РЭШ
Российская экономическая школа

XVIII СЕМИНАР ПО ЭКОНОМИЧЕСКИМ ИССЛЕДОВАНИЯМ

«Искусственный интеллект и экономика»

Дискуссия к докладу С. Селезнева и А. Елисеева

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

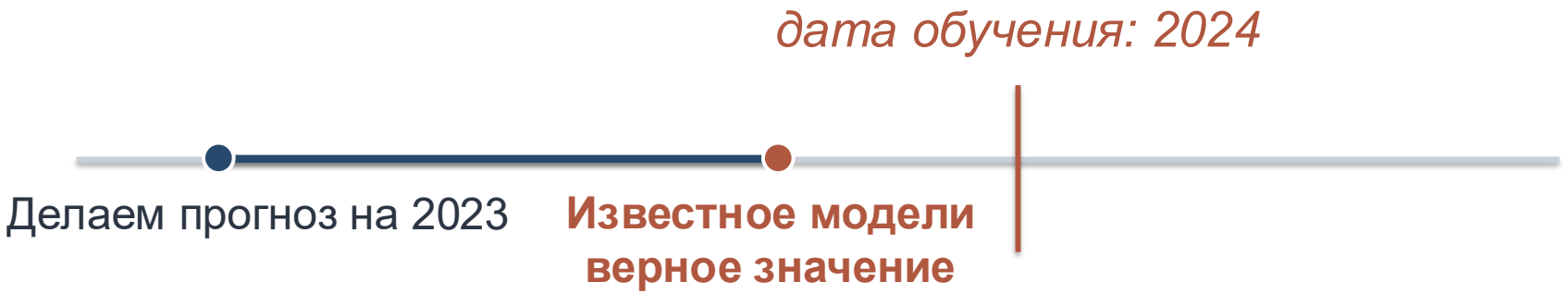
Дискуссант: Марат Салихов

30 июня 2026 года

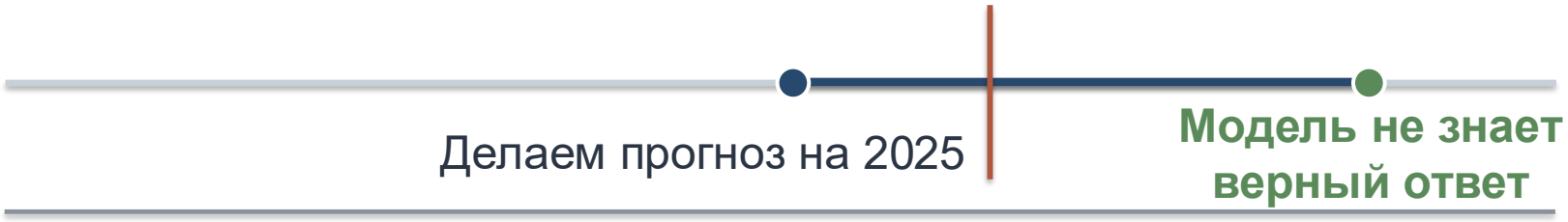
ГЛАВНАЯ ПРОБЛЕМА

*Мы проверяем точность прогностических моделей, сравнивая их прогнозы с реальными результатами на исторических данных. С прогнозами больших языковых моделей возникает проблема: она может **подглядывать**, так как ее обучали на данных, включающих в себя реальные результаты.*

