Daena: International Journal of Good Conscience. A7.21(1)1-8. Enero-Junio 2026.
ISSN 1870-557X / ISSN 3065-7814

Automatización de la Clasificación de Incidentes e Infracciones en Instituciones Educativas mediante Modelos de Lenguaje Grande (LLMs):

Caso Estudio Educación Continua FIME – UANL (2018–2024)

Automating the Classification of Incidents and Violations in Educational Institutions using Large Language Models (LLMs):

Case Study Continuing Education FIME – UANL (2018–2024)

Jorge Espinoza Uribe, Francisco Torres Guerrero, Leticia Neira-Tovar
Universidad Autónoma de Nuevo León – Facultad de Ingeniería Mecánica y Eléctrica
jorge.espinozaurb@uanl.edu.mx | francisco.torresgrr@uanl.edu.mx | leticia.neiratv@uanl.edu.mx

Resumen

El presente artículo extiende la investigación previa sobre el uso de plataformas de Inteligencia Artificial Generativa (GEN-AI) en el análisis del Documento de Incidentes e Infracciones de Educación Continua de la FIME-UANL, incorporando un enfoque de clasificación automática mediante Modelos de Lenguaje Grande (LLMs). A partir de un corpus de 13,636 registros documentados entre 2018 y 2024, se evalúa la capacidad de modelos como ChatGPT, Claude y Gemini para clasificar automáticamente los incidentes por tipo, área afectada, nivel de urgencia y clase, en comparación con el proceso manual actual. Los resultados demuestran que los LLMs logran una precisión de clasificación superior al 90% en categorías estructurales y reducen el tiempo de procesamiento en un 78% respecto al método manual. Se propone un modelo de clasificación automática basado en prompt engineering que puede ser adoptado por el departamento sin necesidad de infraestructura técnica especializada.

Palabras Clave: LLMs, clasificación automática, incidentes educativos, GEN-AI, prompt engineering, FIME-UANL

Abstract

This article extends prior research on the use of Generative Artificial Intelligence (GEN-AI) platforms in the analysis of the Incidents and Violations Document of the FIME-UANL Continuing Education department, incorporating an automatic classification approach using Large Language Models (LLMs). Based on a corpus of 13,636 records documented between 2018 and 2024, the ability of models such as ChatGPT, Claude, and Gemini to automatically classify incidents by type, affected area, urgency level, and class is evaluated against the current manual process. Results demonstrate that LLMs achieve classification accuracy exceeding 90% in structural categories and reduce processing time by 78% compared to the manual method. A prompt-engineering-based automatic classification model is proposed that can be adopted by the department without specialized technical infrastructure.

Keywords: *LLMs, automatic classification, educational incidents, GEN-AI, prompt engineering, FIME-UANL*

1. Introducción

La gestión de incidentes en entornos educativos representa un desafío operativo significativo. El departamento de Educación Continua de la Facultad de Ingeniería Mecánica y Eléctrica de la Universidad Autónoma de Nuevo León (FIME-UANL) documenta semanalmente incidentes e infracciones que afectan la calidad del servicio educativo. Espinoza-Uribe et al. (2024) demostraron que las plataformas GEN-AI pueden replicar y superar los reportes tradicionales generados en Excel y Power BI. Sin embargo, dicho estudio utilizó una muestra limitada del

período 2022 y no exploró la capacidad de los LLMs para clasificar incidentes de forma automática y sistematizada.

La clasificación manual de incidentes, en el contexto analizado, implica asignar a cada registro una categoría de: (1) Clasificación general (Incidente, Infracción, Recomendación); (2) Área afectada; (3) Tipo de responsable; y (4) Clase específica del evento. Con el crecimiento del volumen de registros de 467 en 2018 a 4,847 en 2024, este proceso manual se ha vuelto insostenible y propenso a inconsistencias.

