



Оцени экономику RAG-систем

Инструкция о том, как предсказывать результаты
и затраты, чтобы не разориться

3 уровня

оценки
экономики

«КОРУС КОНСАЛТИНГ»

-  Российская IT-компания
-  Автоматизация бизнес-процессов
-  ИТ-инфраструктура
-  ИТ-аутсорсинг
-  Заказная разработка
-  Корпоративный акселератор

2000 год

основание
компании

10+

Компаний
в экосистеме

**СПб, Москва,
Астана**
офисы

2050

завершенных
проектов

1500

сотрудников

9 млрд

выручка за 2024 год

ЭКОСИСТЕМА АВАНДОК

Уникальная российская платформа для управления электронными документами, бизнес-процессами и корпоративным контентом

ДОКУМЕНТЫ

автоматизация
процессов
документооборота
организации

AI

применение ИИ
для автоматизации
задач

ТРЕКЕР

управление
задачами, в том числе
в рамках проектной
деятельности

РЕПОЗИТОРИЙ

место хранения
контента
информационных
систем организации

SERVICE DESK

автоматизация
работы с
клиентскими
обращениями

ДИСК

место хранения
контента, вне
производственных
систем

CDP, EDP, Loyalty

эффективное
использование
данных клиентов и
сотрудников

CRM

управление
взаимоотношениями
с клиентами и
партнерами

ИНСТРУКЦИЯ

Оцените каждый пункт по шкале:

✓ Да (реализовано полностью)

~ Частично (есть, но требует улучшения)

✗ Нет (отсутствует или не реализовано)

Да = 2 балла, Частично = 1 балл, Нет = 0 баллов

Прозрачность и видимость затрат

Понимаем ли мы, куда именно уходят деньги?

Я знаю общий месячный бюджет на работу ИИ-проекта

☐ ✓ ☐ ~ ☐ ✗

У меня есть дашборд или отчет, где видны текущие расходы

☐ ✓ ☐ ~ ☐ ✗

Я понимаю структуру затрат (API, инфраструктура, разработка и т.д.)

☐ ✓ ☐ ~ ☐ ✗

Я могу увидеть, сколько стоит обработка одного запроса/задачи

☐ ✓ ☐ ~ ☐ ✗

Затраты разделены по бизнес-функциям/продуктам/командам

☐ ✓ ☐ ~ ☐ ✗

Я получаю регулярные отчеты о расходах (еженедельно/ежемесячно)

☐ ✓ ☐ ~ ☐ ✗

Итого баллов _____

Контроль расходов и лимиты

Есть ли у нас механизмы предотвращения перерасхода?

Установлены месячные лимиты расходов на проект

☐ ✓ ☐ ~ ☐ ✗

Настроены автоматические уведомления при приближении к лимиту

☐ ✓ ☐ ~ ☐ ✗

Есть процесс согласования превышения бюджета

☐ ✓ ☐ ~ ☐ ✗

Лимиты установлены для разных пользователей/команд отдельно

☐ ✓ ☐ ~ ☐ ✗

Система может автоматически ограничить расходы при превышении

☐ ✓ ☐ ~ ☐ ✗

У меня есть план действий при внезапном росте расходов

☐ ✓ ☐ ~ ☐ ✗

Итого баллов _____

Эффективность использования (промтты/ запросы)

Оптимально ли мы формулируем задачи для ИИ?

Промтты регулярно проверяются и оптимизируются

☒ ☐ ~ ☐ X

Есть библиотека лучших промттов для типовых задач

☒ ☐ ~ ☐ X

Мы измеряем, сколько токенов тратится на разные типы запросов

☒ ☐ ~ ☐ X

Перед отправкой запросы проверяются на чрезмерную длину

☒ ☐ ~ ☐ X

Мы используем короткие промтты там, где это возможно

☒ ☐ ~ ☐ X

Повторяющиеся инструкции вынесены в системные настройки

☒ ☐ ~ ☐ X

Итого баллов _____

Оптимальный выбор моделей

Используем ли мы подходящие модели для каждой задачи?

