

Внедрение GenAI в Enterprise: от экспериментов к надежной реальности

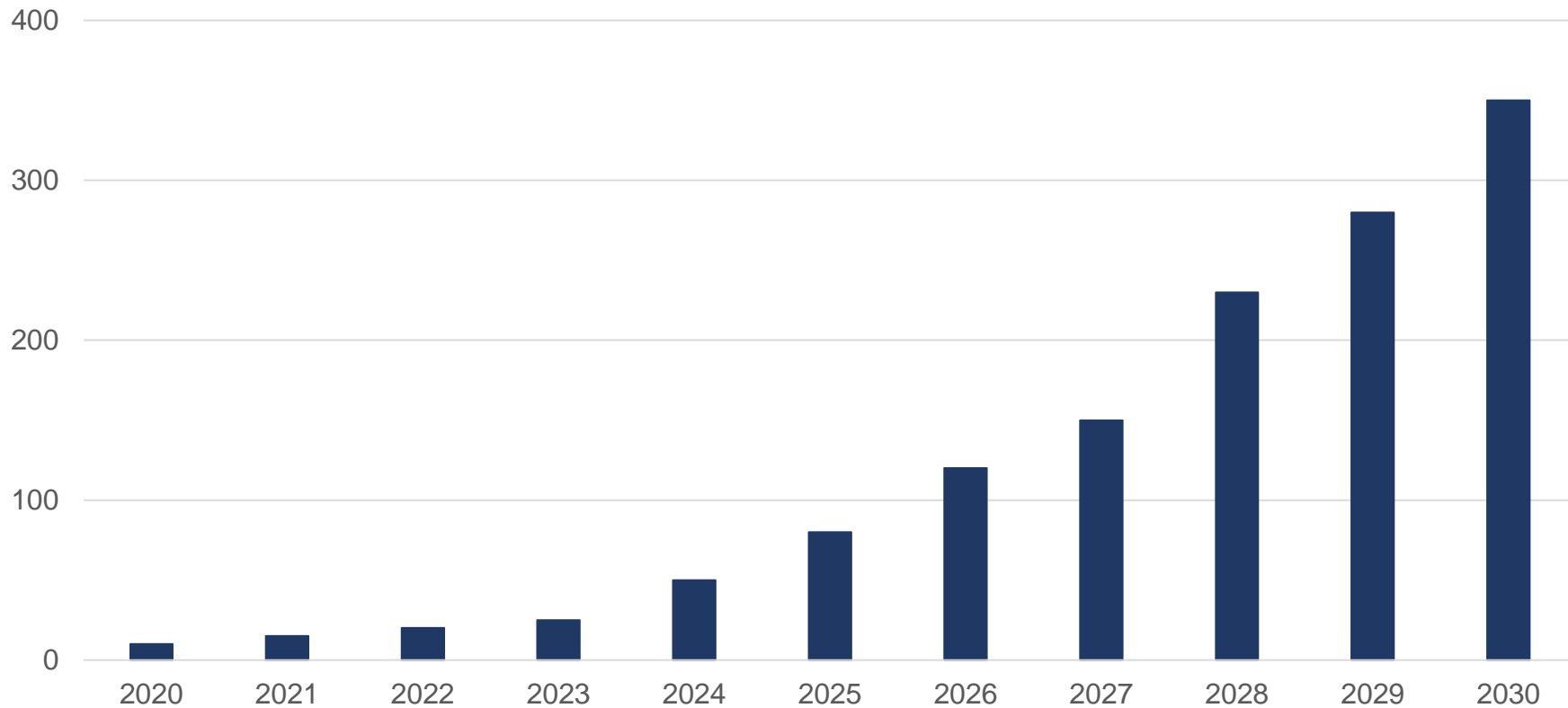


**Кудряшов
Вячеслав**

@SlavaKudryashov

Slava.Kudryashov@gmail.com

Внедрение GenAI в Enterprise: от экспериментов к надежной реальности



* - по данным McKinsey

Внедрение GenAI в Enterprise: от экспериментов к надежной реальности

>1 100

инцидентов с использованием ИИ-
систем зафиксировано
в MIT AI-Incident Database



Goldman
Sachs



Классификатор	Инцидентов шт.
AI system safety, failures & limitations	270
Discrimination & toxicity	200
Human-computer interaction	157
Misinformation	92
Privacy & security	70
etc.	378
ВСЕГО	1 167

