



Банк России

Тесты с подменой дат: можем ли мы доверять внутривыборочной точности LLM в макроэкономическом прогнозировании?

Александр Елисеев, Сергей Селезнев

Департамент исследований и прогнозирования, Банк России*

XVIII семинар по экономическим исследованиям, Северо-Западное ГУ Банка России

Санкт-Петербург, 30 июня 2026 г.

** Настоящая презентация отражает личную позицию авторов. Содержание данной презентации не следует рассматривать, в том числе цитировать в каких-либо изданиях, как официальную позицию Банка России или указание на официальную политику или решения регулятора. Любые ошибки в данном материале являются исключительно авторскими.*

Краткое содержание (1)

О чем эта работа?

- Экономисты начали использовать LLM (вроде ChatGPT, DeepSeek, Qwen, GigaChat и т.п.) для прогнозирования макроэкономических показателей (инфляции, ВВП, % ставок и др.)
- Чтобы понять, насколько хорошо эти модели прогнозируют, их тестируют на исторических данных: «Представь, что сейчас 2019 год, и предскажи рост экономики в 2020-м году»
- Но есть проблема: современные LLM уже видели все данные до 2025-2026 гг. Они знают, что произошло в 2020 году. Поэтому, когда мы просим их сделать «прогноз из прошлого», они могут:
 - а) просто вспоминать ответ из своей памяти, а не предсказывать его логически
 - б) опираться на широкие знания об эпохе, а не только на информацию из промпта
- Мы разработали формальные тесты, цель которых – попытаться ответить на следующий вопрос: «Стоит ли доверять точности макропрогнозов LLM, если они проверяются на данных, которые они уже видели?»

Краткое содержание (2)

Что показали результаты?

- Мы протестировали три модели: Kimi-K2 Instruct, Qwen3 Instruct и Deepseek-V3.1.
- Ни одна из них не прошла сформулированные нами тесты. Это может означать:
 - а) что модели не способны честно «отключить» свои знания о будущем, даже когда их об этом просят в инструкции
 - б) что они явно и неявно используют информацию из своего обучающего датасета, которая соответствует реальному историческому периоду, даже если в тексте промпта этой информации нет

Краткое содержание (3)

Почему это важно?

- Эмпирические работы хвалят LLM за высокую точность макропрогнозов на истории. Наша работа сигнализирует: данная «точность» не обязательно означает, что модели будут способны так же хорошо прогнозировать на новых данных в будущем
- Предложенные нами тесты очень просты в реализации и опираются на небольшой набор предпосылок. Их легко интегрировать в свои эксперименты, чтобы проверить свойства исторических прогнозов LLM
- Нельзя так просто взять и использовать исторические прогнозы/оценки, полученные с помощью «массовых» LLM. Нужны дополнительные проверки, либо же следует использовать специально разработанные методы и модели, позволяющие получить изначально более надежные исторические прогнозы/оценки из LLM



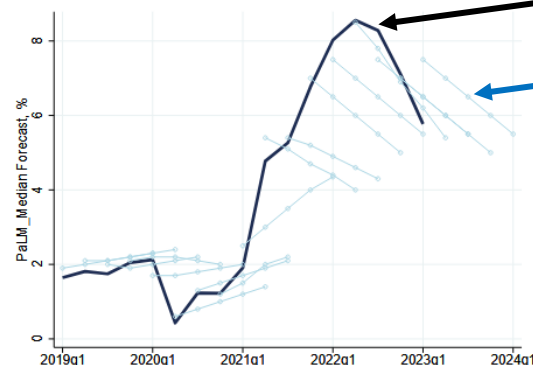
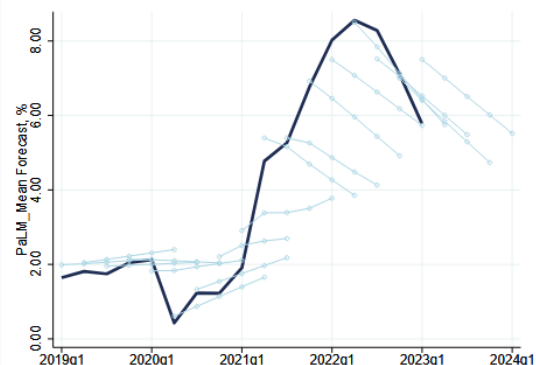
Банк России

1

МОТИВАЦИЯ ИССЛЕДОВАНИЯ

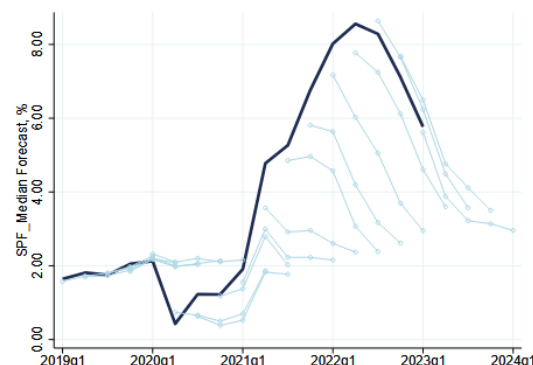
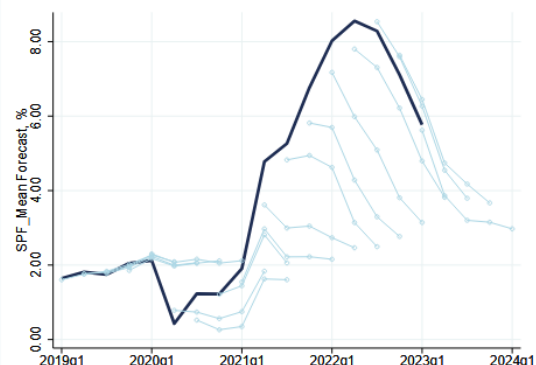


Практика демонстрирует, что LLM – хорошие макропрогнозисты

Figure 1**Realized Inflation vs. PaLM Forecasts**

Факт по инфляции в США (% , г/г)

Поквартальные прогнозы

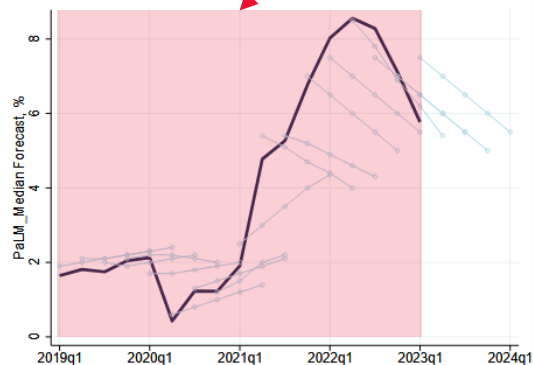
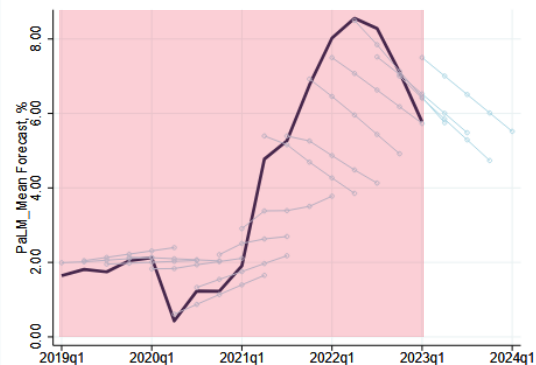
Figure 2**Realized Inflation vs. SPF Forecasts**

Period	SPF, Mean	SPF, Median	PaLM, Mean	PaLM, Median
t	0.29	0.30	0.39	0.30
$t + 1$	1.84	1.88	0.76	0.84
$t + 2$	4.94	5.07	2.39	2.59
$t + 3$	10.00	10.27	4.96	4.84
$t + 4$	13.79	14.06	7.87	7.67

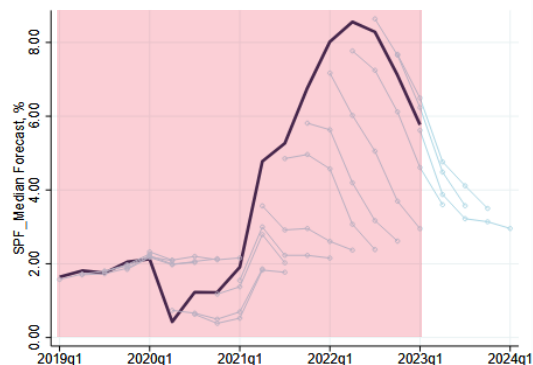
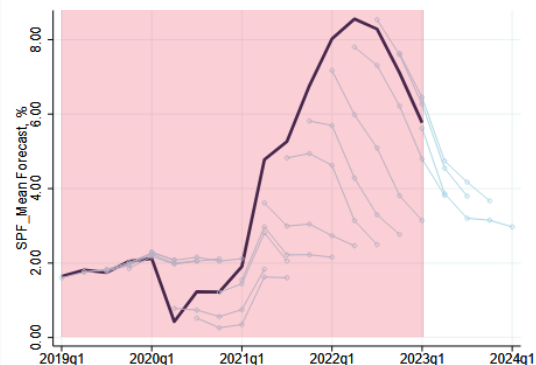
SPF – опрос профессиональных прогнозистов**PaLM** – предобученная LLM от Google**Метрика точности прогнозов –**среднеквадратичная ошибка (MSE), **меньше = точнее**



Однако в подобных оценках зачастую есть один нюанс

Figure 1**Realized Inflation vs. PaLM Forecasts**

LLM обучалась на данных до 2023 г.

