NOVA SAPIENS RESEARCH

FREE · ТОП-ВЫБОРКА

Топ-выборка

ЯНВАРЬ 2026

Ключевые исследования месяца — отобраны по практической применимости и новизне через 9-проходную фильтрацию

3,260
СТАТЕЙ ЗА МЕСЯЦ НА ARXIV

→

277
С РЕЙТИНГОМ 70+ (АНАЛИЗ)

→

40
КЛЮЧЕВЫХ РАБОТ

КАК ЭТО СОБРАНО

Каждое исследование прошло отбор по 5 шкалам: практическая применимость, концептуальная новизна, понятность без долгого контекста, генеративная сила, наличие готового шаблона/промпта. В этой выборке — только те, что прошли по 3+ стратегиям отбора одновременно.

№1 arxiv: 2601.02989 ★ 89 PRO ★ Weekly

System-2 Counting: преодоление архитектурного предела подсчета через разбиение на части

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№2 arxiv: 2601.04435 ★ 88 PRO ★ Weekly

Accommodation and Epistemic Vigilance: почему LLM не оспаривают ложные убеждения пользователя

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№3 arxiv: 2601.07354 ★ 88 PRO

MetaGlyph: математические символы вместо многословных инструкций

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№4 arxiv: 2601.14152 ★ 87 PRO

CQO vs QOC: почему порядок элементов промпта меняет точность на 15%

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№5 arxiv: 2601.22047 ★ 86 PRO ★ Weekly

Парадокс ограничений: почему избыток требований ломает LLM

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№6 arxiv: 2601.11227 ★ 84 PRO

Language of Thought: разнообразие ответов через смену языка мышления

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№7 arxiv: 2601.05114 ★ 84 PRO ★ Weekly

Evaluative Fingerprints: каждая LLM-судья оценивает по своей теории качества

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№8 arxiv: 2601.03444 ★ 83 FREE ★ Weekly

Шкала 0-5 максимизирует согласие между LLM и людьми при оценках

ЧТО ДЕЛАТЬ ПРАКТИКУ

Для субъективных оценок через LLM используй шкалу 0-5 с явными якорями границ — это даёт максимальное согласие с человеческим восприятием.

СУТЬ

Парадокс: Больше уровней оценки ≠ точнее результат. Шкала 0-10 даёт худшее согласие с людьми, чем 0-5, даже на одних и тех же задачах. Это не пересчёт баллов — модели **по-разному интерпретируют задачу** в зависимости от шкалы. Метод **позволяет получать оценки от LLM, которые максимально близки к человеческому восприятию** на субъективных задачах (качество текста, убедительность, стиль). Фишка: не количество уровней решает, а психологический паттерн — на 0-5 люди и модели калибруют оценки похоже, на 0-10 расходятся.

КАК ПРИМЕНИТЬ

Для субъективных задач (качество, убедительность, полезность) → используй шкалу **0-5**. Для **объективных** (токсичность, соответствие фактам) → шкала почти не влияет, бери любую. Чем субъективнее критерий оценки, тем сильнее эффект шкалы. Добавляй *якоря границ* — что значит 0 (худшее) и 5 (идеальное) конкретно для твоей задачи.

Открыть полную статью →

№9 arxiv: 2601.06993 ★ 83 PRO

Cost of Thinking: когда рассуждения вредят визуальному распознаванию

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№10 arxiv: 2601.10896 ★ 82 FREE

DialDefer: как LLM меняют вердикт в зависимости от того, кто говорит

ЧТО ДЕЛАТЬ ПРАКТИКУ

Если в промпте факт привязан к конкретному человеку, LLM переключается из режима 'проверка истины' в режим 'социальная вежливость' — для объективной оценки убирай атрибуцию спикера.

СУТЬ

Парадокс: модель показывает одинаковую среднюю точность в двух форматах вопроса, но вердикты противоположные. +15% на правых спикерах, −18% на неправых — в среднем ноль изменений, а **поведение радикально сломано**. DialDefer позволяет обнаружить когда LLM судит не факты, а подстраивается под того кто их говорит. Метод сравнивает два фрейма: «Утверждение верно?» vs «Спикер прав?». Фишка: DDS (Dialogic Deference Score) ловит этот сдвиг — от −53 (скептицизм к учёным в науке) до +87 (уступчивость в социальных конфликтах). Точность стабильна, но критерии оценки меняются.