Внедрение GenAI в Enterprise:
от экспериментов к надежной реальности

**Факторы, которыми
нельзя пренебрегать**

1. «Грязные» данные

Внедрение GenAI в Enterprise:
от экспериментов к надежной реальности

**Факторы, которыми
нельзя пренебрегать**

1. «Грязные» данные
2. Недетерминированность результатов

Внедрение GenAI в Enterprise:
от экспериментов к надежной реальности

**Факторы, которыми
нельзя пренебрегать**

1. «Грязные» данные
2. Недетерминированность результатов
- 3. Отсутствие
идемпотентности**

Внедрение GenAI в Enterprise:
от экспериментов к надежной реальности

**Факторы, которыми
нельзя пренебрегать**

1. «Грязные» данные
2. Недетерминированность результатов
3. Отсутствие идемпотентности
4. «Разорванная»
трассировка

Внедрение GenAI в Enterprise:
от экспериментов к надежной реальности

**Факторы, которыми
нельзя пренебрегать**

1. «Грязные» данные
2. Недетерминированность результатов
3. Отсутствие идемпотентности
4. «Разорванная» трассировка
- 5. Ручной откат изменений**

Внедрение GenAI в Enterprise:
от экспериментов к надежной реальности

**Факторы, которыми
нельзя пренебрегать**

1. «Грязные» данные
2. Недетерминированность результатов
3. Отсутствие идемпотентности
4. «Разорванная» трассировка
5. Ручной откат изменений
- 6. Еще одна... единая точка отказа**

Внедрение GenAI в Enterprise:
от экспериментов к надежной реальности

**Факторы, которыми
нельзя пренебрегать**

1. «Грязные» данные
2. Недетерминированность результатов
3. Отсутствие идемпотентности
4. «Разорванная» трассировка
5. Ручной откат изменений
6. Еще одна... единая точка отказа
- 7. «Неочевидность»
мониторинга**

Внедрение GenAI в Enterprise: от экспериментов к надежной реальности

Факторы, которыми нельзя пренебрегать

1. «Грязные» данные
2. Недетерминированность результатов
3. Отсутствие идемпотентности
4. «Разорванная» трассировка
5. Ручной откат изменений
6. Еще одна... единая точка отказа
7. «Неочевидность» мониторинга
8. ...

Внедрение GenAI в Enterprise: от экспериментов к надежной реальности



Внедрение GenAI в Enterprise: от экспериментов к надежной реальности



Внедрение GenAI в Enterprise:
от экспериментов к надежной реальности

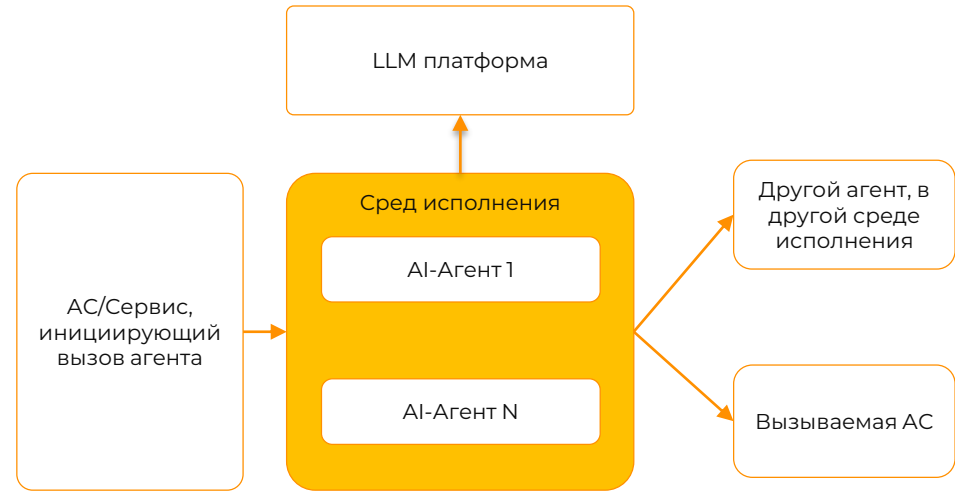
Паттерны, снижающие технологические риски

Позволяющие бесшовно для клиентов, запускать GenAI

1. Уникальная идентификация выполняемых агентами операций
2. HealthCheck агентов
3. Трейс действий агента
4. Предотвращение зацикливаний и чрезмерных взаимодействий агентов
5. Retry
6. Bypass на стороне AC
7. StopEvent для агентов разной критичности
8. Проверка готовности к использованию новой версии LLM
9. ...

