TEMADEV NOTES · STRATEGY B2B 2026

GigaChat или Claude — не тот вопрос

Почему реальный выбор B2B-команды в 2026 году — это выбор между моноархитектурой и роутером.

АПРЕЛЬ 2026 · ДОКУМЕНТ АРХИТЕКТУРНОГО СИНТЕЗА

ПАРАДИГМА

Сравнение LLM-провайдеров — это ловушка закупочного мышления

Выбор между GigaChat, YandexGPT и Claude по цене токена и набору фич воспроизводит ошибку обзоров BPM-систем десятилетней давности. Команды пытаются ответить на закупочный вопрос («какого вендора выбрать»), игнорируя архитектурный — «как устроен наш стек, чтобы вендор был заменим».

ВЗГЛЯД 2024: ЗАКУПКИ

	GigaChat	Yandex	Claude
Контекст (токен)	32k	16k	100k
Стоимость (ввод)	₽0.02	₽0.015	$0.008
Finetuning	✅	✅	❌
Скорость (RPS)	Высокая	Средняя	Высокая

ВЗГЛЯД 2026: АРХИТЕКТУРА

Вендор как заменяемый адаптер

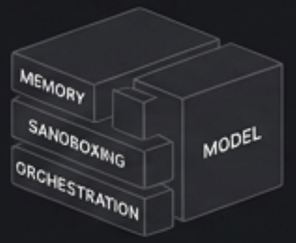
РЫНОЧНЫЙ КОНТЕКСТ · АПРЕЛЬ 2026

Провайдеры продают не модели. Они продают стеки.

В апреле 2026 года произошел одновременный шифт. Anthropic, OpenAI, Сбер и Яндекс перестали конкурировать просто весами моделей. Они поглотили execution layer, превратив agent harness в managed-продукты. Стек стал неделимым.

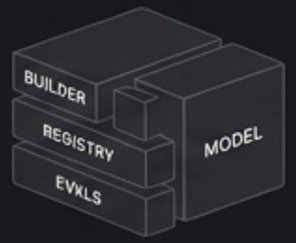
Anthropic

Managed Agents (persistent memory, sandboxing, multi-agent orchestration).



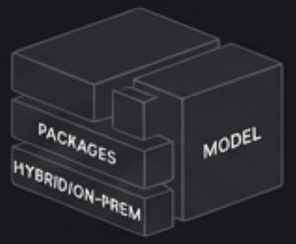
OpenAI

AgentKit (Agent Builder, Connector Registry, Evals).



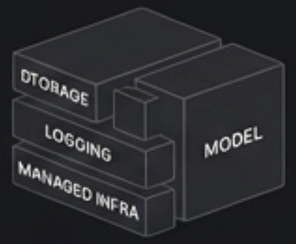
Сбер

GigaChat API (public token packages, hybrid/on-prem deployment).



Yandex Cloud

Foundation Models (Object Storage, Logging, managed infra).

