

✓selecty

MCP vs CLI

Инженерный подход
к оптимизации AI-workflow



Александр Тихомиров

Android-разработчик

6+

лет в коммерческой
Android-разработке

В SberDevices

занимался разработкой
системных приложений
для Android

В Selecty работаю над приложением одного из ведущих
банков РФ



О чем поговорим?

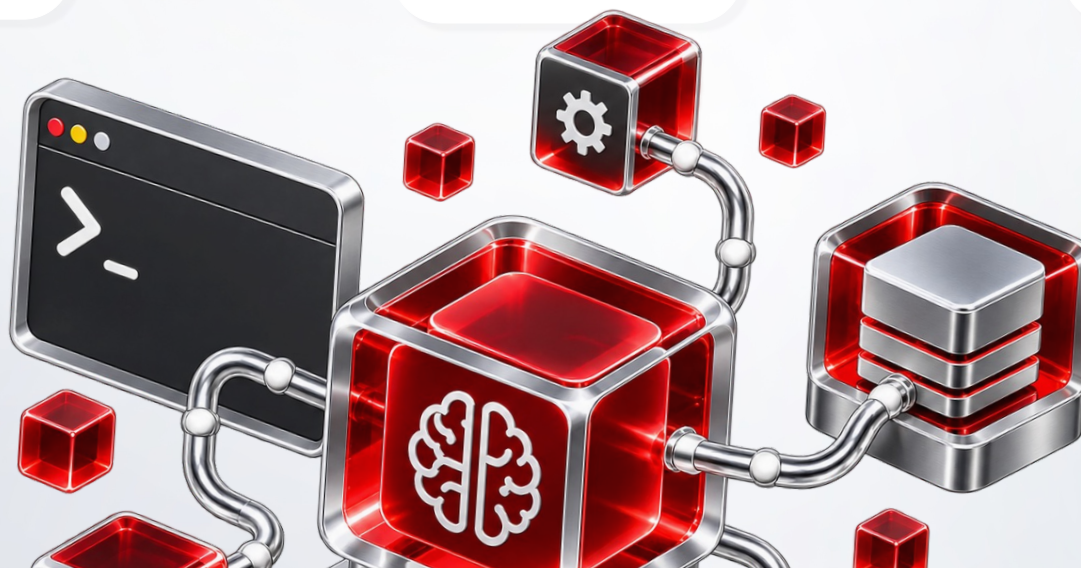
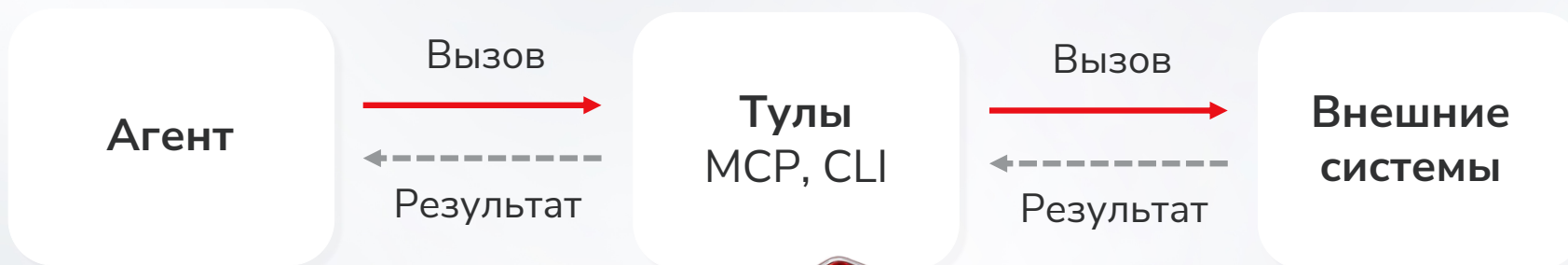


- 01 MCP и скиллы поверх CLI
- 02 Методика сравнения
- 03 Сравнение разных конфигураций в Codex
- 04 Практические выводы