Внедрение GenAI в Enterprise: от экспериментов к надежной реальности

Внедрять GenAI
решения **будет легче** при
реализации доработок на
среде исполнения агентов



ПРОБЛЕМЫ

ПАТТЕРНЫ 01-08

ПРИМЕРЫ

Внедрение GenAI в Enterprise: от экспериментов к надежной реальности

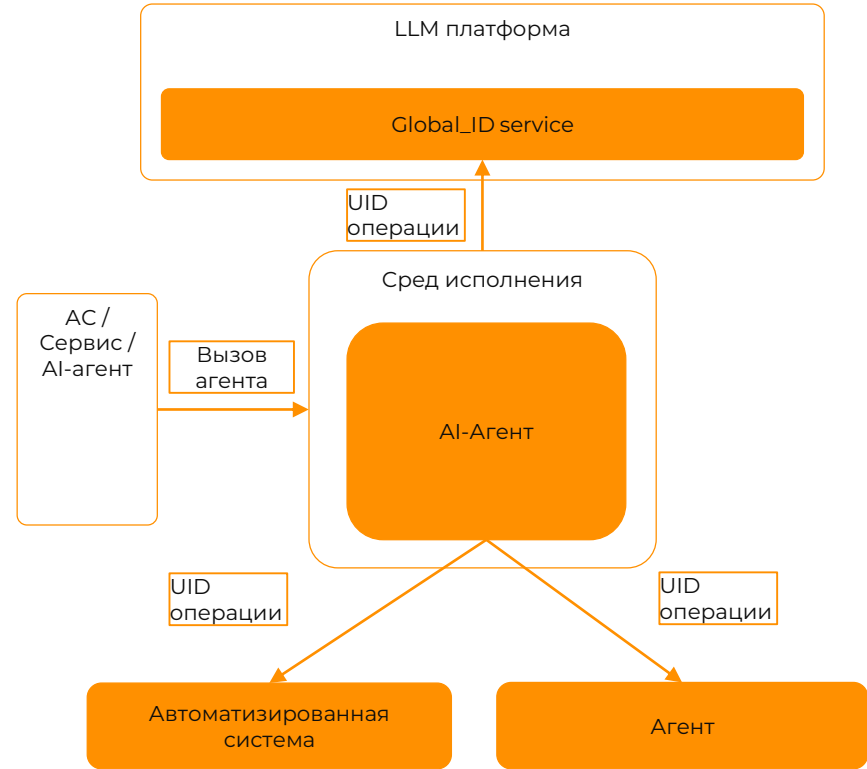
01. Уникальная идентификация выполняемых AI-агентами операций

Проблема

Неразличимость операций агентов ведет к коллизиям, неотслеживаемости действий и невозможности атрибуции в распределенной среде

Паттерн

Присваиваем каждой операции AI-агента уникальный идентификатор (UID) для однозначного определения её принадлежности к родительской операции



Внедрение GenAI в Enterprise: от экспериментов к надежной реальности

02. Предоставление статуса работоспособности AI-агентов (HealthCheck)

Проблема

Неспособность автоматически детектировать и изолировать AI-агенты с нарушенной функциональностью из-за отсутствия предметно-ориентированных проверок работоспособности

Паттерн

Внедряем HealthCheck для AI-агентов, проверяющий:
Техническую доступность (readiness/liveness probe)
Прикладную работоспособность (функциональные проверки, определенные командой, не менее трех))

