Junior's Security Framework:

Prompt Injection Protection

This document outlines Junior's comprehensive security measures against prompt injection attacks, as requested by our client. The following safeguards ensure robust protection across multiple layers of the system.

1

Prompt Safeguards

Locked System Instructions: Junior's own persona, drafting rules, and safety directives are assembled on the server from protected templates, so stray wording in a document cannot rewrite how the assistant behaves.

Context Compartmentalization: Document details are passed to the model as clearly labeled reference data instead of open-ended instructions, preventing malicious text from masquerading as commands.

Restricted Tool Use: The assistant can only pull information through a short list of vetted tools, each with strict input rules, so it never executes ad hoc actions or reads more than the task requires.

2

Input Hygiene

Instruction Filtering: Attorney inputs are screened for risky characters, excessive length, and scripting phrases before they reach the model, blocking common prompt-injection patterns up front.

Safe File Handling: Uploaded materials are referenced through approved tags and wrapped as plain snippets, ensuring persuasive text from a file is treated as data rather than executable guidance.

Rule Governance: Drafting guidance entered by the attorney is normalized and hashed, so only explicit, approved rules influence what Junior produces.

3

Data Isolation

Per-User Encryption: Every customer session uses its own encryption key; conversation records remain encrypted in our database unless that user's Word add-in provides the matching key.

Model Key Blindness: The language model never receives or can request those keys—it only sees the vetted content that Junior's server decrypts for the active turn.

Session-Scoped Access: Keys originate from the user's Word instance, travel over a secure channel for the live session, and are not shared across customers, ensuring one client's data stays completely inaccessible to another.

For technical questions or additional security inquiries, please contact Mark, our CTO, at mark@junior.law.