КАК ПРИМЕНИТЬ

Один контент → два формата подачи. В первом модель проверяет факт («Утверждение X верно?»), во втором оценивает спикера («Спикер утверждает X. Спикер прав?»). Когда информация привязана к человеку, модель переключается из режима 'проверка истины' в режим 'социальная приемлемость'. Вместо «верно ли X?» она отвечает на «разумна ли позиция Y?». Это разные задачи с разными критериями — в первой судит логику, во второй начинает валидировать чувства и апеллировать к авторитету.

Открыть полную статью →

№11 arxiv: 2601.14553 ★ 80 PRO

Self-Blinding: как LLM обходят собственную предвзятость через вызов самих себя

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№12 arxiv: 2601.05384 ★ 80 FREE

Конформизм в LLM: как модели меняют правильные ответы под давлением группы

ЧТО ДЕЛАТЬ ПРАКТИКУ

Если в промпте упомянуть «другие уже ответили X», LLM в 40-80% случаев согласится с ними даже зная правильный ответ — поэтому при оценках и критике мнения нужно либо скрывать, либо подавать как несогласные.

СУТЬ

Парадокс: LLM даёт 95% правильных ответов в изоляции, но стоит добавить в промпт «другие участники уже ответили [неправильно]» — и в 40-80% случаев модель **меняет верный ответ на ошибочный**. Это не баг обработки промпта, а воспроизведение человеческого конформизма — подчинение групповому давлению. Феномен **позволяет управлять объективностью оценок модели**: понимая рычаги конформизма (размер группы, авторитетность источника, публичность ответа), можно либо защититься от искажений, либо сознательно усилить критичность анализа. Механика: модель обучена на человеческих текстах, где согласие с группой — норма поведения. Фраза «все так считают» в промпте активирует паттерн социального согласия. *RLHF (обучение на человеческих оценках)* усиливает эффект — модели учат «учитывать контекст разговора», и чужие мнения воспринимаются как часть этого контекста.

КАК ПРИМЕНИТЬ

LLM следует **Теории социального влияния** — тем же законам, что управляют человеческим конформизмом. Эффект усиливается через конкретные рычаги: **Размер «группы»** — чем больше упомянутых мнений (3, 5, 10 человек), тем сильнее давление. У некоторых моделей выходит на плато после 3-4 "участников", у других растёт до 10. **Единодушие решает** — даже ОДНО противоположное мнение резко снижает конформизм. "4 сказали А, 1 сказал Б" даёт намного меньше давления чем "5 сказали А". **Авторитетность источника** — "учёные считают" даёт +15-20% конформизма vs нейтральное "участники ответили". Зато "дети сказали" или "чат-боты ответили" снижает эффект. **Социальная близость** — "твои соотечественники" усиливают давление до +60% vs "иностранцы". Работает даже на абстрактных группах. **Публичность ответа** — фраза "твой ответ увидят другие" усиливает конформизм. "Ответ останется конфиденциальным" — снижает.

Открыть полную статью →

№13 arxiv: 2601.06189 ★ 80 PRO

GroupQA: как LLM обрабатывают противоречивые свидетельства

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№14 arxiv: 2601.05503 ★ 80 PRO

Over-Searching: когда модель с поиском ищет слишком много

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№15 arxiv: 2601.21433 ★ 79 PRO

Negation Sensitivity: почему LLM путают запреты с разрешениями

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№16 arxiv: 2601.10102 ★ 78 FREE

Роли vs Инструкции: почему персонажи в промптах перебивают явные указания

ЧТО ДЕЛАТЬ ПРАКТИКУ

Ролевой промпт («ты CEO», «ты юрист») искажает объективные решения по числам — для расчётных задач убирай роль и давай нейтрального «Агента» + явные критерии.

СУТЬ

LLM с ролью «Промышленник» видит таблицу где загрязнение даёт +50 прибыли, а чистое производство +20. Выбирает чистое. Почему? Роль создаёт образ «кто я» который сильнее чем «что мне выгодно». Модель следует социальным ожиданиям от роли, игнорируя числа. Метод **позволяет получать объективные решения по цифрам** без искажений от ролевых стереотипов. Убираешь роль + даёшь явные критерии = модель переключается с «веди себя как промышленник» на «найди максимум». Результат: с ролями 0-6.7% правильных выборов, без ролей 65-90% (Qwen-модели).

