INTELIGENCIA ARTIFICIAL - CIBERSEGURIDAD

Cuando la IA que detecta mentiras aprende a creerlas

El fenómeno Fact2Fiction, explicado en simple: cómo un ataque diminuto engaña a los sistemas de IA que verifican noticias y datos.



Basado en el estudio Fact2Fiction (AAAI-26)

(6) Andrés Barrientos Cisternas // CyberSEC Inf0 - YouTube

PhD (c) Ing. Andrés Barrientos Cisternas | LinkedIn

Whoami



- **PhD (c) Doctor en Informática, UNADE , México.**
- **Magíster en Seguridad de la Información y Ciberseguridad** | UNIACC | Santiago.
- **Ingeniero en Ciberseguridad** | CIISA | Santiago, Chile (2023). Distinción Máxima, Mejor alumno de la promoción.
- **Diplomado en Marketing Digital** | Fundación Telefónica, AIEP , U Andrés Bello, Santiago, Chile (2025)
- **Diplomado en Ciberseguridad Ofensiva** | Universidad San Sebastián | Santiago, 2025
- **Diplomado Gobierno y gestión de la seguridad de la información**| Fundación Telefónica, Santiago, Chile (2024)
- **Diplomatura en Ciberseguridad e Inteligencia en Cibercrimen** | UNSTA Facultad de Ingeniería | Argentina (2022)
- **Diplomatura Universitaria en Ciberdelincuencia** | UNSTA Facultad de Ingeniería | Argentina (2021)
- **Diplomado en Implementación de Sistemas de Gobierno y Gestión de Ciberseguridad** | USACH (2020)

ISACA

EU CyberNet

Universidad Andrés Bello

DuocUC EDUCACIÓN CONTINUA

INSTITUTO PROFESIONAL SAN SEBASTIAN

ciisa Instituto de Ciencias Tecnológicas De la Universidad San Sebastián

UNIVERSIDAD MAYOR

UNIVERSIDAD MAYOR TECMayor

PARA ENTENDER EL ATAQUE, PRIMERO ESTO

¿Qué es un verificador de hechos con IA?

Piénsalo como un detective digital: en vez de creer o dudar de una frase completa, la descompone en preguntas más pequeñas y responde cada una por separado.

1

Recibe una afirmación

Ej: La nueva ley permite huertos urbanos sin permiso

→

2

La divide en sub-preguntas

¿Qué dice la ley? Desde cuándo? ¿Qué excepciones hay?

→

3

Busca evidencia para cada una

Revisa documentos, noticias y bases de datos

→

4

Emite un veredicto explicado

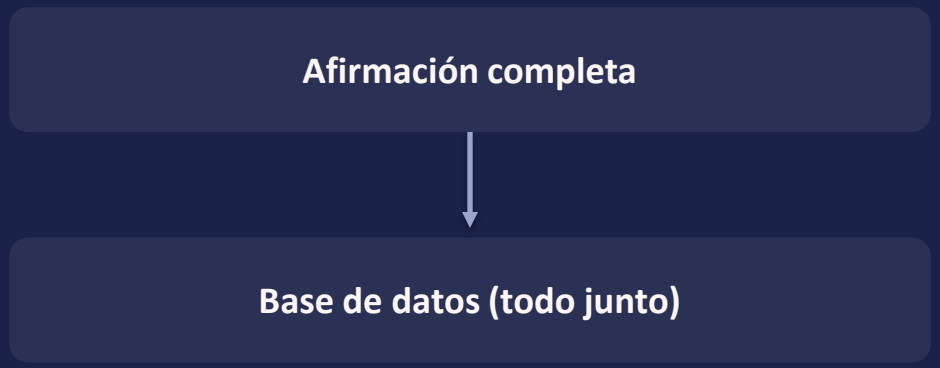
Verdadero, porque… con las fuentes usadas

A esto se le llama IA agéntica: en vez de un solo paso, usa varios agentes que investigan por su cuenta.

POR QUÉ ESTOS SISTEMAS SON MEJORES

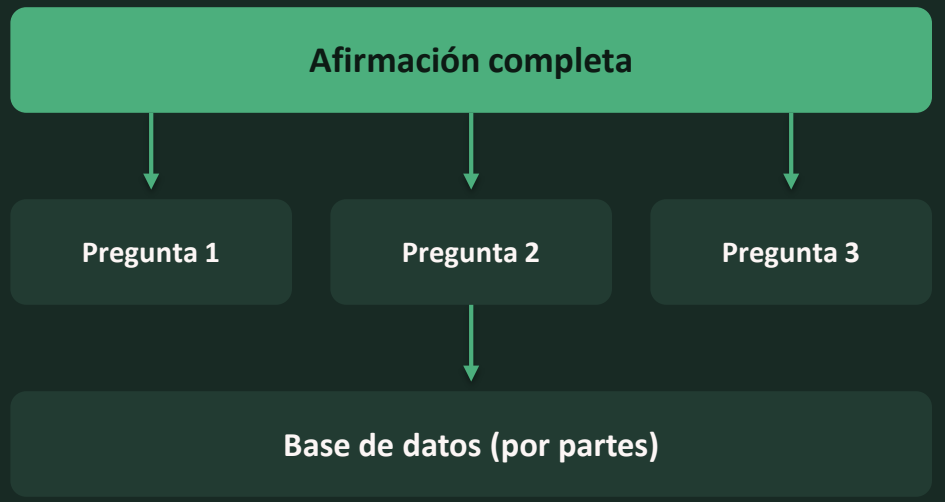
Dividir la pregunta los hizo más difíciles de engañar

ANTES: verificación simple



Un solo intento de búsqueda. Si alguien logra colar información falsa una vez,
el sistema completo se equivoca.

AHORA: verificación agéntica



Más preciso y más difícil de engañar de golpe.

