Audion Office OCR AI - руководство пользователя

Содержание

Audion Office OCR AI превращает офисные документы, PDF, изображения и сканы в независимый `DocumentModel`, а затем сразу формирует выбранные офисные, архивные и машинные результаты. Markdown остаётся одним из экспортов для аудита и LLM, но не является центром системы.

Запуск

```
launcher_gui.cmd
```

Обычная работа идёт через GUI. Проектные папки:

```
input\      исходные документы
output\     готовые пользовательские результаты
config\     настройки, промпты, API-ключи
docs\       документация проекта
docs\PDF\   PDF-версии документации
logs\       журналы запусков
report\     служебные отчёты
workspace\  временные рабочие материалы и проверки
cache\      OCR/preprocess cache
```

Для CLI/FZF используйте `launcher_project.cmd` или `launcher_project_ru.cmd`. Они запускают тот же manifest/service-слой, что и GUI: OCR Brick, Workbench resolver и офисные сборщики не имеют отдельной CLI-копии backend. Канонический Workbench использует названия `Источник`, `Добавить файл...`, `Назначение`, `Сбросить`, `Удалить`, `Список`.

`<имя>.document` — внутренний канонический пакет OCR. Он сохраняет исходник, изображения страниц, OCR-кандидаты, координаты, таблицы, порядок чтения, confidence, результаты сверки и ручные исправления. Полнота означает отсутствие потерь исходных данных, а не обещание абсолютной точности распознавания.

Главное меню

`Офисный текст без OCR`
Для DOCX/XLSX/PPTX/PDF/TXT/CSV/HTML, где текст уже можно извлечь без чтения пикселей.

`Локальный OCR`
Бесплатный OCR на этой машине. Основной рекомендуемый движок - Tesseract. Surya оставлен как медленный optional-режим, когда жалко денег на API и есть время.

`Платный API OCR`
OCR/vision-модели через Yandex, xAI, Mistral OCR 4, Gemini и OpenAI. Движок выбирается одной строкой компактных кнопок; выбранный движок определяет набор параметров ниже. Здесь же доступна управляемая проверка вторым проходом: для Mistral можно выбрать `Нет`, `Tesseract`, `Yandex` или `Yandex + Tesseract`.

В `Локальном OCR` и `Платном API OCR` сразу под движком находится общий блок `Готовые файлы`. Рабочий дефолт — только `DOCX`. Две строки прямоугольных checkbox-чипов образуют ровную сетку 4×2:

- офисные форматы: `DOCX`, `XLSX`, `Searchable PDF`, `ODT`;

- форматы аудита и разработки: `Markdown`, `OCR JSON`, `HTML`, `Проверка`.

Форматы выбираются независимо и могут формироваться одновременно. Сетка лежит на отдельном затемнённом фоне; повторная подпись `Выходные форматы` не показывается. Нижний поясняющий блок сохранён. Полный ZIP-архив поддерживается backend для совместимости, но не показывается в основном OCR-интерфейсе. Каждый отмеченный результат сразу собирается из сохранённого `DocumentModel` без повторного OCR.

В секции `Результат` кнопки контракта `Текст + боксы` и `Раскладка` делят всю строку на две равные колонки. Контракт определяет полноту данных OCR, а не конечный офисный формат.

В блоке `Обслуживание` есть подтверждаемая кнопка `Очистить DocumentModel`. Она удаляет внутренние пакеты `*.document`, `*.document.json` и `*.verification.json` из управляемых `output`, `report` и `workspace`, но сохраняет DOCX, XLSX, Markdown, PDF, изображения, исходные файлы и уже собранные full archive ZIP.

`Тест качества OCR`
Проверка одной страницы перед массовым прогоном. Можно сравнить raw/clean и локальные/API-движки.

`Проверка реквизитов`
Проверяет точные юридические и финансовые строки: номера контрактов, ИКЗ, даты, суммы и похожие поля. Инструмент не переписывает OCR самовольно, а пишет отчёт match/mismatch/uncertain.

`Таблицы`
Три операции раскрыты прямо в одном окне без вложенной навигации: извлечение координатных строк в XLSX, проверка Markdown-таблиц и сборка XLSX. Для линованных сканов основной OCR-backend дополнительно восстанавливает физическую сетку, проверяет число столбцов и сравнивает провайдеров внутри подтверждённых ячеек.

`DEV Markdown PDF`
Сборка dark/light PDF из Markdown через Chromium. По умолчанию выбран `docs\PDF`. Режим `PDF рядом с MD` создаёт одну подпапку `PDF` в каждой папке с Markdown, а `Зеркало в Назначение` формирует отдельное дерево через текущую папку Workbench.

