CNEWS
FORUM КЕЙСЫ
ОПЫТ ИТ-ЛИДЕРОВ

Практика промышленного применения LLM

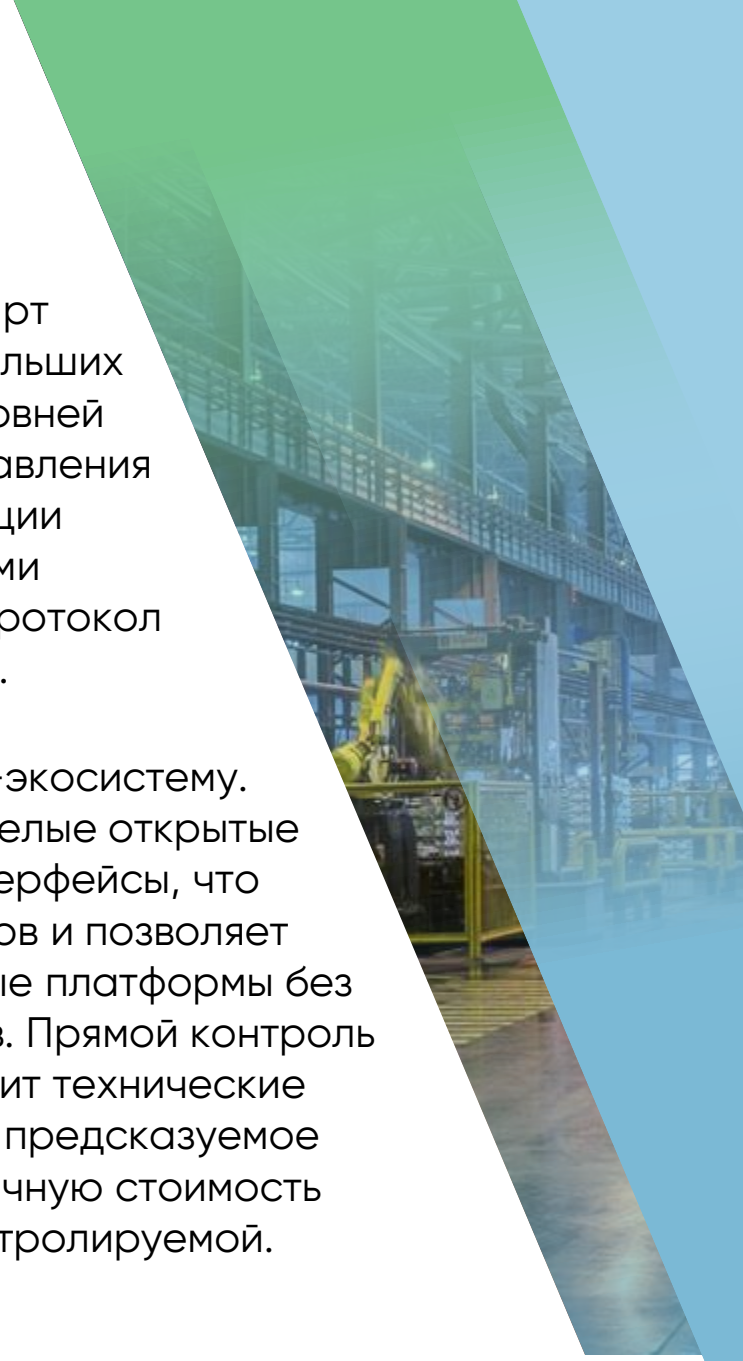
Иван Казарин

Директор направления операционного развития искусственного интеллекта, Департамент технологий искусственного интеллекта, РУСАЛ

Контекст

В мировой индустрии сформировался общепринятый архитектурный стандарт для промышленных систем на базе больших языковых моделей. Он строится из уровней или функциональных плоскостей: управления доступом, оркестрации LLM (реализации бизнес-логики), интеграции с внешними системами через унифицированный протокол MCP и плоскости инференса моделей.

Стандарт опирается на Open Source-экосистему. Для каждой плоскости существуют зрелые открытые решения и стандартизированные интерфейсы, что минимизирует зависимость от вендоров и позволяет проектировать и внедрять собственные платформы без зависимости от внешних поставщиков. Прямой контроль над технологическим стеком переводит технические риски под управление, обеспечивает предсказуемое масштабирование и делает долгосрочную стоимость владения системой прозрачной и контролируемой.

