1. VLM

1.1 Почему я выбрала эту модель.

В ходе сравнительной оценки трёх моделей (Llava, Gemma3 и Qwen) было установлено, что именно Qwen демонстрирует лучшие показатели по всем ключевым метрикам (см. таблицу 1). По результатам экспериментов Qwen показала F1 = 0.888, что является самым высоким значением среди всех протестированных моделей, а также Excess = 0.027, что отражает минимальное количество лишних распознаваний. Кроме того, время работы Qwen составило около одной минуты, что быстрее, чем у Llava и Gemma3, где выполнение занимало от полутора до двух минут. Эти данные подтверждают, что Qwen обеспечивает более надёжное и точное выполнение задачи при меньших затратах времени, и именно поэтому она была выбрана в качестве основной модели.

Таблица 1 – Сравнение альтернатив VLM

Модель	F1	Excess	Время работы
Llava	0.78	0.108	1.5–2 мин
Gemma3	0.611	0.33	1.5–2 мин
Qwen	0.888	0.027	~1 мин

1.2 Trade-offs: качество vs latency

Качество: Qwen обеспечивает наилучший баланс – высокая точность (F1) и минимальный шум (Excess). Llava и Gemma3 уступают по обоим показателям.

Latency: Локальные модели зависят от мощности компьютера. Время работы Llava и Gemma3 составило 1.5–2 минуты, тогда как Qwen быстрее (~1 мин).

Таким образом, Qwen одновременно превосходит альтернативы по качеству и по времени отклика, что делает её наиболее подходящей моделью для использования в локальном окружении.

2. LLM

2.1 Почему я выбрала эту модель

В отчёте для LLM-задач использовалась модель Mistral API. Она была выбрана, потому что обеспечивает стабильное качество генерации текста, низкую задержку и регулярные обновления. В отличие от локальных моделей, запуск через API снимает зависимость от мощности оборудования и гарантирует предсказуемую производительность.

Таблица 2 – Сравнение альтернатив LLM

Модель	Соответствие метрикам	Время работы	Стоимость
Gemma3	70%	~1 мин	–
Mistral API	85%	10-20 сек	~1500 токенов

В качестве метрик использовалось соответствие предпочитаемым времени готовки, сложности, калорийности и соответствие диете.

2.2 Trade-offs: качество vs стоимость vs latency

Качество: Mistral API обеспечивает высокую точность и связность ответов, что подтверждается результатами отчёта.

Стоимость: использование API имеет фиксированную тарификацию за токены (в моем случае примерно 1500 токенов за одну генерацию). Это делает расходы предсказуемыми, но дороже, чем локальный запуск.

Latency: в проведенных экспериментах Mistral API показала низкое время отклика – в среднем около 10 секунд, что значительно быстрее, чем локальные модели.