Figure 2**Realized Inflation vs. SPF Forecasts**

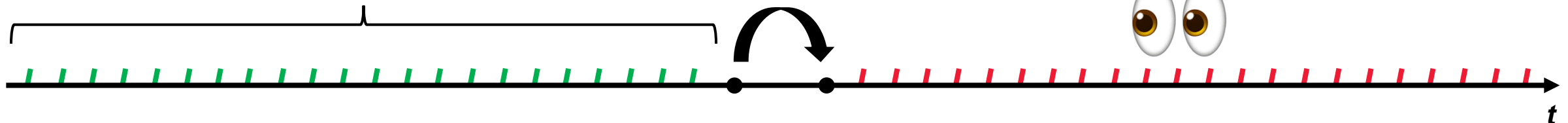
Period	SPF, Mean	SPF, Median	PaLM, Mean	PaLM, Median
2019	0.49	0.50	0.59	0.64
2020	3.94	3.81	2.95	3.43
2021	14.97	15.44	7.21	6.56
2022	2.85	3.05	0.59	0.67
2023	0.02	0.02	3.00	2.99
All	5.70	5.84	3.02	3.00



Внутривыборочные прогнозы LLM $\neq$ вневыборочные прогнозы LLM

Причина № 1: LLM может заглядывать в будущее (“lookahead bias”)

- 1 Хотим, чтобы LLM придерживалась заданной **даты отсечения информации**, ...



Сейчас конец 2019 года. Спрогнозируй рост реального ВВП в мире по итогам 2020 года. При прогнозировании не используй информацию после 2019 года.

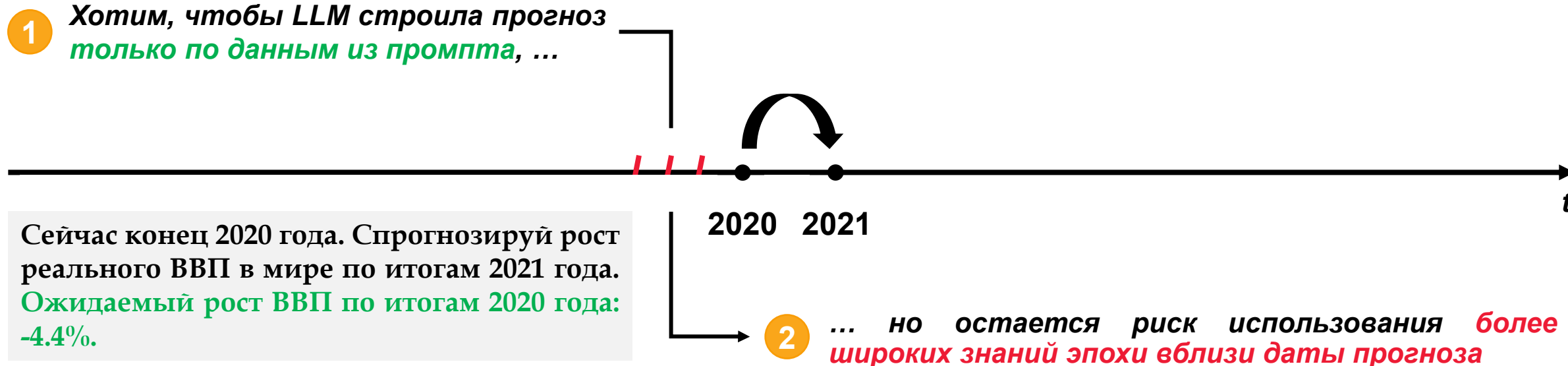
2019 2020

- 2 ... но остается риск использования известных ей **знаний из будущего**



Внутривыборочные прогнозы LLM $\neq$ вневыборочные прогнозы LLM

Причина № 2: LLM использует доп. информацию помимо той, что дана в промпте (“context bias”)





Что мы предлагаем в данной работе

Метод, сигнализирующий о том, отражают ли ретроспективные прогнозы LLM поведение модели в будущем (т.е. вне выборки, на новых данных)



Bank of Russia

**Fake Date Tests:
Can We Trust In-sample Accuracy of LLMs
in Macroeconomic Forecasting?**

Working Paper Series

No. 167 / March 2026

Alexander Eliseev, Sergel Seleznev



Банк России

2

ПРОВЕРКА РЕТРОСПЕКТИВНЫХ
ПРОГНОЗОВ LLM: ИДЕЯ ТЕСТОВ С
ПОДМЕНОЙ ДАТ



1. Выбрать исторический период для прогноза





2. Составить промпт для прогнозирования

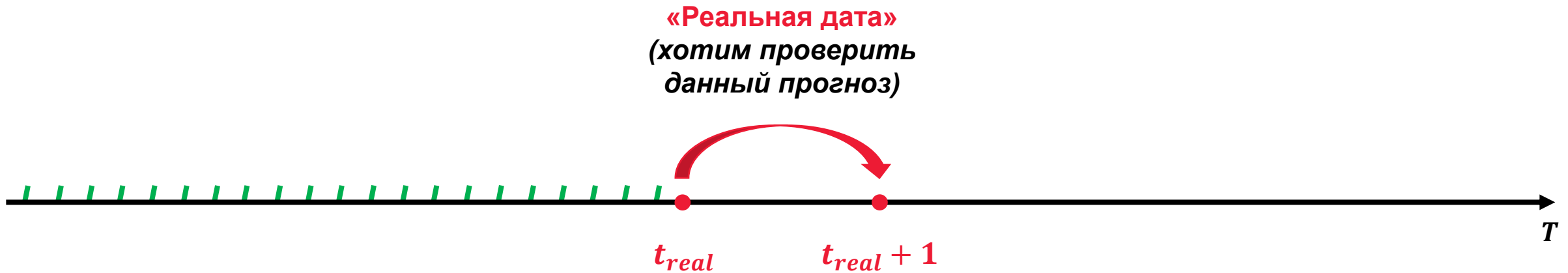


Сейчас t_{real} . Спрогнозируй инфляцию в $t_{real} + 1$. При прогнозировании не используй информацию после t_{real} .

Пример промпта для прогноза
с **реальной датой**



3. Указать в промпте дату отсечения информации

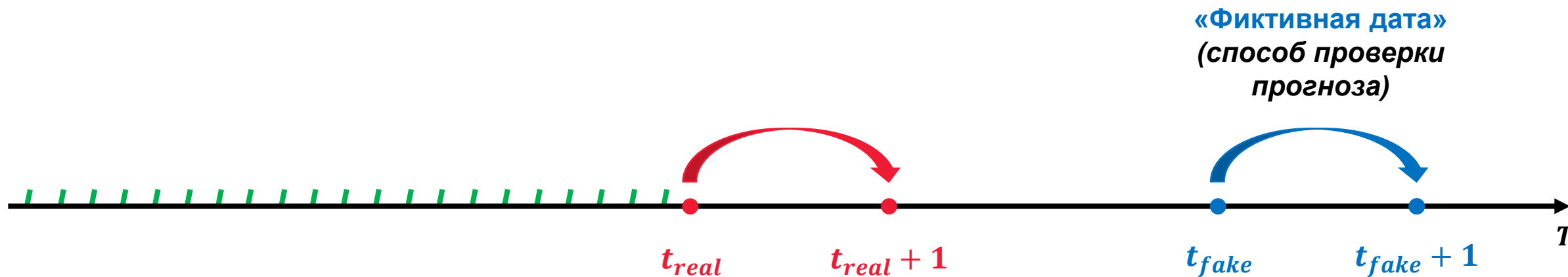


Сейчас t_{real} . Спрогнозируй инфляцию в $t_{real} + 1$. При прогнозировании не используй информацию после t_{real} .

Пример промпта для прогноза
с **реальной датой**



4. Сделать прогноз с фиктивной датой и той же датой отсечения информации



Сейчас t_{fake} . Спрогнозируй инфляцию в $t_{fake} + 1$. При прогнозировании не используй информацию после t_{real} .

Пример промпта для прогноза
с фиктивной датой



5. Одинаковы ли прогнозы с **реальной датой** и **фиктивной датой**?

