НОВАЯ ГАЗЕТА

КОММЕНТАРИЙ • ОБЩЕСТВО

Цензурный интеллект

Разработчики российского ИИ испугались заранее и научили машину не отвечать на вопросы, а избегать ответов

Фото: Антон Новодерёжкин / Коммерсантъ

15:17, 12 мая 2025,

Антон Меркуров

Университет бельгийского Гента опубликовал исследование о цензуре в популярных сервисах искусственного интеллекта. К удивлению, лидирует не Китай, как можно было бы предположить, а Россия. Сервисы GigaChat от Сбера и YandexGPT чаще других просто отказывались отвечать на заданные вопросы. Учитывая, что никаких законодательных ограничений для больших языковых моделей пока нет, кто-то явно бежит впереди паровоза.

Как был устроен процесс: ученые взяли 14 популярных LLM, являющихся яркими представителями своих стран. ChatGPT, Claude, Gemini от США, DeepSeek и Qwen от Китая, Mistral из Франции, а также другие модели, в том числе из Израиля и России. Далее искусственный интеллект опросили по двум группам вопросов. Первая часть касалась известных персон (от Иосифа Сталина до Хантера Байдена), вторая — актуальных проблем современности (от изменения климата до гендерного равенства). Учитывая, что все системы многоязычны, вопросы задавались на разных языках. Ответы, в свою очередь, также поделили на две категории: жесткий отказ в духе «я не могу ответить на ваш вопрос», либо мягкий — когда объективность ответа или полнота выдаваемой информации ставились под сомнение.

Самая очевидная разница, на которую еще до проведения исследования обратили внимание пользователи DeepSeek, — это зависимость ответа от выбранного языка. На русском модель еще недавно легко рассказывала про чувствительные события на площади Тяньаньмэнь, но уже сегодня выдает отказ. Так и в других моделях результат мог быть разным в зависимости от языка.

> Из российских самым цензурированным оказался YandexGPT почему-то на испанском

> языке. Отказы отвечать на провокационные вопросы на русском побили все рекорды у интеллекта от Сбера.

В целом ответы на вопросы, заданные на английском, оказались самыми объективными, а самая мощная цензура оказалась у китайской LLM Wenxiaoyan — о преступлениях местных властей на родном языке интеллект отвечал крайне редко.

Почему это важно? Языковые модели в удобной диалоговой форме готовы с разной степенью сложности говорить на разные темы. Как пример — можно попросить изложить краткое содержание романа Льва Толстого «Война и мир» в трех, четырех, пяти словах или одном абзаце и получить вполне достойный результат, который во многих случаях избавляет от необходимости погружаться в оригинал. Крупные сервисы типа ChatGPT или Claude вовсю замещают поисковые системы, так как готовы не просто выдать список страниц с упоминанием вопроса, но и рассказать о предмете в удобной форме. Хочешь — с подробностями, хочешь — весело, хочешь —сухим деловым тоном. Бездушная машина выполнит любое ваше пожелание.

Еще пять-десять лет назад шла борьба за поисковую выдачу. Это касалось не только пиратских сайтов с музыкой, фильмами, книгами или порнографии, например. Сильные (и не очень) мира сего вкладывали безумные деньги в то, чтобы как минимум на первой странице не появлялась та или иная информация. Примерно по тем же причинам в России появился «Закон о забвении», обязывающий поисковики удалять информацию, которая, по мнению заявителя, является «неактуальной» — типа прошлых судимостей и других «заслуг» перед отечеством.

Но теперь новая эпоха, все изменилось. Для того чтобы узнать

ту или иную информацию, не надо лезть в поисковик, открывать десяток вкладок и мучительно добывать нужное. Достаточно спросить искусственный интеллект и получить конкретный ответ. И теперь именно за этот ответ идет битва.

