

Видео-рекомендер отвечает на два вопроса — что зритель может смотреть и что показать первым — и держится на том, что он оптимизирует. Заложите двухэтапную воронку, три стратегии релевантности, решения cold start и метрику, которая держит честным, до начала. Возможности вендоров в 2026 меняются — сверяйте вживую.

1 · ДВУХЭТАПНАЯ ВОРОНКА (чтобы масштабировалось)

- ☐ **Candidate generation** — дешёвая широкая сеть по всему каталогу; возвращает сотни (миллионы → сотни).
- ☐ **Ranking** — дорогая модель оценивает только шорт-лист и упорядочивает десятки (сотни → десятки).
- ☐ **Зачем делить** — оценить весь каталог на зрителя на загрузку не влезает в ~200 мс бюджета экрана.
- ☐ **Смешивайте источники** — collaborative + content-based + редакция дают один шорт-лист; ranking сортирует.

2 · ТРИ СТРАТЕГИИ РЕЛЕВАНТНОСТИ (планируйте гибрид)

- ☐ **Collaborative filtering** — «похожие на вас смотрели это»; нужно поведение; слеп к новым тайтлам.
- ☐ **Content-based** — «похоже на то, что вы смотрели»; работает на метаданных; тянет новые, но повторяет.
- ☐ **Гибрид** — оба, чтобы закрыть слепые зоны друг друга; к нему приходит почти каждая реальная система.
- ☐ **Метаданные — топливо** — content-based рекомендации не лучше, чем чистый структурированный каталог метаданных.

3 · РЕШЕНИЯ COLD START (день, когда система не знает ничего)

- ☐ **Новый зритель** — выбор вкусов в онбординге, общая популярность и контекст (страна, устройство, время).
- ☐ **Новый тайтл** — метаданные рекомендуют его с первого дня; collaborative filtering не может.
- ☐ **Новая платформа** — сначала редакция + content-based; к collaborative — по мере накопления данных.
- ☐ **Планируйте прогрев** — рекомендатель на старте и через год — это разные системы.

4 · МЕРЯЙТЕ WATCH TIME, НЕ КЛИКИ (честная цель)

- ☐ **Клики врут** — ранжирование по кликам продвигает «обманные видео ... (кликбейт)» (YouTube, RecSys 2016).
- ☐ **Честная метрика** — ожидаемый watch time, досмотры и возврат через неделю предсказывают удержание.
- ☐ **Докажите** — каждое изменение выходит за экспериментом по удержанию, не по тапам.
- ☐ **Помните о границах** — у данных просмотра есть граница приватности; рекомендация за спиннером теряет зрителя.

ПОРЯДОК ДЕЙСТВИЙ — ДВА ЭТАПА, ГИБРИД, ПЛАН ПРОГРЕВА И WATCH TIME КАК ЦЕЛЬ

Рекомендательная система — это в основном обвязка вокруг модели, а не сама модель. По порядку. Первое: разделите работу на два этапа — дешёвый candidate generation, который сужает весь каталог до нескольких сотен тайтлов, затем дорогой ranking, который оценивает только этот шорт-лист; это единственный способ оставаться быстрым на большом каталоге, и так делают YouTube, Netflix и Amazon. Второе: планируйте гибрид — collaborative filtering находит неожиданное, но слеп к новым тайтлам, content-based тянет новые, но повторяет, так что нужны оба, а content-based рекомендации не лучше ваших метаданных. Третье: проектируйте под cold start с первого дня — грейте новых зрителей онбордингом и популярным, тяните новые тайтлы метаданными, а новую платформу запускайте на редакции и content-based, пока не накопится поведение. Четвёртое и главное: оптимизируйте ranking на ожидаемый watch time и возврат, никогда на клики — система на кликах учится кликбейту — и доказывайте каждое изменение экспериментом по удержанию. Возможности вендоров и архитектуры моделей (переход к foundation-модели в 2025, two-tower retrieval) меняются — сверяйте перед планированием.