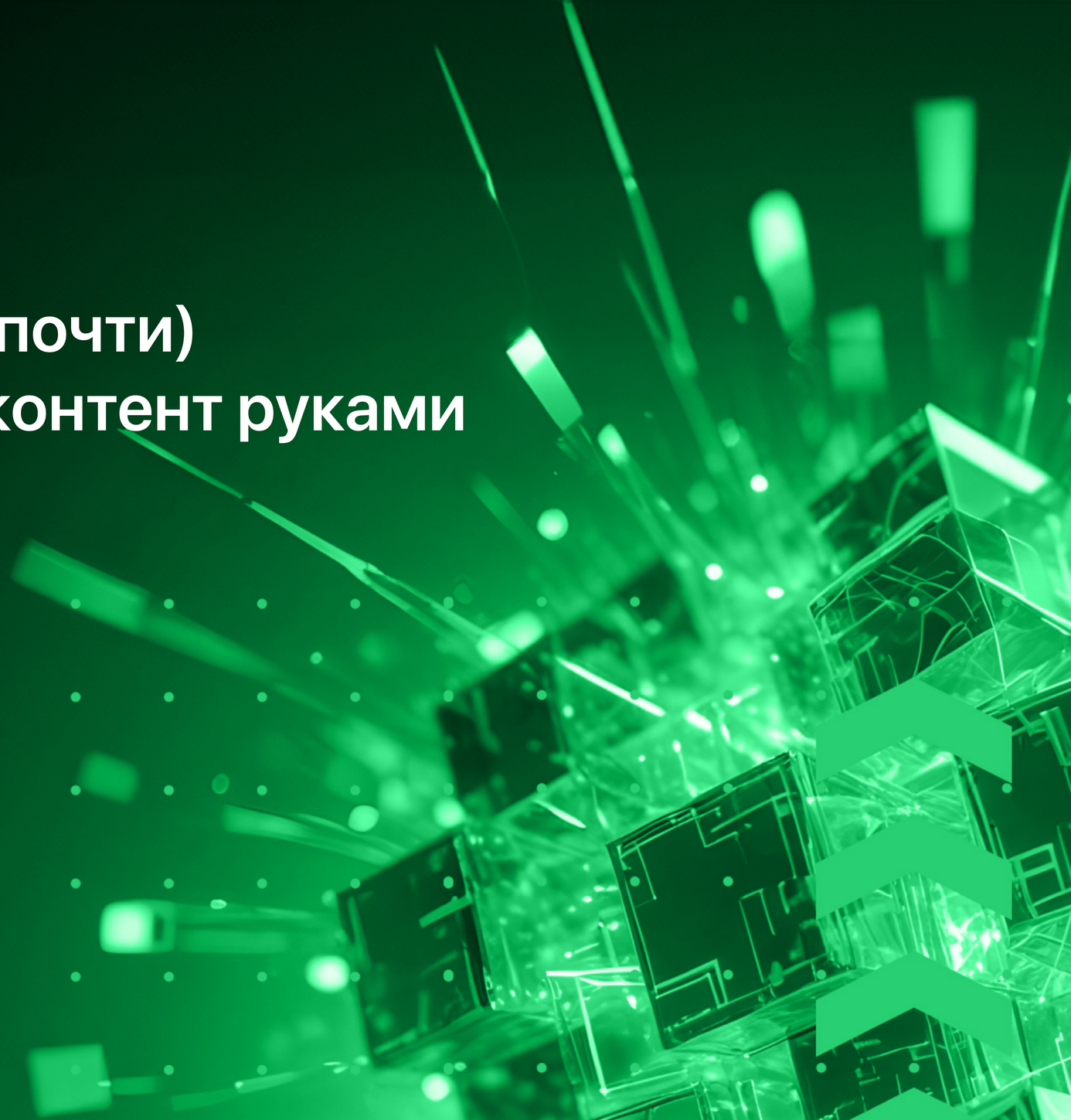
CoNews FORUM

ИИ-оптимизация: как мы (почти) перестали модерировать контент руками

Дмитрий Дуда



Почему модерация необходима

«Выберу.ру» — один из крупнейших финансовых маркетплейсов России.