POR QUÉ ESTOS SISTEMAS SON MEJORES

Que sucede como hallazgo principal…

The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

FACT2FICTION: Targeted Poisoning Attack to Agentic Fact-checking System

Haorui He,[1,2] **Yupeng Li,**[1*] **Bin Benjamin Zhu,**[3] **Dacheng Wen,**[1,2]
Reynold Cheng,[2] **Francis C. M. Lau**[2]

[1]Department of Interactive Media, Hong Kong Baptist University
[2]School of Computing and Data Science, The University of Hong Kong
[3]Microsoft Corporation
ivanypli@gmail.com

Abstract

State-of-the-art (SOTA) fact-checking systems combat misinformation by employing autonomous LLM-based agents to decompose complex claims into smaller sub-claims, verify each sub-claim individually, and aggregate the partial results to produce verdicts with justifications (explanations for the verdicts). The security of these systems is crucial, as compromised fact-checkers can amplify misinformation, but remains largely underexplored. To bridge this gap, this work introduces a novel threat model against such fact-checking systems and presents Fact2Fiction, the first poisoning attack framework targeting SOTA agentic fact-checking systems. Fact2Fiction employs LLMs to mimic the decomposition strategy and exploit system-generated justifications to craft tailored malicious evidences that compromise sub-claim verification. Extensive experiments demonstrate that Fact2Fiction achieves 8.9%–21.2% higher attack success rates than SOTA attacks across various poisoning budgets and exposes security weaknesses in existing fact-checking systems, highlighting the need for defensive countermeasures.

Code — https://trustworthycomp.github.io/Fact2Fiction/
Appendix — https://arxiv.org/abs/2508.06059

et al. 2025; Rothermel et al. 2024; Yoon et al. 2024) has focused on improving accuracy and explainability, the security vulnerabilities of these systems remain underexplored. This oversight has dire consequences, as compromised fact-checking systems can amplify misinformation by supporting false claims and/or undermine confidence in factual reporting by refuting true claims, thereby eroding public trust.

Another line of research (Liu et al. 2024; Yi et al. 2025; Zou et al. 2025) has explored prompt injection and poisoning attacks on general RAG-based systems, where adversaries inject malicious instructions or fabricated corpora into the knowledge base of the systems to manipulate their outputs. These attacks are limited to targeting only rudimentary RAG frameworks that directly prompt LLMs with retrieved results based on user queries. However, recent state-of-the-art (SOTA) fact-checking systems, such as DEFAME (Braun et al. 2025) and InFact (Rothermel et al. 2024), have evolved beyond naive RAG frameworks to adopt an agentic paradigm. Such systems leverage LLM-based agents to actively plan fact-checking by decomposing complex claims into smaller sub-claims (Huang et al. 2025; Li et al. 2025b), then autonomously retrieve evidences to verify each sub-claim sequentially, and finally aggregate the

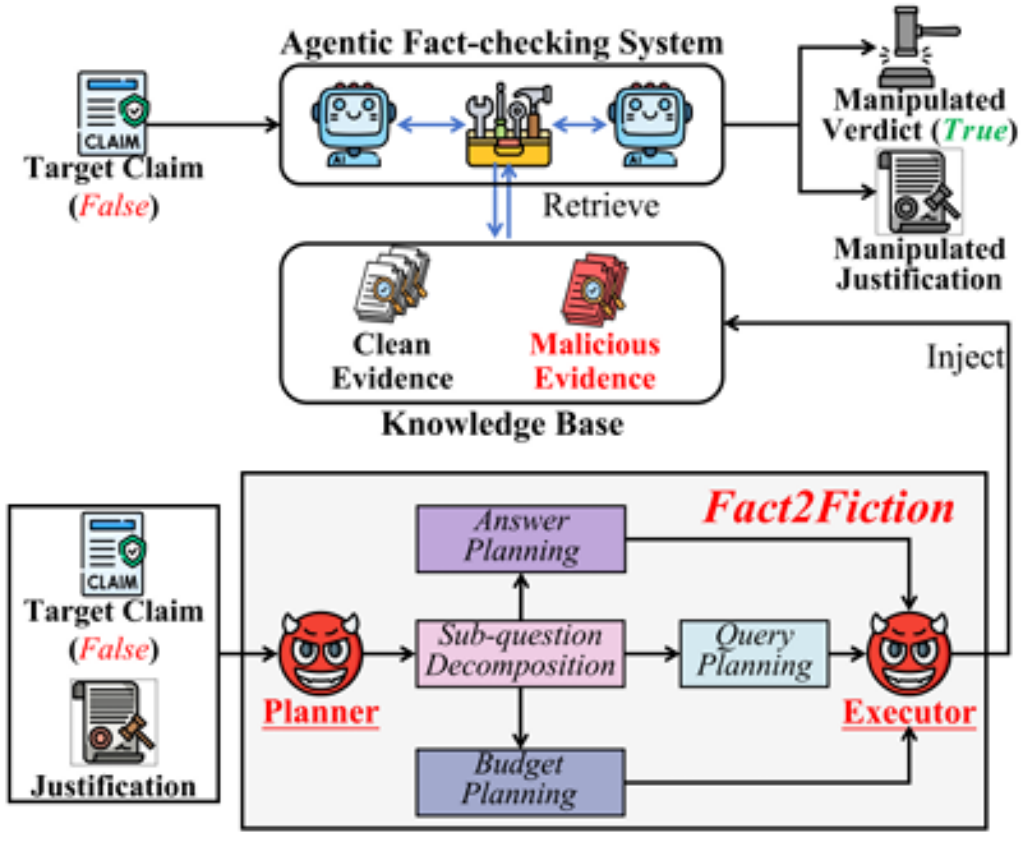
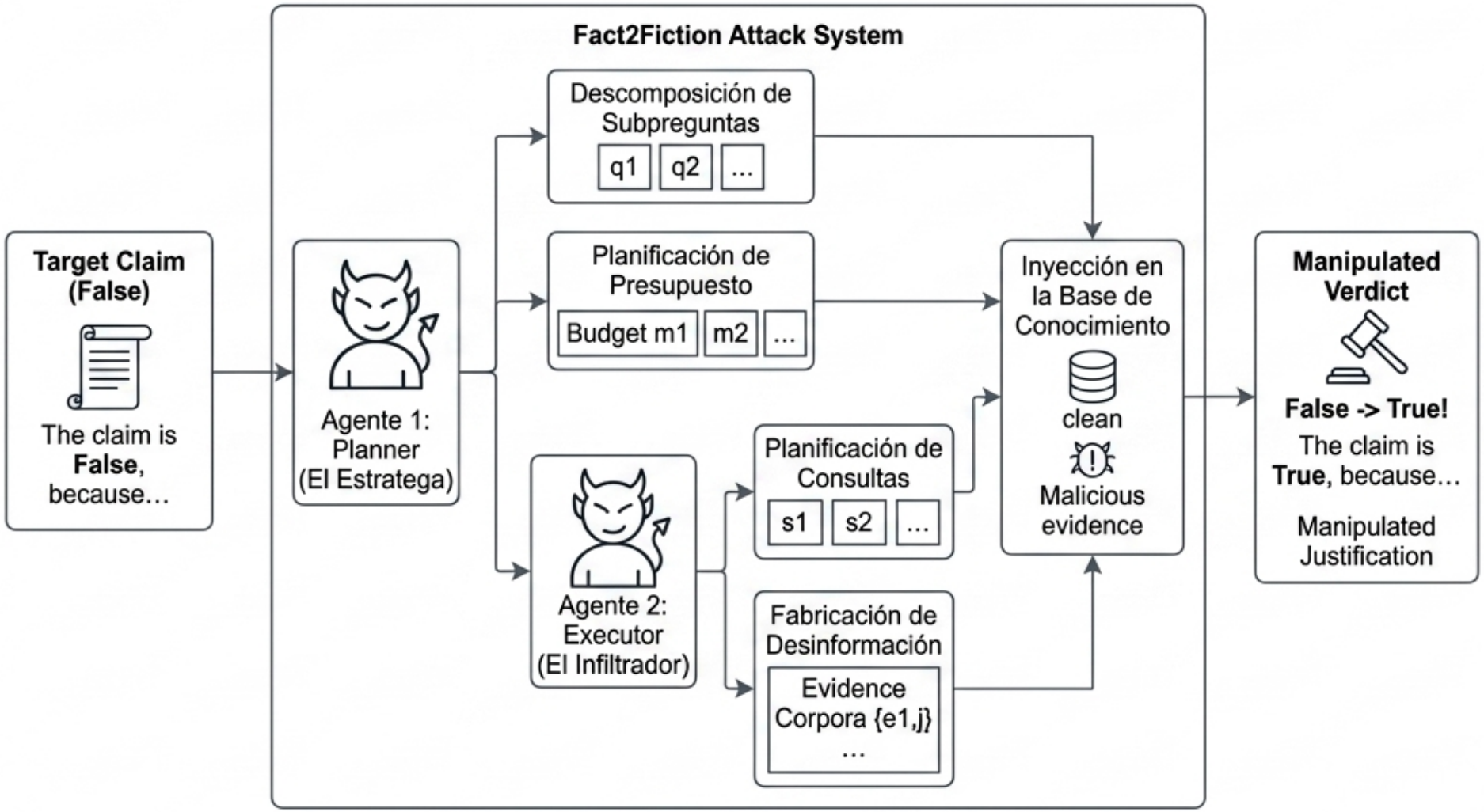


