

Mask Gate — качество детекции: замер, артефакт, ограничения

Mask Gate · trust pack · версия 1.1 · дата 25.10.2026 · проект документа, публикация — после подтверждения владельца

Версия: 1.1 · **Дата:** 25.10.2026 · **Статус:** проект документа, публикация — после подтверждения владельца.
Документ: trust pack, файл 4 из 4. Соседние: «Архитектура безопасности», «Обработка ПДн и 152-ФЗ», «Поручение и субпроцессоры».

Что это. Формальная фиксация того, **как** измеряется качество детектора, **какие числа** получились в последнем замере и **чего эти числа не значат**. Источник чисел — артефакт замера `api/quality_report.json` (только чтение); публичная витрина — страница `/quality`. Числа в этом документе не переписаны «на глаз», а перенесены из артефакта с указанием даты и команды замера.

1. Артефакт и происхождение чисел

Поле	Значение
Артефакт	<code>api/quality_report.json</code> (в репозитории, не редактируется вручную)
Дата замера (<code>generated_at</code>)	<code>2026-10-25</code>
Схема артефакта	<code>schema_version: 3</code>
Команда воспроизведения	<code>python scripts/quality_report.py</code>
Источник замера	<code>python -m eval.run_eval</code>
Чистота дерева при замере	<code>source_tree_clean: true</code> (замер не делался на грязном дереве)
Отпечаток чисел	<code>reproducibility.eval_sha256:</code> <code>6f5bee933501ee215492bfb75dd3100fb865c7756940a042e094733e6356b0c6</code>
Дайджест корпуса	<code>e8cb9e4b30ea...</code> (sha256 корпуса — в артефакте, блок <code>digests</code>)

Артефакт также хранит sha256 **по каждому файлу**, влияющему на детекцию (`digests.detection_files`) — поэтому подмена корпуса или кода детектора без пересборки артефакта видна: `python scripts/quality_report.py --check-digests` сверяет дайджесты за секунды (без прогона эвала).

2. Корпус

Параметр	Значение
Файл	<code>eval/corpus/corpus.jsonl</code>
Версия корпуса	<code>4.5</code>
Записей	<code>1 723</code>
— positive	<code>1 480</code>

Параметр	Значение
— negative	177
— adversarial	52
— ner	14
Размеченных значений (labels_structured)	1 615

Корпус **синтетический**: записи собирает генератор корпуса, значения сделаны так, чтобы проходить наши же валидаторы. В нём есть adversarial-блок — 52 записи с обфускацией и кодированием (гомogliфы, percent-encoding, разбиение и т.п.). Документы клиентов в замер **не входят**.

3. Результат замера

Метрика	Значение артефакта	Как считается
TP (верные находки)	1 612	пересечение отрезков с разметкой, по категориям
FP (ложные)	1	находка есть, разметки нет
FN (пропуски)	3	разметка есть, находки нет
Recall	0,998142... (в публичной витрине: 99,81 % , проценты усечены вниз)	$TP / (TP + FN) = 1\,612 / 1\,615$
Precision	0,999380... (в публичной витрине: 99,93 % , усечено вниз)	$TP / (TP + FP) = 1\,612 / 1\,613$

Остаточный риск в терминах корпуса: **3 пропуска на 1 615 размеченных значений $\approx 0,19 \%$; 1 ложная находка на 1 613 находок $\approx 0,06 \%$** . Это характеристика **нашего корпуса**, а не гарантия для произвольного текста (см. раздел 6).

3.1. По уровням (tier) категорий

Tier	Категорий	TP	FP	FN
1 (критичные)	7	288	0	0
2	9	290	0	0
3	10	342	1	3
4	21	692	0	0
Итого	47	1 612	1	3

3.2. Пороги гейта в CI (не «красивые числа», а условие сборки)

Проверка	Порог в артефакте
Recall по всем категориям	$\geq 0,85$
Recall tier 1	$\geq 0,97$
Precision tier 1	$\geq 0,85$
Recall tier 2	$\geq 0,95$

Проверка	Порог в артефакте
Recall tier 3	≥ 0,75
Precision по всем категориям	≥ 0,80
NER: recall / precision	≥ 0,75 / ≥ 0,70

Если замер опускается ниже порога — CI краснеет; «подкрутить» числа в артефакте вручную нельзя без `--check-digests` (см. 1).

