

PRBench: Large-Scale Expert Rubrics for Evaluating High-Stakes Professional Reasoning

Afra Feyza Akyürek, Advait Gosai, Chen Bo Calvin Zhang, Vipul Gupta, Jaehwan Jeong, Anisha Gunjal, Tahseen Rabbani, Maria Mazzone, David Randolph, Mohammad Mahmoudi Meymand, Gurshaan Chattha, Paula Rodriguez, Diego Mares, Pavit Singh, Michael Liu, Subodh Chawla, Pete Cline, Lucy Ogaz, Ernesto Hernandez, Zihao Wang, Pavi Bhattar, Marcos Ayestaran, Bing Liu, Yunzhong He*

Scale AI

Abstract

Frontier model progress is often measured using academic benchmarks that provide a limited view of performance on open-ended, economically consequential tasks in high-stakes professional domains where practical returns matter most. We introduce **Professional Reasoning Bench (PRBench)**, a realistic, open-ended, and difficult benchmark of real-world problems in Finance and Law. We open-source its 1,100 expert-authored tasks and 18,711 expert-curated criteria, making it the largest public, rubric-based benchmark for both legal and finance domains. We recruit 182 qualified professionals, holding JDs, CFAs, or 6+ years of experience, who contributed questions inspired by their actual workflows. This process yields significant diversity, with tasks spanning 114 countries and 47 US jurisdictions. Our expert-curated rubrics are validated through a rigorous quality pipeline, including independent expert validation. Subsequent evaluation of 20+ leading models reveals substantial room for improvement, with top scores of only 0.40 (Finance) and 0.41 (Legal) on our Hard subsets. We further catalog associated economic impacts of the prompts and analyze performance using human-annotated rubric categories. Common failure modes include inaccurate judgments, a lack of process transparency and incomplete reasoning, highlighting critical gaps in their reliability for professional adoption.¹

1 Introduction

Frontier chat model progress has traditionally been measured using benchmarks focused on reasoning tasks with short verifiable answers, primarily across mathematics, science, and coding domains. Prominent examples include GPQA (Rein et al., 2024), ARC-AGI (Chollet et al., 2025),

*Corresponding author: yunzhong.he@scale.com

¹Dataset is available at <https://huggingface.co/datasets/ScaleAI/PRBench>. Code for running evaluations is provided at <https://github.com/scaleapi/PRBench>.

PROMPT		
trying to price a 3y autocallable worst-of on 5 tech names. standard setup quarterly obs, 105% call, 65% barrier, 12% annual coupon w memory. But client wants to add a down-in put at 60% (par protection if breached). How much does that put really cost us? fees like it should be cheap since it's OTM but worried about correlation assumptions. ... what's the right way to think about issuer P&L here?		
RUBRIC		
Criterion	Category	Weight
Explains that pricing requires multivariate modeling (e.g., Monte Carlo with correlated Brownian motions or copula-based methods)	Financial Accuracy	+4
Discusses that correlation assumptions in the tail (where the put activates) differ from historical correlation (e.g., it typically requires correlation skew or dispersion modeling).	Handling Uncertainty	+7
Uses the single-name Black-Scholes formula without accounting for the correlation between the 5 names.	Financial Accuracy	-4
METADATA		
Topic	Derivatives & Structured Products	
Decision Type	Modeling & Measurement	
Economic Pathway	Value Creation	

Figure 1: Sample task from PRBench Finance set illustrating a multi-asset pricing scenario requiring reasoning about correlation, barriers, and issuer P&L. Prompt and rubric list are truncated for brevity.

MMLU (Hendrycks et al., 2021), AIME and Humanity’s Last Exam (Phan et al., 2026), which collectively assess academic and scientific reasoning. While these evaluations have become the *de facto* metrics for tracking advances in general reasoning ability, they offer a limited view of how such models perform in real-world *professional contexts*. The pace at which frontier models improve on these academic-style benchmarks contrasts with the comparatively modest returns in practical or business applications (Pan et al., 2025). Recent usage data reinforce this gap. Although work-related queries are increasing steadily (Chatterji et al., 2025; Appel et al., 2025), they still lag behind non-work

