🛡️ Hallazgo: Acceso no autenticado y endpoints inefectivos en evaluator.py

Severidad

Moderada – Aunque no es posible modificar la personalidad o el comportamiento del agente IA debido a que el prompt estático se carga desde un archivo interno, el endpoint carece de autenticación y permite generar PDFs y acceder a información de otros usuarios sin restricción. Esto abre la puerta a corrupción de PDFs, posibles cargas maliciosas en documentos y exposición de información sensible. El impacto real es limitado, pero el riesgo conceptual y la violación de con dencialidad lo elevan a severidad moderada.

1. Descripción del fallo

El archivo `evaluator.py y instruction.py` expone varios endpoints que permiten:

- Evaluar contenido mediante un agente IA.
- Cambiar la personalidad del agente.
- Restaurar la personalidad a valores por defecto.

Sin embargo:

✅ La personalidad REAL del agente **no se obtiene desde la base de datos**

La IA usa un *prompt estático* embebido en un archivo `.py` del proyecto.

× Pero los endpoints que permiten modificar la personalidad **no requieren autenticación**

Cualquier usuario externo puede llamar:

- `/instruction GET`
- `/instruction POST cambia instruciones de MIA
- `/evaluate`

…y la API lo aceptará, aunque **estos cambios NO afectan al funcionamiento real de la IA**, porque el agente carga su con guración desde un archivo local del backend.

El resultado es:

→ **Los endpoints quedan expuestos, pero no tienen efecto real.**

---

2. Impacto

Impacto real bajo, pero con riesgo conceptual:

- Un atacante puede ejecutar comandos sobre endpoints sensibles **sin autenticarse**.

- Puede generar confusión al administrador y registros falsos.
- Puede provocar un **DoS lógico** (sobrecarga de peticiones innecesarias).
- Puede inducir a pensar que la IA cambió su personalidad cuando no es así.
- Se revela un diseño inseguro: endpoints críticos sin autenticación.

No existe riesgo de Prompt Injection persistente, porque:

- La IA **no usa la base de datos** para cargar su personalidad.
- El prompt fijo en el archivo `.py` no puede ser modificado por la API.

Por tanto, **no se puede alterar realmente el comportamiento del agente**.

3. Evaluación de grupos sin autenticación

El endpoint `/evaluate` **no requiere autenticación ni autorización**, lo que significa que **cualquier usuario externo puede invocar el servicio** para:

- Solicitar evaluaciones de cualquier grupo (`group_id`).
- Generar PDFs de evaluación de otros usuarios.
- Acceder a información parcial de grupos, aunque no sean los propietarios.

Riesgos asociados:

- **Exposición de información sensible**: nombres de estudiantes, cédulas, programas y evaluaciones.
- **Corrupción o manipulación de PDFs**: un atacante podría inyectar contenido malicioso o información falsa en los PDFs generados.
- **Uso indebido del servicio**: un actor externo podría automatizar peticiones masivas, generando carga en el sistema y confundiendo registros.
- **Fuga de información indirecta**: aunque la IA no cambia su comportamiento, los datos de entrada (conversaciones) podrían contener información sensible y ser devueltos en el PDF.

> Nota: Todas las recomendaciones deben aplicarse primero en un entorno de prueba antes de pasar a producción para evitar corrupción de datos.