SupportAgent – Система автоматической обработки обращений в поддержку

Модель для классификации и генерации ответов:

- **Модель**: qwen2:7b-instruct-q4_K_M;
- **Параметры**: 7B параметров (квантована до 4 бит)
- **Качество**: высокое качество генерации на русском языке, способность к пониманию контекста и выполнению инструкций
- **Память**: ~4GB RAM (благодаря квантованию)
- **Лицензия**: Apache 2.0 - разрешает коммерческое использование

Модель эмбеддингов для семантического поиска:

- **Модель**: cointegrated/rubert-tiny2
- **Параметры**: 29M параметров
- **Назначение**: Создание векторных представлений текстов на русском языке
- **Память**: ~150MB RAM

Модель детекции токсичности:

- **Модель**: sismetanin/rubert-toxic-detection
- **Параметры**: 178M параметров
- **Назначение**: Определение токсичности текста
- **Память**: ~450MB RAM

MCP-серверы и инструменты

Сервер обработки документов (Document Processor):

Функции:

- Извлечение текста из PDF документов
- Автоопределение типа PDF (текстовый/сканированный)
- Использование OCR для сканированных документов
- Пакетная обработка файлов

Параметры:

- `file_path`: путь к PDF файлу
- `use_ocr`: использование OCR (true/false)
- `force_method`: принудительный метод извлечения

Ограничения:

- Максимальный размер файла: 50MB
- Поддерживаемые форматы: PDF
- Максимальное количество страниц: 100

Сервер классификации обращений (Classification Server):

Функции:

- Классификация текста на 5 категорий
- Определение уверенности предсказания
- Проверка токсичности текста
- Пакетная обработка запросов

Параметры:

- `text`: текст обращения
- `include_toxicity`: включить проверку токсичности

Ограничения:

- Длина текста: до 8192 символов
- Поддерживаемые категории: login_issue, technical_issue, account_update, billing_issue, feature_request

Сервер семантического поиска (Embedding Search Server):

- **Функции**:
  - Добавление документов в векторную базу;
  - Семантический поиск похожих обращений;
  - Сохранение и загрузка эмбеддингов;
  - Расчет косинусного сходства;
- **Параметры**:
  - text: текст для добавления или поиска;
  - top_k: количество возвращаемых результатов;
  - min_similarity: минимальный порог сходства.
- **Ограничения**:
  - Максимальное количество документов: 10,000;
  - Максимальная длина текста: 512 токенов.

Модель данных

Обращение (SupportRequest):

python

```
{
   "id": "уникальный_идентификатор",
   "user_id": "идентификатор_пользователя",
   "text": "Текст обращения",
   "category": "login_issue",
   "priority": "high",
   "status": "new",
   "created_at": "2025-01-20T10:00:00Z",
   "toxicity_score": 0.05,
   "confidence": 0.92,
   "metadata": {
      "source": "web_form",
```

```
        "attachments": ["file1.pdf"]
    }
}
```


Действие (SupportAction):


```
{
    "id": "уникальный_идентификатор",
    "request_id": "идентификатор_обращения",
    "action_type": "classify",
    "agent_id": "system",
    "timestamp": "2025-01-20T10:01:00Z",
    "details": {
        "model_used": "qwen2:7b",
        "processing_time": 0.45
    }
}
```


Ответ (SupportResponse):


```
{
    "id": "уникальный_идентификатор",
    "request_id": "идентификатор_обращения",
    "agent_id": "ai_agent",
    "text": "Текст ответа",
    "response_time": 2.34,
    "satisfaction_score": 0.0,
    "is_ai_generated": True,
    "created_at": "2025-01-20T10:01:00Z"
}
```