`Инструменты проекта`
Установка и проверка окружения: Tesseract, Real-ESRGAN, Surya/llama.cpp, модели API, Gemini Batch, smoke-тесты и статус проекта. Здесь же находится экспертная `Пересборка старого Markdown` в DOCX/PDF/PPTX/XLSX для архивного compatibility-процесса; это не основной OCR-

путь. Здесь и в `Локальный OCR` показывается компактный аппаратный бейдж с рекомендацией по GPU/CPU-пути.

OCR

Локальный OCR

Выбирайте `Tesseract`, если нужен быстрый бесплатный OCR, координаты слов, проверка реквизитов или searchable-PDF/text-layer workflow.

Выбирайте `Surya`, если нужен аккуратный локальный разбор сложной страницы и время не критично. Surya optional и может быть тяжёлым по зависимостям.

Платный API OCR

`Yandex` полезен для русского OCR с координатами и режимами `page`, `table`, `markdown`, `handwritten` и другими.

Для буквального OCR первым пробуйте xAI `4.20 fast` (`grok-4.20-non-reasoning-latest`). Режим `4.20 quality` (`grok-4.20-reasoning-latest`) медленнее и предназначен для сложных таблиц, фрагментированных и тяжёлых сканов. xAI вызывается обычным non-streaming `chat/completions`: это исключает зависимость от нестабильного streaming-маршрута. На временных HTTP/сетевых сбоях backend выполняет до четырёх попыток с задержкой, строго проверяет UTF-8 и сохраняет успешные страницы в возобновляемом кэше. Чекбокс `Проверка русского OCR` по умолчанию выключен; при включении reasoning-модель перепроверяет только подозрительные страницы, а слишком сильные правки отклоняются и исходный вариант остаётся в метаданных.

`Mistral OCR 4` сохраняет собственную структуру страницы и табличные attachments. Для тяжёлых русских таблиц включайте второй проход `Yandex`, `Tesseract` или их комбинацию: backend сверяет текст внутри подтверждённой геометрии ячеек, а неоднозначные расхождения оставляет для проверки.

`Gemini` и `OpenAI` доступны как дополнительные vision OCR варианты и для статистики качества. Для Gemini можно выбрать тариф `Standard` или `Flex`: Flex дешевле, но может дольше ждать свободную мощность. Для больших несрочных прогонов используйте `Gemini Batch OCR` в инструментах проекта: сначала собирается JSONL/manifest, затем job отправляется в Google отдельным переключателем, а готовый результат забирается через `Забрать Gemini Batch`.

Подробности по режимам Gemini и проверенному dry-run: `docs\GEMINI_BATCH_FLEX.md`.

Gemini Standard, Flex и Batch

`Standard` - обычный интерактивный запрос. Это безопасный выбор для одной страницы, проверки качества и ручных экспериментов.

`Flex` - интерактивный запрос с `service_tier=flex`. Он дешевле, но может ждать свободную мощность заметно дольше. В GUI это поле `Gemini tier`; для Flex таймаут увеличен.

`Batch` - асинхронный режим для больших несрочных прогонов. `Gemini Batch OCR` по умолчанию только готовит `workspace\gemini_batch\...\requests.jsonl` и `manifest.json`; отправка в Google включается отдельным переключателем `Отправить в Google`. `Забрать Gemini Batch` проверяет job и скачивает Markdown.

Проверено на `input\Вопрос 239.pdf`, страницы `6-8`, `Raw`, `DPI 300`, `gemini-3.5-flash`, без отправки в Google: создано 3 JSONL-запроса с ключами `239_page_0006`, `239_page_0007`, `239_page_0008`.

Очистка скана

В OCR-окнах секция `Препроцессинг` содержит одну строку профилей. Повторная подпись `Очистка скана` скрыта, а нижнее пояснение сохранено:

- `Авто` - рекомендуемый режим; настройки берутся под выбранный движок.
- `Без очистки` - отправить страницу как есть.
- `Тяжёлый скан` - сильнее убрать JPEG-шум, поднять контраст и резкость.
- `Цифры` - профиль для номеров, дат, сумм и реквизитов.
- `Ручной` - открыть старые подробные настройки в Advanced.

Ключи API

Ключи хранятся в `config\`. Добавление и удаление ключей делается из GUI рядом с полем провайдера. Удаление ключа должно подтверждаться предупреждением. Не кладите реальные ключи в публичные архивы.

Перед массовым прогоном

Сначала запускайте `Тест качества OCR` на одной странице. Для юридических документов дополнительно запускайте `Проверка реквизитов`. Затем выберите движок, профиль очистки и нужные готовые файлы. После OCR проверяйте DOCX/XLSX и отчёты в `report`; Markdown нужен только тогда, когда он выбран как отдельный аудит-формат.