds on and extends these four strands of literature. While most existing studies focus on a single type of document (e.g., central bank statements, news headlines, or corporate filings), we apply LLMs to complex, long-form, and semi-structured documents—IMF staff reports—that synthesize macroeconomic, financial, and policy analysis. Furthermore, unlike studies focused purely on classification or prediction, we conduct a comprehensive evaluation of an LLM's ability to perform a real-world, multi-faceted analytical task that involves a combination of data extraction, qualitative rating, and justification, using human evaluation as a robust benchmark. Consequently, the analytical task presented in this study differs significantly from the classification or extraction tasks common in the existing economic literature. Unlike sentiment analysis, which focuses on the tone of specific segments, or forecasting, which relies on pattern recognition in data series, the macrofinancial evaluation requires a holistic and subjective assessment of logical consistency. An evaluator must trace narrative threads across the entire document, for instance, determining whether the systemic risks identified in the 'Risk Assessment' section are adequately integrated into the 'Policy Advice' section later in the report. This requires synthesizing information dispersed across long contexts and applying judgment to evaluate the quality of the analysis, rather than simply verifying the presence of specific keywords or facts.

Section III. Data, GPT Models, and Methodology

Data

Our main dataset consists of staff reports published between 2016-2024, for which a human expert rating is available. Under Article IV of the IMF's Articles of Agreement, the IMF holds bilateral discussions with members, usually every year, on economic developments and current and planned policies. These discussions are presented to the Executive Board for consideration in a staff report. The report, together with the Executive Board's views are published afterwards with the member's consent. One of the main elements of Article IV staff reports is the identification of key macrofinancial linkages. It should present a view of systemic risks, consider feedback effects between financial conditions and the economy, and discuss appropriate policy recommendations such as macroprudential policies to address the identified vulnerabilities (per the 2022 Guidance Note for Surveillance Under Article IV Consultations) (text figure 1).

Text Figure 1. Integration of macrofinancial analysis in staff reports



Source: IMF staff

In this study, we put different LLMs to the task of answering a questionnaire assessing the degree of macrofinancial coverage in a given staff report. Using Article IV staff reports (SRs) over 2016-2024 (except during the pandemic years 2020-21), we fed the universe of published staff reports to the LLMs using refined prompts, including definitions of macrofinancial coverage (Table 1). Given the volume of reports and the granularity of the analytical task, a manual approach was infeasible, so we developed a programmatic workflow to automate the analysis. This was achieved using an Application Programming Interface (API) to interact directly with a GPT model hosted on the Microsoft Azure platform, enabling the systematic processing of each report. We gave instructions to the LLM, using a range of GPT models (see Box 1, we tested GPT–4o, GPT-o1, GPT-4.1, and GPT-5), to give a rating from 1 to 4 (where intermediate ratings in 0.1 increments are possible) and provide a justification to detailed questions about how well the report covers the following three areas : (i) macrofinancial analysis and coverage in the baseline, (ii) assessment of systemic risks and vulnerabilities, and (iii) mapping from risks and vulnerabilities to policy advice. A group of economists answered the exact same questionnaire. We also instructed the LLMs to answer 40 binary (Yes/No) questions, allowing us to compare its accuracy on these binary questions alongside the more nuanced qualitative ratings. Although the human-benchmarked binary questions are available only over more recent years only (2022-2024). We included definitions of macrofinancial linkages and gave some guidance to the LLM of what to look for in each question. We also provided a prompt to the LLM giving general instructions about how to perform the review and generate an excel file with a certain structure. The prompt used was a standard one, without giving context to the LLM about its role, simply asking “your task is to read the pdf staff report”. This is a novel contribution to the related literature both in its application to staff reports and in the creation of a novel LLM-answers dataset.

Table 1. Number of reports by year and Income group

Income Group	2016	2017	2018	2019	2022	2023	2024	Total
AE	10	20	17	18	25	34	34	**158**
EM	15	31	36	30	32	54	61	**259**
LIC	10	8	10	16	14	30	38	**126**
Total	**35**	**59**	**63**	**64**	**71**	**118**	**133**	**543**

Box 1. Differences between the models

GPT-4o (omni). GPT-4o, when released in May 2024, was considered a major leap forward in multimodal capabilities across text, voice, and vision. According to OpenAI it natively ingests text, images, audio, and video, aiming for human-like response times. It serves as a good and versatile general model with instruction-following but without an explicit, slower “reasoning mode.” These design goals—speed and natural interaction—align with the model’s positioning in our study as a capable generalist rather than a deep-reasoner.[7]

GPT-4 (mini). Released in July 2024, this model is designed as a cost-efficient, lightweight alternative to the flagship models. It prioritizes speed and affordability while retaining the same context window and multimodal capabilities as GPT-4o. However, as a smaller model, it is generally less capable in complex reasoning tasks compared to its larger counterparts. This aligns with our findings where its performance mirrored the general tendencies of GPT-4o, proving less effective at the nuanced macrofinancial evaluation than the reasoning-optimized or long-context models.

OpenAI o1 (reasoning-optimized). The o-series (o1, o1-mini) is not simply a bigger model but rather it is trained to *think before answering*, using additional computational resources to work through multi-step problems. This deliberate step improves complex reasoning and debugging but costs time and tokens (its internal “reasoning” also consumes context). These properties plausibly explain why o1 outperforms on our harder rating task despite not being the fastest model. [8]

GPT-4.1 (long-context). GPT-4.1, released in April 2025, emphasizes long-context comprehension, instruction following, and tool use. The API version supports up to 1 million tokens and is explicitly described as better at *using* that large context. This is relevant for our task, which feeds full Article IV PDFs: 4.1 is built to digest long documents with fewer misses from chunking. [9]

GPT-5 (unified router + deeper mode). GPT-5, released in August 2025, is presented as a *unified* system that knows when to respond quickly and when to think longer when prompts are complex, or when the user asks it to “think hard.” [10]

[7] OpenAI. “Hello GPT-4o.” Multimodal, low-latency design and capabilities. https://openai.com/index/hello-gpt-4o/?utm_source=chatgpt.com

[8] OpenAI. “Learning to reason with LLMs / Introducing o1.” Deliberate reasoning before answering.

[9] OpenAI. “Introducing GPT-4.1 in the API / Models: GPT-4.1.” 1M-token context and long-context reliability.

[10] OpenAI. “Introducing GPT-5 / GPT-5 System Card / GPT-5 in ChatGPT.” Unified router and Thinking mode

Methodology I. Metrics

We compared the LLM outputs to the human outputs in different ways, including estimating the accuracy, precision, recall, F1 score, exact match, and near match[11]. To grasp the degree of divergence between LLM and human answer, we estimated a near match, which allows a ± 1 difference in rating level. We categorized the differences in ranges of 0.5 points and show the percentage of reports whose ratings fall within these ranges. These results are presented in bar charts in the next section. We also present tables showing an exact match measure to check the percentage of reports where the LLM and human rating match exactly.

We define accuracy as the proportion of correct predictions made by the LLM out of all predictions. For rating questions, we use the following formula to estimate it.

Accuracy = (Number of correct predictions) / (Total number of predictions) eq. (1)
= (TP + TN) / (TP + TN + FP + FN)

Where:
- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

We categorized both human and LLM answers as positive (giving it a value of 1) if their rating is above or equal to 3 points and negative if below that threshold (giving it a value of 0). The number of reports classified as TP (TN) then would be when both answers were categorized as 1 (0). FP is when human answer is categorized as 0 and LLM as 1, and vice versa for FN. Since we have three main areas and an overall rating (average of the three), we estimate the accuracy in each of these. A correct prediction would be one that correctly identifies a true positive or a true negative, meaning a highly (low) rated report by human is also highly (low) rated by the LLM.

For binary questions, since there are no positives or negatives, we simply estimate it as equation (1), where the number of correct predictions mean an exact match between human and LLM answer. Therefore, when we mention the accuracy rate in the context of binary questions, we mean an exact match rate.

We define precision as the number of LLM's positive predictions that were actually correct. This is done only for rating questions, since binary questions do not have a positive/negative categorization.

Precision = TP / (TP + FP) eq. (2)

We define recall as the number of actual positive cases the LLM successfully caught. This is done only for rating questions, since binary questions don't have a positive/negative categorization.

[11] In selecting models, we focused on the GPT-family variants that are currently deployed and accessible to authors. These models differ not only in scale but also in architecture and training objective. Exact parameter sizes for these models are not publicly disclosed, so our comparison should be interpreted as a practical assessment of the main production models available to staff, rather than a size-controlled benchmark of architectures. As a limitation to our study, benchmarking against other proprietary frontier models (e.g., Claude, Gemini) or open-source solutions (such as Llama or Mistral) was not feasible due to infrastructure constraints but remains an important avenue for future research to assess cost-efficiency and performance trade-offs.

Recall = TP / (TP + FN) eq. (3)

We also use an F1 score, which balances precision and recall and it's useful to summarize both. This is done only for rating questions, since binary questions don't have a positive/negative categorization.

F1 = 2 * (Precision * Recall) / (Precision + Recall) eq. (4)

While accuracy provides a general sense of performance, it can be misleading if the classes in the dataset are unbalanced (i.e., if there is a disproportionately high number of positive ratings compared to negative ones). In such cases, a model could achieve high accuracy simply by predicting the majority class (optimistic bias) without truly distinguishing between reports. Therefore, we utilize these additional metrics to capture specific nuances of the model's reliability. Precision is critical in this context as it measures the model's trustworthiness—specifically, how often the LLM avoids 'hallucinating quality' (i.e., rating a weak report as high). Recall, conversely, measures the model's coverage, telling us if the LLM is failing to identify reports that are actually high-quality (False Negatives). The F1 Score provides the harmonic mean of the two, offering a single metric that penalizes extreme values in either direction, which is essential for comparing models that might prioritize caution (high precision) over coverage (high recall).