Если мы прогнозируем на дату до **отсечки обучения**, модель может подглядывать



Честные прогнозы возможны только **после отсечки**

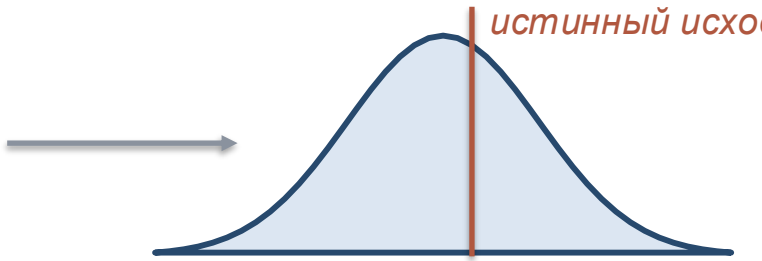


КАК ТЕСТИРУЕТСЯ ПОДГЛЯДЫВАНИЕ

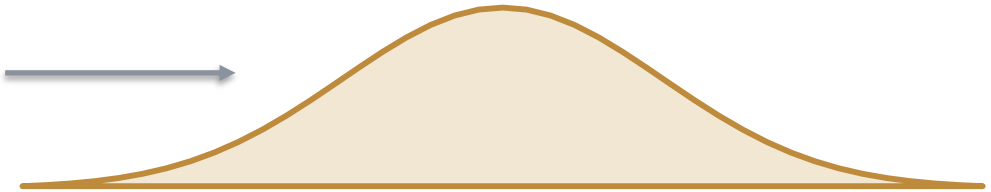
Меняем дату, на которую просим прогноз,
и корректируем исторические данные

Генерируем 100 прогнозов

Промпт с реальной датой
15 мая 2021



Промпт с подложной датой
15 февраля 2030



Тест Колмогорова-Смирнова
Если распределения различаются, то это сигнал, что модель подглядывает

Результаты
Ни одна модель не прошла тест. Мы сделали небольшую репликацию с новой моделью и получили схожие результаты

АЛЬТЕРНАТИВНЫЕ ТЕСТЫ

Gao-Jiang-Yan (июнь 2026). *Detecting Lookahead Bias in LLM Forecasts.*

1. Смотрят на помесячные прогнозы фондового рынка.
2. Измеряют показатель вероятности подглядывания (LAP, ПВП) – на вход отправляют только название фирмы и дату, на выходе измеряют вероятности модели для трех возможных исходов: вверх, вниз, неизвестно (непосредственно через логиты на последнем слое модели). Берут сумму вероятностей «вверх» и «вниз».
3. До отсечки обучения: ПВП положительный. После: нулевой.
4. В предсказательной регрессии доходности акций тестируют равенство нулю для взаимодействия между словесным прогнозом и ПВП. Если корреляция между предсказанием и реальным исходом становится сильнее при высоком ПВП, то у нас есть утечка.

Этот подход может дополнить подход с подложными датами – если хотя бы один тест выдал положительный результат, мы можем подозревать подглядывание.

ПОЧЕМУ БЫ ПРОСТО НЕ ПОПРОСИТЬ?

РЭШ

Может быть, можно указать в промпте, чтобы утечки не было?

Wu, Xi and Xie (апрель 2026): *LLM Survey Framework: Coverage, Reasoning, Dynamics, Identification*

Они используют большие языковые модели, чтобы симулировать опросы домохозяйств об инфляционных ожиданиях на определенную дату – в промптах просят игнорировать знание определенных событий и получают, что сгенерированные результаты похожи на исторические данные опросов.

Противоречит ли это статье? Скорее нет – результаты Wu, Xi и Xie как раз и говорят о том, что модель сохраняет результаты опросов. Может ли это помочь? Интересно протестировать.

Мы не можем гарантировать того, что промпты заблокируют «подглядывание», так как большая языковая модель – это черный ящик

НЕ ОБУЧАТЬ МОДЕЛИ НА НОВЫХ ДАННЫХ

РЭШ

Капсульные LLM: просто обучить на данных, например, до 2015 года.

В работе цитируются похожие статьи: например, ChronoGPT. Такие модели раньше были недостаточно мощными.

Сейчас ситуация меняется. Недавно вышла модель Talkie-13b, обученная на данных до 1930 года. Еще одно семейство: Point-in-Time модели.

Kelly, Malamud, Schwab, Xu (май 2026): Scaling Point-in-Time Language Models

https://huggingface.co/Diamegs/PIT-4B-201511
запустить не так просто, как, например, модели с BotHub или OpenRouter

НЕ ТОЛЬКО ПОДГЛЯДЫВАНИЕ

Большие языковые модели – это черный ящик, и смещение из-за подглядывания – это одна из многих проблем.

Mazeika et al. (2025): у больших языковых моделей можно оценить систему ценностей. Они, например, ценят жизнь людей из разных стран по-разному.

Turpin et al. (2023): цепочка размышлений модели не соответствует реальным процессам внутри модели.

Mechanistic interpretability (Neel Nanda): мы можем частично понять, как устроен черный ящик, экспериментируя с промптами и наблюдая, как модель на них реагирует на уровне весов – а затем использовать это для управления моделью (например, фреймворк Dwarfstar позволяет таким образом управлять моделью DeepSeek V4).