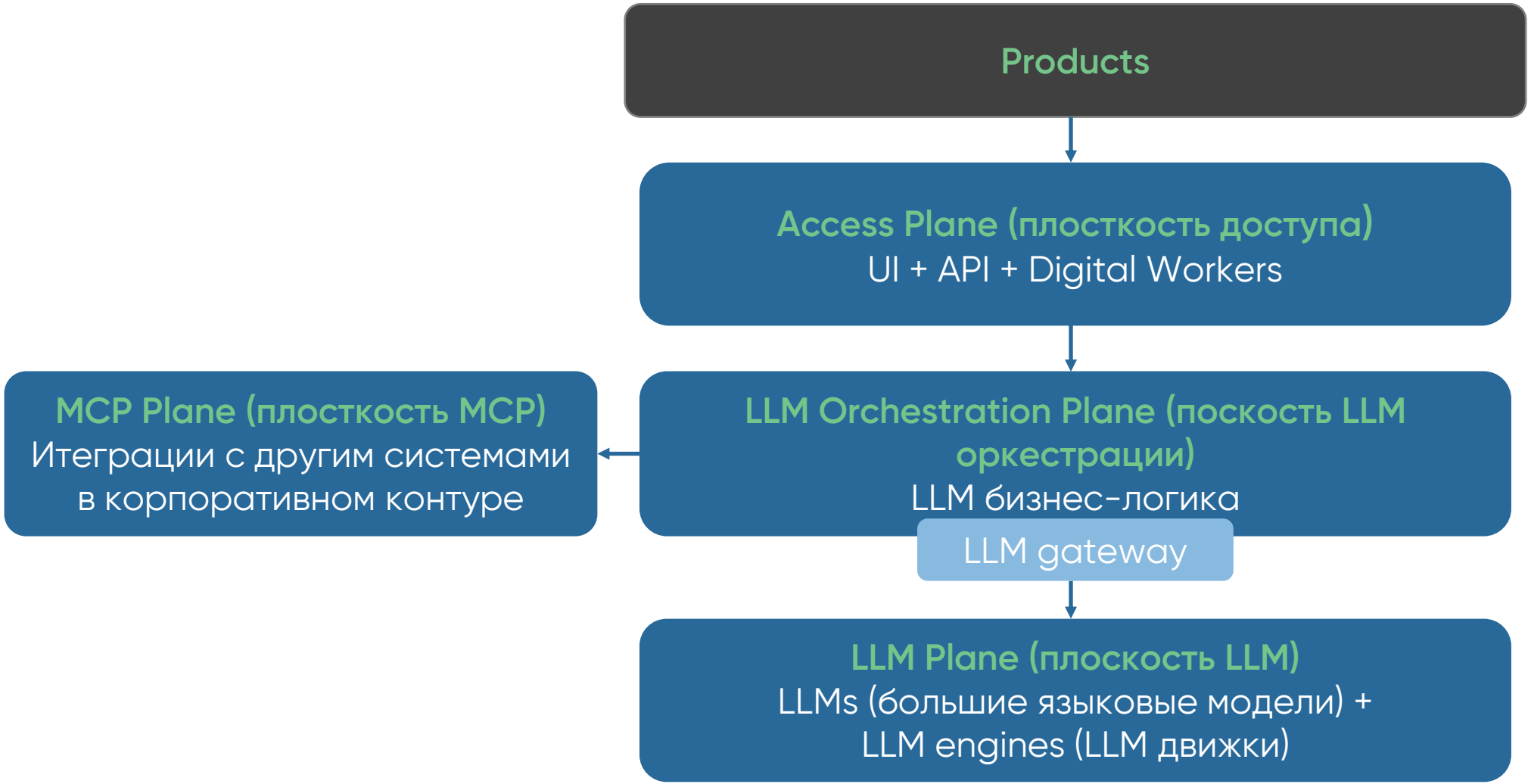


Данная архитектура позволяет запускать, перестраивать или заменять GenAI-сервисы за несколько дней или недель

