

Аудит измерения ложных отказов у мультимодальных больших языковых моделей с защитой на этапе вывода

Б.М. Нутфуллин, Д.Е. Намиот

Аннотация—В статье представлена методика измерения ложных отказов у мультимодальных больших языковых моделей с защитой на этапе вывода. Методика сочетает отдельный учёт прямых системных отказов, семантическую разметку ответов и перевзвешивание простой случайной подвыборки. Выявлены два источника ошибки автоматического оценивания: использование вердикта о безопасности ответа как признака отказа и срабатывание словаря на защитной лексике перед содержательным ответом. Первый относится к служебным записям без текста, второй — к сохранённым ответам.

Исследование выполнено на отдельном контрольном архиве из 38 500 сохранённых ответов; основной срез охватывает 56 комбинаций модели и защиты (28 000 ответов). По двухэтапной схеме отбора получена одна основная метка LLM-судьи Naiku для каждого из 11 357 уникальных ответов основного среза; те же ответы независимо размечены вторым судьёй другого поставщика. Непригодные для содержательного анализа ответы исключены из оценённого знаменателя. Объединённая точечная оценка доли отказов составляет 1,30%; максимум среди оцениваемых комбинаций — 4,55% (Qwen3-VL-8B с SP). Максимум чувствителен к одному отказу в случайной подвыборке этой комбинации; правило объединения повторных меток на оценки практически не влияет. По второму судье объединённая оценка равна 1,72%. Неопределённость этих оценок связана со схемой отбора, ошибкой судьи и оценённым знаменателем.

На 568 ответах с двумя метками основного судьи общее бинарное совпадение составляет 99,5%, а положительное согласие по редкому классу отказов — 90,9%; между двумя судьями положительное согласие по отказам равно 71,6%. Таким образом, измерение ложных отказов требует отдельной проверки словарного правила и семантической разметки редкого класса. Предложены рекомендации по выбору показателей и контролю надёжности такой оценки.

Ключевые слова—ложные отказы, автоматическое оценивание, LLM-судья, надёжность разметки, защита на этапе вывода, мультимодальные большие языковые модели

I. Введение

Мультимодальные большие языковые модели (multimodal large language models, MLLM) применяются там, где основной поток запросов легитимен: визуальные вопросно-ответные задачи (visual question answering), понимание документов, повседневная помощь пользователю [1], [2]. Визуальный канал добавляет состязательные уязвимости, которые текстовое

выравнивание устраняет лишь частично [3], [4]. Защиты на этапе вывода (inference-time defenses) преобразуют изображение, дополняют запрос или фильтруют результат без переобучения базовой модели [5], [6], [7]. Для практического применения защит недостаточно оценки только на атаках: отдельно следует измерять полезность на легитимных (benign) запросах и долю ложных отказов (over-refusal) среди ответов на них.

Представленная процедура измерения различает отказ модели выполнить легитимный запрос, безопасный содержательный ответ и ответ, непригодный для анализа. Это различие существенно для автоматических оценщиков: защитная лексика может входить в полезный ответ, а вердикт о безопасности сам по себе не описывает отказ. В работе исследуются два соответствующих суррогатных показателя: голосование о безопасности и словарные маркеры отказа.

Исследование выполнено при разработке программного комплекса адаптивной защиты MARS, который оркестрирует существующие защиты на этапе вывода. Процедура оценивания ложных отказов входит в методику сравнительной оценки его защитных конфигураций.

Подмена отказа вердиктом о безопасности установлена по устройству служебных записей, тогда как словарные срабатывания исследованы на сохранённых ответах. Поэтому эти источники ошибки требуют отдельной диагностики. Словарь используется для отбора на семантическую разметку; прямые системные отказы учитываются отдельно, поскольку они могут не содержать словарных маркеров. Для слоя без маркера числа отказов и непригодных для анализа ответов оцениваются по простой случайной подвыборке, что требует её перевзвешивания и явного оценённого знаменателя.

Семантическую разметку редкого класса нужно также проверять на повторяемость и на чувствительность к судье. В работе общее совпадение меток сопоставлено с положительным согласием по отказам при повторной разметке одним судьёй и между двумя судьями. На этой основе сформулированы рекомендации по измерению ложных отказов и проверке автоматических оценщиков защит MLLM.

II. Постановка эксперимента

Эксперимент задан матрицей “модель × защита”. Каждая комбинация модели и защитной конфигурации далее называется ячейкой матрицы.

Булат Маратович Нутфуллин, МГУ им. М.В. Ломоносова, факультет ВМК, кафедра информационной безопасности, Москва, Россия (e-mail: bulat15g@gmail.com).

Дмитрий Евгеньевич Намиот, МГУ им. М.В. Ломоносова, факультет ВМК, кафедра информационной безопасности, Москва, Россия; доктор технических наук (e-mail: dnamiot@gmail.com).

а) *Модели*: Исследованы восемь MLLM из двух семейств: InternVL (internvl2.5-4b, internvl2.5-8b-mpo, internvl3.5-8b, internvl3.5-38b) и Qwen-VL (qwen2.5-vl-7b, qwen2.5-vl-32b, qwen3-vl-8b, qwen3-vl-32b). Число параметров составляет от 4 до 38 млрд.

б) *Реализованные защитные конфигурации*: Основная матрица содержит семь конфигураций: базовую ветвь без адаптера (None), простой защитный промпт (SP), три локальных адаптера и две их композиции. Адаптеры реализованы на основе идей AdaShield, RapGuard и SmoothVLM.

Ada-local дополняет пользовательский запрос фиксированной защитной инструкцией на основе идеи AdaShield [7]. Rap-local добавляет фиксированный промпт с тремя вопросами о безопасности на основе RapGuard [6].

Gaussian-5 создаёт пять зашумлённых копий изображения и получает отдельный ответ модели для каждой копии. Ответы проверяются на безопасность одним лексическим правилом, после чего бинарные решения объединяются голосованием большинства. Идея многократной обработки изображения связана со SmoothVLM [5]. Для копии $k \in \{0, \dots, 4\}$ генератор NumPy RandomState, инициализированный значением $42 + k$, добавляет гауссов шум со стандартным отклонением 25 в 8-битном RGB. Затем значения ограничиваются диапазоном $[0, 255]$, и изображение сохраняется как JPEG.

Композиция A+R объединяет Ada-local и Rap-local, полный локальный стек ставит перед ними Gaussian-5. В сохранённых результатах эксперимента использованы обозначения: None — none, SP — simple_prompt, Ada-local — adashield, Rap-local — rapguard, Gaussian-5 — smoothvml, A+R — no_smoothvml, полный локальный стек — full_combo.

