

ПРИЛОЖЕНИЕ №1.EVO.11.2
к Договору**ОПИСАНИЕ И УСЛОВИЯ ПРЕДОСТАВЛЕНИЯ УСЛУГИ**
«EVOLUTION FOUNDATION MODELS»**1. ПРЕДМЕТ СОГЛАШЕНИЯ**

- 1.1. Evolution Foundation Models (далее – Услуга) – это облачное решение, которое предоставляет возможность онлайн-обращения к ML-моделями (Большим языковым моделям (БЯМ)) посредством API и пользовательского интерфейса в ЛК.
- 1.2. Услуга реализована на оборудовании, принадлежащем Исполнителю, и средствами системы виртуализации собственной разработки (в т. ч. на базе компонентов с открытым исходным кодом). Ресурсом Услуги является ML-модель/БЯМ.
- 1.3. Функциональные возможности:
- Выполнение запросов к Большим языковым моделям для получения сгенерированного контента.
- 1.4. Услуга состоит из следующих компонентов:
- сервис-контроллер для управления Услугой из личного кабинета Облачной Платформы.
- 1.5. Для оказания Услуги Исполнитель использует БЯМ сторонних разработчиков с открытым исходным кодом, развернутые в Инфраструктуре Cloud.ru, и предоставляемые Заказчику по мере обращения к соответствующей БЯМ. Перечень и количество БЯМ может варьироваться в зависимости от изменений модели лицензирования БЯМ правообладателем (разработчиком); ограничений установленных действующим законодательством; выходом новых моделей с открытым исходным кодом.
- Функциональные возможности (например, ранкирование / ранжирование, преобразование одних типов данных в другие, генерация контента и пр., предоставляемые каждой конкретной БЯМ описываются в соответствующем разделе в ЛК.
- 1.6. При потреблении Услуги Заказчик использует Токены. Потребление входящих и генерируемых Токенов не является симметричным / синхронным (количество использованных входящих Токенов при обращении к БЯМ будет отличаться от генерируемого (ответа БЯМ Заказчику)) и отличается в зависимости от используемой Заказчиком БЯМ.
- 1.7. Услуга оказывается Заказчику с учётом следующих ограничений:
- 1.7.1. Ограничений, описанных в разделе 2 Приложения №3 к Договору;
- 1.7.2. Ограничений, установленных правообладателем (разработчиком) БЯМ в лицензионном соглашении¹;

2. БАЗОВАЯ ФУНКЦИОНАЛЬНОСТЬ И ПАРАМЕТРЫ РЕСУРСОВ УСЛУГИ

Услуга	Тарифицируемые Ресурсы	Метрика	Допустимые значения
Foundation Models Входные / Генерируемые запросы	БЯМ 32 Миллиарда параметров 1 000 000 входных Токенов	Шт	1 000 000
	БЯМ 32 Миллиарда параметров 1 000 000 генерируемых Токенов	Шт.	1 000 000
	БЯМ 70 Миллиарда параметров 1 000 000 входных Токенов	Шт.	1 000 000
	БЯМ 70 Миллиарда параметров 1 000 000 генерируемых Токенов	Шт	1 000 000
	БЯМ 24 Миллиарда параметров 1 000 000 входных Токенов	Шт	1 000 000
	БЯМ 24 Миллиарда параметров 1 000 000 генерируемых Токенов	Шт	1 000 000

¹Каждая конкретная БЯМ, используемая для оказания Услуги, имеет в своём описании в ЛК текст лицензионного соглашения правообладателя (разработчика), ограничения которого распространяются на Заказчика.