Methodology II. Correlations

We also run OLS regressions to estimate what characteristics are associated with higher LLM and human ratings. Since our sample is an unbalanced panel with gaps (given that we do not cover two COVID-related years), and the country sample varies over time, we opted for a pooled OLS specification with year fixed effects (FE).

$$\boldsymbol{Rating_{it} = \delta + \alpha_1 * X_{it} + \beta * macro_{it} + \alpha_2 * Y_i + \theta_t + e_{it}} \quad \text{eq. (5)}$$

Rating is either the human or the LLM overall average rating at time t for country report i. Xit is a set country report characteristics: a dummy capturing whether country i had a recent[12] Financial Sector Assessment Program (FSAP) at time t, a dummy capturing whether the report went through a specialized macrofinancial review during its writing, a dummy with the participation in the writing of the report by a macrofinancial economist, a dummy for whether the report has a well-articulated view of systemic risk in year t, and a variable which is the sum of binary values in year t[13]. These are all factors that, in our prior, should tend to increase the rating. We also include a dummy capturing whether the country had a program at time t, which in our prior would tend to decrease the rating given the number of program issues the report needs to cover and limited space in the report. We also have time-invariant dummies Yi: one dummy for each income group (AE, EM, and LIC) and a dummy capturing whether the country experienced a banking crisis[14]. The macro variables are the level of government debt in percent of GDP and private sector credit to GDP level. The regressions used clustered standard errors at the country level for robustness. Country fixed effects were only included in some specifications to control for unobserved inherent country characteristics but were excluded from most

[12] In the last 5 years.

[13] A higher value in the sum of binary answers indicates the report was more comprehensive in covering different angles of macrofinancial issues such as different sectoral vulnerabilities, structural vulnerabilities, emerging macrofinancial issues, FSAP integration, analytical tools, etc.

[14] Using Laeven and Valencia (2018) database. We define banking crises as a dummy equal one if it experienced a banking crisis over 2003-2018.

specifications especially those with country report dummy variables (Yi) due to collinearity with the country FE. Year FE ($\boldsymbol{\theta_t}$) were included to control for time-varying developments affecting all countries.

We also estimated logit and probit regressions analyzing the factors that could be associated with a higher likelihood of a match between LLM and human answers.

$\boldsymbol{Match_i = \delta + \alpha * X_i + e_i}$ eq. (6)

Match is a dummy variable taking the value of 1 if the LLM answer matches the human answer, and 0 otherwise for country i. Xi is a vector of variables including: the model type (one dummy variable per model GPT-4o, GPT-o1, GPT-5, and GPT-4.1), topic of the question (e.g. general information, emerging issues, FSAP integration, macrofinancial coverage rating questions, vulnerabilities that are structural in nature, etc.), type of question (open-ended, rating question, factual question), complexity of the question (simple or complex), and one dummy per income group (AE, EM, LIC). Our prior is that more advanced models, factual questions, simple questions, and general information topics would tend to be associated with a higher likelihood of a match.

Section IV. Empirical Results across GPT Models

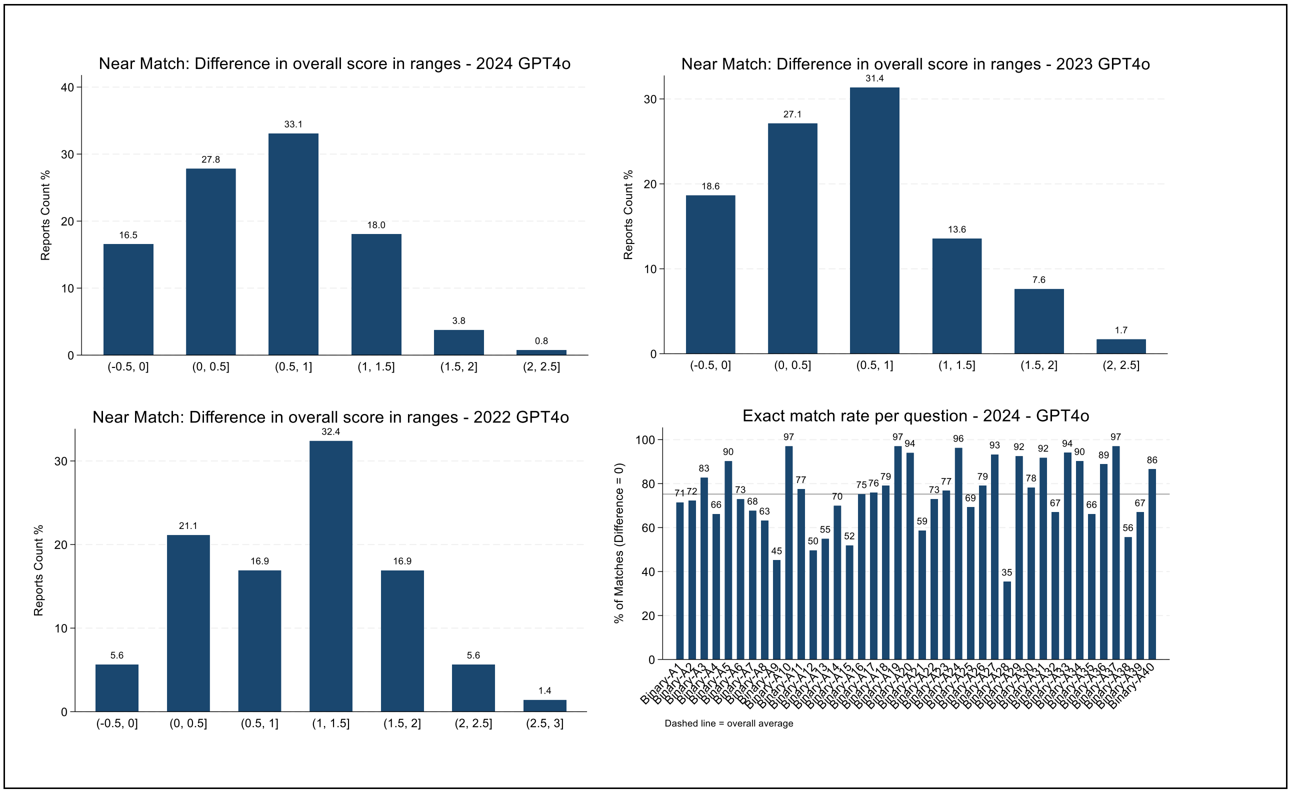
This section presents the empirical findings from our analysis, comparing the assessments generated by different LLMs against the benchmark of human economists' ratings. We begin with the results from the non-reasoning general model GPT-4o, then discuss how performance evolved with the introduction of the more advanced GPT-4.1 and reasoning-capable models such as GPT-o1 and GPT-5. The analysis covers both ratings and binary questions, aiming to provide a picture of the models' current capabilities and limitations in this specific task.

I. Initial Observations from the GPT-4o Model

Our first set of analyses used the GPT-4o model with the original prompt. The results from this phase appeared to show that the LLM assigned higher and more frequent positive ratings compared to the human reviews. This is suggested in the differences in rating distribution charts (see Figure 2 top row), where the fraction of LLM ratings is heavily skewed to the right.

When looking at accuracy with this initial setup, the performance was modest. The results on Annex I, table 1 show that the overall accuracy for rating questions hovered between 19% and 59% from 2016 to 2024. The F1 score, which considers both precision and recall, was also in a similar range, reaching 72% in 2024 but being considerably lower in prior years. In terms of how closely the LLM's ratings matched the human ones, the analysis suggested room for improvement. The near match charts (figure 1, first row) indicate that for the overall rating in 2024 and in 2023, the LLM’s assessment was within one point of the human rating in about 77% of cases (and within 0.5 points for 44.4% of the reports). However, for earlier years like 2022, this figure was much lower, at only 43.6%, suggesting that prompt and model refinements might be needed to achieve greater consistency. On the binary questions, this model performed relatively better, with an estimated exact match rate of 75 percent on average across questions for 2024 (figure 1, last row, right chart), with similar results for 2022-23. Although, with some heterogeneity in the exact match rate across questions.

Figure 1. GPT-4o Model results



II. Performance Gains with Model and Prompt Refinements

Following the initial tests, we moved to smarter models available to us at the time of this study (GPT-5, GPT-4.1, and GPT-o1) and introduced an updated prompt[15]. The results from this phase suggest a noticeable improvement in the model's performance. Annex I Table 2 provides a direct comparison across models, indicating that for most years, the GPT-o1 and GPT-4.1 and GPT-5 models achieved higher accuracy in rating questions than the GPT-4o model, although GPT-5 was not as accurate as the first two over 2016-23. For instance, for reports from 2022, accuracy appeared to jump from 27% with GPT-4o to 56-59% with the o1 and 4.1 models (44% in GPT-5). The improvement in recent models was even more pronounced for recent years, with overall accuracy for 2023 and 2024 reaching 74% and 75%, respectively in the o1 model (Annex I, Table 2).