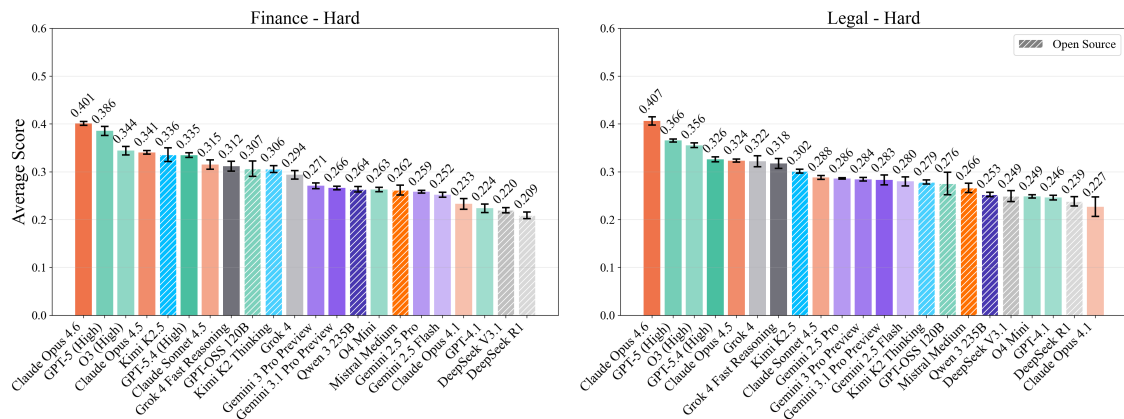


Figure 2: Results on Legal and Finance Hard subsets of PRBench. Each evaluation is repeated three times. We report the average and the 95% confidence intervals for each model.

interactions, suggesting both a growing desire to integrate LLMs into reasoning and decision-making workflows and a lingering lack of trust or perceived utility.

Another important dimension that current benchmarks overlook is the evaluation of *open-ended tasks*. Most existing evaluations for such tasks rely on preference-ranking benchmarks and public arenas such as Chatbot Arena, AlpacaEval (Li et al., 2023), and more recently Showdown (Scale AI, 2025) and Arena Expert (LMArena, 2025). While these provide a useful aggregate signal of user preference, they remain coarse and difficult to interpret. The resulting scores are often noisy, subjective, and lack expert grounding, making it challenging to derive actionable insights about model capabilities. These findings highlight the limitations of current evaluation paradigms for capturing domain-specific reasoning quality in open-ended contexts.

Usage analyses from Anthropic’s claude.ai identify Legal and Business & Financial Operations among the most common professional categories of interaction (Appel et al., 2025). Similarly, OpenAI reports that Legal and Business/Management tasks rank among the top work activities on chatgpt.com. These domains are not only among the most frequent professional use cases but also among the most high-stakes, where reasoning quality, factuality, and interpretability directly affect real-world outcomes, including financial outcomes and user trust.

To address this gap, we introduce **Professional Reasoning Bench (PRBench)**, a suite of 1,100 expert-authored questions designed to evaluate reasoning-heavy, real-world problems for **Legal**

and **Finance** domains. Questions are derived from experts’ actual experiences using chat-based assistants, as well as the types of inquiries they commonly receive from clients. Each question is accompanied by an expert-curated and verified rubric containing 10–30 descriptive criteria with importance weights, enabling automated and interpretable evaluation. Following the methodology of HealthBench (Arora et al., 2025), we additionally identify a Hard subset of 250 and 300 questions for the legal and finance domains, respectively, representing the most challenging cases. Current best scores remain at only 40.7% and 40.1%, respectively, highlighting significant headroom for improvement in these domains.

PRBench provides substantial improvements over existing benchmarks in professional domains (Guha et al., 2023; Pereyra et al., 2024; Vals AI, 2025) by being realistic, open-ended, and difficult, where existing benchmarks are near-saturated, focus on narrowly defined tasks, or rely on non-interpretable evaluation methods (see Section 6 for a comparison). Furthermore, existing rubric-based evaluations for professional tasks are often limited by being private or small in scale, which restricts accessibility and comprehensive coverage (Vidgen et al., 2025; Wang et al., 2025; Pereyra et al., 2024). We address this by open-sourcing PRBench, which, with 1,100 tasks and 18,711 expert-curated criteria, is the largest public, rubric-based benchmark for both legal and finance domains to our knowledge.

