

ДЛЯ СТО, ИТ-АРХИТЕКТОРОВ И СПЕЦИАЛИСТОВ ИБ

LLM для on-premise 2026: что реально поставить в контур

Открытые, коммерческие и российские модели на одном развороте: лицензия, доступность весов, память под модель, качество на русском и пригодность к закрытому контуру. Плюс расчёт видеопамяти, критерии выбора, типовые конфигурации железа и чек-лист перед пилотом.

12 позиций

Расчёт памяти

8 критериев

3 конфигурации

Чек-лист на 10 пунктов

Данные — по состоянию на 25 сентября 2026 года. Версии моделей, лицензии и форматы весов меняются за месяцы: перед внедрением сверяйтесь с карточкой модели на первоисточнике. Окончательный выбор делайте по пилоту на собственных документах, публичный лидерборд для этого не годится.

Как пользоваться

СОДЕРЖАНИЕ

- | | |
|-----------------------------------|------------------------------------|
| 1. Сравнительная таблица моделей | 5. Три типовые конфигурации |
| 2. Сколько памяти нужно модели | 6. Что выбирать под разные задачи |
| 3. Ориентиры и поправки к расчёту | 7. Чек-лист перед пилотом |
| 4. Восемь критериев выбора | 8. Подобрать модель под ваш контур |

Документ читается в два прохода. Сначала фильтр: таблица в разделе 1 показывает, у каких моделей открыты веса и какая у них лицензия. Всё, что не проходит по этим двум столбцам, для закрытого контура выпадает сразу, остальные столбцы для такой модели можно не читать.

Затем расчёт. Раздел 2 даёт формулу памяти и пример по шагам, раздел 3 — ориентиры по флагманам и две поправки, которые сдвигают результат на целый класс конфигурации. Раздел 5 показывает, во что этот расчёт превращается в железе: один сервер, узел на восемь ускорителей или кластер.

Разделы 6 и 7 нужны перед пилотом: подбор кандидатов под задачу и чек-лист из десяти пунктов. Его удобно пройти вместе с ИБ и юристом: половина пунктов касается лицензии, эксплуатации и того, как веса попадают в контур.

Для заполнения чек-листа от руки используйте печатную версию: светлый фон и увеличенные поля.

1 Сравнительная таблица моделей

Столбец «Память с запасом» посчитан по методике из раздела 2: веса в указанном формате × 1,4; под цифрой — класс конфигурации для ускорителей 80–141 ГБ. Для MoE-моделей в расчёт идут все параметры, активные на память не влияют. Цифры ориентировочные, ГБ и ТБ десятичные.

