OPTIMIZING NODE LATENCY: RESPONSE TIMES FOR RETENTION

In digital sales, every 100ms of delay costs money. If your AI takes 5 seconds to think, the user has already closed the tab.

PARALLEL PROCESSING

We optimize latency by using parallel API calls. While the LLM is generating a greeting, the system is simultaneously pulling CRM data in the background. This creates a "Zero-Lag" feel.

STREAMING TEXT

We implement token-streaming where possible, allowing the user to begin reading the response as it is being generated. This psychological trick reduces "perceived latency" and keeps the user engaged with the node.