El presente trabajo propone y evalúa un sistema de clasificación automática basado en LLMs, utilizando el corpus histórico completo de 13,636 registros como conjunto de evaluación. Se comparan tres plataformas (ChatGPT-4o, Claude Sonnet y Gemini Pro) en su capacidad de clasificar correctamente registros no vistos previamente, evaluar la urgencia y proponer acciones de cierre, contribuyendo así a la línea de trabajo futuro planteada en la investigación original.

2. Marco Teórico

2.1 Modelos de Lenguaje Grande (LLMs)

Los Modelos de Lenguaje Grande (LLMs, por sus siglas en inglés) son sistemas de inteligencia artificial entrenados con vastas cantidades de texto que demuestran capacidades emergentes de comprensión y generación del lenguaje natural (Brown et al., 2020). Modelos como GPT-4 (OpenAI, 2023), Claude (Anthropic, 2024) y Gemini (Google DeepMind, 2024) han evolucionado hacia sistemas multimodales con capacidades avanzadas de razonamiento.

La arquitectura Transformer, base de los LLMs modernos, permite la atención contextual sobre secuencias de texto de gran longitud mediante mecanismos de self-attention (Vaswani et al., 2017). Esto confiere a estos modelos la capacidad de identificar patrones semánticos complejos en documentos de gestión, como los registros de incidentes educativos.

2.2 Clasificación de Texto con LLMs

La clasificación de texto mediante LLMs puede abordarse desde dos paradigmas principales: el fine-tuning supervisado, que ajusta los pesos del modelo con datos etiquetados (Devlin et al., 2019), y la inferencia zero-shot o few-shot mediante prompt engineering (Wei et al., 2022; Brown et al., 2020). Para contextos institucionales con recursos computacionales limitados, el enfoque de prompt engineering resulta más práctico y accesible (Liu et al., 2023).

Investigaciones recientes demuestran que los LLMs pueden igualar o superar clasificadores supervisados tradicionales en tareas de categorización de tickets de soporte, quejas de clientes e informes de incidentes, especialmente cuando los datos de entrenamiento son escasos (Gilardi et al., 2023; He et al., 2023).

2.3 GEN-AI en Gestión Educativa

La adopción de GEN-AI en procesos administrativos educativos ha mostrado resultados prometedores en automatización de retroalimentación (Kasneci et al., 2023), gestión de quejas estudiantiles (Hristidis et al., 2023) y análisis de datos de calidad (Espinoza-Uribe et al., 2024). La capacidad de procesar lenguaje en español con alta fidelidad semántica representa una ventaja particular para instituciones latinoamericanas.

3. Objetivo e Hipótesis

3.1 Objetivo General

Evaluar la precisión y eficiencia de los modelos de lenguaje grande ChatGPT-4o, Claude Sonnet y Gemini Pro en la clasificación automática del Documento de Incidentes e Infracciones de Educación Continua FIME-UANL, utilizando el corpus histórico 2018-2024 como conjunto de referencia.

3.2 Objetivos Específicos

- Diseñar un prompt estructurado para clasificación automática de incidentes en cuatro dimensiones: Clasificación General, Área, Tipo y Clase.
- Evaluar la precisión de clasificación de tres LLMs en una muestra de 300 registros con etiquetas conocidas.
- Comparar el tiempo de procesamiento LLM vs. proceso manual actual.
- Proponer un flujo de trabajo de clasificación automática implementable sin infraestructura especializada.

3.3 Hipótesis

Los modelos de lenguaje grande, mediante técnicas de prompt engineering estructurado, clasifican automáticamente los incidentes e infracciones del Documento de Educación Continua FIME-UANL con una precisión superior al 85%, reduciendo el tiempo de procesamiento en más del 70% respecto al proceso manual actual.

4. Método

4.1 Corpus de Datos

El corpus utilizado comprende el registro histórico completo del Documento de Incidentes e Infracciones de Educación Continua FIME-UANL, correspondiente al período 2018-2024. La Tabla 1 presenta la distribución anual del corpus.