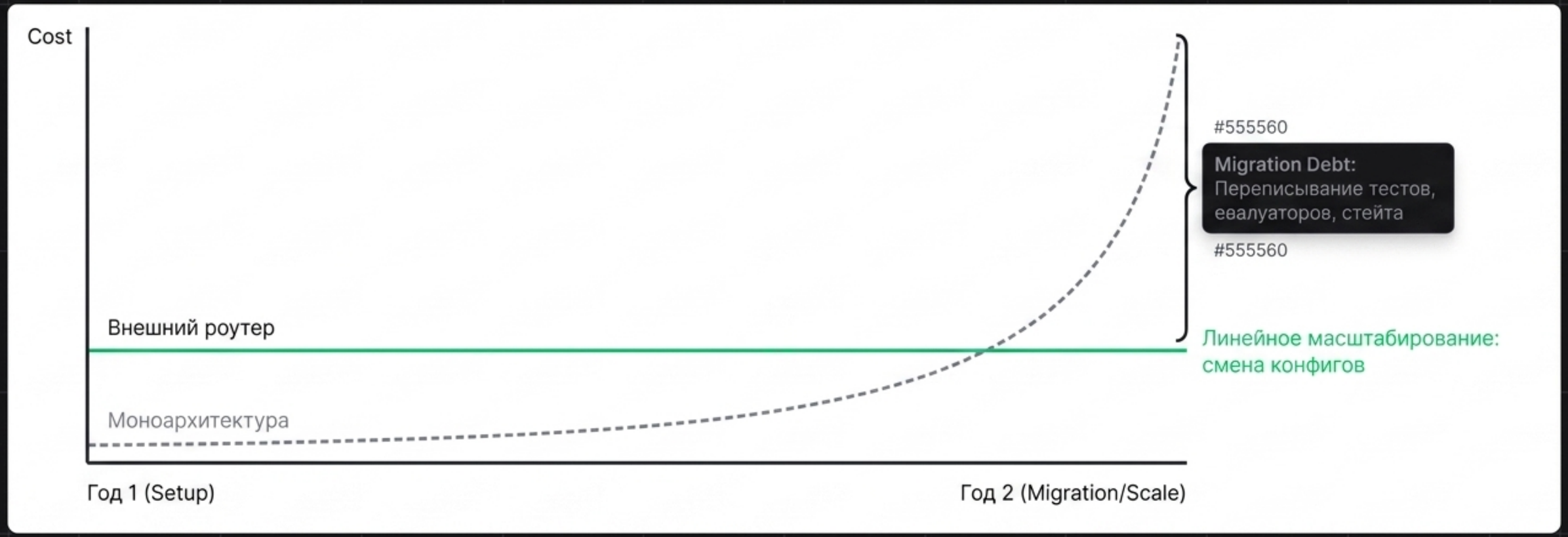


Цена ошибки смещается с цены токена на стоимость переезда между стеками.

ЭКОНОМИКА АРХИТЕКТУРЫ

Иллюзия первого года: монолит берет в долг

Базовый сценарий 2026: команда собирает всё (парсинг, классификацию, tool calling) на одном SDK провайдера. Это экономит инженерные часы на старте. Но дешевизна первого года оплачивается стоимостью миграции на втором. Долг миграции конечен и математически измерим.



ТОПОЛОГИИ В СРАВНЕНИИ

Моноархитектура vs. Внешний роутер

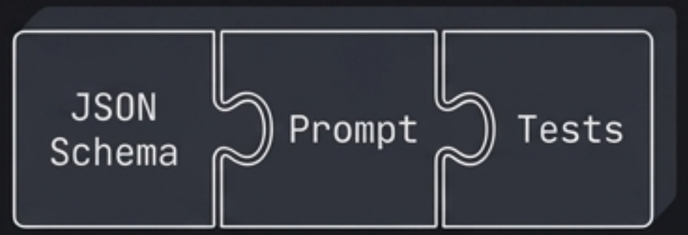
	[Монолит]	[Роутер]
Оркестрация & Состояние	Зашиты в managed-продукт вендора (SDK)	Владение состоянием на стороне продукта
Масштабирование (Год 2)	Смена архитектуры при смене вендора	Смена конфига (смена адаптера)
Изоляция 152-ФЗ	Один комплаенс-периметр для всего продукта	Маршрутизация на уровне отдельного запроса
SLA	Жестко равен SLA провайдера	Функция политики маршрутизации (fallback)
Скрытые риски	Vendor Lock-in на уровне execution loop	Стоимость поддержки собственного evaluator-харнесса

СТЫК ОШИБКИ 1 / 4

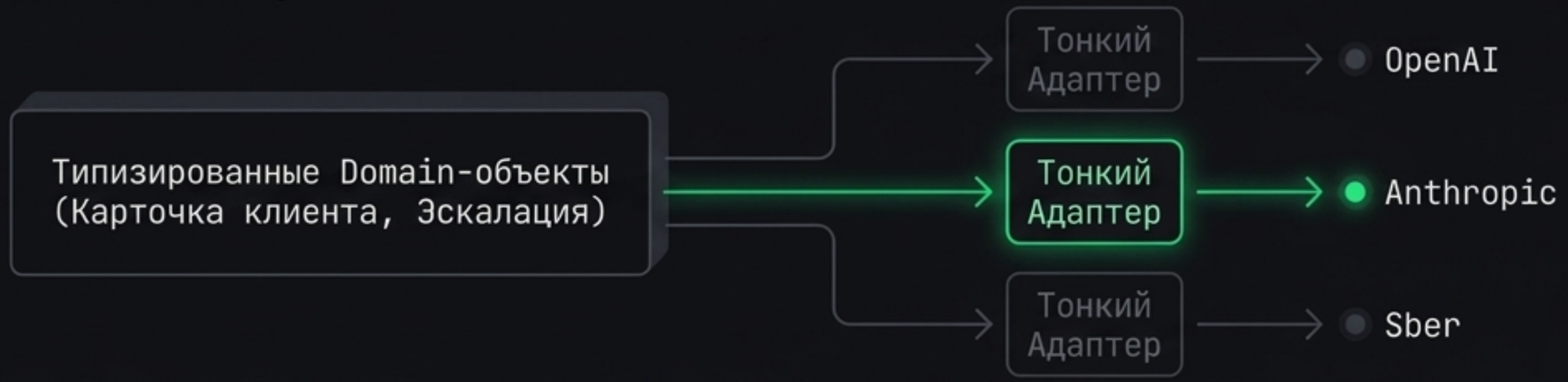
Structured Output: Формат проникает в каждый prompt