Для простых задач используются более дешевые модели

☒ ☐ ~ ☐ X

Есть понимание, когда можно использовать более легкие версии моделей

☒ ☐ ~ ☐ X

Мы сравниваем стоимость и качество разных моделей

☒ ☐ ~ ☐ X

Проводится регулярный анализ: какая модель для какой задачи оптимальна

☒ ☐ ~ ☐ X

Мы рассматриваем альтернативные модели и провайдеров

☒ ☐ ~ ☐ X

Решение о выборе модели принимается на основе анализа цена/качество

☒ ☐ ~ ☐ X

Итого баллов _____

Эффективность использования (промтты/ запросы)

Оптимально ли мы формулируем задачи для ИИ?

Промтты регулярно проверяются и оптимизируются

☒ ☐ ☐

Есть библиотека лучших промттов для типовых задач

☒ ☐ ☐

Мы измеряем, сколько токенов тратится на разные типы запросов

☒ ☐ ☐

Перед отправкой запросы проверяются на чрезмерную длину

☒ ☐ ☐

Мы используем короткие промтты там, где это возможно

☒ ☐ ☐

Повторяющиеся инструкции вынесены в системные настройки

☒ ☐ ☐

Итого баллов_____

Оптимальный выбор моделей

Используем ли мы подходящие модели для каждой задачи?

Для простых задач используются более дешевые модели

☒ ☐ ☐

Есть понимание, когда можно использовать более легкие версии моделей

☒ ☐ ☐

Мы сравниваем стоимость и качество разных моделей

☒ ☐ ☐

Проводится регулярный анализ: какая модель для какой задачи оптимальна

☒ ☐ ☐

Мы рассматриваем альтернативные модели и провайдеров

☒ ☐ ☐

Решение о выборе модели принимается на основе анализа цена/качество

☒ ☐ ☐

Итого баллов_____

Повторное использование результатов (кэширование)

Избегаем ли мы лишних повторных запросов?

Реализовано кэширование для одинаковых или похожих запросов

☒ ☐ ☒

Результаты типовых задач сохраняются и переиспользуются

☒ ☐ ☒

Мы понимаем, какой процент запросов можно закэшировать

☒ ☐ ☒

Настроено время жизни кэша в зависимости от типа данных

☒ ☐ ☒

Мы отслеживаем эффективность кэширования (hit rate)

☒ ☐ ☒

Итого баллов_____

Группировка запросов (батчинг)

Обрабатываем ли мы задачи эффективными партиями?

Похожие задачи группируются и обрабатываются вместе

☒ ☐ ☒

Используется batch API для массовых операций

☒ ☐ ☒

Есть понимание, какие задачи можно обрабатывать в режиме батчинга

☒ ☐ ☒

Измеряется экономия от батчинга по сравнению с обычными запросами

☒ ☐ ☒

Итого баллов_____

Оптимизация инфраструктуры

Эффективно ли используются технические ресурсы?

Инфраструктура масштабируется в зависимости от нагрузки

☐ ✓ ☐ ~ ☐ X

Неиспользуемые ресурсы автоматически отключаются

☐ ✓ ☐ ~ ☐ X

Проводится регулярный аудит инфраструктуры на избыточность

☐ ✓ ☐ ~ ☐ X

Используются резервированные мощности или spot-инстансы где возможно

☐ ✓ ☐ ~ ☐ X

Мы понимаем разницу в стоимости on-demand vs reserved ресурсов

☐ ✓ ☐ ~ ☐ X

Итого баллов _____

Измерение эффективности

Понимаем ли мы ценность, которую получаем за деньги?