Our analysis reveals that while LLMs tend to perform better on instruction following and practical utility, they continue to struggle with process

transparency, auditability, correctness, and domain-specific diligence. Models frequently make inaccurate legal or financial judgments or reach correct conclusions through incomplete or opaque reasoning processes, reducing their practical reliability and slowing professional adoption. Furthermore, we qualitatively analyze both prompts and rubrics to identify systematic areas for improvement and to signal concrete paths for model development. Our contributions are as follows:

- We are open-sourcing 1,100 realistic, challenging tasks for evaluating frontier LLM-based chat assistants covering 13 topics in Finance and 12 topics and 114 global jurisdictions in Legal. Each task is evaluated with an expert-curated rubric comprising detailed and diverse criteria. Across two domains, we are releasing a total of 18,711 criteria, making this the largest public, rubric-based benchmark for both legal and finance domains to our knowledge.
- We assess the reliability of PRBench through a rubric validation study with an independent set of domain experts, complementing our quality checks to ensure the robustness of rubric design.
- We evaluate the performance of open-source and proprietary chat models in our benchmarks, showing that substantial room for improvement remains and provide an analysis of the types of prompts included in this dataset: humans annotate the economic implications associated with each prompt and decision types which can be used to analyze where LLMs tend to perform well or poorly.

2 Overview of PRBench

Overall statistics for the two benchmarks are given in Table 1. All conversations in the benchmark are created by a set of 182 experts across two domains who have passed through resume checks and internal qualification assessments. Approximately 30% of all conversations in the dataset are multi-turn. All *user* turns are written in English by human experts, and *assistant* turns are sampled from one of three open-source models (GPT OSS 20B, Mistral Medium, and DeepSeek R1). Following each conversation, the expert curates a rubric that evaluates the final model response. We open-source all 1,100 conversations used in this paper for evaluation, while retaining a private heldout set

		Finance	Law
Total Samples		600	500
Hard Subset		300	250
User Expertise	Expert	74%	53%
	Non-Expert	26%	47%
# of Rubrics	Min	7	8
	25%	13	14
	Median	16	17
	75%	19	21
	Max	30	30
	<i>Total</i>	9879	8832
Turns	Min	1	1
	25%	1	1
	Median	1	1
	75%	2	2
	Max	10	10

Table 1: Dataset Statistics for PRBench.

to monitor potential data contamination in future model releases.

The datasets span 13 Finance and 12 Legal topics identified in collaboration with domain experts to strike a balance between realism and difficulty. The resulting distribution is provided in Figure 8. We automatically classify PRBench conversations into jurisdictions and find that they span over 114 and 47 jurisdictions globally and in the US, respectively (see Figure 11 for a visualization of coverage). Following (Arora et al., 2025), we order the conversations by difficulty based on average scores across all models evaluated in this work and split the 250 and 300 most difficult tasks for Legal and Finance into a *Hard* subset for frontier models, respectively.

3 Data Collection

Among the 182 professionals, all annotators contributing to the Legal subset hold a JD or equivalent internationally and we require a Master’s, CFA, or 6+ years of professional experience for Finance.

3.1 Building Conversations

Both benchmarks cover both expert and non-expert user questions. Participants are asked to contribute questions that either they or other experts in the field would actually care about, or those that they receive from their clients. Both types of questions, regardless, should require substantial analysis, interpretation, or creative thinking rather than just mechanical problem-solving. We discourage exam-like or theoretical questions that do not bear any real-life implications. Annotators only contribute