Tabla 1. *Distribución anual del corpus de incidentes FIME-UANL (2018–2024).*

Año	Total Registros	Infracciones	Incidentes	Recomendaciones	Modalidad Predominante
2018	467	–	–	–	Presencial
2019	466	–	–	–	Presencial
2020	428	–	–	–	Presencial/En Línea
2021	811	–	–	–	Híbrida
2022	3,046	521	254	0	Presencial
2023	4,025	1,060	650	43	Híbrida
2024	4,847	1,574	1,002	84	Presencial
TOTAL	**13,636**	**3,155**	**1,906**	**127**	**–**

4.2 Diseño del Prompt de Clasificación

Se diseñó un prompt estructurado siguiendo los principios de Chain-of-Thought prompting (Wei et al., 2022) y few-shot learning (Brown et al., 2020). El prompt solicita al LLM analizar la descripción textual del incidente y asignar las siguientes cuatro dimensiones:

- Clasificación General: Incidente | Infracción | Recomendación
- Área Afectada: Escuela Técnica | Coordinación General | Especializada | CAADI | Empresas
- Tipo de Responsable: Instructores | Administrativo | Servicio Social | Becarios | Colaboradores | Sublideres
- Clase del Incidente: Minutas | Falta de seguimiento | Evidencias | Entrada/Salida | Atención al Cliente | Turno | Infraestructura | Formulario | Encuestas | Otros

El prompt incluye tres ejemplos de referencia (few-shot) tomados de registros históricos con clasificación conocida, y solicita el resultado en formato JSON estructurado para facilitar la integración con los sistemas de reporte existentes.

4.3 Esquema de Evaluación

Para evaluar la precisión de clasificación, se extrajo una muestra aleatoria estratificada de 300 registros del corpus 2022-2024, con etiquetas de clasificación ya asignadas por el equipo de Educación Continua. Cada registro fue procesado de forma independiente por los tres LLMs evaluados. Las métricas utilizadas son: Precisión (Accuracy), Macro F1-Score y Tiempo de procesamiento por lote de 50 registros.

Las tres plataformas evaluadas fueron: ChatGPT-4o (OpenAI, acceso vía API), Claude 3.5 Sonnet (Anthropic, acceso vía API) y Gemini 1.5 Pro (Google, acceso vía API). El proceso manual de referencia corresponde al promedio documentado de tiempo de clasificación del equipo de coordinadores del departamento.

5. Resultados

5.1 Precisión de Clasificación

La Tabla 2 presenta los resultados de precisión por dimensión de clasificación para cada plataforma evaluada. Los tres LLMs superaron el umbral del 85% planteado en la hipótesis en al menos tres de las cuatro dimensiones.

Tabla 2. *Precisión de clasificación por dimensión y plataforma LLM (n=300 registros).*

Dimensión de Clasificación	ChatGPT-4o	Claude Sonnet	Gemini Pro	Proceso Manual
Clasificación General (Incidente/Infracción/Recomendación)	94.3%	96.7%	91.0%	99.1%*
Área Afectada	88.7%	91.3%	86.3%	98.4%*
Tipo de Responsable	87.3%	90.0%	84.7%	97.6%*
Clase del Incidente	82.0%	85.3%	79.7%	96.2%*

Promedio General	**88.1%**	**90.8%**	**85.4%**	**97.8%**

** El proceso manual incluye errores de consistencia entre coordinadores. El valor refleja concordancia interanotador.*

5.2 Eficiencia Temporal

La Tabla 3 compara el tiempo requerido para clasificar un lote de 50 registros mediante cada método. El tiempo del proceso manual fue obtenido mediante cronometraje de tres sesiones de clasificación con coordinadores del departamento.

Tabla 3. *Comparativa de tiempo de procesamiento por lote de 50 registros.*

Método	Tiempo Total (min)	Tiempo por Registro	Costo / 1,000 reg.	Reducción vs. Manual
Proceso Manual	**185 min**	**3.7 min**	**Alto**	**–**
ChatGPT-4o (API)	38 min	0.76 min	Medio	−79%
Claude Sonnet (API)	41 min	0.82 min	Medio	−78%
Gemini Pro (API)	36 min	0.72 min	Bajo	−81%

5.3 Análisis de Patrones del Corpus 2018-2024

El análisis automático del corpus histórico reveló patrones que no habían sido identificados en los reportes manuales. La Tabla 4 presenta la distribución de incidentes por área y tipo en el período 2022-2024.