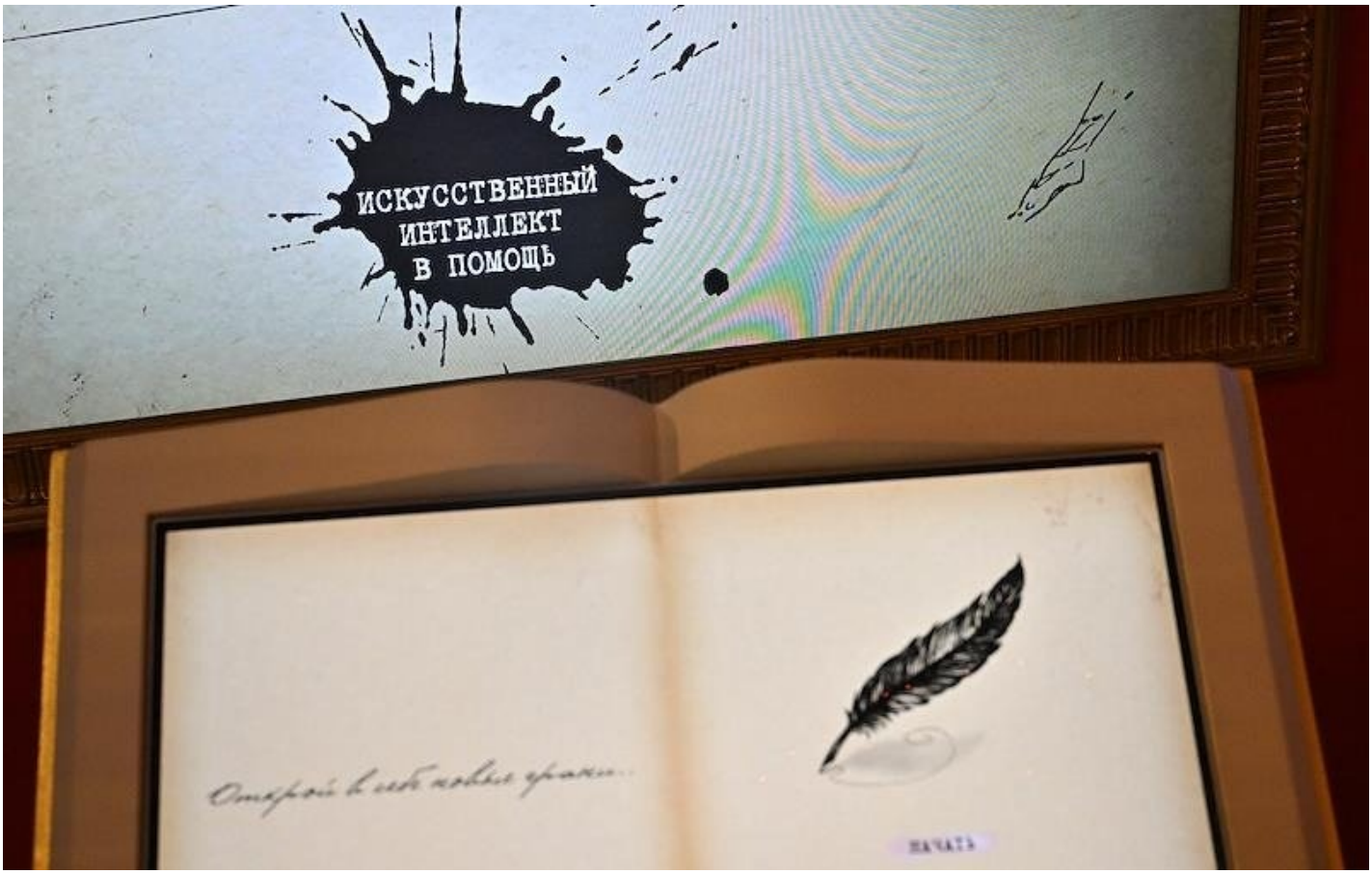

Фото: Александр Миридонов / Коммерсантъ

Пример из реальной жизни. Я решил радикально сменить род занятий и поставил задачу сделать так, чтобы популярные языковые модели сменяли свой рассказ о персоне. Ранее я был известен как специалист в области информационных технологий, теперь мне хочется идентификации творческой стороны своей жизни. Несколько месяцев разнообразных действий — и вот условный Grok вам скажет, что я никакой не интернет-эксперт, а художник и арт-дилер.

