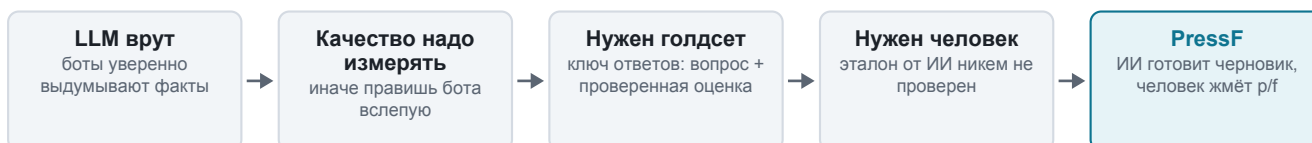


PressF

Проверяем, говорит ли ваш бот правду: ИИ-судья сверяет ответы с вашей документацией — вы подтверждаете одной кнопкой.

Зачем это существует

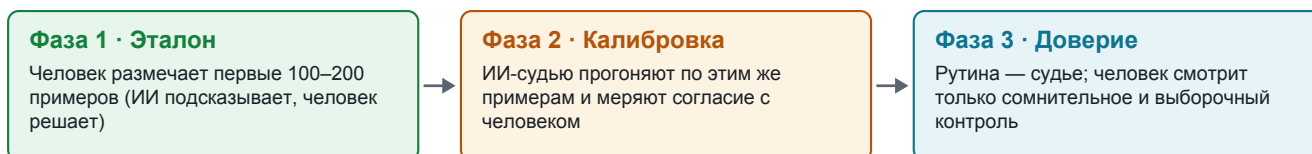
Компании подключают LLM к своим документам (RAG-боты, агенты), а LLM генерирует правдоподобный текст, а не правду: если ответа в документах нет, бот уверенно его выдумает. Классика — бот Air Canada изобрёл правило возврата билетов, и суд обязал компанию его исполнить. Отсюда цепочка, на которой стоит продукт:



Кому нужно: любой команде, у которой LLM отвечает людям по документам — боты поддержки, внутренние базы знаний, юридические и медицинские ассистенты. Чем дороже цена вранья, тем нужнее.

Методология: три фазы доверия

Голдсет — это ключ ответов к экзамену: набор «вопрос → ответ бота → оценка», где каждой оценке можно верить, потому что её подтвердил человек. Путь к автоматике всегда проходит три фазы:



Ключевой вопрос фазы 2: «судья согласился с человеком в N%». 95% — рутину можно отдавать судье. 70% — судью чинить (модель посильнее, гайдлайны чётче). PressF делает фазу 1 быстрой (судья готовит черновик и цитаты — человек только решает: 10–20 секунд на пример вместо 2–4 минут), фазу 2 — автоматической (отчёт согласия), и помогает жить в фазе 3 (сомнительное — первым).

Схема приложения

Пять экранов, один линейный путь. Никаких терминалов, конфигов и слова «проект». Правило каждого экрана: сначала «что это» и «зачем», потом одна большая кнопка.

Постоянный элемент: карта станций

С момента старта и до конца сверху окна висит карта всего пути. Пройденные станции зелёные, текущая подсвечена. Человек всегда знает, где он и сколько осталось.

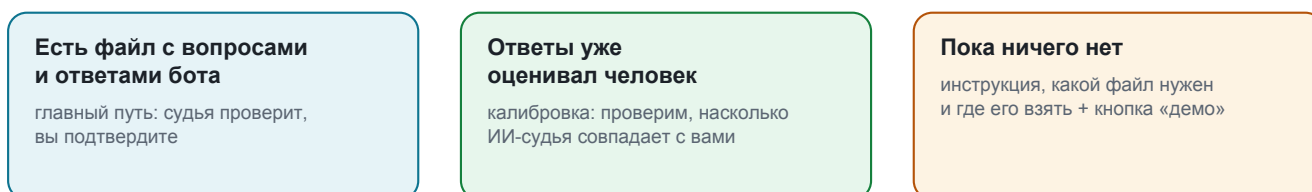


Экран 0 · Дом — «Мои проверки»

Одна фраза, объясняющая всё приложение (см. подзаголовок этого документа), кнопки «Попробовать на примере» (готовое демо — сразу к карточкам) и «Новая проверка», ниже — список прошлых проверок: «Бот поддержки · 3 мая · 100 ответов · 24 плохих · продолжить». Клик — возвращаешься туда, где остановился.

Экран 1 · «Что у вас есть?» — вход в новую проверку

Один вопрос и три карточки. Выбор карточки — это выбор сценария, но человек об этом не знает:



1 · Станция «Данные»

Две зоны загрузки: «файл с вопросами и ответами» и «документация, по которой бот должен отвечать» (папка с файлами). Колонки определяются автоматически; если неясно — человек тыкает в заголовок колонки, а не заполняет форму. Нашлась колонка с оценками pass/fail — приложение само предложит импортировать их и проверить судью (сценарий калибровки).

2 · Станция «Проверка судьёй»

Одна большая кнопка: «Проверить 100 ответов — примерно \$0.40». Цена всегда до запуска, запуск всегда с подтверждением. Выше — карточка «что сейчас произойдёт» в три строки. Во время работы — живой прогресс: «Проверено 34 из 100 · 8 выглядят враньём». Нет API-ключа — дружелюбный экран «вставьте ключ, вот где взять, хранится только на вашем компьютере».