Считается стоимость одного пользовательского запроса/действия

☐ ✓ ☐ ~ ☐ X

Отслеживается ROI (возврат инвестиций) от ИИ-проекта

☐ ✓ ☐ ~ ☐ X

Есть метрики качества работы ИИ (точность, релевантность)

☐ ✓ ☐ ~ ☐ X

Бизнес-эффект измеряется и сопоставляется с затратами

☐ ✓ ☐ ~ ☐ X

Мы понимаем юнит-экономику нашего ИИ-продукта

☐ ✓ ☐ ~ ☐ X

Регулярно сравниваем показатели с предыдущими периодами

☐ ✓ ☐ ~ ☐ X

Итого баллов _____

Процессы и управление

Есть ли у нас регулярные процессы управления экономикой?

Проводятся регулярные ревью расходов (минимум ежемесячно)

☒ ☐ ☐

Есть назначенный ответственный за управление затратами

☒ ☐ ☐

Команда обучена лучшим практикам оптимизации затрат

☒ ☐ ☐

Есть документированный процесс оптимизации затрат

☒ ☐ ☐

Решения об изменениях принимаются на основе данных

☒ ☐ ☐

Есть культура cost-awareness в команде проекта

☒ ☐ ☐

Итого баллов_____

Защита от рисков и аномалий

Защищены ли от неожиданностей и злоупотреблений?

Есть мониторинг аномальных скачков расходов

☒ ☐ ☐

Настроены алерты на подозрительную активность

☒ ☐ ☐

Доступ к дорогим операциям ограничен и контролируется

☒ ☐ ☐

Есть процесс расследования случаев перерасхода

☒ ☐ ☐

Предусмотрен резервный бюджет на непредвиденные расходы

☒ ☐ ☐

Итого баллов_____

УРОВЕНЬ ЗРЕЛОСТИ ВАШЕГО ПРОЦЕССА

80-110 баллов – Высокий уровень зрелости

Отлично! У вас выстроены процессы управления экономикой ИИ-проекта. Сфокусируйтесь на тонкой настройке и распространении лучших практик

60-79 баллов – Средний уровень зрелости

Хорошая база! Основные процессы есть, но есть зоны для улучшения. Приоритизируйте блоки с наименьшими баллами

40-59 баллов – Базовый уровень

Есть с чем работать. У вас есть отдельные элементы контроля, но нужна системная работа. Начните с блоков 1, 2 и 8 (прозрачность, контроль, метрики)

0-39 баллов – Начальный уровень

Высокий риск перерасхода. Срочно требуется внимание к управлению затратами. Начните с базовой прозрачности (блок 1) и установки лимитов (блок 2)

РЕКОМЕНДАЦИИ ПО ПРИОРИТИЗАЦИИ УЛУЧШЕНИЙ

Определите 3 блока с наименьшим количеством баллов – это ваши приоритеты для работы:

Место	Блок	Баллы	Приоритет
1			🔴 Высокий
2			🟡 Средний
3			🟢 Базовый

БЫСТРЫЕ ПОБЕДЫ (QUICK WINS)

Начните с этих простых действий, которые дадут быстрый эффект:

- ✓ Подключите базовый дашборд затрат (если нет)
- ✓ Установите месячный лимит расходов с уведомлениями
- ✓ Проведите аудит 10 самых частых промптов на оптимизацию
- ✓ Проверьте, используете ли вы самые дешевые модели для простых задач
- ✓ Назначьте ответственного за мониторинг расходов
- ✓ Настройте еженедельный отчет по расходам

МОЙ ПЛАН ДЕЙСТВИЙ

Заполните на основе
результатов чек-листа:

**1. Что начну делать в первую
неделю:**

**2. Что реализую в первый
месяц:**

**3. Долгосрочные цели
(3 месяца):**

КОНТАКТЫ



Алексей Борщов

Менеджер по продуктам
искусственного
интеллекта

✉ AlBorshchov@korusconsulting.ru

✈ [@Borshchov_Aleksey](#) 🖱