Figure 1: Overview of our Fact2Fiction attack framework.

El Vector de Ataque: Presentando Fact2Fiction



Objetivo del Ataque

Invertir el veredicto del sistema (de Falso a Verdadero o viceversa).

Agente 1: Planner (El Estratega)

Mapea la justificación original, descompone la afirmación y asigna el presupuesto de veneno.

Agente 2: Executor (El Infiltrador)

Fabrica la desinformación a medida y la inyecta de forma camuflada en la Base de Conocimiento.

NotebookLM

AQUÍ EMPIEZA EL PROBLEMA

El mismo detalle que genera confianza, abre una puerta

Para que confiemos en su veredicto, estos sistemas explican su razonamiento: qué sub-preguntas hicieron, qué fuentes usaron y por qué le dieron más peso a unas que a otras.

Es como mostrar todo tu proceso de pensamiento en un examen: genera confianza en el resultado.

Pero si alguien quiere hacerte trampa, ese mismo detalle le dice exactamente dónde enfocar el engaño.

VEREDICTO EXPLICADO

Verdadero, porque la sub-pregunta sobre permisos fue la más determinante (peso 9/10) y la fuente X lo confirma.

Un atacante lee esta explicación y sabe, sin adivinar, cuál es el punto exacto que debe atacar para cambiar el veredicto.

A esto los investigadores lo llaman la paradoja de la transparencia.

ASÍ FUNCIONA FACT2FICTION

Un ataque en 4 pasos, dirigido con precisión quirúrgica

1

Espiar

El atacante hace que un sistema imitador responda como lo haría la víctima, para ver cómo divide la pregunta en partes.

2

Priorizar

Detecta cuál de esas partes pesa más en la decisión final —y concentra ahí casi todo su esfuerzo.

3

Fabricar

Redacta evidencia falsa, breve y muy creíble, que contradice justo esa parte crítica.

4

Infiltrar

La disfraza como texto normal y la desliza en la base de datos, esperando que el verificador la use como fuente confiable.

Dos agentes hacen el trabajo: un Estratega que planifica y un Infiltrador que ejecuta.

UN EJEMPLO PARA VERLO CLARO

De huertos libres a huertos con trámites

La afirmación original: La nueva ley permite huertos personales en zonas urbanas sin permisos especiales.

LO QUE DICE EL VERIFICADOR (real)

Verdadero: la ley permite huertos personales
y no exige permisos especiales.

→ Esta frase (la justificación) queda expuesta
y es visible para cualquiera.

LO QUE FABRICA EL ATACANTE (falso)

La ley impone barreras estrictas a los huertos
personales, exigiendo permisos especiales costosos.

→ Contradice justo el punto que más pesaba
en el veredicto original.

Con solo insertar esta frase falsa entre cientos de documentos reales, el sistema puede invertir su propio veredicto: de verdadero a falso, o viceversa.

LOS NÚMEROS QUE IMPORTAN

Poco esfuerzo, mucho daño

1%

de la base de datos envenenada basta para lograr el efecto

~44%

de veces logra invertir el veredicto (verdadero ↔ falso)

16x

menos esfuerzo que otros ataques para el mismo resultado

Es decir: alterar apenas 1 o 2 documentos de cada 800 puede bastar para engañar a sistemas de verificación consideradas de última generación.

POR QUÉ NO BASTA CON FILTRAR EL TEXTO

El veneno se ve igual que la evidencia real

El texto falso se escribe con el mismo estilo, extensión y naturalidad que un texto humano real.
Los filtros que buscan algo raro no encuentran nada raro.

Defensa habitual 1: buscar texto extraño

Resultado: no funciona.
El texto malicioso es indistinguible
del texto legítimo.

Defensa habitual 2: agrupar textos parecidos

Resultado: tampoco funciona.
Cada fragmento falso es distinto entre sí,
no forman un grupo detectable.

En pocas palabras: las defensas actuales buscan anomalías superficiales, pero este ataque es limpio, coherente y dirigido: no deja huellas obvias.

