

同じ長文を毎回読み直している

- システム指示
- ツール定義
- 会話履歴
- 大きな文書

Prompt Cachingの仕組み

- 共通の接頭辞を再利用
- 新しい末尾だけを追加計算
- モデル出力はキャッシュしない

効かせる入力設計

- 固定情報を先頭へ
- 可変情報を末尾へ
- 順序と内容を安定させる

注意点

- ヒットは保証されない
- TTLと最小長
- データ保持要件を確認

実務で測る

- cached tokensを監視
- レイテンシと費用を比較
- RAG・要約と組み合わせる