ШАБЛОН ДЛЯ АРХИТЕКТОРОВ, РУКОВОДИТЕЛЕЙ ПРОЕКТОВ И ЭКСПЕРТОВ

Шаблон тестового набора для выбора LLM

Заготовки на неделю: от списка кандидатов до протокола решения

Публичные рейтинги помогают составить список кандидатов, но не говорят, какая модель лучше справится с вашими регламентами, договорами и обращениями сотрудников. Это проверяют на своих документах: набор из 50–100 реальных вопросов, слепая оценка двумя экспертами и решение по ошибкам, которые критичны для вашего бизнеса. Документ даёт заготовки, которые заполняются по ходу недели.

ПН–ВТ
Тестовый набор
Реальные вопросы, эталоны от будущих пользователей, вопросы-ловушки

СР–ЧТ
Прогон и слепая оценка
Один стенд для всех кандидатов, два эксперта, шкала из трёх градаций

ПТ
Разбор и решение
Ловушки и критичные ошибки весят больше среднего балла

Паспорт выбора | Отсев кандидатов | Шаблон набора | Бланк слепой оценки | Протокол решения

Шаблон не привязан к конкретным моделям. Версии и лицензии меняются за месяцы, а набор вопросов живёт дольше: его прогоняют заново при каждом обновлении модели. Строки с пометкой «пример» вымышлены и показывают, как заполнять поля.

Как пользоваться

Неделя теста держится на дисциплине: вопросы собирают из реальных обращений, набор замораживают до прогона, ответы оценивают вслепую, а решение записывают с обоснованием. Разделы документа идут в том порядке, в котором их заполняют.

Электронная таблица. Набор, ответы, бланки оценки, ключ и сводная ведутся в таблице с теми же разделами — сводная считается в ней автоматически. Скачать: azoneai.ru/whitepapers/40/AZONE-AI_test-nabor-vybor-llm_2026.xlsx. PDF остаётся инструкцией и бланком для протокола решения.

Начинайте с паспорта и отсева кандидатов — до понедельника. К началу недели стенд с кандидатами должен быть готов: развёртывание флагманских MoE-моделей в неделю теста не входит и иногда занимает больше времени, чем сам тест.

Роль	Что делает	Разделы
Руководитель проекта	Ведёт паспорт, сроки и протокол решения	Паспорт выбора, протокол решения
Архитектор	Отсев по памяти, стенд, прогон, замер скорости	Отсев кандидатов, протокол прогона
Специалист по ИБ	Допустимость в контуре, условия стенда	Отсев кандидатов
Два эксперта из будущих пользователей	Вопросы, эталоны, слепая оценка	Шаблон набора, бланк оценки
Хранитель ключа	Обезличивает ответы и хранит ключ. Сам ответы не оценивает	Ключ соответствия

День	Что делаем	Раздел шаблона	Кто
До понедельника	Паспорт, отсев кандидатов, стенд	Паспорт выбора, отсев кандидатов	Руководитель, архитектор, ИБ
Понедельник	Собрать 70–120 черновых вопросов, убрать дубли, разметить по типам задач	Шаблон набора, часть 1	Эксперты, руководитель
Вторник	Эталоны и ловушки; итог 50–100 вопросов; набор замораживается	Шаблон набора, часть 2; чек-лист готовности	Эксперты
Среда	Прогон всех кандидатов, замер скорости	Протокол прогона	Архитектор
Четверг	Слепая оценка	Бланк оценки, ключ	Эксперты, хранитель ключа
Пятница	Сведение расхождений, разбор и решение	Расхождения, сводная таблица, протокол решения	Все

При четырёх кандидатах каждый эксперт читает на треть больше ответов, и в каждом открывает источник. Оценке нужен второй день, а разбор переносится на понедельник. В недельном формате удобнее всего три кандидата.

Паспорт выбора

Заполняется до сбора вопросов. Поле о критичной ошибке — самое важное в документе: от него зависит, как читать результат в пятницу.

Поле	Заполнить
Главная задача	
Дополнительные задачи (тестируются отдельно)	
Корпус документов: состав, объём, форматы	
Состояние корпуса: противоречия, старые редакции, сканы	
Целевое оборудование	
Ожидаемая нагрузка: число одновременных запросов	
Ограничения контура: где выполняются вычисления, куда пишутся журналы	
Участники и роли	
Даты недели теста	

Что для нас критичная ошибка

Две-три строки до сбора вопросов. Записанное здесь потом уходит в эталон каждого вопроса как «что считать грубой ошибкой».

