Fase 1: Planificación y Preparación de Datos

1.1. Definir el Objetivo y Alcance (Detallado): El objetivo principal de este proyecto es desarrollar un Tutor Inteligente Interactivo basado en RAG que actúe como una herramienta de apoyo integral para estudiantes que se preparan para los exámenes de ingreso a la universidad en Cuba. Este tutor estará especializado en las asignaturas de Matemática, Español-Literatura e Historia.

1.1.1. Objetivo General:

Proveer una plataforma conversacional y dinámica que facilite el estudio autónomo y la comprensión profunda de los contenidos académicos clave para los exámenes de ingreso a la universidad, simulando la interacción con un tutor humano y adaptándose a las necesidades del estudiante.

1.1.2. Objetivos Específicos:

Asistencia en Contenidos:

- Responder Preguntas Conceptuales: Clarificar definiciones, teoremas, reglas gramaticales, eventos históricos, figuras literarias, etc., de manera precisa y comprensible.
- Explicar Temas Complejos: Desglosar conceptos difíciles en pasos más simples, utilizando analogías y ejemplos relevantes para cada asignatura.
- Proporcionar Ejemplos Ilustrativos: Ofrecer ejemplos concretos para cada concepto o procedimiento, especialmente en Matemática y Español-Literatura.

Soporte en Resolución de Ejercicios:

- Guiar en la Resolución de Problemas: No solo dar la respuesta, sino guiar al estudiante paso a paso a través de la lógica de resolución de ejercicios matemáticos, análisis literarios o interpretación de eventos históricos.
- Ofrecer Ejercicios de Práctica: Generar o recuperar ejercicios relevantes basados en el tema que el estudiante está estudiando o en sus áreas de debilidad identificadas.
- Corregir y Proporcionar Retroalimentación: Evaluar las respuestas del estudiante a ejercicios y proporcionar retroalimentación constructiva, identificando errores y sugiriendo mejoras.

Interacción y Personalización:

- Mantener Contexto Conversacional: Recordar el historial de la conversación para ofrecer respuestas más pertinentes y evitar repeticiones.
- Adaptar el Nivel de Explicación: Ajustar la complejidad del lenguaje y los ejemplos según el nivel de comprensión aparente del estudiante.

- Identificar Áreas de Debilidad (a futuro): Aunque no en la fase inicial, un objetivo a largo plazo es identificar automáticamente las áreas donde el estudiante necesita más refuerzo basándose en sus interacciones y respuestas.

Acceso al Conocimiento:

- Acceso a Información Curada: Asegurar que toda la información provenga de fuentes académicas verificadas y relevantes para el currículo universitario cubano.
- Capacidad de Actualización Dinámica: Integrar mecanismos para actualizar la base de conocimiento con nuevos materiales o revisiones del currículo, o para buscar información complementaria cuando el conocimiento interno no sea suficiente.

1.1.3. Alcance del Proyecto (Fase Inicial):

El alcance de la fase inicial del desarrollo del tutor inteligente se centrará en:

- *Asignaturas Cubiertas:* Matemática, Español-Literatura e Historia, con enfoque en los temas principales y más frecuentes en los exámenes de ingreso a la universidad cubana.
- *Tipos de Interacción:*
  - Preguntas directas sobre conceptos y hechos.
  - Solicitudes de explicación o ejemplos.
  - Presentación y corrección de ejercicios simples y de complejidad media.
  - Capacidad de seguir un hilo conversacional básico sobre un tema.
- *Fuentes de Conocimiento:* Prioritariamente textos estructurados de libros de texto oficiales y guías de estudio preuniversitarias cubanas.
- *Tecnología Base:* Implementación de un sistema RAG avanzado, con énfasis en un módulo de recuperación robusto, un módulo de generación coherente y la integración de un Grafos de Conocimiento básico para cada asignatura.
- *Crawler Automatizado:* Implementación de la fase inicial del crawler para la ingesta y vectorización de la base de conocimiento curada. La fase dinámica del crawler se planificará, pero su implementación completa puede ser escalonada.
- *Evaluación:* Desarrollo de un pipeline de evaluación inicial basado en métricas de rendimiento y la participación de un grupo reducido de evaluadores humanos (profesores, estudiantes piloto).

No Incluido en la Fase Inicial (Potenciales Fases Futuras):

- Personalización profunda: No se incluirá un perfil de estudiante persistente ni una adaptación curricular compleja al inicio.
- Generación de exámenes completos: Se enfocará en ejercicios individuales o series cortas, no en exámenes simulacro completos.

- Análisis de rendimiento predictivo: No se realizarán análisis avanzados para predecir el rendimiento del estudiante.
- Multilingüismo: Solo español.
- Integración con plataformas académicas externas: No se conectará con sistemas de gestión de aprendizaje (LMS) existentes.

Preparación de Datos

1.2. Identificar Fuentes de Conocimiento Externas (Enfoque en Cuba)

Para asegurar la pertinencia y precisión de la información de nuestro tutor inteligente en el contexto de los exámenes de ingreso a la universidad en Cuba, es crucial identificar las fuentes de conocimiento primarias y secundarias más relevantes y autorizadas.

