

ШАБЛОН ДЛЯ ИТ, ИБ, ЗАКУПОК И ВЛАДЕЛЬЦЕВ ПРОЦЕССОВ

Шаблон ТЗ на локальную LLM

Техническое задание без противоречий: формулировки, поля для заполнения, самопроверка перед публикацией закупки.

1 Разрядность и ускоритель

2 Время ответа и цепочка

3 Бюджет и конфигурация

4 Пилот и весь штат

5 Аттестация и сроки

Как пользоваться шаблоном

Шаблон удобнее заполнять не по порядку. Начните с разделов 2 и 3: сценарии и нагрузка определяют всё остальное. Потом раздайте разделы по таблице ниже и соберите всех на одну встречу, пройдя Приложение А вслух. Большинство противоречий обнаруживается как раз на этой встрече, когда ИТ впервые видит сумму, а финансы — конфигурацию.

Делать это лучше до публикации закупки: изменения после неё ограничены 44-ФЗ и 223-ФЗ. Порядок этапов и формулировки для конкретной закупки согласуйте со службой закупок и юристом.

Кто что заполняет

Раздел шаблона	Кто заполняет	С кем сверить
1. Общие сведения	Руководитель проекта	Служба закупок
2. Назначение и сценарии	Владелец процесса	ИТ
3. Профиль нагрузки	Владелец процесса и ИТ	—
4. Требования к модели	ИТ	Служба ИБ, юрист
5. Требования к качеству	Владелец процесса	ИТ
6. Требования к производительности	ИТ	Владелец процесса
7. Среда исполнения и защита информации	Служба ИБ	ИТ
8. Этапы работ	Руководитель проекта	Служба ИБ, служба закупок
9. Приёмка	ИТ	Владелец процесса
10. Обновление модели и сопровождение	ИТ	Служба ИБ
11. Права на результаты	Юрист	ИТ
12. Бюджет	Финансы	ИТ

Как читать шаблон

- Текст в рамке с пометкой **Формулировка для ТЗ** можно переносить в документ как есть, подставив свои значения.
- _____ — поле, которое заполняет заказчик.
- Буквы Р, N, K, X, T1, T2, T3, D — числа, которые заказчик задаёт сам. Их расшифровка ниже.
- Подсказки — **Зачем**, **Типичная ошибка**, **Сроки** — в ТЗ не переносят.
- Номер противоречия в скобках отсылает к Приложению А: там перечислены все пять и вопросы самопроверки по каждому.

Переменные

Обозначение	Что означает	Единица
P	Минимальная доля правильных ответов на тестовом наборе	%
N	Число вопросов в тестовом наборе	вопросов
K	Число одновременных сессий, при котором проверяются требования	сессий
X	Минимальная скорость генерации на одну сессию	токенов в секунду
T1	Предельное время до первого токена для справочных запросов	секунд
T2	Предельное время выполнения многошагового сценария	секунд
T3	Срок этапа аттестации	недель
D	Срок передачи исходных данных заказчиком на этапе аттестации	рабочих дней

1

Общие сведения

ФОРМУЛИРОВКА ДЛЯ T3

Наименование работ: _____

Основание для проведения работ: _____

Заказчик: _____

Место размещения оборудования: _____

Контур размещения: ☐ изолированный, без доступа к интернету ☐ с ограниченным доступом ☐ иное: _____

Ответственное лицо заказчика: _____

Зачем: в изолированный контур веса модели заносят на носители, а движки собирают заранее и фиксируют версии. Это недели работы, которые исполнитель заложит в срок, только если знает о них.

2

Назначение и сценарии

ФОРМУЛИРОВКА ДЛЯ T3

Система предназначена для _____.

Перечень сценариев использования приведён в таблице 2.1. Требования к производительности (раздел 6) применяются к каждому сценарию в соответствии с его типом.

Таблица 2.1. Сценарии использования

№	Сценарий	Тип: справочный / многошаговый	Подразделения	Источники данных	Внешние системы, к которым обращается сценарий
1					
2					
3					
4					
5					

Справочный запрос — поиск по документам и один вызов модели. Многошаговый сценарий — цепочка из нескольких вызовов модели, поиска и обращений к другим системам, в которой каждый шаг зависит от результата предыдущего.