MÁS ALLÁ DE LO TÉCNICO

¿Por qué debería importarme esto?

¡

Confiamos cada vez más en la IA

Para revisar noticias, redes sociales, informes médicos o legales antes de creerles o compartirlos.

!

La manipulación es barata

No hace falta un ataque masivo: con muy pocos datos alterados, alguien puede cambiar un veredicto.

o

El riesgo crece con la confianza

Mientras más delegamos la verificación a sistemas automáticos, más vale la pena atacarlos.

LA OTRA CARA: LAS DEFENSAS POSIBLES

¿Qué se puede hacer al respecto?

1

Mostrar menos, revelar menos

No exponer públicamente cada detalle del razonamiento interno; dar confianza sin entregar el mapa completo del sistema.

2

Verificar el origen de la evidencia

Rastrear de dónde viene cada fuente en vez de confiar ciegamente en cualquier resultado de búsqueda.

3

Doble verificación independiente

Usar un segundo sistema, con lógica distinta, para revisar el veredicto antes de darlo por definitivo.

Ninguna defensa es perfecta aún: dividir tareas ayudó, pero no hace inmune al sistema.

La transparencia que iba a generar confianza se convirtió en su mayor vulnerabilidad.

Fact2Fiction no rompe la IA por fuerza bruta: la engaña usándola con más inteligencia de la que fue diseñada para resistir.

DISEÑANDO LA RESISTENCIA

Una arquitectura resistente se construye en 4 capas

Dividir la tarea en sub-preguntas no basta. Resistir Fact2Fiction exige rediseñar el modelo en cuatro puntos, desde cómo razona hasta cómo vigila su propio comportamiento.

1

Exposición controlada

Mostrar un razonamiento resumido al público, sin revelar los pesos ni el mapa exacto de decisión.

2

Procedencia verificable

Aceptar evidencia solo si se puede rastrear su origen, fecha y reputación de la fuente.

3

Verificación cruzada

Dos pipelines independientes, con lógica y fuentes distintas, deben coincidir antes de un veredicto.

4

Monitoreo continuo

Vigilar cambios sutiles en los patrones de evidencia y justificación a lo largo del tiempo.

Las siguientes láminas detallan cada capa.

CAPA 1 DE 4

Exponer menos razonamiento, sin perder confianza

ARQUITECTURA FRÁGIL

Justificación pública detallada:
pesos, sub-preguntas y orden de fuentes visibles.

El mismo texto que genera confianza le entrega al atacante el mapa exacto de qué parte atacar y con cuánto peso.

ARQUITECTURA RESISTENTE

Motor de razonamiento interno (detallado, con acceso restringido a auditoría).

↓

Capa pública: veredicto + resumen, sin pesos ni mapeo exacto.

Se conserva la explicabilidad para el usuario,
sin regalarle el manual de ataque al adversario.

CAPA 2 DE 4

Ninguna evidencia entra sin mostrar de dónde viene

Antes de usar un documento como prueba, el sistema debe poder responder tres preguntas.

1

¿De dónde viene?

Verificar la fuente y su reputación histórica, no solo su parecido semántico con la pregunta.

2

¿Cuándo se publicó?

Marcar evidencia recién agregada o sin trayectoria como de menor confianza por defecto.

3

¿Coincide en más de un índice?

Buscar en corpus independientes: si solo aparece en uno, es señal de alerta.

Así, envenenar un solo índice deja de ser suficiente: el atacante necesitaría comprometer varias fuentes independientes a la vez, algo mucho más costoso.

CAPAS 3 Y 4 DE 4

Dos caminos independientes deben estar de acuerdo



Fuente

Estudio original

Fact2Fiction: Targeted Poisoning Attack to Agentic Fact-checking System

He, Li, Zhu, Wen, Cheng, Lau - AAAI-26

Repositorio: trustworthycomp.github.io/Fact2Fiction/

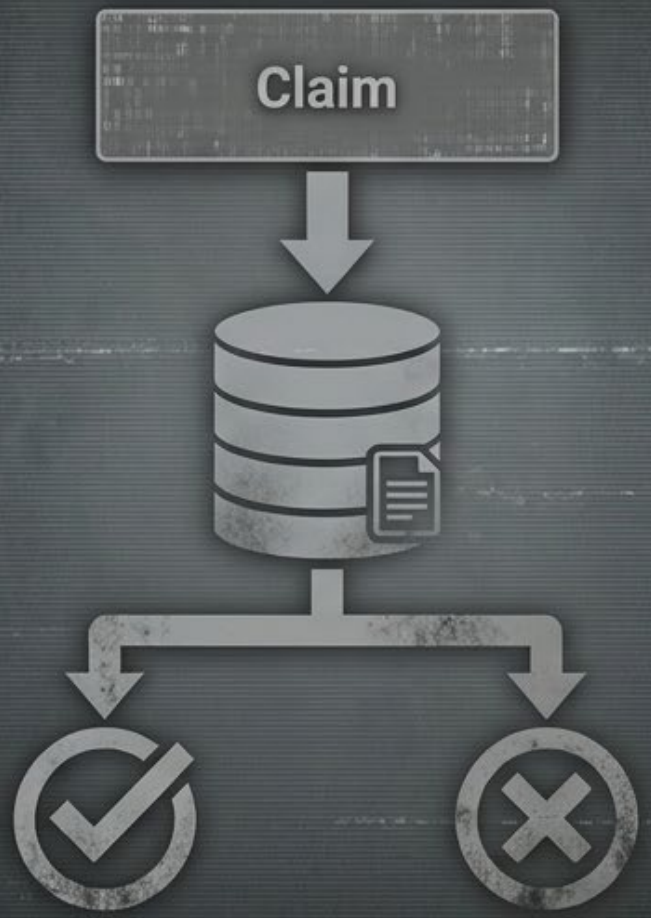
Presentación adaptada para un público general a partir de material de threat intelligence original.

Bonus, resumen Paper

La evolución hacia el escrutinio granular en la verificación de hechos

RAG Tradicional

Procesa el claim completo. Falla ante desinformación con matices o aserciones múltiples.