1.2.1. Fuentes Primarias (Oficiales y Curriculares):

Estas son las fuentes más importantes, ya que definen el currículo y los estándares de los exámenes.

- *Programas de Estudio Oficiales del Ministerio de Educación Superior (MES) / Ministerio de Educación (MINED):*
    - Para cada asignatura (Matemática, Español-Literatura, Historia), los programas de estudio detallan los contenidos específicos que se evalúan en los exámenes de ingreso. Son la base para la selección y priorización de la información.
    - Formato: Generalmente PDFs, documentos Word, o HTML en sitios web oficiales.
- *Libros de Texto de Preuniversitario (Ediciones Nacionales):*
    - Los libros de texto cubanos utilizados en la educación preuniversitaria (10mo, 11no, 12mo grados) son la fuente fundamental de los contenidos. Es vital usar las ediciones más recientes y reconocidas por el sistema educativo cubano.
    - Ejemplos Notables:
        - Matemática: Textos de Álgebra, Geometría, Trigonometría y Cálculo introductorio (ej., los de la serie para preuniversitario).
        - Español-Literatura: Libros de gramática, ortografía, redacción y antologías literarias cubanas y universales que forman parte del programa.
        - Historia: Libros de Historia de Cuba e Historia Universal adaptados al currículo cubano.
    - Formato: Principalmente libros físicos, que requerirán digitalización (OCR) o acceso a versiones digitales si existen.
- *Guías y Folletos de Preparación para Exámenes de Ingreso:*

    - Publicaciones específicas (a menudo del MINED o editoriales académicas cubanas) que contienen resúmenes, ejercicios resueltos y exámenes de años anteriores.
    - Formato: PDFs o documentos impresos.

- *Archivos de Exámenes de Ingreso de Años Anteriores:*

    - Los exámenes pasados son una fuente invaluable para entender el formato de las preguntas, los temas recurrentes y el nivel de dificultad. Las soluciones oficiales también son críticas.
    - Formato: PDFs o imágenes de exámenes escaneados.

1.2.2. Fuentes Secundarias (Complementarias y de Apoyo):

Estas fuentes pueden enriquecer la base de conocimiento y proporcionar perspectivas adicionales.

- Publicaciones Académicas Cubanas (si son accesibles): Artículos, ensayos o tesis de universidades cubanas que profundicen en temas específicos del currículo.
- Enciclopedias y Diccionarios de Referencia (Online o Digitales): Fuentes generales para clarificar conceptos o términos que puedan aparecer en las preguntas o respuestas.
- Bases de Datos de Ejercicios y Problemas: Si existen repositorios digitales de ejercicios de matemáticas o análisis de textos con soluciones, serían muy valiosos.
- Grafos de Conocimiento Pre-existentes (si son adaptables): Aunque construiremos los nuestros, si existen KGs generales (ej., Wikidata para historia, bases de datos de terminología matemática) que puedan ser filtrados y adaptados al contexto cubano, podrían acelerar el proceso.
- Recursos Web Confiables: Sitios web de instituciones educativas o de divulgación científica cubanas que ofrezcan explicaciones adicionales, ejercicios o recursos multimedia. (Usados con precaución y validación).

1.2.3. Consideraciones Adicionales sobre las Fuentes en el Contexto Cubano:

- *Acceso:* La principal limitación puede ser el acceso digital a todo el material. Muchos libros y guías aún predominan en formato físico, lo que requerirá un proceso de digitalización (escaneo y OCR).
- *Autoría y Validación:* Dada la necesidad de precisión para exámenes, priorizaremos siempre las fuentes oficiales del MINED/MES o de autores y editoriales académicas reconocidas en Cuba.
- *Actualización:* Es vital asegurar que las ediciones de los libros de texto y los programas de estudio sean los más recientes, ya que el currículo puede variar.

1.3. Preparar la Base de Conocimiento (Detallado)

La preparación de la base de conocimiento es el pilar de la calidad del sistema RAG. Un conocimiento bien curado y estructurado es esencial para la precisión de las respuestas del tutor.

1.3.1. Proceso de Recopilación y Digitalización:

- *Inventario de Fuentes:* Crear un inventario exhaustivo de todos los libros, guías, programas de estudio y exámenes de ingreso identificados.
- *Adquisición:* Adquirir las versiones físicas o digitales de estas fuentes.
- *Digitalización (si es necesario):*
    - Para materiales impresos: Escanear a alta resolución (DPI) y utilizar software de Reconocimiento Óptico de Caracteres (OCR) para convertir las imágenes en texto editable y buscable. Es crucial usar OCR de alta calidad para minimizar errores.
    - Validar la calidad del OCR, corrigiendo posibles errores de transcripción.

1.3.2. Procesamiento del Lenguaje Natural (NLP) y Limpieza de Datos:

Una vez que tengamos el texto digital, aplicaremos técnicas de NLP para limpiarlo, normalizarlo y prepararlo para la indexación y la construcción del grafo.