Зачем (противоречие 2): от типа сценария зависит, какое требование ко времени ответа к нему применимо. Один лимит на все сценарии многошаговые не выдержат.

Типичная ошибка: сценарий описан как «ассистент по документам», а на деле предполагает обращение к учётной системе и СЭД. Внешние системы стоит перечислить сразу: их время ответа исполнитель не контролирует.

3

Профиль нагрузки

ФОРМУЛИРОВКА ДЛЯ ТЗ

Число пользователей:
— зарегистрированных в системе: _____
— активных в течение рабочего дня: _____
— работающих одновременно в пиковый час: _____
Число запросов в пиковый час: _____
Параметры сценариев приведены в таблице 3.1. Объёмы входных и выходных токенов указываются отдельно для одного обращения к модели и для пользовательского запроса целиком.

Таблица 3.1. Параметры сценариев

№ сценария	Вызовов модели на запрос: среднее / максимум	Входных токенов на один вызов	Выходных токенов на один вызов	Входных токенов на запрос целиком	Выходных токенов на запрос целиком	Длина документов в контексте
1	/					
2	/					
3	/					
4	/					
5	/					

Зачем (противоречие 4): оборудование считают по числу одновременно работающих пользователей, а в ТЗ чаще стоит число учётных записей. Разница кратная.

Типичная ошибка (противоречие 2): объём токенов указан «на запрос», и непонятно, к чему он относится. Мы видели ТЗ, где стояло около десятка вызовов модели и несколько сотен выходных токенов при лимите примерно в десять секунд. Если токены относились к каждому вызову, одна генерация занимала под минуту. Выяснилось это только при расчёте коммерческого предложения.

4 Требования к модели

ФОРМУЛИРОВКА ДЛЯ ТЗ

Языковая модель должна соответствовать следующим требованиям:

- основной язык запросов и ответов: _____;
- длина контекста: не менее _____ токенов;
- лицензия допускает коммерческое использование модели в контуре заказчика для целей, указанных в разделе 2;
- запросы, документы и ответы не передаются за пределы контура;
- модель разворачивается и работает без доступа к интернету;
- статус модели по 243-ФЗ: ☐ не требуется ☐ требуется: _____ (заполняется при наличии обоснования).

Конкретную модель предлагает исполнитель с обоснованием выбора. Соответствие модели требованиям к качеству подтверждается на тестовом наборе (раздел 5). Формат весов определяет исполнитель.

Зачем: название модели, выбранное на этапе ТЗ, устаревает раньше, чем заканчивается закупка: открытые линейки обновляются за месяцы. Требования к свойствам модели переживут смену поколения.

Типичная ошибка (противоречие 1): требование «запуск модели в BF16 без квантования». На тех же ускорителях модель в BF16 оставляет меньше памяти под одновременные сессии, генерирует медленнее, а иногда не помещается вовсе. У части актуальных моделей родной формат — FP8, и BF16-версия нужна им прежде всего для дообучения. Качество защищают требованием из раздела 5.

Типичная ошибка: требовать статус по 243-ФЗ в ТЗ, которое исполняется до 1 марта 2027 года. Нормы о статусах вступают в силу с этой даты, и раньше подтвердить статус нечем.

5 Требования к качеству

ФОРМУЛИРОВКА ДЛЯ ТЗ

Качество ответов проверяется на тестовом наборе заказчика из N вопросов. Доля правильных ответов — не ниже P%.

Тестовый набор формирует заказчик. Набор содержит вопросы из реальной работы подразделений, эталонные ответы и документы, на которые эти ответы опираются. В набор включаются:

- типовые вопросы;
- трудные вопросы, ответ на которые требует нескольких документов;
- вопросы, правильный ответ на которые — «в предоставленных документах этого нет».

Открытая часть набора (_____ % вопросов) передаётся исполнителю на этапе _____.

Закрытая часть (_____ % вопросов) предъявляется только на приёмочных испытаниях.

Правильность ответа оценивает комиссия заказчика по критериям:

_____.

При использовании формата весов с пониженной разрядностью исполнитель на приёмке подтверждает, что качество на тестовом наборе не ниже порога P%.

Зачем: без тестового набора требования к качеству не проверить, и приёмка сводится к впечатлениям. Собирает набор заказчик. Больше никому: только он знает, какой ответ правильный.