Модель	Лицензия	Веса	Параметры	Память с запасом и класс	Русский	Комментарий
GigaChat 3.5 Ultra Сбер	MIT	Открыты (HF, GitVerse)	432B / 28B активных, MoE	≈605 ГБ (FP8) узел на 8 ускорителей	Высокий	Выпущена 6 июля 2026, родной формат FP8; есть BF16, GGUF и базовый чекпоинт для дообучения
GigaChat 3 Lightning Сбер	MIT	Открыты (HF, GitVerse)	10B / 1,8B активных, MoE	≈28 ГБ (BF16) одна карта	Высокий	Компактная MoE, контекст до 256K
DeepSeek V4 Flash DeepSeek	MIT	Открыты (HF)	284B / 13B активных, MoE	≈225 ГБ (родной формат) сервер на 2–4 ускорителя	Хороший	Финальные веса 31.07.2026; эксперты в FP4, остальное в FP8; контекст до 1M
DeepSeek V4 Pro DeepSeek	MIT	Открыты (HF)	1,6T / 49B активных, MoE	≈1,2–1,25 ТБ (родной формат) кластер	Хороший	Финальные веса 13.08.2026; на одном узле 8 × 141 ГБ работает на пределе, с ограничением контекста
Qwen3.8-27B Alibaba	Apache 2.0	Открыты (HF, ModelScope)	27B, плотная	≈38 ГБ (FP8) / ≈76 ГБ (BF16) одна-две карты	Хороший	Выпущена 14.08.2026; есть официальная FP8-версия; принимает изображения и видео
Qwen3.8-2.4T Alibaba	Собственная лицензия Qwen3.8-Max	Открыты (HF)	2,4T / 95B активных, MoE	≈3,4 ТБ (FP8) кластер	Хороший	Лицензия с условиями для крупных AI-сервисов: юристу читать до выбора
Alice AI Foundation 80B Яндекс	Apache 2.0	Открыты (HF)	80B / 3B активных, MoE	≈224 ГБ (BF16) сервер на 2–4 ускорителя	— (базовая модель)	Претрейн без инструктивного обучения, для контура — только после дообучения; экспериментальная, выпущена в сентябре 2026
YandexGPT 5 Lite Яндекс	Собственная лицензия Яндекса	Открыты (HF)	8B, плотная	≈22 ГБ (BF16) одна карта	Высокий	Pretrain и Instruct, контекст 32K, дообучение через LoRA; условия лицензии — по карточке
Alice AI LLM Яндекс	Проприетарная	Закрыты	—	—	Высокий	Флагманская линейка Яндекса, API в Yandex AI Studio; развёртывание в контуре заказчика — по договору с вендором
Llama 3.3 70B / Llama 4 Meta	Llama Community License	Открыты (HF)	70B, плотная (Llama 4 — MoE)	≈98 ГБ (FP8) / ≈196 ГБ (BF16) для 70B одна-две карты в FP8	Средний	Порог 700 млн пользователей в месяц; при распространении — пометка «Built with Llama», производная модель должна называться с префиксом «Llama»; открытой в строгом смысле не считается
T-lite / T-pro Т-Технологии	Apache 2.0 (по карточке версии)	Открыты (HF)	От класса 8B	От ≈22 ГБ (8B, BF16)	Высокий	Дообучены на русском поверх Qwen; размер и лицензию сверять по конкретной версии
Mistral Mistral AI	Apache 2.0 для части линейки	Открыты частично	По версии	По версии	Средний	Часть моделей закрыта; лицензию смотреть по конкретной версии

Оценка русского — качественная, по данным карточек моделей и опыту пилотов; для выбора нужен замер на своих документах.

DeepSeek V4.1 Flash вышла 10 сентября 2026 года с новой архитектурой. В таблицу она не вошла: практика развёртывания только складывается, а часть её параметров, по сторонним разборам, может храниться вне памяти ускорителя.

Про GigaChat и YandexGPT. У обеих линеек есть корпоративные варианты развёртывания в инфраструктуре заказчика. Для GigaChat, кроме того, открыты веса флагмана и компактной модели под MIT.

2 Сколько памяти нужно модели

Расчёт укладывается в одну строку:

$$\text{Память} = \text{параметры} \times \text{байт на параметр} \times 1,4$$

Байт на параметр: BF16 — 2, FP8 — 1, 4 бита — около 0,5 плюс небольшие накладные расходы. Запас 1,4× закрывает KV-кэш, буферы движка и промежуточные вычисления при типовом корпоративном профиле: 20–30 одновременных сессий с окном до 32K токенов, из которых длинные документы одновременно держат немногие.

Главное правило для MoE

Маршрутизатор выбирает экспертов заново для каждого токена на каждом слое, и при двух десятках сессий за один шаг генерации задействуется почти весь их набор. Поэтому в памяти ускорителей должны лежать все веса модели. Активные параметры определяют скорость генерации и нагрузку на вычислители, на объём памяти они не влияют.

Посчитать память по активным — типовая ошибка в техзаданиях.

Пример: GigaChat 3.5 Ultra в FP8

- 1 Веса.** 432 млрд × 1 байт ≈ 432 ГБ.
- 2 С запасом.** 432 × 1,4 ≈ 605 ГБ.
- 3 Число карт.** 605 / 80 ≈ 7,6; 605 / 96 ≈ 6,3; 605 / 141 ≈ 4,3. Тензорный параллелизм обычно делит модель на 2, 4 или 8 частей, поэтому во всех трёх случаях выходит восемь карт.
- 4 Сверка.** Примеры запуска в карточке модели рассчитаны на восемь ускорителей. Восемь карт по 80 ГБ работают впритык, восемь по 141 ГБ оставляют запас под рост нагрузки.