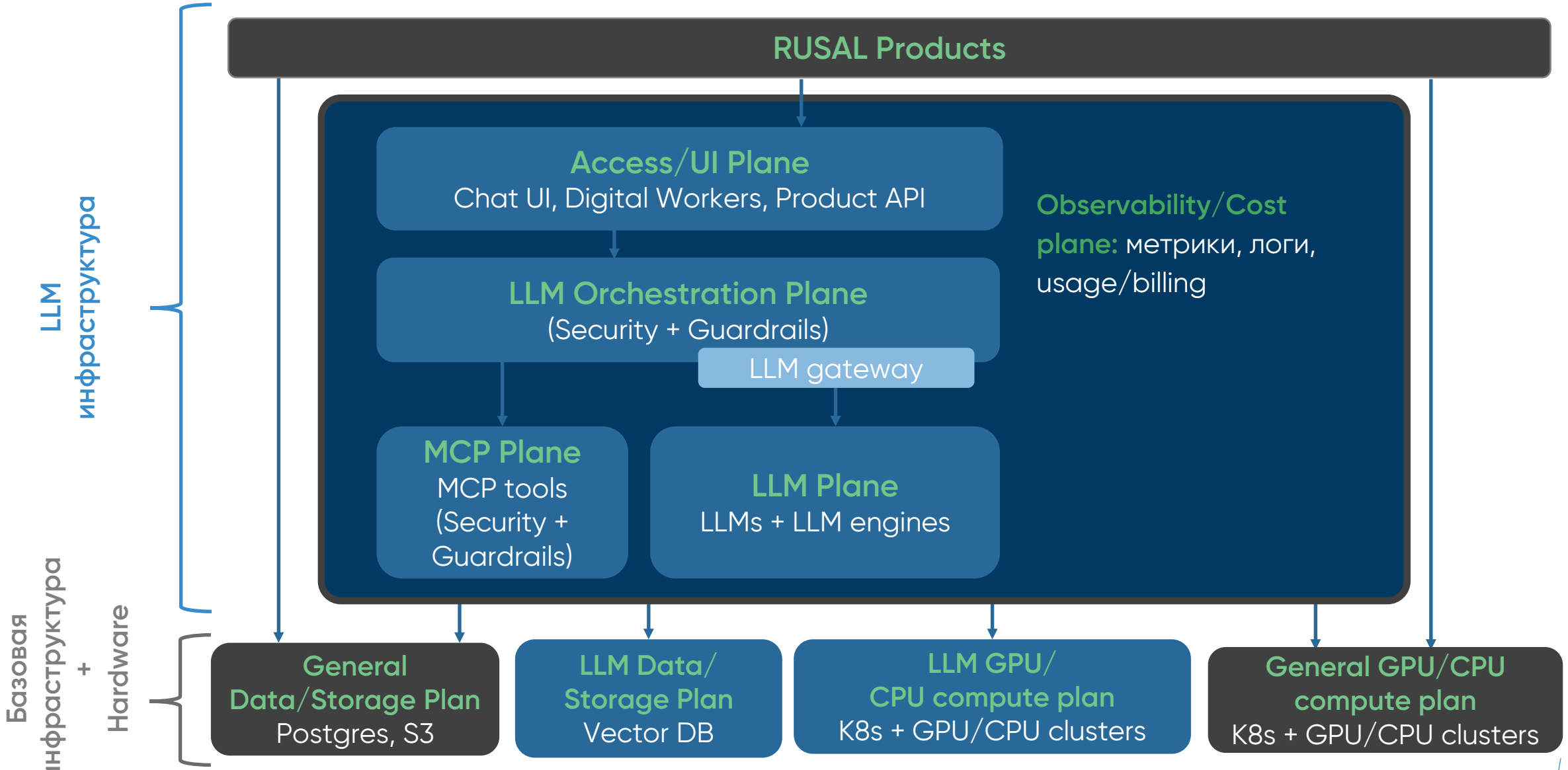
Реальная ценность при промышленном применении LLM создается на уровне платформы, инфраструктуры и команды

Компания, способная поддерживать и развивать такую архитектуру, получает устойчивое операционное преимущество и долгосрочный стратегический актив

Ядро архитектуры стандарта построения промышленных систем на базе LLM



Реализация в РУСАЛе



Единый корпоративный хаб для работы с LLM

Доступ к моделям и сервисам через привычный чат

В одном интерфейсе собраны все доступные для корпоративных пользователей LLM-модели и GenAI-сервисы, опубликованные сотрудниками компании. Сотруднику не обязательно изучать API или настраивать среду: достаточно зайти в знакомый веб-портал, выбрать нужную модель или сервис из каталога и начать диалог.