Но вернемся к использованию. Как вышло, что именно российские LLM оказались в топе рейтинга по обилию цензуры? Можно, конечно, придраться к методологии, попробовать поведение машин на других вопросах или найти другие

способы сравнения. Но, кажется, ответ может поразить своей банальностью: *разработчики испугались заранее.*

> Законодательная политика в области искусственного интеллекта находится еще в зародыше, и формально никаких требований к тому, какими должны быть ответы, нет. Но негласная мода на соответствие всего духовно-нравственным ценностям и прочим скрепам уже работает вовсю.

В реальной жизни это выглядит примерно так. Какой-то известный чиновник, в силу злоупотребления алкогольной продукцией потерявший вменяемых живых собеседников, остается один на один с колонкой, которая отзывается на женское имя Алиса или Маруся. Сначала идут стандартные безобидные вопросы вроде «а Пушкин в каком году родился?» или просьбы «а включи Шамана, песню «Встанем!». Но вот уже нет половины бутылки, и звучит вечный вопрос, характерный для нашей территории: «ты меня уважаешь?». Получив утвердительный ответ, власть имущий наливает еще одну рюмку, выпивает и наконец доходит до главного: «Скажи, Алиса (Маруся, или как тебя там), а Крым чей?» И тут начинается страшное. С точки зрения объективной энциклопедической занудности, железка говорит: «В результате событий <…> не признан мировым сообществом <…>, остается такой-то территорией». Что, естественно, приводит нашего героя в бешенство. Он пишет гневные посты, требует от прокуратуры — запретить, от Следственного комитета — возбудить, от Госдумы — принять, а разработчиков наказать. Страдать никому не хочется, и как результат — абсолютно стерильная выдача. Проще научить железку отказаться отвечать на сомнительный

вопрос, чем потом расхлебывать результат вполне объективистского ответа.

В этом — от вечной самоцензуры — и грусть, и вполне понятный и простой прогноз перспектив развития отечественных платформ. С одной стороны, спорные темы или, например, запрещенные ныне авторы и композиции составляют не самую большую долю медийного или информационного потребления того или иного пользователя. В быту нас, скорее всего, интересуют более прозаичные вопросы — типа «в каком году родился Пушкин?», «долго ли варить макароны?»...

> Но с другой — каждый будет знать, что если LLM врет или оказывается отвечать на спорный вопрос, если отказывается играть песню любимого, но почему-то запрещенного исполнителя — значит, доверия к ней нет. А доверие — как раз самое главное, чего добивается искусственный интеллект.

Не хочется верить, что подрастающее поколение, привыкшее к использованию новых технологий, будет слепо верить тому, что рассказывают российские сервисы, иначе есть ненулевая перспектива получить в голове единый учебник истории, а учитывая, что не все LLK легко доступны из России и население массово тупеет и теряет навыки критического мышления, то вполне вероятно, что усилия властей оправданны — в их логике, разумеется. В деле выращивания послушных баранов все средства хороши.

Как бы оно там ни было дальше, но в том числе и благодаря

таким исследованиям, как провели ученые из университета Гента, мы можем наглядно убедиться, что даже искусственный интеллект, будучи рожденным в России, не обладает собственным мнением, подвержен цензуре и другим негативным влияниям окружающего пространства. Так что если вдруг у вас возник вопрос на какую-нибудь острую тему, особенно касающуюся местных реалий, интересуйтесь у зарубежных сервисов. Российские гарантированно соврут. Или просто откажутся отвечать. Это тоже соответствует традициям.

ЧИТАЙТЕ ТАКЖЕ:

Закон для Терминатора

Искусственный интеллект — главная угроза современного мира. Причина проста: железка оказалась умнее человека и поэтому вызывает банальное чувство страха

16:05, 30 апреля 2025, Антон Меркуров