

Памятка по оценке model-visible output в coding agents

Короткий чек-лист для команд, которые внедряют или оценивают агентные инструменты разработки. Его можно вынести в <downloads/codex-model-visible-output-checklist.pdf>.

Перед внедрением hooks

- Разделить raw tool result, model-visible output и audit/transcript events.
- Проверить, какие инструменты возвращают длинные или шумные результаты.
- Определить, какие hooks могут менять вывод, видимый модели.
- Зафиксировать порядок применения hooks и правило для конфликтов.
- Не считать filtering-for-model полноценной redaction-системой.

При работе с логами и тестами

- Сохранять путь к полному артефакту или исходному логу.
- Передавать модели статус, ключевые ошибки и affected files.
- Не удалять диагностические строки без понятного правила.
- Сравнивать исходный typed result и подготовленный model-visible output при отладке.
- Проверять, не ухудшилось ли качество следующих шагов агента после сжатия вывода.

Для безопасности и аудита

- Проверить, где сохраняется сырой output помимо контекста модели.
- Описать доступ к UI, transcript, events и внутренним логам.
- Логировать, какой hook изменил output и почему.
- Не обещать отсутствие утечки, если контролируется только model-visible слой.
- Проверять конфликты между пользовательскими hooks и организационными политиками.