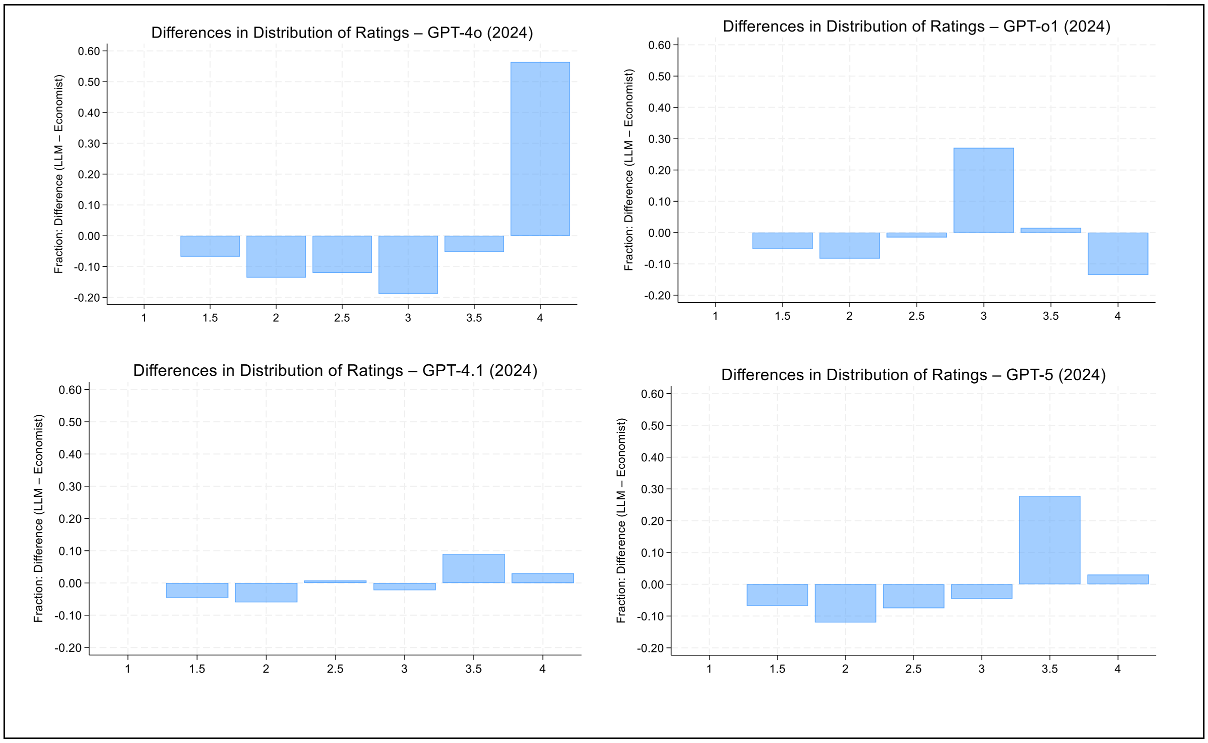
The difference in distribution of ratings also appeared to show more alignment with human assessments, although a tendency toward optimism remained. Figure 2 shows that the GPT-o1 and 5 models still exhibited an optimistic bias, this manifested as a clustering in the upper-middle range of the scale rather than at the ceiling. Unlike GPT-4o, the advanced models were less likely than humans to assign the absolute highest ratings (as shown by the negative bar in the right-hand side of the distribution for GPT-o1), but they significantly under-assigned lower ratings, resulting in a higher average. GPT-4.1 was the most balanced among the models we tried. In general, the more advanced models showed a distribution less compressed at the highest end of the scale compared to the GPT-4o results. This could suggest that the more advanced model, combined

[15] The prompt was updated to include examples of economists' write-ups in ratings questions for lower and higher ratings. This is called "few-shot" prompting, where the prompt includes examples of both high-quality and low-quality report excerpts, paired with the "correct" human rating and justification. We also allowed the LLM to give a longer justification and asked it to include a confidence score for its own answers. More tailoring was introduced for specific binary questions, particularly for those with lower match rate that tend to be more open-ended.

with a better prompt, might be slightly more discerning in its assessments. We also tried a GPT-4.1-mini in 2024 (Annex I, figure 1) and found that results resemble those of the GPT-4o model, with a skew to the right of LLM ratings.

A deeper look into other metrics beyond accuracy for the GPT-o1 model reveals a few interesting patterns. The model's recall—its ability to identify reports that humans rated highly—was consistently very strong, often above 90% (Annex I, Table 3). This suggests that the model is unlikely to miss a report that is genuinely well-regarded by a human expert. The precision, which captures the number of positive predictions that were actually correct, was more varied but also improved significantly in recent years, reaching 71% in 2024 (Annex I, Table 3). Consequently, the overall F1 score showed a strong upward trend, signaling a generally positive performance reaching 80% in 2024. Annex I table 4 and table 5 show similar results for GPT-4.1 and GPT-5 model. This may indicate that recent models are becoming more balanced in their ability to both identify positive cases and avoid false positives.

Figure 2. All GPT models – Differences in the Distribution of Overall Ratings

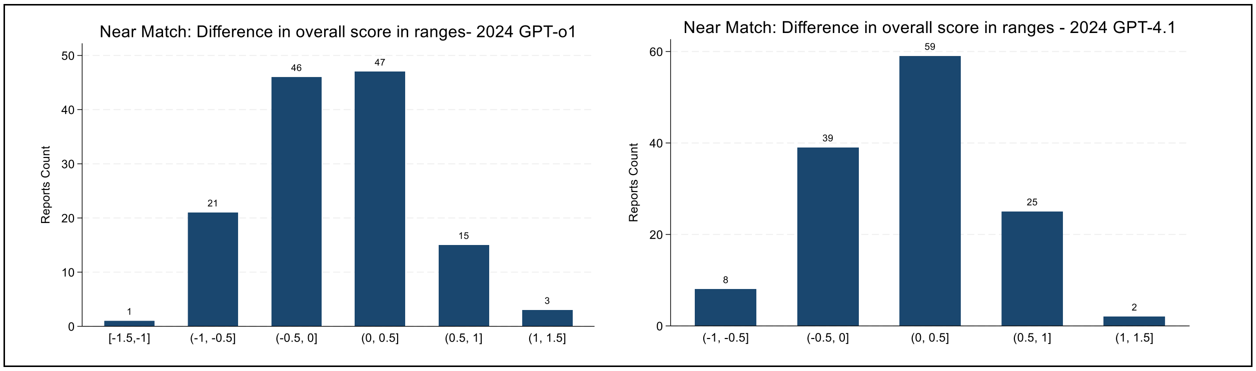


III. The Nature of Model-Human Agreement

Beyond simple accuracy, we examined the degree of agreement between the LLM and human ratings. The "exact match" rate—where the LLM and human give the exact same rating—remained low, hovering around 0-5% for all models in the overall rating in 2024 (Annex I, Table 6). This is perhaps not surprising given that decimal differences in rating choices in the three rating questions matter for the overall score.

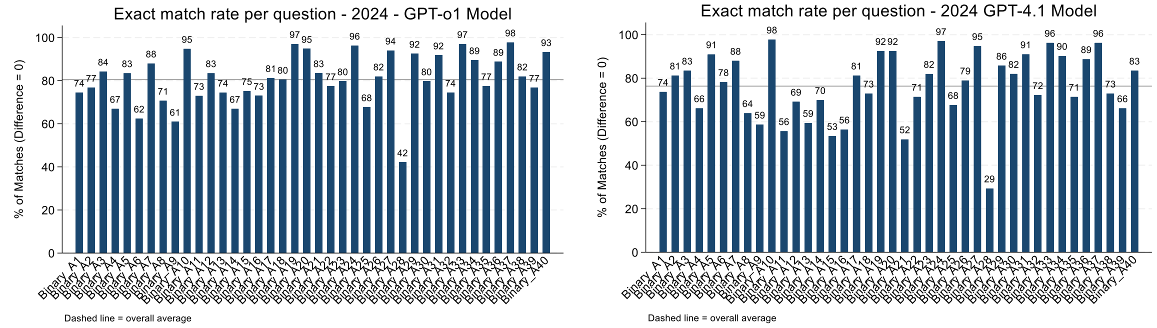
A more practical measure may be the "near match" rate. Figure 3 shows the difference between LLM and human ratings in ranges of 0.5 points for two of our most recent GPT models. For 2024, in about 70% of cases, the difference was between 0 and 0.5 points in the GPT-o1 model and 74% in GPT-4.1 (and 66% in GPT-5, see annex I, figure 5a for the latter[16]), and in about 97%-98% of cases, the difference was within 1 point for both models (while it was 91% for GPT-5). This suggests a reasonable level of agreement. We repeated this analysis by each of the three areas (macrofinancial coverage in the baseline, risk assessment, and policies) and results are similar. We also investigated the near match by income group and our findings indicate that the LLM differs slightly more from economists ratings by 0.5 points or more in emerging market and low income countries (Annex I, Figure 2).

Figure 3. Near match rate in GPT-o1 and GPT-4.1 models



The analysis of binary (Yes/No) questions offers another perspective, as these are typically more factual. Here, the LLM's performance appeared to be quite strong. Figure 4 indicates an average exact match rate of 80.6% across all binary questions for the 2024 reports in the GPT-o1 model and 76.3% in the GPT-4.1 model (Annex I, figure 5a shows GPT-5 results at 81.3% on average). Performance was also solid in previous years, with average exact match rate of around 77% for 2022 and 2023 (Annex I, Figure 3), 71.5% in GPT-4.1 model (Annex I, figure 4), and 79.2% in GPT-5 (annex I, figure 5a). This suggests that for fact-based extraction tasks, the LLM may be a highly reliable tool.

Figure 4. Exact match rates in binary questions in GPT-o1 and GPT-4.1 models



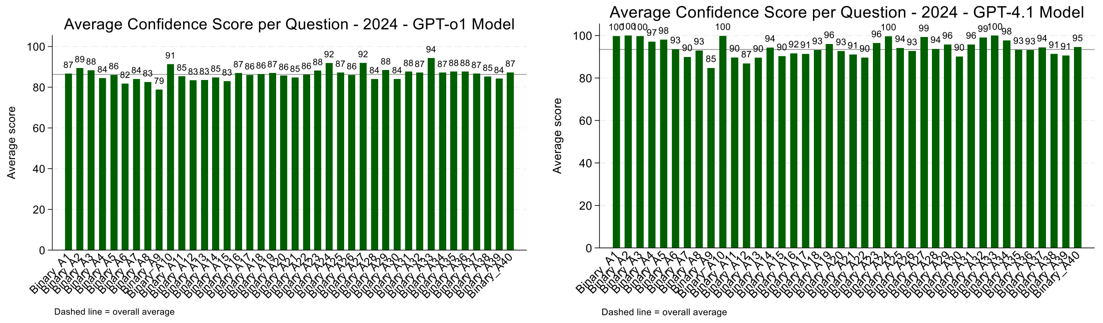
Notes: Binary (Yes/No) questions have been numbered from 1 to 40 given their confidential nature in the internal Fund survey. The binary questions ask about the reports' systemic risk assessment, FSAP integration, coverage of emerging macrofinancial issues, use of analytical tools, sources of financial sector vulnerabilities, and identified areas for policy recommendations.

[16] See Annex I figure 5b for results using GPT-5 high effort, which show similar results to the medium effort (default) setting.

IV. Exploring a Confidence Score Mechanism

To better handle uncertainty, we modified the prompt to ask the LLM to provide a confidence score from 0 to 100 for each of its answers. This might allow reviewers to filter results in the future and prioritize manual validation on answers where the model itself signals lower confidence, which could be helpful for triaging reports. Based on the GPT-o1 and GPT-4.1 models, we found the confidence to be around 81 percent on average (86% for the GPT-4.1 model and 88.4% for the GPT-5 model). The LLM signaled slightly lower confidence around open-ended questions, where the exact match rate is lower. In contrast, high confidence was shown in questions asking about issues typically covered in staff reports such as whether the report identifies banking sector-related vulnerabilities, financial integrity issues (AML/CFT), identifies regulation and supervision policies, and whether it uses FSIs.

