Metodología híbrida para el aprendizaje y la evaluación de modelos de lenguaje bajo esquemas de incertidumbre difusa

1st Juan F. Gaitán-Guerrero
Departamento de Informática
Universidad de Jaén
Jaén, España
jgaitan@ujaen.es

2nd Carmén Martínez-Cruz
Departamento de Informática
Universidad de Granada
Granada, España
cmcruz@ugr.es

3rd Macarena Espinilla
Departamento de Informática
Universidad de Jaén
Jaén, España
mestevez@ujaen.es

4th David Díaz-Jiménez
Departamento de Informática
Universidad de Jaén
Jaén, España
ddjimene@ujaen.es

5th José L López
Departamento de Informática
Universidad de Jaén
Jaén, España
llopez@ujaen.es

***Abstract*—Este artículo propone una metodología híbrida para el entrenamiento y evaluación de modelos de generación de lenguaje natural en dominios con presencia de datos temporales cuyo procesamiento y descripción están caracterizados por incertidumbre y ambigüedad. El enfoque central se basa en la lógica difusa como herramienta transversal: en la generación y validación del conjunto de datos, en la guía del modelo mediante ingeniería de prompts, y en la evaluación del rendimiento a través de una metodología de evaluación objetiva. La propuesta se valida en el dominio de la monitorización continua de glucosa en pacientes con diabetes, demostrando mejoras significativas en precisión, claridad y seguridad de las salidas generadas por modelos de la familia GPT.**

***Index Terms*—Lógica difusa, GPT, Series temporales, Protoformas, Ingeniería de prompts, IoT, Fine-tuning, Resúmenes lingüísticos, Metodología de evaluación.**

I. Introducción

La interpretación automática de datos provenientes de sensores en tiempo real representa uno de los grandes retos en la inteligencia artificial aplicada. En dominios críticos como la salud, donde los datos suelen ser complejos, ruidosos e inciertos, resulta esencial contar con metodologías robustas que permitan a los modelos generar interpretaciones confiables y comprensibles para los usuarios finales.

En la actual era de la Inteligencia Artificial, muchas personas ya usan modelos de lenguaje. Esto coincide con las limitaciones del sistema sanitario, donde las citas son breves y falta una visión clara de la evolución de pacientes crónicos monitorizados. Ante ello, tecnologías accesibles como las de OpenAI —con más de 400 millones de usuarios— abren la posibilidad de que tanto profesionales como pacientes analicen datos temporales sin conocimientos técnicos, gracias a interfaces amigables.

Tradicionalmente, la lógica difusa ha sido crucial para representar conocimiento experto bajo incertidumbre, ofreciendo una solución interpretable para traducir series temporales numéricas en lenguaje natural. Como consecuencia, en este trabajo proponemos una dualidad entre ambos paradigmas, pero donde la lógica difusa cumple un rol central en tres niveles:

1) En la generación de un conjunto de entrenamiento etiquetado con descripciones lingüísticas expertas.
2) En el diseño de prompts que guían al modelo en función de conocimiento clínico impreciso.
3) En la evaluación de los resultados generados por los modelos mediante criterios lingüísticos definidos en la metodología.

II. Metodología general

La metodología se estructura en cuatro fases principales: (1) adquisición y preprocesamiento de los datos, (2) modelado experto y generación del dataset, (3) ingeniería de prompts y entrenamiento, y (4) evaluación.

A. Fase 1: Arquitectura de adquisición y preprocesamiento

Se emplea una arquitectura IoT centrada en sensores como el Freestyle Libre 3, que proporciona valores de glucosa cada 5 minutos. Los datos son recolectados mediante una app móvil, almacenados en una base de datos MongoDB y posteriormente procesados para su análisis.

Este resultado ha sido parcialmente respaldado por el proyecto PID2021-127275OB-I00 y el proyecto PID2021-126363NB-I00 financiados por el MICIU/AEI/10.13039/501100011033 y por "ERDF A way of making Europe", y por el proyecto PDC2023-145863-I00 financiado por el MICIU/AEI/ 10.13039/501100011033 y por la "European Union NextGenerationEU/PRTR".

Este resultado es un resumen del trabajo publicado con referencia doi.org/10.1016/j.cmpb.2025.108878

Cada serie diaria consta de hasta 288 muestras, que son segmentadas mediante el algoritmo de Ramer-Douglas-Peucker para simplificar su representación y facilitar la detección de patrones significativos como picos, valles o tendencias.

B. Fase 2: Modelado experto y generación del dataset

Esta es la fase clave, donde interviene el modelado experto mediante lógica difusa para transformar los datos numéricos en lenguaje humano interpretable. Se definen las siguientes categorías lingüísticas:

- Valores de glucosa: muy baja, baja, normal, alta, muy alta.
- Tendencias: descendente, estable, ascendente.
- Momentos del día: madrugada, mañana, tarde, noche.
- Cuantificadores: pocos, muchos, casi todos.