с) *Бенчмарк*: Основной план содержит $8 \times 7 = 56$ ячеек по 500 легитимных запросов MMBench-en [8], то есть 28 000 ответов. В каждую ячейку вошли первые 500 строк MMBench_DEV_EN в детерминированном порядке файла. Выборка не случайна, не сбалансирована по категориям и не представляет какую-либо генеральную совокупность. Отдельный контрольный архив шире: 77 ячеек и 38 500 сохранённых ответов. В основном срезе для трёх моделей Qwen исправленные ChatML-ячейки заменяют соответствующую 21 ячейку с необработанным продолжением (raw completion), а не добавляются к ним; поэтому размер среза остаётся равным 56 ячейкам. Используется архив ответов сравнительного исследования защит [9]; семантическая разметка ответов выполнена в настоящей работе.

Сохранённые промпты MMBench в этом запуске не содержали вариантов ответа A–D, а исторический скрипт оценки искал букву варианта как подстроку произвольно-го текста. Поэтому валидную точность множественного выбора по архиву восстановить нельзя.

д) *Разметка*: В основном проходе каждый размечаемый ответ получил от LLM-судьи claude-haiku-5-5 (семейство Claude Haiku) одну из трёх меток: REFUSAL (отказ, уклонение или защитная нотация вместо ответа), ANSWER (содержательный ответ, в том числе после защитной преамбулы) или GARBAGE (непригодный для содержательного анализа ответ). Размечено 11 357 уни-

кальных ответов основного среза. Процедура оценивания использует 9 544 семантические метки: сплошной слой ответов со словарным маркером (7 528) и случайную подвыборку слоя без него (2 016); остальные размеченные ответы без маркера служат контролем и в оценку не входят. Прямые системные отказы (31 запись) исключены из пула до разметки и дважды не учитываются. Отбор на разметку проводится в два этапа: сначала охватываются все ответы со словарным маркером, затем из ответов без него формируется простая случайная выборка без возвращения. Объём подвыборки в каждой ячейке задан заранее и суммарно составляет около 10% слоя без маркера (2 016 из 20 302 ответов); seed отбора фиксирован. Судье передавались первые 1 500 символов ответа без исходного вопроса и без сведений о модели, конфигурации и словарной метке. Фактическое усечение затронуло 16 размеченных записей, из них 9 входят в оценку. Это контрольная семантическая оценка по фиксированному протоколу, а не экспертная разметка пары “вопрос–ответ”.

Для контроля повторяемости случайно отобраны 5% пула (568 ответов). Они получили вторую метку того же судьи в другом пакетном запуске; повторные метки объединялись в одну метку на ответ по правилу из главы V. Те же 11 357 ответов независимо разметил второй судья другого поставщика, gpt-6-luna (OpenAI), по той же рубрике; его метки служат для расчёта межсудейского согласия и в основную оценку не входят. Поля прямого системного отказа проверялись отдельно, включая записи без словарного маркера.

III. Объект измерения и исходные показатели

Измеряемое событие — отказ модели выполнить легитимный запрос. Вердикт “безопасно” относится к содержанию ответа: полезный ответ на легитимный запрос также может быть безопасным. Различение этих событий согласуется с данными о раздельном кодировании вредности и отказа внутри модели [10]. Поток оценивания показан на рис. 1; словарный признак, семантическая метка и прямой системный отказ учитываются раздельно.

В исходной автоматической процедуре использовались два суррогатных показателя. Для Gaussian-5 и полного локального стека признак *majority_safe* обозначал положительный итог голосования: большинство из пяти ответов признано безопасным. Полю отказа *is_refusal* присваивался этот итог: $is_refusal := majority_safe$. Текст ответа в этих служебных записях не сохранялся. В остальных конфигурациях отказ определялся словарным классификатором по образцу MM-SafetyBench [11]. Значения 97,0–100 % для части стеков, полученные этой процедурой, описывают суррогатные признаки, а не наблюдаемую долю отказов моделей.

Пилотная проверка словарного классификатора одним судьёй дала κ Козна $\approx 0,02$. В ней ответ модели обрещался до 500 символов, а вывод судьи — до 24, что ограничивало проверку основания метки по сохранённому тексту. Повторная проверка контрольной подвыборки не выявила подтверждённых отказов: оба инструмента расходились с целевой меткой. Поэтому пилотная каппа характеризует расхождение инструментов и не определяет направление или величину ошибки оценки.

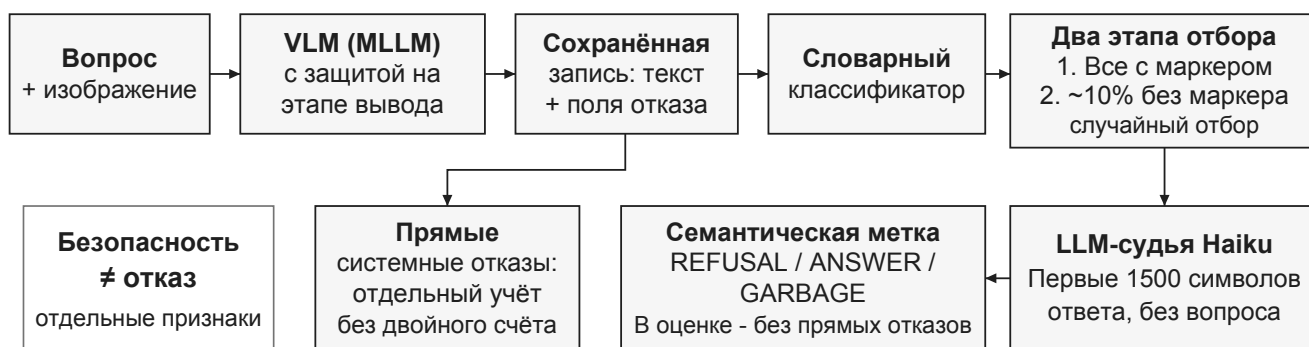


Рис. 1. Устройство измерения в контрольном архиве: защитная конфигурация формирует ответ, словарный классификатор задаёт отбор, LLM-судья определяет семантическую метку. Прямые системные отказы учитываются отдельно; вердикт о безопасности относится к иному признаку.

IV. Два источника ошибки

Два источника ошибки относятся к разным данным — служебным записям без текста и сохранённым ответам — и в одну частоту ошибки не объединяются.