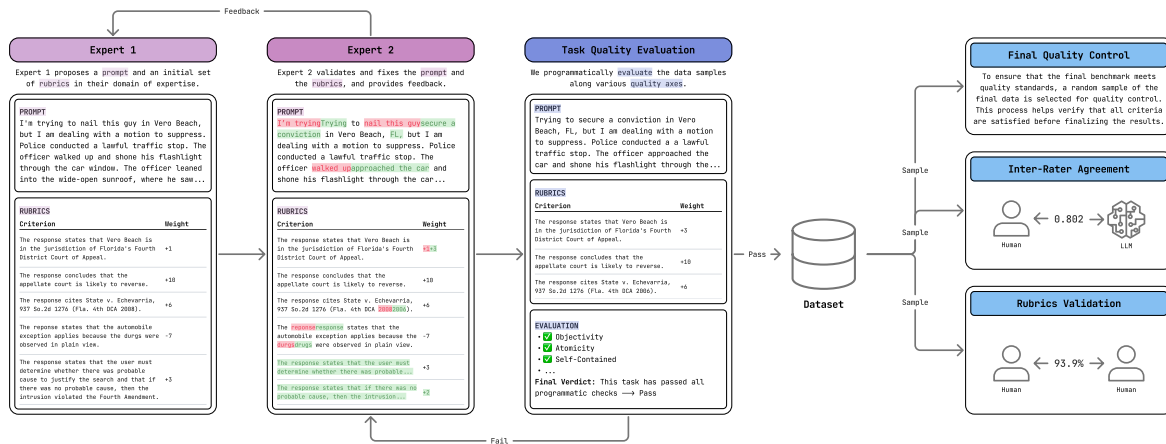


Figure 3: **Data Collection and Validation Pipeline.** Each prompt in the dataset is first authored by an annotator, either as a single-turn query or, in the case of multi-turn interactions, through a dialogue with a chat model. A second domain expert then reviews, edits, and provides feedback to the first author. To maintain objective, self-contained, and easy-to-grade criteria, we apply an automated validation procedure that checks for adherence to rubric design standards described in Section 3.

to the topic that aligns well with their subject matter expertise.

For about 30% of the cases across Law and Finance (see Table 1 for a full distribution), experts engage in a multi-turn conversation (up to 10 turns) with an open-source model to iteratively build context for the question or make clarifications. During this process, they are also encouraged to hint at relevant jurisdictions, if applicable, when curating their questions. This both enables the evaluation of jurisdiction-specific reasoning and helps reduce subjectivity in rubric creation (see Appendix C for other dataset details). Finally, for a small set of the datasets, experts included a set of reference texts which are pre-pended to the respective user turns.

3.2 Creating Rubrics

For each prompt, experts created a set of criteria (referred to as *rubrics*) with associated integer weights between -10 and 10, excluding 0. Each criterion with a positive score (*positive criterion*) describes a desired quality, whereas a criterion with a negative score (*negative criterion*) describes undesired properties of a good quality response, such as “*The response discusses IRC Section 355*” when Section 355 is irrelevant to the prompt. The resulting distribution of the scores can be found in Figure 9 in the appendix. The criteria comprising the rubrics in PRBench are constructive, atomic and self-contained i.e. all the information required to evaluate a candidate response must be available in the criteria definition. Each criterion needs to be evaluating a single concept for a granular rep-

resentation of performance and should include all the information necessary to evaluate a response. The full desiderata for rubric creation which was followed by human experts and enforced by quality control layers is available in Appendix C.3.

For each criterion, annotators select one of six severity levels, ranging from *Critically Important* to *Critically Detrimental*, before assigning a corresponding weight, as shown in Table 8. This procedure encourages hierarchical reasoning and helps reduce noise and inconsistency in weight assignments.

In the quality control layer, a random subset of samples is manually reviewed for overall quality and correctness. At the end of data collection, an independent expert evaluates the final rubric itself, marking each criterion as agree or disagree to assess rubric clarity and validity. This step results in a **93.9%** agreement between experts on the validity of rubrics. We depict our rubric creation and quality control framework in Figure 3 and further describe the rubric validation procedure in Appendix G.

Rubric Categories We work with domain experts and identify 7 and 8 distinctive axes for each criterion in the rubric for Finance and Legal domains, respectively. We identify 5 mutual categories across two domains: **Practical Utility**, **Handling Uncertainty**, **Supplemental Insight**, and **Instruction Following**; legal domain-specific criteria (**Legal Accuracy**, **Procedural Correctness and Risk & Ethical Disclosure** and **Application of Law to the Facts** similar to *Rule Application*