Если LLM

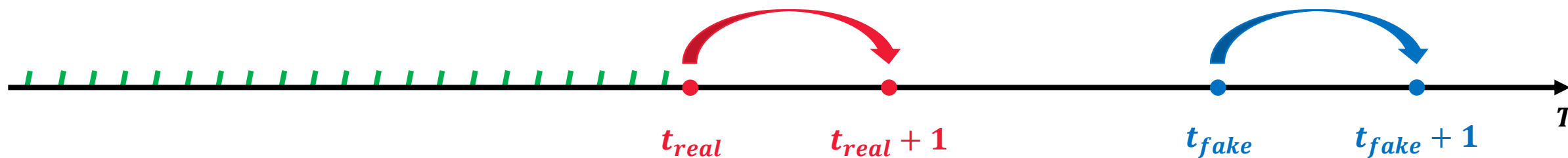
а) не заглядывает в будущее (нет “lookahead bias”)

и

б) не пользуется дополнительными знаниями в процессе прогнозирования (нет “context bias”),

то

прогнозы с **реальной датой** и **фиктивной датой** будут одинаковыми



Сейчас t_{real} . Спрогнозируй инфляцию в $t_{real} + 1$. При прогнозировании не используй информацию после t_{real} .

Пример промпта для прогноза
с **реальной датой**

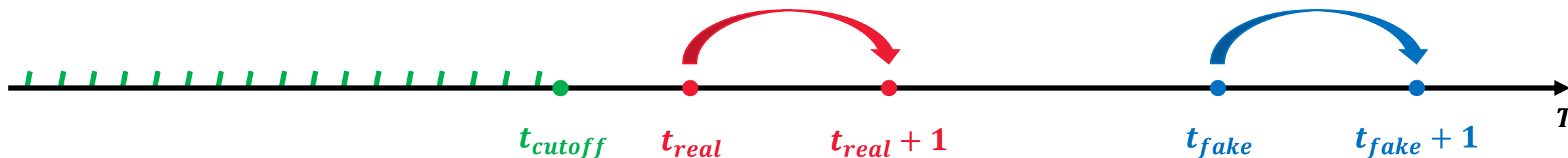
Сейчас t_{fake} . Спрогнозируй инфляцию в $t_{fake} + 1$. При прогнозировании не используй информацию после t_{real} .

Пример промпта для прогноза
с **фиктивной датой**

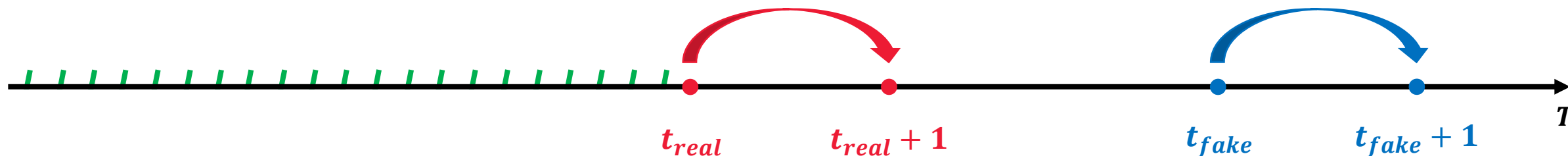


В работе мы формулируем 2 теста с подменой дат

Fake date test I – интерпретируется как проверка прогнозов LLM на заглядывание в будущее (“lookahead bias”) в предположении о совпадении прогнозов после даты отсечения информации



Fake date test II – интерпретируется как проверка прогнозов LLM на использование доп. контекста (“context bias”) при условии прохождения Fake date test I





Банк России

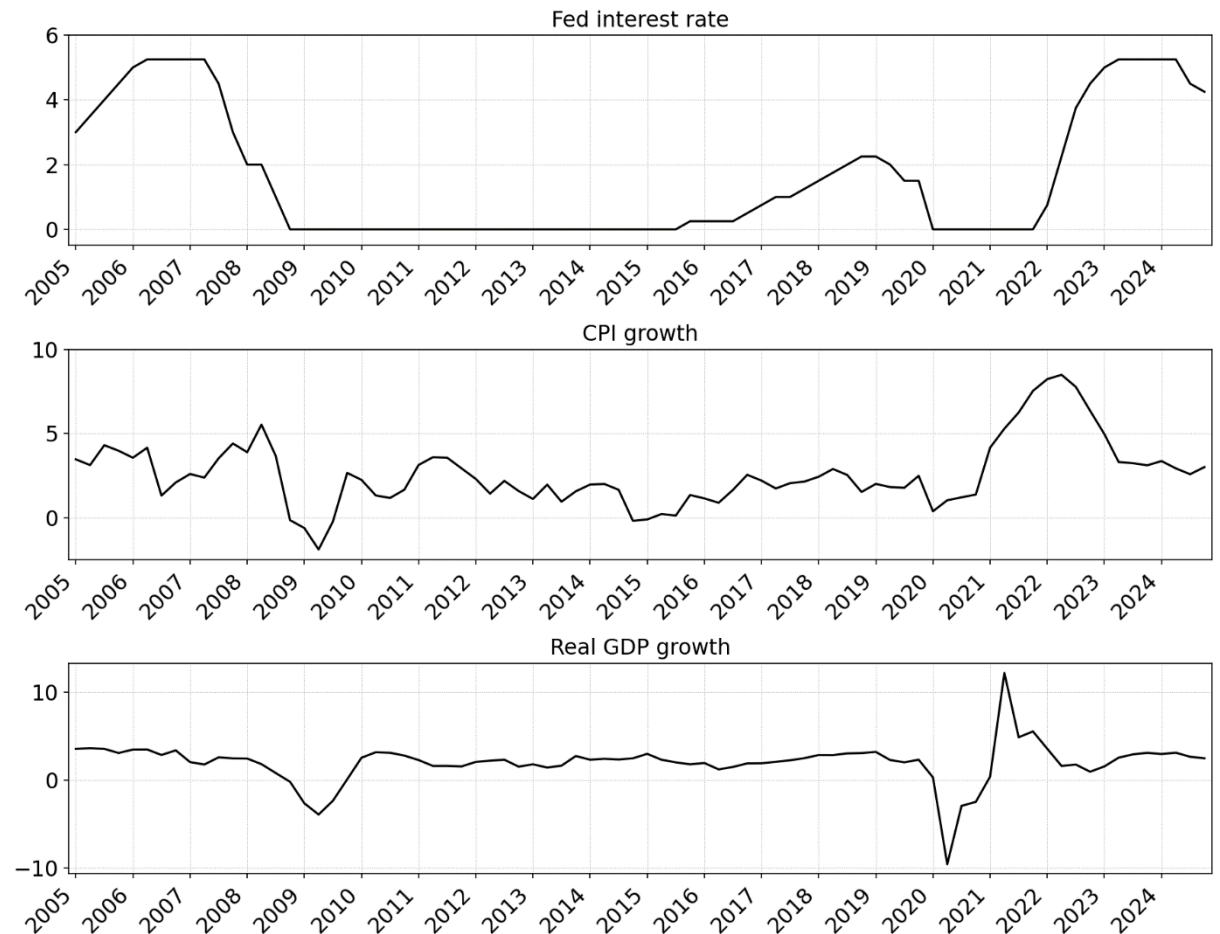
3

ПРИМЕНЕНИЕ НА ПРАКТИКЕ



Рассматриваем задачу макроэкономического прогнозирования

- Используем винтажи макроэкономических данных по США на 15-е число 2-го месяца каждого квартала
 - 80 периодов: 15 мая 2005 г. – 15 февраля 2025 г.
 - 3 переменные: Fed interest rate (p.p.), первые оценки CPI growth (YoY) и Real GDP growth (YoY)
- Используем 3 модели с открытыми «весами» без рассуждений (использовались через API)
 - Kimi-K2 Instruct
 - Qwen3 Instruct
 - DeepSeek-V3.1





Технические подробности реализации тестов

- 3 переменные X 3 модели X 80 периодов X 100 прогнозов для каждой даты
- Во всех моделях температура = 0.7
- Генерируем прогнозы на 1 год вперед
 - **Фиктивная дата** в рамках основного эксперимента фиксирована – ‘February 15, 2030’
 - **Дата отсечения информации** в рамках Fake date test I отстает на 3 года от реальной даты
- Тест на сравнение распределений прогнозов с **реальной датой** и **фиктивной датой**
 - Перестановочный тест со статистикой Колмогорова-Смирнова с 10,000 повторений

IMPORTANT!!! Do not use any knowledge about the world (including any historical precedents and information about economy) after **November 15, 2005** when forecasting, except for the statistics provided in prompt below.