Промышленная мощность On-Premise

За простым интерфейсом диалога работает выделенный уровень LLM на базе GPU-кластера. Платформа обеспечивает эффективную агрегацию вычислительных мощностей и динамическое распределение нагрузки, что особенно важно в периоды пиковых нагрузок. Бизнес получает производительность LLM корпоративного уровня без необходимости самостоятельной настройки инфраструктуры и управления контейнеризацией.

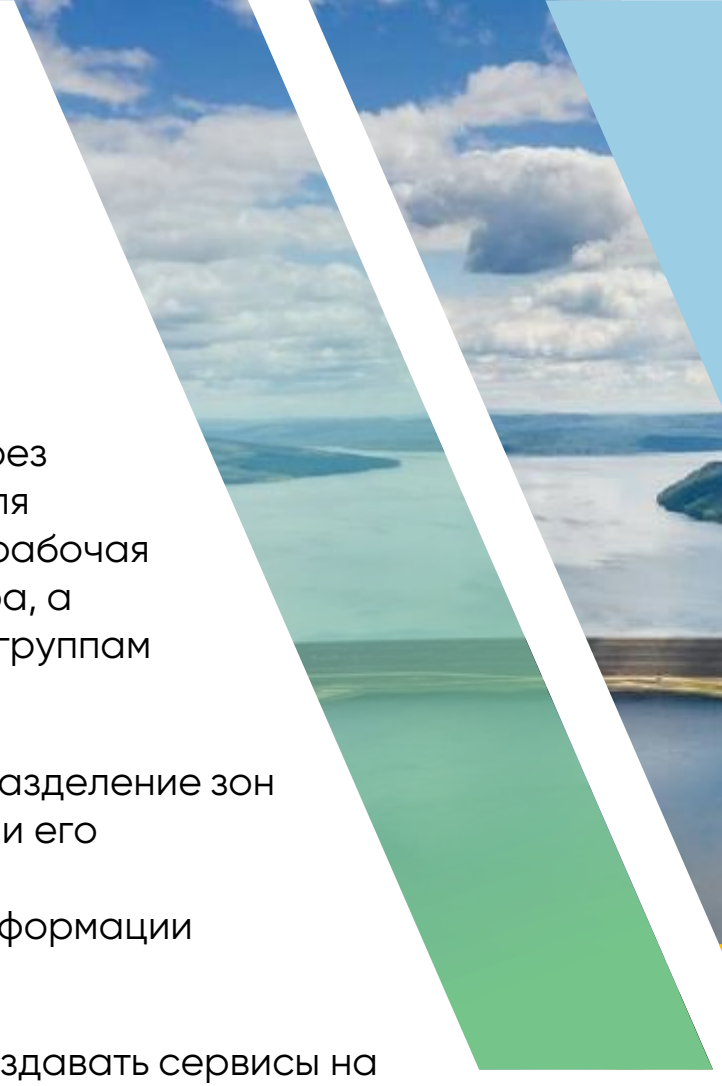
Универсальный доступ регламентирован, каждый запрос проходит централизованную проверку: корпоративная аутентификация, внутренние контрольные механизмы для работы с чувствительной информацией, применение гранулярных политик доступа (RBAC) и фиксация метрик потребления

Сотрудник формулирует задачу на естественным языке, а платформа гарантирует сохранение периметра конфиденциальности, логирование и прозрачную структуру затрат

100%

корпоративных пользователей имеют доступ к LLM-платформе

Управляемые RAG-сервисы



Проектно-ориентированная модель

Доступ к RAG-сервисам осуществляется через единый пользовательский веб-интерфейс. Для каждого проекта формируется выделенная рабочая группа с назначением Knowledge-менеджера, а доступ к сервису предоставляется целевым группам через централизованную систему прав.

Это гарантирует изоляцию данных, четкое разделение зон ответственности между созданием контента и его потреблением, а также предотвращает несанкционированное распространение информации внутри контура.

Knowledge-менеджер имеет возможность создавать сервисы на базе существующих моделей, задавать системные промты по умолчанию, управлять параметрами инференса, а также контролировать и наполнять базы знаний для сервисов. Такое разделение ролей позволяет бизнес-подразделениям настраивать поведение инструментов LLM-платформы под специфику предметной области без дополнительного привлечения специалистов или пересборки инфраструктурных конвейеров.

