

ОПИСАНИЕ И УСЛОВИЯ ПРЕДОСТАВЛЕНИЯ УСЛУГИ
«EVOLUTION AI AGENTS»

1. ОБЩАЯ ИНФОРМАЦИЯ И ОПИСАНИЕ УСЛУГИ

- 1.1. Evolution AI Agents (далее – Услуга) – это облачное решение для разработки, развертывания и эксплуатации автономных AI-агентов и мультиагентных систем в единой среде. Услуга поддерживает полный цикл работы с агентами: от создания и запуска до мониторинга и масштабирования.
- 1.2. Услуга реализована на оборудовании, принадлежащем Исполнителю, средствами системы виртуализации собственной разработки (в т.ч. на базе компонентов с открытым исходным кодом). Ресурсом Услуги является виртуальные мощности, описанные в п. 2.1. Приложения, необходимые для функционирования AI-агента.
- 1.3. Функциональные возможности:
- развертывание автономных AI-агентов и мультиагентных систем;
 - развертывание агентов с использованием собственных или готовых Docker-образов из Evolution Artifact Registry (Приложение №1.EVO.6 к Договору);
 - интеграция с сервисом Evolution Foundation Models (Приложение №1.EVO.11.2 к Договору) для использования готовых Больших языковых моделей (БЯМ) с тарификацией по токенам;
 - интеграция с сервисом Evolution ML Inference (Приложение №1.EVO.11.1 к Договору) для запуска собственных ML-моделей в качестве основы для агентов;
 - автоматическое масштабирование модели в зависимости от нагрузки;
 - управление и контроль доступа к агентам, мультиагентным системам и **Model Context Protocol (MCP)**¹;
 - подключение к базам знаний через сервис Evolution Managed RAG (Приложение №1.EVO.20. к Договору) для повышения точности ответов
 - организация взаимодействия между агентами через протокол A2A (Agent-to-Agent) и подключение к внешним источникам данных через протокол MCP¹.
 - мониторинг и сбор статистики по потреблению ресурсов модели;
 - встроенная телеметрия и трассировка для мониторинга производительности и анализа принятия решений агентами;
 - управление конфигурацией агентов, MCP¹;
- 1.4. Услуга состоит из следующих компонентов:
- сервис-контроллер для управления Услугой из Личного кабинета Облачной Платформы;
 - компонент сбора данных мониторинга;
 - компонент, отвечающий за управление жизненным циклом продукта;
 - компонент, отвечающий за сбор тарификационных данных;
 - интеграция с услугами Cloud.ru:
 - Evolution Artifact Registry (Приложение №1.EVO.6 к Договору);
 - Evolution ML Inference (Приложение №1.EVO.11.1 к Договору);
 - Evolution Foundation Models (Приложение №1.EVO.11.2 к Договору);
 - Evolution Managed RAG (Приложение №1.EVO.20. к Договору);
 - Платформа Cloud.ru Evolution, обеспечивающая выбор Виртуальных машин для разворачивания программного обеспечения.
- 1.5. Обеспечение защиты инфраструктуры.
Защита Инфраструктуры облачной Платформы обеспечивается на следующих уровнях:
- на физическом уровне;
 - на сетевом уровне;
 - на инфраструктурном уровне;
 - обеспечение защиты от несанкционированного доступа к Виртуальным машинам, на которых функционирует сервис;
 - антивирусная защита Виртуальных машин, на которых функционирует сервис;
 - периодическая проверка на соответствие требованиям Информационной безопасности (в том числе с использованием сканеров безопасности) и обновление образов Виртуальных машин, используемых сервисом, и установленной на них Операционной системой;
 - мониторинг и реагирование на Инциденты Информационной безопасности, возникающие при функционировании сервиса;
 - межсетевое экранирование сетевых потоков сервиса средствами Платформы Cloud.ru Evolution.

¹ См. подробнее https://cloud.ru/docs/ai-agents/ug/topics/concepts_protocols-mcp

2. БАЗОВАЯ ФУНКЦИОНАЛЬНОСТЬ И МЕТРИКИ УСЛУГИ

2.1. Параметры Ресурсов Услуги:

Таблица 1. Параметры предоставляемых Ресурсов Услуги

Услуга	Тарифицируемый Ресурс	Метрики	Кратность
vCPU	Выделенный ресурс vCPU	Количество (шт.)	1
		Время работы (час)	1
	Динамический ресурс vCPU	Объём (ГБ)	1
		Время работы (час)	1
RAM	Выделенный ресурс RAM	Количество (шт.)	1
		Время работы (час)	1
	Динамический ресурс RAM	Количество (шт.)	1
		Время работы (час)	1

3. ТАРИФИКАЦИЯ УСЛУГИ

- 3.1. Для данной Услуги используется Динамическая тарификация (Pay as you go).
- 3.2. Динамическая тарификация предполагает оплату пула ресурсов (см. п. 2.1. Приложения) по факту их потребления Заказчиков в течение Отчетного периода.
- 3.3. Окончательная стоимость Услуги в Отчётном периоде формируется в соответствии с тарифами, установленными в Приложении №7.EVO.22. к Договору.
- 3.4. Тарифицируется потребление ресурсов зависимых сервисов, которые используются AI-агентами для своей работы. Объекты тарификации включают, но не ограничиваются:

- Использование моделей из Evolution Foundation Models: тарификация осуществляется в соответствии с условиями предоставления услуги Evolution Foundation Models (Приложение №1.EVO.11.2.), как правило, на основе количества обработанных токенов;
- Использование собственных моделей через Evolution ML Inference: тарификация осуществляется в соответствии с условиями предоставления услуги Evolution ML Inference (Приложение №1.EVO.11.1. к Договору). Объектами тарификации являются вычислительные ресурсы (vRAM GPU, Инстансы), хранение модели в кэше и количество запросов к модели;
- Хранение Docker-образов в Evolution Artifact Registry (Приложение №1.EVO.6. к Договору): тарификация осуществляется в соответствии с условиями предоставления соответствующей услуги;
- Использование других сервисов Платформы Cloud.ru Evolution, задействованных в работе агентов.