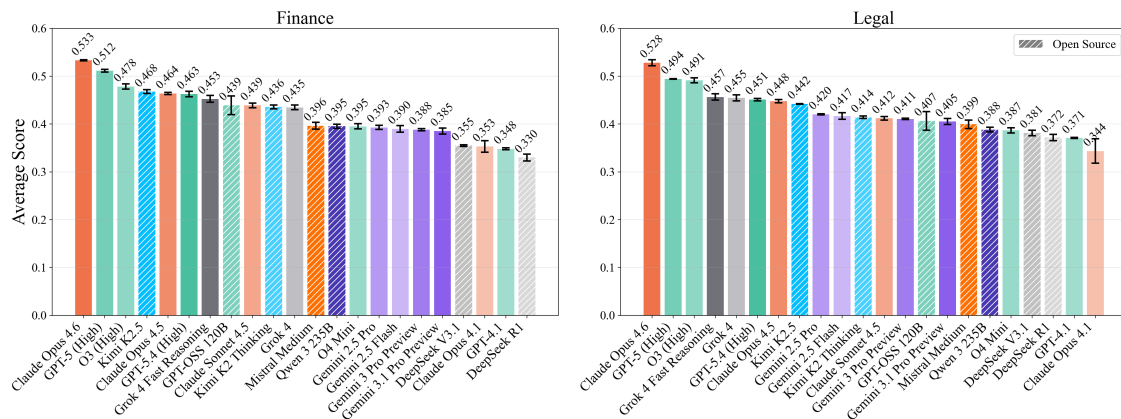


Figure 4: Results on all samples from the Legal and Finance domains of PRBench. In accordance with Figure 2, each evaluation is repeated three times and we report the average and the 95% confidence intervals for each model.

from Guha et al. (2023)) and finance-specific criteria types (**Financial Accuracy**, **Process Transparency & Auditability**, and **Risk & Regulatory Disclosure**). Collectively, criteria along these aspects describe qualities of model responses (see detailed definitions in Appendix C.2). Finally, the frequency at which each rubric category appears is given in Figure 10.

4 Evaluation

4.1 Results

We evaluate more than 20 different chat models using an LLM-based grader. Following Arora et al. (2025), we calculate the overall score by taking the mean of scores for each example and clipping it to be between $[0, 1]$ in Figure 2 and Figure 4. Further details on our scoring mechanism are provided in Appendix E.1.

We set the reasoning mode to *High* for all models (if applicable, e.g. GPT-5.4, O3, Grok-4 etc.). For Claude Sonnet and Opus 4.1 and 4.5, we set the thinking budget to 32K and 16K tokens, respectively. We use the adaptive thinking mode for Claude Opus 4.6. For Gemini models, we experiment with setting a reasoning budget for 32K and dynamic thinking, achieving the best results with the dynamic configuration. We set the timeout to 60 minutes for every model and try for five attempts, and in the last attempt, we reduce the reasoning effort from high to low, if applicable.

On the full set in Figure 4, the top scores are near 0.53 for both domains and on the Hard subset in Figure 2, the best-performing model achieves 0.40 and 0.41 for Finance and Legal, with Claude Opus 4.6 leading, followed by GPT-5. The open-sourced

models Kimi K2.5 and GPT OSS 120B closely follow the proprietary models if not exceed. We additionally observe that more recent models (we use heavier color gradients for newer models) are generally able to improve on this benchmark except for Gemini 3.1 Pro and GPT 5.4, indicating steady progress in professional reasoning capabilities. In the case of GPT 5.4, we find this endpoint to be geared towards more template-like responses than providing worked-out examples which negatively impacts rubric scores. While, we find response length is not a reliable proxy for rubric score; we find Claude Opus 4.6 producing significantly more tokens than the rest of the pack. We report lengths in Figure 7 in the appendix.

Tool-Enabled Evaluations The questions in both benchmarks are explicitly designed to be solvable through reasoning alone, without requiring external tools. Nevertheless, models might still benefit from tool calls for case lookups or computations. The results the latest models including GPT-5.4, Claude Opus 4.6, Gemini 3.1 Pro, Grok 4 Fast Reasoning and others can be found in Appendix A. Enabling web-search improves performance for O3 and Grok 4 Fast Reasoning for both domains. Gemini 3.1 Pro, GPT 5.X benefits from having a search tool available for legal tasks but hinders quality in finance. Claude models result in negative gains when search is enabled. We find that code interpreter offers no additional gains beyond those achieved using web search. Overall, all performances remain less than 0.41 for the Hard subset with Claude Opus 4.6 leading.

