Разработка программного прототипа автоматической генерации и структурирования знаний и навыков с использованием LLM

Заказчик: ТОО «2дэй академи».

Сфера применения: Образование (EdTech).

Технологическое направление: Искусственный интеллект.

Бюджет: 200 000 тенге.

Срок реализации: 20 рабочих дней.

Стадия: Разработка программного прототипа (MVP).

1. Описание проблемы

Ручное формирование структурированных перечней знаний и навыков по различным профессиям и предметным областям требует значительных временных затрат.

Необходимо автоматизировать процесс получения, классификации и структурирования знаний и навыков с использованием больших языковых моделей.

2. Ожидаемый результат

Рабочее самостоятельное программное приложение, позволяющее пользователю ввести название профессии или предметной области и получить структурированную базу знаний и навыков с указанием логических зависимостей между ними.

Допускается реализация в виде консольного приложения на Python с сохранением результатов в локальные JSON-файлы.

Прототип должен запускаться на компьютере заказчика без необходимости развёртывания отдельной серверной инфраструктуры.

3. Сценарии использования

Сценарий 1. Формирование базы знаний по профессии

Пользователь вводит название профессии «Финансовый аналитик».

Приложение обращается к LLM, определяет основные категории знаний, формирует перечень атомарных знаний и практических навыков, классифицирует их и сохраняет результаты.

Сценарий 2. Формирование базы знаний по предметной области

Пользователь вводит название предметной области «Математическая статистика».

Приложение формирует перечень основных понятий, определений, методов и навыков, определяет базовые зависимости между ними и сохраняет результаты.

Сценарий 3. Расширение существующей базы

Пользователь повторно запускает генерацию для ранее обработанного направления.

Приложение загружает существующую базу, формирует дополнительные единицы знаний и навыков, проверяет их на дублирование и добавляет новые записи.

Сценарий 4. Экспорт результатов

Пользователь получает сформированную базу знаний и связей в формате JSON для последующего анализа или использования во внешних информационных системах.

4. Функциональные требования

Программный прототип должен обеспечивать:

1. Ввод названия профессии, специализации или предметной области.
2. Автоматическое формирование знаний и навыков посредством LLM.
3. Классификацию полученных единиц по категориям.
4. Декомпозицию сложных понятий на самостоятельные единицы знаний и навыков.
5. Определение базовых связей между единицами.
6. Выявление дублирующихся записей на основе нормализации текста и базового семантического сопоставления.
7. Повторное расширение ранее сформированной базы.
8. Сохранение и экспорт результатов в JSON.
9. Проверку соответствия результатов установленной структуре данных.
10. Вывод информации об ошибках обработки и возможности повторного запуска.

5. Входные данные

Обязательный параметр:

- название профессии, специализации или предметной области.

Дополнительные параметры:

- количество генерируемых единиц;
- тип генерируемых единиц: знания, навыки или оба типа.

Язык генерации в рамках MVP — русский.

6. Выходные данные

Каждая единица знаний или навыков должна содержать:

- уникальный идентификатор;
- наименование;
- краткое описание;
- тип: знание или навык;
- категорию;
- предметную область;
- идентификаторы связанных единиц.

Пример результата:

```
{
  "id": "K001",
  "name": "Дисконтирование денежных потоков",
  "description": "Приведение будущих денежных потоков к текущей стоимости",
  "type": "knowledge",
  "category": "Финансовая математика",
  "domain": "Финансовый анализ",
  "dependencies": ["K002", "K003"]
}
```

Пример является предварительным. Окончательная JSON-схема согласовывается на первом этапе разработки.

Все результаты генерации считаются кандидатами в знания до проведения экспертной проверки.

7. Интеграции

В рамках MVP предусматривается интеграция только с API одной большой языковой модели.

Допускается использование OpenAI GPT-4.1 mini, Google Gemini Flash или аналогичной модели по согласованию с заказчиком.

API-ключ и параметры подключения должны задаваться через конфигурацию приложения.

Расходы на использование LLM при разработке и приёмочном тестировании включены в общий бюджет проекта в пределах 10 000 тенге и оплачиваются исполнителем.