Типичная ошибка: весь набор передан исполнителю в начале проекта. Система настраивается под известные вопросы, и приёмка показывает качество на наборе, а не в работе. Закрытая часть это исправляет. Форма набора — в Приложении Б.

6 Требования к производительности

ФОРМУЛИРОВКА для ТЗ

Для справочных запросов (поиск и один вызов модели) время до первого токена — не более $T1$ секунд для 95% запросов при K одновременных сессиях.

Скорость генерации — не менее X токенов в секунду на одну сессию при K одновременных сессиях.

Для многошаговых сценариев время выполнения — не более $T2$ секунд для 95% сценариев каждого согласованного типа. Время ответа внешних систем в это значение не включается и фиксируется в журнале отдельно.

Типы многошаговых сценариев и значения $T2$ для каждого из них:

Замер проводится на стенде заказчика по методике, согласованной сторонами до начала испытаний.

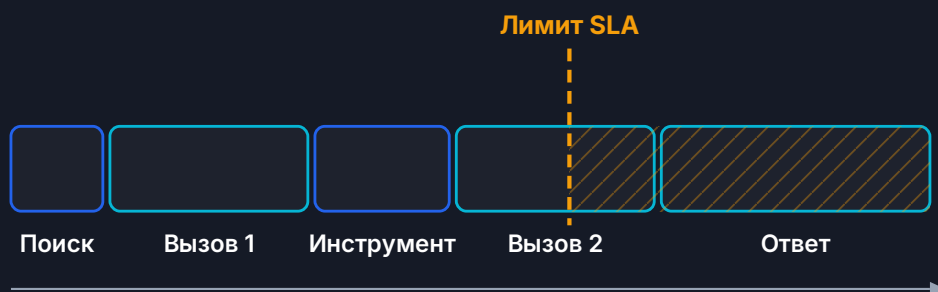


Рисунок 1. В многошаговом сценарии время шагов складывается, и цепочка выходит за общий лимит.

Зачем (противоречие 2): в многошаговом сценарии время шагов складывается — поиск, вызов модели, обращение к внешней системе, второй вызов, итоговый ответ. Дополнительные ускорители увеличивают прежде всего число одновременных сессий, одну последовательную цепочку они почти не ускоряют. Не знаете, каким должно быть $T2$? Зафиксируйте его по итогам нагрузочного теста на пилоте.

Типичная ошибка: «время ответа — не более столько-то секунд» без перцентиля и без уточнения, что считать временем ответа. При потоковой выдаче пользователь видит первые слова задолго до конца генерации, и для него важнее время до первого токена. Без перцентиля один медленный запрос в пиковый час проваливает испытания целиком.

7

Среда исполнения и защита информации

ФОРМУЛИРОВКА для ТЗ

Требования к операционной системе: _____

Обоснование требования: _____

Требования к оборудованию (в том числе к происхождению): _____

Обоснование требования: _____

Исполнитель до поставки подтверждает совместимость драйверов ускорителей и предлагаемой версии программного обеспечения для запуска модели с указанной операционной системой. Способ подтверждения: ☐ протокол испытаний на стенде исполнителя ☐ иное:

_____.

Ответственность за совместимость несёт _____.

Система ведёт журнал обращений и ответов модели. Состав журнала: _____. Срок хранения: _____. Журнал передаётся в _____.

Аттестация объекта информатизации: ☐ требуется ☐ не требуется. Решение службы ИБ от _____ № _____.

Зачем: требование к сертифицированной ОС и требование к актуальным версиям движков по отдельности выполнимы. Вместе — только после проверки, и провести её нужно до поставки: на монтаже будет поздно.

Типичная ошибка: аттестация вписана в ТЗ по аналогии с соседним проектом. Нужна ли она, определяется категорией обрабатываемой информации и требованиями к конкретной системе, и решение принимается до ТЗ вместе со службой ИБ.

8

Этапы работ

ФОРМУЛИРОВКА для ТЗ

Работы выполняются в этапы, приведённые в таблице 8.1. Каждый следующий этап начинается после подписания акта о завершении предыдущего.