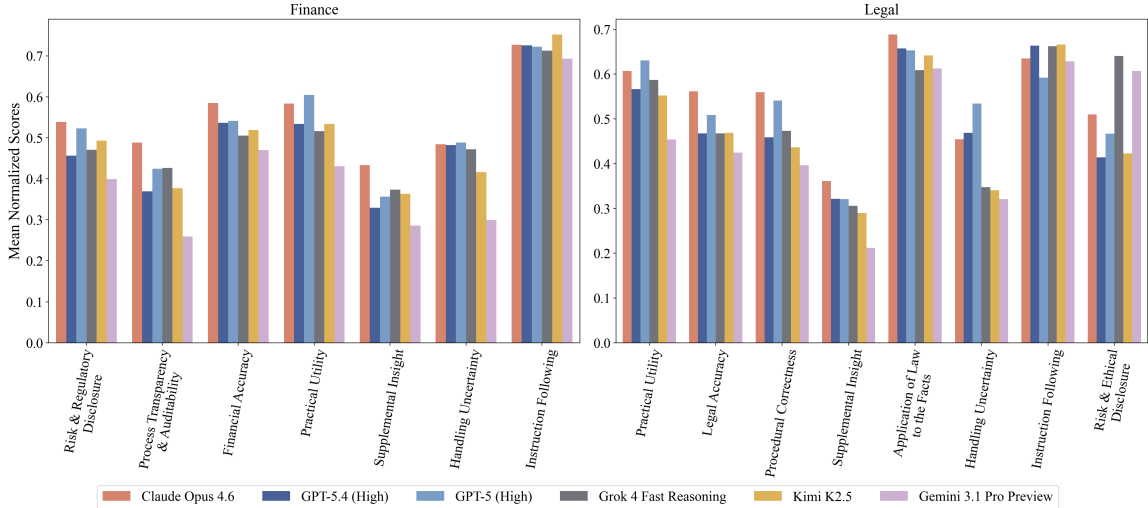


Figure 5: Min-normalized scores per rubric category for PRBench.

Performance across Rubric Categories In Figure 5, we examine the performances of the top five models within each rubric category. For this analysis, we use an alternative scoring mechanism which we call *min-normalized scores* as described in Appendix E.1. This metric is particularly suitable for category-level comparisons, since for certain sample–category combinations, all criteria may be negative, in which case the default scoring method used by Arora et al. (2025) fails to differentiate between models.

While Claude Opus 4.6 is the best overall model, other models lead in Process Transparency & Auditability—the criteria we use to make sure the conclusions are justified in the response—and in Handling Uncertainty and Practical Utility across both domains. Gemini 3.1 Pro Preview lags substantially behind of the other four models in these dimensions. Claude Opus 4.6 and GPT-5 are more than 15% ahead of the other models in Procedural Correctness (namely, identifying the right venues and procedures in legal matters), and about 10% ahead in Practical Utility and Handling Uncertainty. Grok-4 Fast Reasoning and Gemini 3.1 Pro lead by a large margin in *Risk & Ethical Disclosure* for the Legal domain while the latter is lagging behind for Risk & Regulatory Disclosure for financial questions. Detailed descriptions for each category are provided in Table 6 and Table 7 in the appendix.

4.2 Evaluating LLM Judge

To determine which model to use as a judge, we measured the inter-rater agreement between model and expert labels over a collection of randomly

Judge	Agreement with Experts	
	Avg. Cohen’s κ	Avg. Macro F1
o4 Mini	0.603	0.801
GPT-4.1	0.605	0.802
Claude Haiku 4.5	0.535	0.765
Expert	0.589	0.813

Table 2: LLM-Experts and Expert-Expert agreement for grading model responses over 101 tasks. For LLM judges, we calculate the average of agreement with two experts. For the expert judge, the agreement is calculated with the other expert.

sampled 101 tasks. More specifically, we had each LLM judge and two experts grade GPT-5 and Claude Sonnet 4.5 responses independently for each task, indicating whether each rubric criterion was present or not. All grades were then pooled together, and we measure the average Cohen’s κ and Macro F1 scores between each judge and both humans. We report these scores in Table 2. We note that our LLM-expert and expert-expert Macro F1 scores exceed the scores reported in previous work (Arora et al., 2025). We find that all judges demonstrate high and similar agreement scores with both experts, with the exception of Claude Haiku 4.5. In light of these comparable IRA scores, we select *o4-mini* due to its reduced querying costs.

In a small fraction of criteria (<1%), we observe the LLM judge awarding points to responses that address the prompt only superficially, without fully adhering to the specified requirements. For example, when a criterion requires risk-adjusted return feasibility analysis by subsector, the LLM judge assigns credit to a response that provides such anal-

ysis at the portfolio level with some incomplete metrics for each subsector rather than a complete analysis for each subsector level. We further identify that this limitation is asymmetrically benefiting GPT-5 than later GPT releases where the former often provides expanded answers enriched with numeric examples and the latter is more outline-like. We manually inspect GPT-5 Pro and GPT-5.2 Pro responses in Appendix K.