Function Calling у OpenAI, tool use у Anthropic и интерфейсы Сбера/Яндекса решают одну задачу, но имеют разные имена полей и поведение. Зашивая формат провайдера, вы кодируете его во все промпты, тесты, валидаторы и ретраи.

[Ловушка формата]



[Типизированные Domain-объекты]



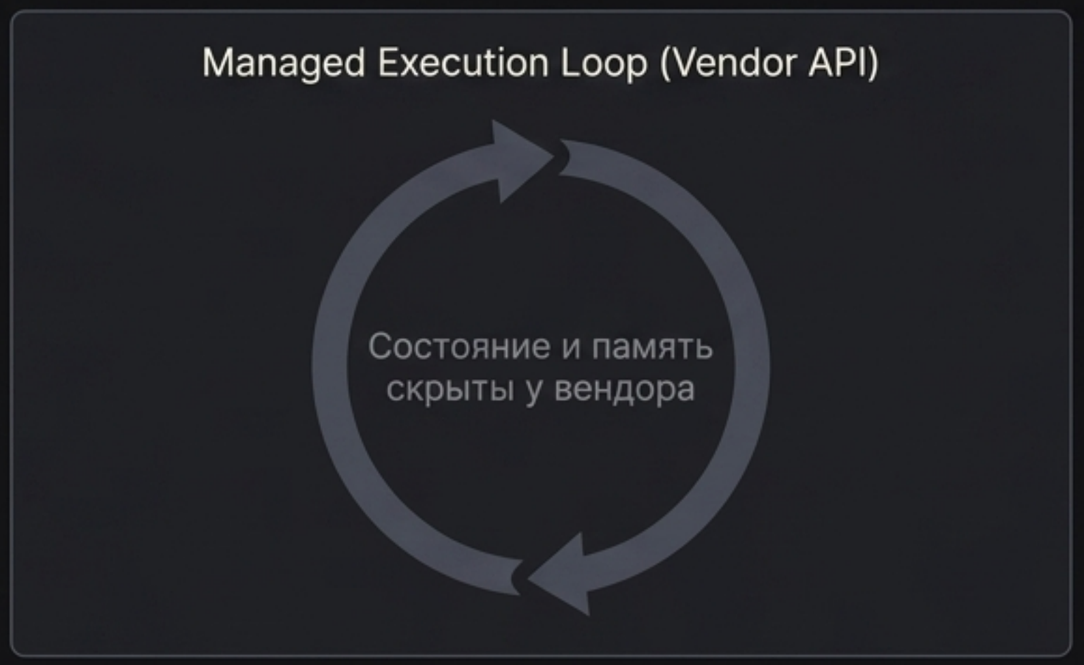
Миграция меняет только адаптер, а не бизнес-логику.

СТЫК ОШИБКИ 2 / 4

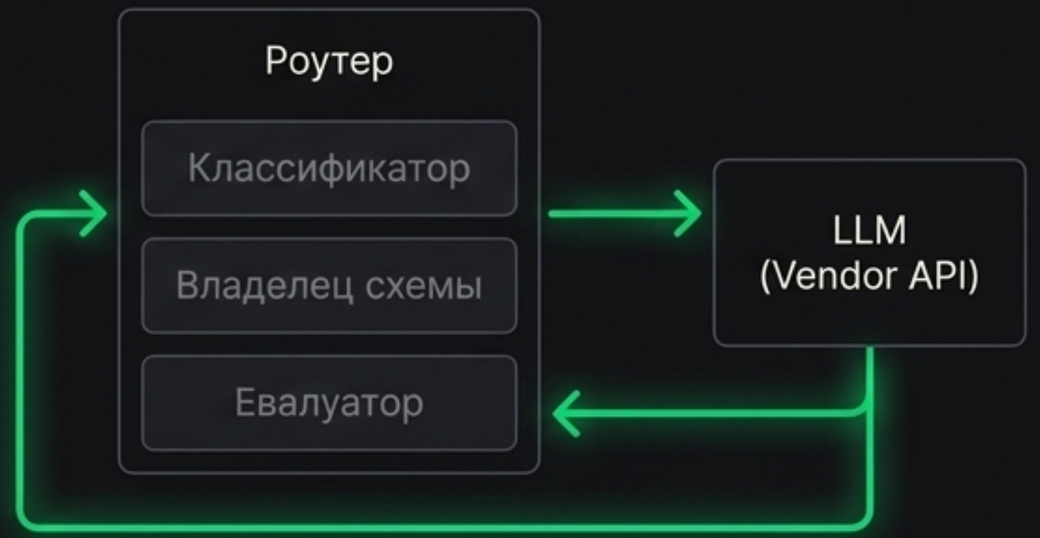
Tool Use: Кто владеет циклом исполнения?