Агенту нужны внешние тулы



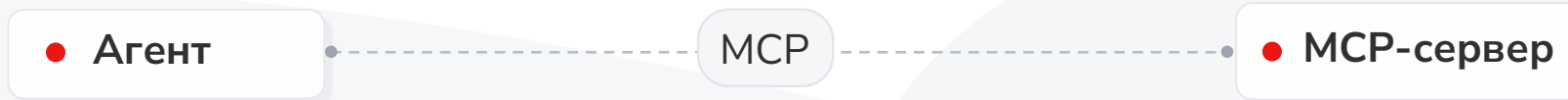
Через **тулы** агент получает данные и выполняет действия в других системах



Как работает MCP



Протокол взаимодействия AI-приложений с внешними инструментами

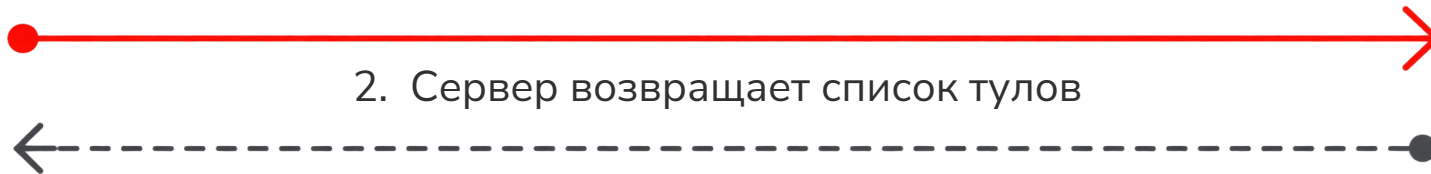


01 Список тулов



1. Агент запрашивает список тулов

2. Сервер возвращает список тулов

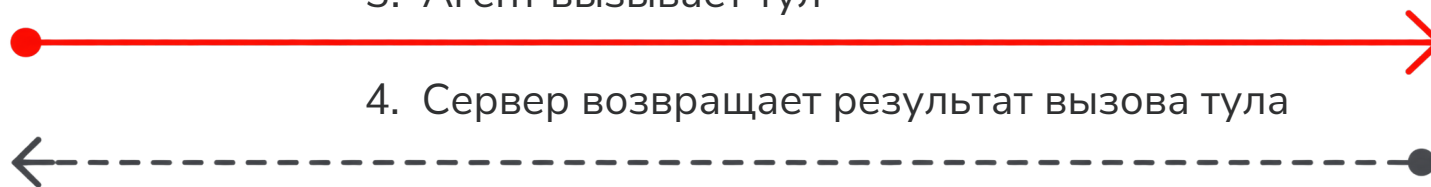


02 Вызов тула



3. Агент вызывает тул

4. Сервер возвращает результат вызова тула



Список тулов MCP сервера



Пример одного тула • JSON сокращен

```
{  
  "name": "jira-jira_get_issue",  
  "title": "Get Issue",  
  "description": "...", // Что делает тул и когда его вызывать  
  "inputSchema": {  
    ... // Параметры вызова: типы, ограничения, обязательные поля  
  },  
  "outputSchema": {  
    ... // Формат результата: поля ответа и их типы  
  }  
}
```