Входит в перечень социально значимых информационных ресурсов Минкомсвязи РФ — **повышенные обязательства по контролю за публикуемым контентом.**

7 млн
пользователей в месяц

до 40
материалов в день

до 1000
комментариев в день

30% комментариев требуют фильтрации

Содержат нецензурную брань, политику, упоминания тем, освещение которых ограничено действующим законодательством

Минусы ручной модерации

До 35 часов в неделю

на ручную модерацию комментариев и статей

Конфликты

при пропуске нежелательных комментариев

Юридические риски

при публикации запрещённого контента

Потеря пользователей

из-за долгого ожидания модерации

Ручная модерация — это затратно, рискованно и влияет на рост аудитории

Автоматизация помогает сохранить качество, безопасность и лояльность пользователей

Задача проекта

Создать систему на базе искусственного интеллекта, которая:

- анализирует каждый комментарий в режиме реального времени
- проверяет каждую публикуемую статью
- выявляет контент, недопустимый для размещения на ресурсе
- автоматически публикует контент, прошедший модерацию



Технологическая структура ИИ-модератора

Для модерации используется языковая модель — **Выберу ИИ**.

LLM

калиброванная под домены, связанные с деятельностью маркетплейса

RAG

внешние источники данных о запрещенных организациях и физлицах

API-интеграция

с CMS сайта: анализ запускается автоматически при отправке комментария или сохранении статьи

Админ-панель

для модераторов: очередь пограничных случаев, статистика, управление категориями фильтрации

Модерация комментариев

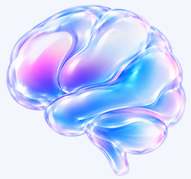


до 10 секунд

Скорость принятия решения
После модерации
автоматическая публикация

ИИ-модель проверяет каждый комментарий на:

- оскорбления и персональные нападки;
- ненормативную лексику;
- спам и рекламные публикации;
- обсуждение политических тем в формате конфликтного дискурса;
- упоминания тем, освещение которых ограничено действующим законодательством РФ;
- призывы к незаконным действиям.



Модель не ищет стоп-слова — она анализирует семантику и контекст высказывания, понимает сарказм, косвенные формулировки и завуалированные попытки обойти фильтр.

Модерация статей

Анализ материалов, подготовленных редакцией к публикации:

- **Сканирование текста** на упоминания тем, ограниченных законодательством РФ;
- **Выделение** потенциально проблемных фрагментов;
- **Формирование отчёта** для редактора с указанием причин срабатывания;
- **Повторная проверка** после внесения изменений.



Финальное решение остаётся за редактором

ИИ выступает как инструмент предварительного контроля, а не заменяет редакционное решение.



Базовый алгоритм

Направленный ациклический граф:

1

Чанкинг

Разделение исходного текста на фрагменты

2

Определение сущностей

Выделение значимых объектов и упоминаний

3

Определение намерения

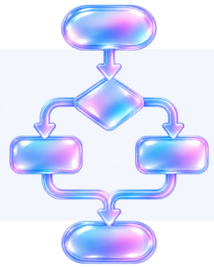
Понимание цели и смысла высказывания

4

Проверка тематик

Анализ соответствия тематическим ограничениям

Единый алгоритм последовательно анализирует текст на каждом этапе модерации



Обогащение контекста



Результаты внедрения ИИ-модерации

Экономия рабочего времени в неделю **до 35 часов**

Покрытие запрещенного контента составляет **98%**

Рост среднего числа комментариев на статью **↑ 27%**

Пользователи знают, что их голос будет услышан сразу

Полное покрытие **100%**

Обрабатывается 100% пользовательских материалов

Комплаенс

публикуемый контент соответствует требованиям законодательства



Масштабируемость и потенциал развития

Возможность добавления новых категорий фильтрации

Мониторинг настроений пользователей
За счет классификации тональности комментариев можно определить что интересно аудитории, а что нет

Применимость в других отраслях
Модульная архитектура позволяет адаптировать систему для любого медиа-ресурса, интернет-магазина с отзывами, социальной платформы или корпоративного портала



Выберу

Спасибо
за внимание!