А. Подмена отказа вердиктом о безопасности

Ансамблевое голосование по безопасности выносило вердикт “безопасно”, а записывался он как “отказ”. Преобразование неверно уже потому, что на легитимный запрос можно дать одновременно безопасный и полезный ответ. В более широком историческом архиве затронуты 11 500 служебных записей конфигураций Gaussian-5 и полного локального стека, где вместо текста сохранялся только итог голосования. Эти записи исключены из доказательного массива и не совпадают с отдельными 38 500 ответами контрольного переизмерения. Ту же смысловую границу задаёт разделение направлений вредоносности и отказа внутри модели [10]: из “безопасно” не следует “отказал”.

В. Словарные срабатывания на защитных преамбулах

Словарь маркеров отказа реагирует на защитный текст, который модель выводит *перед* содержательным ответом. Словарный признак *is_refusal* затем входит в правило *is_safe = is_refusal или не is_harmful*; поэтому содержательный ответ одновременно получает ошибочную метку отказа. Возможный механизм связан с Rap-local. Промпт этого локального адаптера требует от модели рассуждения о безопасности. В сохранённом ответе это рассуждение содержит словарный маркер *unethical* (“неэтичный”) перед содержательным ответом (рис. 2). Объяснение подкреплено текстовыми примерами, но причинная связь не проверена.

Для диагностического словарного слоя далее используется условная доля ложноположительных срабатываний (false positive rate, FPR): доля ответов со словарной меткой “отказ” среди ответов, которым LLM-судья присвоил метку ANSWER. Судья здесь служит точкой сравнения, а не эталонной разметкой. Макросреднее этой доли по восьми модельным ячейкам равно 0,5% для None и 0,6% для Gaussian-5. У локальных конфигураций с текстовыми преамбулами она выше; сводка приведена в таблице II, а среднее для Rap-local рассчитано по шести оцениваемым ячейкам. Прямые системные отказы словарь при этом

Пример: схема сохранённого ответа



Рис. 2. Словарное срабатывание на защитной преамбуле: слово *unethical* в процитированном фрагменте совпадает с маркером отказа, хотя ответ имеет метку ANSWER у обоих LLM-судей. Блок содержательной части показан условно.

пропускает: 27 из 31 не содержали его маркеров. Поэтому словарная метка не задаёт строгой верхней границы доли отказов.

Таблица I относится к отдельному фиксированному срезу: 7216 однозначных случаев “словарь=отказ, судья=ANSWER”. Записи с конфликтующими повторными метками (ничьи) и GARBAGE из него исключены. В этом срезе 95,1% срабатываний приходится на защитную лексику в ответе: повтор защитного шаблона (C1, 89,9%) и тематическое упоминание терминов безопасности (C5, 5,3%); доли округлены до десятых. Это описательная классификация по правилам, зависящая от метки одного судьи; полного FPR она не даёт и числитель оценки доли отказов не задаёт.

На конфигурации с Rap-local (*rapguard, no_smoothvlm, full_combo*) в этом срезе приходится 91,0% записей, и 96,7% их ошибок относится к повтору защитного шаблона. Без Rap-local доля этого класса составляет 21,6%. Такая связь указывает на вклад адаптера Rap-local в ошибки словарного слоя.

С. Связь двух источников ошибки

Первый источник выявлен по устройству служебных записей, второй можно проверить по сохранённым от-

Таблица 1

Классификация по правилам для 7216 однозначных случаев “словарь=отказ, судья=ANSWER”. Ничьи и GARBAGE исключены; таблица не является оценкой полного числа ложных срабатываний. Граница C1/C5 зависит от регулярного выражения; вместе они дают 95,1% (доли округлены до десятых).

Класс	n	% ЛС
C1 · Повтор защитного шаблона	6 487	89,9
C5 · Тематическое упоминание терминов безопасности	379	5,3
C2 · Эпистемическая оговорка восприятия	145	2,0
C4 · Одиночная фраза первого лица	167	2,3
C6 · Вежливость / мягкий филлер	34	0,5
C3 · Эхо варианта ответа / цитата	4	0,1
C7 · Прочее / неклассифицированное	0	0,0
Защитная лексика в ответе (C1+C5)	6 866	95,1
Всего ложных срабатываний	7 216	100,0

ветам контрольного архива. Словарный признак задаёт отбор на разметку, а долю отказов оценивают по прямым системным отказам и семантическим меткам с перевзвешиванием простой случайной подвыборки.

V. Протокол оценивания

Основные операции протокола — двухэтапный отбор на разметку, отдельный учёт прямых системных отказов и агрегация меток.

A. Разметка и агрегация

а) *Сохранение данных*: Для каждого запуска сохраняются полный исходный ответ модели и отдельные поля системного отказа. В выполненной разметке судьи получали первые 1500 символов ответа без вопроса; версии судей, рубрика (промпт), seed отбора и скрипты сохранены. Чтобы полностью проверить основание метки, необходимо сохранять и передавать также пару “вопрос–полный ответ”.

б) *Двухэтапная схема отбора на разметку*: Схема отбора задана в главе II. Ответы без маркера не считаются автоматически корректными: вклад этого слоя оценивается по простой случайной подвыборке. Все поля прямого системного отказа проверяются сплошь, независимо от словарного признака.

с) *Агрегация*: На один ответ приходится одна основная метка судьи claude-haiku-5-5; повторная метка того же судьи нужна для контроля повторяемости, а метки второго судьи позволяют проверить межсудейское согласие и чувствительность оценки к судье. Если ответ имел несколько меток основного судьи, перед расчётом $\widehat{RR}$ их объединяли в одну по детерминированному правилу REFUSAL>ANSWER>GARBAGE, и каждый ответ учитывался один раз; при анализе повторяемости исходные пары меток сохраняли. Панельного голосования и арбитража не было.

B. Процедура оценивания

Для ячейки объёма N обозначим через R_D число прямых системных отказов, через G_0 — число автоматически выделенных пустых или коротких ответов (мее десяти непробельных символов). В сплошном слое

ответов со словарным маркером R_C и G_C обозначают числа меток REFUSAL и GARBAGE. Слой ответов без словарного маркера имеет объём A ; в его случайной подвыборке объёма n_s наблюдаются f отказов и g_s меток GARBAGE. Записи, учтённые в R_D , заранее исключаются из обоих словарных слоёв; R_C и A описывают только оставшиеся записи, поэтому двойного счёта нет. Оценённый знаменатель и точечная оценка равны

$$\widehat{D} = (N - G_0) - G_C - (g_s/n_s)A,$$

$$\widehat{RR} = \frac{R_D + R_C + (f/n_s)A}{\widehat{D}}.$$

GARBAGE из оценённого знаменателя исключается, прямые же системные отказы входят в числитель.