Figure 5. Confidence score in Binary questions in GPT-o1 and GPT-4.1 models



V. Exploring Correlations

We also conducted some correlation analysis to see if certain characteristics were associated with the ratings assigned by the LLM as presented in equation 5. Table 2 presents the results of these regressions (Annex I, tables 7 and 8 present similar results obtained using GPT-4.1 and GPT-o1). The results might suggest, for instance, a positive correlation between a country having a recent Financial Sector Assessment Program (FSAP) and the LLM's overall rating. Similarly, characteristics like being an Emerging Market or a Low-Income Country appeared to be negatively correlated with the LLM's rating, which aligns with the patterns observed in human ratings. Having had a banking crisis over 2003-2018, higher values in binary questions (indicating more coverage of macrofinancial issues), a well-articulated view of systemic risk, a specialized macrofinancial review during the production of the report and participation in the team by a financial sector-expert economists seems to have a positive correlation with the LLM’s overall rating. As a comparator exercise, we also run these regressions for human ratings (Table 3), obtaining similar results. This type of analysis could be a fruitful area for future research with a more comprehensive dataset to better understand the factors that might influence both human and machine-generated assessments.

Table 2. Factors associated with LLM average overall ratings over 2016-2024 using GPT-5 model

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
VARIABLES	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM
Recent FSAP	0.2870***										0.0270
	(0.038)										(0.033)
Combined Article IV / Program		-0.2460***									-0.1116***
		(0.045)									(0.034)
Macrofinancial review			0.2647***								0.0385
			(0.043)								(0.037)
Participation of specialized economists				0.2073***							0.1080***
				(0.041)							(0.028)
General government gross debt, percent of GDP					-0.0006						-0.0004
					(0.001)						(0.000)
Credit-to-GDP, percent						0.0000***					0.0000
						(0.000)					(0.000)
Bank Crisis							0.1689***				-0.0240
							(0.049)				(0.046)
Cumulative Binary values								0.0533***			0.0366***
								(0.003)			(0.003)
Well-articulated view of Systemic Risk									0.7656***		0.4113***
									(0.073)		(0.065)
Income Group = 2, EME										-0.1929***	-0.0959**
										(0.047)	(0.047)
Income Group = 3, LIC										-0.4077***	-0.0976*
										(0.049)	(0.053)
Observations	544	541	544	541	494	456	544	544	544	544	476
R-squared	0.213	0.094	0.185	0.126	0.792	0.777	0.086	0.554	0.219	0.225	0.610
Country FE	No	No	No	No	Yes	Yes	No	No	No	No	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. countries	182	181	182	181	140	132	182	182	182	182	154

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

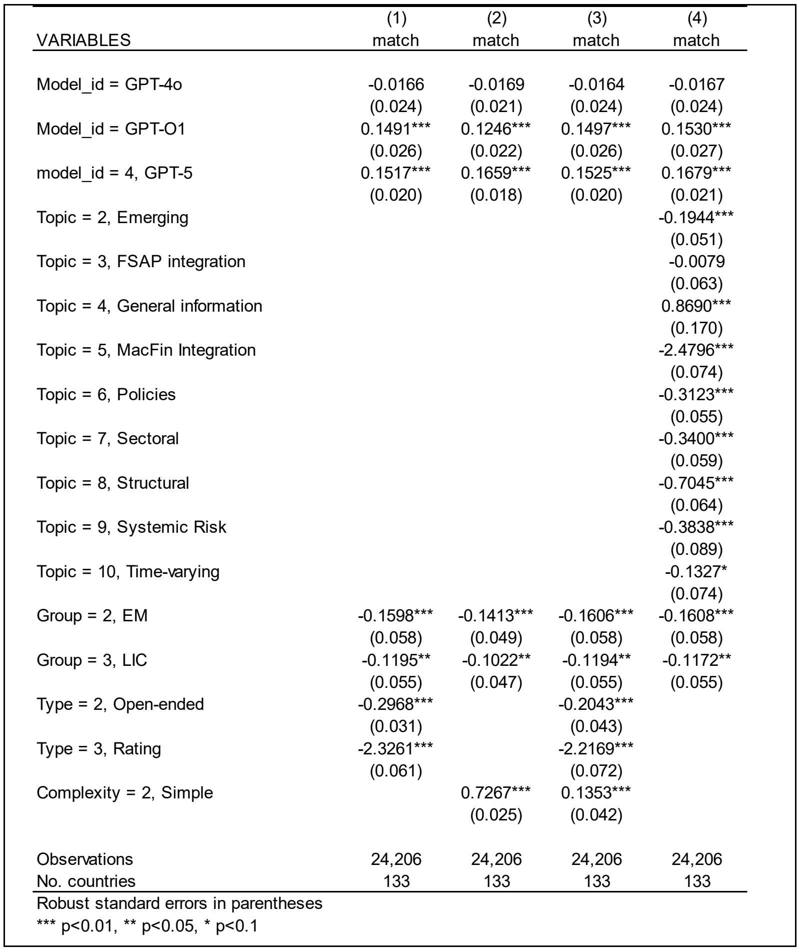
Table 3. Factors associated with Human average overall ratings over 2016-2024

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	Overall	Overall	Overall	Overall	Overall	Overall	Overall	Overall	Overall	Overall	Overall
VARIABLES	Economist	Economist	Economist	Economist	Economist	Economist	Economist	Economist	Economist	Economist	Economist
Recent FSAP	0.4133***										-0.0156
	(0.070)										(0.066)
Combined Article IV / Program		-0.4003***									-0.0380
		(0.083)									(0.097)
Macrofinancial review			0.5025***								0.0260
			(0.078)								(0.075)
Participation of specialized economists				0.3172***							0.1354**
				(0.083)							(0.068)
General government gross debt, percent of GDP					-0.0041						0.0007
					(0.003)						(0.001)
Credit-to-GDP, percent						0.0001***					-0.0000***
						(0.000)					(0.000)
Bank Crisis							0.2522**				-0.0593
							(0.101)				(0.075)
Cumulative Binary values								0.0809***			0.0563***
								(0.004)			(0.006)
Well-articulated view of Systemic Risk									0.5811***		0.1858**
									(0.088)		(0.072)
Income Group = 2, EME										-0.4944***	-0.3221***
										(0.084)	(0.092)
Income Group = 3, LIC										-0.7695***	-0.3627***
										(0.081)	(0.113)
Observations	546	542	546	542	497	458	546	546	322	546	271
R-squared	0.200	0.143	0.240	0.160	0.677	0.698	0.134	0.386	0.240	0.288	0.607
Country FE	No	No	No	No	Yes	Yes	No	No	No	No	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. countries	183	182	183	182	141	132	183	183	176	183	145

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

We also estimated probit (table 4) regressions as in equation 6 to estimate the factors associated with a higher probability of a match between LLM and human answers. Logit regression results are presented in Annex I, table 9.

Table 4. Factors associated with the probability of a match between LLM and human ratings in 2024

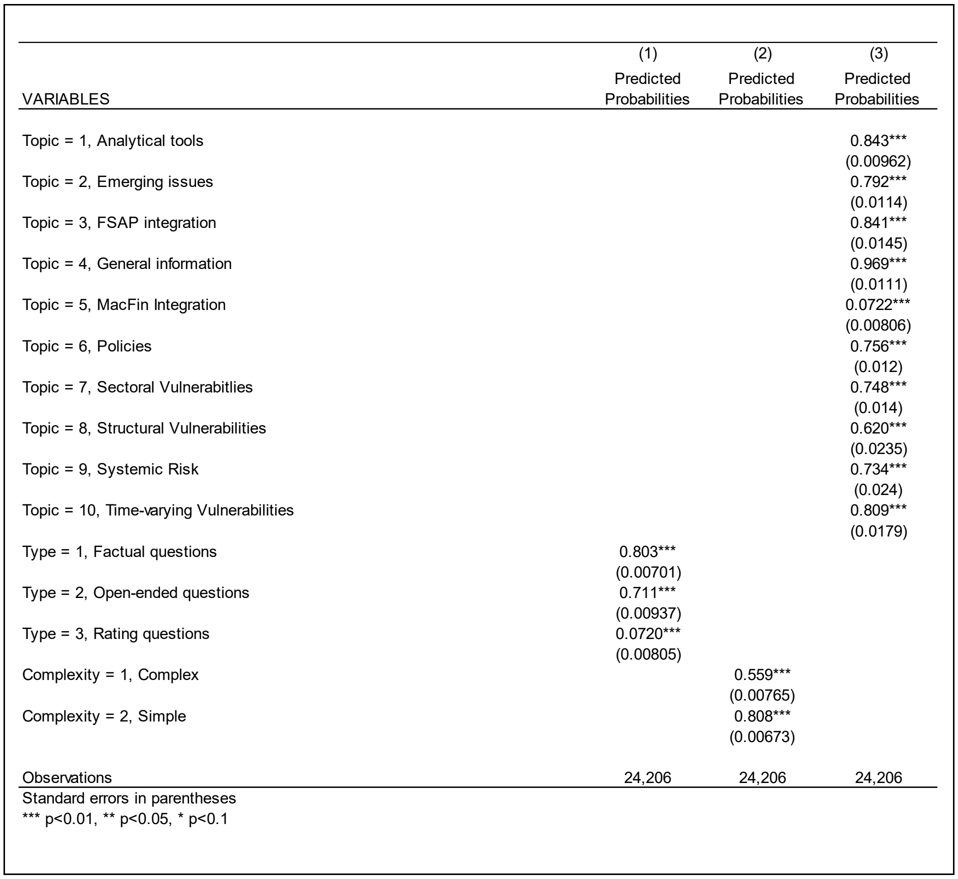


VARIABLES	(1) match	(2) match	(3) match	(4) match
Model_id = GPT-4o	-0.0166	-0.0169	-0.0164	-0.0167
	(0.024)	(0.021)	(0.024)	(0.024)
Model_id = GPT-O1	0.1491***	0.1246***	0.1497***	0.1530***
	(0.026)	(0.022)	(0.026)	(0.027)
model_id = 4, GPT-5	0.1517***	0.1659***	0.1525***	0.1679***
	(0.020)	(0.018)	(0.020)	(0.021)
Topic = 2, Emerging				-0.1944***
				(0.051)
Topic = 3, FSAP integration				-0.0079
				(0.063)
Topic = 4, General information				0.8690***
				(0.170)
Topic = 5, MacFin Integration				-2.4796***
				(0.074)
Topic = 6, Policies				-0.3123***
				(0.055)
Topic = 7, Sectoral				-0.3400***
				(0.059)
Topic = 8, Structural				-0.7045***
				(0.064)
Topic = 9, Systemic Risk				-0.3838***
				(0.089)
Topic = 10, Time-varying				-0.1327*
				(0.074)
Group = 2, EM	-0.1598***	-0.1413***	-0.1606***	-0.1608***
	(0.058)	(0.049)	(0.058)	(0.058)
Group = 3, LIC	-0.1195**	-0.1022**	-0.1194**	-0.1172**
	(0.055)	(0.047)	(0.055)	(0.055)
Type = 2, Open-ended	-0.2968***		-0.2043***	
	(0.031)		(0.043)	
Type = 3, Rating	-2.3261***		-2.2169***	
	(0.061)		(0.072)	
Complexity = 2, Simple		0.7267***	0.1353***	
		(0.025)	(0.042)	
Observations	24,206	24,206	24,206	24,206
No. countries	133	133	133	133