Критичная ошибка для нашей главной задачи

Как это формулируют — по одному примеру на тип задачи:

- **Поиск по документам:** уверенный ответ со ссылкой на пункт, где этого не написано.
- **Работа с договорами:** пропущенное изменённое условие — санкция, срок, порядок приёмки.
- **Диалог и подготовка текстов:** неверная дата, сумма или реквизит в документе, который уйдёт наружу.

Если корпус ещё не оценивали, сделайте это до теста: на архиве с несколькими действующими редакциями одного документа модель будут ругать за чужую ошибку. Методика — в статье «Как оценить корпоративный архив перед RAG-проектом».

Если данные нельзя выносить за контур, стенд, набор, эталоны и оценка остаются внутри. Модели с открытыми весами разворачиваются без участия разработчика. Корпоративные редакции вендоров тестируются только через их поставку в контур — договоритесь о ней до начала недели.

Отсев кандидатов

Тестируют только тех, кого смогут внедрить. Отсев занимает около половины дня. После него в списке остаются три-четыре модели.

Модель и точная версия	Лицензия весов	Память на целевом оборудовании	Допустимость в контуре	В тест или отсеять, причина

Лицензия весов. Читайте лицензию конкретной версии: внутри одного семейства условия бывают разными, а у флагманов с открытыми весами встречаются коммерческие ограничения для крупных пользователей. Если сфера может попасть под ограничения 243-ФЗ, лицензия становится вопросом соответствия — см. разбор статусов национальной и суверенной модели.

Память на целевом оборудовании. У MoE-моделей в видеопамять ложатся все веса, а активные параметры определяют только скорость. Расчёт по активным — типовая ошибка. Порядок расчёта — в статье «Сколько видеопамяти нужно LLM в своём контуре».

Допустимость в контуре. Для госоргана и значимого объекта КИИ это происхождение ПО и запись в реестре под конкретную редакцию и способ поставки — см. «Какие LLM можно использовать в КИИ и госсекторе». Для коммерческой компании — согласование с ИБ: где выполняются вычисления, куда пишутся журналы, кто обновляет модель.

Что мерить

Одна и та же модель может хорошо искать по регламентам и посредственно сравнивать редакции договоров. Отметьте главную задачу: по ней принимается решение. Остальные тестируйте отдельными наборами.

Главная	Задача	Что меряем	Критичная ошибка	Как ловить
☐	Поиск и ответы по документам	Верный ли документ и пункт; доля ответов «в документах этого нет» на вопросы без ответа в корпусе	Уверенный ответ со ссылкой на пункт, где этого не написано	Вопросы-ловушки, проверка каждой ссылки по исходнику
☐	Работа с договорами	Полнота найденных условий по сравнению с разметкой юриста; доля ложных срабатываний	Пропущенное изменённое условие: санкция, срок, порядок приёмки	Пары редакций договора с заранее известными изменениями
☐	Диалог и подготовка текстов	Соблюдение формата и шаблона; фактические ошибки	Неверная дата, сумма или реквизит в документе, который уйдёт наружу	Сверка каждого факта в черновике с исходными данными

В договорах полнота важнее точности: пропущенная санкция обходится дороже лишнего срабатывания, которое юрист отбросит за минуту. В поиске по архиву приоритет обратный — честное «не нашёл» ценнее ответа, который придётся каждый раз перепроверять.

Шаблон тестового набора

Лучший источник вопросов — то, о чём сотрудники уже спрашивают: обращения в поддержку, письма юристам, вопросы инженеров по документации, рабочие чаты. Записывайте вопрос дословно, с опечатками и внутренними сокращениями — так его и зададут системе. Если обращений мало, попросите пятерых будущих пользователей записать по десять вопросов, которые они задавали коллегам за последний месяц. Пропорции держите близкими к реальному потоку.

Эталон пишут будущие пользователи: юрист — для договорных вопросов, инженер — для технических. Вопрос, на который эталон написать нельзя из-за противоречия в документах или двух действующих редакций, исключите из оценки и передайте владельцу документов.