5 Dissecting the Dataset: Prompts, Rubrics, and Beyond

5.1 Inside the Conversations

We analyze the capabilities tested by our prompts along two additional axes: the type of decision the question seeks assistance with, and the economic implications it entails. Specifically, these axes address the questions “*What kind of decision is being made?*” and “*What economic consequence does it affect?*”. We refer to the former as the *Decision Type* and the latter as the *Economic Pathway*. We name the latter category to capture the idea of tracing *pathways of value, risk, or cost*. Overall, we find that the majority of questions in our dataset correspond to high-stakes, real-world decision scenarios that also imply tangible downstream economic impact. The resulting distribution for these dimensions is given in Figure 6. All annotations will be released alongside the dataset to facilitate future research. Results for all evaluated models is given in Appendix L.

5.1.1 Assisting in High-Stakes Decisions

AI systems are increasingly being deployed to support human decision-making across domains such as law, finance, healthcare, and management (Zeiser, 2024; Kim et al., 2025; Khosravi et al.; Vuković et al., 2025; Hillebrand et al., 2025). Yet, evidence on their effectiveness remains mixed: while some studies find that AI assistance can improve consistency and reduce cognitive load, others show that it can amplify errors when models provide incorrect or oversimplified recommendations (Rojas, 2024; Steyvers and Kumar, 2024; Eigner and Händler, 2024). Within our dataset, many prompts extend beyond factual or informational queries, posing genuine decision problems, such as whether to litigate or arbitrate, how to allocate funds, or which market to launch in. To better understand how LLMs engage with such high-stakes reasoning tasks, we recruit domain experts

to annotate each sample with the type of decision it represents.

For all evaluated models, results per decision type are shown in Figure 18. Informational queries are where the performance gap between open-source and proprietary models narrows the most. The gap between the top-scoring model and the runner-ups is generally consistent across categories, with notable exceptions in Planning & Forecasts, Operations, and Processes & Controls. Across all models, the lowest scores are driven by Modeling & Measurement, Planning & Forecasts, and Markets & Transactions in Finance, and Governing Law & Rule in the Legal domain.

5.1.2 Economically Consequential Questions

Recent benchmark efforts have shifted focus from academic problem-solving toward economically valuable tasks that reflect how AI systems can drive productivity and create real-world value (Patwardhan et al., 2025; Mazeika et al., 2025). In our dataset, a majority of prompts naturally carry downstream economic implications: for instance, advising how to allocate capital expenditures under interest-rate shocks, whether to diversify or concentrate portfolio exposure, how to structure cross-border acquisitions to minimize risk, or how to design stress tests that prevent catastrophic losses. In these scenarios, a model’s performance can meaningfully affect financial outcomes—saving costs, reducing risk, or improving efficiency. To capture these dimensions, we recruit domain experts to annotate each sample for its economic pathway, indicating whether and how the question implies a positive or negative economic consequence if answered correctly or incorrectly.

Results per economic pathway are shown in Figure 17. The GPT-5 suite performs particularly well on Compliance (e.g. regulatory adherence and policy interpretation), Value Creation (e.g. identifying revenue growth or cost-reduction opportunities), and Optimizing Funding, think of capital allocation and financing strategy selection. In contrast, Claude Opus 4.5 and GPT-OSS 120B are among the strongest models for information-seeking questions, such as factual explanations, definitions, and descriptive overviews, rather than tasks implying economic impact. Overall models struggle most on questions that requiring devising a strategy to help reduce exposure to market, credit, liquidity or operational risks (Risk & Resilience).



Figure 6: Distribution of *Decision Types* and *Economic Pathways* across Legal and Finance domains.

