Настройка интеграции с языковыми моделями

Памятка Администратора

Версия 5.4 и выше

г. Самара, 2026

Оглавление

Введение

Взаимодействие с языковой моделью на данный момент доступно через:

1. Выражение «Вызов языковой модели» (спец.шаг «Вычисление выражения») для пользователей с ролью Администратор;
2. ИИ-ассистент;
3. Чат-бот.

Использование языковой модели позволяет разнообразить и упростить работу с макросами. Языковые модели могут сгенерировать описание, переписать текст, сделать суммаризацию, найти опечатки в тексте, сформатировать ответы в определенном формате и многое другое. Качество ответа нейросети напрямую зависит от точности переданной в API инструкции (системный промпт и т.п.).

Возможно использовать любую доступную языковую модель, одобренную службой информационной безопасности Вашей организации:

- локально развернутую модель;
- облачную модель по действующему контракту (Яндекс, Сбер и др);
- модель в рамках партнёрского предложения MWS GPT (тестовый период 2 месяца, лимит 5 млн токенов).

  Для использования необходимо предварительно подписать отдельное соглашение, после чего будут предоставлены токены и адреса подключения.

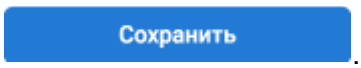
1. Сущность «Языковая модель»

Создание «Языковой модели» доступно в меню «Администрирование» – «Базовые настройки» – «Языковые модели».

1.1. Добавление языковой модели

Для добавления языковой модели необходимо выполнить следующие действия:

1. Открыть раздел меню «Администрирование» – «Базовые настройки» – «Языковые модели».
2. Нажать кнопку 

.
3. В открывшемся окне заполнить данные:
   - «Название» (произвольное удобочитаемое имя).
   - «Тип модели» – выбрать доступный тип из списка:
     - YandeGPT;
     - GigaChat (Сбер);
     - Llama.cpp-совместимая;
     - Vllm;
     - Ollama;
     - Groovy скрипт.
4. Указать параметры подключения для выбранного типа модели:
   - «Адрес API» (адрес сервиса или локального сервера модели);
   - «API-ключ» (или иной токен/учетные данные, предусмотренные провайдером модели).
5. Нажать 

.

После сохранения языковая модель станет доступна для выбора в выражении «Вызов языковой модели».

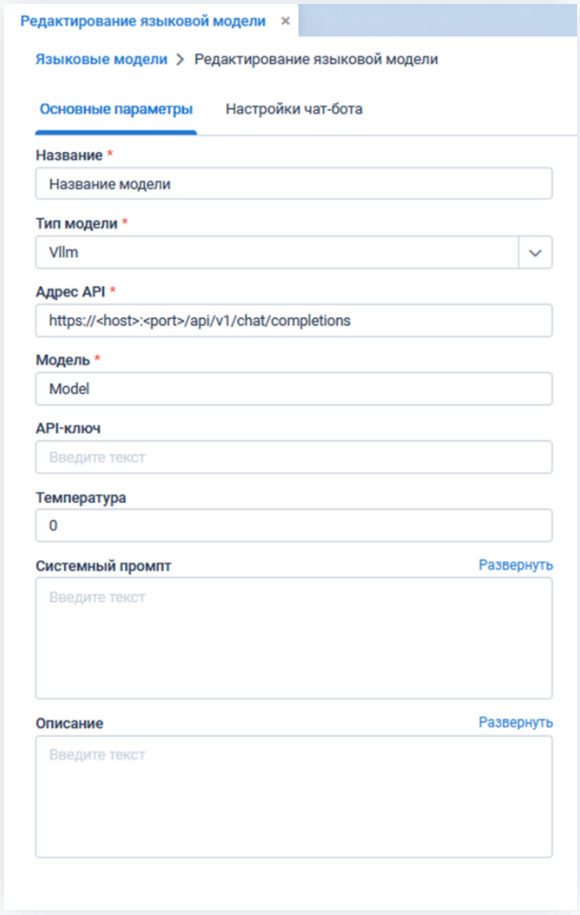


Рисунок 1. Пример редактирования языковой модели

1.2. Общие настройки языковой модели

На экране редактирования языковой модели доступны поля, представленные в таблице ниже.

