Latency Budgets for AI Assistants in User-Facing Products

Whitepaper - portfolio demo document.

A demo-safe paper that frames response time as a product decision, not only a systems metric.

Demo-safe	Whitepaper	Replaceable asset

ABSTRACT

Abstract

Users experience assistants as a conversation, so latency shapes trust, iteration speed, and perceived quality. A practical latency budget should account for routing, retrieval, model time, tool calls, and presentation.

CORE FRAMEWORK

Core framework

A useful budget separates the stack into stages so teams can measure regressions and make tradeoffs explicitly.

- Perceived latency: how quickly the interface responds to user intent.
- Model latency: time spent generating the answer.
- Tooling latency: retrieval, APIs, and orchestration.
- Recovery budget: what happens when a step fails or times out.

EXAMPLE BUDGET

Example budget

Stage	Target	Owner
UI acknowledgment	< 150 ms	Frontend
Routing + retrieval	< 450 ms	Backend
Initial answer token	< 1.2 s	Model / orchestration
Tool fallback / retry	< 900 ms	Workflow

DESIGN IMPLICATIONS

Design implications

Latency budgets are also interface decisions.

- Show progress states early.
- Decompose slow workflows into visible stages.
- Use partial streaming when quality remains acceptable.
- Log per-stage timings so hiring teams can discuss tradeoffs concretely.