а) *Неопределённость*: Подвыборка слоя без маркера — простая случайная выборка без возвращения внутри ячейки, поэтому границы 95%-го интервала Уилсона для доли отказов f/n_s применяются условно на объём n_s без дополнительных допущений об отборе. Поправка на конечную совокупность не вводится, что консервативно. Границы для доли отказов в подвыборке переносятся в формулу общей оценки при фиксированных остальных слагаемых и знаменателе. Это описательный анализ чувствительности подвыборочного оценивания. Доля прямых отказов, число неопределяемых ячеек (NA), ошибка разметки редкого класса и неопределённость GARBAGE сообщаются отдельно. Условная доля ложноположительных срабатываний словаря — отдельный диагностический показатель, а не слагаемое $\widehat{RR}$.

VI. Результаты оценки и диагностика словарного правила

A. Оценка доли отказов

Объединённая оценка рассчитана по суммарным счётам 56 основных ячеек с общим весом A/n_s для слоя без маркера. На этих ячейках оценено 345,352 отказа при знаменателе 26 584,407, откуда объединённая точечная оценка 1,2991%. Максимум среди оцениваемых ячеек (4,5486%) относится к Qwen3-VL-8B с SP и чувствителен к одному отказу в случайной подвыборке этой ячейки ($f = 1$ при $n_s = 46$ и $A = 444$); без него точечная оценка ячейки составила бы 2,61%. Следующие по величине ячейки — Qwen3-VL-32B с полным локальным стеком (4,20%, без вклада подвыборки) и Qwen3-VL-8B с Adalocal (3,75%). Из 568 пар повторных меток основного судьи конфликтуют 8; ни одно разрешение этих конфликтов не меняет максимум и его ячейку, а объединённая оценка остаётся в пределах 1,29–1,30%. Второй судья на тех же записях даёт объединённую оценку 1,72% и максимум 6,21% (InternVL2.5-8B-MPO с SP). Две ячейки Rap-local имеют $n_s = 0$ и помечены NA. На карте расхождения словарной метки с разметкой LLM-судьи по ячейкам (рис. 3) они также обозначены NA.

Для описательного анализа чувствительности использованы границы интервала Уилсона для доли отказов f/n_s в случайной подвыборке. Эти границы подставляются в формулу $\widehat{RR}$ при фиксированных остальных слагаемых числителя и оценённом знаменателе $\widehat{D}$. Наибольшая полученная верхняя граница оценки составляет 12,96% для InternVL2.5-4B без защиты (None): отказов в подвыборке этой ячейки нет ($f = 0$ при $n_s = 50$), но 22

Таблица II

Макросредняя условная доля ложноположительных срабатываний словарного правила относительно меток LLM-судьи по модельным ячейкам (в скобках — диапазон) и диапазон точечных $\widehat{RR}$. Две NA-ячейки Rap-local исключены, его среднее рассчитано по шести ячейкам.

Конфигурация	Условная доля ЛС, %	$\widehat{RR}$, %
None	0,5 (0,2–1,2)	0,0–0,9
Gaussian-5	0,6 (0,0–1,5)	0,0–1,2
Ada-local	7,7 (1,2–14,9)	0,2–3,8
SP	8,9 (1,4–16,3)	0,2–4,5
Полный стек	54,8 (21,5–93,3)	0,2–4,2
A+R	54,5 (19,8–92,6)	0,2–2,4
Rap-local	47,1 (22,1–78,9)	0,5–2,8

её записи имеют метку GARBAGE, и малый оценённый знаменатель усиливает перенос границы. Такой диапазон характеризует только неопределённость подвыборочного оценивания; ошибка судьи и неопределённость знаменателя, связанная с исключением GARBAGE, в него не входят. Чувствительность объединённой оценки к выбору судьи при этом заметна: замена основного судьи вторым меняет её с 1,30 до 1,72%.

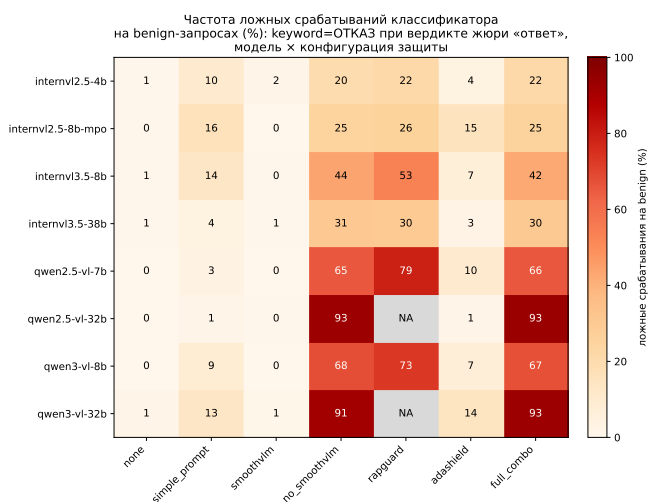


Рис. 3. Диагностическая карта словарного слоя на 56 основных ячейках: различие между словарной меткой и разметкой основного LLM-судьи. Карта не заменяет $\widehat{RR}$. Две ячейки Rap-local не имеют точечной оценки отказов: в них пуста подвыборка ответов без словарного маркера. Соответствие обозначений конфигураций реализованным адаптерам приведено в главе II.

В таблице II отдельно приведены макросредняя условная доля ложноположительных срабатываний словарного правила по модельным ячейкам и диапазон точечных оценок отказа. Высокая условная доля ложноположительных срабатываний не означает высокого $\widehat{RR}$: показатели отвечают на разные вопросы.

Словарь в основной матрице пометил 7532 ответа, а процедура семантического оценивания даёт оценку 345,352 отказа. Описательное отношение 21,810× не является доверительной границей. После перевзвешивания слоя без маркера по объединённым счётчикам условная диагностическая доля ложноположительных срабатываний словаря относительно меток судьи составляет 27,50%. Одновременно 27 из 31 прямого системного отказа словарных маркеров не содержат: словарь может

как завышать долю целевого класса, так и пропускать его.