Для проектов на базе RAG также доступна функция создания отдельных коллекций в системе векторного хранения, с возможностью тонкой настройки параметров поиска

> 20 RAG-сервисов

реализовано и эксплуатируется на базе LLM-платформы

API-контур для подразделений и сервисов

Шлюз и ключи для сотрудников и сервисов

Сотрудники профильных подразделений разработки и аналитики данных используют API-ключи для интеграции ИИ-ассистентов непосредственно в рабочие инструменты и процессы повседневной работы. Бизнес-подразделения применяют единый API-шлюз для обращений профильных ИТ-сервисов. Это формирует двухвекторный подход: от индивидуального повышения продуктивности до совершенствования бизнес-сервисов подразделений и корпоративных информационных систем, как ERP или СЭД.

Data-driven стратегия

LLM-платформа функционирует как управляемая фабрика интеграций для сервисов, где приоритет отдается проектам с оцененным экономическим эффектом. Все API-вызовы маршрутизируются с централизованной аутентификацией через плоскость оркестрации. Платформа непрерывно собирает метрики потребления, включая стоимость токенов. Это позволяет предотвратить перегрузку кластера и формировать портфель сервисов, опираясь на объективные данные об их экономической эффективности и операционной стабильности.

> 60 GenAI-сервисов

в эксплуатации на LLM-платформе с настроенной API-интеграцией

~ 30 GenAI-сервисов

отправляют более 100 запросов в среднем за сутки, значительная часть из них – более 1 000 запросов

Зона создания агентов

Сотрудники могут прототипировать диалоговые сценарии, проектировать многошаговые рабочие процессы и готовить решения к интеграции. Работа пользователей осуществляется в рамках установленных стандартов безопасности и доступа.

LLM-платформа предоставляет всем корпоративным пользователям рабочую зону на базе CPU compute для самостоятельного создания ИИ-агентов.

Инструмент ориентирован не только на AI-инженеров, но также является драйвером развития экспертизы внутри подразделений, стимулируя бизнес-команды самостоятельно осваивать навыки проектирования решений для своих задач

Опытное использование агентской экосистемы

РУСАЛ

В выделенной среде авторизованные сотрудники (разработчики и инженеры) самостоятельно проектируют сложные мультиагентные сценарии, настраивают маршрутизацию задач между взаимодействующими узлами, обрабатывают корпоративные документы и реализуют сценарии дополнения ответов внутренней базой знаний.

Такая архитектурная гибкость позволяет выстраивать глубоко интегрированные технологические процессы без привязки к сторонним оркестраторам, ускорять подготовку точных ответов за счет встроенного RAG-контура с гибридным поиском и подключать решения к целевым системам через нативную поддержку протокола MCP.

Безопасность обеспечивается строгой изоляцией контура, выделением workspace для каждой команды, гранулярной моделью прав на уровне БД и многоуровневыми механизмами защиты, включая фильтрацию опасных промптов и шифрование чувствительных данных.

Дорожная карта формирования бизнес-эффекта

РУСАЛ

Бизнес-эффект от проектов с GenAI-сервисами возникает в результате прохождения этапов:

Шаг 0 – Доступность

LLM-платформа предоставляет корпоративным пользователям доступ к моделям и инструментам. Разрозненные эксперименты заменяются стандартизированным безопасным рабочим контуром.

Шаг 1 – Применение на практике

Подразделения начинают применять LLM-модели и инструменты платформы в ежедневных задачах: настраивают персональные ассистенты, используют готовые сервисы, интегрируют ИИ-решения в процессы работы.

Шаг 2 – Фиксация полезности

Команды фиксируют практическую отдачу: насколько быстро решаются типовые запросы, снижается нагрузка на сотрудников, какой прирост скорости дают инструменты на конкретных операциях.

Шаг 3 – Оценка экономического эффекта

Проекты с GenAI-сервисами переводят в финансовые метрики. На основе оценки экономического эффекта рассчитывается ROI, оптимизируется бюджет на GPU-ресурсы и принимается решение о развитии и тиражировании.

> 100 GenAI-сервисов

в эксплуатации на LLM-платформе для инициатив с понятной полезностью или оцененным экономическим эффектом

> 20 млн руб./год

в среднем от проекта с использованием GenAI-сервисов с оцененным экономическим эффектом

РУСАЛ

СПАСИБО!

Казарин Иван Владимирович

Директор направления операционного развития искусственного интеллекта

@ikazx
Ivan.Kazarin@rusal.com