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

To get the average predicted probability of a match based on the control variables such as the complexity of the question and topics, we estimated the margins, which convert regression coefficients from Table 4 into estimated probabilities (Annex I, table 10 presents margin results from logit regressions). Table 5 shows that factual questions have a close to 80% probability of a match in 2024. Simple questions had a similar probability. In contrast, ratings questions and complex questions have a lower chance of a match. Questions from different topics had a similar probability in the range of 60%-97%, except for ratings questions which have a lower probability (around 7%) in 2024 as already pointed out.

Table 5. Margins (Probit) in 2024



	(1)	(2)	(3)
VARIABLES	Predicted Probabilities	Predicted Probabilities	Predicted Probabilities
Topic = 1, Analytical tools			0.843*** (0.00962)
Topic = 2, Emerging issues			0.792*** (0.0114)
Topic = 3, FSAP integration			0.841*** (0.0145)
Topic = 4, General information			0.969*** (0.0111)
Topic = 5, MacFin Integration			0.0722*** (0.00806)
Topic = 6, Policies			0.756*** (0.012)
Topic = 7, Sectoral Vulnerabitlies			0.748*** (0.014)
Topic = 8, Structural Vulnerabilities			0.620*** (0.0235)
Topic = 9, Systemic Risk			0.734*** (0.024)
Topic = 10, Time-varying Vulnerabilities			0.809*** (0.0179)
Type = 1, Factual questions	0.803*** (0.00701)		
Type = 2, Open-ended questions	0.711*** (0.00937)		
Type = 3, Rating questions	0.0720*** (0.00805)		
Complexity = 1, Complex		0.559*** (0.00765)	
Complexity = 2, Simple		0.808*** (0.00673)	
Observations	24,206	24,206	24,206

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In summary, the empirical results seem to point to a technology that is rapidly improving in its ability to assist with macrofinancial analysis. While initial models showed modest performance and a strong optimistic tendency, more advanced models like GPT-o1, GPT-5, and GPT-4.1, when paired with refined prompts, appear to deliver a higher level of accuracy and closer alignment with human experts. The models seem particularly effective at objective, binary questions. However, some degree of optimistic bias appears to persist, and exact agreement on ratings remains low, suggesting that the most effective use of this technology is likely as a supportive tool for, rather than a replacement of human expertise.

Section V. Discussion of Results and LLM issues

The evaluation presented in the preceding sections offers some initial insights into the capabilities of LLMs when applied to the specialized domain of macrofinancial analysis. Our findings suggest that while these models could be a useful tool for assisting economists, their application may benefit from an awareness of their potential limitations and systematic behaviors. This section discusses some of the primary issues that appear to emerge from our results—reproducibility, potential bias, and the interpretation of specialized information—and attempts to place them within the context of the models' overall performance. It also outlines a possible agenda for future research aimed at refining this technology into a more consistent analytical support tool.

I. Performance, Accuracy, and Reproducibility

A key observation from this paper is that certain GPT models appear to achieve useful -but still limited-levels of accuracy in analyzing IMF staff reports, particularly on structured, fact-based questions. For binary (Yes/No) questions, our analysis for the 2022-2024 period points to an average exact match rate of approximately 72-81% for some of the more advanced models tested. The F1 scores for the ratings questions also show a generally positive performance, reaching around 80% in 2024 with the GPT-o1 model. These results could suggest that LLMs might be able to replicate a portion of the manual review process, possibly offering some efficiency gains.

Furthermore, our results seem to show a performance evolution across model versions. As detailed in Table 2 (Annex I), the transition from the GPT-4o model to the GPT-o1 and GPT-4.1 models appeared to yield an improvement in accuracy across most years of our sample. This may highlight that model selection could be an important factor.[17] For instance, the higher accuracy we observe for o1 (75% in 2024) and GPT-5 (72%) relative to 4o (59%) is consistent with their *reasoning* capabilities (o1) and router-to-thinking behavior (GPT-5). Moreover, GPT-4.1's relatively strong performance (71% accuracy in 2024) is comparable to o1 and GPT-5 despite lacking a separate reasoning mode: fewer misses from truncation, better retrieval across long inputs. GPT-4.1 is explicitly tuned for long-context use, making it a safe choice in our process of loading large reports end-to-end. When a rating requires weighing multiple passages, o1's deliberate step may reduce logical slips; GPT-5 in default mode, automatically escalates effort only when needed. By contrast, 4o prioritizes responsiveness and may under-invest compute on borderline cases.

Across 2022–24, all three advanced models have very high recall (often between 90–100%), with precision improving into the mid-60s/low-70s by 2024, yielding F1 between 77–80%[18]. High recall reflects models' consistent ability to correctly identify reports that feature strong macrofinancial analysis, ensuring that genuine examples are rarely missed. Meanwhile, the improvement in precision—measuring the accuracy of those positive classifications—suggests that newer models are better at filtering out weak signals. The o1 model's training to 'think before answering' likely aids this discernment, reducing false positives, while GPT-5's router achieves a similar effect on complex prompts.

The pace of development in this area might mean that the capabilities and limitations observed are a moving target, possibly requiring periodic re-evaluation as new models become available.

This leads to the methodological question of reproducibility. Given the probabilistic nature of LLMs, a concern might exist that the same prompt and model could yield different results on subsequent runs. While perfect one-to-one identity is not always expected, our experience in this study indicated a high degree of practical consistency. In repeated tests using the same model, prompt, and temperature settings, we observed that the outputs were very similar, with an average 88% consistency in answers[19]. This level of stability could suggest that for the purpose of a large-scale review exercise, the results may be sufficiently reproducible to be useful. At the same time, this non-deterministic characteristic might imply that for certain applications, a validation

[17] Another possibility is that this improvement over time could be driven by the evolution of Article IV staff reports themselves, specifically the mainstreaming following the 2017 policy review and the 2022 Guidance note. However, this is a hypothesis.

[18] F1 is the harmonic mean of precision and recall (best = 1, worst = 0); values of 0.77–0.80 indicate that the models achieved a balanced performance on the rating task: they were catching most actual true positives (high recall) without too many false positives (decent precision). Because F1 is a balance, both precision and recall had to be reasonably high to land near ~0.8.

[19] A representative and well-balanced sample of 35 AEs, EMs, and LICs was used for this test. Using the GPT-4.1 model, the average match rate between the LLM's first answer and the LLM's newer answer was 93% in binary questions and the accuracy was 83% for rating questions. Using the GPT-4o model, those figures are 87% and 88% respectively.

framework, perhaps involving multiple runs or human review of the generated output, could be beneficial. This slight variability seems to support the role of the LLM as a supportive tool rather than a final arbiter.

II. Apparent Bias and a Tendency Toward Optimism

While the accuracy metrics seem promising, our analysis may point to a tendency toward over-optimism in the LLM's ratings. Across the years, country income groups, and model types we examined, the LLMs tended to assign higher average ratings for macrofinancial integration than the human economists serving as our benchmark. The charts displaying the differences in rating distributions (Figure 2) illustrate this phenomenon. The presence of positive bars in the upper-middle range indicates that the LLM assigns these scores more frequently than economists, while the negative bars at the lower end suggest the LLM is less inclined to provide highly critical assessments compared to the more evenly distributed human ratings.

The source of this apparent optimistic tendency could be multifaceted. It might be an artifact of the models' underlying training data, which may contain a higher proportion of neutral or positive-toned text. It could also stem from the "alignment" process used by developers to make models helpful, which may inadvertently discourage critical assessments. In the context of our study, the LLM might be giving significant weight to the mere presence of certain keywords or sections (e.g., a discussion of systemic risk) without fully evaluating the depth or quality of that discussion—a task that human experts may be better equipped to handle. Furthermore, this divergence may stem from a fundamental difference in evaluation styles: while human reviewers may implicitly 'grade on a curve,' naturally aiming for a normal distribution of scores relative to the peer group, the LLM evaluates each document in isolation against the prompt's criteria, lacking the internal mechanism to normalize the distribution across the sample.

The potential implications of such a tendency are worth considering. If used without careful oversight, an LLM leaning toward optimism could create a misleading impression of the quality of a staff report, possibly failing to flag documents with weaker macrofinancial analysis. This observation would seem to support the case for keeping a human involved in the review process. The LLM's output might best be treated as a preliminary assessment that could benefit from human validation and contextual interpretation.

This tendency also seems to interact with other factors, such as country characteristics. Our results suggest that the LLM mimics the human pattern of assigning higher ratings to Advanced Economies (AEs) compared to Emerging Markets (EMs) and Low-Income Countries (LICs). However, the gap in ratings between these country groups appears to be narrower in the LLM's assessments. This could indicate that the optimism is a broad effect that compresses the overall variance of the ratings, potentially making the LLM less sensitive to factors that lead to greater rating dispersion among human reviewers. While our efforts to refine prompts and use more advanced models appeared to improve overall accuracy, they did not seem to fully eliminate this tendency, suggesting it could be a persistent challenge.