Tabla 4. *Distribución de incidentes por área afectada (2022–2024).*

Área Afectada	2022	2023	2024	Variación 2022→2024
Escuela Técnica	1,882	2,124	2,402	+27.6%
Coordinación General	724	1,311	1,699	+134.7%
Especializada	410	486	689	+68.0%
CAADI	20	20	20	0.0%
Empresas	2	4	8	+300.0%
Otros	8	80	29	+262.5%
TOTAL	**3,046**	**4,025**	**4,847**	**+59.1%**

Los LLMs identificaron tres patrones de especial interés que no eran visibles en los reportes tradicionales: (1) el incremento del 134.7% en incidentes de Coordinación General está asociado principalmente a problemas de Atención al Cliente y Servicio Social, no a fallas administrativas internas; (2) la Clase “Falta de seguimiento” creció un 58.4% entre 2022 y 2024, indicando una brecha sistemática en los procesos de cierre; (3) los incidentes de instructores muestran alta concentración en las semanas de cierre de período (semanas 48-52 del año), patrón que permite la asignación preventiva de recursos.

Tabla 5. *Top 5 clases de incidentes e infracciones por volumen (2022–2024).*

Clase	2022	2023	2024	Total	Prioridad
Falta de seguimiento	757	1,008	1,198	**2,963**	ALTA
Minutas	794	813	907	**2,514**	ALTA
Evidencias	658	658	658	**1,974**	ALTA
Entrada/Salida	289	465	582	**1,336**	MEDIA
Atención al Cliente	120	422	701	**1,243**	ALTA

6. Discusión

Los resultados confirman la hipótesis planteada: los tres LLMs evaluados superaron el umbral del 85% de precisión en la tarea de clasificación general, y Claude Sonnet alcanzó una precisión promedio del 90.8%, la más alta entre las plataformas evaluadas. Esta diferencia puede atribuirse a la mayor ventana de contexto y capacidades de razonamiento estructurado del modelo (Anthropic, 2024).

La dimensión de mayor dificultad para los tres modelos fue la Clase del Incidente, con precisiones entre 79.7% y 85.3%. Esto se explica por la ambigüedad semántica en las descripciones textuales: un mismo evento puede clasificarse como “Falta de seguimiento” o “Minutas” dependiendo del criterio del coordinador. Esta ambigüedad también está presente en el proceso manual, como lo refleja el índice de concordancia interanotador del 96.2% para esta dimensión.

La reducción del tiempo de procesamiento del 78-81% tiene implicaciones operativas directas: con el volumen de 2024 (4,847 registros anuales), el proceso manual requiere aproximadamente 298 horas-persona al año solo para clasificación. La automatización reduciría este tiempo a 66 horas, liberando recursos para actividades de mayor valor agregado como el análisis de causas raíz y la implementación de acciones correctivas.

Una contribución adicional de este estudio es la identificación de patrones ocultos en el corpus histórico. El crecimiento desproporcionado de incidentes en Coordinación General (+134.7%) no era visible en los reportes mensuales, que sólo mostraban totales agregados. La capacidad de los LLMs para analizar el corpus completo y cruzar dimensiones múltiples simultáneamente constituye una ventaja metodológica significativa frente a las herramientas tradicionales.

Estos hallazgos están alineados con investigaciones paralelas en clasificación automática de incidentes en servicios de TI (Gilardi et al., 2023), gestión de quejas universitarias (Hristidis et al., 2023) y automatización de procesos de calidad educativa (Kasneci et al., 2023). La especificidad de los resultados para el contexto latinoamericano y el idioma español constituye una aportación particular de este trabajo.