- *Extracción de Texto y Limpieza:*
    - Eliminar encabezados, pies de página, números de página, marcas de agua, y otros elementos no textuales que puedan ser ruido.
    - Tratar saltos de línea y formatos inconsistentes.
    - Identificar y separar secciones importantes (capítulos, lecciones, ejercicios, soluciones).
- *Normalización del Texto:*
    - Tokenización: Dividir el texto en palabras y oraciones.
    - Lematización/Radicalización: Reducir las palabras a su forma base (ej., "corriendo" a "correr") para mejorar la consistencia en las búsquedas.
    - Remoción de Stop Words: Eliminar palabras comunes sin significado semántico fuerte (ej., "el", "la", "de") si no son relevantes para la recuperación.
    - Manejo de Variantes: Normalizar términos que puedan tener múltiples formas (ej., "Marxismo-Leninismo" vs. "Marxismo Leninismo", "2da Guerra Mundial" vs. "Segunda Guerra Mundial").
- *Detección y Extracción de Entidades y Relaciones (NER y RE):*
    - Asignatura de Historia: Identificar nombres de personas, fechas, eventos, lugares geográficos, organizaciones, leyes, tratados.
    - Asignatura de Español-Literatura: Identificar autores, obras literarias, movimientos literarios, figuras retóricas, personajes.
    - Asignatura de Matemática: Identificar teoremas, fórmulas, conceptos matemáticos, tipos de problemas, nombres de matemáticos.
    - Extracción de Relaciones: Identificar las conexiones entre las entidades (ej., "X nació_en Y", "Evento_A causó Evento_B", "Obra_Z escrita_por Autor_W", "Fórmula_F pertenece_a

Teorema_T"). Esto es crucial para la construcción del grafo de conocimiento.

1.3.3. Representación del Conocimiento: Construcción del Grafo de Conocimiento (KG):

La creación de un KG será fundamental para la capacidad del tutor de razonar y proporcionar explicaciones estructuradas, especialmente en Matemáticas e Historia.

- *Diseño del Esquema/Ontología:*
    - Definir las clases de entidades (ej., Persona, Evento, ConceptoMatemático, ObraLiteraria) y los tipos de relaciones (ej., participa_en, causa, define, es_prerrequisito_de, escrito_por, pertenece_a).
    - Asegurar que el esquema sea lo suficientemente flexible para abarcar las tres asignaturas.
- *Poblado del Grafo:*
    - Utilizar la información extraída por NER y RE para poblar el grafo con nodos (entidades) y aristas (relaciones).
    - Se puede usar un enfoque semi-automatizado: extracción automática con revisión y curación manual por expertos en el dominio (profesores, estudiantes avanzados) para asegurar la precisión.
    - Integrar metadatos cruciales: asignaturas, nivel de dificultad (ej., básico, intermedio, avanzado), tipo de contenido (definición, ejemplo, ejercicio, solución).
- *Almacenamiento del Grafo:* Utilizar una base de datos de grafos (ej., Neo4j, ArangoDB) para almacenar el KG, lo que permitirá consultas eficientes sobre relaciones complejas.

1.3.4. Segmentación Estratégica (Chunking) para el Recuperador:

Una vez que el texto esté limpio y las entidades/relaciones extraídas, procederemos a la segmentación para el módulo recuperador.

- *Segmentación basada en Contenido:*
    - No solo por longitud: chunking por párrafos, secciones, o incluso por la unidad de significado más pequeña (ej., una definición completa, un teorema, la explicación de un paso en un problema).
    - Sobrecarga de Contexto: Cada chunk incluirá cierto contexto circundante para asegurar que la información no esté aislada.
- *Etiquetado de Chunks (Metadatos):*
    - Cada chunk debe ser etiquetado con metadatos como:
        - asignatura: Matemática, Español-Literatura, Historia.

        - tema_principal: Álgebra, Geometría, Gramática, Ortografía, Revolución Cubana, etc.
        - subtema: Ecuaciones cuadráticas, pronombres, Guerra de los Diez Años.
        - tipo_de_contenido: definición, ejemplo, ejercicio, solución, explicación_teórica, biografía, análisis_obra.
        - dificultad_estimada: básico, intermedio, avanzado.
        - fuente_original: (ej., "Libro de Matemática 12mo Grado, pag 45").
    - Estos metadatos serán esenciales para la recuperación filtrada y la personalización de las respuestas del tutor.

1.3.5. Consideraciones de Curación para la Calidad:

- *Validación de Datos:* Un equipo de expertos en cada asignatura (profesores o estudiantes universitarios avanzados) deberá validar la precisión y relevancia de los datos extraídos y las relaciones en el KG.
- *Resolución de Ambigüedades:* Identificar y resolver ambigüedades terminológicas, especialmente en Español-Literatura (ej., diferentes interpretaciones de una figura retórica) o Historia (fechas exactas de eventos).
- *Normalización de Ejercicios y Soluciones:* Asegurar un formato consistente para ejercicios y sus soluciones para facilitar la corrección automática y la guía paso a paso.

Esta fase es intensiva en datos y en el uso de NLP, pero es la que garantizará que nuestro tutor inteligente tenga una base de conocimiento robusta y precisa para ser efectivo en la preparación de los exámenes de ingreso.