Таблица 1

Поле	Описание	Примечание
Название*	Имя языковой модели	Указывается пользователем
Тип модели*	Выбирается один из доступных типов	YandexGPT,GigaChat (Сбер), Groovy-скрипт, Llama.cpp-совместимая, Vllm, Ollama

Адрес API*	URL для обращения к модели	Различается в зависимости от типа модели
Модель*	Указывается используемая модель	Отображается для типов моделей Llama.cpp-совместимая, Vllm, Ollama
URI модели YandexGPT	Уникальный идентификатор модели	Отображается при выборе типа модели «YandexGPT», идентификатор модели, которая будет использоваться для генерации ответа
Тип модели GigaChat (Сбер)	Определяет используемый тип модели	Автоматически проставляется значение по умолчанию
Groovy-скрипт	Поле отображается при выборе типа «Groovy-скрипт»	Используется для написания кастомных скриптов
Температура	Настройка вариативности ответов	Чем выше значение, тем более случайными будут ответы
Системный промпт	Текстовый запрос, задающий контекст работы модели	Поле необязательное, можно заполнить при создании выражения с типом «Вызов языковой модели» в соответствующем поле
Описание	Текстовое описание модели	Указывается пользователем

2. Настройки для конкретных моделей

Примечание:

За API-токеном/ключом необходимо обратиться напрямую к поставщику выбранной модели.

2.1. YandexGPT

Для работы с YandexGPT необходимо получить Адрес API, id каталога (указывается в поле «URI модели YandexGPT»: gpt://<идентификатор_каталога>/yandexgpt/<версия вызываемой модели>) и API-токен.

Для корректной работы языковых моделей YandexGPT необходимо прописать в файле «\tomcat\conf\app-core\local.app.properties» следующий параметр – «API-ключ» для YandexGPT: «thesis.macro.llm.YandexGPT.apiKey=<API токен>».

2.2. GigaChat (Сбер)

Для работы с GigaChat (Сбер) необходим персональный токен (можно получить его на сайте https://developers.sber.ru/portal/products/gigachat-api). После регистрации нужно сгенерировать токен и получить значение из поля «Авторизационные данные» (https://developers.sber.ru/docs/ru/gigachat/individuals-quickstart).

Для подключения к API будет необходимо также установить сертификаты Минцифры в cacerts (скачать их из https://www.gosuslugi.ru/crt и добавить их через keytool).

Для корректной работы языковых моделей GigaChat необходимо прописать в файле «\tomcat\conf\app-core\local.app.properties» следующий параметр:

- «thesis.macro.llm.GigaChat.scope=GIGACHAT_API_PERS» – scope для GigaChat, может быть GIGACHAT_API_PERS (физ. лица) или GIGACHAT_API_CORP (юр. лица);
- «thesis.macro.llm.GigaChat.authUrl=https://ngw.devices.sberbank.ru:9443/api/v2/oauth PERS» – url авторизации для GigaChat;

- «thesis.macro.llm.GigaChat.accessTokenExpirationTimeMs=#{25601000}» – время жизни токена GigaChat (25 минут по умолчанию);
- «thesis.macro.llm.GigaChat.credentials=<Авторизационные данные>» – API-ключ для GigaChat.

2.3. Llama.cpp

Для работы с языковыми моделями, запущенными через LLAMA.cpp, API-ключ и название модели заполнять необязательно, необходим только Адрес API (через kobold.cpp для модели адрес будет такой: http://localhost:5001/v1/chat/completions). Для тестирования необходим софт, который позволяет вызывать модель (https://github.com/LostRuins/koboldcpp), и модель (при тестировании использовалась mistral-7b-instruct-v0.2.Q8_0.gguf).

2.4. Ollama

Для работы с языковыми моделями, запущенными через Ollama, необходимо установить версию для соответствующей операционной системы (https://ollama.com/). После установки нужно выбрать модель для загрузки, скопировать строку для запуска модели и выполнить её в командной строке. Ollama загрузит и установит модель.

При создании языковой модели в меню «Администрирование» – «Базовые настройки» – «Языковые модели» необходимо указать Адрес API, по умолчанию для Ollama используется http://localhost:11434/api/generate, также необходимо указать выбранную модель в поле «Модель».