Когда агент работает в цикле, вопрос в том, где живет состояние и кто управляет памятью между шагами. Если вы строите flow внутри managed-конструкции провайдера, переезд потребует переписывания всей логики оркестрации.

[Монолит]



[Роутер]

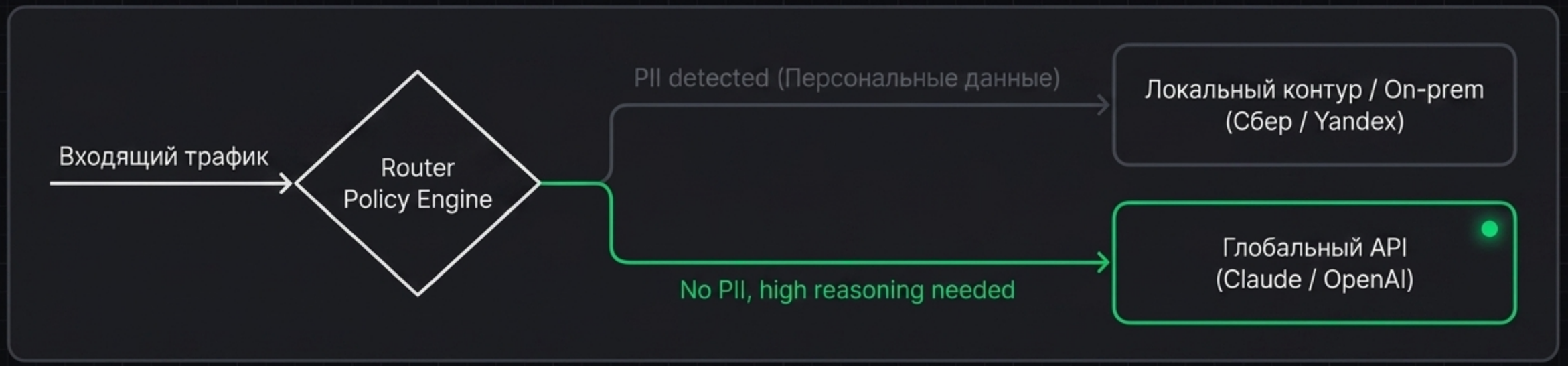


Роутер управляет шагами. Вендор только вычисляет.

СТЫК ОШИБКИ 3 / 4

152-ФЗ как архитектурный атрибут запроса

152-ФЗ — это не фильтр «какого провайдера выбрать», а параметр маршрутизации. Монолит загоняет весь продукт в единый периметр. Роутер решает это на уровне классификатора.



152-ФЗ — это ось маршрутизации, а не ограничение продукта.

SLA: Деградация провайдера против деградации продукта

В монолитной архитектуре SLA провайдера — это жесткий потолок вашего продукта.
Роутер делает SLA продукта функцией вашей политики переключений (fallback).

Моноархитектура

Product dies with the provider.