63

тула в виде JSON

106000

СИМВОЛОВ

Популярный
Jira MCP

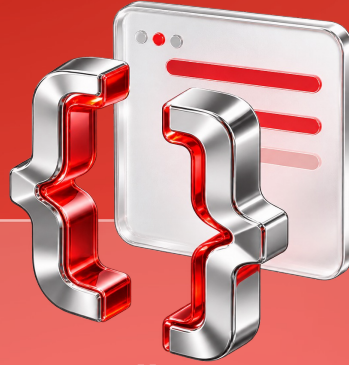


Ненужные тулы тоже занимают контекст



Большой каталог

Сервер предоставляет много тулов



Лишний контекст

В контекст попадают даже ненужные описания



✓selecty



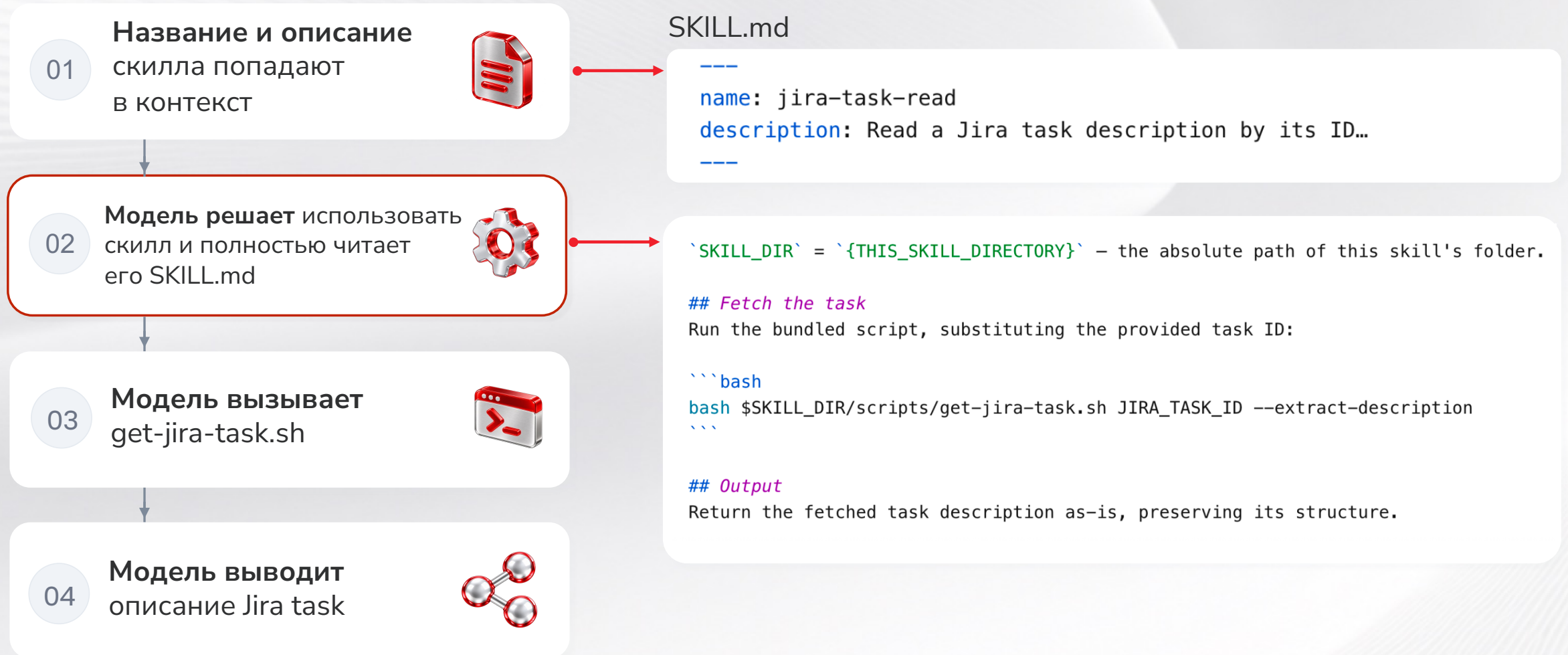
Стоимость и время

Могут вырасти стоимость и время ответа



Отключайте ненужные тулы MCP

Как работает Skill



Как скиллы устроены в Codex



Каталог скиллов формирует клиентское приложение **Codex**



```
## Skills
```

```
A skill is a set of...
```

```
### Skill roots
```

- r0 = /Users/user1/.codex/skills/.system
- r1 = /Users/user1/project/agents/skills

```
### Available skills
```

- imagegen: Generate or edit raster images... (file: r0/imagegen/SKILL.md)
- openai-docs: Use for Codex pricing... (file: r0/openai-docs/SKILL.md)

```
...
```