- 3.5. Величина ежемесячного платежа за пользование Услугой определяется в соответствии с фактическим потреблением Ресурсов. Доступные Ресурсы описаны в п. 2.1. Приложения, методика расчета - в примере ниже:

Итоговая стоимость =

vRAM ГБ * Цена 1 ГБ vRAM за выделенный ресурс * суммарное время жизни агентов
+ кол-во vCPU * Цена 1 vCPU за выделенный ресурс * суммарное время жизни агентов
+ vRAM ГБ * Цена 1 ГБ vRAM за динамический ресурс * суммарное время жизни динамического ресурса
+ кол-во vCPU * Цена 1 vCPU за динамический ресурс * суммарное время жизни динамического ресурса

Пример расчета работы агента:

- 1 час работал выделенный Ресурс (2 ГБ RAM, 1 vCPU)
- 0.5 часа работал динамический ресурс с теми же характеристиками, из-за того, что была высокая нагрузка на пользовательский сервис и была необходимость в масштабировании (масштабировался до двух агентов), первый агент - тарифицируется как выделенный ресурс, второй тарифицируется как динамический ресурс.

Вычислим цену за использование агента:

$(2 \times A_2 \times 1) = \text{vRAM, выделенный ресурс}$
 $(1 \times B_2 \times 1) = \text{vCPU, выделенный ресурс}$
 $(2 \times A_1 \times 0.5) = \text{vRAM, динамический ресурс}$
 $(1 \times B_1 \times 0.5) = \text{vCPU, динамический ресурс}$

Где:

- 2 ГБ — объём vRAM
- B_2 — цена 1 vCPU/час выделенного ресурса
- A_2 — цена 1 ГБ vRAM/час выделенного ресурса
- 1 ч — время работы выделенного ресурса
- A_1 — цена 1 ГБ vRAM/час динамического ресурса
- B_1 — цена 1 vCPU/час динамического ресурса
- 0.5 ч — время работы динамического ресурса

4. ИНЫЕ УСЛОВИЯ, ПРИМЕНИМЫЕ К УСЛУГЕ

- 4.1. Возможные виды подключения / изменения / отключения Услуги:
- посредством совершения действий в Личном кабинете.
- 4.2. Возможный порядок расчётов по Услуге:
- предоплата²;
 - постоплата³;
- 4.3. Возможные способы оплаты / порядок пополнения баланса:
- в безналичном порядке на основании выставленного Исполнителем счёта⁴;
 - оплата посредством электронных средств платежа⁵.
- 4.4. Требования к инфраструктуре Заказчика:
- наличие доступа в Интернет.

5. ОСОБЕННОСТИ УРОВНЯ ПРЕДОСТАВЛЕНИЯ УСЛУГИ

- 5.1. В соответствии с пп. 1.1.4. Приложения №2.0. к Договору вносятся следующие уточнения в уровень предоставления Услуги, действующий в отношении услуг Платформы Cloud.ru Evolution по умолчанию:
- 5.1.1. Показатель Доступности Услуги⁶ и Размеры компенсаций Заказчику за нарушение Доступности Услуг (п. 2.1.2. Приложения №2.EVO.0. к Договору) указаны в Таблице 2:

Таблица 2. Показатели доступности Услуги

Доступность Ресурса за Отчетный период	Размер компенсации от стоимости Ресурса за Отчетный период
≥ 99,7%	Компенсация не предоставляется
< 99,7%	10%
< 99,0 %	20%
< 95,0 %	30%

- 5.2. Для Услуги устанавливаются следующие особенности определения уровня Доступности:
- 5.2.1. Доступность рассчитывается отдельно для каждого Ресурса Услуги (п. 1.2. Приложения)
- 5.2.2. Недоступностью Услуги являются:
- Невозможность обращения в AI-агента или MCP-сервер в течении 5 (пяти) минут по причинам, зависящим от Cloud.ru, при условии, что сервис настроен согласно документации.
- 5.2.3. Компенсация выплачивается пропорционально объёму недоступных Ресурсов Услуги, т.е. Компенсация за нарушение целевых показателей Доступности Услуги рассчитывается согласно количеству недоступных Ресурсов.
- 5.3. Во всём остальном в части уровня предоставления Услуги применимы положения Приложения №2.EVO.0. к Договору.

² Является способом по умолчанию для Заказчиков-физических лиц

³ Является способом по умолчанию для Заказчиков-юридических лиц и ИП

⁴ Является способом по умолчанию для Заказчиков-юридических лиц и ИП

⁵ Является способом по умолчанию для Заказчиков-физических лиц

⁶ Целевые (гарантированное) значение Доступности