3. Вызов языковой модели

Выражение «Вызов языковой модели» позволяет использовать языковую модель в шагах макроса. При выборе этого типа выражения необходимо заполнить область «Модель» языковой моделью и область «Сообщение пользователя», она предназначена для отправки пользовательских сообщений к модели в контексте запроса из системного промпта.

Область «Системный промпт» необязательна и может быть заполнена как в языковой модели, так и непосредственно в выражении. Если промпт задан в языковой модели, то он будет дополняться к промпту выражения.

Промпт, заданный в пункте меню «Администрирование» – «Базовые настройки» – «Языковые модели», будет применяться ко всем запросам, которые эта модель будет отправлять (выражение, ИИ-ассистент, Чат-бот).

Область «Макс. токенов в ответе» заполнена по умолчанию значением «1000».

Примечание:

При редактировании параметра «Макс. токенов в ответе» необходимо учесть, что значение указывается в Мегабайтах.

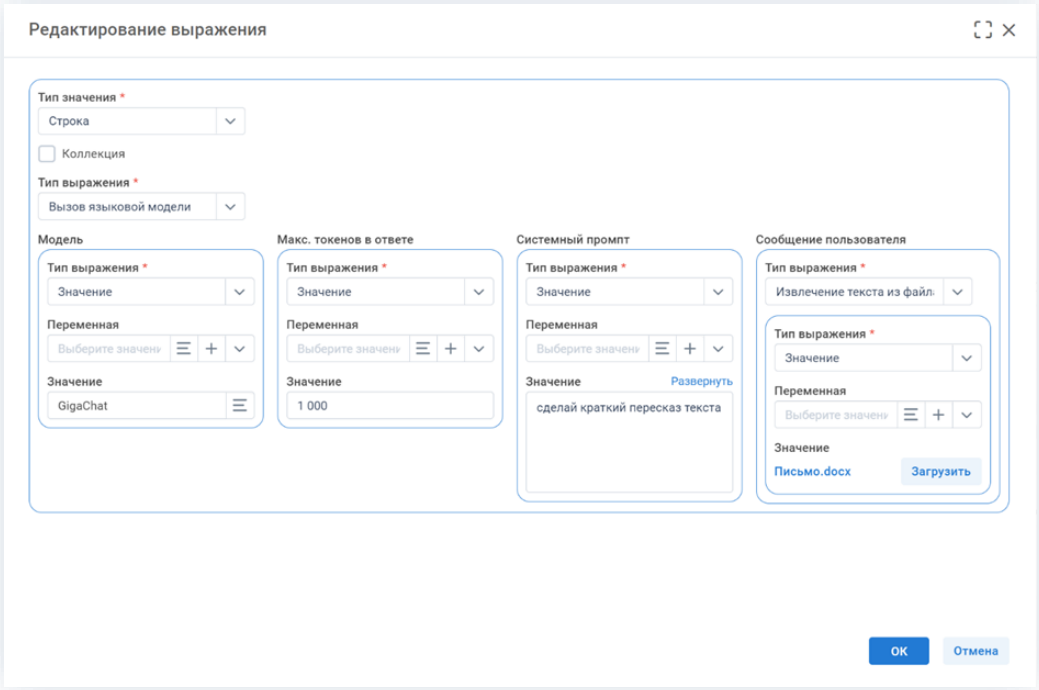


Рисунок 2. Пример редактирования выражения

Пример использования выражения:

Допустим, необходимо заполнить многотекстовое поле в карточке кратким содержанием документа.

Это можно сделать через тип выражения «Вызов языковой модели» со следующими параметрами с использованием суммаризации текста документа:

- указать нужную языковую модель;
- в поле «Макс. токенов в ответе» проставить необходимое значение или оставить значение по умолчанию;
- в области «Системный промпт» указать поведение Системы (или указать при создании языковой модели), в данном случае это краткий пересказ текста файла;
- в области «Сообщение пользователя» выбрать тип выражения «Извлечение текста из файла» и загрузить необходимый файл/выбрать переменную, если использовался перехват выгрузки файла.

Результат модели будет сохранен в переменную.