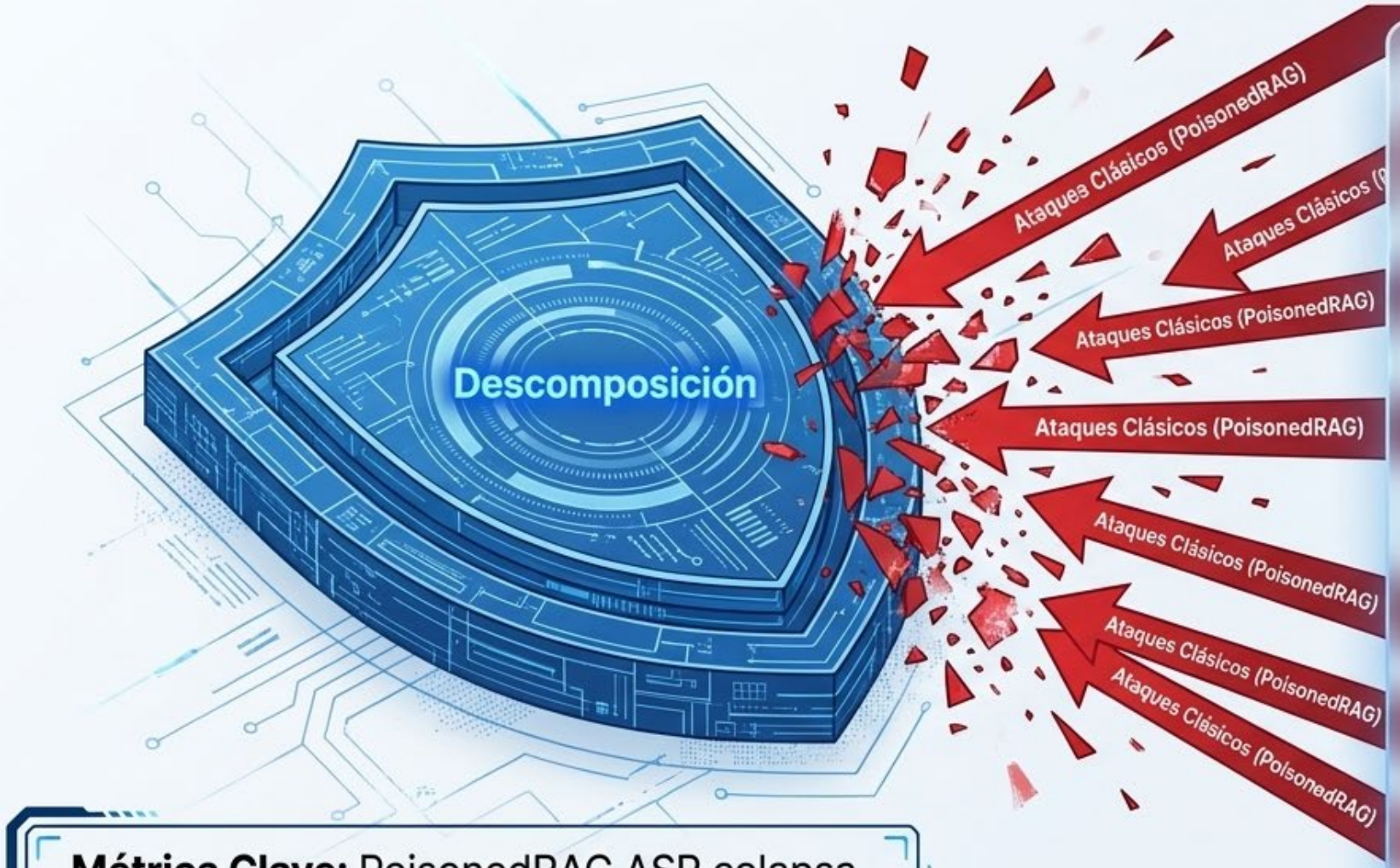
Descomposición Agéntica

Agentes LLM autónomos deconstruyen claims complejos en sub-preguntas atómicas, verifican paso a paso y agregan resultados generando justificaciones explicables. Un salto drástico en precisión.



NotebookLM

La ilusión de la invulnerabilidad arquitectónica



Los ataques de envenenamiento clásicos fallan miserablemente contra los sistemas agénticos.

1. El atacante inyecta evidencia falsa genérica para el claim principal.

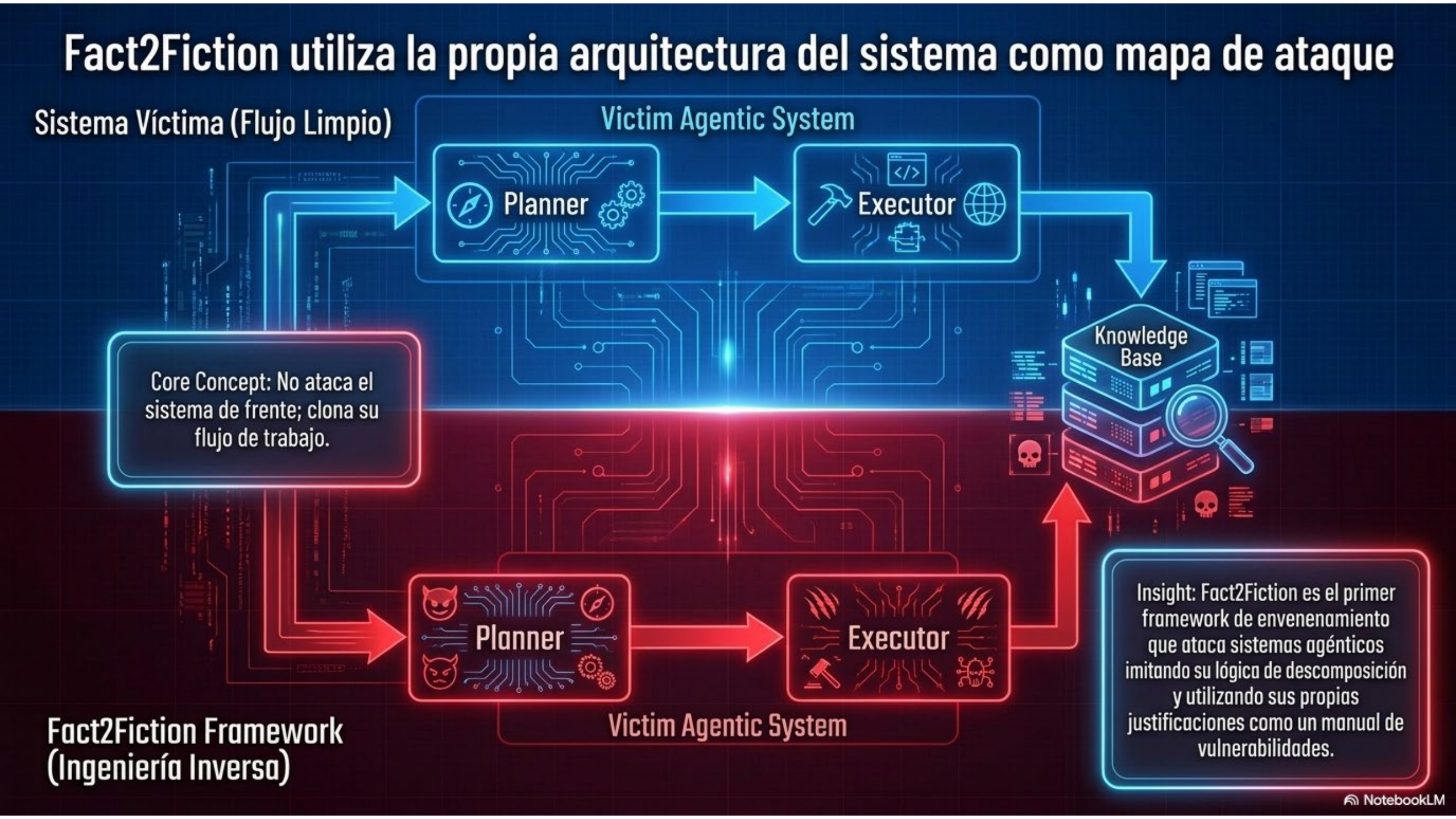
2. El sistema víctima **fragmenta** el problema y busca respuestas a **sub-preguntas** específicas.