Интеграция с образовательными платформами, внешними базами данных и другими информационными системами в рамках данного этапа не предусматривается.

8. Состав поставки

По завершении работ исполнитель передаёт заказчику:

- полный исходный код приложения;
- используемые промпты и шаблоны запросов к LLM;
- файл зависимостей программного обеспечения;
- шаблон конфигурации подключения к LLM;
- JSON-схему выходных данных;
- демонстрационные наборы знаний и навыков;
- инструкцию по установке и запуску;
- инструкцию по использованию;
- отчёт о результатах тестирования.

Переданный исходный код должен позволять заказчику самостоятельно запускать, изменять и развивать приложение.

9. Измеримые критерии приёмки

Критерий	**Требование**
Количество пилотных предметных областей	2
Количество единиц на одну область	Не менее 30
Общее количество знаний и навыков	Не менее 60
Количество логических связей	Не менее 10 на область

Доля экспертно подтверждённых единиц	Не менее 80% по каждой области
Доля корректных логических связей	Не менее 80% по каждой области
Допустимая доля дубликатов	Не более 10%
Соответствие JSON-схеме	100% записей
Целостность ссылок между единицами	100%
Повторное расширение базы	Не менее 5 новых корректных единиц в одной из пилотных областей без перезаписи существующих записей
Самостоятельный запуск заказчиком	Обязателен

10. Методика приёмочного тестирования

Заказчик совместно с исполнителем определяет две пилотные предметные области.

Исполнитель запускает генерацию и предоставляет полученные результаты в формате JSON.

Заказчик проводит экспертную оценку всех сформированных единиц и связей.

Единица считается корректной, если:

- соответствует заданной предметной области;
- не содержит существенных фактических ошибок;
- представляет самостоятельное знание или навык;
- имеет корректное наименование и описание;
- правильно классифицирована.

Доля корректных единиц рассчитывается как отношение количества единиц, одобренных экспертом, к общему количеству сгенерированных единиц.

Доля дубликатов рассчитывается как отношение количества избыточных семантически эквивалентных записей к общему количеству записей.

Корректность связей определяется отношением количества подтверждённых экспертом связей к общему количеству сформированных связей.

Проверка JSON-схемы и целостности ссылок выполняется автоматически.

При несоответствии критериям исполнитель устраняет выявленные недостатки в рамках согласованного бюджета.

11. Границы MVP

В объём текущего этапа не входят:

- разработка полноценного веб-интерфейса;
- графическая визуализация и редактор графа;
- специализированная графовая база данных;
- обработка книг, документов и видеоматериалов;
- поддержка казахского и других языков;
- обучение собственных моделей искусственного интеллекта;
- автоматическая проверка достоверности по внешним источникам;
- разработка публичного интеграционного API;
- интеграция с образовательными платформами;
- промышленное развёртывание и масштабирование.

Цель первого этапа — проверка технологической возможности автоматизированного формирования структурированной базы знаний и навыков.

12. Бюджет и этапы реализации

Этап	Результат	Срок	Стоимость
1. Проектирование	Структура данных, JSON-схема, согласованные сценарии	4 рабочих дня	40 000 ₸
2. Разработка	Рабочее приложение, генерация, классификация,	10 рабочих дней	100 000 ₸

	связи, дедупликация и экспорт		
3. Приёмка	Тестирование, исправление замечаний, документация и передача исходного кода	6 рабочих дней	60 000 ₸
Итого		**20 рабочих дней**	**200 000 ₸**

Бюджет обоснован ограниченным объёмом прототипа, использованием готовых LLM и существующих программных библиотек, отсутствием необходимости разработки графического интерфейса и серверной инфраструктуры.

Экспертную оценку сформированных знаний обеспечивает заказчик.

13. Договорные условия и передача прав

Работы выполняются на основании договора подряда с поэтапной оплатой после приёмки результатов каждого этапа.

Исключительные имущественные права на специально созданные в рамках договора исходный код, документацию и иные охраняемые результаты разработки передаются заказчику после полной оплаты на условиях договора.

Сторонние программные компоненты используются в соответствии с их лицензиями.

Исполнитель обязан обеспечить конфиденциальность предоставленной внутренней документации заказчика.