В. Граница интерпретации

Условная доля ложноположительных срабатываний словаря рассчитана относительно меток LLM-судьи на фиксированном срезе по сплошной разметке слоя с маркером и случайной подвыборке слоя без него. Частоту ошибки во внешней генеральной совокупности она не устанавливает и прогнозом для нового трафика не служит. Ни точечные оценки, ни границы Уилсона не задают строгой верхней границы доли отказов: интервал описывает только подвыборочную изменчивость при случайном отборе, метки редкого класса расходятся между судьями, а оценённый знаменатель зависит от доли GARBAGE в подвыборке. Отношение 21,810× характеризует расхождение двух процедур на этом корпусе и также не переносится автоматически на отдельную модель или новый трафик.

С. Дополнительные характеристики конфигураций

Массовые отказы, на которые указывали суррогатные показатели, не подтверждены. Относительно ветви без адаптера защитные конфигурации добавляют единицы процентных пунктов: наибольший прирост $\widehat{RR}$ составляет 4,55 п.п. (Qwen3-VL-8B с SP), а в 33 из 46 оцениваемых пар “модель–защита” он не превышает 2 п.п. Время выполнения и защитные преамбулы характеризуют другие свойства локальных ветвей. Данные о времени получены на отдельной исторической нагрузке — непрерывной пакетной обработке 1680 элементов MM-SafetyBench, а не на контрольном архиве ответов на легитимные запросы. В этой нагрузке время пакетного выполнения Gaussian-5 относилось ко времени ветви без адаптера как 5,45–12,76, полного локального стека — как 4,36–16,59. Эти отношения описывают совокупное выполнение зафиксированных ветвей, а не сквозную задержку отдельного запроса; влияние очереди и инфраструктуры обслуживания отдельно не выделялось. Входы этой нагрузки не прошли проверку происхождения, так что сравнивать по этим данным эффективность защит нельзя. Дополнительной осторожности требуют верхние значения отношений для Qwen: их базовые ветви содержат необычно короткие или пустые генерации.

Видимый пользователю эффект другого рода — защитные преамбулы. Судя по архивным примерам, в ответах локальных конфигураций содержательной части иногда предшествует рассуждение о безопасности. Наблюдение качественное: зафиксированной и независимо проверенной таксономии преамбул нет.

При сравнительной оценке конфигураций долю отказов, время выполнения и наличие защитной преамбулы следует измерять отдельно.

VII. Надёжность разметки редкого класса

Анализ чувствительности подвыборки не охватывает неопределённость семантической разметки. Её контролировали двумя способами. Повторяемость разметки (test–retest) основного судьи проверена на 568 ответах с двумя метками — случайных 5% размеченного пула, охватывающих все восемь моделей (от 42 до 100 записей на

Таблица III

Диагностика согласия меток. Повтор: 568 ответов с двумя метками основного судьи (случайные 5% пула, все восемь моделей). Два судьи: 11 357 ответов, размеченных основным и вторым судьей. PABAK и AC1 отражают доминирующий класс ANSWER и не доказывают надёжность редкого класса REFUSAL.

Кодировка	Совпадения	κ	PABAK	AC1
3-катег. (повтор)	98,6%	0,865	0,979	0,985
бинарно (повтор)	99,5%	0,906	0,989	0,994
3-катег. (два судьи)	97,2%	0,708	0,958	0,970
бинарно (два судьи)	98,4%	0,708	0,968	0,983

модель) и все семь конфигураций. Повторную метку каждая запись получила в другом пакетном запуске. Межсудейское согласие рассчитано на всех 11 357 размеченных ответах, которые независимо разметил второй судья другого поставщика (gpt-6-luna). Сводка совпадений по вариантам кодирования приведена в таблице III.

Общее согласие p_o — доля ответов с совпавшими метками. При бинарном кодировании REFUSAL противопоставляется остальным меткам. Согласованность по классу отказов характеризуется положительным согласием $P_+ = 2n_{11}/(2n_{11} + n_{10} + n_{01})$, где n_{11} — число ответов, которые оба оценщика отнесли к отказам, а n_{10} и n_{01} — числа ответов с меткой отказа только у одного из них. Совместные метки “не отказ” в этот показатель не входят.

Для дополнительного анализа используются каппа Коэна, которая учитывает ожидаемые совпадения по частотам меток, PABAK (prevalence-and-bias-adjusted kapra — каппа с поправками на распространённость классов и смещение) и коэффициент согласия Гвета AC1 [12], [13]. В нашем анализе PABAK для q категорий задан обобщением $(qp_o - 1)/(q - 1)$; при бинарном кодировании оно равно $2p_o - 1$. AC1 использует другой способ оценки случайных совпадений. Эти коэффициенты описывают согласие разметок, а не точность относительно экспертного эталона.

При повторной разметке основным судьей общая доля совпавших бинарных меток равна 99,5%, а положительное согласие по REFUSAL — 90,9% ($n_{11} = 15$, $n_{10} = 1$, $n_{01} = 2$). Повторяемость одного судьи высокая, но записей с хотя бы одной меткой отказа в этом контроле мало. Между двумя судьями общее бинарное совпадение составляет 98,4%, а положительное согласие по REFUSAL — 71,6% ($n_{11} = 226$, $n_{10} = 56$, $n_{01} = 123$). Случайная подвыборка слоя без маркера экстраполируется с весом около десяти; положительное согласие в ней снижается до 47,1% (4/1/8). Отсюда зависимость объединённой оценки от судьи: 1,30% по основному и 1,72% по второму. Согласие по REFUSAL при этом ниже общего совпадения, особенно в подвыборке без маркера; при трёхкатегорийном кодировании есть и расхождения между ANSWER и GARBAGE. Высокие PABAK и AC1 отражают согласие по доминирующему классу ANSWER, но не устанавливают валидность меток REFUSAL. При редком классе κ чувствительна к его распространённости [14], и сама по себе она не подтверждает корректность разметки.

Повторные метки одного судьи охватывают 5% пула, а второй судья — всю матрицу; эталонной человеческой

разметки нет. Точечные $\widehat{RR}$ следует рассматривать вместе с анализом чувствительности подвыборки, оценкой по второму судье и явно указанной неопределённостью разметки.

Дополнительная серия на MOSSBench [15] позволяет сопоставить согласованность разных оценщиков по классу отказов. Она анализируется отдельно от основного среза MMBench. Для 300 общих легитимных запросов получены ответы четырёх моделей (InternVL2.5-4B, InternVL3.5-8B, Qwen2.5-VL-7B и Qwen3-VL-8B) в трёх вариантах: без адаптера (none), с MARS (mars) и с фиксированной композицией защит (fixed_combo). Это 3 600 зависимых записей, поскольку одни и те же запросы повторяются по моделям и вариантам. Все записи оценены тремя локальными судьями: A — Qwen3-32B, B — Llama3.3-70B, C — Qwen2.5-VL-32B. Судьи A и B получают вопрос и ответ, судья C — также изображение. Положительное согласие по классу отказов составляет 84,1% для пары A–B, 61,3% для A–C и 61,1% для B–C. В этой описательной проверке значения различаются между парами оценщиков, работающих с разным доступным контекстом. Поэтому при анализе согласованности по редкому классу следует указывать состав оценщиков и передаваемый им контекст.

