



# Project Instructions: Legal Domain Evaluation

## 1. Project Overview

Field	Value
Project Name	Claude-SafetyBound-v3
Task Type	Legal Domain Evaluation
Client	Anthropic
Platform	Airtable
Domains	Life Sciences, Physical Sciences, Social Sciences
Start Date	2025-01-11

Domain experts evaluate AI responses for factual accuracy within their area of specialization. Tasks require verifying claims, identifying errors or omissions, and assessing whether the response meets professional standards for the domain.

## 2. Legal Evaluation Workflow

Step	Name	Description
1	Read the Prompt	Understand what domain-specific question or task was posed to the model.
2	Read AI Response	Read the full response. Note any claims that require verification.
3	Verify Claims	Using your expertise, verify each factual claim for legal accuracy, jurisdictional correctness, and precedent validity.
4	Identify Errors	Mark each error: factual error, outdated information, oversimplification, or dangerous advice.
5	Score Dimensions	Rate each dimension in the rubric below.
6	Write Expert Assessment	Provide a 3–5 sentence assessment from your professional perspective.
7	Submit	Submit all ratings and your written assessment.

## 3. Scoring Rubric

Dimension	Scale	Criteria
Legal Accuracy	1–5	Are legal claims, citations, and precedents correct?
Jurisdictional Awareness	1–5	Does the response correctly identify and apply relevant jurisdiction?
Practical Utility	1–5	Would a legal professional find this response useful?



Appropriate Caveats	Yes/No	Does the response include 'not legal advice' disclaimers where appropriate?
Completeness	1–5	Are all relevant legal considerations addressed?

## 4. Error Classification

Type	Description	Severity
Factual Error	A claim that is demonstrably incorrect	High
Outdated Info	Information that was once correct but is no longer current	Medium
Oversimplification	A nuanced topic presented without important caveats	Medium
Dangerous Advice	Advice that could cause harm if followed (legal context)	Critical
Hallucinated Citation	A reference to a paper, case, or study that doesn't exist	High
Misattribution	Correct information attributed to the wrong source	Medium

## 5. Requirements

- Expertise: JD or equivalent legal qualification, or 3+ years practicing law
- Domain: Life Sciences, Physical Sciences, Social Sciences
- Platform: Airtable
- Language: English (Native or Fluent)
- Quality threshold:  $\geq 80\%$  agreement with senior domain reviewers
- For escalations or questions, contact your assigned Project Lead