Instructivo del Proyecto Final (Parte 1): Investigación crítica, razonamiento de chatbots LLMs

Prof. Dr. Manuel David Morales

Duración Total: 1 hora 30 minutos (sesión única)

Matemáticas II, Licenciatura en IA y Ciencia de Datos, Semestre 2025B

I. Introducción y Objetivos de la Actividad

Tema: *Prompt Engineering Crítico* (PE-C) sobre integración y derivación multivariable.

Objetivo principal: Utilizar su conocimiento avanzado de Matemáticas II para **falsificar** la respuesta de la IA. Es decir, demostrar que, a pesar de su complejidad lingüística, el `chatbot` es susceptible de cometer errores de rigor, precisión o incoherencia en operaciones de alto nivel.

Objetivos específicos:

- **Diagnóstico:** Aislar el punto exacto donde la IA comete un error conceptual o algebraico.
- **Rigor Formal:** Justificar su refutación usando terminología precisa de Matemáticas II.
- **Uso Ético:** Tratar a la IA como un asistente que requiere validación humana.

IMPORTANTE: Plataforma de IA. Para asegurar la consistencia metodológica de la investigación, **todos los grupos utilizarán la plataforma Gemini de Google** para la totalidad de las actividades. Acceder al sitio: `https://gemini.google.com`.

II. Metodología de Trabajo y Asignación de Roles

Para maximizar la eficiencia en la sesión de 90 minutos, cada equipo debe establecer inmediatamente los siguientes roles técnicos. **La participación de todos es obligatoria.**

Rol Técnico	Tarea Principal	Relevancia en el PE-C
1. Prompt Engineer	Responsable de redactar todos los `prompts` de Primer y Segundo Orden (falsificación). Debe definir el Rol y las Restricciones de Salida del `chatbot`.	Lidera la estrategia de diálogo para exponer la debilidad del LLM.
2. Analista Algorítmico	Resuelve los problemas analíticamente (a mano) en paralelo con la IA. Responsable de identificar y aislar el paso algebraico o conceptual exacto donde el LLM comete un error.	Provee la base de verdad y diagnostica la falla del LLM.
3. Curador del Reporte y Presentador	Documenta el proceso (capturas de `prompts` y respuestas). Redacta el Diagnóstico Crítico final y el resumen conceptual. Responsable de la presentación final de 3 minutos.	Asegura la coherencia del entregable y comunica los hallazgos del grupo.
4. Auditor de Rigor	(Obligatorio para grupos de 4+) Supervisa la coherencia entre la formulación del `Prompt Engineer` y los cálculos del `Analista`. Revisa y asegura la correcta notación formal (símbolos, ecuaciones, variables, etc.).	Asegura el rigor formal y la consistencia lógica interna del proceso de falsificación.
5. Especialista en Meta-Data	(Optativo para grupos de 5) Se encarga de la captura uniforme de todas las interacciones (`prompts`, respuestas) y asesora al grupo en el `prompt` de desafío **ético/contextual**.	Facilita la recolección de datos de investigación y el cumplimiento del uso ético de la IA.

III. Protocolo de Prompt Engineering Crítico (PE-C)

Siga este flujo de trabajo sistemático para ambas actividades. Documente cada paso.

1. **Definición del Rol:** Configure al `chatbot` (ej., "Eres un tutor purista de cálculo integral/diferencial llamado Bernoulli...").
2. **Prompt de Primer Orden (Inducción de Falla):** Introduzca el problema y exija que el `chatbot` muestre su **proceso paso a paso completo**[1].
3. **Respuesta del LLM:** Registre la respuesta del `chatbot`.

[1] **Instrucción crítica:** Para inducir la falla, el `Prompt Engineer` debe intencionalmente **omitir una restricción de rigor** o incluir una directriz que lleve a una decisión algebraica o conceptual sub-óptima. El objetivo es que la IA revele su lógica, incluso si es defectuosa.

4. **Detección de Falla:** El `Analista Algorítmico` encuentra la discrepancia entre el cálculo manual y la solución del `chatbot`.

5. **Prompts de Segundo Orden (Falsificación):** El `Prompt Engineer` diseña posteriores preguntas iterativas que cuestionan directamente el paso donde se detectó el error (ej., pregunta por una igualdad o la justificación de un signo).

6. **Diagnóstico Crítico (Grupal):** Redacten la conclusión final: *¿Por qué falló la IA y cómo se relaciona esto con la debilidad de los LLMs en el razonamiento simbólico?*

IV. Actividad 1: Integración Crítica (Tiempo Estimado: 35 min)

El Problema de Desafío

Utilicen la metodología PE-C para resolver y criticar la solución de la siguiente integral, cuya resolución requiere **Sustitución Trigonométrica** y, posteriormente, **Integración por Partes** (IBP):

$$\int x^2 \sqrt{1 - x^2}\, dx$$

Puntos de Ataque Estratégico (Diseño del Prompts):

- **Exploración de Rigor (Prompt de primer orden):** El `Prompt Engineer` debe solicitar la solución e intencionalmente **omitir la restricción** de usar identidades trigonométricas o la constante de integración, buscando un error de secuencia lógica o rigor.

- **Aislamiento de la Falla (Prompts de segundo orden):** Una vez detectado el error del LLM (ej., un fallo en la IBP o en la identidad de reducción de potencia), diseñen un `prompt` de segundo orden que aísle ese paso y obligue a la IA a **justificar la identidad trigonométrica correcta** que debió usar.