Модель видит название, описание и путь; полный текст читает **по необходимости**

Скиллы поверх CLI



01



SKILL.md

Короткая инструкция по работе со скриптом



02



scripts/script.sh

Оборачивает взаимодействие с общим CLI



03



CLI Tool

Общий инструмент для широкого спектра задач

jira-cli

открытый
неофициальный клиент



glab

официальный
CLI GitLab



jira-task-read

Читает задачу Jira

gitlab-mr-comments

Читает открытые комментарии к MR

gitlab-mr-create

Создает MR из текущей ветки

Сравним на задачах

Jira MCP и специализированные
скиллы поверх jira-cli



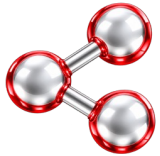
Скрипт бенчмарк



Выходные данные бенчмарка



Что считает **скрипт**



Токены

Количество потраченных токенов



Запросы к модели

Количество шагов агента



Время

Время работы процесса Codex

Общая статистика по всем запускам

Как сравниваем



Одна конфигурация

Codex 0.153.4
gpt-6-astra • medium

15 запусков

Каждый кейс
проверяется серией
прогонов

Разбор по трейсам **session.json**

Выбираем прогон
с медианным
расходом
токенов



01 Расход вхолостую



Reply with one word: hello



MCP

Подключен Jira MCP

Запрос не требует вызова инструментов



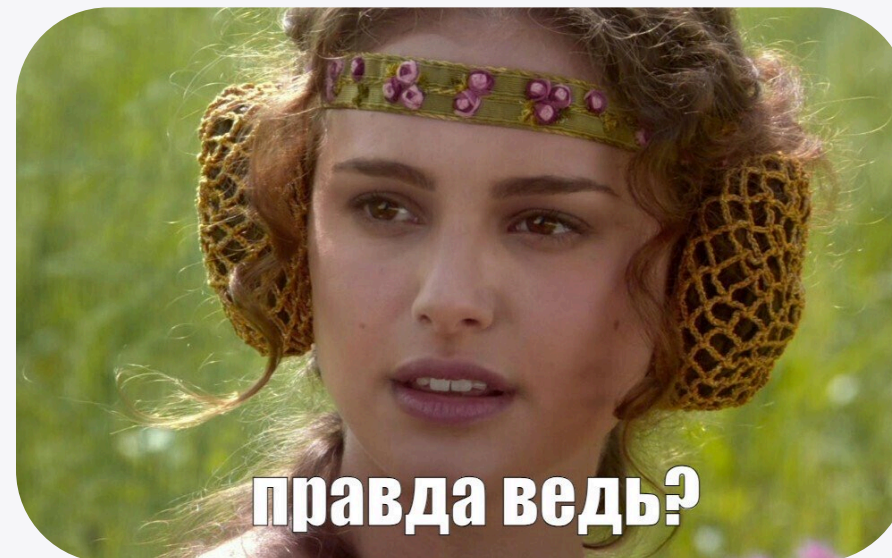
CLI

Доступен jira-task-read

Скилл также не используется

Какой вариант потратит больше токенов?

МСР должен потратить больше токенов, так как у него много тулов



Нет

Сравнение результатов



Кейс	Расход токенов	Время	Итерации
MCP	12 400	6,46 с	1
CLI	12 505	6,08 с	1

+0,85%

токенов у CLI по сравнению с MCP



Трейсы



MCP

01 · Начальный контекст
Каталог пяти системных
скиллов

02 · Финальный ответ: hello

Без вызовов инструментов

CLI

01 · Начальный контекст
Каталог пяти системных скиллов

+ **описание** `jira-task-read`

02 · Финальный ответ: hello

Без чтения SKILL.md

+105 токенов у CLI
разница связана с описанием скилла `jira-task-read`

→ Почему в варианте с MCP не подгрузились тулы?

Ленивая подгрузка MCP тулов в Codex selecty

01 Модель получает указание

Прочитай Jira-задачу RC-80899

02 Поиск тулов

Запрос

```
ALL_TOOLS.filter(jira|issue)
```

Результат

```
{  
  "name": "mcp__jira__jira__jira__get_issue",  
  "description": "Get details of a Jira issue..." // Полное описание тула и его параметров  
}, ...
```

03 Вызов тула

Запрос

```
tools.mcp_jira_get_issue({ issue_key: "RC-80899", fields: "summary,description" })
```

Результат

```
{ "summary": "Название Jira задачи", "description": "Описание Jira задачи" }
```

Выводы

**В Codex подключенные
неиспользуемые MCP
не добавляют накладных
расходов, в отличие
от скиллов**



02 Читаем задачу Jira



Read the title (summary) and description of issue RC-80899 and return both unchanged

Вернуть заголовок и описание RC-80899 без изменений



Сравнение результатов



Кейс	Расход токенов	Время	Итерации
MCP	80 751	17,65 с	4
CLI	52 887	21,99 с	4
MCP с 1 тулом	48 949	14,60 с	3