Внедрение GenAI в Enterprise: от экспериментов к надежной реальности

03. Трейс действий AI-агента

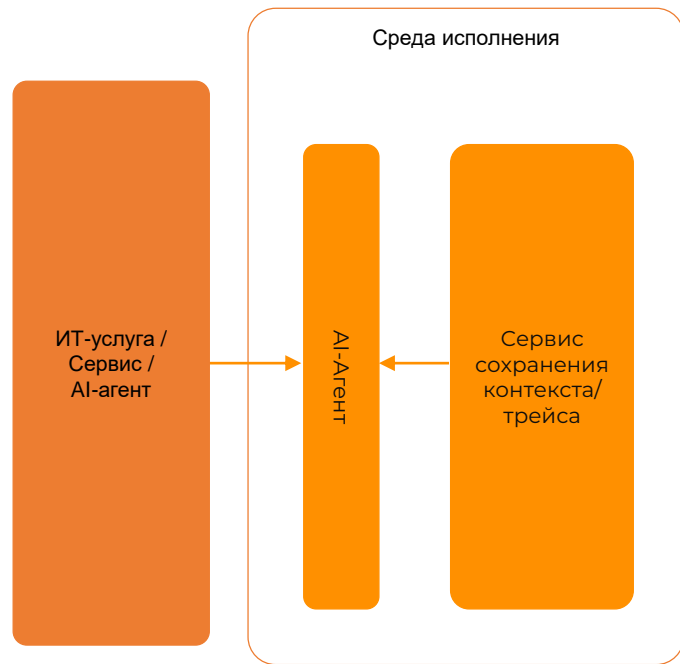
Проблема

Отсутствие детерминированной записи шагов AI-агента делает невозможным аудит, отладку и восстановление после сбоев, особенно при нелинейной логике работы

Паттерн

Сохраняем (в AEF) полный трейс* работы AI-агента: от вызова до завершения обработки запроса

**В трейс сохраняется информация о входящем запросе (в том числе, но не ограничиваясь, UID родительской операции и агента), вызовах сервисов/API/других агентов/LLM, признаки возможности отката выполняемых изменений (если вызов на изменение), получаемые ответы, выполняемые действия, возвращаемый результат, UID агента, UID операции, контрольные параметры выполнения задачи агентом (кол-во хопов, TTL, StopEvent), исполняемый json*



Внедрение GenAI в Enterprise: от экспериментов к надежной реальности

04. Предотвращение зацикливаний и чрезмерных взаимодействий AI-агентов

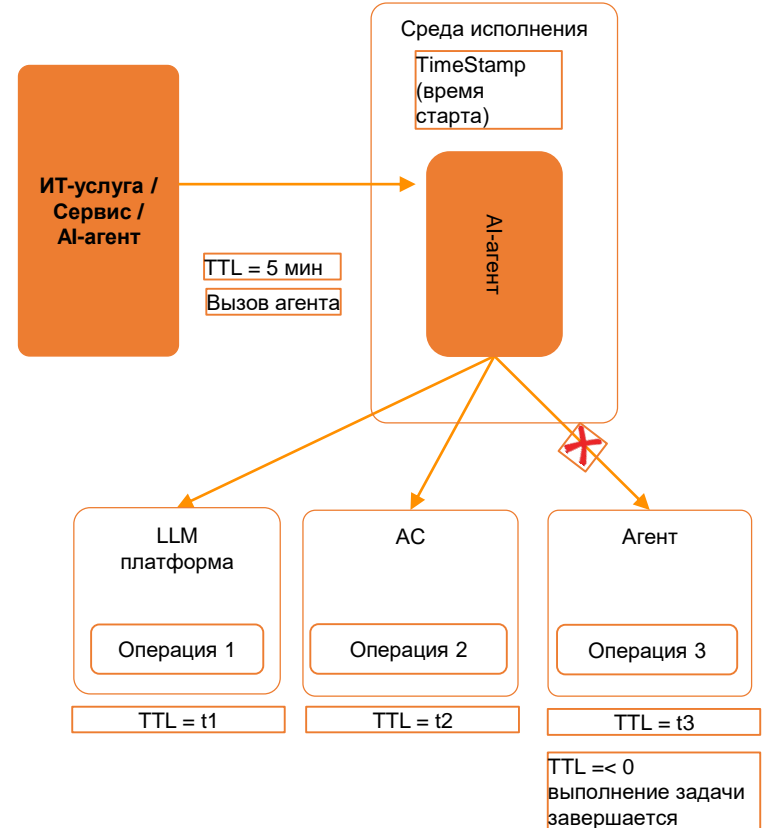
Проблема

Отсутствие гарантированного прерывания AI-агентов при выходе за нормативные границы провоцирует критические инциденты: бесконечные циклы, ресурсные коллапсы и распространение некорректных состояний в мультиагентной среде

Паттерн

Ограничиваем работу AI-агента контрольными параметрами:

- Внедряем параметры в логику AI-агента и внешние сервисы
- Прерываем выполнение задачи при срабатывании хотя бы одного параметра (TTL, StopEvent...)
- Уведомляем сервис-инициатор с указанием причины прерывания



