

Памятка для тестирования DeepSeek V4 перед внедрением

PDF-файл для [downloads/checklist.pdf](#): короткая рабочая памятка, которую можно приложить к внутреннему тесту DeepSeek V4, чтобы не смешивать свойства открытых весов, официального API и поведения конкретного endpoint.

Model ID и миграция

- Проверить, используются ли `deepseek-chat` и `deepseek-reasoner` в коде, конфигурациях, low-code сценариях и документации.
- Заменить legacy-имена на явные model ID для V4-Flash или V4-Pro.
- Отдельно зафиксировать режим Thinking или Non-Thinking для каждого теста.
- Проверить дату deprecation legacy-имен: 24 июля 2026 года, 15:59 UTC, по документации DeepSeek.

Паспорт API-теста

- Сохранять prompt, полный ответ и параметры запроса.
- Фиксировать endpoint, model ID, режим, дату и время теста.
- Записывать latency, output speed и повторяемость результата.
- Не делать вывод о GA-релизе только по изменению поведения API.

Экономика контекста

- Разделить сценарии, где нужен V4-Pro, и сценарии, где достаточно V4-Flash.
- Оценить долю cache hit и cache miss в реальном приложении.
- Проверить, не ломаются ли кэши из-за мелких изменений системных инструкций.
- Считать стоимость на типовой рабочей нагрузке, а не на одном демонстрационном prompt.

Проверяемость и governance

- Разделять тесты открытых весов, официального API и внешних endpoints.
- Для открытых весов фиксировать checkpoint, inference engine, quantization и конфигурацию оборудования.
- Для чувствительных данных проверить endpoint, правила обработки, журналирование и возможность локального deployment.
- Не переносить выводы о DeepSeek R1 или общих китайских моделях на V4 без отдельного источника.