

ТЕХНИЧЕСКОЕ ЗАДАНИЕ / ЗАДАЧИ ИКТ / MVP И ПИЛОТ

Разработка высокопроизводительной системы мультязычной кластеризации информационных событий Центральной Азии

Заказчик	ТОО «IMAS GROUP»
Предмет	Разработка работающего MVP независимого программного модуля событийной кластеризации с API и пилотной апробацией
Формат	Задачи ИКТ: проектирование → разработка MVP → измеримые испытания → пилот
Статус документа	Версия для повторной подачи на Astana Hub. Сроки и бюджет — предлагаемые ориентиры, требуют внутреннего утверждения.
Основание	Технологическая задача iMAS на площадке Astana Hub; комментарий модерации по предыдущей заявке

Публичная постановка практической разработки. Не содержит внутренней архитектуры, конфиденциальных данных и схем производственной платформы iMAS для открытого поиска технологического решения. Не содержит описания действующей инфраструктуры, схем данных или конфиденциальных клиентских материалов.

1. Вводные разделы

№	Элемент шаблона	Содержание и требования
1.1.1	Наименование	Разработка высокопроизводительной системы мультязычной кластеризации информационных событий Центральной Азии (программный MVP с API и испытаниями).
1.1.2	Краткая характеристика области применения	Медиамониторинг и аналитика открытых информационных потоков: материалы СМИ, социальные сети, заголовки, короткие посты и другие текстовые публикации. Поддержка разнородного и многоязычного контента.
1.2.1	Документ-основание	Объявление технологической задачи категории «Задачи ИКТ» на Astana Hub. Основание доработки — замечания модерации о конкретизации пилота, сроков, бюджета, KPI, ответственного и технологии.
1.2.2	Дата утверждения ТЗ	Проект для повторной подачи. Дата утверждения: [указать после согласования заказчиком].
1.2.3	Орган утверждения	ТОО «IMAS GROUP»; утверждает уполномоченный представитель заказчика. Ответственный за коммуникацию: [ФИО, должность, корпоративный e-mail, телефон].
1.3.1	Наименование объекта	Самостоятельный программный компонент кластеризации публикаций по информационным событиям с API и возможностью интеграции в медиаплатформу.
1.3.2	Заказчик / разработчик	Заказчик — ТОО «IMAS GROUP». Ответственный за задачу: [ФИО, должность, e-mail, телефон — указать в карточке портала]. Исполнитель определяется по итогам рассмотрения предложений.

№	Элемент шаблона	Содержание и требования
1.3.3	Сроки	Предлагаемый календарный срок MVP и пилота — 16 недель от подписания договора и предоставления согласованного тестового набора: 3 + 6 + 4 + 3 недели по этапам. Точные сроки требуют утверждения.
1.3.4	Финансирование	Бюджет на MVP и пилот определяется после утверждения заказчиком и оценки трудозатрат. До утверждения конкретного объёма финансирования сумма не заявляется; форма вознаграждения — по договорённости с исполнителем, порядок платежей и права на результаты оформляются договором.
1.4.1	Перечень терминов и сокращений	Инфоповод/событие — конкретный факт, описываемый несколькими публикациями; кластер — связанный набор публикаций; дубль — фактически повторённый материал; MVP — минимально жизнеспособный работающий модуль; API — программный интерфейс; LLM — большая языковая модель; baseline — базовое решение для сравнения; P95/P99 — процентиля времени обработки.
1.5.1	Объект	Поток мультязычных публикаций, содержащих повторные сообщения, перепечатки, переводы, дополнения и самостоятельные тематически близкие события.
1.5.2	Цель	Разработать автономный MVP, который в непрерывном потоке присваивает публикации идентификатор существующего события либо создаёт новое событие, связывает межязыковые описания, различает похожие самостоятельные события и передаёт результат через API.
1.5.3	Методы	Предложить и реализовать работоспособный многоэтапный способ сопоставления: поиск кандидатов, языково-независимые/семантические признаки, временные и событийные ограничения, повторная оценка сложных случаев. Исполнитель вправе выбрать модели и технологии, подтверждая выбор экспериментами.
1.5.4	Новизна	Предмет технологической разработки — доступный для внедрения потоковый модуль для региональных языков Центральной Азии с контролем ошибочных слияний событий, версионированием кластеров и экономической обработкой. Абсолютная научная новизна межязыкового анализа не заявляется; инновационность решения оценивается вручную по сравнению с существующими методами.
1.5.5	Эффективность	Ожидается автоматизация событийной аналитики. Целевые ориентиры пилота: pairwise precision $\geq 0,90$ и F1 $\geq 0,85$ на независимой разметке; ложные объединения среди специально подобранных пар разных событий $\leq 5\%$; обработка 1 млн сообщений/сутки при зафиксированной конфигурации; сравнение полной стоимости с baseline. Пороги должны подтверждаться испытаниями, а не декларацией.