Внедрение GenAI в Enterprise: от экспериментов к надежной реальности

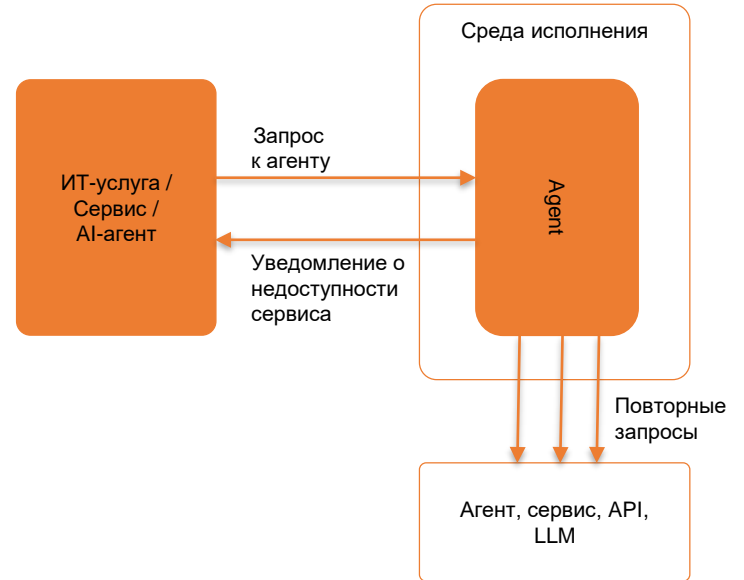
05. Retry

Проблема

В мультиагентной архитектуре неограниченные обращения к деградировавшему сервису создают лавинообразную нагрузку, ведущую к каскадным отказам

Паттерн

Устанавливаем retry - ограничения повторных вызовов сервисов (обращения к другим AI-агентам, LLM, API,...) при их временной недоступности



Внедрение GenAI в Enterprise: от экспериментов к надежной реальности

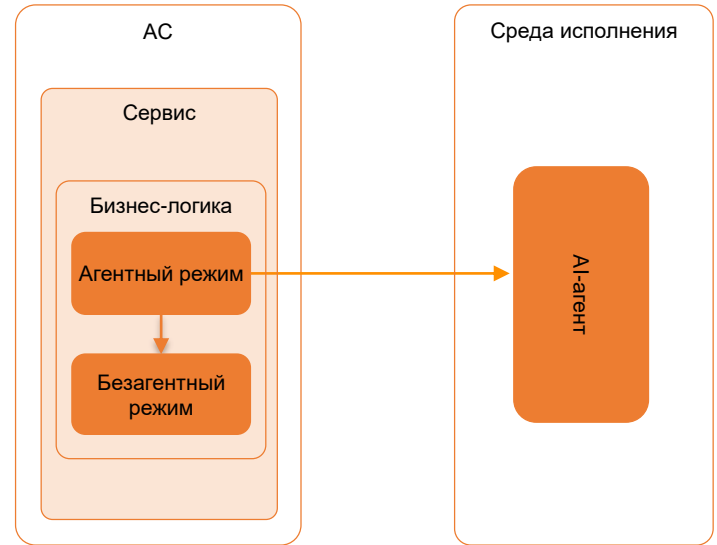
Об. ВыPass на стороне АС (Безагентный режим) и наличие Business Continuity Plan (BCP)

Проблема

«Паралич» ИТ-услуг при отказах AI-агентов из-за отсутствия автоматизированных механизмов быстрого переключения на гарантированные резервные сценарии, ведущий к нарушению SLA и бизнес-потерям

Паттерн

Реализуем в вызывающей AI-агента ИТ-услуге механизм позволяющий продолжать операцию в безагентном режиме – без участия AI-агента, используя резервные механизмы/сценарии (заранее описанные и согласованные в BCP)



