EGE-Checker — Дорожная карта

23 марта — 24 мая 2025 · Разработка скилла для OpenClaw (Qwen3 / GPT-4o)

Цветовые обозначения фаз

Ф1 Базовый прототип	Ф2 Распознавание	Ф3 Стабильность оценки	Ф4 Интеграция

Хронология по неделям

23.03	30.03	06.04	13.04	20.04	27.04	04.05	11.05	18.05
Прототип	Прототип	Распознавание	Распознавание	Стабильность	Стабильность	Стабильность	Интеграция	Интеграция

Ф1 Базовый прототип

23 марта — 5 апреля

Задача	Неделя	Детали
Скилл: OCR + оценка К1–К12 (русский)	Нед. 1	*SKILL.md с инструкцией: читать рукопись через vision + исходный текст задания, загружать критерии, последовательно проходить К1–К12, выдавать баллы и комментарии.*
Скилл: аудирование (английский)	Нед. 1–2	*Агент получает аудиофайл + ключи (текст или фото). Транскрибирует аудио, извлекает ответы ученика, сверяет с ключами по блокам 1/2-9/10-18, выводит таблицу баллов.*
Сборка тестового набора данных	Нед. 2	*10–15 сочинений с известными баллами от экспертов (ФИПИ, YouTube-разборы). 5–10 аудиозаписей с ключами. Без этого набора нечем измерять прогресс.*

M1	5 апреля	Прототип работает на тестовом наборе. Есть первые цифры расхождения с эталоном.

Ф2 Улучшение распознавания

6 апреля — 19 апреля

Задача	Неделя	Детали
Итерации промпта OCR для рукописи	Нед. 3	*Тест OCR на 15 бланках разного качества: хороший/плохой почерк, поворот, тень. Метрика: % верно распознанных слов. Добавить обработку зачёркнутых мест и маркер [?] для неразборчивых фрагментов.*
Качество распознавания аудио-ответов	Нед. 3	*Тест на 10 записях: тихий голос, шум, быстрая речь. Ключевая проблема: "two/too/2" — решается prompt hint с описанием формата ответов. Метрика: % верно извлечённых ответов по заданиям.*
Локальный STT как fallback	Нед. 4	*Если нет внешнего Whisper — поднять faster-whisper локально. Совместим с тем же API форматом. Сравнить качество и latency с облачным вариантом на том же тестовом наборе.*

М2	19 апреля	OCR точность >90% на чистых бланках. Аудио: верно извлекаются ответы в >85% случаев.

Ф3 Стабильность оценки — 20 апреля — 10 мая

Задача	Неделя	Детали
Измерение дисперсии: 10 прогонов одной работы	Нед. 5	*Прогнать каждое из 10 тестовых сочинений 10 раз. Записать баллы по каждому критерию. Вычислить std. Цель: std < 0.3 по каждому К. Найти самые нестабильные критерии (обычно К2, К5, К6).*
temperature=0 + структурированный вывод	Нед. 5–6	*Установить temperature=0 в запросах к LLM. Перейти на JSON Schema если модель поддерживает. Добавить chain-of-thought в промпт: "сначала выпиши цитаты, затем примени критерий, затем выставь балл".*
Few-shot примеры в критериях К2, К5, К6	Нед. 6	*К2 — самый субъективный (0–6 баллов). Добавить в russian-essay-criteria.md: "вот фрагмент с пояснением → К2=4", "вот пересказ без анализа → К2=0". Аналогично для К5 и К6. Перемерить std после добавления примеров.*
Калибровка по эталонным работам	Нед. 7	*5 работ с официальными баллами экспертов ФИПИ. Найти систематические расхождения (например, модель всегда завышает К4). Добавить корректирующие инструкции в SKILL.md для каждого проблемного критерия.*

М3	10 мая	10 прогонов одной работы → итого расходится не более ±1 балла. Дисперсия по К2 < 0.5.

Ф4 Интеграция и финализация — 11 мая — 24 мая

Задача	Неделя	Детали
Упаковка и документация скилла	Нед. 8	*Финальная версия SKILL.md с обновлёнными промптами и few-shot примерами в references/. Добавить раздел "версия критериев" (год ФИПИ). Написать README. Упаковать .skill файл.*
Интеграция с OpenClaw / решение CORS	Нед. 8	*Выяснить формат скиллов в OpenClaw и адаптировать. Решить CORS для запросов к llm.lambda.coredump.ru: nginx reverse proxy с нужными заголовками или серверный вызов изнутри OpenClaw.*
Финальное тестирование на новых работах	Нед. 9	*5 новых сочинений + 3 новые аудиозаписи (не использовавшиеся в разработке). Сравнить с эталоном. Если расхождение > 2 баллов — доработать. Цель: ±2 балла от эксперта на 80% работ.*

М4	24 мая	Скилл задеплоен в OpenClaw. Точность ±2 балла от эксперта. Стабильность 10/10 прогонов.

Ключевые технические решения

- Русский модуль: проверяющий присылает агенту сканы бланков + исходный текст задания + тему. Без исходного текста невозможно проверить К1, К2, К3, К12.
- Английский модуль: проверяющий присылает агенту аудиофайл + ключи (текстом или фото). Агент сам транскрибирует и сверяет — никакого отдельного HTML-инструмента.
- Стабильность достигается через temperature=0, chain-of-thought промпт и few-shot примеры в файлах критериев. Скилл не привязан к конкретной модели — работает с Qwen3, GPT-4o и другими multimodal моделями.
- Главный риск: субъективность К2 (комментарий, 0–6 баллов) — требует наибольшего внимания в фазе 3.

Дата создания: 24.03.2026