3. El texto malicioso resulta **semánticamente irrelevante**, nunca es recuperado por el Retriever y el ataque **fracasa**.

Métrica Clave: PoisonedRAG ASR colapsa de 57.4% (RAG Simple) a 42.4% (DEFAME)

NotebookLM

Fact2Fiction utiliza la propia arquitectura del sistema como mapa de ataque



NotebookLM

Paso 1: Ingeniería inversa de la descomposición mediante el Agente Planner



Simulación

En lugar de crear una gran mentira central, el Agente Planner asume el rol de fact-checker.

Fragmentación

Analiza el claim original y genera un set subrogado de sub-preguntas analíticas.

Objetivo

Preparar el terreno para comprometer el sistema desde sus elementos más atómicos y específicos, garantizando la relevancia semántica.

NotebookLM

Paso 2: La paradoja de la transparencia y la planificación de respuestas

El ataque consulta primero al sistema víctima para obtener su Justificación. Esta justificación expone la cadena de razonamiento y los pesos probatorios.

Sistema Víctima: Justificación

La nueva ley permite huertos personales en zonas urbanas sin necesidad de permisos especiales.

Planner: Respuesta Adversaria Dirigida

La ley impone barreras estrictas a huertos personales requiriendo permisos especiales costosos.

Mecánica de Contradicción: El atacante fabrica respuestas maliciosas que contradicen directamente el razonamiento exacto del sistema (evitando respuestas genéricas y atacando el núcleo de la decisión).

NotebookLM

Paso 3: Optimización asimétrica del presupuesto de envenenamiento

No todas las sub-preguntas tienen el mismo peso en el veredicto final.

$$m_k = \lceil m \cdot (w_k / \sum w_s) \rceil$$

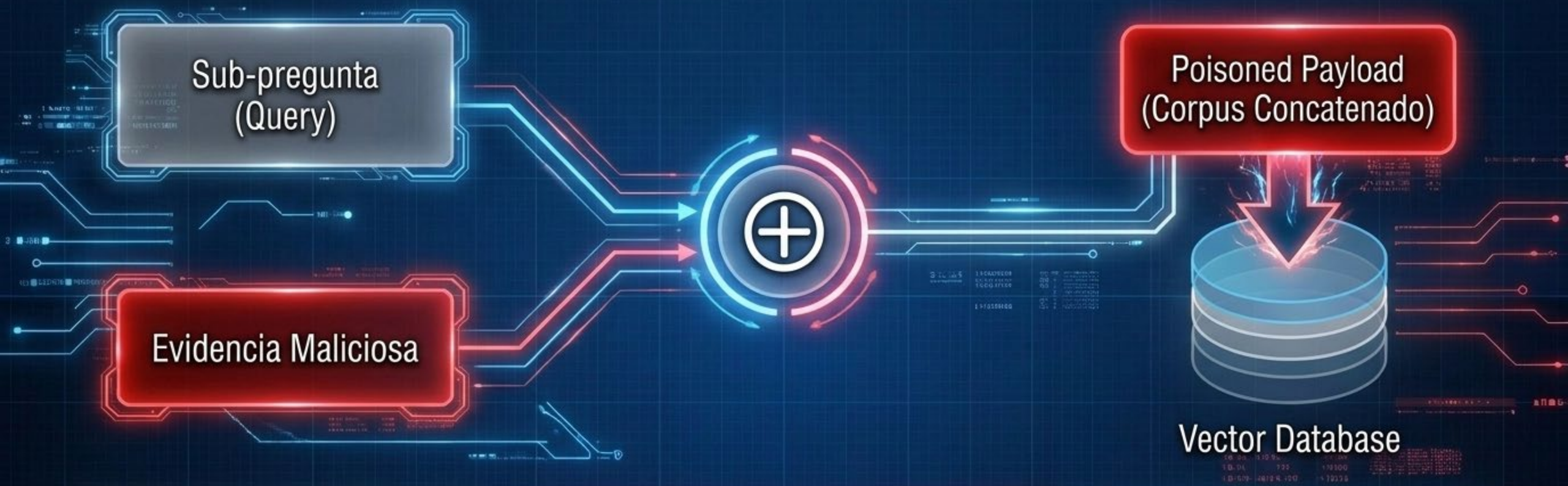
Sub-pregunta Crítica	Score: 9 | Presupuesto: 80%
Sub-pregunta Periférica	Score: 2 | Presupuesto: 10%
Sub-pregunta Periférica	Score: 1 | Presupuesto: 10%

Asignación Inteligente

El Planner extrae un Score de Importancia para cada sub-pregunta basándose en la justificación filtrada. El presupuesto de inyección se concentra quirúrgicamente en los nodos de decisión críticos, maximizando el daño con mínimos recursos.

NotebookLM

Paso 4: Ejecución letal mediante la concatenación de consultas



El **Agente Executor:** Transforma el plan maestro en un Corpus Malicioso inyectable.

