Artificial Intelligence Applied to Cryptographic Reviews

Daniel Calvache [1], Marcelo Cipriano [1], Edith García [1], Ariel Maiorano [1], Eduardo Malvacio [1]

[1] Facultad de Ingeniería del Ejército (FIE), Universidad de la Defensa Nacional (UNDEF)
{dcalvache,mcipriano,egarcia,maiorano,emalvacio}@fie.undef.edu.ar

Abstract. This short paper presents an ongoing research line on the use of artificial intelligence for reviewing drafts, manuscripts, preprints and, in general, arguments that include results, processes, mechanisms or partial demonstrations in the context of cryptology. The practical question is whether LLM-based assistants can help detect errors, unjustified steps, hidden assumptions, and weak reductions. The main objective is to evaluate the contribution of these models as part of the expert human validation process. We consider, on the one hand, adversarial- review case studies that show useful support in technical reasoning [8], and, on the other hand, benchmark results that still reveal limitations in strict claim-to-evidence validation [6]. Based on this, we propose a reproducible workflow with structured and iterative prompting. In the experimental stage, we will run a publicly replicable protocol using the OpenAI API with a frontier model (GPT-5.4 Pro with maximum reasoning effort) over a paper previously published by the authors [7].

Keywords: cryptographic review · large language models · claim-evidence reasoning · adversarial prompting · security proof auditing

Inteligencia Artificial Aplicada a Revisiones Criptográficas

Resumen. Este artículo corto presenta una línea de investigación en desarrollo sobre el uso de inteligencia artificial para revisar borradores, manuscritos de autor, preprints y, en general, argumentos que incluyan resultados, procesos, mecanismos o demostraciones parciales en el contexto de la criptología. La pregunta central es si asistentes basados en modelos de lenguaje pueden ayudar a detectar errores, pasos no justificados, supuestos implícitos y reducciones débiles. El objetivo principal es evaluar el aporte de estos modelos como parte del proceso de validación humana experta. Consideramos, por un lado, evidencia de revisión adversarial con apoyo útil en razonamiento técnico [8] y, por otro, evaluaciones con benchmarks que aún muestran limitaciones para validar afirmaciones bajo criterios estrictos de evidencia [6]. En base a esto, proponemos un flujo reproducible con prompting estructurado e iterativo. En la etapa experimental, realizaremos una prueba replicable usando la API de OpenAI con un modelo de frontera (GPT-5.4 Pro con razonamiento máximo) sobre un artículo previamente publicado por los autores [7].

Palabras clave: revisión criptográfica · modelos de lenguaje · razonamiento afirmación-evidencia · prompting adversarial · auditoría de pruebas de seguridad

1 Introducción

La revisión de argumentos criptográficos exige verificar consistencia lógica, validez de reducciones de seguridad y correspondencia entre hipótesis y conclusiones, tanto en pruebas como en postulados, proposiciones, lemas, etc. Esta carga de trabajo es costosa y suele depender de revisiones iterativas entre coautores y, en etapas posteriores, entre pares y revisores. En este contexto, la IA generativa aparece como soporte potencial para revisar justificaciones intermedias, aunque su uso sin controles puede introducir errores difíciles de detectar.

Este trabajo plantea una línea aplicada de revisión criptográfica asistida por IA, orientada a apoyar al revisor humano sin reemplazar su juicio experto.

2 Antecedentes

Son relevantes los avances recientes sobre el uso de IA en validación científica, aunque evidencia disponible todavía muestra resultados mixtos. En [6], los autores construyen una evaluación sistemática para medir si los modelos pueden decidir correctamente cuándo una afirmación científica está respaldada por evidencia específica. Sus resultados indican brechas importantes en razonamiento de soporte, especialmente cuando la relación afirmación–evidencia requiere inferencias no triviales.

Desde otra perspectiva, [8] reporta estudios de caso de investigación asistida por modelos avanzados y constituye la referencia principal de este trabajo. En particular, el caso criptográfico allí referido, sobre SNARGs, se plantea como un escenario de auditoría técnica de una demostración compleja, donde el modelo no actúa como oráculo de corrección, sino como un generador sistemático de objeciones. El flujo reportado puede resumirse en cuatro etapas: (i) reconstrucción de la estructura de la prueba (supuestos, lemas intermedios y conclusión), (ii) búsqueda de posibles puntos frágiles en reducciones y dependencias entre afirmaciones, (iii) formulación de contraargumentos o hipótesis de fallo sobre pasos concretos, y (iv) validación humana de cada observación para separar hallazgos útiles de falsos positivos.

Como otro ejemplo complementario, en línea con el anterior, [2] presenta experimentos tempranos de aceleración científica con modelos GPT-5, mostrando colaboración efectiva en tareas de apoyo matemático y científico bajo supervisión humana. Este otro antecedente, desde parte del ecosistema de OpenAI y con modelos frontera, refuerza la hipótesis de que estos sistemas pueden aportar valor en razonamiento técnico cuando se integran en flujos verificables por expertos.