-34,51% токенов у CLI по сравнению с MCP

-39,38% токенов у MCP с 1 тулом по сравнению с MCP

Трейсы: цена поиска тулов



MCP

4 ответа

Начальный контекст

Поиск тулов

Широкий поиск в ALL_TOOLS



Повторный поиск тулов

Описание jira_get_issue



Вызов Jira

Только summary и description



Финальный ответ

Заголовок и описание задачи

CLI

4 ответа

Начальный контекст

Чтение скилла

SKILL.md



Чтение исходника

get-jira-task.sh



Запуск скрипта

get-jira-task.sh



Финальный ответ

Заголовок и описание задачи

Перерасход возникает до чтения задачи

Поиск в ALL_TOOLS добавляет описания Jira и других инструментов. Чтение скилла и исходника CLI добавляет сравнительно мало текста

Трейсы: цена поиска тулов



MCP с 1 тулом

3 ответа

Начальный контекст

Поиск тулов

Широкий поиск в ALL_TOOLS



Повторный поиск тулов

Описание jira_get_issue

Вызов Jira

Только summary и description



Финальный ответ

Заголовок и описание задачи

CLI

4 ответа

Начальный контекст

Чтение скилла

SKILL.md



Чтение исходника

get-jira-task.sh



Запуск скрипта

get-jira-task.sh



Финальный ответ

Заголовок и описание задачи

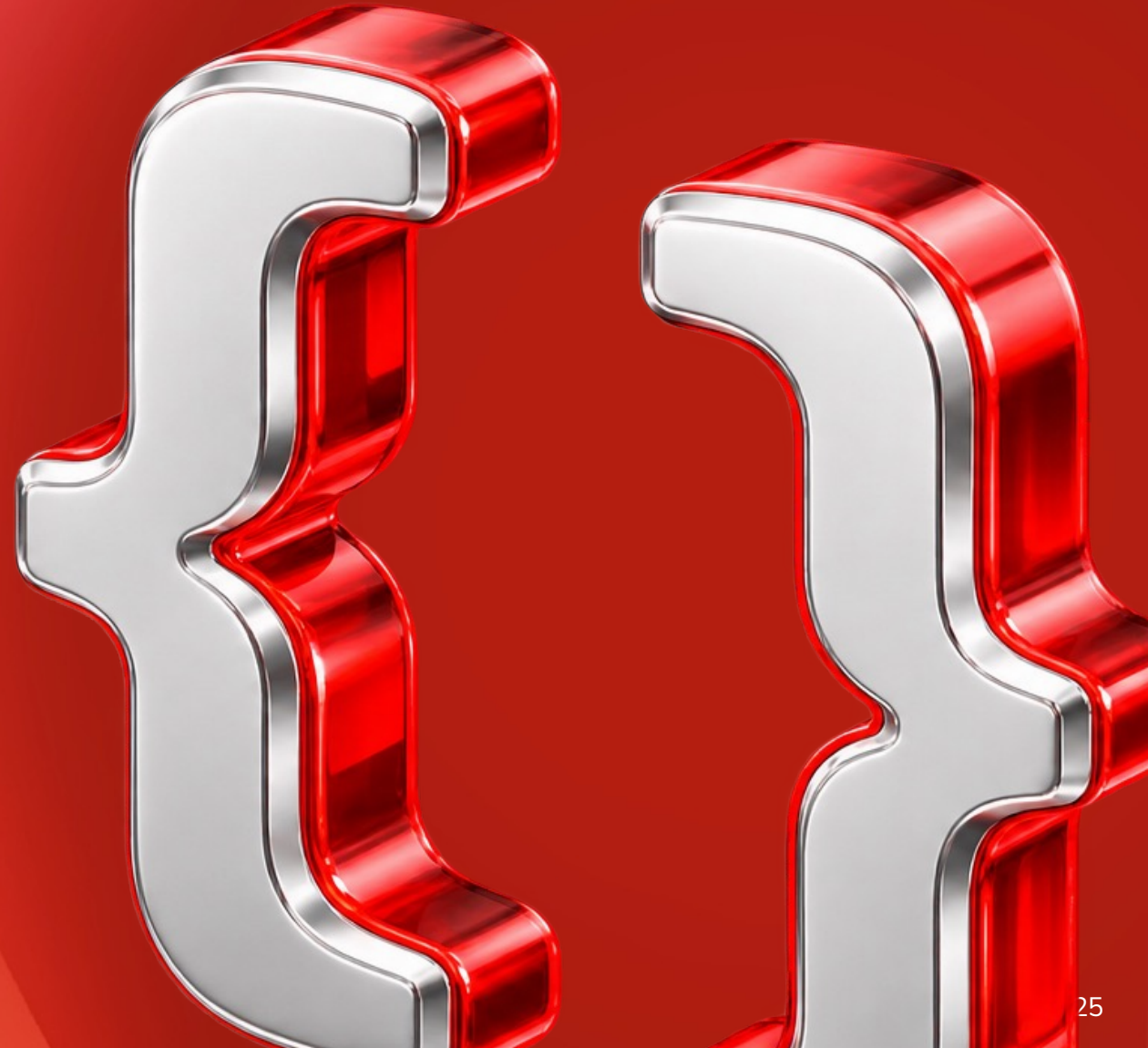
На один ответ модели меньше

При одном доступном туле повторный поиск не нужен: 3 ответа модели у MCP вместо 4

Выводы

Ленивая загрузка не устраняет
стоимость поиска инструментов

Узкий набор тулов сокращает
