Performance Considerations

Speed vs quality

- More layers means more latency. If you need fast responses, waiting on three consecutive prompts to complete is probably a bad solution
- But parallelization can help. Some layers don't depend on each other and can run simultaneously. If your reasoning layer and your memory retrieval layer both only need the user input, run them in parallel
- Performance can be helped and hurt by layering. Adding a layer isn't always the answer. Sometimes consolidating two weak layers into one strong prompt improves both speed and quality

Hallucination Considerations

Avoiding architectures that tend to compound hallucinations

- In multilayered systems, hallucinations compound. One layer's mistake becomes the next layer's input. If your reasoning layer hallucinates a fact and your content layer writes it beautifully, you've just produced confidently wrong output. The more layers, the more opportunities for errors to slip through and get amplified
- Correction layers should come before memory consolidation. If you don't catch errors before they enter your system's permanent memory, those minor mistakes slip into history and slowly expand. They reintroduce themselves ad infinitum, compounding with each cycle until your system's "source of truth" is corrupted