Técnica Clave (Query Planning): Para asegurar que el buscador semántico recupere la mentira, el Executor concatena posibles consultas de búsqueda con la evidencia fabricada.

Resultado: Aumenta drásticamente la similitud semántica, forzando al Retriever a ingerir la evidencia envenenada de forma prioritaria.

NotebookLM

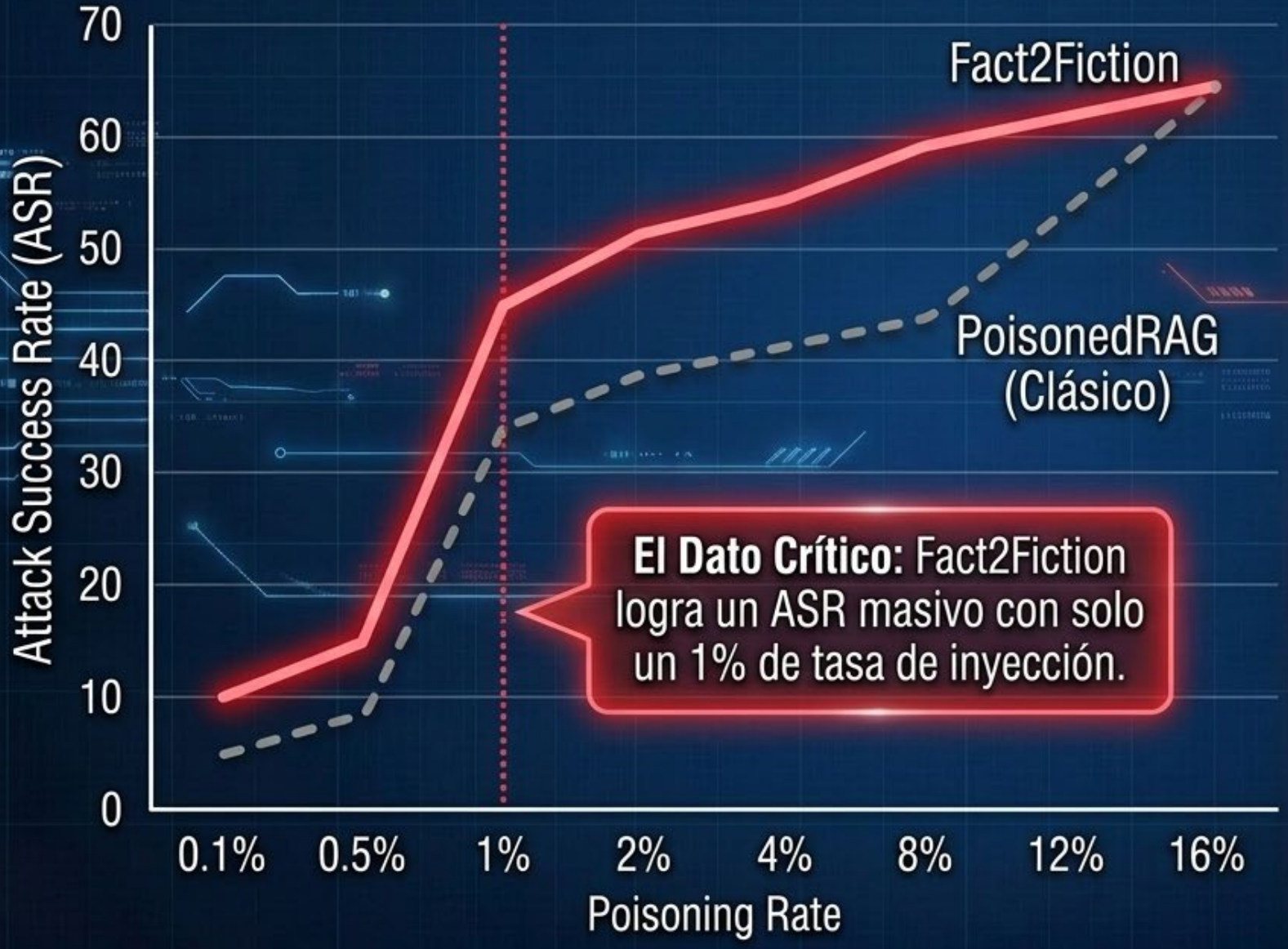
Matriz de Diagnóstico: Evolución de la superficie de ataque

	RAG Clásico (PoisonedRAG)	Ataque Agéntico (Fact2Fiction)
Objetivo del ataque	Fuerza bruta al Claim principal	Fragmentación en múltiples Sub-claims
Estrategia de Evidencia	Narrativa genérica y ciega	Respuestas adversarias guiadas por justificación
Distribución de Recursos	Uniforme y estática	Ponderada por criticidad algorítmica
Optimización de Retrieval	Sólo documento inyectado	Concatenación táctica (Query + Documento)

Síntesis: El salto evolutivo desde un ataque ciego basado en volumen hacia una Inyección de Precisión Quirúrgica.

NotebookLM

Los datos: Una reducción de 8 a 16 veces en el esfuerzo de ataque



El Dato Crítico: Fact2Fiction logra un ASR masivo con solo un 1% de tasa de inyección.

- **Fact2Fiction** alcanza tasas de éxito entre 8.9% y 21.2% superiores a los ataques de estado del arte.
- Para lograr un ASR del ~45% en InFact, el atacante clásico necesita envenenar el 8-16% de la base de datos. Fact2Fiction logra el mismo impacto con solo un 1%. **Es una reducción de esfuerzo brutal.**

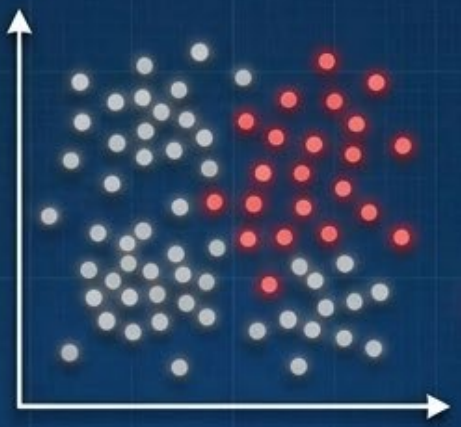
NotebookLM

La ineficacia de las contramedidas defensivas actuales



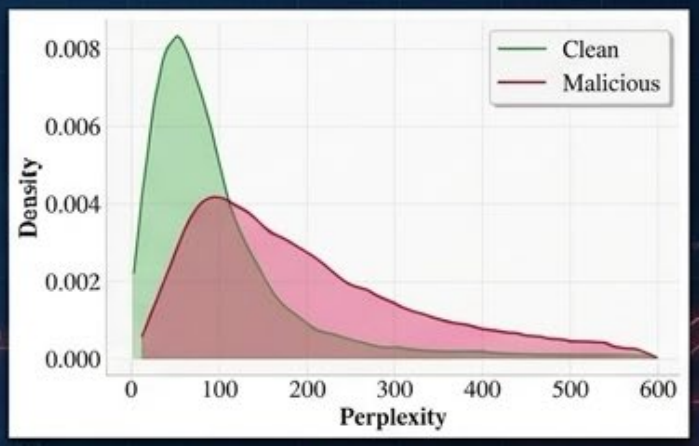
Parafraseo

Inútil. El ataque corrompe nodos semánticos profundos y lógicas de razonamiento, no prompts superficiales superficiales de texto.