Таблица 8.1. Этапы работ

№	Этап	Содержание	Срок	Результат
1	Пилот	Подразделения: _____ Сценарии: _____	_____	Протокол пилота с измеренными метриками
2	Развёртывание промышленной конфигурации	Конфигурация рассчитывается по профилю нагрузки, измеренному на этапе 1	_____	Акт о завершении развёртывания
3	Аттестация объекта информатизации (если требуется)	См. формулировку ниже	Т3 недель	Проведённые аттестационные испытания и решение по их итогам
4	Масштабирование (при необходимости)	Подразделения: _____	_____	_____

ФОРМУЛИРОВКА ДЛЯ Т3: МЕТРИКИ ПЕРЕХОДА ОТ ПИЛОТА

Переход к этапу 2 осуществляется при одновременном выполнении условий:

- доля правильных ответов на тестовом наборе — не ниже P%;
- с системой регулярно работает не менее _____% пилотной группы;
- профиль нагрузки измерен и зафиксирован в протоколе пилота: пиковое число сессий, длина запросов и документов в контексте.

Если условия не выполнены, стороны согласуют _____.

ФОРМУЛИРОВКА ДЛЯ Т3: ЭТАП АТТЕСТАЦИИ

Этап 3. Аттестация объекта информатизации. Начинается после подписания акта о завершении этапа 2 (развёртывание). Срок этапа — Т3 недель. Исходные данные предоставляет заказчик по перечню из приложения к Т3 в течение D рабочих дней с начала этапа; при задержке срок этапа продлевается на период задержки. Результат этапа — проведённые аттестационные испытания и решение по их итогам.

Зачем (противоречие 4): пилот и работа на весь штат — разные этапы. До пилота неизвестны ни реальная нагрузка, ни то, какая модель нужна.

Типичная ошибка: «по результатам пилота принимается решение о масштабировании» без порогов. Фраза ничего не фиксирует, и решение о переходе превращается в спор о впечатлениях.

Сроки (противоречие 5). По оценке AZONE-AI, документация на объект — около восьми недель, аттестация — ещё три-пять. Испытания идут на системе в рабочей конфигурации, поэтому раньше развёртывания их не начать. Втиснуть всё это в короткий общий срок контракта не получится даже при добросовестной работе. Положительный результат аттестации заранее гарантировать нельзя, и обещать его в Т3 не стоит.

Типичная ошибка: пилот запланирован на реальных данных, для которых нужна аттестация, а сама аттестация стоит этапом 3. Тогда пилот работает на обезличенных данных или пилотный стенд сразу входит в аттестуемый контур. Решают это до ТЗ.

9 Приёмка

ФОРМУЛИРОВКА ДЛЯ ТЗ

Приёмочные испытания проводятся на стенде _____ в присутствии представителей заказчика: _____ и исполнителя: _____.

Методика испытаний согласуется сторонами не позднее _____ до начала испытаний.

В состав испытаний входят:

- проверка качества на тестовом наборе, включая закрытую часть (раздел 5);
- проверка производительности по показателям раздела 6;
- проверка журналирования (раздел 7);
- _____.

По итогам испытаний оформляется _____, подписываемый _____.

Типичная ошибка: производительность проверяют при одной сессии. Цифры выходят красивые и ничего не говорят о пиковом часе — требования раздела 6 проверяются при К одновременных сессиях.

10 Обновление модели и сопровождение

ФОРМУЛИРОВКА ДЛЯ ТЗ

Решение о переходе на новую версию модели принимает _____.

Перед переходом новая версия проверяется на тестовом наборе раздела 5. Переход допускается при доле правильных ответов не ниже Р%.

Повторную приёмку проводит _____.

Для аттестованного сегмента смена версии модели рассматривается как изменение программной среды, которое может потребовать повторной оценки. Порядок согласования: _____.

Срок сопровождения: _____. Состав сопровождения: _____.

Зачем: за срок сопровождения у выбранной модели почти наверняка выйдет новая версия. Без описанного порядка её либо не ставят вовсе, либо ставят без проверки. Второе хуже.

11 Права на результаты

ФОРМУЛИРОВКА ДЛЯ ТЗ

Наборы данных, подготовленные для дообучения модели, принадлежат _____.

Веса модели и адаптеры, полученные в результате дообучения, принадлежат _____.

Исполнитель ☐ вправе ☐ не вправе использовать наборы данных, веса и адаптеры, указанные выше, в других проектах.