Como publicación más reciente, [4] reporta resultados de Google DeepMind con un agente basado en Gemini para el desafío FirstProof, indicando que resolvió 6 de 10 pruebas del benchmark. Este resultado aporta también evidencia complementaria de que los modelos frontera pueden ofrecer apoyo útil en tareas de razonamiento formal, siempre que el proceso incorpore criterios de evaluación explícitos y revisión experta.

En conjunto con la evidencia de [6] y con reportes sobre límites de revisores automáticos para detectar razonamientos defectuosos [3], estos resultados sugieren que la utilidad práctica aumenta cuando la IA se usa para generar hipótesis de fallo y no como verificador final autónomo.

En síntesis, estos antecedentes sugieren que la principal contribución de la IA está en ampliar la capacidad de revisión crítica (supuestos implícitos, saltos argumentales y justificaciones débiles), mientras que la decisión final de validez se mantiene bajo control de revisores con conocimiento especializado.

Aunque puede entenderse como un aspecto o aplicación diferente, también resultan relevantes los avances recientes en revisión de código fuente de alto impacto en seguridad, con reportes de hallazgos significativos en proyectos críticos cuando se combina automatización con validación experta [1,5]. En relación directa con este trabajo, estos antecedentes son importantes porque sugieren una continuidad metodológica: si un modelo puede interpretar definiciones criptográficas formales y razonar sobre sus propiedades (requerimientos, restricciones o límites, etc.), entonces probablemente también pueda asistir en la revisión de código fuente para verificar si esas definiciones y mecanismos están correctamente implementados.

3 Línea de investigación

La línea de investigación se centra, de manera práctica, en usar la IA como apoyo para detectar saltos lógicos y justificaciones incompletas en pruebas criptográficas, postulados y proposiciones, manteniendo trazabilidad entre afirmaciones, lemas y supuestos de seguridad, y comparando esos resultados con una revisión humana.

Como prueba aplicada, realizaremos un experimento con un modelo LLM de frontera, el mejor disponible al momento de la confección de este trabajo. Se trata de un modelo públicamente accesible, aunque no gratuito, de OpenAI. La configuración elegida es: plataforma OpenAI API, modelo `gpt-5.4-pro`, esfuerzo de razonamiento `xhigh` (nivel máximo). El objetivo es analizar un artículo publicado previamente por los autores [7] y emplearlo como caso de prueba metodológica.

El diseño contempla reconstruir una prueba central del artículo original, ejecutar un protocolo de revisión asistida por IA (modo revisor crítico y modo verificador de supuestos), y comparar contra la versión inicial para identificar correcciones, justificaciones fortalecidas o falsos positivos. Las salidas se registrarán junto con parámetros de ejecución para facilitar su replicabilidad por terceros.

Como parte de este protocolo, utilizaremos prompts explícitos de revisión adversarial, tomando como base el enfoque de prompting reportado en [8]. Un ejemplo inicial sería: “Actuá como un revisor adversarial de artículos académicos relativos a criptografía. Dado el siguiente artículo (el archivo que ha sido subido), realizá una revisión inicial estrictamente objetiva, enfocándote sólo en la detección de errores o problemas.” Luego, iterativamente, continuaría: ”Corregí tu primera revisión revisando y criticando rigurosamente tus propios hallazgos.", ”Realizá una segunda ronda de revisión, incorporando tus correcciones, para refinar la revisión y considerar al artículo como un todo.", y por último ”En base a lo anterior, generá una revisión final como devolución o feedback para los autores.”

Esta prueba permitirá cuantificar utilidad práctica (tiempo de revisión, costo y cobertura de hallazgos) y riesgos (alucinaciones argumentales u objeciones infundadas), generando evidencia para iterar el protocolo.

Resultados de la prueba de ejemplo

Los parámetros utilizados en platform.openai.com fueron: *model: gpt-5.4-pro-2026-03-05, text.format: text, eflort: xhigh, verbosity: low, summary: concise y store: true.*

El tiempo total de ejecución fue de aproximadamente 50 minutos, incluyendo uno o dos minutos luego de cada respuesta sólo para confirmar que la ejecución hubiera sido exitosa. El costo total fue de aproximadamente 16 dólares estadounidenses.

Concretamente Los resultados incluyeron por ejemplo lo que entendió como errores en la presentación el algoritmo (textualmente):

> ... falta la parte esencial de la especificación del cifrador: carga de clave, carga de IV, fase de inicialización/warm-up.

Lo que podría considerarse estrictamente correcto, aunque las dos primeras se darían por sobre-entendido. De forma similar identificó potenciales problemas en el pseudocódigo incluido en el artículo, también correctas si se considerara estrictamente.