	БЯМ 7 Миллиарда параметров 1 000 000 входных Токенов	Шт	1 000 000
	БЯМ 7 Миллиарда параметров 1 000 000 генерируемых Токенов	Шт	1 000 000
	БЯМ 600 миллионов параметров 1 000 000 входных Токенов	Шт	1 000 000
	БЯМ 600 миллионов параметров 1 000 000 генерируемых Токенов	Шт	1 000 000
	БЯМ 200 миллионов параметров 1 000 000 входных Токенов	Шт	1 000 000
	БЯМ 200 миллионов параметров 1 000 000 генерируемых Токенов	Шт	1 000 000

3. ТАРИФИКАЦИЯ УСЛУГИ

- 3.1. Для данной Услуги используется Динамическая тарификация (Pay as you go). Заказчик начинает платить за потребляемые Токены.
- 3.2. Динамическая тарификация предполагает оплату 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.11.2.
- 3.4. Объекты тарификации:
- Количество входных Токенов
 - Количество генерируемых Токенов
- 3.5. Специфика тарификации Услуги: в связи с тем, что стоимость 1 Токена (как генерируемого, так и входящего) значительно меньше, чем применяемая по Договору валюта (рубли), для удобства восприятия Заказчика, цены указаны за 1 000 000 токенов.
- В каждом Отчётном периоде Заказчик оплачивает только фактическое потребленное количество Токенов.
- В этой связи округление потребления Токенов в рублевом эквиваленте осуществляется вверх (до «ближайшей копейки») в соответствии с примером расчёта, указанным ниже.
- 3.6. Пример расчёт

$$\text{Стоимость} = \text{Количество Токенов входных} * \text{Стоимость Токенов входных} + \text{Количество Токенов генерируемых} * \text{Стоимость Токенов генерируемых}$$

Далее полученное число округляется до «ближайшей копейки» вверх.

4. ИНЫЕ УСЛОВИЯ, ПРИМЕНИМЫЕ К УСЛУГЕ

- 4.1. Возможные виды подключения / изменения / отключения Услуги:
- 4.1.1. Посредством совершения действий в Личном Кабинете.
- 4.2. Возможный порядок расчётов по Услуге:
- Предоплата²;
 - Постоплата³;
- 4.3. Возможные способы оплаты / порядок пополнения баланса:
- 4.3.1. В безналичном порядке на основании выставленного Исполнителем счёта;
- 4.3.2. оплата посредством электронных средств платежа.
- 4.4. Требования к инфраструктуре Заказчика:
- 4.4.1. Наличие доступа в Интернет.
- 4.5. **Особенность доступности Услуги:** особенности уровня предоставления услуги
- 4.6. В соответствии с пп. 1.1.4. вносятся следующие уточнения в уровень предоставления Услуги, действующий в отношении услуг Evolution по умолчанию (Приложения №2.EVO.0. к Договору).

² Является способом по умолчанию для физических лиц, присоединившихся к условиям Договора/Оферте путём акцепта (п. 1.5. Договора/Оферты)

³ Является способом по умолчанию для юридических лиц

- 4.7. Для Услуги устанавливаются следующие особенности определения уровня Доступности:
- 4.7.1. Доступность конкретной БЯМ из существующих в принципе / тех, что были доступны Заказчику когда-то и не доступны (не размещены у Исполнителя) в конкретный момент – не относится к доступности Услуги (см. подробнее п. 1.4. и 1.5. Приложения).
К Доступности Услуги относится возможность обращения Заказчика к БЯМ, размещенным в ЛК.
- 4.7.2. Доступность рассчитывается отдельно для каждого Ресурса Услуги (п. 1.2. Приложения);
- 4.7.3. Недоступностью Услуги является ситуация, при которой ML-модель/БЯМ не принимает запросы/события и не дает ответа в течение 5 (пяти) и более минут по причинам, зависящим от Cloud.ru
- 4.7.4. Компенсация выплачивается пропорционально объёму недоступных Ресурсов Услуги, т.е. Компенсация за нарушение целевых показателей Доступности Услуги рассчитывается согласно количеству недоступных Ресурсов.
- 4.8. Во всём остальном в части уровня предоставления Услуги применимы положения Приложения №2.EVO.0. к Договору.