2. Назначение, цели, объект

№	Элемент шаблона	Содержание и требования
2.1.1	Назначение системы	Работающий MVP: принимает новости и посты, определяет их принадлежность к событию, создаёт новые события, сохраняет

№	Элемент шаблона	Содержание и требования
		стабильные event_id, историю объединений/разделений и предоставляет API для результатов. Отдельный полноценный интерфейс iMAS и замена имеющейся инфраструктуры в предмет MVP не входят.
2.1.2	Показатели достижения цели	Целевые измеримые KPI (предварительные, уточняемые до старта): pairwise precision $\geq 0,90$; pairwise F1 $\geq 0,85$; ложное слияние на hard-negative парах $\leq 5\%$; обработка ≥ 1 млн сообщений в сутки на задокументированной конфигурации; P95 времени обработки публикации ≤ 5 с без времени ожидания в очереди; тест масштабирования до 5 млн/сутки; полная цена 1 млн сообщений и сравнение с baseline. Метрики отдельно по семи языкам, с фиксацией объёма размеченной выборки.
2.2	Назначение разработки	Практическая ИКТ-разработка: выпустить автономный MVP и проверить возможность пилотного подключения к медиаплатформе через документированный API. Сравнение алгоритмов — часть инженерной валидации продукта, а не единственный результат.
2.3	Характеристика предмета	Динамические потоки с повторами, переводами, короткими формами, неполными данными, смешением языков, разными вариантами имен и географических названий. Одно событие развивается во времени; схожие темы могут включать разные самостоятельные происшествия.
2.4	Существующее состояние	Открытые решения семантического поиска и кластеризации существуют. Практический недостаток для требуемого сценария — сочетание сложных межъязыковых различий, высокой цены ошибок объединения событий и стоимости потоковой обработки. Промышленная архитектура iMAS в открытой задаче не раскрывается.
2.5	Научная цель / гипотеза	Техническая гипотеза для сравнительной проверки: каскадная архитектура (быстрый поиск кандидатов → контекстная межъязыковая проверка → выборочная обработка неоднозначных случаев) способна обеспечить точность не ниже установленного KPI при более низкой стоимости и задержке, чем сплошная LLM-обработка. Гипотеза не подменяет требование работающего MVP; результат эксперимента может быть и отрицательным.

3. Требования к результату / системе

№	Элемент шаблона	Содержание и требования
3.1.1	Структура и функционирование	Предлагаемая функциональная архитектура MVP: 1) потоковый/пакетный приём текста с id и временными метками; 2) нормализация и определение/использование языка; 3) многоязыковое представление текста; 4) поиск кандидатов в индексе событий; 5) решение «существующий/новый инфоповод»; 6) хранилище кластеров, идентификаторов, версий и истории merge/split; 7) API результатов; 8) мониторинг. Исполнитель предлагает собственную реализацию компонентов. Это целевая схема нового модуля, а не описание действующей инфраструктуры iMAS.
3.1.2	Показатели назначения	Обязательный выпуск MVP с потоковым и пакетным режимами, API, демонстрацией объединения одного события на разных языках и разделения тематически похожих событий. Основная испытательная нагрузка: 1 млн документов в сутки; перспектива масштабирования —