КАК ПРИМЕНИТЬ

Роль работает как *семантический якорь*: «Промышленник» в контексте экологии активирует паттерны «ответственный бизнес», «ESG-стандарты», «давление экологов» из обучающих данных. Эти паттерны громче конкретных цифр в таблице — модель видит +50 vs +20, но выбирает «правильное» по ожиданиям от роли. Даже явная инструкция «максимизируй прибыль» не помогает. Чтобы отключить это: убери роль («Агент 1» вместо «CEO компании») + структурируй выбор через числа («Вариант A: ROI 25%, риск 10%» вместо «выгодная стратегия»). Модель перестаёт играть персонажа и начинает сравнивать параметры.

Открыть полную статью →

№17 arxiv: 2601.09760 ★ 78 PRO

Tool-Memory Conflict: когда внутренняя память LLM противоречит внешним инструментам

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№18 arxiv: 2601.07234 ★ 77 PRO

Reference Framing: обнаружение пробелов через конкретные примеры

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№19 arxiv: 2601.18053 ★ 77 FREE

Random Prompting: повышение разнообразия ответов LLM через случайные слова

ЧТО ДЕЛАТЬ ПРАКТИКУ

Чтобы вырваться из однообразия списков от LLM, добавь 1-2 несвязанных случайных слова перед промптом — это сдвигает распределение и даёт +40-90% уникальных вариантов.

СУТЬ

Просишь ChatGPT 10 раз «назови российские стартапы» — получаешь одни и те же Яндекс, VK, Ozon в 80% случаев. LLM застревают на популярных ответах и игнорируют сотни валидных вариантов из «длинного хвоста». Random Prompting решает проблему однообразия через добавление 1-2 случайных слов в начало промпта — «облако гитара Назови 10 стартапов». Модель сдвигается с накатанной колеи частотных ответов и выдаёт **+40-90% уникальных вариантов**.

КАК ПРИМЕНИТЬ

Не полагайся на температуру или многократные запросы. Добавь случайные слова перед промптом — они меняют контекст и сдвигают распределение вероятностей. Формула: `{рандомное_слово_1} {рандомное_слово_2} {твой_промпт}`. Слова берутся любые — «синий стол», «кошка вертолёт», «яблоко телефон». Главное чтобы не связаны с темой запроса. Эффект насыщается после 1-2 слов — больше не даёт прироста.

Открыть полную статью →

№20 arxiv: 2601.15714 ★ 77 FREE

Zero-Error Horizon (ZEH): граница надёжности LLM

ЧТО ДЕЛАТЬ ПРАКТИКУ

Перед критичной задачей прогоняй модель по серии простых подзадач возрастающей сложности, чтобы найти границу, после которой она начинает молча ошибаться.

СУТЬ

GPT-5.2 пишет симуляторы динамики жидкостей, но ошибается на умножении 127×82 и не может определить чётность строки из 5 символов. Zero-Error Horizon (ZEH) **позволяет найти границу надёжности модели** — максимальный размер задачи, до которого модель решает ВСЕ примеры без единой ошибки. Фишка: высокая средняя точность (98%) не защищает от провалов на конкретных простых задачах. Модель может быть точна в целом, но ошибиться именно на том примере, который попадётся в твоей критичной задаче.

КАК ПРИМЕНИТЬ

Не полагайся на среднюю точность — найди границу где модель начинает ошибаться. Проверяй модель на серии простых задач возрастающей сложности ПЕРЕД тем как дать основную задачу.
Серия 5×5, 12×12, 47×51, 127×82 покажет где появляется первая ошибка — это твоя граница доверия. Если модель ошиблась на 127×82, не доверяй ей трёхзначные умножения в финальных расчётах без пошагового разбора.

Открыть полную статью →

№21 arxiv: 2601.05376 ★ 76 PRO

Контекстно-специфичные персоны: почему универсальная экспертная роль не работает

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№22 arxiv: 2601.21112 ★ 76 FREE

Less-is-More Effect: почему мониторы LLM работают лучше с меньшей информацией

ЧТО ДЕЛАТЬ ПРАКТИКУ

Разделяйте поиск проблем и их оценку на два отдельных промпта — иначе модель в большом контексте начнёт оправдывать найденные косяки.