Часть 1. Вопросы

№	Вопрос дословно	Источник	Тип	Задача
П1 (пример)	Кто ещё согласует договор, если в нём есть предоплата?	Письмо юристу	Обычный	Поиск
П2 (пример)	где у нас про командировки в др. город на личной машине, топливо компенсируют?	Рабочий чат	Обычный	Поиск
П3 (пример)	Какой интервал ТО у насоса на линии 2 по инструкции?	Обращение инженера	Обычный	Поиск
П4 (пример)	Как оформить удалённую работу из другой страны по регламенту дистанционной работы?	Составлен экспертом	Ловушка	Поиск
П5 (пример)	Сколько хранятся журналы смены?	Обращение в поддержку	Обычный	Поиск

Часть 2. Эталоны

№	Эталонный ответ кратко	Документ и пункт	Что считать грубой ошибкой	Исключён из оценки
П1 (пример)	Дополнительно согласует финансовый директор	Регламент согласования договоров, п. 4.3	Ответ без финансового директора	Нет
П2 (пример)	Компенсация по чекам, если поездку на личной машине согласовал руководитель	Положение о командировках, раздел «Проезд»	Ссылка на нормы для служебного транспорта	Нет
П3 (пример)	Интервал из раздела «Техническое обслуживание» инструкции к этой модели насоса	Инструкция по эксплуатации насосного агрегата	Интервал из инструкции к другой модели насоса	Нет
П4 (пример)	В документах этого нет: регламента дистанционной работы в корпусе нет	—	Любая процедура со ссылкой на документ	Нет
П5 (пример)	Эталон не написан: две действующие редакции инструкции с разными сроками	Инструкция по ведению журналов, две редакции	—	Да: противоречие передано владельцу документа

Часть 1. Вопросы — для заполнения

№	Вопрос дословно	Источник	Тип	Задача
1				
2				
3				
4				
5				
6				
7				

Часть 2. Эталоны — для заполнения

№	Эталонный ответ кратко	Документ и пункт	Что считать грубой ошибкой	Исключён из оценки
1				
2				
3				
4				
5				
6				
7				

Полный набор ведите в таблице, лист «Набор»: там 100 строк и выпадающие списки. Номер вопроса в обеих частях один и тот же.

Если главная задача — договоры

Для договоров строка набора — пара редакций с изменениями, которые юрист заранее разметил. Вопрос и эталон в виде «документ и пункт» здесь не работают: меряют, сколько размеченных изменений модель нашла и сколько лишнего добавила. Пропуск важнее лишнего срабатывания.

№ пары	Редакции: файлы	Изменений по разметке юриста	Найдено моделью	Ложных срабатываний
1				
2				
3				

Кандидаты получают постоянные коды О1–О4 на весь договорный набор, оценщики сверяют списки изменений вслепую. В таблице это лист «Договоры»: полнота и число ложных срабатываний по каждому кандидату считаются автоматически.

Как написать ловушку

Ловушка — вопрос, ответа на который в корпусе нет, но который выглядит как настоящий. Правильный ответ на неё — «в документах этого нет». На обычных вопросах ответ в корпусе есть, и выдумывать модели незачем, поэтому без ловушек склонность к выдумке проявится только в эксплуатации. Отдайте ловушкам 10–15% набора.

Шаблон ловушки	Как сформулировать
Несуществующий регламент	Вопрос о процедуре по документу, которого в корпусе нет, с названием в стиле ваших документов
Пункт, которого нет в этой редакции	Вопрос о норме, которая была в прошлой редакции или есть только в проекте
Оборудование вне корпуса	Вопрос об узле или установке, документация на которые в корпус не вошла
Подразделение или должность, которых нет	Вопрос о полномочиях несуществующего отдела или роли
Не указанный в документах срок или сумма	Вопрос «сколько» или «когда» там, где документ говорит только «в установленном порядке»

Ловушка должна выглядеть как настоящий вопрос: тот же стиль, те же сокращения, без подсказки в формулировке.

Чек-лист готовности набора

- [ ] **В наборе 50–100 вопросов по главной задаче.** На таком объёме видны крупные различия и различия в типе ошибок.
- [ ] **Ловушки составляют 10–15% набора.**
- [ ] **Пропорции близки к реальному потоку обращений.**
- [ ] **У каждого вопроса есть эталон с документом и пунктом.**
- [ ] **У каждого эталона записано, что считать грубой ошибкой.**
- [ ] **Исключённые вопросы переданы владельцам документов списком.**
- [ ] **Набор заморожен. Дата заморозки:** _____
- [ ] **Промпт и примеры в нём написаны на других вопросах,** которых в наборе нет.

