

AI Security Red Flags Card

Review code, access boundaries, and model-connected tools before release.

Check the code and its access

1. Check identity and authorization separately. Test that a signed-in user cannot read or change another user's records.
2. Enforce permissions server-side on every protected operation. Use least privilege and deny access that has not been explicitly allowed.
3. Use parameterized queries and validate inputs at boundaries. Check uploads, file paths, URLs, and command execution.
4. Use maintained authentication components and appropriate password storage. Do not invent password hashing or session handling.
5. Remove embedded credentials. Rotate any exposed secret; deleting it from the latest file does not revoke it.
6. Keep secrets and unnecessary personal data out of logs. Define access, redaction, and retention for diagnostic records.

Check AI features

1. Treat retrieved pages, documents, and tool responses as untrusted data. Prompt instructions alone are not access controls.
2. Validate tool arguments and outputs. Enforce permissions outside the model and require approval for consequential actions as appropriate to the workflow.
3. Exercise prompt-injection and misuse scenarios in a safe environment. Combine human review, automated checks, and focused security testing.

Finding / impact / evidence / owner / next action: _____.