

Памятка по проверке DeepSeek API-интеграции

Отдельный PDF для [downloads/checklist.pdf](#): короткая памятка для команды, которая подключает DeepSeek и хочет снизить риск ошибок из-за ручных лимитов и устаревших alias-ов.

Реестр моделей

- Проверить, где хранится список model id и alias-ов.
- Зафиксировать context length и max output для каждой используемой модели.
- Добавить цены input cache hit, input cache miss и output.
- Добавить concurrency limits, если они влияют на нагрузку.
- Отметить дату депрекации legacy-имён deepseek-chat и deepseek-reasoner.

Предвалидация

- Считать токенный бюджет до отправки запроса.
- Резервировать место под ответ.
- Ограничивать слишком длинные документы и историю диалога.
- Выбирать chunking, summarization или retrieval до запуска модели.
- Показывать пользователю предупреждение при слишком дорогой операции.

Финансы и эксплуатация

- Считать input и output отдельно.
- Логировать фактический расход токенов по моделям.
- Тестировать cache miss сценарии.
- Настроить лимиты на пользователя или проект.
- Проверить fallback на более дешёвую модель, если сценарий это допускает.

Агентные цепочки

- Проверить размер tool schemas и function arguments.
- Ограничить объём данных, возвращаемых инструментами в контекст.
- Тестировать неполный JSON и обрыв ответа.
- Логировать шаг, на котором агент упирается в контекст или max output.
- Не запускать длинную цепочку без проверки бюджета на первом шаге.