Внедрение GenAI в Enterprise: от экспериментов к надежной реальности

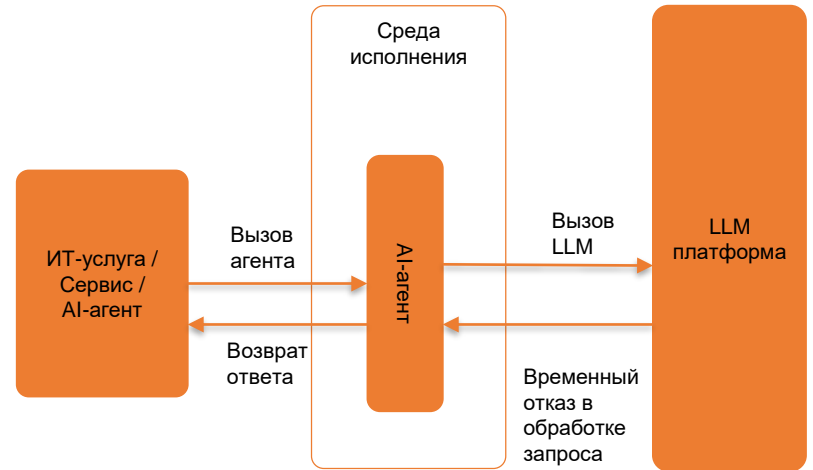
07. StopEvent для AI-агентов разной критичности

Проблема

Отсутствие избирательного ограничения запросов к базовым LLM-сервисам при пиковых нагрузках или сбоях, может приводить к коллапсу высокочритичных бизнес-процессов

Паттерн

- Реализуем на стороне AI-агентов корректную обработку сообщений типа «отказ в обслуживании» от LLM
- Реализуем на стороне LLM возможность различать запросы от AI-агентов разного уровня критичности



Внедрение GenAI в Enterprise: от экспериментов к надежной реальности

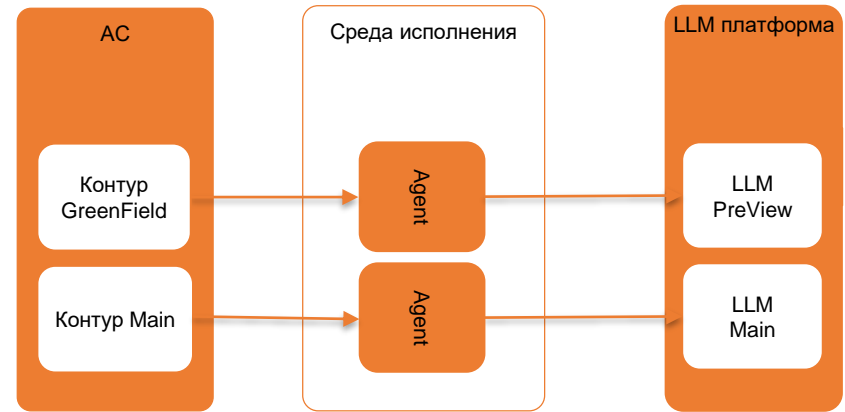
08. Проверка готовности к использованию новой версии LLM платформы

Проблема

Прямой перенос непроверенных обновлений LLM-модели в промышленный контур может спровоцировать массовые сбои зависимых AI-агентов из-за скрытых несовместимостей

Паттерн

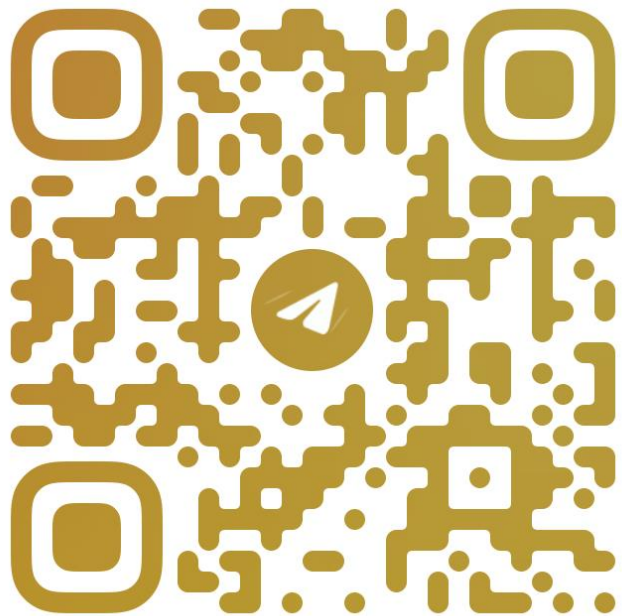
После обновления версии LLM-модели на инсталляции PreView в промышленном контуре необходимо выполнять проверку работы агентов на лояльных потребителях



Внедрение GenAI в Enterprise:
от экспериментов к надежной реальности

от «экспериментальной технологии» к «Enterprise инфраструктуре»

- Отказы не приводят к катастрофическим последствиям
- Поведение агентов полностью предсказуемо
- Изменения вводятся без нарушения SLA



Канал про надёжность:
@Simple_Reliability

ВОПРОСЫ