контекст и может убрать лишние
шаги



02 Выполняем workflow



- 01 Перевести задачу в статус **Development**
- 02 Добавить метку **ai_agent**
- 03 Вернуть описание задачи

MCP

Шаги через инструменты Jira MCP

CLI

Всё взаимодействие с Jira — в скрипте

Общий CLI

Jira CLI без документации в SKILL.md

Общий подробный CLI

Jira CLI с документацией в SKILL.md

Validation проверяет статус и метку; cleanup возвращает задачу в исходное состояние

Сравнение результатов



Кейс	Расход токенов	Время	Итерации
МСР	114 029	31,71 с	5
Общий CLI	87 154	38,71 с	6
Общий подробный CLI	77 715	35,52 с	5
CLI	39 058	28,89 с	3

-65.75% токенов у CLI по сравнению с МСР

Трейсы: цена поиска инструментов



MCP

5 ответов

Начальный контекст

Чтение скилла SKILL.md

Поиск инструментов jira|transition|issue

Получить схему jira_get_issue

Получить доступные статусы

Перевести в Development

Прочитать метки

Добавить метку ai_agent

Прочитать обновлённую задачу

Финальный ответ

Заголовок и описание задачи

CLI

3 ответа

Начальный контекст

Чтение скилла
SKILL.md

Запуск скрипта
start-task.sh

Финальный ответ

Заголовок и описание задачи

Больше обращений к модели x больше контекста в последующих обращениях

В историю добавляются описания инструментов и промежуточные данные Jira

Вызовы MCP могут объединяться в functions.exes: один вызов инструмента ≠ один ответ модели

