Precedent Finder Pipeline - Technical Summary (One Page)
Source baseline: /app/api/search/route.ts + /lib/pipeline/*
Generated: 2026-02-14

1) API Entry, Guardrails, and Caching
- POST /api/search is the single search entrypoint; query length floor is ~12 chars.
- Non-debug requests apply IP rate limiting (default 40 req / 60s window).
- Response cache key includes normalized query + policy seed; default TTL is 300s.
- In-flight lock/coalescing serves concurrent duplicates from shared cache when possible.
- Routing is server-only in this path; status can be completed/partial/blocked/no_match.

2) Intent + Planning Layer
- Intent extraction builds domains, issues, statutes, procedures, actors, court hint, date window.
- Bedrock reasoner pass-1 (bounded, fail-open) outputs proposition sketch + query variants.
- Deterministic planner always runs and grounds/backs up reasoner output.
- Canonical rewrite V2 can synthesize prioritized strict+broad variants with token constraints.
- Retrieval provider is selected by RETRIEVAL_PROVIDER (auto/indiankanoon_html/serper).

3) Retrieval Scheduler Mechanics
- Phase order: primary -> fallback -> rescue -> micro -> revolving -> browse.
- Default phase limits: 2/2/1/1/1/1, global budget: 8 attempts, max elapsed: 9000 ms.
- Adaptive scheduler reorders variants using utility, case-like yield, challenge, timeout rates.
- Per-attempt timeout is dynamically bounded (cap ~3500 ms) by remaining request budget.
- Query signatures prevent duplicate fetches; candidate provenance tracks strict/relaxed hit origin.
- Stop reasons include enough_candidates, budget_exhausted, blocked, completed.

4) Verification, Scoring, and Proposition Gating
- Candidates are classified as case/statute/noise/unknown before ranking.
- Verifier hydrates shortlisted docs (concurrency default 4) with retry/fallback handling.
- Score combines lexical overlap, context matches, proposition coverage, court/citation signals.
- Penalties apply for polarity mismatch, contradictions, missing interactions, low-quality provenance.
- Proposition gate enforces hook groups, relations, outcome polarity, role-chain constraints.
- Results split into exact_strict, exact_provisional (confidence-capped), and near_miss tiers.

5) Resilience, Pass-2, and Always-Return
- Conditional reasoner pass-2 runs once when exact quality/coverage is insufficient and budget remains.
- Timeout recovery can expand deterministic budget (extended_deterministic mode).
- Always-return guarantee may trigger extra backfill attempts when tier counts are sparse.
- Stale-similarity cache fallback and synthetic advisory fallback avoid empty final payloads.
- pipelineTrace returns planner/scheduler/retrieval/classification/verification diagnostics.

Technical intent: bounded latency, transparent ranking rationale, fail-open behavior, and explicit degradation paths under source blocking, rate limits, or LLM unavailability.