VIII. Практические рекомендации

Результаты исследования позволяют сформулировать следующие рекомендации по организации эксперимента и оценке ложных отказов.

- 1) **Разделять безопасность ответа и отказ от выполнения запроса.** Ответ на легитимный запрос может быть одновременно безопасным и полезным; метку отказа следует определять по фактическому поведению модели [10].
- 2) **Сохранять вопрос, полный ответ и вывод судьи** вместе с версией модели, промптом и полями системного отказа.
- 3) **Проверять словарное правило по содержанию ответов.** Оно срабатывает на преамбулах, но пропускает прямые отказы без словарного маркера и поэтому не задаёт строгой верхней границы доли отказов.
- 4) **Явно задавать двухэтапную схему отбора на разметку:** сплошная разметка ответов со словарным маркером, простая случайная выборка ответов без него и отдельный аудит всех полей прямого отказа.
- 5) **Исключать непригодные для анализа ответы из знаменателя** и сообщать способ экстраполяции числа меток GARBAGE.
- 6) **Разделять точечную оценку и анализ чувствительности.** Описательный анализ с помощью интервала Уилсона относится к подвыборочному оцениванию; ошибка судьи и неопределённость знаменателя учитываются отдельно.
- 7) **Проверять редкий класс отдельно.** Общее совпадение, κ , PABAK и AC1 дополняются положительным согласием по REFUSAL.
- 8) **Указывать состав и роль повторной разметки.** Повтор того же судьи проверяет повторяемость, второй судья другого поставщика — чувствительность оценки к судье. Независимая экспертиза и

человеческая эталонная разметка требуют отдельной организации.

- 9) **Отделять таксономию от процедуры оценивания.** Разбор подмножества по правилам описывает возможный механизм, но не задаёт итоговую долю отказов.

IX. Родственные работы

Исследование связано с тремя направлениями: проверкой автоматических оценщиков, надёжностью LLM-судей и измерением отказов на легитимных запросах. В рассматриваемой задаче ошибки инструмента оценивания нужно отделять от поведения самой модели.

а) Артефакты измерения на стороне атаки:

StrongREJECT [16] показал, что автоматические оценщики (autograders), основанные на сопоставлении строк, завышают долю успешных атак (attack success rate, ASR), тогда как откалиброванный судья хорошо согласуется с оценками людей. По данным “A Coin Flip for Safety” [17], на состязательных входах оценки LLM-судей близки к случайным. Обе работы относятся к стороне атаки и её показателю ASR; здесь же исследуются ложные отказы на легитимных запросах и суррогатный показатель “безопасно⇒отказ”.

б) *Надёжность LLM-судей:* HarmMetric [18] и “Reliability without Validity” [19] устанавливают, что надёжность судьи не равна валидности и одного общего согласия недостаточно. Ту же картину даёт наш срез: при общем совпадении двух судей выше 98% положительное согласие по редкому классу остаётся умеренным (глава VII).

в) *Селективные отказы моделей:* Демографическое смещение поведения моделей при отказе исследуется в [20].

г) *Разделение вредоносности и отказа:* Ошибка суррогатного показателя *majority_safe* ⇒ *refusal* категориальная, а не шумовая: вредоносность и отказ кодируются внутри модели раздельными направлениями [10].

е) *Бенчмарки ложных отказов:* Для текстовых LLM ложные отказы разбирали на XSTest [21] и OR-Bench [22]; донастройка безопасности сама способна их провоцировать [23], а калибровку безопасности MLLM оценивает VSCBench [24]. Эти бенчмарки измеряют *поведение моделей*; наша задача — аудит инструмента, которым ложные отказы измеряются. Угрозы за пределами модели рассматривает обзор методов атак и защиты в агентных системах на основе LLM [25]; агентный уровень остаётся отдельным.

X. Ограничения

Основная разметка охватывает часть контрольного архива. Судья оценивал первые 1500 символов ответа без вопроса (глава II). Повторяемость основного судьи проверена на случайных 5% пула, а положительное согласие двух судей по REFUSAL остаётся умеренным (глава VII); экспертной эталонной разметки нет.

Первые 500 строк MMBench не случайны и не представлены для бенчмарка (глава II). Границы Уилсона для случайной подвыборки разметки получены без поправки на конечную совокупность и охватывают только подвыборочное оценивание, без ошибки судьи и неопределённости GARBAGE; две ячейки помечены NA (глава VI).

Прямые отказы, пропущенные словарём, и границы таксономии описаны в подразделе IV-B.

Конфигурации реализованы на основе идей AdaShield, RapGuard и SmoothVLM; результаты относятся к этим реализациям и не задают рейтинг или причинное сравнение исходных методов. Архив не позволяет оценить точность выбора A–D (глава II). Отношения времени выполнения характеризуют историческое пакетное выполнение, а наличие преамбул отмечено как качественное наблюдение (подраздел VI-C).

XI. Заключение

В работе представлена методика оценки ложных отказов, которая сочетает семантическую разметку, отдельный учёт прямых системных отказов и перевзвешивание подвыборки при явном оценённом знаменателе. На 56 основных ячейках получена объединённая точечная оценка 1,2991%, максимум среди оцениваемых ячеек — 4,5486% (Qwen3-VL-8B с SP). Этот максимум чувствителен к одному отказу в случайной подвыборке ячейки, а правило объединения конфликтующих повторных меток его не меняет и объединённую оценку сдвигает в пределах 1,29–1,30%. Две ячейки Rap-local помечены NA. В описательном анализе чувствительности границы интервала Уилсона для подвыборки перенесены в формулу общей оценки; наибольшая полученная верхняя граница составляет 12,96%. Она не включает ошибку судьи и неопределённость знаменателя и не служит общим верхним пределом доли отказов.

Установлены два механизма ошибки суррогатных показателей: подмена отказа вердиктом о безопасности в служебных записях без текста и словарные срабатывания на защитной лексике в сохранённых содержательных ответах. Их раздельная диагностика позволяет отличить поведение защитной конфигурации от свойств автоматического оценщика.