3 · Станция «Твоё слово» — главный экран приложения

Одна карточка на весь экран. Мнение судьи — человеческой фразой, доказательства — дословными цитатами с именем файла-источника. При первом входе обучающая плашка: «Вы — учитель, судья — ваш ассистент. Он уже всё проверил, вы подтверждаете или отменяете».

Карточка 12 из 100 · согласие с судьёй 92%

Вопрос: Какой лимит запросов на базовом тарифе?

Ответ бота: «Лимит составляет 1000 запросов в час, при превышении вернётся ошибка 429.»

Судья: противоречит документации (уверенность 0.99)
бот сказал 1000 — в документах написано 600

«Базовый тариф позволяет не более 600 запросов в час на один API-ключ»
rate-limits.md

Хороший P

Плохой F

Не могу решить S

Отменить (U) всегда на виду · сомнительные карточки идут первыми

4 · Станция «Итог»

Отчёт словами, не таблицей: «Вы проверили 100 ответов: 71 правдивый, 24 плохих, 5 неясных. Судья совпал с вами в 92% — ему можно доверять рутинную проверку». Три действия: «Посмотреть 8 ответов, где вы разошлись с судьёй», «Сохранить результат», «Проверить новые ответы». И подсказка на будущее: «Обновите бота — прогоните эти же вопросы снова и увидите, стал ли он лучше».

Сквозные принципы

- Термины — с подсказками, а не под запретом. Слова «судья», «галлюцинация», «голдсет» используются, но каждое подчёркнуто пунктиром: навёл — два предложения объяснения. Новичок не спотыкается, джуниор выучивает язык индустрии.
- Режим обучения («под капотом»). Выключен по умолчанию. Включил — на каждой станции разворачивается панель: какая команда выполнялась (lazy check), какой файл записался (verdicts.jsonl — открыть), как судья пришёл к вердикту (утверждения → цитаты → категория). Приложение выращивает пользователя до понимания внутренностей и работы с CLI.
- Правило трёх вопросов. Каждая станция отвечает «что это», «зачем» и «как это работает» до того, как просит что-то сделать.
- Деньги — всегда явно. Любое действие, тратящее деньги (запуск судьи), показывает смету и просит подтверждение. Отмена решения — всегда одна кнопка.

Меню задач судьи и дорожная карта

«Факт-чек по базе» — только одна из задач, под которые бывают готовые судьи. Архитектура PressF (разбить на части → найти доказательства → вердикт → человек подтверждает) натягивается почти на каждую. В конфиге уже лежит поле task — задел под вторую задачу.

Задача судьи	Что проверяет	Статус для PressF
Про правду		
Faithfulness / факт-чек по базе	Не выдумал ли бот: каждое утверждение против документации, с дословными цитатами	Есть — ядро продукта
Качество поиска (retrieval)	Принёс ли поиск чанки, из которых вообще можно ответить. Половина провалов RAG — мусор из поиска, а не враньё модели	Приоритет №3: доказательства уже собираются, осталось судить их отдельно
Полнота ответа	Ответ правдив, но целиком ли отвечает? (в базе три условия — бот назвал одно)	Кандидат позже
Про правила		
Соблюдение политик	Ответ против правил компании: «не обещать возвраты», «не давать мед. советов». Та же механика, «базой» служит документ политик	Приоритет №2: почти нулевая новая механика, платёжеспособный рынок (юристы, банки)
Инструкции и формат	Соблюдён ли JSON, длина, язык, структура	Кандидат позже — скучно, но ломается постоянно
Безопасность	Утечки персональных данных, провокации, prompt injection	Кандидат позже
Про сравнение		
Pairwise-судья	Два ответа на один вопрос (старый и новый бот): какой лучше и почему	Приоритет №1: замыкает регрессию «обновил бота — стало лучше?» и подбирает chosen для DPO-пар; переиспользует ревью-экран и отчёт согласия
Другие жанры — сознательно не берём		
Оценка агентов	Траектория действий: те ли инструменты, достиг ли цели	Не берём: другая архитектура, туда уходят большие платформы
Код, суммаризация, перевод, text-to-SQL	Свои специализированные проверки	Не берём: другие продукты

Порядок приоритетов

- №1 Pairwise — замыкает регрессию (самый ценный повторяющийся сценарий: голдсет делается не ради разметки, а ради сравнений «было/стало») и переиспользует всё: тот же ревью-экран (карточка с двумя ответами — «левый / правый / ничья»), тот же отчёт согласия.
- №2 Политики — почти нулевая новая механика: политики компании подключаются как та же «база знаний», цитаты те же. Рынок платёжеспособный: юристы, банки, медицина.
- №3 Retrieval — дёшево (доказательства уже собираются при факт-чеке, надо только судить их отдельно) и даёт диагностику «чини поиск, а не промпт».

Позиция на рынке

Большие платформы (LangSmith, Braintrust, Langfuse) покрывают всё это функционально, но они тяжёлые, облачные и командные. Библиотеки метрик (Ragas, DeepEval) дают судью без петли человеческого ревью.

Универсальные разметчики (Argilla, Label Studio) не знают, что такое RAG-факт-чек. Ниша PressF — пересечение: локальный, бесплатный, специализированный, с готовым судьёй и дешёвой человеческой проверкой — и единственный, кто учит методологии, а не предполагает её знание. Сила — в узости: задачи, где доказательство — это цитата из документа.

pressf · июль 2026 · «Lazy goldset annotation: agent does the work, you press F»