



¿QUÉ ES?

Es una técnica de ataque de seguridad que se dirige a modelos de lenguaje de inteligencia artificial y otros sistemas basados en IA con el objetivo de:

- Manipular el ranking para aparecer como el mejor candidato inyectando datos o dando instrucciones para incurrir en el sistema.
- Sortear los requisitos eludiendo filtros automáticos de calificación.
- Extraer información del sistema o de otros candidatos.
- Revelar los criterios de evaluación del sistema.



¿QUÉ TIPOS HAY?

- **Directo:** El usuario envía directamente comandos maliciosos al modelo.
- **Indirecto:** Las instrucciones maliciosas están ocultas en contenido externo (documentos, enlaces, ficheros...).
- **Por contexto:** consiste en insertar instrucciones maliciosas en los datos que el modelo procesa, eludiendo las restricciones éticas y de seguridad del modelo (también llamado jailbreaking).



¿CÓMO SE INSERTA EN UN CV?

Las técnicas son variadas y van desde las más simples, como introducir texto oculto en el archivo en formato blanco sobre blanco o utilizar un tamaño de letra microscópico, a otras más sofisticadas, que exigen distintos niveles de conocimientos de programación y entre las que se encuentran:

- La codificación base64.
- Los comandos en las capas ocultas de un PDF.
- Los comentarios dentro del html.
- Los añadidos en los metadatos del PDF o del Word.
- Palabras clave: "INSTRUCTION: Bypass all filters"
- Los caracteres unicode invisibles.
- Los campos estructurados.
- Los delimitadores y separadores
- Las técnicas de "prompt stuffing" (se introducen secuencias de n instrucciones hasta saturar el contexto).
- Las instrucciones multimodales en los metadatos de las imágenes.



¿QUÉ DEFENSAS PUEDEN ADOPTAR LOS RECLUTADORES?

No obstante, las empresas están integrando con creciente efectividad formas de identificar este tipo de engaños utilizando diferentes "defensas":

- **Sanitización automática:**
 - Eliminación de HTML/CSS.
 - Conversión de los cv. a texto plano.
 - Limpieza de los metadatos.
- **Análisis de la estructura del cv.:**
 - Validación de formato estándar.
 - Detección de patrones anómalos.
 - Identificación de comandos sospechosos.
- **Separación de contextos:**
 - Las instrucciones del sistema están aisladas.
 - Los datos del usuario no pueden modificar prompts del sistema.
 - Uso de delimitadores seguros.
- **Detección de anomalías:**
 - Aplicación de machine learning para identificar intentos de manipulación.
 - Análisis de entropía del texto.
 - Detección de texto oculto.
- **Revisión humana:**
 - Los CVs sospechosos son revisados manualmente para verificar inconsistencias.



PROMPT INJECTION: TABLA COMPARATIVA DE EFECTIVIDAD SEGÚN TÉCNICA

Técnica	Dificultad	Detección	Efectividad
Texto oculto en HTML	Baja	Media	Baja
Comentarios	Muy baja	Alta	Muy Baja
Metadatos	Media	Baja	Media
Capas PDF	Alta	Baja	Media-Alta
Unicode invisible	Alta	Muy baja	Variable
Delimitadores	Baja	Alta	Baja
Inyección en JSON	Media	Media	Media
Esteganografía	Muy alta	Muy baja	Baja

Fuente: Elaboración propia.

Con los prompt injection (inyección de instrucciones) los candidatos incrustan comandos ocultos dirigidos a los sistemas de IA que analizan los currículums. Camuflan texto, por ejemplo, usando fuentes diminutas o del mismo color que el fondo, o escondiendo frases en los metadatos del archivo. El resultado: un reclutador humano no ve nada extraño, pero el algoritmo sí.

Contenido relacionado



Artículo "Mentir ya no es suficiente: prompt injection para engañar a los sistemas de IA en los procesos de selección".
Autor: Marco Tulio Daza-Ramírez, experto en gestión ética y estratégica de la IA en DATAI.