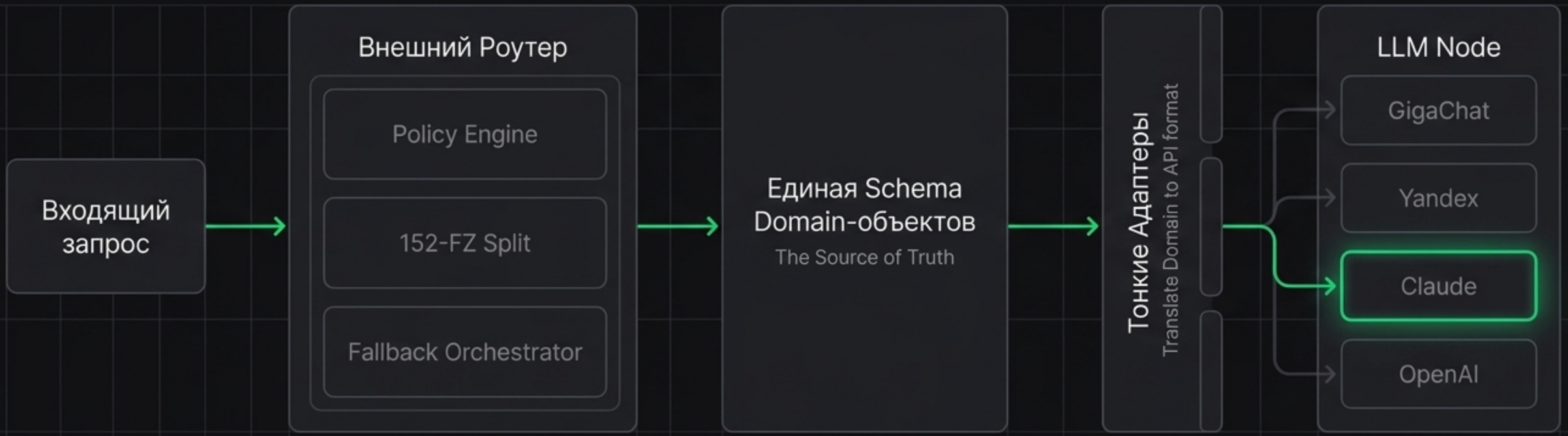
Роутерная топология

Latency p95 поднимается, контур не останавливается.

АРХИТЕКТУРНЫЙ СИНТЕЗ

Minimal Viable Architecture (MVA) 2026

Минимальная архитектура, которая делает любого вендора полностью заменяемым. Роутер поглощает функции классификации, типизации и оркестрации, оставляя LLM-провайдеру роль взаимозаменяемого вычислительного узла.



Провайдер — это исполнитель. Роутер — это источник правды.

ИНФРАСТРУКТУРА НЕЗАВИСИМОСТИ

Evaluator-харнесс: Цена вашей автономности

Роутер бесполезен без механизма тестирования. Поддержка двух адаптеров дешевле одного жесткого формата только тогда, когда у вас есть evaluator-харнесс на 200–500 типичных кейсов с эталонными ответами.

```
[EVAL] Structured Output (145 cases) -> Pass
[EVAL] Tool Use Chain (80 cases) -> Pass
[EVAL] PII Isolation (152-FZ) -> Pass
```

Смена провайдера должна давать численный ответ за одну ночь: лучше, хуже, на каких подмножествах. Остальное — ощущения.

ПЛЕЙБУК ДИАГНОСТИКИ LOCK-IN

Чек-лист для инженерных руководителей

Для CTO

Тест: В скольких местах кода/тестов зашит формат текущего провайдера?

Симптом: Десятки мест на одну задачу = монолит.

Решение: Вынести формат в адаптер, типизировать domain-объекты.

Для Head of Product

Тест: Если провайдер удалит фичу или поднимет цену на 30%, что перестанет работать через неделю?

Симптом: Продукт зависит от roadmap вендора сильнее, чем от своего.

Для Tech Lead

Тест: Можете ли вы прогнать новую модель за ночь и получить численный delta-отчет?

Симптом: Решения о миграции принимаются на «ощущениях».

Решение: Инвестировать недели (а не дни) в evaluator-харнесс.

ИНДИКАТОРЫ РЫНКА

Сигналы B2B-рынка в 2026 году

Тренд перехода от монолита к роутерам проверяется тремя рыночными сигналами. Как только они становятся массовыми, моноархитектура окончательно становится анахронизмом.

1. Managed-роутеры в GA	Появление внешних сервисов с единым API, принимающих на себя логику маршрутизации запросов (commoditization of routing).
2. Публичные кейсы миграции	Официальные tech-блоги команд о переводе production-нагрузок между РФ и зарубежными моделями без потери качества.
3. Возврат на монолит (Контр-сигнал)	Команды, публично отказывающиеся от роутера из-за операционной сложности (маркирует границу применимости роутера для простых задач).

СТРАТЕГИЧЕСКИЙ ВЫВОД

Закупочный вопрос не определяет второй год. Архитектурный — определяет.

Моноархитектура не дешевле — она просто берет в долг у вашего второго года. Инвестиции в единую доменную схему, адаптеры и evaluator-харнесс — это цена, которую B2B-бизнес платит за то, чтобы провайдеры оставались лишь взаимозаменяемыми вычислительными мощностями.