III. Interpretation of Nuanced Information

Beyond quantitative measures, a challenge may lie in the LLM's interpretation of specialized instructions. In our work, we did not observe significant instances of outright factual fabrication, but we did notice cases that could be interpreted as misinterpretations of open-ended questions. Here, the LLM might generate justifications that, while referencing the text correctly, did not seem to fully align with the contextual intent of the question. A review of the justifications provided by the LLM was a part of our analysis, and these instances of interpretation differences are worth noting.

Specific examples from our binary question analysis might illustrate this. For "Binary-A28," which asks if the report covers financial sector policies not included in a predefined list, the LLM frequently answered "yes" by citing policies related to fintech, deposit insurance, payment systems, financial inclusion, crypto regulation, or gender in finance. While these policies were indeed mentioned in the report, human economists generally did not consider these specific wording to qualify as “substantive discussion of other financial policies in the report” and answered "no." The LLM, perhaps lacking this specific contextual framework, may have adopted a more literal interpretation. Similarly, for "Binary-A15," which assesses coverage of climate-related financial risks, the GPT-4o model sometimes returned a positive answer if there was any mention of the country's susceptibility to climate change, which could be a lower threshold than a human expert might apply. However, the newer models distinguish better than GPT-4o, the fact that climate policies can be discussed extensively but not the financial sector risks associated with it (correctly assigning a zero in that question). In the instances where the newer models answer affirmatively, they mention as justification the discussion of green finance, climate vulnerability assessment of banks, or NPLs being low due to a climate event. In fact, for this question, the unmatched cases tend to be cases of false negatives (i.e. the newer models answers “No” to question of whether the report covered climate-related financial risks, while the economist considered it covered).

These examples could suggest two things. First, they may point to the "brittleness" of an LLM's understanding in highly specialized contexts. It can identify text but may find it difficult to weigh its significance or interpret its intended meaning within the specific framework of our assessment questionnaire. Second, they may highlight the value of including a justification feature in the prompt design. A simple "yes" or "no" from the model would have been harder to validate and, in these cases, might have been less informative. The requirement to provide a direct quote and justification acted as a way for the model to "show its work," allowing for human oversight to catch these interpretative differences. It seems to be in these moments of ambiguity where the LLM's performance can diverge from human assessment, which could support its role as an assistant for initial processing rather than as a substitute for expert judgment.

IV. Potential Future Research and Refinements

The findings and potential limitations identified in this study could inform future work in this area. One goal might be to refine the methodology to address some of the observed shortcomings and improve the LLM’s utility as an analytical assistant. The following steps represent some possible avenues for future research:

- *Further Prompt Refinements:* This appears to be a practical area for continued improvement. Future work could explore role-based prompting where we would instruct the LLM to adopt a specific role such as an expert economist in macrofinancial issues. This technique might help to better align the LLM with the expectations of human economists, and could potentially tailor better the LLM’s responses.
- *Calibration and Contextual Benchmarking*: Future studies could attempt to simulate 'grading on a curve' by moving beyond isolated document processing. This could involve batch-processing multiple reports to allow the model to rank them relatively or using Retrieval-Augmented Generation (RAG) to provide the model with a dynamic library of 'gold standard' historical reports to use as comparative references.
- *A Closer Look at Error Types*: A more formal effort could be made to categorize the types of divergence between LLM and human answers. By creating a typology—for example, (i) factual differences, (ii) optimistic rating differences, and (iii) different interpretations of open-ended questions—one might develop targeted strategies for each. This could also help in identifying which questions on the survey may be less "LLM-friendly." Furthermore, efforts to decompose the differences

between LLM and economists' answers may be helpful to assess whether LLMs are useful in providing a second opinion.

- *Expanding the Data Sample:* To allow for more extensive quantitative analysis, it might be useful to extend the assessment period back to 2008 and before. This could create a more continuous panel dataset, which might in turn allow for regression analysis to better understand the drivers of LLM ratings. However, human ratings are not available over that period.

Overall, this study suggests that LLMs may have reached a point where they can be helpful in performing certain complex, domain-specific textual analysis tasks. They might have the potential to increase the efficiency and consistency of some large-scale review exercises. However, their use would likely benefit from an understanding of their potential limitations, including tendencies toward bias and difficulties with deep contextual reasoning. The path forward may not be about replacing human experts, but rather about developing effective human-machine collaboration systems where the LLM could act as a capable, but supervised, partner.

Section VI. Conclusions

This study sets out to evaluate the extent to which large language models (LLMs) can support the resource-intensive task of reviewing macrofinancial issues in IMF Article IV staff reports. By systematically comparing model outputs against expert economist benchmarks across both factual and nuanced dimensions, we provide evidence on the opportunities and limitations of current-generation LLMs for macrofinancial surveillance.

On one hand, LLMs—especially more advanced models such as GPT-o1, GPT-4.1, and GPT-5—demonstrate relatively stronger performance on fact-based, binary questions, with accuracy rates in the range of 76–81%. This suggests clear potential for these models to handle routine extraction of objective information. Moreover, even in rating questions, model assessments frequently fell within a near-match range relative to human reviewers.

On the other hand, important limitations remain. A consistent upward bias in ratings indicates that models tend to provide more favorable assessments than human counterparts, underscoring the risk of systematic optimism. Additionally, the models struggle with nuanced, open-ended questions where context, interpretation, and judgment are essential. These shortcomings make it clear that LLMs cannot yet substitute for expert economic analysis. Human oversight remains indispensable to ensure rigor, context-sensitivity, and the correction of biases.

Interpreting these divergences requires nuance. While LLMs exhibit a tendency toward optimism—likely an artifact of alignment training designed to produce helpful and non-confrontational responses—it is important to recognize that the human benchmark itself is not a context-free objective standard. Human reviewers operate with private information sets, which may not be explicitly detailed in the report's text. Therefore, the divergence in ratings may partially reflect the LLM's reliance solely on the written document, whereas human experts incorporate a broader, unobserved information set. Future research could further shed light on these dynamics by attempting to disentangle model bias from the implicit broader knowledge held by economists.

Taken together, our findings suggest that LLMs could become a valuable complement to human reviewers, helping save time and improve consistency in assessments without replacing expert judgment.

Deployed as preliminary secondary reviewers or providers of second opinions, LLMs could enhance consistency and reduce staff workload on routine elements of surveillance. Continued advances in model development, combined with careful prompt design, hold promise for deepening the contribution of LLMs.

Annex I. Additional Tables and Figures

Table 1. Accuracy, Precision, Recall, and F1 score in Model GPT-4o

year	Accuracy	Precision	Recall	F1 score
2016	47%	46%	100%	63%
2017	28%	27%	100%	43%
2018	28%	28%	100%	44%
2019	19%	18%	92%	30%
2022	27%	25%	100%	40%
2023	53%	50%	100%	66%
2024	59%	57%	100%	72%

Table 2. Accuracy in rating questions across GPT models

year	GPT-4o	GPT-o1	GPT-4.1	GPT-5 (default effort)
2016	47%	50%	50%	46%
2017	28%	46%	46%	36%
2018	28%	49%	49%	38%
2019	19%	44%	44%	34%
2022	27%	56%	59%	44%
2023	53%	74%	74%	61%
2024	59%	75%	71%	72%

Table 3. Accuracy, Precision, Recall, and F1 score in Model GPT-o1

year	Accuracy	Precision	Recall	F1 score
2016	50%	48%	94%	64%
2017	46%	33%	94%	48%
2018	49%	36%	100%	53%
2019	44%	24%	92%	38%
2022	56%	35%	100%	52%
2023	74%	65%	93%	77%
2024	75%	71%	90%	80%

Table 4. Accuracy, Precision, Recall, and F1 score in Model GPT-4.1

year	Accuracy	Precision	Recall	F1 score
2016	57%	52%	88%	65%
2017	54%	35%	81%	49%
2018	63%	44%	100%	61%
2019	55%	28%	92%	43%
2022	59%	37%	100%	54%
2023	74%	66%	91%	76%
2024	71%	68%	88%	77%

Table 5. Accuracy, Precision, Recall, and F1 score in Model GPT-5 (medium effort)

year	Accuracy	Precision	Recall	F1 score
2016	46%	46%	100%	63%
2017	36%	30%	100%	46%
2018	38%	32%	100%	48%
2019	34%	22%	100%	36%
2022	44%	30%	100%	46%
2023	61%	55%	95%	69%
2024	72%	67%	97%	79%

Table 6. Exact match rate across GPT models

year	GPT-4o	GPT-o1	GPT-4.1	GPT-5
2016	6%	12%	12%	0%
2017	3%	5%	5%	0%
2018	2%	2%	2%	2%
2019	2%	2%	2%	0%
2022	1%	1%	1%	0%
2023	6%	4%	0%	0%
2024	5%	5%	0%	0%

Figure 1. Differences in Distribution of ratings in GPT-4.1-mini model – 2024

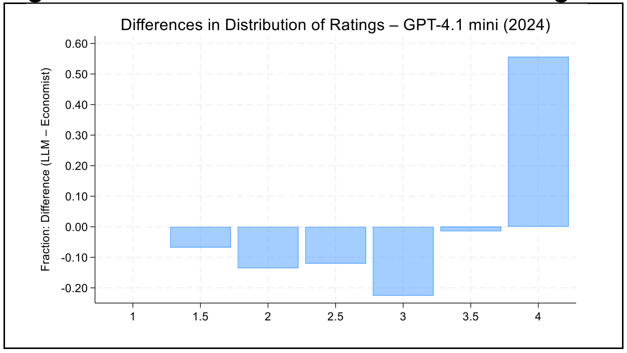


Figure 2. Near match rate by income group and area of assessment in GPT-O1 Model – 2024