Ваша модель: параметры _____ × байт _____ × 1,4 = _____ ГБ · память одной карты _____ ГБ · карт _____

3 Ориентиры и поправки к расчёту

Модель	Всего / активных	Веса в FP8	Веса в BF16	Опубликованные веса	Ориентир с запасом	Класс конфигурации
GigaChat 3.5 Ultra	432 млрд / 28 млрд	≈ 432 ГБ	≈ 864 ГБ	FP8 (родной), BF16, GGUF	≈ 605 ГБ (FP8)	Узел на 8 ускорителей; в BF16 — кластер
DeepSeek V4 Flash	284 млрд / 13 млрд	≈ 284 ГБ	≈ 568 ГБ	Эксперты FP4, остальное FP8; ≈ 160 ГБ	≈ 225 ГБ (родной формат)	Один сервер на 2– 4 ускорителя
DeepSeek V4 Pro	1,6 трлн / 49 млрд	≈ 1,6 ТБ	≈ 3,2 ТБ	Эксперты FP4, остальное FP8; ≈ 0,87–0,9 ТБ	≈ 1,2–1,25 ТБ (родной формат)	Кластер; узел 8 × 141 ГБ — на пределе
Плотная модель 20– 40B (например, Qwen3.8-27B)	20–40 млрд, активны все	≈ 20–40 ГБ	≈ 40–80 ГБ	BF16, FP8, квантованные сборки	≈ 28–56 ГБ (FP8), ≈ 56–112 ГБ (BF16)	Один сервер, 1– 2 ускорителя*

Ориентировочно, по состоянию на 25 сентября 2026 года. Класс конфигурации — для ускорителей H100 (80 ГБ), RTX PRO 6000 Blackwell (96 ГБ) и H200 (141 ГБ). Объём весов DeepSeek V4 — по размеру файлов чекпоинтов, округлённо. * Для плотных моделей на длинных документах KV-кэш считается отдельно, см. ниже.

Запас с избытком

У моделей со сжатым или линейным вниманием KV-кэш заметно компактнее классического. GigaChat 3.5 Ultra использует гибридное внимание (MLA и GatedDeltaNet), и KV-кэша на токен ей нужно примерно вчетверо меньше, чем GigaChat 3.1. DeepSeek V4 сжимает KV-кэш по длине последовательности и хранит его в FP8. Для этих моделей множитель 1,4× даёт запас сверху, но срезать его до нуля нельзя: память нужна движку под буферы и модули спекулятивного декодирования.

Запаса может не хватить

У плотной модели с классическим вниманием на длинных документах KV-кэш бывает больше самих весов. Формула на токен: $2 \times \text{число слоёв} \times \text{число KV-голов} \times \text{размерность головы} \times \text{байт на значение}$; все величины есть в config.json модели.

Условная плотная модель на 32 млрд параметров (64 слоя, 8 KV-голов, размерность 128, BF16) тратит 256 КБ на токен. 25 сессий по 32К — около 215 ГБ KV-кэша при 64 ГБ весов. Для такой модели множитель 1,4× покрывает в среднем около 4К токенов на сессию — для чата хватает, для договоров целиком нет.

Формат весов и ускоритель

Аппаратная поддержка FP8 есть у поколений Ada (L40S), Hopper (H100, H200) и Blackwell, FP4 — только у Blackwell. На Hopper эксперты DeepSeek V4 в FP4 обрабатываются через преобразование на лету или используется полностью FP8-сборка, которая почти вдвое тяжелее. У A100 нет аппаратной поддержки FP8. Какой режим поддерживает выбранная версия движка на выбранной карте, проверяется до закупки.

Подробный разбор методики — в статье [«Сколько видеопамати нужно LLM в своём контуре»](#).

4 Восемь критериев выбора

Критерии неравнозначны. Первые два работают как фильтр: без открытых весов и приемлемой лицензии модель выпадает из списка, остальные пункты для неё можно не смотреть. Дальше идёт взвешивание компромиссов между теми, кто фильтр прошёл.

1. Доступность весов ФИЛЬТР

Без открытых весов закрытый контур невозможен в принципе.