№	Элемент шаблона	Содержание и требования
		5 млн/сутки, проверяемая нагрузочным сценарием. Объёмы не характеризуют текущие показатели iMAS.
3.1.3	Надёжность	Идемпотентность повторной подачи, гарантированное восстановление подтверждённого состояния после перезапуска, журнал операций merge/split, обработка запоздавших сообщений и накопленных очередей. Временный отказ отдельного обработчика не должен приводить к неконтрольной потере материалов.
3.1.4	Безопасность	Аутентификация к API, разграничение доступа к тестовым данным, управление секретами, безопасное журналирование, отсутствие передачи конфиденциальных материалов третьим сторонам без отдельного разрешения.
3.1.5	Эргономика и техническая эстетика	Достаточен минимальный интерфейс/страница проверки результатов с просмотром событий, связанных сообщений, причины решения и уверенности. Основное требование — машиночитаемые API и поддержка быстрых пользовательских выборов на стороне принимающей платформы.
3.1.6	Эксплуатация и обслуживание	Инструкция установки, конфигурирования и запуска; процедуры обновления модели и отката, резервного копирования состояния, обработки ошибок и повторного запуска; показатели загрузки CPU/GPU/RAM и ёмкости хранилища.
3.1.7	Защита от несанкционированного доступа	Контроль API-ключей/ролей, ограничение доступа к результатам и журналам, отсутствие секретов в исходном коде и открытых репозиториях; демонстрация проверки прав на тестовом окружении.
3.1.8	Сохранность при авариях	Подтверждённые входные события и результаты должны восстанавливаться после перезапуска; тестируются повторная доставка, прерывание обработки, сбой зависимого сервиса, восстановление очереди и состояния кластеров.
3.1.9	Защита от внешних воздействий	Для программного компонента — устойчивость к сетевым задержкам, пикам входящего потока, тайм-аутам зависимостей, повреждённым/пустым данным и временной нехватке ресурсов. Требования к специализированному физическому оборудованию не предъявляются.
3.1.10	Патентная чистота	Исполнитель декларирует права на результаты, соблюдает лицензии библиотек, датасетов и моделей; перечень сторонних компонентов и ограничений коммерческого/локального применения передаётся заказчику. Распределение прав на РИД — отдельным договором.
3.1.11	Стандартизация / унификация	Документированные контракты API и форматы обмена (например, JSON/JSONL), идентификаторы публикаций и событий, версии моделей, единая кодировка Unicode, машиночитаемые коды ошибок, воспроизводимое контейнерное развёртывание по согласованию.
3.1.12	Дополнительные требования	Нужны инструкции установки, диагностики и API; 1 демонстрационное занятие для технической команды заказчика. Отдельно предоставить протокол ручной экспертной оценки инновационности относительно минимум двух существующих подходов.
3.2.1	Функции и критерии отказов	Входные данные: item_id, text/title, published_at, source_id, language (если доступен); потоковый и пакетный приём. Результат: event_id, item_id, decision_type (новое/существующее событие, дубль, обновление), confidence, версия, дата принятия решения. Реализовать

№	Элемент шаблона	Содержание и требования
		late arrivals, merge/split с журналом, повторную доставку без неконтролируемых дублей, ошибки API. Подтвердить основные бизнес-сценарии на контрольных примерах.
3.3.1	Информационное обеспечение	Испытательные данные — открытые, синтетические или отдельно разрешённые обезличенные. Предлагаемый минимум для независимой оценки: 10 000 публикаций и не менее 500 событий, все 7 языков, не менее 1 000 трудноразличимых отрицательных пар. Объём редких языков, репрезентативность и независимость от обучения фиксируются отдельно; при недостатке подтверждённых данных метрики показываются с оговорками.
3.3.2	Лингвистическое обеспечение	Обязательно: казахский, кыргызский, узбекский, таджикский, туркменский, русский, английский; кириллица, латиница и применимые варианты арабской графики; транслитерация, смешанные языки, имена, аббревиатуры и географические названия. Расширяемость на каракалпакский, уйгурский и иные языки; качество по каждому языку измеряется отдельно.
3.3.3	Программное обеспечение	Выбор языка разработки, фреймворка, индекса и моделей свободный. Обязательны работающий компонент, исходный код или передаваемый исполняемый артефакт по согласованной лицензии, API-спецификация, автоматические тесты, повторяемая сборка и инструкция запуска.
3.3.4	Техническое обеспечение	Ресурсы и требования к оборудованию для заявленного throughput подтверждаются тестами. Представить измерения CPU/GPU/RAM, размера индекса, задержек P50/P95/P99 и стоимости обработки 1 млн сообщений, включая внешние API при их использовании.
3.3.5	Организационное обеспечение	Заказчик утверждает набор данных и KPI до независимой оценки. Исполнитель предоставляет руководителя проекта и контактную точку. Команда iMAS назначает ответственного: [ФИО / контакты]. Производственные данные и внутренние сервисы не открываются участникам на этапе публичного отбора.
3.4	Требования к программному изделию	MVP устанавливается в независимом тестовом контуре и демонстрирует полный цикл «публикация → событие → обновление → API». Для пилота поддерживается интеграция через документированный контракт без прямого доступа к внутренним таблицам действующей платформы.
3.5	Информационная безопасность	Тестовые данные без разглашения коммерческой тайны и персональных данных; соблюдение прав доступа, политика хранения и удаления материалов, указание внешних сервисов обработки, аудит зависимостей и лицензий.
3.6	Промышленная безопасность / ESG	Физические промышленные риски неприменимы к данному программному модулю. Экономическая и ресурсная эффективность обязательна: оценить энергозатратные вычисления косвенно через CPU/GPU-время, стоимость обработки и долю тяжёлых моделей.