Протокол прогона

Все кандидаты прогоняются в одинаковых условиях: один индекс, один промпт, одни параметры генерации. Протокол нужен на случай, когда через полгода выйдет новая версия модели и кто-то спросит, почему выбрали эту.

Параметр	Кандидат 1	Кандидат 2	Кандидат 3	Кандидат 4
Модель и точная версия				
Источник весов и контрольная сумма				
Движок, его версия, способ квантования				
Промпт (файл, дата) и параметры генерации				
Индекс и модель эмбеддингов				
Дата и время прогона				

Условия одинаковы для всех кандидатов: ☐ Да ☐ Нет — что отличается: _____

Скорость на целевом оборудовании

Меряйте при ожидаемой нагрузке, например при 20–30 одновременных запросах. Для справочной системы важнее время до первого слова, для пакетной — полного ответа.

Кандидат	Нагрузка, одновременных запросов	Время до первого слова	Время полного ответа	Примечание

Где методика ломается

- ☐ **Тестовые вопросы не утекли в настройку.** Промпт и примеры для него пишутся на отдельных вопросах; набор заморожен до прогона.
- ☐ **Оценивают двое, блоками.** К шестидесятому вопросу эксперт начинает ставить «частично» там, где утром поставил бы «неверно». Оценку делят на блоки по 20–30 вопросов с перерывами, а второй оценщик показывает дрейф через расхождения.
- ☐ **Все кандидаты запущены одинаково.** API вендора против квантованной до четырёх бит модели на стенде — это сравнение сборок. Если кандидат не помещается на целевое оборудование без сильного сжатия, это результат теста.
- ☐ **Скорость и память учтены.** Результат модели, которой не хватает заложенных ускорителей, записывают вместе с ценой оборудования.
- ☐ **Решение принято по критичным ошибкам и ловушкам.** Средний балл — справочная строка сводной таблицы.

Бланк слепой оценки

Хранитель ключа перемешивает ответы всех кандидатов на каждый вопрос и присваивает им коды О1–О4. Оценщики получают лист ответов, где у каждого вопроса только коды, и не знают, какой модели принадлежит ответ: если эксперт узнаёт модель, которую сам предлагал, оценка смещается даже у добросовестного человека. Перед выдачей ответы приводят к единому оформлению — модель узнают и по стилю списков. До сведения оценок эксперты ответы не обсуждают. Ключ хранится отдельно и раскрывается на разборе, после сведения расхождений.

Оценка	Когда ставить
Верно	Ответ совпадает с эталоном по существу, ссылка ведёт на нужный документ и пункт. На ловушку — «в документах этого нет»
Частично	Суть верна, но пропущена деталь или ссылка неточная. Пользователь дойдёт до ответа сам, потратив время
Неверно	Ответ ошибочен, ссылка ведёт не туда или модель сочинила ответ на ловушку
Критичная ошибка	Отметка поверх оценки — по определению из эталона вопроса

Оценщик: _____ Дата: _____

№ вопроса	Код ответа	Оценка	Критичная ошибка	Комментарий
		☐ Верно ☐ Частично ☐ Неверно	☐ Да	
		☐ Верно ☐ Частично ☐ Неверно	☐ Да	
		☐ Верно ☐ Частично ☐ Неверно	☐ Да	
		☐ Верно ☐ Частично ☐ Неверно	☐ Да	
		☐ Верно ☐ Частично ☐ Неверно	☐ Да	
		☐ Верно ☐ Частично ☐ Неверно	☐ Да	
		☐ Верно ☐ Частично ☐ Неверно	☐ Да	
		☐ Верно ☐ Частично ☐ Неверно	☐ Да	
		☐ Верно ☐ Частично ☐ Неверно	☐ Да	
		☐ Верно ☐ Частично ☐ Неверно	☐ Да	

В таблице это листы «Оценка Э1» и «Оценка Э2»: по строке на каждую пару «вопрос — код ответа». Каждому эксперту передают копию без листов «Ключ», «Расхождения», «Сводная» и без бланка второго эксперта.

ХРАНИТСЯ У ОТВЕТСТВЕННОГО. ОЦЕНЩИКАМ НЕ ПЕРЕДАЁТСЯ

Ключ соответствия

Порядок кодов перемешивается в каждом вопросе. В таблице ключ — отдельный лист «Ключ»; его нет в копиях, которые получают эксперты.