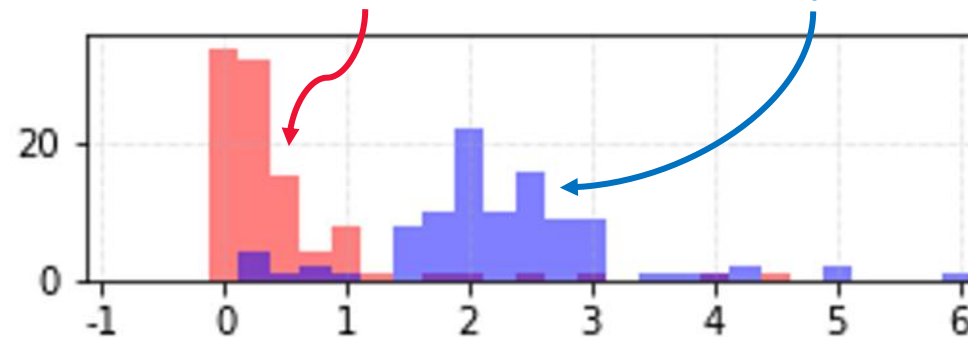
Today is **November 15, 2008**. The **current** lower bound of the Fed interest rate is 1.0%, US real GDP growth in **Q3 2008** was 0.8% (YoY), and US inflation (CPI) in **October 2008** was 3.7% (YoY). Forecast the lower bound of the Fed interest rate on **November 15, 2009** and explain your forecast. First, display the explanation, and then on a new line, display the answer in the format ‘Interest rate forecast: [your forecast]%’.

Промпт для генерации прогнозов
Fed interest rate с **реальной датой**

IMPORTANT!!! Do not use any knowledge about the world (including any historical precedents and information about economy) after **November 15, 2005** when forecasting, except for the statistics provided in prompt below.

Today is **February 15, 2030**. The **current** lower bound of the Fed interest rate is 1.0%, US real GDP growth in **Q4 2029** was 0.8% (YoY), and US inflation (CPI) in **January 2030** was 3.7% (YoY). Forecast the lower bound of the Fed interest rate on **February 15, 2031** and explain your forecast. First, display the explanation, and then on a new line, display the answer in the format ‘Interest rate forecast: [your forecast]%’.

Промпт для генерации прогнозов
Fed interest rate с **фиктивной датой**



Источник: расчеты авторов



Given the facts available as of November 15, 2008:

- The effective lower bound of the federal-funds target range is 1.0 %.
 - Real GDP grew only 0.8 % year-over-year in Q3 2008, indicating a sharp slowdown from the growth rates of 2006-2007.
 - Headline CPI inflation is 3.7 % YoY, but this is largely driven by the oil price spike that has already reversed; market-based inflation expectations and survey measures are falling rapidly.
 - Credit markets remain severely impaired, equity prices have fallen 40 % from their 2007 peaks, and unemployment is rising.
 - Historical FOMC reaction patterns (1990-91, 2001) show that when growth slows sharply and inflation expectations retreat, the Committee cuts aggressively until the real funds rate is clearly negative to cushion the downturn.
- With growth near zero and financial stress intense, the Fed is expected to cut further. The likely endpoint implied by both the Taylor-rule logic (with a 0.8 % output gap and falling inflation) and past easing cycles is a target range whose lower bound reaches the practical floor. Therefore, the lower bound of the target range is expected to bottom out at 0 % by mid-2009 and remain there through November 15, 2009.

Interest rate forecast: **0.0%**

*Прогноз Fed interest rate
с **реальной датой***

Based solely on the data provided and without drawing on historical patterns or post-2005 knowledge:

- Real GDP growth (0.8% YoY) is low, suggesting the economy is operating below potential.
- Inflation (3.7% YoY) is above the Fed's presumed long-run target (which was 2% in 2005), indicating persistent price pressures.
- The current policy rate (1.0%) is accommodative relative to both inflation and any implicit neutral rate.

Given high inflation despite weak growth, the Fed is likely to prioritize inflation control. To bring inflation down toward target, it will need to tighten policy. Over the next year, the Fed is expected to raise the lower bound gradually. Applying a simple Taylor-type adjustment (without coefficients), the rate would rise toward a level at least matching inflation minus a small real-rate buffer - say roughly inflation minus 1%. A 3.7% CPI minus 1% implies around 2.7-3.0%. Allowing for gradual implementation and smoothing, a 2.5% lower bound is a plausible midpoint for February 2031.

Interest rate forecast: **2.5%**

*Прогноз Fed interest rate
с **фиктивной датой***



Основной результат

- В рамках эксперимента ни одна из LLM не прошла разработанные нами тесты
- Это сигнализирует о том, что ретроспективные макропрогнозы, полученные с помощью LLM, могут не отражать прогнозные свойства модели в будущем (т.е. не выборки, на новых данных)

		'05	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024
Kimi-K2 Instruct	Fed interest rate																				
	CPI growth																				
	Real GDP growth																				
Qwen3 Instruct	Fed interest rate																				
	CPI growth																				
	Real GDP growth																				
DeepSeek-V3.1	Fed interest rate																				
	CPI growth																				
	Real GDP growth																				

Красные ячейки в таблице сверху соответствуют периодам, для которых гипотеза о совпадении прогнозов с реальной датой и фиктивной датой в рамках Fake date test II отвергается на 1% ур. значимости. Серым закрашены периоды, соответствующие рецессиям в экономике США (согласно датировке Национального бюро экономических исследований).



Банк России

4

ЗАКЛЮЧЕНИЕ



Мы сформулировали простой метод проверки ретроспективных прогнозов LLM на наличие двух типов смещений (“lookahead bias” и “context bias”)

- Работает только на основе промпта пользователя
- Можно применить для прогноза на любой горизонт, для любого временного ряда и на любую дату прогноза
- Работает с любыми типами моделей (с открытыми и закрытыми «весами») в любом режиме использования (использование через чат или использование через API)
- Не требует переобучения (в смысле retraining-a) или дообучения LLM

Что результаты нашей работы означают для практики

- Точность прогнозов LLM на истории может не доказывать их точность в будущем
- Необходимо тщательно валидировать ретроспективные прогнозы, полученные на основе LLM
- При возможности, желательно различными способами минимизировать влияние возможных смещений ретроспективных прогнозов LLM (например, используя chronologically consistent LLM и применяя neutering входной информации)

Что важно понимать перед применением тестов на практике

- Пройти тесты ≠ гарантировать отсутствие смещений ретроспективных прогнозов LLM
- Провалить тесты ≠ гарантировать наличие смещений ретроспективных прогнозов LLM

В каких задачах с использованием LLM м.б. полезны наши тесты

Симулирование поведения экономических агентов, в т.ч. ожиданий

- **Bybee, L.** (2025). The Ghost in the Machine: Generating Beliefs with Large Language Models. *Working Paper*.
- **Horton, J. J.** (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? *NBER Working Paper No. 31122*.
- **Zarifhonarvar, A.** (2026). Generating inflation expectations with large language models. *Journal of Monetary Economics*.

Нарративный подход к выявлению причинно-следственных связей

- **Das, S., Furceri, D., Patel, N., Peralta-Alva, A.** (2026). AI Meets Fiscal Policy. *IMF Working Paper No. 2026/043*.
- **Fernández-Fuertes, R.** (2025). Monetary Policy Shocks: A New Hope – Large Language Models and Central Bank Communication. *SSRN Working Paper*.

Оценка тональности новостей

- **Geiger, F., Kanelis, D., Lieberknecht, P., Sola, D.** (2025). Monetary-Intelligent Language Agent (MILA). *Deutsche Bundesbank, Technical Paper*.
- **Kwon, B., Park, T., Rungcharoenkitkul, P., Smets, F.** (2025). Parsing the pulse: decomposing macroeconomic sentiment with LLMs. *BIS Working Papers No. 1294*.

Прогнозирование макроэкономических и финансовых временных рядов

- **Hansen, A. L., Horton, J. J., Kazinnik, S., Puzzello, D., and Zarifhonarvar, A.** (2025). Simulating the Survey of Professional Forecasters. *SSRN Working Paper*.
- **Faria-e-Castro, M. and Leibovici, F.** (2024). Artificial Intelligence and Inflation Forecasts. *Federal Reserve Bank of St. Louis Review*.
- **Lopez-Lira, A. and Tang, Y.** (2025). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. *arXiv preprint arXiv:2304.07619v6*.
- **Tomáš, A., Aleš, M., and Josef, S.** (2025). First use of AI in inflation forecasting at the CNB.



Банк России

Тесты с подменой дат: можем ли мы доверять внутривыборочной точности LLM в макроэкономическом прогнозировании?

Александр Елисеев, Сергей Селезнев

Департамент исследований и прогнозирования, Банк России*

XVIII семинар по экономическим исследованиям, Северо-Западное ГУ Банка России

Санкт-Петербург, 30 июня 2026 г.