A partir de estas categorías se construyen protoformas lingüísticas, siguiendo la propuesta de Marín y Sánchez [1], de la forma:

Durante la mañana, muchos valores de glucosa fueron altos.

Cada protoforma se evalúa mediante su grado de verdad (DoT), calculado con funciones de pertenencia trapezoidales, y se filtran aquellas que no superan un umbral mínimo. También se aplican reglas de calidad textual para garantizar que las descripciones sean claras, no redundantes y respeten la coherencia temporal.

El dataset final se organiza en formato JSONL utilizando el sistema de roles estructurados de OpenAI [2]: el rol *system* para proporcionar las instrucciones iniciales y el contexto de la tarea, el rol *user* especificando los datos de la serie temporal, y el rol *assistant* con la salida esperada (el resumen modelado).

C. Fase 3: Ingeniería de prompts guiada por lógica difusa

La guía del modelo durante la generación textual es fundamental para reducir errores semánticos y alucinaciones. Para ello, se diseña un prompt estructurado en cuatro partes:

1) **Encabezado**: presentación del objetivo y estructura de los datos.
2) **Dominio**: contexto clínico relevante (por ejemplo, identificar hipoglucemias).
3) **Instrucción**: rol deseado (ej. endocrinólogo experto), nivel de detalle y periodo a analizar.
4) **Conocimiento**: explicitación de etiquetas difusas y cuantificadores que debe usar.

Esta estructura, inspirada en el enfoque prompt-as-prefix (PaP), facilita que el modelo interiorice las reglas del dominio y genere descripciones que reflejan el conocimiento experto. A través del rol *system* mencionado anteriormente, se introduce esta información, estableciendo un proceso de experimentación a través de la variación del contenido.

D. Fase 4: Evaluación

La evaluación de los resultados generados se basa en la metodología propuesta en [3], que propone una serie de criterios objetivos para valorar descripciones generadas automáticamente, Las dimensiones evaluadas son:

- **Calidad de la información**: precisión, relevancia, cobertura, corrección.
- **Capacidad de razonamiento**: comprensión y lógica de la salida.
- **Calidad comunicativa**: claridad del lenguaje, accesibilidad al usuario y uso de referencias temporales.
- **Seguridad del contenido**: ausencia de afirmaciones infundadas o no derivables de los datos.

Cada criterio es cuantificado a partir de la comparación con descripciones expertas, obviando la necesidad de evaluadores expertos y la subjetividad implícita a esta actuación, siendo una evaluación objetiva, y ampliando además la metodología [3] a escenarios con datos temporales.

III. Resultados

Se aplicó esta metodología para entrenar y evaluar versiones de GPT-3.5, GPT-4o y GPT-4o-mini. Los mejores resultados se obtuvieron con GPT-4o, utilizando la serie original (no segmentada) y prompts enriquecidos con conocimiento difuso, alcanzando un 96% de conformidad con los criterios.

Los modelos sin ajuste fino o sin guía explícita tendieron a producir descripciones imprecisas, redundantes o incluso alucinaciones. El uso sistemático de lógica difusa en el dataset y en el prompt redujo significativamente estos errores.

IV. Discusión

El estudio demuestra que la lógica difusa no solo es útil para representar conocimiento incierto, sino que constituye una herramienta poderosa para estructurar datasets, guiar el comportamiento generativo de modelos y evaluar su rendimiento. Este enfoque híbrido resulta especialmente valioso en contextos clínicos donde los errores pueden tener consecuencias serias.

Además, el diseño del sistema permite escalar la metodología a otros dominios como monitoreo cardíaco, análisis de actividad física o predicción meteorológica, siempre que exista una base de conocimiento experto susceptible de representación difusa.

V. Conclusiones

Este trabajo propone una metodología robusta, replicable y segura para entrenar modelos de lenguaje bajo esquemas de incertidumbre. A través del uso transversal de la lógica difusa, desde la generación del dataset hasta su evaluación, se logra alinear modelos como GPT con los objetivos y limitaciones del razonamiento experto humano.

References

[1] N. Marín and D. Sánchez, "On generating linguistic descriptions of time series," *Fuzzy Sets and Systems*, vol. 285, pp. 6–30, 2016.

[2] OpenAI, "Fine-tuning - openai api." https://platform.openai.com/docs/guides/fine-tuning, 8 2024. (Accessed on 11/10/2024).

[3] T. Y. C. Tam, S. Sivarajkumar, S. Kapoor, A. V. Stolyar, K. Polanska, K. R. McCarthy, H. Osterhoudt, X. Wu, S. Visweswaran, S. Fu, *et al.*, "A framework for human evaluation of large language models in healthcare derived from literature review," *NPJ digital medicine*, vol. 7, no. 1, p. 258, 2024.