4. Как замер устроен технически

Параметр движка в замере	Значение
Политика	<code>policy.yaml</code> , версия политики 2
NER (Natasha/имена)	выключен (<code>enable_ner: false</code>)
Порог score	0,5
Порог энтропии	4,0

Метод: каждое размеченное значение корпуса сверяется с находками детектора **по пересечению отрезков текста**, отдельно по каждой категории; из TP/FP/FN считаются recall и precision. Проценты на публичной странице **усечены вниз** до двух знаков, а не округлены — округление вверх завышало бы публичное число; рядом с процентами показаны счётчики TP/FP/FN.

5. Чего эти числа не значат

- **Это не гарантия 100 % обнаружения.** 3 пропуска в корпусе — это факт; на незнакомых форматах, языках, вёрстке и качестве сканов пропусков может быть больше. Остаточный риск лежит на пользователе и закрывается процессом (fail-closed, отказ от обработки неопределённого, проверка результата).
- **Корпус синтетический.** Ничьи реальные документы в замере не участвуют; значения сгенерированы.
- **Есть adversarial-блок** — 52 записи с обфускацией: он показывает устойчивость к известным приёмам обхода, а не ко всем возможным.
- **NER в этом замере выключен.** Записи `ner` присутствуют в корпусе, но движок в замере не использует NER-модель; распознавание имён через NER в эти числа не входит.
- **Латентность не публикуется.** В артефакте `latency.published: false` с причиной: воспроизводимого замера латентности в репозитории нет, а число зависело бы от железа.
- **OCR не публикуется.** `ocr.published: false`: OCR-путь (скан → распознавание → детекция) меряется отдельным мини-корпусом сканов и требует движка tesseract.
- **Нет сравнения с конкурентами.** Мы не утверждаем «лучше/хуже» других инструментов: сравнительного замера на общем корпусе не проводилось.
- **Число категорий:** в артефакте измерено **47** категорий; в `config/policy.yaml` на 25.10.2026 —

48. Расхождение — из-за методики: замер считает категории, представленные в размеченном корпусе; одна категория политики без размеченных значений в замер не попала. Это ограничение замера, а не «47 vs 48»

в тексте: перечень категорий для клиента — на `/kakie-dannye-nahodim`.

6. Как проверить

Действие	Команда
Сверить дайджесты корпуса и детектирующих файлов (без прогона)	<code>python scripts/quality_report.py --check-digests</code>
Пересобрать артефакт после правок	<code>python scripts/quality_report.py</code>
Прогнать замер	<code>python -m eval.run_eval</code>
Посмотреть публичную витрину	страница <code>/quality</code> (числа читаются из артефакта, не из текста страницы)
Проверить отпечаток чисел	<code>reproducibility.eval_sha256</code> — sha256 канонического JSON чисел секции <code>eval</code> ; метод и команда описаны в артефакте

Замер от **25.10.2026**; при изменении политики, детекторов или корпуса артефакт пересобирается, и числа в этом документе должны быть обновлены вместе с версией документа.

7. Что не заявляется

- нет утверждения «детектор находит 99,9 % ПДн в любых документах»;
- нет утверждения «ложных срабатываний нет» (1 FP в замере — есть);
- нет утверждения «сервис заменяет DLP-систему» — не заменяет;
- нет сертификата или независимого аудита качества — только воспроизводимый замер на открытом артефакте.

Ошибки детектора можно сообщить через форму `/report` (она собирает только структурные признаки, без свободного текста) или на `security@maskgate.ru`.

Документ не является юридическим заключением. Сертификатов, независимого аудита и SLA у сервиса нет — это прямо указано в тексте.
Вопросы: `security@maskgate.ru`