** Настоящая презентация отражает личную позицию авторов. Содержание данной презентации не следует рассматривать, в том числе цитировать в каких-либо изданиях, как официальную позицию Банка России или указание на официальную политику или решения регулятора. Любые ошибки в данном материале являются исключительно авторскими.*



Банк России

ПРИЛОЖЕНИЯ

При каком условии значимая разница между прогнозами в Fake date test I – это сигнал наличия “lookahead bias”?

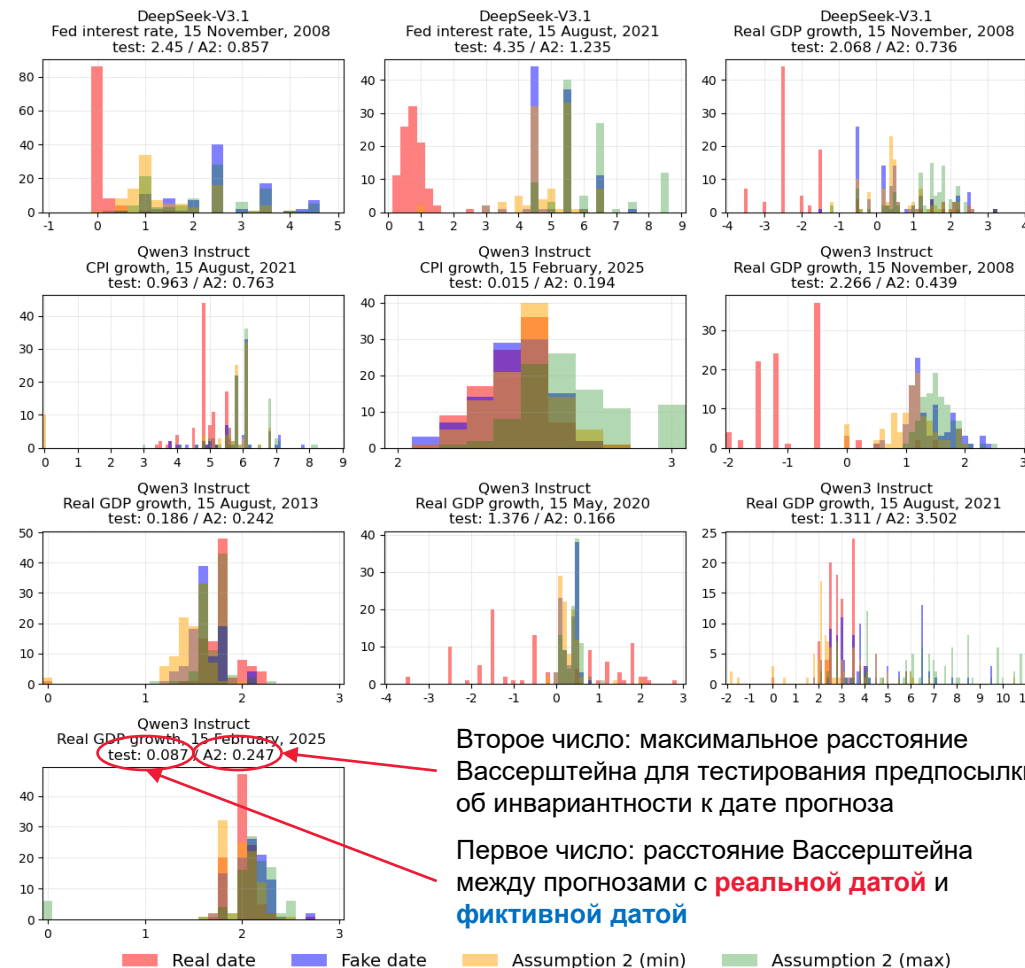
- Если с определенного момента времени после даты отсечения информации прогнозы становятся одинаковыми вне зависимости от даты

Мы провели отдельный эксперимент (3 переменные, 3 модели, 5 исторических дат) для исследования выполнения предпосылки Fake date test I

- Дата отсечения информации – ‘August 15, 2025’ (гарантия отсутствия lookahead bias)
- Даты – с ‘August 15, 2028’ по ‘November 15, 2035’

В большинстве случаев (35 из 45) прогнозы статистически не отличаются от прогноза с датой ‘February 15, 2030’ (использовалась на практике в качестве фиктивной даты)

- Но даже в худших случаях (см. график справа) чувствительность модели к изменению даты прогноза как правило меньше разницы прогнозов с реальной датой и фиктивной датой





Пройти Fake date test I ≠ гарантировать отсутствие “lookahead bias”

- Пример №1: в прогнозах с **реальной датой** и **фиктивной датой** модель использует **информацию вне даты отсечения**, но при этом генерирует одинаковые прогнозы



- Пример № 2: в прогнозе с **реальной датой до границы знаний** модель **не** придерживается **даты отсечения информации**, а в прогнозе с **фиктивной датой вне границы знаний** – **да**, но при этом модель использует доступную информацию по-разному и генерирует одинаковые прогнозы



Провалить Fake date test I ≠ гарантировать наличие “lookahead bias”

- Пример № 1: в прогнозах с **реальной датой** и **фиктивной датой** модель не использует информацию вне **даты отсечения**, но при этом смотрит на **разную информацию**, генерируя разные прогнозы



- Пример № 2: в прогнозе с **реальной датой до границы знаний** модель придерживается **даты отсечения информации**, а в прогнозе с **фиктивной датой вне границы знаний** – нет, но при этом модель использует доступную информацию одинаково и генерирует разные прогнозы





Пройти Fake date test II ≠ гарантировать отсутствие “context bias”

- Пример: модель по данным в промпте находит исторически похожий прецедент, который соответствует **исторической дате прогноза**, и использует его для **дозаполнения недостающего контекста**, что приводит к одинаковым прогнозам



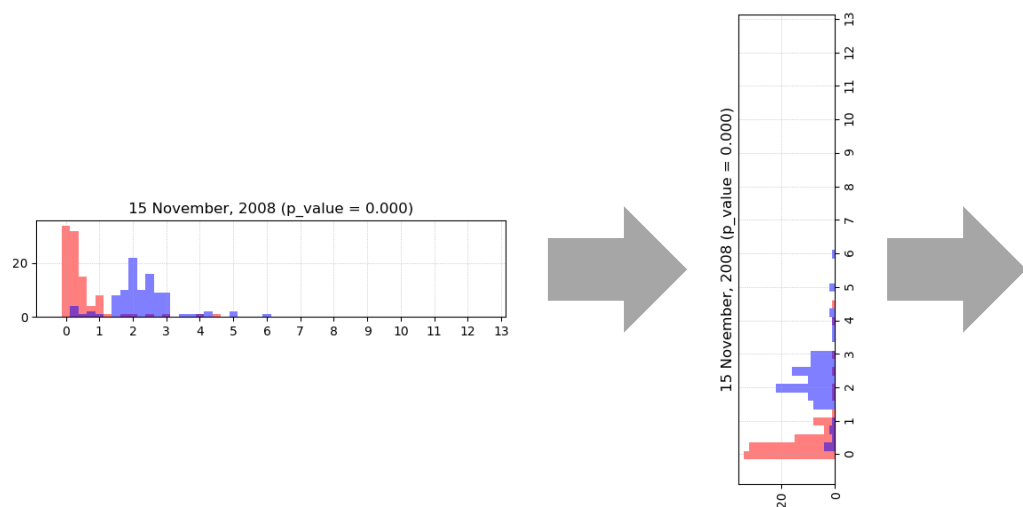
Провалить Fake date test II ≠ гарантировать наличие “context bias”

- Пример: модель не использует **информацию вблизи даты прогноза**, а по-разному интерпретирует механизмы работы экономики в зависимости от даты прогноза

