Описание программного обеспечения

1. Общие сведения

Полное наименование: Программный комплекс «ЦСТ.Локальный LLM-шлюз».

Сокращенное наименование: ПК «ЦСТ.Локальный LLM-шлюз».

Рабочее наименование версии: Local LLM.

Правообладатель: Общество с ограниченной ответственностью «Центр Суперкомпьютерных Технологий» (ООО «ЦСТ»), ИНН 5044124034, КПП 504401001, ОГРН 1225000001445. Юридический адрес: 141554, Московская область, г. Солнечногорск, д. Мелечкино, д. 50.

Версия веб-приложения: 0.2.0.

Версия шлюза моделей: 2.1.0.

2. Назначение

Программное обеспечение предназначено для организации диалогового доступа пользователей к большим языковым моделям, развернутым в локальной или контролируемой инфраструктуре организации.

ПО предоставляет веб-интерфейс с учетными записями и историей диалогов, подключается к серверу модели через OpenAI-совместимый программный интерфейс и, при использовании собственного шлюза, управляет очередями запросов и переключением нескольких моделей на общей вычислительной платформе.

3. Область применения

ПО может применяться в корпоративных, образовательных и исследовательских информационных системах, где требуется централизованный доступ к локально развернутым языковым моделям, хранение истории взаимодействия и разграничение прав пользователей.

Конкретная языковая модель и аппаратная платформа выбираются эксплуатирующей организацией. ПО не является нейросетевой моделью и не включает в себя веса модели.

4. Пользователи

В ПО предусмотрены две роли:

Пользователь работает с чатами, выбирает доступную модель, отправляет сообщения, получает потоковые ответы и управляет собственной историей.

Администратор выполняет функции пользователя, создает и блокирует учетные записи, меняет роли и пароли, просматривает статистику и журнал событий, а в управляемом режиме контролирует состояние шлюза и переключает модели.

Отдельно выделяется системный администратор ОС. Он устанавливает и обновляет ПО, настраивает сервер модели, резервное копирование, HTTPS, службы и средства наблюдения.

5. Реализованные функции

- аутентификация по имени пользователя и паролю;
- роли пользователя и администратора;
- создание, просмотр, переименование и удаление чатов;
- хранение истории диалогов в локальной базе SQLite;
- выбор языковой модели из списка, полученного от сервера или заданного настройками;
- передача контекста диалога на сервер модели;
- потоковое отображение ответа по мере генерации;
- отдельное отображение рассуждений модели, если сервер передает reasoning_content или теги <think>;
- остановка текущего ответа путем разрыва клиентского запроса;
- отображение времени до первого токена, длительности и скорости ответа при наличии необходимых данных;
- безопасное отображение ограниченного набора Markdown без выполнения кода модели в браузере;
- администрирование учетных записей;
- статистика обращений к моделям и журнал событий;
- резервное копирование и восстановление базы данных;
- проверка состояния приложения и подключенного сервера моделей;
- необязательное управление очередями, переключением и восстановлением серверов вывода через собственный шлюз;
- экспорт эксплуатационных метрик шлюза в формате Prometheus.

6. Состав продукта

Собственный исходный код включает:

- серверное веб-приложение app/;

- статический веб-интерфейс static/;

- шлюз моделей gateway/;

- инструменты резервного копирования, восстановления и проверки tools/;

- примеры конфигурации и служб deploy/;

- документацию docs/.

Не входят в состав собственного исходного кода:

- веса и конфигурации языковых моделей;

- Docker и контейнерные образы сред вывода;

- Caddy и иные обратные прокси;

- CPython и сторонние Python-пакеты;

- драйверы, библиотеки и утилиты производителей ускорителей;

- операционная система и systemd.

Указанные компоненты используются как внешняя среда и распространяются на условиях своих правообладателей.

7. Варианты работы

7.1. Прямой режим

Веб-приложение подключается непосредственно к серверу, реализующему необходимые методы OpenAI API. Панель аппаратной телеметрии и переключения моделей отключена. Этот режим не требует применения собственного шлюза.

7.2. Управляемый режим

Между веб-приложением и серверами вывода работает собственный шлюз. Он предоставляет единый OpenAI-совместимый адрес, хранит очереди запросов в оперативной памяти, запускает серверы моделей по декларативным профилям, переключает активную модель и восстанавливает сервер после сбоя.

Проверенный профиль управляемого режима использует четыре логических устройства STC/Teco AC700, контейнер tecovLLM 3.2.0 и модели Qwen3-8B и Qwen3-14B. Эта конфигурация является вариантом развертывания, а не обязательным свойством продукта.

8. Краткий текст для карточки реестра

Программный комплекс «ЦСТ.Локальный LLM-шлюз» предназначен для организации диалогового доступа пользователей к большим языковым моделям, развернутым в инфраструктуре организации. ПО предоставляет веб-интерфейс с учетными записями, ролями и историей диалогов, потоково отображает ответы моделей и поддерживает подключение к OpenAI-совместимым серверам. Необязательный шлюз управляет очередями запросов, запуском, переключением и восстановлением серверов моделей на общей вычислительной платформе.

9. Классификация

Основной класс: 05.13 Средства интеллектуальной обработки информации и интеллектуального анализа бизнес-процессов.

Дополнительный класс: 02.06 Серверное и связующее программное обеспечение.

Коды ОКПД2: 58.29.29, 62.01.29 (подлежат финальному подтверждению при заполнении формы).

Признак отношения к сфере искусственного интеллекта: Да, поскольку ПО предоставляет пользователям функции на основе языковых моделей.