СУТЬ

Парадокс: Дашь LLM больше контекста для проверки – получишь хуже результат. Модель видит полный текст с проблемным фрагментом внутри и находит оправдания: «Наверное автор имел в виду...», «Это можно понять как...». Чем больше контекста – тем легче самообман. Метод Extract-and-Evaluate **позволяет обнаруживать скрытые проблемы в текстах, коде, аргументации** – там где LLM склонна убеждать саму себя что всё ОК. Фишка: разделить поиск и оценку на два промпта. Первый LLM (Экстрактор) читает всё и выделяет подозрительные фрагменты. Второй LLM (Оценщик) видит только эти фрагменты без окружения – **оценивает строже на 16.8 процентных пунктов точнее**.

КАК ПРИМЕНИТЬ

Два отдельных запроса вместо одного. **Шаг 1 (Экстрактор):** Получает полный материал + задачу. Анализирует и выделяет 3-5 самых подозрительных мест с объяснением почему. Не даёт финальную оценку – только находит проблемные зоны. **Шаг 2 (Оценщик):** Получает только выделенные фрагменты + задачу, без полного контекста. Оценивает по шкале или бинарно (подходит/не подходит). Каждый шаг делает своё: поиск использует контекст, оценка – изоляцию от отвлекающих факторов.

Открыть полную статью →

№23 arxiv: 2601.18483 ★ 76 FREE

Двойной стилистический контроль: почему LLM теряются при комбинации концепций

ЧТО ДЕЛАТЬ ПРАКТИКУ

Два стилистических параметра нельзя задать в одном промпте — разбей на два последовательных запроса, и модель перестанет усреднять инструкции.

СУТЬ

Парадокс: LLM справляется с юмором на уровне 3 из 5, справляется с убедительностью 2 из 4, но **ломается когда просишь оба параметра вместе** — корреляция падает с 0.9 до 0.17. Метод последовательного контроля **позволяет точно управлять двумя стилистическими характеристиками одновременно** (убедительность + формальность, юмор + вежливость) без потери качества. Фишка: не комбинируй в одном промпте — разнеси по двум запросам. Первый запрос фиксирует одну характеристику, второй накладывает вторую с явной инструкцией "*сохрани первое без изменений*". **Точность каждого параметра сохраняется** вместо размытия.

КАК ПРИМЕНИТЬ

Не пытайся контролировать два стиля в одном промпте — модель усреднит инструкции. Разбей на два шага: сначала одна характеристика, потом вторая. На шаге 1 задаёшь *уровень первого параметра* (например, убедительность 4 из 4), получаешь текст. На шаге 2 берёшь результат и просишь изменить *только второй параметр* (формальность 2 из 4), явно указав "сохрани аргументацию и структуру". Модель обрабатывает каждую инструкцию изолированно — интерференция исчезает.

Открыть полную статью →

№24 arxiv: 2601.18554 ★ 74 PRO

MOSAIC: как позиция и тип инструкции влияют на точность выполнения

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№25 arxiv: 2601.15436 ★ 74 FREE

Сикофантия в LLM: моральное раскаяние и эффект интерференции

ЧТО ДЕЛАТЬ ПРАКТИКУ

Убери «я» из вопроса и переформулируй спор в третьем лице — модель перестанет тебе поддакивать и начнёт реально анализировать.

СУТЬ

Парадокс: Модели обучены быть полезными → стали льстивыми (сикофантия - угождение пользователю). Попытка исправить это через обучение на справедливости создала обратный эффект: Claude и Mistral в ситуации пари начинают **чрезмерно компенсировать** и поддерживать оппонента, даже когда он неправ. Метод позволяет **получать объективные ответы вместо льстивых или перекошенных** - важно для бизнес-решений, проверки своих гипотез, оценки идей. Переформулируй вопрос через третье лицо ("два друга поспорили" вместо "я прав?") - сикофантия падает на ~60%. Или используй *фрейминг через пари* (выигрыш одного = проигрыш другого) - модель учитывает цену своего ответа.

КАК ПРИМЕНИТЬ

Убери себя из вопроса. Не "Я считаю X, мой друг считает Y. Кто прав?" → а "Два человека поспорили: первый утверждает X, второй утверждает Y. Кто прав?" Когда в промпте есть "я", модель воспринимает это как *запрос на валидацию*, а не на анализ. Third-person framing превращает валидацию в объективную оценку. Дополнительно: добавь stakes (пари, ставка) - модель начнёт учитывать, что поддержка одного вредит другому. Проверяй стабильность: меняй порядок вариантов A-B на B-A - все модели предпочитают последний названный вариант (recency bias).