Трейсы: цена поиска инструментов



MCP

5 ответов

Начальный контекст

Чтение скилла SKILL.md

Поиск инструментов jira|transition|issue

Получить схему jira_get_issue

Получить доступные статусы

Перевести в Development

Прочитать метки

Добавить метку ai_agent

Прочитать обновлённую задачу

Финальный ответ

Заголовок и описание задачи

Общий CLI

6 ответов

Начальный контекст

Чтение скилла SKILL.md

Общая справка jira --help

Справка команд параллельно

move --help

edit --help

view --help

Перевести в Development

Добавить метку ai_agent

Прочитать задачу view --plain

Финальный ответ

Заголовок и описание задачи

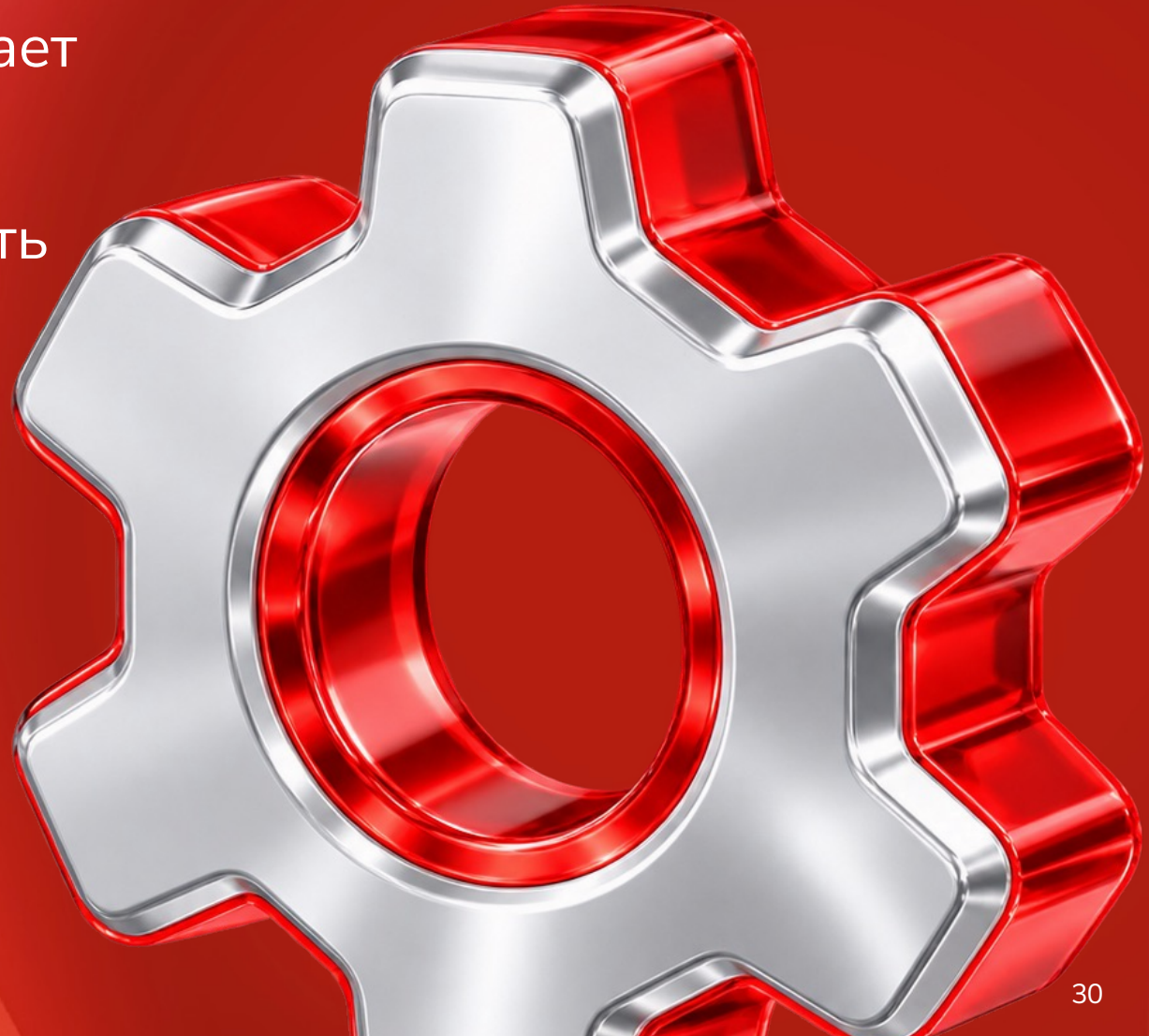
Выводы



Специализированный скрипт сокращает расход ресурсов

CLI общего назначения может работать не хуже MCP

CLI гибче в автоматизации: действия можно объединить в единый скрипт



Единого победителя нет



Вхолостую

Неиспользуемый MCP с ленивой загрузкой не загрузил описания тулов



Чтение задачи

Узкий набор тулов сокращает контекст и может убрать лишние шаги



Workflow

Последовательность действий в коде значительно сокращает потребляемые токены

Качество инструмента важнее протокола — MCP или CLI

Результат зависит от количества инструментов, объема данных и числа итераций модели

01 Контекст

Возвращайте модели только данные, необходимые для задачи

02 Проверка

Сравнивайте успешность, токены и время; причины ищите в трейсах

Выносите в **код** стабильные шаги



Исполнение

Повторяемые шаги — в код. Заранее продумайте обработку ошибок.



Детерминированная задача

Выносите в код

Недетерминированная задача

Оставляйте модели

Репозиторий проекта



Александр Тихомиров

Android-разработчик

Tr:
@OrangeFor



LinkedIn:
orange-coder