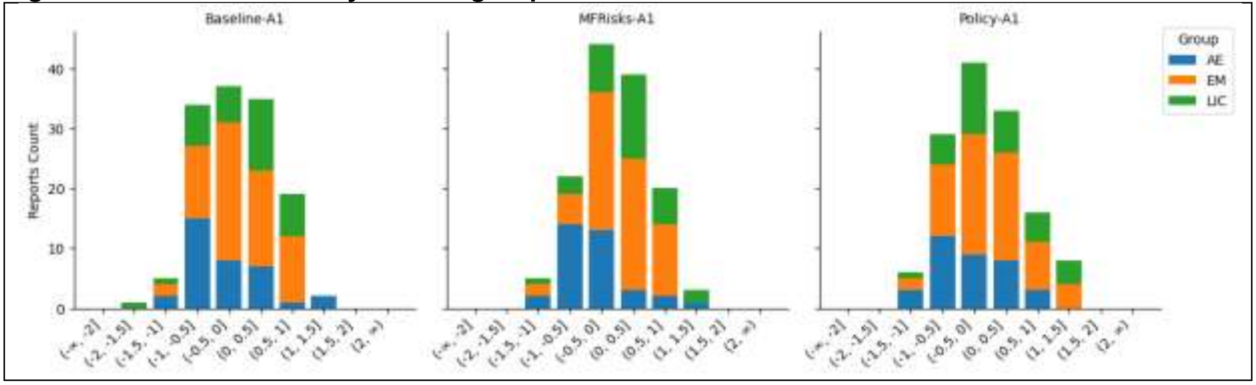


Figure 3. Exact match rate in binary questions in GPT-O1 Model – 2023 and 2022

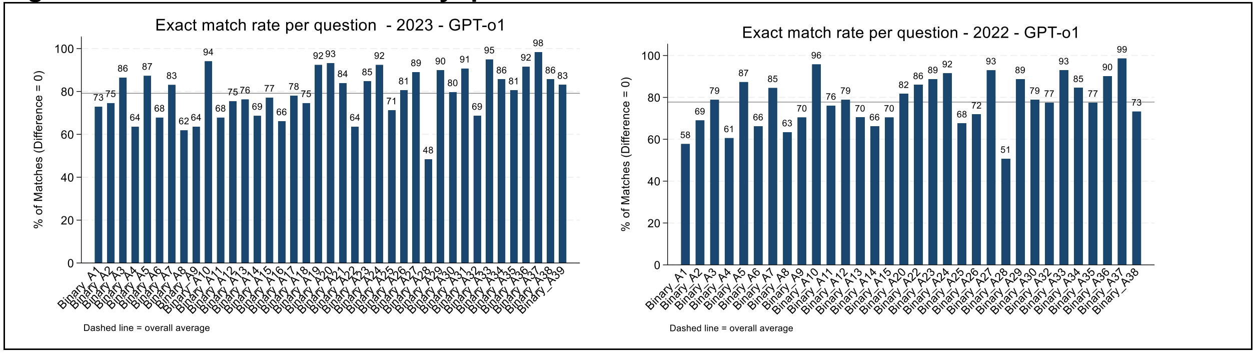


Figure 4. Exact match rate in binary questions in GPT-4.1 Model – 2023 and 2022

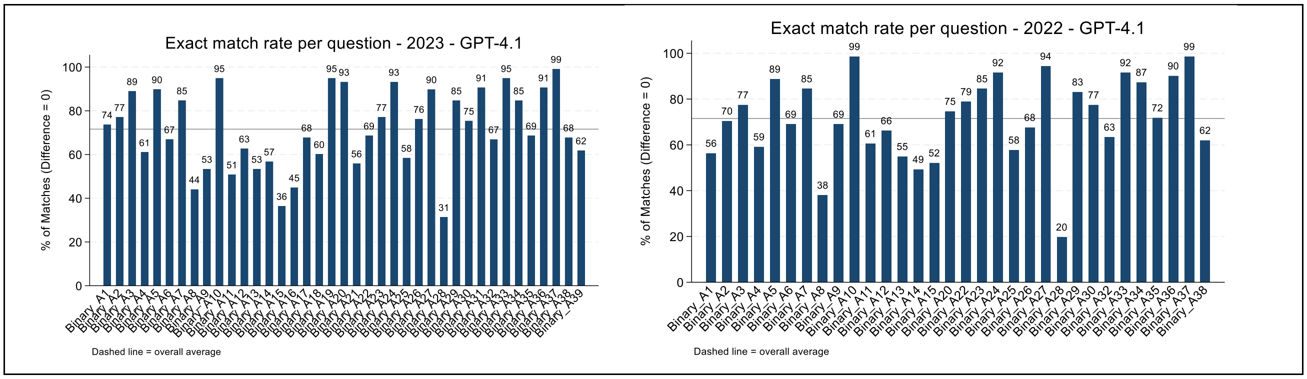


Figure 5a. Match rate in Binary questions, Distribution of Differences, Confidence in GPT-5 Model (medium effort) – 2023 and 2024

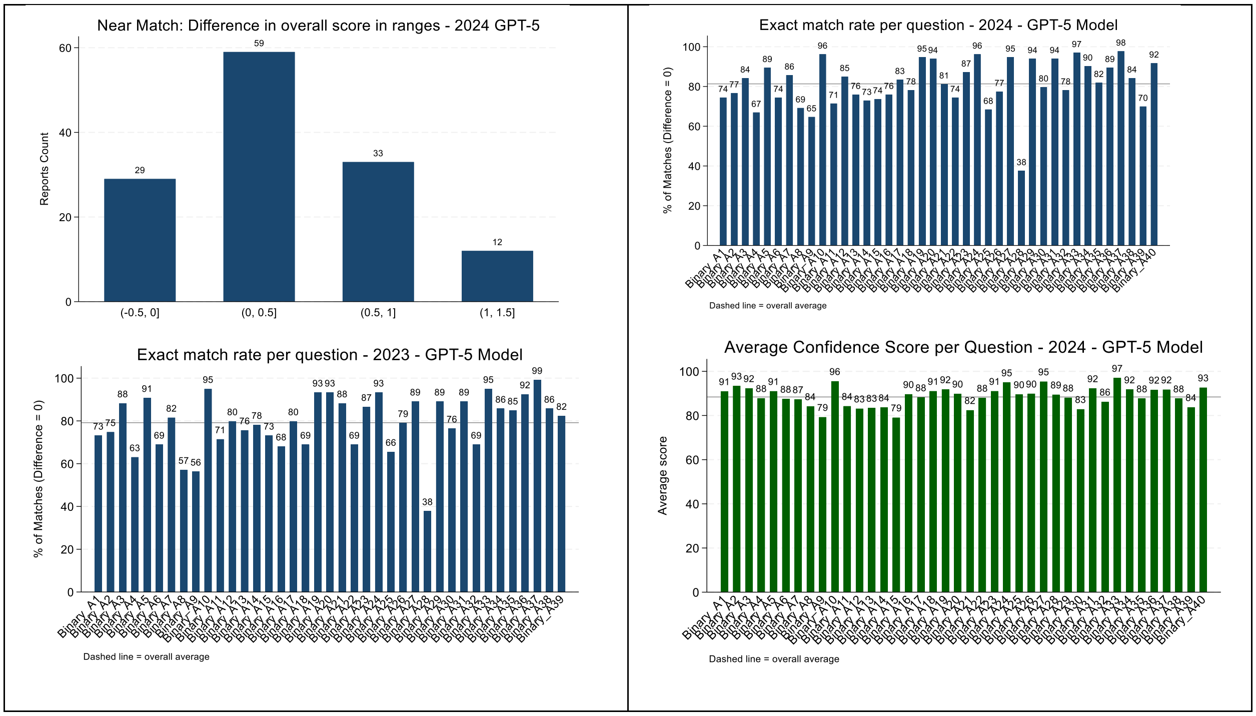


Figure 5b. Match rate in Binary questions, Distribution of Differences, Confidence in GPT-5 Model (high effort) – 2023 and 2024

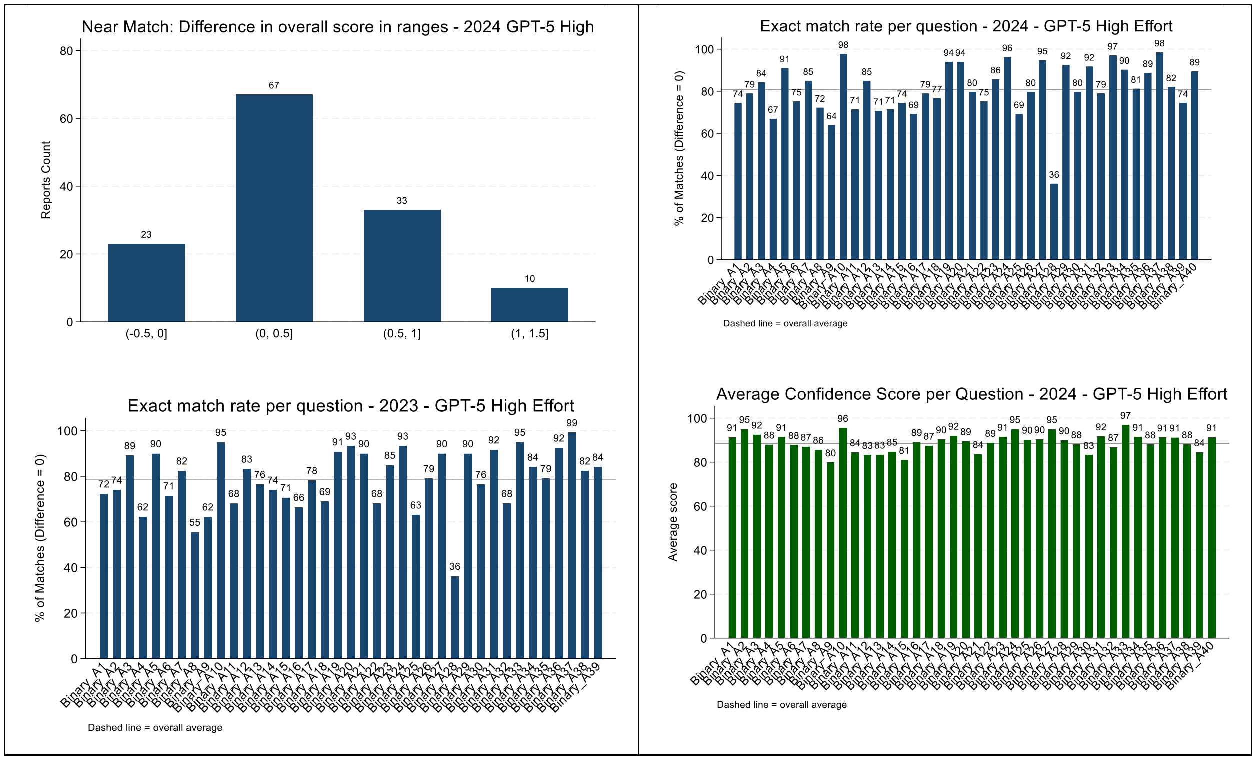


Table 7. Factors associated with LLM average overall ratings over 2016-2024 using GPT-4.1 Model