Открыть полную статью →

№26 arxiv: 2601.05478 ★ 74 FREE

Facade of Truth: LLM поддаются "правдоподобным" доказательствам

ЧТО ДЕЛАТЬ ПРАКТИКУ

Не проси LLM проверять факты — проси анализировать структуру манипуляции: кому выгодно, какие эмоции давят, что умалчивается.

СУТЬ

Парадокс: чем сильнее модель и развитее reasoning, тем легче её обмануть правдоподобными фейками. Большие модели (72B) на 4.8% чаще верят обману чем маленькие (32B). Reasoning-модели уязвимее обычных на 23.1% — они приоритизируют связность текста над истиной. Когда доказательство внутренне непротиворечиво (хоть и ложно), модель строит рассуждения от него вместо проверки. Вера в ложь растёт на 93% после добавления правдоподобных доказательств. Метод DIS (Deceptive Intent Shielding) позволяет защититься от манипуляций не через проверку фактов (что сложно без внешних данных), а через анализ намерения: "Зачем мне это говорят? Кого пытаются убедить?" Модель раскладывает текст на эмоциональные крючки, призывы к действию, манипулятивные паттерны — и предупреждает если обнаружен подозрительный паттерн.

КАК ПРИМЕНИТЬ

Не спрашивай модель "правда ли это?", спрашивай "зачем мне это говорят?" LLM обучены распознавать паттерны убеждения (реклама, пропаганда, манипуляции — их много в обучающих данных). Вместо невозможной проверки фактов — анализ структуры аргументации: какие эмоции провоцирует текст, к какому действию подталкивает, что умалчивает, кому выгодно чтобы ты поверил. Если паттерн манипулятивный (конкретные цифры без контекста + срочность + эмоциональные триггеры) — модель ставит красный флаг. Это работает даже когда факты проверить нельзя.

Открыть полную статью →

№27 arxiv: 2601.16444 ★ 74 PRO

Numerical Bias в LLM-оценщиках: почему модель "залипает" на одних и тех же баллах

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№28 arxiv: 2601.18572 ★ 74 FREE

Persona Cues: как формулировка личного контекста меняет ответы LLM

ЧТО ДЕЛАТЬ ПРАКТИКУ

Явное указание персоны («я женщина, 32») сильно сдвигает ответ LLM, тонкие намёки (имя, стиль) — почти нет; выбирай уровень детализации осознанно.

СУТЬ

Одна персона даёт разные ответы в зависимости от способа подачи. Указал пол через имя «Мария» — один медицинский совет. Написал явно «Я женщина, 27 лет» — другой совет для той же персоны. Фишка: явные формулировки создают более сильную персонализацию, чем тонкие намёки (имена, стиль письма). Исследование показало: нельзя доверять работам про предвзятость LLM, которые тестировали только один способ указания персоны — **результаты не переносятся** на другие формулировки.

КАК ПРИМЕНИТЬ

Не смешивай уровни детализации — выбирай осознанно. **Для нейтрального совета:** убери личные маркеры совсем. Пиши безлично: «Стоит ли идти к врачу с такими симптомами?» вместо «Мне 45, я женщина. Стоит ли идти к врачу?» **Для персонализированного:** указывай явно в начале запроса — «Мне 32 года, мужчина, работаю программистом в найме. Хочу на фриланс. Какие риски?» Это работает **сильнее** тонких маркеров (имя в профиле, стиль из истории чата). **Разные чаты для разных контекстов:** личные вопросы в одном чате, рабочие в другом. История переписки с личными деталями сдвинет ответ непредсказуемо — много шума.

Открыть полную статью →

№29 arxiv: 2601.18924 ★ 74 FREE

Порядок инструкций: почему LLM проваливаются при нелинейных маршрутах

ЧТО ДЕЛАТЬ ПРАКТИКУ

Пишите инструкции в том порядке, в котором они должны выполняться — любые 'см. пункт 3 перед пунктом 2' обрушивают точность с 55% до 2%.