4. Работы, сроки, приёмка

№	Элемент шаблона	Содержание и требования
4.1.1	Стадии, исполнители, документы	Предлагаемый график 16 недель: Этап 1 (недели 1–3) — согласование сценариев, тестовых данных, KPI, 2 baseline и архитектуры MVP; отчёт и план работ. Этап 2 (недели 4–9) — работающий модуль, API, индекс/состояние событий, поддержка повторов и поздних данных; демонстрация. Этап 3 (недели 10–13) — независимая оценка качества по языкам, нагрузочные и отказные испытания, стоимость и сравнение с baseline; отчёт. Этап 4 (недели 14–16) — пилот на согласованном потоке в изолированном окружении, исправление дефектов, документация, решение о внедрении.
4.2.1	НИР → ОКР → ОПЭ / внедрение	Тип задачи — ИКТ, результат — MVP. Проверка сравнительной гипотезы выполняется внутри этапов, не требует заявления фундаментальной научной новизны. После успешного MVP возможны пилот и отдельный договор на промышленную интеграцию.
4.3.1	Контроль и приёмка	Приёмка MVP: 1) независимый запуск из инструкции; 2) демонстрация API, event_id, поздних сообщений, merge/split и повторной доставки; 3) результаты размеченного закрытого теста; 4) pairwise precision $\geq 0,90$ и F1 $\geq 0,85$, false-merge rate $\leq 5\%$ на hard negatives; 5) throughput ≥ 1 млн/сутки на описанной конфигурации и P95 ≤ 5 с без очереди; 6) протокол аварийных сценариев; 7) отчёт об инфраструктурных затратах и ручная оценка инновационности. Если пороги недостижимы, оформляется анализ причин и решение сторон по доработке, а не фиктивный акт приёмки.
4.4.1	Технико-экономические показатели	Для каждого теста фиксируются датасет, алгоритм, конфигурация CPU/GPU/RAM, версии моделей, pairwise precision/recall/F1, ошибки merge/split и показатели по языкам; производительность, P95, стоимость 1 млн материалов, доля обращений к LLM и цена относительно двух baseline. Сценарии: 100 тыс./1 млн/5 млн сообщений в сутки; 5 млн — масштабирование, не безусловная производительность одной машины.
4.5.1	Порядок оплаты	Форма вознаграждения: «По договорённости с исполнителем; рассматривается оплачиваемая разработка MVP с пилотной апробацией». Конкретный бюджет указывается только после решения заказчика, этапные платежи и интеллектуальные права определяются договором.

5. Подготовка объекта и документирование

№	Элемент шаблона	Содержание и требования
5.1	Подготовка к вводу в действие	Заказчик предоставляет согласованный тестовый корпус и возможность изолированного пилота через API/файловый обмен без доступа исполнителя к production-инфраструктуре. Исполнитель обеспечивает развёртывание, тестирование, мониторинг и повторяемый запуск. Финальное внедрение — по отдельному решению.
5.2.1	Комплект документации	Передаются исходный код/развёртываемый артефакт согласно договору, архитектура нового модуля, схема входа/выхода, OpenAPI или эквивалент, инструкции, тесты, отчёты качества/производительности/затрат, матрица ограничений по языкам,

№	Элемент шаблона	Содержание и требования
		протокол пилота и ручная сравнительная оценка инновационности; при необходимости оформление по ГОСТ 34.201 и 7.32-2017.
5.3.1	Источники разработки	Настоящее ТЗ, публичная заявка, открытые опубликованные исследования и готовые программные решения (в т.ч. multilingual retrieval/event matching), открытые или согласованные датасеты. Результаты сравнения оформляются без заявления исключительной новизны алгоритма.

6. Квалификационные требования

№	Элемент шаблона	Содержание и требования
6.1.1	Материально-техническая база	Исполнитель способен создать, показать и протестировать программный MVP; желательно практическое владение NLP, потоковой обработкой и проверкой качества. Допускаются IT-компании, команды и стартапы, предлагающие существующую технологию с требуемыми доработками.
6.2	Разрешительные документы	Специальные лицензии для разработки программного модуля не предопределяются данным ТЗ. Исполнитель обязан иметь законные права на используемые данные, ПО и модели и выполнять требования законодательства, применимые к выбранному решению.
6.3	Гарантийные обязательства	В рамках этапа пилота исполнитель устраняет воспроизводимые дефекты, препятствующие согласованным KPI, и передаёт актуальные инструкции и тесты. Дополнительная гарантия, SLA и коммерческая поддержка устанавливаются отдельным договором.

Приложение А. Матрица проверки MVP

Показатель	Предлагаемый критерий приёмки
Точность	Pairwise precision $\geq 0,90$, pairwise F1 $\geq 0,85$; закрытая разметка с независимыми событиями.
Границы событий	Ложное объединение на сложных отрицательных парах $\leq 5\%$; отдельно по типам ошибок.
Производительность	Нагрузка ≥ 1 млн сообщений/сутки на описанной конфигурации; сценарий роста до 5 млн/сутки.
Задержка	P95 ≤ 5 секунд на этап решения без времени в очереди; задержка очереди оценивается отдельно.
Стабильность	Повторная доставка, поздние материалы, перезапуск и восстановление без потери подтверждённых результатов.
Экономика	Стоимость 1 млн публикаций, CPU/GPU/RAM и внешние API; сравнение минимум с двумя baseline.