

Агентная платформа: рантайм и модель безопасности

Внутренняя платформа, на которой LLM-агенты исполняют бизнес-сценарии под контролем исполняемой политики безопасности.

Период: 07/2026 – наст. время

Роль: разработчик рантайма и модуля безопасности в команде платформы

Исходный код закрыт

ЗАДАЧА

Сценарии агентов длятся часами и днями и по ходу ждут подтверждения человека. Перезапуск воркера не должен откатывать их назад. При этом по каждому вызову нужно видеть, кто его инициировал, в какой зоне он выполняется и какое правило его разрешило. До этого модель доступа существовала только в виде документа.

РЕШЕНИЕ

- Сценарий описывается снимком и компилируется в граф LangGraph, который исполняет Temporal. Каждый узел работает отдельной activity, поэтому падения и ожидание подтверждений не прерывают сценарий.
- Компиляция реализована чистой функцией: из одного снимка всегда получается один и тот же граф. Иначе Temporal при воспроизведении истории падает с ошибкой недетерминизма.
- В состоянии графа и в истории хранятся только ссылки на артефакты, сами данные туда не попадают.
- Ошибки разделены на три класса. Временные повторяются, фатальные и отказы политики не повторяются, а бизнес-результат вроде расхождения данных оформлен обычным переходом в графе.
- Модуль безопасности описывает зоны, роли, классы данных, уровни автономии и абсолютные запреты кодом на Python. Функция decide() проверяет каждый вызов в семь шагов в фиксированном порядке и возвращает правило, на котором основано решение.
- Сервисы на Java обращаются к модулю по HTTP с mTLS. Без записи в журнал решений, куда можно только добавлять, решение не выдаётся.

РЕЗУЛЬТАТЫ

Состояние на 09/2026, подтверждено кодом и тестами

- Детерминизм компиляции проверяется тестом между процессами: один снимок даёт один и тот же хеш в разных процессах.
- Рантайм: 176 тестов в standalone-версии, 58 тестов в репозитории платформы.
- Модель безопасности в коде: 8 зон и 10 разрешённых переходов, матрица ролей 12 × 9, 22 вида действий, 10 абсолютных запретов, уровни автономии L0–L3 с аварийным отключением.
- Модуль безопасности: 145 тестов, проверка типов туру в режиме strict.
- Следующий шаг: подключение модуля безопасности к основному сервису платформы.

ТЕХНОЛОГИИ

Python · Temporal · LangGraph · gRPC · HTTP с mTLS · Policy as code · mypy strict · pytest