2. Лицензия ФИЛЬТР

Арасhe 2.0 и MIT против собственных лицензий с порогами и условиями. У части открытых флагманов, включая Llama и Qwen3.8-2.4T, лицензия собственная.

3. Русский язык

Деловой текст, терминология, юридические формулировки.

4. Память и железо

Считается по всем весам модели и её родному формату, с запасом под одновременные сессии и длину документов.

5. Стоимость эксплуатации

Полная стоимость владения железом и людьми на горизонте 1–2 лет, цена токена здесь не показатель.

6. Совместимость с RAG

Длина контекста, следование инструкциям, стабильность формата вывода.

7. Поддержка дообучения

LoRA и QLoRA, доступность базового чекпоинта и тренировочного стека.

8. Зрелость для эксплуатации

Поддержка в vLLM и SGLang в основной ветке, мониторинг, сообщество, регулярность обновлений.

5 Три типовые конфигурации

Классы построены по той же методике расчёта памяти. Какая модель попадает в какой класс, зависит от её формата весов и объёма памяти выбранного ускорителя.

Small Пилот и ограниченный RAG

Модели	GigaChat 3 Lightning, YandexGPT 5 Lite, компактные T-lite и Qwen.
Железо	Один сервер, один ускоритель на 48–96 ГБ.
Плюсы	Быстрый старт, недорого, модель легко обновить.
Ограничения	Небольшая пропускная способность; на сложных запросах компактная модель ошибается чаще.
Что проверить	Хватает ли качества на ваших документах при таком размере модели.

Medium Корпоративный ассистент

Модели	Qwen3.8-27B, T-pro, Llama 3.3 70B в FP8; DeepSeek V4 Flash — на верхней границе класса.
Железо	Один сервер на 2–4 ускорителя, отдельный узел под векторную БД.
Плюсы	Баланс качества и стоимости, реальная многопользовательская нагрузка.
Ограничения	Нужен MLOps — мониторинг, обновления, дежурство. Для плотной модели на длинных документах память под KV-кэш считается отдельно.
Что проверить	Кто эксплуатирует контур после пилота и по каким SLA.

Enterprise Флагманские модели

Модели	GigaChat 3.5 Ultra — узел на 8 ускорителей; DeepSeek V4 Pro и Qwen3.8-2.4T — кластер.
Железо	Узел на 8 ускорителей по 80–141 ГБ или несколько узлов с межузловой сетью InfiniBand или RoCE, резервирование.
Плюсы	Максимальное качество, несколько сценариев на одной платформе.
Ограничения	Дорого по железу и людям. Плотный узел на 8 ускорителей предъявляет требования к стойке, питанию и охлаждению, которые часто выясняются только на этапе размещения.
Что проверить	Готовность серверной и команды; как веса объёмом от 0,4 до 2,4 ТБ заносятся в контур, проверяются и хранятся.

6 Что выбирать под разные задачи

Класс модели определяется задачей. Под классификацию обращений нет смысла поднимать МоЕ на сотни гигабайт: компактная модель справится, и на потоке она обойдётся кратно дешевле.

Задача	Подход	Кандидаты	Комментарий
Документооборот, извлечение реквизитов	RAG + дообучение под формат	GigaChat 3 Lightning / Qwen3.8-27B / T-pro	Важна стабильность формата вывода, размер модели вторичен
Классификация обращений	Малая модель + few-shot или дообучение	GigaChat 3 Lightning / YandexGPT 5 Lite	Массовая задача, чувствительна к стоимости на потоке
Корпоративный поиск	RAG с упором на эмбединги и reranker	Любая с хорошим следованием контексту	Качество определяет в первую очередь ретривер
Диалоговый ассистент	RAG + длинный контекст	Qwen3.8-27B / GigaChat 3.5 Ultra	Нужны уверенный русский и вызов инструментов
Генерация и ревью кода	Специализированная или сильная общая модель	Qwen3.8-27B / DeepSeek V4 Flash	Русский вторичен, решает качество кода
Суммаризация длинных документов	Длинный контекст + разбиение на фрагменты	DeepSeek V4 Flash / GigaChat 3.5 Ultra	Сжатое внимание экономит память на длинном контексте
Договорная аналитика	RAG + дообучение на разметке	GigaChat 3.5 Ultra / Qwen3.8-27B + LoRA	Цена ошибки высокая, ответ проверяет юрист
ИБ-аналитика журналов	Компактная модель + строгие промпты	Qwen3.8-27B / GigaChat 3 Lightning	Закрытый контур обязателен, телеметрия недопустима

Почти в каждой строке задача решается связкой «модель + RAG». Качество ответа там упирается в поиск и состояние архива: если нашлась устаревшая версия документа, флагман ответит так же неверно, как компактная модель.