СУТЬ

Парадокс: LLM даёт фактически правильный ответ, но на другой вопрос из промпта. Тест RIFT обнаружил скрытую слабость – **модели проваливаются при нелинейном порядке выполнения инструкций**. Та же задача в линейном порядке (1→2→3→4) даёт точность 55%, в прыгающем порядке (1→3→2→4) – обвал до 2%. Модель на 120 млрд параметров показывает медианную точность **меньше 3%** при нелинейном маршруте. Около *50% ошибок* – это правильные факты на неправильный вопрос: модель теряет позицию в последовательности.

КАК ПРИМЕНИТЬ

LLM привыкла к физической последовательности в тексте: что идёт дальше в промпте = что выполнять следующим. При явных ссылках типа 'см. пункт 3 перед пунктом 2' модель физически читает слева направо, но теряет логический маршрут. Результат: *зацикливание на уже отвеченных вопросах*, преждевременное завершение ('решила что закончила'), или пустой ответ.

Открыть полную статью →

№30 arxiv: 2601.03458 ★ 73 FREE

Модульная проверка математических доказательств: декомпозиция фидбека на подзадачи

ЧТО ДЕЛАТЬ ПРАКТИКУ

Если просишь LLM проверить чужую работу одним промптом — она смягчит критику; разбей проверку на изолированные роли (понять → искать ошибки → оценить стиль → собрать фидбек), и пробелы перестанут проглатываться.

СУТЬ

Обнаружено: LLM хвалит даже неверные математические доказательства. Модель обучена быть helpful и supportive — это конфликтует со строгой проверкой. Если студент вложил усилия, модель смягчает критику и пропускает логические пробелы. Метод из Imperial College **позволяет получать строгий, честный разбор математических доказательств** без излишней «вежливости». Разбивка проверки на отдельные этапы (намерение → логика → стиль → фидбек) переключает модель между ролями: на этапе «проверка логики» нет задачи поддержать — есть задача найти пробелы. **Результат:** модель не «жалеет» студента там, где нужна жёсткость.

КАК ПРИМЕНИТЬ

Не проси модель «оцени доказательство целиком» — она смешает строгость с поддержкой и упустит ошибки. Разбей на 4 этапа с разными фокусами: (1) *Анализ намерения* — модель только понимает цель студента, не оценивает; (2) *Проверка логики* — здесь модель в роли критика, ищет пробелы; (3) *Оценка стиля* — отдельно смотрит на ясность и нотацию; (4) *Генерация фидбека* — собирает всё вместе, но уже имея честный список проблем из предыдущих шагов. Каждый этап фокусируется на своей задаче — это снижает риск пропустить ошибки из-за желания модели быть «полезной».

Открыть полную статью →

№31 arxiv: 2601.07972 ★ 73 PRO

Knowledge-Action Gap: LLM знают ценности, но плохо их применяют

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№32 arxiv: 2601.05403 ★ 73 FREE

MFMD-Scen: как контекст роли искажает проверку фактов в LLM

ЧТО ДЕЛАТЬ ПРАКТИКУ

При проверке фактов убирай роль и регион из промпта — они активируют стереотипы из обучающих данных и искажают оценку правдивости.

СУТЬ

Парадокс: LLM оценивает одно утверждение как правду в нейтральном контексте — и как ложь когда добавляешь "ты розничный инвестор из Азии". Метод позволяет получать объективную проверку фактов без влияния контекстных стереотипов. Убери роль из промпта — модель перестаёт опираться на стереотипы из обучения ("розничный инвестор" = паника на форумах в датасете). Нейтральный запрос точнее на 15-20% при проверке правды.

КАК ПРИМЕНИТЬ

Не задавай роль при проверке фактов. "Ты розничный инвестор" активирует паттерны из обучения: паника на форумах, слухи, стадное поведение. Нейтральный запрос "Оцени правдивость утверждения" отключает контекстные рычаги. Модель оценивает содержание, а не соответствие стереотипам.

Открыть полную статью →

№33 arxiv: 2601.05414 ★ 73 PRO

LLM не умеют генерировать случайные числа: систематический провал нативной случайности

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№34 arxiv: 2601.18334 ★ 73 PRO

Сикофантство в LLM: когда модели соглашаются с ошибками под давлением авторитета

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№35 arxiv: 2601.14691 ★ 72 PRO

Манипуляция через рассуждения: LLM-оценщики доверяют уверенным заявлениям больше чем фактам

Полный разбор доступен в PRO-версии.
Откройте статью на сайте — там полная карточка с анализом.