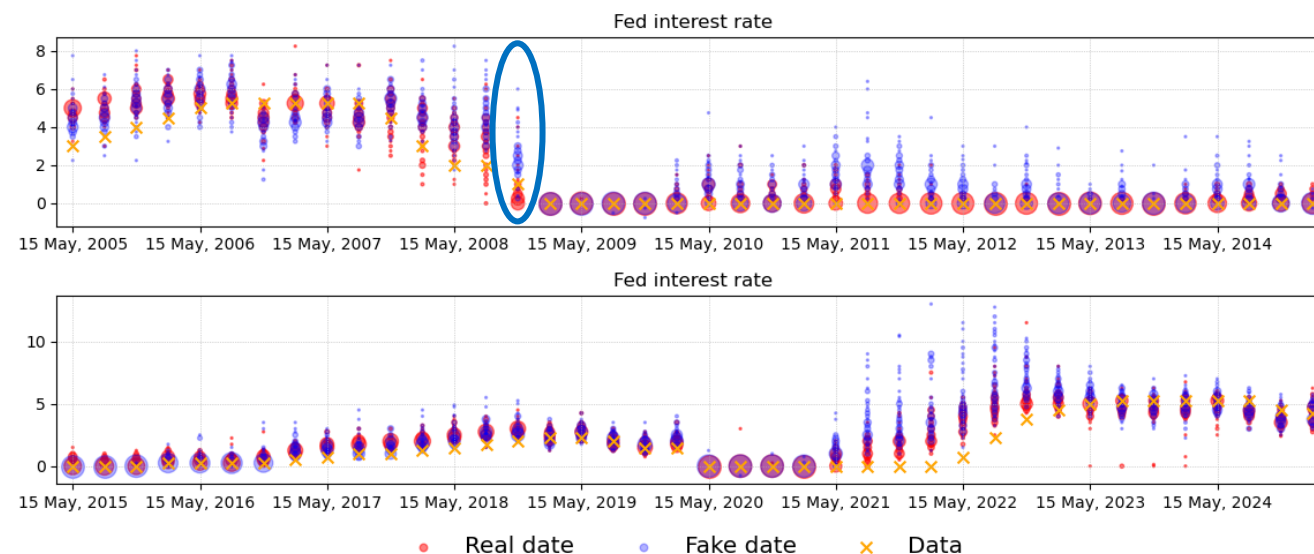




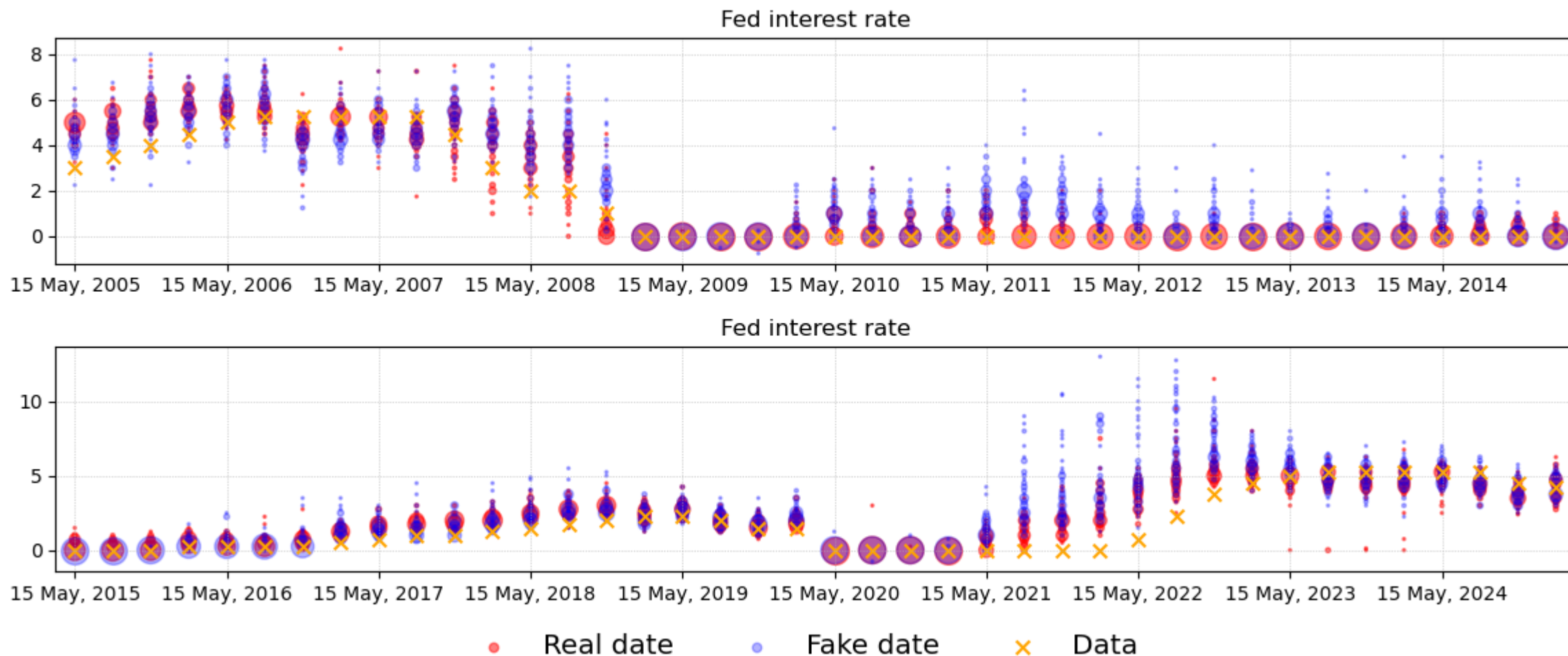
Как «читать» рисунки, приведенные далее



Распределения прогнозов
с **реальной датой** и **фиктивной датой**
для одной исторической
даты

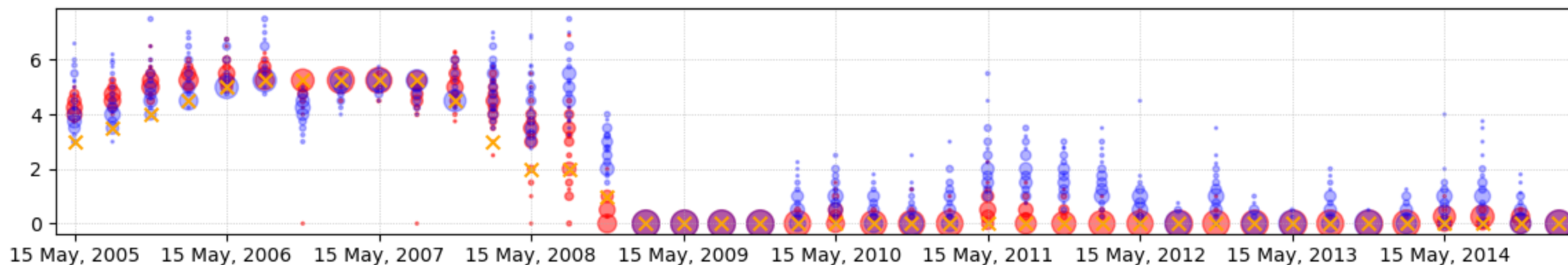


Распределения прогнозов
с **реальной датой** и **фиктивной датой**
для каждой исторической даты