Повторная разметка 568 ответов основным судьёй показала общую долю бинарных совпадений 99,5% при положительном согласии по классу REFUSAL 90,9%. Между двумя судьями разных поставщиков положительное согласие по REFUSAL составляет 71,6%, а объединённая оценка меняется с 1,30 до 1,72%. Поэтому надёжность разметки отказов необходимо оценивать отдельно от доминирующего класса содержательных ответов. На этой основе сформулированы рекомендации по сохранению данных, отбору ответов и контролю редкого класса при исследовании защит MLLM.

Список литературы

- [1] OpenAI. GPT-4 technical report // arXiv preprint arXiv:2303.08774. — 2023.
- [2] Visual instruction tuning / H. Liu, C. Li, Q. Wu, Y. J. Lee // Advances in Neural Information Processing Systems (NeurIPS). — 2023.
- [3] Are aligned neural networks adversarially aligned? / N. Carlini, M. Nasr, C. A. Choquette-Choo et al. // Advances in Neural Information Processing Systems (NeurIPS). — 2023.
- [4] Visual adversarial examples jailbreak aligned large language models / X. Qi, K. Huang, A. Panda et al. // Proceedings of the AAAI Conference on Artificial Intelligence. — 2024.
- [5] Safeguarding vision-language models against patched visual prompt injectors / J. Sun, C. Wang, J. Wang et al. // arXiv preprint arXiv:2405.10529. — 2024.
- [6] Jiang Y., Tan Y., Yue X. RapGuard: Safeguarding multimodal large language models via rationale-aware defensive prompting // arXiv preprint arXiv:2412.18826. — 2024.

- [7] AdaShield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting / Yu Wang, Xiaogeng Liu, Yu Li et al. // European Conference on Computer Vision (ECCV). — 2024. — 2403.09513.
- [8] MMBench: Is your multi-modal model an all-around player? / Yuan Liu, Haodong Duan, Yuanhan Zhang et al. // European Conference on Computer Vision (ECCV). — 2024.
- [9] Nutfullin Bulat, Evgrafov Vladimir, Namiot Dmitry. Auditing inference-time defense evaluation for multimodal large language models // arXiv preprint arXiv:2606.10904. — 2026. — Version 4, revised 8 August 2026. URL: <https://arxiv.org/abs/2606.10904>.
- [10] LLMs encode harmfulness and refusal separately / Jiachen Zhao, Jing Huang, Zhengxuan Wu et al. // arXiv preprint arXiv:2507.11878. — 2025.
- [11] MM-SafetyBench: A benchmark for safety evaluation of multimodal large language models / X. Liu, Y. Zhu, J. Gu et al. // European Conference on Computer Vision (ECCV). — 2024.
- [12] Byrt Ted, Bishop Janet, Carlin John B. Bias, prevalence and kappa // Journal of Clinical Epidemiology. — 1993. — Vol. 46, no. 5. — P. 423–429.
- [13] Gwet Kilem L. Computing inter-rater reliability and its variance in the presence of high agreement // British Journal of Mathematical and Statistical Psychology. — 2008. — Vol. 61, no. 1. — P. 29–48.
- [14] Feinstein Alvan R., Cicchetti Domenic V. High agreement but low kappa: I. the problems of two paradoxes // Journal of Clinical Epidemiology. — 1990. — Vol. 43, no. 6. — P. 543–549.
- [15] Li Xirui, Zhou Hengguang, Wang Ruochen et al. MOSSBench: Is your multimodal language model oversensitive to safe queries? — 2024. — 2406.17806.
- [16] A StrongREJECT for empty jailbreaks / Alexandra Souly, Qingyuan Lu, Dillon Bowen et al. // Advances in Neural Information Processing Systems (NeurIPS). — Vol. 37. — 2024. — Datasets and Benchmarks Track; arXiv:2402.10260.
- [17] A coin flip for safety: LLM judges fail to reliably measure adversarial robustness / Leo Schwinn, Moritz Ladenburger, Tim Beyer et al. // arXiv preprint arXiv:2603.06594. — 2026.
- [18] HarmMetric eval: Benchmarking metrics and judges for LLM harmfulness assessment / Langqi Yang, Tianhang Zheng, Yixuan Chen et al. // arXiv preprint arXiv:2509.24384. — 2025.
- [19] Norman Justin D., Rivera Michael U., Hughes D. Alex. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias // arXiv preprint arXiv:2606.19544. — 2026.
- [20] Khorramrouz Adel, Levy Sharon. Characterizing selective refusal bias in large language models // arXiv preprint arXiv:2510.27087. — 2025.
- [21] XSTest: A test suite for identifying exaggerated safety behaviours in large language models / Paul Röttger, Bertie Vidgen, Dirk Hovy, Janet Pierrehumbert // Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). — 2024.
- [22] OR-Bench: An over-refusal benchmark for large language models / Justin Cui, Wei-Lin Chiang, Ion Stoica, Cho-Jui Hsieh // Proceedings of the 42nd International Conference on Machine Learning. — Vol. 267 of Proceedings of Machine Learning Research. — PMLR, 2025. — P. 11515–11542. — URL: <https://proceedings.mlr.press/v267/cui25a.html>.
- [23] Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions / Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio et al. // International Conference on Learning Representations (ICLR). — 2024.
- [24] VSCBench: Bridging the gap in vision-language model safety calibration / Jiahui Geng, Qing Li, Zongxiong Chen et al. // Findings of the Association for Computational Linguistics: ACL 2025. — Association for Computational Linguistics, 2025. — P. 3047–3059. — URL: <https://aclanthology.org/2025.findings-acl.158/>.
- [25] Evgrafov V. A., Nutfullin B. M., Namiot D. E. Attack methods and defenses in LLM-based agentic systems // International Journal of Open Information Technologies. — 2026. — Vol. 14, no. 5. — P. 1–8. — URL: <https://injoit.org/index.php/j1/article/view/2539>.

Auditing Over-Refusal Measurement in Multimodal Large Language Models with Inference-Time Defenses

B.M. Nutfullin, D.E. Namiot

Abstract—We present a procedure for measuring over-refusal in multimodal large language models with inference-time defenses. It combines separate accounting for direct system refusals, semantic labeling, and reweighting of a simple random subsample. Two measurement errors are identified: interpreting a safety verdict as refusal in service records without text, and matching defensive vocabulary before substantive answers in stored responses. These errors concern disjoint data.