V. Actividad 2: Derivación Multivariable Compleja (Tiempo Estimado: 35 min)

El Problema de Desafío

Sea la función de costo $J = F(u, v)$, donde $u(x, y) = x^3 + e^y$ y $v(x, y) = \ln(x^2 y)$.

Utilicen la metodología PE-C para calcular y criticar la derivada parcial con respecto a x:

$$\frac{\partial J}{\partial x}$$

Puntos de Ataque Estratégico (Diseño del Prompts):

- **Exploración de Rigor (Prompt de primer orden):** El `Prompt Engineer` debe exigir que la IA muestre primero la **fórmula general de la Regla de la Cadena** y luego calcule los componentes internos, buscando que la IA **confunda o omita** algún término en la complejidad de la composición.

- **Aislamiento de la Falla (Prompts de segundo orden):** Una vez que el `Analista Algorítmico` aísle el error (ej., una derivada parcial interna incorrecta o un error de signo), el `Prompt Engineer` debe diseñar un `prompt` que obligue al `chatbot` a **justificar la contribución de cada rama** de la Regla de la Cadena, forzando la corrección o el reconocimiento del error.

VI. Presentación de Hallazgos y Debate (Últimos minutos)

Duelo de Diagnóstico Crítico (DDC): Al finalizar las actividades, el profesor seleccionará uno o dos grupos al azar. El `Curador del Reporte y Presentador` comenzará la exposición.

- **Tiempo:** Máximo **3 minutos** (rigurosamente cronometrados).
- **Contenido (Foco en el Diagnóstico):**
    1. La **falla conceptual** más interesante que encontraron.
    2. El **Diagnóstico Crítico** basado en Cálculo II.
    3. La **efectividad** del `prompt` de falsificación utilizado.

Mecanismo de Verificación Individual: Durante el debate, el profesor se reservará el derecho de dirigir preguntas específicas a cualquier miembro del equipo (ej., `Analista Algorítmico`, `Prompt Engineer`) relacionadas con su tarea de rol, para verificar su contribución directa y dominio del contenido. La respuesta individual será la base para el puntaje de la rúbrica DDC.

VII. Rúbricas de Evaluación

La evaluación se estructura en dos componentes principales: el rigor técnico del trabajo grupal (Entregable) y la capacidad de síntesis/argumentación individual (DDC).

VII.A. Rúbrica para el Entregable Final (Nota Grupal)

Esta rúbrica mide la calidad de la documentación, el rigor matemático en el diagnóstico y la aplicación sistemática de la metodología de `Prompt Engineering Crítico` (PE-C).

Criterio	**Bajo (1-2)**	**Aceptable (3)**	**Sobresaliente (4-5)**
Rigor de Prompt Engineering Crítico (PE-C)	Los `prompts` son ambiguos o carecen de roles y contexto. No se evidencia intención de falsificación.	Los `prompts` definen rol e intentan iteración básica. El formato de salida es inconsistente.	Se utiliza el PE-C de forma sistemática; se incluyen `prompts` de falsificación y control de salida estructurado.
Análisis Crítico / Falsificación	El grupo acepta la respuesta del `chatbot` o identifica el error superficialmente.	Se detecta el error, pero la refutación carece de rigor matemático o es incompleta.	Identifica el punto exacto de la falla conceptual (ej. secuencia de sustituciones, Regla de la Cadena). La refutación es exhaustiva.
Profundidad Conceptual Matemática	La justificación final reproduce la explicación de la IA o contiene errores conceptuales graves.	La justificación es correcta, pero la conexión entre el concepto y su aplicación en la crítica de la IA/ML es débil.	Construye una argumentación sólida y original, explicando por qué el error del LLM es predecible en el contexto de la IA y la Ciencia de Datos.
Organización y Comunicación del Reporte	El reporte es confuso y desorganizado. No se utiliza notación formal adecuada.	El reporte está bien organizado. La comunicación es clara, pero la integración de evidencia es inconsistente.	Muestra una organización lógica excelente. El flujo de trabajo (`prompt`-respuesta-diagnóstico) es claro y la exposición es fluida.

VII.B. Rúbrica para el Duelo de Diagnóstico Crítico (DDC) y Verificación Individual

Esta rúbrica evalúa la **capacidad de síntesis del grupo** (a través del presentador) y el **dominio individual** al responder preguntas dirigidas a su rol (ej., `Analista Algorítmico`, `Prompt Engineer`).

Criterio	**Insuficiente/Aceptable (1-3)**	**Excelente (4-5)**
Dominio de la Refutación (Individual)	La respuesta es confusa o incompleta, sin utilizar la terminología técnica correcta asociada a su rol (ej., el `Analista` no puede aislar el paso algebraico).	La persona cuestionada demuestra un dominio convincente. Utiliza terminología precisa para explicar su contribución específica (diseño de `prompt`, cálculo, o curación).
Capacidad Argumentativa y Evidencia	La argumentación es básica; no utiliza evidencia (`prompt` o cálculo) para sostener el diagnóstico o la corrección.	La argumentación es aguda. Se selecciona información clave (evidencia) para acentuar el diagnóstico y defender la refutación con solidez.
Claridad y Concisión (Grupal)	La exposición excede el tiempo límite (3 min) o el lenguaje es desorganizado.	La exposición es clara, concisa y organizada. El mensaje principal se comunica de forma efectiva dentro del límite de tiempo.