28.05.2025 Планы

а, вот краткий и стратегический анализ текущих **планов по проекту Librarian AI** и **предлагаемые приоритеты** на ближайшие этапы:

**🔍 Обзор текущего состояния**

Вы реализовали:

* ✅ **Обработку файлов** (в том числе .zip/.rar/.tar)
* ✅ **Асинхронную очередь задач (Celery + Redis)** с контролем /status/{task\_id}
* ✅ **Потоковую загрузку больших файлов**
* ✅ **Генерацию мультиперспективных summary** с кастомными ролями и разными LLM (включая локальные модели)
* ✅ **LLMRouter** с поддержкой OpenRouter, GigaChat, Mistral, Dolphin
* ✅ **Архитектуру ядра с CLI, API и Telegram-адаптерами**

**🧭 Текущая фаза: Архитектурная стабилизация и расширение ИИ-ядра**

**📌 Краткие планы:**

| **Направление** | **Конкретные задачи** |
| --- | --- |
| ✅ **Универсальный LLMRouter** | ✔ GigaChat, ✔ Mistral, ✔ OpenRouter, ✔ Dolphin |
| 🧠 **Анализ и логические выводы** | Создание librarian\_ai.py, поддержка Chain-of-Thought |
| 📚 **База знаний** | Хранение чанков, графов, сущностей, feedback |
| 📈 **Панель мониторинга** | Статистика загрузок, LLM-ответов, кэширование |
| 🔐 **API-защита** | Внедрение API-ключей / JWT для безопасности |
| 🧪 **Тестирование** | Unit/Integration/Load тесты, мониторинг Celery/Redis |
| 🤖 **Игровые модели расследований** | Модели типа RedTeam vs BlueTeam, симуляции преступлений, выявление улик |
| ☁️ **Облачная стратегия** | Возможность запуска моделей в режиме "Low RAM", затем масштабирование с GPU |

**🧱 Архитектура моделей: приоритеты**

| **Категория** | **Цель** | **Приоритет** |
| --- | --- | --- |
| 🆓 Бесплатные LLM | GigaChat, OpenRouter, YandexGPT | ✅ Использовать сейчас |
| 🧠 Локальные LLM | Mistral 7B + дообучение на кейсах | 🚧 План на будущее с GPU |
| 🕵️ Табу-free модели | Dolphin, LM Studio | ✅ Уже подключено, расширить |
| 🧰 Fine-tuning | Обучение на спец. текстах (расследования) | 🚧 Позже, с финансированием |

**💼 Коммерческий потенциал**

| **Подход** | **Применение** | **Доход** |
| --- | --- | --- |
| 📦 SaaS API | Анализ документов, построение графов, детекция угроз | 💸 B2B (юристы, ИБ-компании) |
| 🧑‍💼 Виртуальные эксперты | Юрист, Психолог, Следователь | 💸 Платные консультации |
| 🎮 Образовательные симуляции | Расследование преступлений | 💡 Интерактивные курсы |
| 📖 Анализ книг, статей, архивов | Summary + знания | 💰 Подписка |

**🔧 Следующие шаги (по приоритету)**

1. **librarian\_ai.py** — логический Reasoner: выводы, гипотезы, действия (на основе графа и сущностей)
2. **Визуализация графа знаний** — через web/panel и CLI
3. **JWT-защита API** — разделение публичных и внутренних LLM-запросов
4. **Telegram-интерфейс** — быстрый анализ документов из телефона
5. **Интерактив CLI** — выбор модели, роли, режим анализа
6. **Кэширование LLM-ответов** (по prompt + hash)
7. **Расширение моделей (DeepSeek, Yi, LMStudio)**

**🧠 Заключение: перспектива**

Вы уже построили:

* Умную, модульную платформу
* С возможностью интеграции как SaaS, так и автономных приложений
* С гибкой архитектурой: локальные + облачные модели

👉 Ваш проект **готов к монетизации и масштабированию**, как только будет:

* Визуальный интерфейс (дешёвый MVP на Streamlit/FastAPI+Vue)
* Безопасность и стабильность (защита API, обработка ошибок)
* Первая версия Reasoner (логические выводы)

⚙️ Готов поддержать:

* Разработку librarian\_ai.py (Reasoning + выводы)
* Интеграцию новых LLM (DeepSeek, Yi, LMStudio)
* Подготовку MVP для инвесторов или клиентов