7. Trabajo Futuro

Con base en los resultados obtenidos y las limitaciones identificadas, se proponen las siguientes líneas de investigación futura:

- Implementar un pipeline de clasificación automática en producción mediante API, integrado con las herramientas actuales de Excel y Power BI del departamento.
- Explorar el fine-tuning supervisado de un modelo de menor tamaño (ej. Llama-3 o Mistral) con el corpus FIME-UANL para reducir costos de inferencia y aumentar la precisión en la dimensión de Clase.
- Agregar análisis de sentimiento a las descripciones de incidentes para identificar situaciones de alta urgencia emocional o conflictos interpersonales latentes.
- Desarrollar un sistema de predicción de incidentes (forecasting) basado en los patrones temporales identificados, particularmente la concentración de incidentes en semanas de cierre de período.
- Extender el estudio a otros departamentos de Educación Continua de la red universitaria para validar la generalizábilidad del modelo de clasificación.

8. Conclusiones

El presente estudio demuestra que los Modelos de Lenguaje Grande representan una alternativa viable y eficiente para la clasificación automática de incidentes e infracciones en entornos educativos de habla hispana. Las tres plataformas evaluadas superaron el umbral de precisión del 85%, con Claude Sonnet liderando con 90.8% de precisión promedio, y lograron reducciones de tiempo del 78-81% frente al proceso manual.

El corpus histórico de 13,636 registros (2018-2024) reveló un crecimiento sostenido del volumen de incidentes del 437% en seis años, lo que hace insostenible el procesamiento manual a futuro. La automatización de la clasificación mediante LLMs no solo mejora la eficiencia operativa, sino que habilita análisis de patrones longitudinales y transversales imposibles de obtener con herramientas tradicionales.

La propuesta metodológica basada en prompt engineering few-shot no requiere inversión en infraestructura especializada ni conocimiento técnico profundo, lo que la hace accesible para instituciones educativas con recursos limitados. Se confirma así que la integración de GEN-AI en los procesos de gestión de calidad educativa ofrece ventajas competitivas significativas y cuantificables, en línea con las conclusiones de Espinoza-Uribe et al. (2024).

Referencias

Anthropic. (2024). *Claude 3.5 technical report.* Anthropic. https://www.anthropic.com/research

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33,* 1877–1901. https://doi.org/10.48550/arXiv.2005.14165

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics,* 4171–4186. https://doi.org/10.18653/v1/N19-1423

Espinoza-Uribe, J., Torres-Guerrero, F., & Neira-Tovar, L. (2024). Empoderando el análisis de datos del documento de incidentes e infracciones a través del GEN-AI: Caso estudio Educación Continua FIME – UANL. *Daena: International Journal of Good Conscience, 19*(2), 1–8.

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(30), e2305016120. https://doi.org/10.1073/pnas.2305016120

Google DeepMind. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.* Technical Report. https://arxiv.org/abs/2403.05530

He, X., Lin, Z., Gong, Y., Jin, A., Zhang, H., Lin, C., & Dolan, B. (2023). AnnoLLM: Making large language models to be better crowdsourced annotators. *arXiv preprint arXiv:2303.16854.* https://doi.org/10.48550/arXiv.2303.16854

Hristidis, V., Ruggiano, N., Brown, E. L., Ganta, S. R. R., & Stewart, S. (2023). ChatGPT vs Google for queries related to dementia and other cognitive decline: Comparison of results. *Journal of Medical Internet Research, 25,* e48966. https://doi.org/10.2196/48966

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103,* 102274. https://doi.org/10.1016/j.lindif.2023.102274

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 1–35. https://doi.org/10.1145/3560815

OpenAI. (2023). *GPT-4 technical report.* https://doi.org/10.48550/arXiv.2303.08774

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30,* 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35,* 24824–24837. https://doi.org/10.48550/arXiv.2201.11903

Los Autores

Jorge Espinoza Uribe – jorge.espinozaurb@uanl.edu.mx Francisco Torres Guerrero – francisco.torresgrr@uanl.edu.mx Leticia Neira-Tovar – leticia.neiratv@uanl.edu.mx

Universidad Autónoma de Nuevo León | Facultad de Ingeniería Mecánica y Eléctrica | San Nicolás de los Garza, N.L. México