Respecto a correcciones que entenderíamos no válidas, pudieron identificarse, por ejemplo:

- Error factual/referencial: Trivium se describe como si generara “at least 2^{64} bits”.
- Espacio de estados no equivalente al cifrador real: el estado inicial arbitrario no se restringe a estados alcanzables por el key/IV loading estándar.
- Inconsistencia en Trivium-Toy (ec. 9): $Y_0 = X_{21} \oplus X_{30} \oplus X_{29}X_{28} \oplus Y_{25}$, pero en pseudocódigo se usa $s55$; si $Y = (s31, \ldots, s58)$, entonces $Y_{24} = s55$ y $Y_{25} = s56$.

4 Conclusiones preliminares y trabajos futuros

Como conclusión preliminar, la literatura sugiere que la IA puede aportar valor en la detección temprana de inconsistencias, pero su desempeño sigue siendo heterogéneo cuando la validez depende de justificaciones específicas. Por ello, la hipótesis central de este proyecto es que el beneficio no proviene de la automatización completa, sino de una colaboración estructurada entre modelo y revisor experto (humano).

Como trabajos futuros se prevén: ampliar la evaluación a múltiples tipos de pruebas criptográficas (por ejemplo argumentos de indistinguibilidad), definir un esquema de reporte replicable para revisiones asistidas, y consolidar recomendaciones metodológicas para equipos de investigación que busquen incorporar IA sin comprometer el rigor formal.

Declaración sobre uso de IA

Durante la preparación de este trabajo, utilizamos diferentes servicios de OpenAI con el fin de mejorar el lenguaje y la legibilidad del manuscrito (y por supuesto la realización de la prueba principal informada en este trabajo). Tras utilizar estos servicios, revisamos y editamos el contenido según fue necesario y asumimos plena responsabilidad por el contenido de la publicación.

Referencias

1. AISLE Research. (2026). What AI Security Research Looks Like When It Works. https://aisle.com/blog/what-ai-security-research-looks-like-when-it-works

2. Bubeck, S., Coester, C., Eldan, R., Gowers, T., Lee, Y. T., Lupsasca, A., Sawhney, M., Scherrer, R., Sellke, M., Spears, B. K., Unutmaz, D., Weil, K., Yin, S., & Zhivotovskiy, N. (2025). Early science acceleration experiments with GPT-5. arXiv preprint arXiv:2511.16072. https://doi.org/10.48550/arXiv.2511.16072

3. Dycke, N., & Gurevych, I. (2025). Automatic Reviewers Fail to Detect Faulty Reasoning in Research Papers: A New Counterfactual Evaluation Framework. arXiv preprint arXiv:2508.21422. https://doi.org/10.48550/arXiv.2508.21422

4. Feng, T., Jung, J., Kim, S., Pagano, C., Gukov, S., Tsai, C., Woodruff, D., Javanmard, A., Mokhtari, A., Hwang, D., Chervonyi, Y., Lee, J. N., Bingham, G., Trinh, T. H., Mirrokni, V., Le, Q. V., & Luong, T. (2026). Aletheia tackles FirstProof autonomously. arXiv preprint arXiv:2602.21201. https://doi.org/10.48550/arXiv.2602.21201

5. Fort, S. (2026). AI found 12 of 12 OpenSSL zero-days (while curl cancelled its bug bounty). LessWrong. https://www.lesswrong.com/posts/7aJwgbMEiKq5egQbd/ai-found-12-of-12-openssl-zero-days-while-curl-cancelled-its

6. Javaji, S. R., Cao, Y., Li, H., Yu, Y., Muralidhar, N., & Zhu, Z. (2025). Can AI Validate Science? Benchmarking LLMs for Accurate Scientific Claim-to-Evidence Reasoning. arXiv preprint arXiv:2506.08235. https://doi.org/10.48550/arXiv.2506.08235

7. Castro Lechtaler, A., Cipriano, M., García, E., Liporace, J., Maiorano, A., & Malvacio, E. (abril 2014). Model Design for a Reduced Variant of a Trivium Type Stream Cipher. Journal of Computer Science & Technology, 14(1), 55–58. Facultad de Informática. ISSN 1666-6038 http://sedici.unlp.edu.ar/handle/10915/34550

8. Woodruff, D. P., Cohen-Addad, V., Jain, L., Mao, J., Zuo, S., Bateni, M., Branzei, S., Brenner, M. P., Chen, L., Feng, Y., Fortnow, L., Fu, G., Guan, Z., Hadizadeh, Z., Hajiaghayi, M. T., JafariRaviz, M., Javanmard, A., S., K. C., Kawarabayashi, K., Kumar, R., Lattanzi, S., Lee, E., Li, Y., Panageas, I., Paparas, D., Przybocki, B., Subercaseaux, B., Svensson, O., Taherijam, S., Wu, X., Yogev, E., Zadimoghaddam, M., Zhou, S., Matias, Y., Manyika, J., Mirrokni, V. (2026). Accelerating Scientific Research with Gemini: Case Studies and Common Techniques. arXiv preprint arXiv:2602.03837. https://doi.org/10.48550/arXiv.2602.03837