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
VARIABLES	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM
Recent FSAP	0.4650***										0.0103
	(0.058)										(0.047)
Combined Article IV / Program		-0.2560***									0.0209
		(0.072)									(0.051)
Macrofinancial review			0.4499***								0.0502
			(0.064)								(0.047)
Participation of specialized economists				0.2494***							0.1034***
				(0.063)							(0.035)
General government gross debt, percent of GDP					-0.0007						-0.0004
					(0.001)						(0.000)
Credit-to-GDP, percent						0.0000***					-0.0000***
						(0.000)					(0.000)
Bank Crisis							0.4034***				0.0256
							(0.062)				(0.046)
Cumulative Binary values								0.0815***			0.0535***
								(0.004)			(0.005)
Well-articulated view of Systemic Risk									1.3336***		0.6002***
									(0.117)		(0.124)
Income Group = 2, EME										-0.4408***	-0.3106***
										(0.056)	(0.050)
Income Group = 3, LIC										-0.8546***	-0.4471***
										(0.058)	(0.064)
Observations	543	539	543	539	494	456	543	543	543	543	475
R-squared	0.209	0.031	0.192	0.061	0.833	0.820	0.110	0.536	0.083	0.371	0.644
Country FE	No	No	No	No	Yes	Yes	No	No	No	No	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. countries	182	181	182	181	140	132	182	182	182	182	154

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 8. Factors associated with LLM average overall ratings over 2016-2024 using GPT-o1 Model

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
VARIABLES	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM	Overall LLM
Recent FSAP	0.2261***										0.0138
	(0.036)										(0.027)
Combined Article IV / Program		-0.1460***									-0.0207
		(0.051)									(0.037)
Macrofinancial review			0.2062***								-0.0239
			(0.040)								(0.030)
Participation of specialized economists				0.2052***							0.0686**
				(0.041)							(0.028)
General government gross debt, percent of GDP					-0.0002						0.0004
					(0.002)						(0.000)
Credit-to-GDP, percent						0.0000***					-0.0000**
						(0.000)					(0.000)
Bank crisis							0.1607***				-0.0335
							(0.037)				(0.026)
Cumulative Binary values								0.0435***			0.0287***
								(0.003)			(0.003)
Well-articulated view of Systemic Risk									0.6035***		0.2373***
									(0.051)		(0.044)
Income Group = 2, EME										-0.1873***	-0.1014***
										(0.038)	(0.032)
Income Group = 3, LIC										-0.3975***	-0.1187***
										(0.043)	(0.042)
Observations	542	538	542	538	493	455	542	542	542	542	474
R-squared	0.121	0.033	0.102	0.094	0.620	0.605	0.052	0.476	0.306	0.188	0.504
Country FE	No	No	No	No	Yes	Yes	No	No	No	No	No
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
No. countries	182	181	182	181	140	132	182	182	182	182	154

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 9. Logit regression in 2024

	(1)	(2)	(3)	(4)
VARIABLES	match	match	match	match
Model_id = GPT-4o	-0.0325	-0.0271	-0.0325	-0.0330
	(0.041)	(0.034)	(0.041)	(0.041)
Model_id = GPT-O1	0.2595***	0.2107***	0.2598***	0.2631***
	(0.046)	(0.037)	(0.046)	(0.047)
model_id = 4, GPT-5	0.2693***	0.2818***	0.2707***	0.2959***
	(0.035)	(0.029)	(0.035)	(0.036)
Topic = 2, Emerging				-0.3475***
				(0.090)
Topic = 3, FSAP integration				-0.0132
				(0.114)
Topic = 4, General information				1.7677***
				(0.386)
Topic = 5, MacFin Integration				-4.2674***
				(0.147)
Topic = 6, Policies				-0.5474***
				(0.097)
Topic = 7, Sectoral				-0.5988***
				(0.104)
Topic = 8, Structural				-1.1972***
				(0.107)
Topic = 9, Systemic Risk				-0.6688***
				(0.152)
Topic = 10, Time-varying				-0.2343*
				(0.132)
Group = 2, EM	-0.2852***	-0.2318***	-0.2856***	-0.2887***
	(0.102)	(0.082)	(0.102)	(0.103)
Group = 3, LIC	-0.2136**	-0.1724**	-0.2139**	-0.2161**
	(0.098)	(0.078)	(0.098)	(0.099)
Type = 2, Open-ended	-0.5083***		-0.3493***	
	(0.053)		(0.074)	
Type = 3, Rating	-3.9929***		-3.8063***	
	(0.124)		(0.139)	
Complexity = 2, Simple		1.2079***	0.2315***	
		(0.043)	(0.073)	
Observations	24,206	24,206	24,206	24,206
No. countries	133	133	133	133

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 10. Logit margin results: conveying the average probability of a match.

	(1)	(2)	(3)
VARIABLES	Predicted Probabilities	Predicted Probabilities	Predicted Probabilities
Topic = 1, Analytical tools			0.843*** (0.00963)
Topic = 2, Emerging issues			0.791*** (0.0114)
Topic = 3, FSAP integration			0.841*** (0.0144)
Topic = 4, General information			0.969*** (0.0112)
Topic = 5, MacFin Integration			0.0717*** (0.00801)
Topic = 6, Policies			0.757*** (0.0119)
Topic = 7, Sectoral Vulnerabilities			0.747*** (0.014)
Topic = 8, Structural Vulnerabilities			0.620*** (0.0236)
Topic = 9, Systemic Risk			0.734*** (0.024)
Topic = 10, Time-varying Vulnerabilities			0.809*** (0.0179)
Type = 1, Factual questions	0.803*** (0.00701)		
Type = 2, Open-ended questions	0.711*** (0.00935)		
Type = 3, Rating questions	0.0716*** (0.00801)		
Complexity = 1, Complex		0.559*** (0.00765)	
Complexity = 2, Simple		0.808*** (0.00675)	
Observations	24,206	24,206	24,206

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

References

Baker, S. R., N. Bloom, and S. J. Davis (2016). "Measuring Economic Policy Uncertainty," The Quarterly Journal of Economics 131(4): 1593–1636. Available at: https://academic.oup.com/qje/article/131/4/1593/2607345

Bergant, K., A. Fernandez, K. Teoh, and M. Uribe (2026). “Expanding the Landscape of Cross-border Flow Restrictions: Modern Tools and Historical Perspectives” NBER Working Paper No. 34615. Available at: https://www.nber.org/papers/w34615

Carriero, A., D. Pettenuzzo, and S. Shekhar (2025). "Macroeconomic Forecasting with Large Language Models." Available at: https://ssrn.com/abstract=4881094

Chen, J., G. Tang, G. Zhou, and W. Zhu (2025). "ChatGPT and DeepSeek: Can They Predict the Stock Market and Macroeconomy?" Available at: https://arxiv.org/abs/2502.10008

Chen, Y., B. T. Kelly, and D. Xiu (2022). "Expected Returns and Large Language Models." Available at: https://ssrn.com/abstract=4416687

Chen, Y., A. Padmanabhan, J. Yang, and K. Z. Watkins (2025). "Total Recall? Evaluating the Macroeconomic Knowledge of Large Language Models," FEDS Paper No. 2025-044. Available at: https://www.federalreserve.gov/econres/feds/files/2025044pap.pdf

Christiano Silva T., K. Moriya, and R. M. Veyrune (2025). "From Text to Quantified Insights: A Large-Scale LLM Analysis of Central Bank Communication," IMF Working Paper No. 2025/109. Available at: https://www.imf.org/en/Publications/WP/Issues/2025/06/06/From-Text-to-Quantified-Insights-A-Large-Scale-LLM-Analysis-of-Central-Bank-Communication-567522

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Available at: https://aclanthology.org/N19-1423/

Fang, H., R. Jia, H. Li, and W. Lu (2025). "Decoding China's Industrial Policies," NBER Working Paper No. 33814. Available at: https://www.nber.org/papers/w33814

Hansen, A. L., and S. Kazinnik (2023). "Can ChatGPT Decipher Fedspeak?" Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4399406

IMF (2017). "Approaches to Macrofinancial Surveillance in Article IV Reports," IMF Policy Paper, Washington D.C. Available at: https://www.imf.org/en/Publications/Policy-Papers/Issues/2017/03/28/approaches-to-macrofinancial-surveillance-in-article-iv-reports

IMF (2022). "Guidance Note for Surveillance Under Article IV Consultations," IMF Policy Paper, Washington D.C. Available at: https://www.imf.org/en/publications/policy-papers/issues/2022/06/23/guidance-note-for-surveillance-under-article-iv-consultations-519916

Jha, M., J. Qian, M. Weber, and B. Yang (2025). "ChatGPT and Corporate Policies," Chicago Booth Research Paper No. 23-15. Available at: https://ssrn.com/abstract=4521096

Korinek, A. (2023). "Generative AI for Economic Research: Use Cases and Implications for Economists," Journal of Economic Literature 61(4): 1281–1317. Available at: https://www.aeaweb.org/articles?id=10.1257/jel.20231736

Kwon, B., T. Park, F. Perez-Cruz, and P. Rungcharoenkitkul (2024). "Large Language Models: A Primer for Economists," BIS Quarterly Review, December. Available at: https://www.bis.org/publ/qtrpdf/r_qt2412b.htm

Laeven, L. and F. Valencia. (2018). “Systemic Banking Crises Revisited”, IMF Working Papers 2018, 206. https://doi.org/10.5089/9781484376379.001

Romer, C. D., and D. H. Romer (1989). "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz," NBER Macroeconomics Annual 4: 121–170. Available at: https://www.nber.org/papers/w2966

PUBLICATIONS

How Effectively Can Current LLMs Analyze Macrofinancial Issues

Working Paper No. WP/2026/035