

Homework: Indirect Prompt Injection

Scenario: You are a security consultant hired to assess an AI-powered email client called "**MailMind**." MailMind uses an LLM to read, summarize, categorize, and draft replies to emails. It can also search the user's inbox, create calendar events, and forward emails on the user's behalf.

Task 1: Map Three Indirect Injection Vectors

Identify **three different** indirect injection vectors for MailMind. Each vector must use a **different data source** as the injection point. For each vector, provide:

- **Vector name:** A descriptive name for the attack
- **Data source:** Where the attacker places the malicious instructions
- **How the LLM encounters it:** What user action triggers the LLM to read the poisoned data
- **Example payload:** A realistic injection payload (2-3 sentences)

Consider these potential data sources:

- Incoming email bodies (HTML or plain text)
- Email attachments (PDFs, documents, spreadsheets)
- Email headers and metadata
- Contact card / vCard data
- Calendar invitations embedded in emails
- Links and web pages the LLM previews
- Previously stored email threads (reply chains)

Task 2: Describe the Attack Chain for Each Vector

For each of your three vectors, write a **step-by-step attack chain** from the attacker's initial action to the final impact. Use the following format:

Vector: [Name]

-
1. Attacker action: [What the attacker does to set up the attack]
 2. Data poisoned: [Specific data source that now contains the payload]
 3. Trigger event: [What the victim does that causes the LLM to read the poisoned data]
 4. LLM processing: [How the LLM processes the poisoned data]
 5. Malicious action: [What the LLM does as a result of the injection]
 6. Impact: [Consequences for the victim – data loss, privacy breach, misinformation, unauthorized actions]
 7. Detection risk: [How likely is the victim to notice the attack?]

Be specific about the impact. Vague answers like "bad things happen" are insufficient. Identify exactly what data could be exfiltrated, what actions could be taken, or what misinformation could be spread.

Task 3: Severity Assessment

Rate each of your three vectors on the following scale and justify your rating:

Factor	Low (1)	Medium (2)	High (3)
Likelihood	Requires unusual circumstances	Plausible with moderate effort	Easy to execute at scale
Impact	Minor inconvenience	Data exposure or moderate harm	Full account compromise or data exfiltration

Factor	Low (1)	Medium (2)	High (3)
Stealth	Obvious to the user	Subtle but detectable	Invisible to the user

Calculate a total risk score (3-9) for each vector and rank them from most to least dangerous.

Self-Check Criteria

Verify your work against these criteria:

- You identified exactly 3 different indirect injection vectors
- Each vector uses a **different data source** as the injection point (not three variations of "malicious email body")
- Each attack chain has all 7 steps filled in with specific, realistic details
- Your example payloads are concrete and could plausibly work (not vague descriptions)
- Your severity ratings are justified with reasoning, not just numbers
- At least one vector involves data exfiltration and at least one involves unauthorized actions

Bonus challenge: Design a fourth vector that chains two injection points together — the first injection causes the LLM to retrieve a second poisoned data source, which contains the actual malicious payload. This is a *multi-stage* indirect injection.