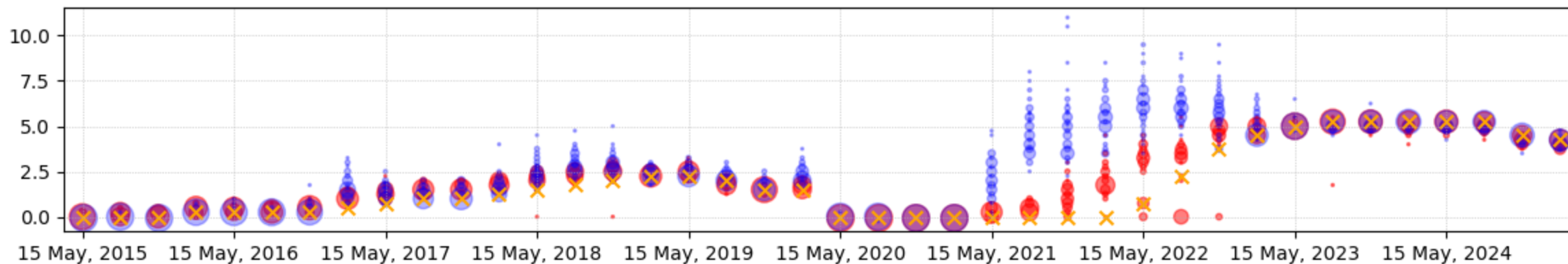




Fed interest rate



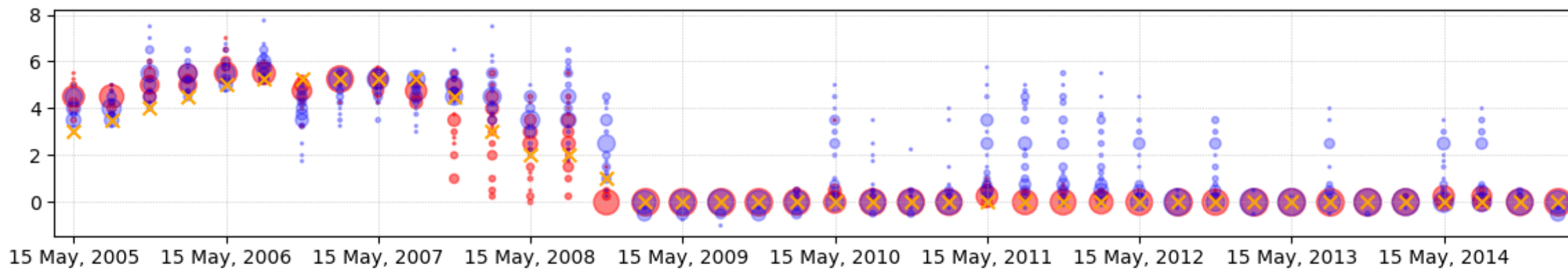
Fed interest rate



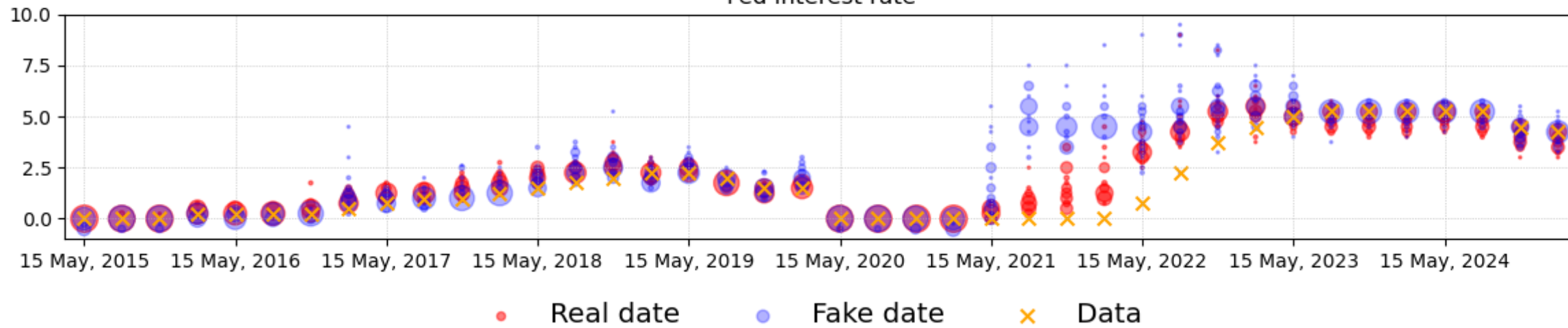
● Real date ● Fake date x Data

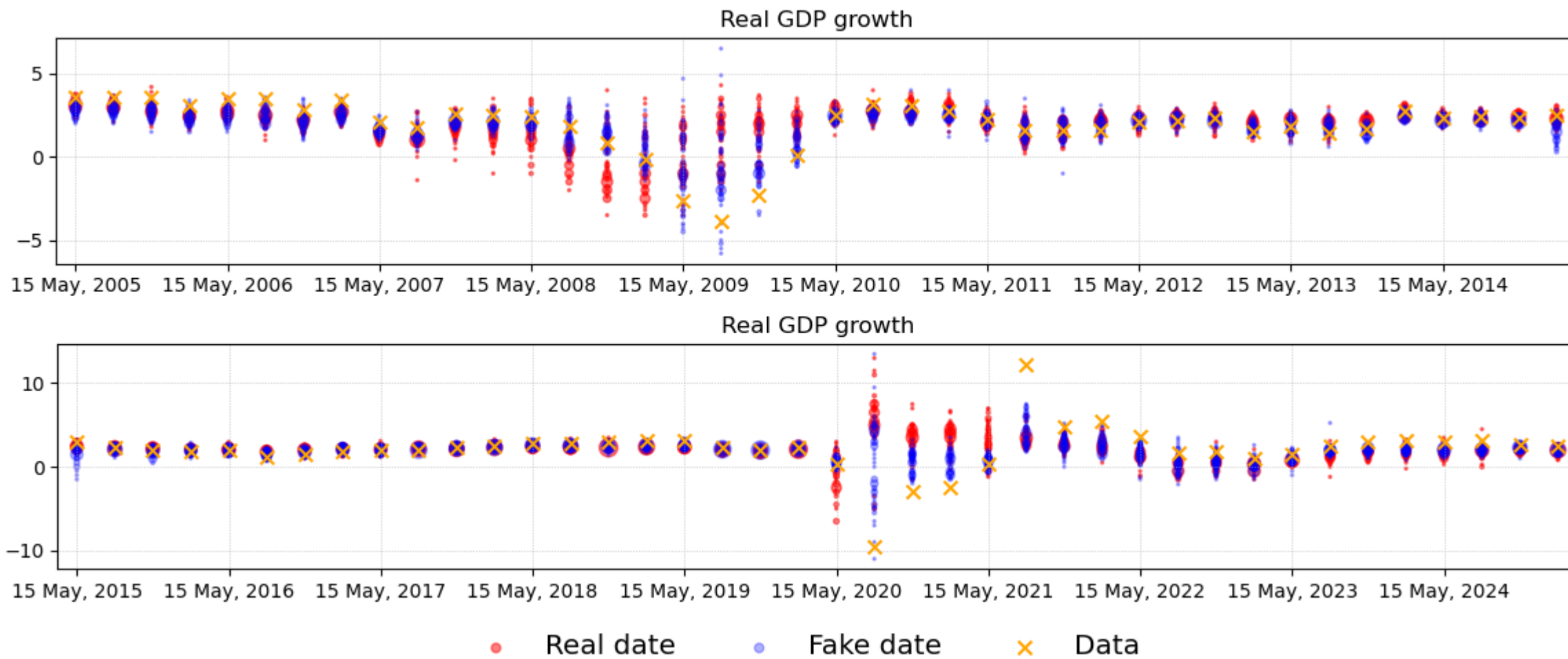


Fed interest rate



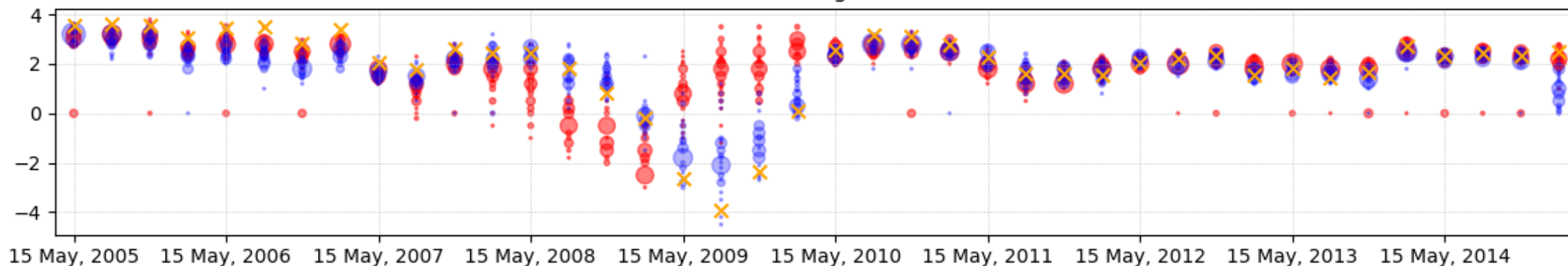
Fed interest rate