7 Чек-лист перед пилотом

Отметьте закрытые пункты. Каждый незакрытый стоит закрыть до закупки оборудования.

- ☐ 1 Веса модели доступны для скачивания и развёртывания в вашем контуре.
- ☐ 2 Лицензию прочитал юрист, особенно пункты о коммерческом использовании, порогах и дообучении.
- ☐ 3 2–3 модели прогнаны на реальной выборке своих документов — с грязными сканами и старыми версиями файлов.
- ☐ 4 Посчитана полная стоимость владения на горизонте 1–2 лет, включая железо, людей и дежурство.
- ☐ 5 Память посчитана по всем весам модели, а не по активным параметрам, с запасом под ваш профиль нагрузки; для плотной модели на длинных документах KV-кэш посчитан отдельно.
- ☐ 6 Родной формат весов (BF16, FP8, FP4) и его поддержка движком на выбранных ускорителях проверены до закупки.
- ☐ 7 Понятно, кто эксплуатирует контур после пилота и по каким SLA.
- ☐ 8 Проверена совместимость с RAG-стеком и решено, нужно ли дообучение или задачу закрывает качественный поиск.
- ☐ 9 Учтены требования ИБ: отсутствие телеметрии, изоляция, аудит, аттестация при необходимости. Веса и сборки движка занесены в контур, сверены по контрольным суммам и зафиксированы по версиям.
- ☐ 10 Заложен регламент пересмотра: кто и как часто сверяет новые релизы и решает, переходить ли на них.

Кандидаты для пилота:

Ответственный за чек-лист и дата:

Подобрать модель и конфигурацию под ваш контур

Таблица помогает сузить список до двух-трёх кандидатов. Дальше решают ваши документы, нагрузка и требования ИБ.

AZONE-AI — направление корпоративных ИИ-решений компании «Азон», российского ИТ-интегратора с лицензиями ФСТЭК, ФСБ и Минобороны. Специализация — ИИ в закрытых и защищённых контурах: там, где данные не могут покинуть периметр.

Экспресс-аудит: сценарии, пиковая нагрузка, длина документов, требования к среде исполнения. По итогам — расчёт конфигурации под выбранную модель и архитектура пилота в вашем контуре, при необходимости — в аттестованной среде.

[Обсудить подбор модели и пилот](#)

[Сколько видеопамати нужно LLM в своём контуре](#)

[Open-source и коммерческие LLM: что выбрать для on-premise](#)

[Сравнение отечественных LLM](#)

Актуальность. Редакция 09.2026, данные по состоянию на 25 сентября 2026 года, плановый пересмотр — 12.2026. Версии моделей, лицензии и форматы весов меняются за месяцы: перед внедрением сверяйтесь с карточкой модели на первоисточнике. Сведения о лицензиях не являются юридическим заключением. Расчёты памяти ориентировочные и не заменяют нагрузочного теста на ваших данных.

Что изменилось в редакции 09.2026. Вместо GigaChat 3.1 Ultra в таблице GigaChat 3.5 Ultra, DeepSeek V4 разделён на Flash и Pro с финальными весами, Qwen обновлён до поколения 3.8. У Яндекса появилась открытая базовая модель Alice AI Foundation. Столбец памяти пересчитан по единой методике, и ей отведён отдельный разворот: в прошлой редакции память у MoE-моделей была указана ориентирами без расчёта. Конфигурации железа и чек-лист приведены в соответствие с этой методикой. Уточнены условия лицензии Llama: с версии 3.1 выход модели можно использовать для обучения других моделей при соблюдении требований к названию.