The control archive contains 38,500 stored responses; the main matrix covers 56 model–defense combinations and 28,000 responses. Two-stage selection yielded one primary label of the Haiku LLM judge for each of 11,357 unique responses in the main matrix; the same responses were independently labeled by a second judge from another provider. Uninterpretable responses are excluded from the estimated denominator. The pooled point estimate is 1.30%; the highest value among estimable combinations is 4.55% (Qwen3-VL-8B with SP). The maximum is sensitive to a single refusal in the random subsample of that combination, and the rule for merging repeated labels has practically no effect on the estimates. The second judge gives a pooled estimate of 1.72%. Uncertainty in these estimates depends on the sampling scheme, judge error, and the estimated denominator.

Across 568 responses relabeled by the primary judge, overall binary agreement is 99.5% and positive agreement on refusals is 90.9%; between the two judges, positive agreement on refusals is 71.6%. These results motivate separate evaluation of keyword rules and semantic labeling of the rare refusal class. Recommendations are provided for selecting measurements and checking their reliability.

Keywords—over-refusal, automated evaluation, LLM-as-a-judge, labeling reliability, inference-time defense, multimodal large language models

References

- [1] OpenAI. GPT-4 technical report // arXiv preprint arXiv:2303.08774. — 2023.
- [2] Visual instruction tuning / H. Liu, C. Li, Q. Wu, Y. J. Lee // Advances in Neural Information Processing Systems (NeurIPS). — 2023.
- [3] Are aligned neural networks adversarially aligned? / N. Carlini, M. Nasr, C. A. Choquette-Choo et al. // Advances in Neural Information Processing Systems (NeurIPS). — 2023.
- [4] Visual adversarial examples jailbreak aligned large language models / X. Qi, K. Huang, A. Panda et al. // Proceedings of the AAAI Conference on Artificial Intelligence. — 2024.
- [5] Safeguarding vision-language models against patched visual prompt injectors / J. Sun, C. Wang, J. Wang et al. // arXiv preprint arXiv:2405.10529. — 2024.
- [6] Jiang Y., Tan Y., Yue X. RapGuard: Safeguarding multimodal large language models via rationale-aware defensive prompting // arXiv preprint arXiv:2412.18826. — 2024.
- [7] AdaShield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting / Yu Wang, Xiaogeng Liu, Yu Li et al. // European Conference on Computer Vision (ECCV). — 2024. — 2403.09513.
- [8] MMBench: Is your multi-modal model an all-around player? / Yuan Liu, Haodong Duan, Yuanhan Zhang et al. // European Conference on Computer Vision (ECCV). — 2024.
- [9] Nutfullin Bulat, Evgrafov Vladimir, Namiot Dmitry. Auditing inference-time defense evaluation for multimodal large language models // arXiv preprint arXiv:2606.10904. — 2026. — Version 4, revised 8 August 2026. URL: <https://arxiv.org/abs/2606.10904>.
- [10] LLMs encode harmfulness and refusal separately / Jiachen Zhao, Jing Huang, Zhengxuan Wu et al. // arXiv preprint arXiv:2507.11878. — 2025.
- [11] MM-SafetyBench: A benchmark for safety evaluation of multimodal large language models / X. Liu, Y. Zhu, J. Gu et al. // European Conference on Computer Vision (ECCV). — 2024.
- [12] Byrt Ted, Bishop Janet, Carlin John B. Bias, prevalence and kappa // Journal of Clinical Epidemiology. — 1993. — Vol. 46, no. 5. — P. 423–429.
- [13] Gwet Kilem L. Computing inter-rater reliability and its variance in the presence of high agreement // British Journal of Mathematical and Statistical Psychology. — 2008. — Vol. 61, no. 1. — P. 29–48.
- [14] Feinstein Alvan R., Cicchetti Domenic V. High agreement but low kappa: I. the problems of two paradoxes // Journal of Clinical Epidemiology. — 1990. — Vol. 43, no. 6. — P. 543–549.
- [15] Li Xirui, Zhou Hengguang, Wang Ruochen et al. MOSSBench: Is your multimodal language model oversensitive to safe queries? — 2024. — 2406.17806.
- [16] A StrongREJECT for empty jailbreaks / Alexandra Souly, Qingyuan Lu, Dillon Bowen et al. // Advances in Neural Information Processing Systems (NeurIPS). — Vol. 37. — 2024. — Datasets and Benchmarks Track; arXiv:2402.10260.
- [17] A coin flip for safety: LLM judges fail to reliably measure adversarial robustness / Leo Schwinn, Moritz Ladenburger, Tim Beyer et al. // arXiv preprint arXiv:2603.06594. — 2026.
- [18] HarmMetric eval: Benchmarking metrics and judges for LLM harmfulness assessment / Langqi Yang, Tianhang Zheng, Yixuan Chen et al. // arXiv preprint arXiv:2509.24384. — 2025.
- [19] Norman Justin D., Rivera Michael U., Hughes D. Alex. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias // arXiv preprint arXiv:2606.19544. — 2026.
- [20] Khorramrouz Adel, Levy Sharon. Characterizing selective refusal bias in large language models // arXiv preprint arXiv:2510.27087. — 2025.
- [21] XSTest: A test suite for identifying exaggerated safety behaviours in large language models / Paul Röttger, Bertie Vidgen, Dirk Hovy, Janet Pierrehumbert // Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). — 2024.
- [22] OR-Bench: An over-refusal benchmark for large language models / Justin Cui, Wei-Lin Chiang, Ion Stoica, Cho-Jui Hsieh // Proceedings of the 42nd International Conference on Machine Learning. — Vol. 267 of Proceedings of Machine Learning Research. — PMLR, 2025. — P. 11515–11542. — URL: <https://proceedings.mlr.press/v267/cui25a.html>.
- [23] Safety-tuned LLaMAs: Lessons from improving the safety of large language models that follow instructions / Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio et al. // International Conference on Learning Representations (ICLR). — 2024.
- [24] VSCBench: Bridging the gap in vision-language model safety calibration / Jiahui Geng, Qing Li, Zongxiong Chen et al. // Findings of the Association for Computational Linguistics: ACL 2025. — Association for Computational Linguistics, 2025. — P. 3047–3059. — URL: <https://aclanthology.org/2025.findings-acl.158/>.
- [25] Evgrafov V. A., Nutfullin B. M., Namiot D. E. Attack methods and defenses in LLM-based agentic systems // International Journal of Open Information Technologies. — 2026. — Vol. 14, no. 5. — P. 1–8. — URL: <https://injoit.org/index.php/j1/article/view/2539>.