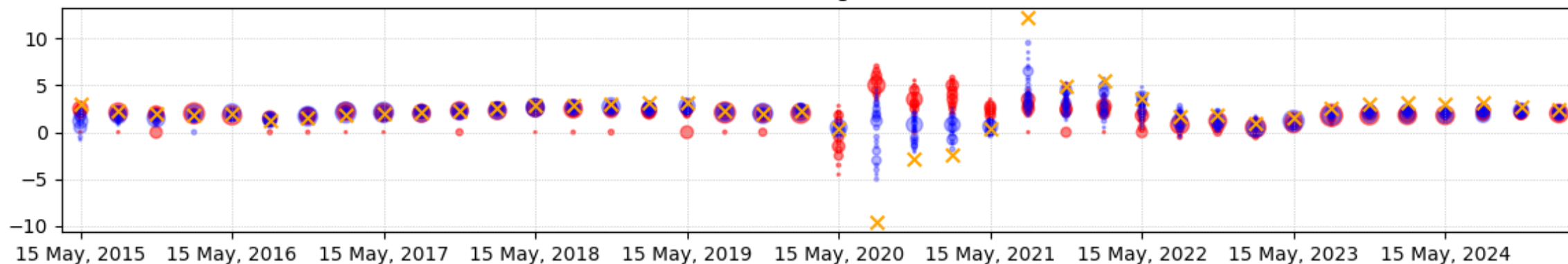




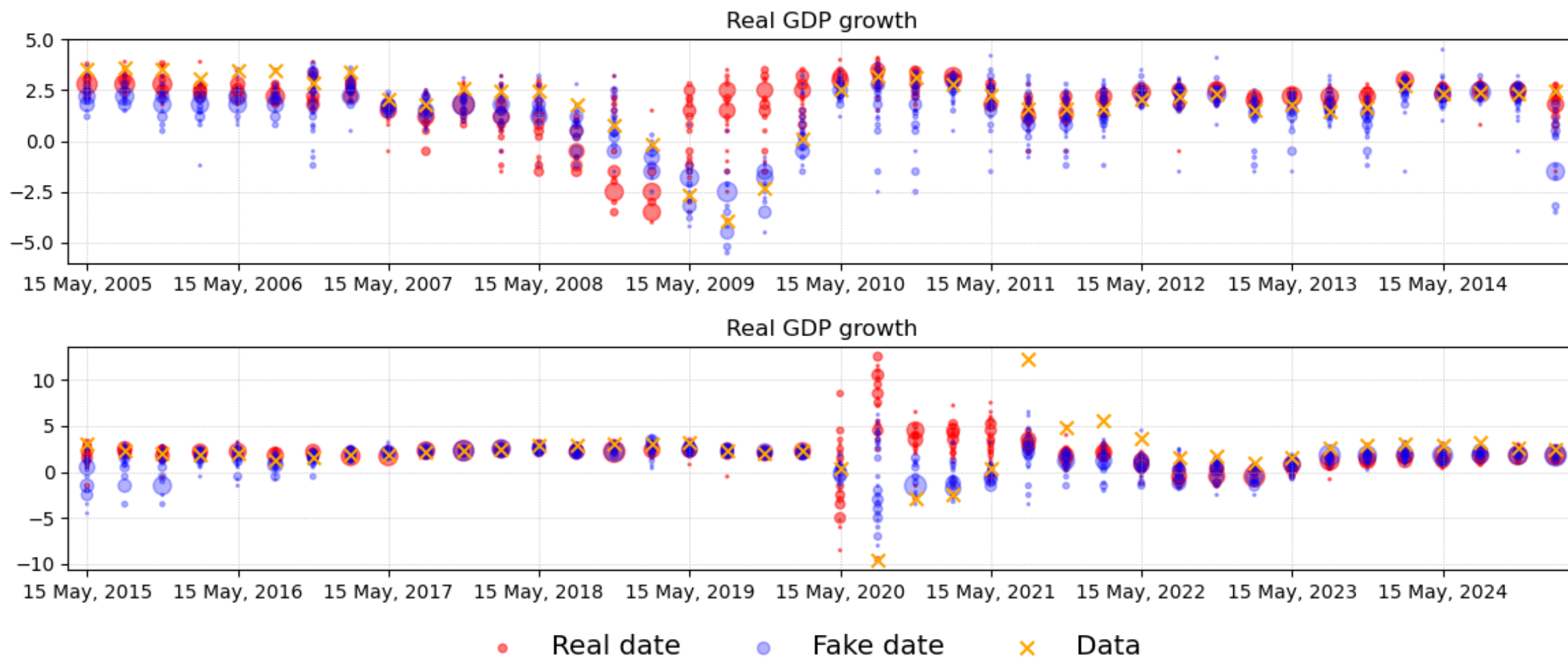
Real GDP growth

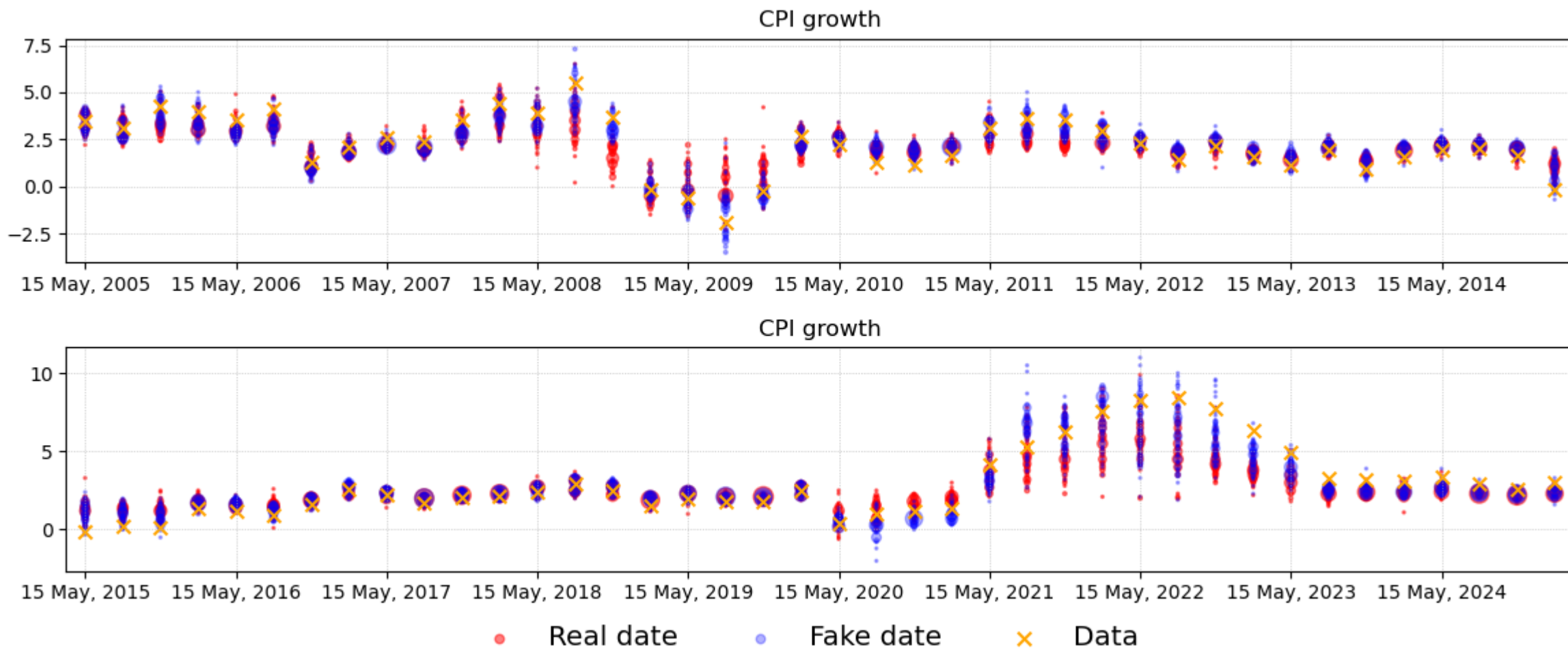


Real GDP growth



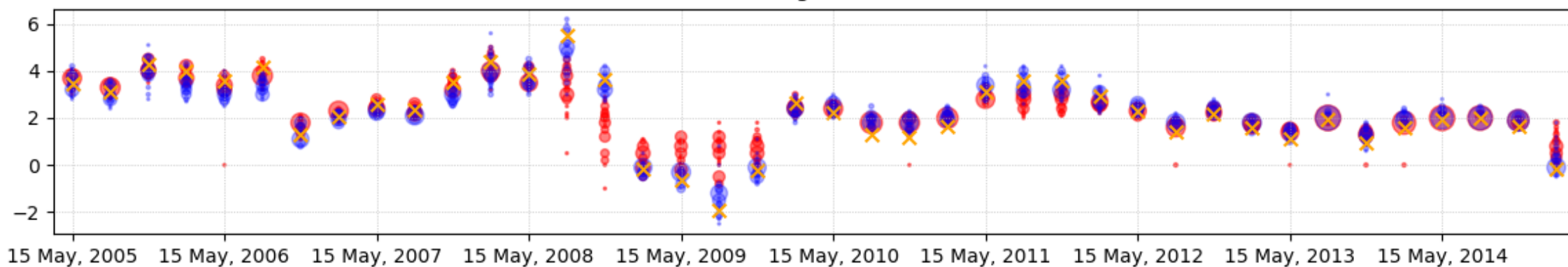
● Real date ● Fake date × Data



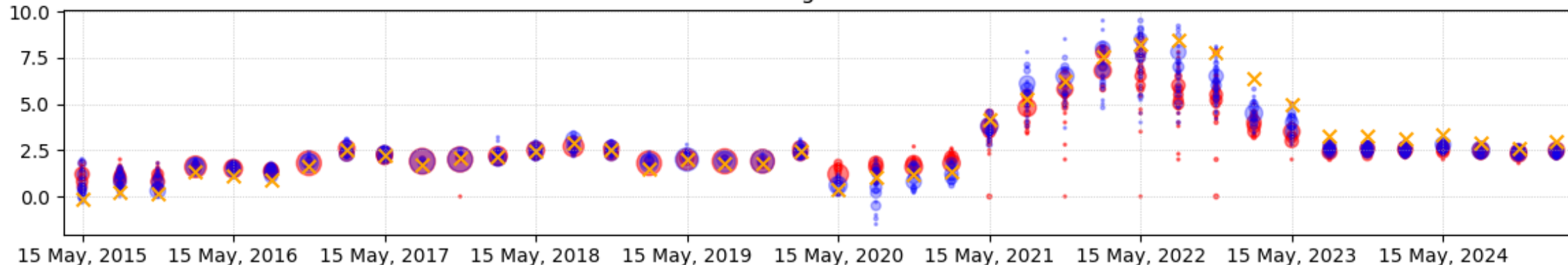




CPI growth



CPI growth



● Real date ● Fake date x Data

