

Вероятные Вопросы для Q&A по Докладу "Компилятор для идей"

Данный список вопросов составлен на основе анализа структуры доклада (58 слайдов) и предположений о типичной осведомленности и фокусе внимания представителей аудитории (Разработчики, Архитекторы, Менеджеры/QA).

Менеджеры и Руководители (Фокус: ROI, ТСО, Стратегия, Риски)

1. Экономика и ROI (Часть 1)

- **Вопрос:** "Это все интересная философия (про 'охоту' и 'фермы'), но вы показали слайд с ростом продуктивности (Слайд 9). Насколько эти +30-50% — реальность, а не хайп? На каких задачах достигается такой рост, и что не учитывается в этих отчетах (например, время на проверку ИИ-кода)?"
- **Цель:** Понять реальный возврат инвестиций (ROI) и избежать покупки "модного" инструмента, который не принесет выгоды.

2. On-Premise vs Cloud (Часть 2)

- **Вопрос:** "Вы упомянули российских провайдеров (Слайд 13) и целый блок про on-premise (Слайды 14-24). Зачем нам вообще 'строить свое' (CapEx), если Yandex/Sber уже предлагают готовые облачные решения? Разве они не решают наши проблемы с безопасностью?"
- **Цель:** Обосновать капитальные затраты на HW и персонал для on-prem в сравнении с готовым Paas.

3. Новые Роли и Обучение (Часть 3, 5)

- **Вопрос:** "Вы показали много сложных техник (CoT, Агенты) и упомянули 'курирование данных' (Слайд 55). Это значит, что нам **нужно нанимать новых людей** (Prompt Engineers / Data Curators)? Или это должны делать мои текущие разработчики вдобавок к своей работе?"
- **Цель:** Понять, как изменится оргструктура и какие новые компетенции необходимо развивать или нанимать.

4. Приоритизация Безопасности (Часть 4)

- **Вопрос:** "Вы упомянули OWASP (Слайд 48) и NIST (Слайд 50). **Какой один самый важный шаг** мы должны сделать завтра для защиты, если мы хотим начать 'Топ-3 задачи' (Слайд 58)? С чего начать?"
- **Цель:** Получить конкретный, приоритизированный action item для митигации рисков.

5. Юридический Риск (Новый)

- **Вопрос:** "У нас много Legacy-кода. Если ИИ-агент (Слайд 46) использует код из нашей кодовой базы для обучения и затем генерирует ответ, **кому принадлежат авторские права** на этот новый код? И как избежать юридических рисков, связанных с копирайтом?"
- **Цель:** Оценка юридических и комплаенс-рисков, связанных с генерируемым кодом.



Разработчики и ML-Инженеры (Фокус: Инструменты, Надежность, Практика)

6. Надежность Формата Вывода (Часть 3)

- **Вопрос:** "Насколько надежным в итоге можно сделать JSON-вывод (Слайд 30) с помощью response_schema? Какой реальный процент сбоев этого формата на 1000 запросов? Не придется ли мне писать 'парсер для парсера'?"
- **Цель:** Выяснить практические проблемы интеграции LLM и ожидания по обработке ошибок.

7. Trade-off Квантования (Часть 2)

- **Вопрос:** "Вы упомянули vLLM, TGI и квантование (Слайды 17-18). Насколько сильно квантование (GGUF) на самом деле снижает точность модели? Есть ли реальный trade-off между скоростью/экономией и качеством ответов?"
- **Цель:** Понять, какой ценой достигается экономия ТCO.

8. Граница Возможностей SLM (Часть 2)

- **Вопрос:** "Вы упомянули SLM (Слайд 16) как 'идеальный инструмент'. Если я возьму 7B-модель (SLM) и 70B-модель (LLM) и задам им одну и ту же задачу по рефакторингу, в чем я на практике увижу разницу (например, в стиле, глубине понимания)?"
- **Цель:** Определить, когда использование маленькой модели становится нецелесообразным.

9. Проверка ИИ-Тестов (Часть 4)

- **Вопрос:** "Кейс с 'Тестированием Legacy' (Слайд 39) — это отлично. Но **кто будет проверять тесты, которые написал ИИ?** Мы не получим просто 'мусор на входе, мусор на выходе' и ложное чувство безопасности?"
- **Цель:** Классический вопрос о доверии к ИИ-коду, особенно к тестам, которые должны быть эталоном качества.

10. Этические Риски (Новый)

- **Вопрос:** "Если мы используем LLM для предварительного ревью (Слайд 43), не начнет ли модель **реплицировать предвзятость** или неоптимальные паттерны из нашего старого кода, тем самым усиливая технический долг, а не исправляя его?"
- **Цель:** Указать на этические и технические риски, связанные с обучением LLM на "грязном" коде.

Архитекторы и Технические Лиды (Фокус: Интеграция, Масштаб, Наблюдаемость)

11. Отладка Агентного Кода (Часть 4)

- **Вопрос:** "Вы упомянули 'Разработку с Агентами' (Слайд 46). Но **как отлаживать** то, что сделал агент? Если он 'починил' 5 файлов, а 'сломал' 2 в другом конце проекта, как я это найду и как обеспечить наблюдаемость (observability)?"
- **Цель:** Определить механизмы контроля и отладки автономных систем.

12. Глубина RAG для Legacy (Часть 4)

- **Вопрос:** "Наш легаси — это 15-летний монолит на C++/Java с миллионами строк. Насколько реально RAG-система (Слайд 40) сможет дать осмысленный ответ по такой кодовой базе, а не просто 'найти по ключевым словам'?"
- **Цель:** Оценка реальной применимости RAG для крупномасштабных, сложных кодовых баз.

13. CI/CD и Автоматизация RAG-Оценки (Часть 3)

- **Вопрос:** "Можно ли инструменты вроде Ragas/DeepEval (Слайд 30) интегрировать в наш CI/CD-пайплайн? Или автоматизированная оценка RAG — это пока утопия, и все равно нужен человек, чтобы ставить 'лайк/дизлайк'?"
- **Цель:** Понять, можно ли автоматизировать контроль качества LLM-приложений.

14. Смешивание LLM (Новый)

- **Вопрос:** "Насколько архитектурно сложно **смешивать несколько моделей**? Например, использовать on-prem SLM для рефакторинга (Слайд 40) и облачный GPT для саммари PR (Слайд 43)? Не возникнет ли проблем с управлением контекстом и скоростью?"
- **Цель:** Оценка сложности построения гетерогенных архитектур (Multi-Model Architecture).

15. Standard Input/Output (Новый)

- **Вопрос:** "Вы упомянули MCP как 'USB-C для ИИ'. Насколько это готовый стандарт, и есть ли уже реальные библиотеки, которые гарантируют этот протокол, чтобы не изобретать собственный парсер и валидатор?"
- **Цель:** Выяснить практический статус готовности ключевых инструментов стандартизации.