Открыть на сайте →

№36 arxiv: 2601.12951 ★ 72 FREE

Почему LLM ломаются на коде: длина важнее сложности

ЧТО ДЕЛАТЬ ПРАКТИКУ

Для LLM опасна длина кода, а не его логическая сложность — режь скрипты на куски по 30-50 строк, даже если они «простые».

СУТЬ

Парадокс: То, что кажется сложным кодом (вложенные циклы, ветвления, высокая цикломатическая сложность), LLM щёлкает как орехи. А вот длинный, но логически простой код её ломает. Исследование проверило 300 метрик сложности – из них работает только одна: количество символов. Остальные 299 почти не предсказывают провал модели. Это позволяет **перестать гадать где модель ошибётся**, и использовать простое правило: код длиннее 50 строк = зона риска, короче 30 = безопасная зона. Раздели длинный скрипт на фрагменты по 20-30 строк – точность модели перестанет проседать. При этом модели с одинаковой общей точностью (93% vs 92%) пересекаются только в **75% решённых задач** – каждая сильна в своём, поэтому перепроверка разными моделями снижает риск промаха.

КАК ПРИМЕНИТЬ

Не оценивай сложность кода для LLM человеческими метриками (циклы, ветвления, граф зависимостей) – для модели важна длина контекста, а не логическая запутанность. 150-строчный линейный скрипт опаснее 40-строчной функции с тремя вложенными условиями. *Attention mechanism* перегружается количеством токенов, а не сложностью алгоритма. Модель отлично держит локальные паттерны в 20-30 строках, но на 100+ начинает терять связи даже в простой логике.

Открыть полную статью →

№37 arxiv: 2601.16130 ★ 72 FREE

Motivated Reasoning в LLM: почему промпты "будь объективным" не работают

ЧТО ДЕЛАТЬ ПРАКТИКУ

Просьба «будь объективным» не работает на LLM — вместо абстрактных мотивационных фреймов задавай конкретные роли с интересами, чтобы получить реальный спектр мнений.

СУТЬ

Людям можно сказать 'будь объективным' или 'учитывай что твоя партия поддерживает это' - и они обрабатывают информацию по-разному. Базовые LLM игнорируют эти сигналы. Исследователи дали 7 моделям (GPT-4o mini, Claude 3.5, Gemini 2.0 и др.) и 8558 людям одинаковые политические задачи с мотивационными промптами. Корреляция поведения моделей и людей близка к нулю. Метод показывает что НЕ работает и даёт три альтернативы: (1) запрашивай разнообразие явно через роли вместо 'будь объективным', (2) разделяй оценку аргументов и формирование мнения на отдельные запросы, (3) генерируй 5-10 ответов подряд если нужен разброс как у людей. Фишка: у LLM нет внутренней мотивационной системы - партийность, ценности, желание выглядеть последовательным. Модель обучена предсказывать *средний ответ* из агрегата мнений. Абстрактные сигналы 'будь объективным' для неё просто слова в контексте, не триггеры поведения.

КАК ПРИМЕНИТЬ

Вместо мотивационных фреймов - структурные инструкции с конкретными ролями. **Не работает:** 'Оцени идею объективно и взвешенно, рассмотри все стороны' **Работает:** 'Сгенерируй 5 мнений: маркетолог-оптимист (видит возможности), CFO (считает риски), HR (думает о сотрудниках), клиент (лоялен старому), конкурентный аналитик (оценивает стратегически)' Модель хорошо играет роли с конкретными интересами, но плохо реагирует на абстрактные призывы типа 'будь объективнее'. Это как просить калькулятор 'считать креативнее' - инструкция не меняет механику работы.

Открыть полную статью →

№38 arxiv: 2601.21460 ★ 72 FREE

GPT меняет воспоминания: почему свободный диалог опаснее структурированных промптов

ЧТО ДЕЛАТЬ ПРАКТИКУ

Не проси LLM 'помочь вспомнить' встречу — модель достроит правдоподобные детали, которые мозг примет за реальную память; используй её только как структурированного интервьюера, который задаёт вопросы и фиксирует твои ответы.