После окончания срока действия договора наборы данных, веса и адаптеры ☐ передаются заказчику ☐ уничтожаются исполнителем с оформлением _____ ☐ иное: _____.

Раздел нужен, если ТЗ предусматривает дообучение. Без этих строк права определяются по умолчанию — и не обязательно в пользу заказчика.

12 Бюджет

Выберите один режим и отметьте его. Выбор делается до утверждения суммы.

ФОРМУЛИРОВКА ДЛЯ ТЗ

☐ **Фиксированный бюджет.** ТЗ описывает требуемый результат: сценарии (раздел 2), профиль нагрузки (раздел 3), качество (раздел 5), производительность (раздел 6). Конфигурацию оборудования предлагает исполнитель и обосновывает расчётом.

☐ **Фиксированная конфигурация.** Конфигурация приведена в приложении _____. Стоимость оценена по коммерческим предложениям не менее _____ поставщиков, с учётом работ, сопровождения и НДС. Дата оценки: _____.

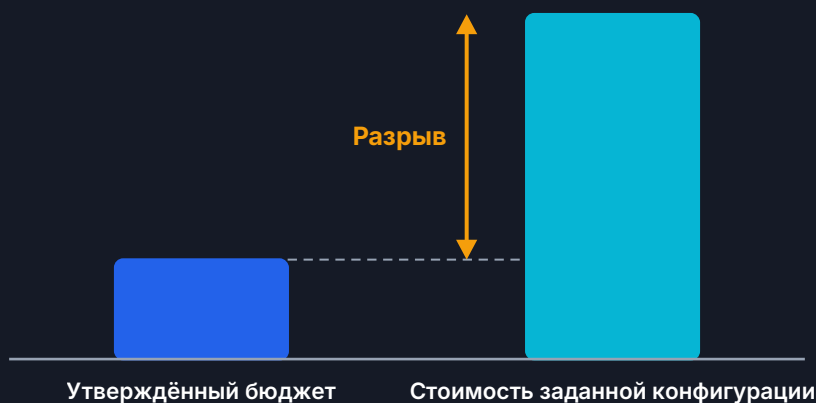


Рисунок 2. Сумма и конфигурация, определённые в разное время, расходятся в разы.

Зачем (противоречие 3): сумму и конфигурацию обычно определяют в разное время и разные люди, и расходятся они в разы. Закупка с таким разрывом либо не собирает заявок, либо расхождение всплывает при исполнении.

Типичная ошибка: сумма посчитана только по оборудованию, а ТЗ включает монтаж, настройку, интеграции, аттестацию и сопровождение. Или посчитана без НДС, который с 1 января 2026 года составляет 22%. Цены на ускорители меняются быстро, поэтому оценку обновляют непосредственно перед расчётом начальной цены.

Приложение А Самопроверка перед публикацией закупки

Пять противоречий, которые мы чаще всего находим в ТЗ на локальную LLM:

- 1. **Разрядность и ускоритель** — «BF16 без квантования» рядом с фиксированным числом карт и требованием к скорости.
- 2. **Время ответа и длина цепочки** — многошаговые сценарии и единый лимит времени; неясно, к чему относятся объёмы токенов.
- 3. **Бюджет и конфигурация** — заданная конфигурация и сумма, которая её не покрывает.
- 4. **Пилотный сервер и весь штат** — сервер на пилотную группу и требование обслуживать всех сотрудников.
- 5. **Аттестация внутри общего срока** — аттестация включена в короткий срок контракта.

Отдельно — совместимость среды исполнения: сертифицированная ОС и оборудование из реестра против актуальных версий движков.

№	Вопрос	Где в шаблоне	Да	Нет
ПРОТИВОРЕЧИЕ 1				
1	Формат весов (BF16, FP8, квантование) не задан требованием, а выбирается исполнителем?	Раздел 4	☐	☐
2	Качество задано порогом Р% на тестовом наборе, и исполнитель подтверждает его на приёмке при любом формате весов?	Раздел 5	☐	☐
ПРОТИВОРЕЧИЕ 2				
3	Для каждого типа сценария указано, к чему относится объём токенов — к одному вызову модели или к запросу целиком?	Раздел 3	☐	☐
4	Требования ко времени заданы отдельно для справочных запросов и многошаговых сценариев, через перцентиль и при заданном числе сессий?	Раздел 6	☐	☐
ПРОТИВОРЕЧИЕ 3				
5	Выбран один режим бюджета: фиксированный бюджет или фиксированная конфигурация?	Раздел 12	☐	☐
6	Сумма посчитана с НДС, работами и сопровождением, и дата оценки не старше _____?	Раздел 12	☐	☐
ПРОТИВОРЕЧИЕ 4				
7	Указаны три разных числа пользователей: зарегистрированные, активные за день, одновременные в пиковый час?	Раздел 3	☐	☐
8	Пилот выделен в отдельный этап, и переход к следующему этапу зависит от метрик с порогами?	Раздел 8	☐	☐