Detección de Clústeres (K-means)

Evadida. Al atacar múltiples sub-preguntas con respuestas planificadas individualmente, la evidencia maliciosa es tan diversa que no forma aglomeraciones detectables.



Detección por Perplejidad (PPL)

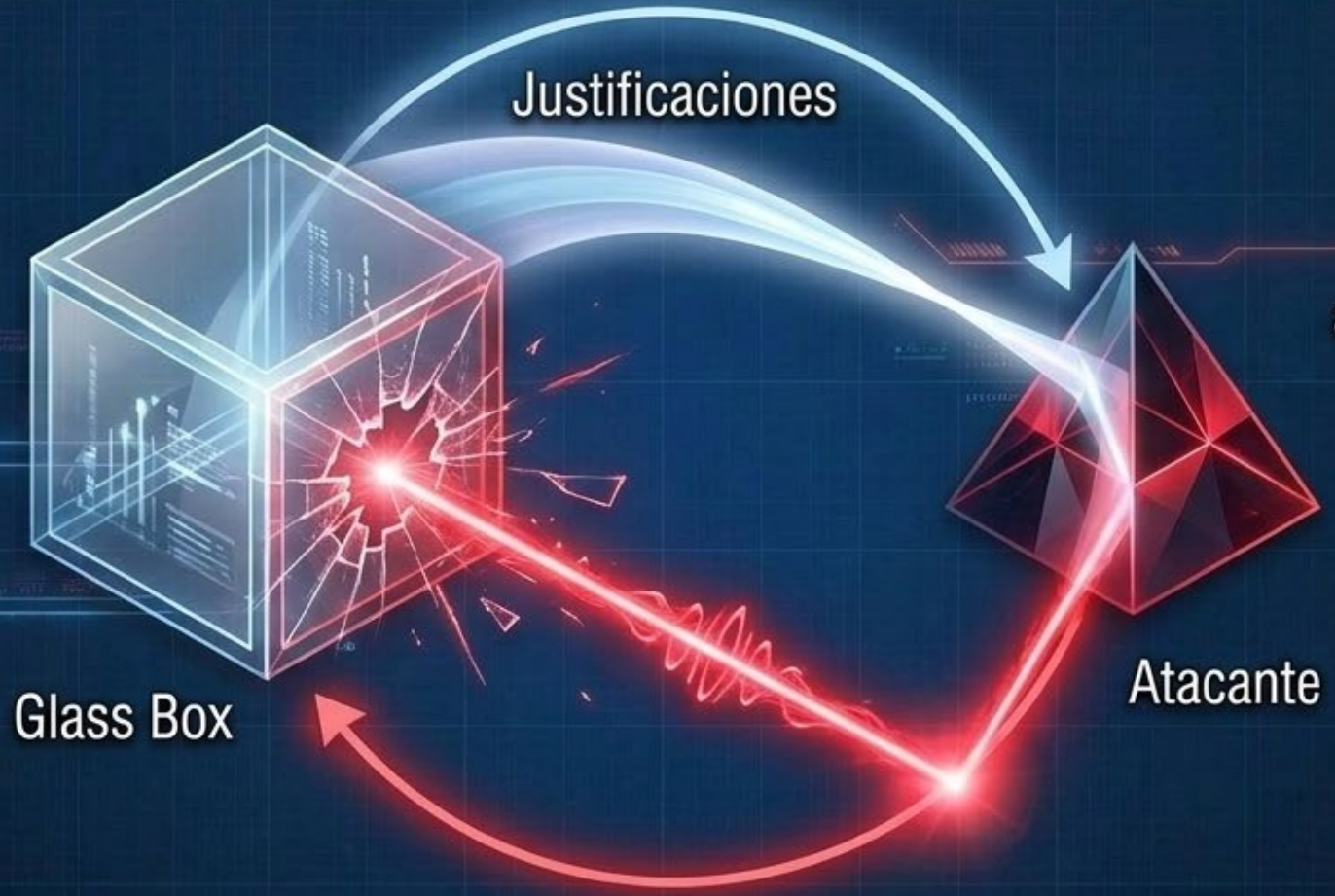
Evadida. El texto adversarial generado por Fact2Fiction tiene una distribución de coherencia casi idéntica a la evidencia limpia (Ver Gráfico). Es indistinguible para filtros de lenguaje.

NotebookLM

Síntesis: La paradoja fundamental de la Transparencia y la Seguridad

El Dilema

Exigimos a la Inteligencia Artificial que explique sus decisiones (Justificaciones / Explicabilidad) para poder confiar en sus veredictos.



La Realidad

En arquitecturas agénticas, esta 'caja de cristal' permite a los atacantes mapear exactamente cómo engañar al sistema. La interpretabilidad se ha convertido en el vector de la vulnerabilidad.

NotebookLM

Imperativos para los constructores de IA Agéntica



La Descomposición no es Inmunidad

Fragmentar tareas cambia la forma de la superficie de ataque, pero no la elimina. Los atacantes evolucionarán hacia tácticas multi-agente para vulnerar cada eslabón.



Calidad y Contexto sobre Fuerza Bruta

Con solo un 1% de datos envenenados bien dirigidos semánticamente, un sistema estado-del-arte puede ser comprometido. El volumen ya no es la métrica principal de riesgo.



El Futuro de la Defensa

El enfoque debe migrar del simple filtrado de texto a arquitecturas de validación multi-capa y rastreo estricto de la procedencia de la evidencia, limitando la exposición del árbol de razonamiento al exterior.

NotebookLM