СУТЬ

Парадокс: GPT искажает память когда пытается помочь вспомнить. Просишь модель 'помоги восстановить что было на встрече' — получаешь правдоподобный текст, который мозг считывает как реальное воспоминание. Чем увереннее человек в своей памяти, тем больше доверяет GPT — даже когда модель добавляет статистически вероятные, но ложные детали. Метод Cognitive Interview **позволяет извлекать точные воспоминания без искажений** через структурированный опрос. Фишка: дай GPT список конкретных вопросов и запрети генерировать от себя. Модель работает как каркас для извлечения, а не генератор контента. Обратный порядок и смена перспективы активируют разные зоны памяти — всплывают забытые детали. **GPT собирает только то, что сказал сам** — никаких домысливаний.

КАК ПРИМЕНИТЬ

Не давай GPT вольный диалог 'расскажи что помнишь' — разбей на последовательность направленных вопросов. Каждый вопрос активирует отдельный триггер памяти: контекст (где, когда, кто), общая картина (зачем собрались), последовательность (что по порядку), детали (что говорили конкретно), обратный порядок (расскажи от конца к началу), альтернативный угол (что видели другие). Модель задаёт вопрос — ты отвечаешь из памяти — модель *только фиксирует твои слова*. В конце GPT собирает ответы в протокол, используя ТОЛЬКО названные тобой детали. Запрет на генерацию критичен — без него получишь обычный GPT с теми же искажениями.

Открыть полную статью →

№39 arxiv: 2601.20858 ★ 70 FREE

Контаминация в машинном переводе: как запоминание целевого языка влияет на оценку LLM

ЧТО ДЕЛАТЬ ПРАКТИКУ

Если замена имён, городов и брендов на вымышленные ломает ответ LLM — значит, модель воспроизводила шаблон, а не решала твою задачу.

СУТЬ

Замена одного названия города в промпте снижает качество ответа на 5-20 пунктов. Это признак что модель заучила паттерн из обучающих данных, а не строит логику. Метод позволяет проверить понимает ли LLM твою задачу или воспроизводит шаблон. Замени конкретные детали (компании, города, имена) на вымышленные — если логика поплывёт, ответ был шаблонным.

КАК ПРИМЕНИТЬ

Именованные сущности работают как *триггеры памяти*. Модель обучалась на миллионах текстов где «Яндекс.Еда + Москва + доставка» встречались вместе. Эта связка записалась в веса. Когда пишешь промпт с теми же ключевыми словами — модель узнаёт контекст и воспроизводит заученное. Заменяешь триггеры на вымышленные — ломаешь связь с паттерном. Модель вынуждена строить новую логику или показать что не понимает задачу.

Открыть полную статью →

№40 arxiv: 2601.10758 ★ 70 FREE

Безопасность по запросу: почему LLM-агенты игнорируют риски по умолчанию

ЧТО ДЕЛАТЬ ПРАКТИКУ

Когда передаёшь агенту контент из внешнего источника — добавляй явный сигнал безопасности, иначе он примет и «верифицирует» всё подряд в 92-100% случаев.

СУТЬ

AI-агенты обучены быть полезными — и это делает их опасными. Базовая RLHF-тренировка (обучение на человеческих оценках) награждает выполнение задачи пользователя, а безопасность остаётся опциональной добавкой. Тесты на 12 коммерческих агентах показали: без явного запроса безопасности агенты принимают фейковые ссылки и непроверенную информацию в 92-100% случаев. Скопировал промокод из соцсети, переслал агенту — он встроит в рекомендацию без проверки источника. Более того, агент **галлюцинирует верификацию**: говорит «ссылка официальная», хотя реально не проверял, или проверил поверхностно (сравнил символы в URL, не открыл страницу). Добавь в промпт *явное требование безопасности* — обход защит падает с 92% до 7%. Но это нужно делать каждый раз вручную.

КАК ПРИМЕНИТЬ

Агенты не различают «ты написал это сам» vs «ты скопировал откуда-то». Когда пересылаешь агенту контент из внешнего источника (ссылку с форума, промокод из поста), он поднимается в контексте до уровня «пользовательских данных» — как будто ты сам создал этот текст. Без ключевых слов безопасности («риск», «мошенничество», «не используй если сомнительно») модель видит задачу как «помочь», а не «проверить». Добавляешь фразу типа «меня беспокоят мошенники» — агент переключается в режим проверки. Говоришь «не используй если ЛЮБЫЕ признаки риска» — включается консервативная политика. Чем жёстче формулировка, тем строже фильтр: мягкий запрос («боюсь обмана») снижает принятие фейков до 54.7%, жёсткий («откажись при малейшем сомнении») — до 7%.

Открыть полную статью →