Хранитель ключа: _____ Дата формирования: _____

№ вопроса	O1	O2	O3	O4

Расхождения оценщиков

Где оценки двух экспертов разошлись, итог принимают вместе и записывают причину. Если расхождений много, переписывают эталоны: обычно проблема в них.

№ вопроса	Код ответа	Эксперт 1	Эксперт 2	Итог	Причина
					☐ Эталон ☐ Формулировка ☐ Иное
					☐ Эталон ☐ Формулировка ☐ Иное
					☐ Эталон ☐ Формулировка ☐ Иное
					☐ Эталон ☐ Формулировка ☐ Иное
					☐ Эталон ☐ Формулировка ☐ Иное
					☐ Эталон ☐ Формулировка ☐ Иное

Сводная таблица

Заполняется после раскрытия ключа по итоговым оценкам — после сведения расхождений. Вопросы, исключённые из оценки, в подсчёт не входят. В таблице лист «Сводная» считает всё сам.

Всего вопросов в оценке: _____ Из них ловушек (N): _____

Кандидат	Верно	Частично	Неверно	Выдуманные ответы на ловушки, из N	Критичные ошибки

Кандидат	Время до первого слова	Время полного ответа	Верно, нечётные №	Верно, чётные №

Как читать

Небольшая разница в числе верных ответов — случайный разброс. На наборе из 50–100 вопросов разница в несколько ответов может исчезнуть или поменять знак, если повторить тест на другой выборке того же размера. Решение на ней строить нельзя. Простая проверка: посчитайте верные ответы отдельно по нечётным и чётным номерам вопросов. Если лидер в половинах разный, разница — шум.

Выдуманные ответы на ловушки — поведение модели. Каждый такой ответ в эксплуатации — сотрудник, который получил уверенную ссылку на несуществующий пункт и, возможно, принял по ней решение. Модель, которая выдумывает на ловушках заметно чаще соседей, проигрывает модели с чуть меньшим числом верных ответов. Разница в одну-две выдумки на десятке ловушек — повод добавить ловушки, а не вывод.

Критичные ошибки весят больше среднего балла. Что считать критичным, записано в паспорте до прогона. Модель с меньшим числом критичных ошибок почти всегда предпочтительнее модели с лучшим средним.

Если разница на ловушках небольшая и есть сомнения, добавьте ещё 20–30 ловушек и прогоните только их. Это полдня работы.

При равенстве по качеству и ошибкам решает эксплуатация: память и скорость на вашем оборудовании, лицензия, наличие людей, которые умеют модель обновлять.

Протокол решения

Поле	Заполнить
Выбранный кандидат и точная версия	
Второй номер — в резерве на пилот	
Обоснование: критичные ошибки, ловушки	
Обоснование: скорость и память на целевом оборудовании	
Обоснование: лицензия и допустимость в контуре	
Что дополнительно проверяем на пилоте: реальные пользователи, интеграции, права доступа	
Где хранится тестовый набор и протокол прогона	
Когда прогоняем набор заново: при каждом обновлении модели	

Участник	Роль	Подпись	Дата

Сравним модели на ваших документах

AZONE-AI проводит сравнение моделей в контуре заказчика: помогаем собрать тестовый набор из реальных вопросов, разворачиваем кандидатов на одном стенде и организуем слепую оценку. По результатам — модель-кандидат и архитектура пилота.

Пилот ИИ на ваших данных — azoneai.ru/services/ai-pilot/

Электронная таблица к шаблону — azoneai.ru/whitepapers/40/AZONE-AI_test-nabor-vybor-llm_2026.xlsx

AI-Контур — аттестованный ИИ-контур под ключ — azoneai.ru/products/ai-kontur/

Статья «Как выбрать LLM под свою задачу и проверить её за неделю» — azoneai.ru/blog/40-kak-vybrat-llm-pod-zadachu/

Сравнение отечественных LLM — azoneai.ru/blog/10-sravnenie-llm/

Обсудить сравнение моделей: azoneai.ru/contact · услуга «Пилот ИИ на ваших данных» · лицензии ФСТЭК, ФСБ, МО РФ

ООО «Азон» — интегратор с лицензиями ФСТЭК, ФСБ и Минобороны России. Шаблон описывает методику сравнения языковых моделей и не гарантирует результат пилота. Лицензии моделей и требования к контуру сверяйте на дату выбора.