№	Вопрос	Где в шаблоне	Да	Нет
ПРОТИВОРЕЧИЕ 5				
9	Решение о необходимости аттестации принято службой ИБ до ТЗ?	Раздел 7	☐	☐
10	Срок аттестации отделён от срока развёртывания и начинается после него?	Раздел 8	☐	☐
СОВМЕСТИМОСТЬ				
11	Для требований к ОС и оборудованию указано обоснование?	Раздел 7	☐	☐
12	Определено, кто и как подтверждает совместимость драйверов и версии движка с ОС до поставки?	Раздел 7	☐	☐
ЧТО ОБЫЧНО ВЫПАДАЕТ				
13	Тестовый набор собран, и часть его закрыта до приёмки?	Раздел 5, Приложение Б	☐	☐
14	Описан порядок перехода на новую версию модели?	Раздел 10	☐	☐
15	Если предусмотрено дообучение, права на данные, веса и адаптеры распределены?	Раздел 11	☐	☐

Любое «нет» в вопросах 1–10 — это противоречие, которое после публикации закупки обойдётся дороже. «Нет» в вопросах 11–15 можно закрыть и позже, но лучше до подписания контракта.

Приложение Б

Форма тестового набора

Тип вопроса: **Т** — типовой, **С** — трудный, ответ требует нескольких документов, **Н** — правильный ответ «в документах этого нет».

Доступ: **О** — открытый, передаётся исполнителю, **З** — закрытый до приёмки.

№	Вопрос	Тип (Т / С / Н)	Эталонный ответ	Документ-источник	Доступ (О / З)
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					

Покрытие набора важнее его размера. Вопросы типа «Н» проверяют то, что пользователи замечают первым: придумывает ли система ответ, когда в документах его нет.

Типичная ошибка: эталонный ответ записан без ссылки на документ. На приёмке становится непонятно, ответ неверен или устарел сам документ.

Приложение В Глоссарий

TTFT (время до первого токена). Сколько пользователь ждёт первого слова ответа. При потоковой выдаче о скорости системы судят по нему.

P95 (95-й перцентиль). Значение, в которое укладываются 95% измерений. Требование через перцентиль допускает единичные медленные запросы и не проваливает испытания из-за одного пика.

Одновременные сессии. Сколько запросов система обрабатывает в один момент. Из этого числа считают оборудование.

Скорость генерации. Токенов в секунду в одной сессии. Падает с ростом числа одновременных сессий, поэтому без K в требовании эта цифра мало что значит.

SLA. Согласованные показатели работы системы — время ответа, скорость, доступность. В шаблоне это требования раздела 6.

KV-кэш. Память ускорителя под уже прочитанный текст каждой сессии. Растёт с длиной контекста и числом сессий; на длинных документах бывает больше самих весов.

BF16 и FP8. Форматы хранения весов модели: два байта и один байт на параметр. От формата зависят объём памяти, скорость генерации и, в разной степени для разных моделей, качество.

Многошаговый (агентный) сценарий. Цепочка последовательных шагов: модель, поиск, внешняя система, снова модель. Время шагов складывается.

Тестовый набор. Вопросы из реальной работы с эталонными ответами и документами-источниками, по которым проверяется качество системы на приёмке и при обновлении модели.

Аттестация объекта информатизации. Проверка системы на соответствие требованиям по защите информации; по итогам выдаётся аттестат соответствия. Нужна не каждой системе: решение зависит от категории обрабатываемой информации.

Проверка ТЗ до публикации

AZONE-AI может проверить ваше ТЗ до публикации закупки. Обычно хватает одной встречи, чтобы понять, какие из пяти противоречий в нём есть.

azoneai.ru/contact/?product=other