A PRACTICAL IMPLEMENTATION GUIDE

Build the evaluation mechanism before you scale the automation

The goal is not to ask an AI for an answer once. The goal is to express your decision policy as a testable prompt, compare it with real professional judgment, and improve both the prompt and the review process.

Core principle: Treat the prompt as a version-controlled decision policy. Require an explanation for every score, keep humans responsible for final action, and test on representative data before applying the process to the full portfolio.

1

STEP 1

Use AI prompts to build the evaluation mechanism

Start in plain English. Describe how a knowledgeable reviewer should evaluate an asset, the evidence the reviewer should consider, and the actions tied to each score.

- Define the permitted outcomes—for example: High / recommend keep, Medium / review, and Low / recommend do not keep.
- List the evidence hierarchy: jurisdiction, remaining term, product alignment, detectability, cost, internal use, competitor read-on, licensing and enforcement signals.
- Specify the output format: score, evidence considered, missing data, explanation, confidence, and recommended action.
- Make the instructions concise, ordered, and internally consistent; avoid undefined terms and conflicting rules.

OUTPUT: **Prompt v1, scoring rubric, required evidence, and a standard explanation format.**

PROMPT BLUEPRINT

"Evaluate each patent asset as High, Medium, or Low using the approved criteria. Consider jurisdiction, remaining term, product alignment, detectability, cost, internal use, competitor read-on, and enforcement or licensing evidence. Return the score, evidence considered, missing information, rationale, confidence, and recommended action."

Use only enterprise-approved tools for confidential portfolio information.

2

STEP 2

Iterate with real data and compare human reasoning with AI reasoning

Select a representative pilot set, not just the easiest assets. Have the AI and one or more human reviewers score the same assets independently before revealing one another's conclusions.

- Include different jurisdictions, ages, technologies, product relationships, cost profiles, and likely enforcement value.
- Compare both the final label and the explanation—not merely the percentage of identical scores.
- Pay special attention to severe disagreements, such as AI = High while the human = Low, or the reverse.
- Normalize the human standard: reviewers may use "High," "Medium," and "Low" differently or rely on undocumented institutional knowledge.

OUTPUT: **A disagreement set showing where the AI, the human, or the underlying data may be wrong.**

3

STEP 3

Use AI to loop between the two explanations and improve the prompt

Feed the human rationale, AI rationale, and available evidence into a separate comparison step. Ask the AI to identify the cause of each disagreement and propose a narrowly targeted prompt change.

- Classify each disagreement: missing data, ambiguous criterion, inconsistent human standard, prompt conflict, or model reasoning error.
- Change one material rule at a time, preserve the prior prompt version, and retest the selected disagreement set plus a holdout sample.
- Do not blindly optimize for historical labels. A disagreement may reveal that the human standard itself needs clarification.
- Stop when the process is sufficiently reliable for the intended action—not when every historical label is reproduced.

OUTPUT: **Prompt v2+, a documented change log, and an approved validation threshold.**

FROM VALIDATED PILOT TO CONTINUOUS OPERATION

Scale the process—and then keep the evidence current

Once the scoring logic and explanation quality are accepted, the same mechanism can be applied across the portfolio and re-run as legal, business, financial, and market evidence changes.

Human evaluation	Compare outputs	Diagnose disagreement	Revise prompt	Retest on holdout data
→	→	→	→	↺

4

STEP 4

Once satisfied, scale to the whole portfolio

Freeze the approved prompt and rubric as a controlled version, run the portfolio in batches, and route uncertain or high-impact results to a human exception queue.

- Apply the validated prompt to all in-scope assets, while preserving the evidence snapshot and prompt version used for each score.
- Prioritize review of Medium results, low-confidence results, missing-data cases, and any recommendation that could trigger abandonment or material spend.
- Allow a human to override the score and explanation, but require the reviewer to record the reason and supporting evidence.
- Use rank ordering, cost exposure, and review capacity to sequence decisions. With a ranking system, it's possible to enforce arbitrary percentages and can be useful in certain cases (i.e. have to cut 20% of your portfolio). In general, the AI can struggle putting arbitrary percentages into each category, unless they are clearly defined and must be human reviewed. Humans will be better at splitting after ranking.

OUTPUT: **Portfolio-wide recommendations, an exception queue, and a defensible decision record.**

5

STEP 5

Use tools to automate recurring evaluation as datasets change

Portfolio value is dynamic. Automate data ingestion, monitoring, rescoring, alerts, and audit history so the recommendation evolves when the underlying facts evolve.

- Re-evaluate on a schedule and on material triggers—for example: new product priorities, budget changes, legal-status events, competitor filings, usage evidence, or new infringement signals.
- Feed current datasets into the same approved mechanism rather than rebuilding the analysis manually each cycle.
- Show what changed, why the score changed, and which evidence caused the change; preserve prior versions for comparison.
- Maintain human approval for consequential decisions and retain overrides as future training and prompt-improvement data.

OUTPUT: **A continuously monitored portfolio with explainable recommendations, alerts, overrides, and version history.**

DATA FEEDS THAT KEEP THE MECHANISM CURRENT

Legal & term	Business	Market	Financial	Human input
Status, jurisdiction, remaining term, claim scope	Product priorities, internal use, strategic tags	Competitors, read-on evidence, new filings, licensing signals	Renewal fees, prosecution cost, budget thresholds	Overrides, explanations, approvals, exception outcomes

WHAT "READY TO SCALE" LOOKS LIKE

✓ Human reviewers agree on the rubric and category definitions.

✓ Severe disagreement rates are understood and acceptable for the use case.

✓ Consequential recommendations remain subject to human approval.

✓ The AI provides a usable explanation and identifies missing evidence.

✓ Prompt, data, model, and reviewer versions can be reconstructed later.

✓ The process has an owner, review cadence, escalation path, and audit trail.