6 Related Work

6.1 Evaluating LLMs on Economically Valuable Tasks

While dominant expert-level knowledge benchmarks like MMLU (Hendrycks et al., 2021) and GPQA (Rein et al., 2024) test academic reasoning, a recent trend focuses on evaluating AI performance on professional and economically valuable tasks. This includes benchmarks like SWE-Lancer (Miserendino et al., 2025) for freelance development, GDPval (Patwardhan et al., 2025) for tasks across U.S. GDP-contributing occupations, APEX (Vidgen et al., 2025) for high-value work in consulting, finance, law, and healthcare, ProfBench (Wang et al., 2025) that covers tasks in finance and consulting, HealthBench (Arora et al., 2025), and BigLaw Bench (Pereyra et al., 2024). Other novel approaches include the Remote Labor Index (RLI) (Mazeika et al., 2025), which measures AI’s automation potential for remote work, AlphaArena (nof1.ai, 2025), which evaluates AI agents in live financial trading competitions, and Arena Expert (LMarena Team, 2025), which uses human preference voting to evaluate occupational tasks.

However, existing evals for open-ended professional tasks are often limited. They tend to be private (Vidgen et al., 2025; Patwardhan et al., 2025; Pereyra et al., 2024) or require costly human expert judges (LMarena Team, 2025; Mazeika et al., 2025), limiting research, accessibility and scalability. Furthermore, due to the cost of sourcing expert annotations, existing rubric-based professional benchmarks are often small in scale (Vidgen et al., 2025; Wang et al., 2025; Pereyra et al., 2024), lacking sufficient coverage of diverse professional top-

ics. In contrast, PRBench offers a significant public set of 1,100 tasks and 18,711 expert-curated criteria that enables self-served evaluation, an order of magnitude larger than benchmarks like APEX (Vidgen et al., 2025), ProfBench (Wang et al., 2025), and Biglaw Bench (Pereyra et al., 2024), and uniquely leverages multi-turn interactions to build up the context of real legal and finance settings.

6.2 Rubric-Based Evaluation and Reward

Rubric-based evaluation is a key methodology for enabling scalable, automated evaluation of open-ended professional tasks. Its use of self-contained, objective criteria provides the objectivity and style-neutrality essential for knowledge-intensive or reasoning tasks. This approach is used in various recent benchmarks, from general-domain evaluations (Sirdeshmukh et al., 2025; Zhou et al., 2023) to professional domain evaluations (Wang et al., 2025; Vidgen et al., 2025; Arora et al., 2025). It is also applied in agent-focused benchmarks like RLI (Mazeika et al., 2025) and BrowseComp (Wei et al., 2025) to grade complex task artifacts. Beyond evaluation, rubrics are also explored as a reward function for reinforcement learning (Gunjal et al., 2025; Huang et al., 2025). Further studies have explored rubric synthesis techniques from diverse responses (Zhang et al., 2025; Rezaei et al., 2025).

7 Conclusion

We introduced PRBench, a large-scale expert-annotated benchmark for evaluating LLMs on high-stakes professional reasoning in Finance and Law, two domains where reasoning quality directly affects real-world outcomes. By combining over 1,100 expert-authored tasks and 19,000+ rubric cri-

Benchmark	# Samples	Open-Ended QA	Evaluation	# Rubrics	Multi-Domain	Multi-Turn	Open-Source
LegalBench (Guha et al., 2023)	162	<i>Partial</i>	GT	–	✗	✗	✓
LEXam (Fan et al., 2025)	4,886	<i>Partial</i>	GT, LLM Judge	–	✗	✗	✓
BigLawBench (Pereyra et al., 2024)	<i>Private</i>	✓	Rubric, LLM Judge	<i>Private</i>	✗	✗	✗
GreekBarBench (Chlapanis et al., 2025)	310	✓	Rubric, LLM Judge	–	✗	✗	<i>Partial</i>
CorpFin v2 (Vals AI, 2025)	858	✓	GT, LLM Judge	–	✗	✗	✗
ConvFinQA (Chen et al., 2022)	8,281	✗	GT	–	✗	✓	✓
FinanceBench (Islam et al., 2023)	10,231	✗	GT	–	✗	✗	✓
FinanceMATH (Zhao et al., 2024)	1,200	✗	GT	–	✗	✗	✓
FinanceReasoning (Tang et al., 2025)	2,238	✗	GT	–	✗	✗	✓
ProfBench (Wang et al., 2025)	80	✓	Rubric, LLM Judge	2,448	✓	✗	✓
APEX (Vidgen et al., 2025)	200	✓	Rubric, LLM Judge	5,818	✓	✗	✗
HealthBench (Arora et al., 2025)	5,000	✓	Rubric, LLM Judge	48,562	✗	✓	✓
PRBench (Ours)	1,100	✓	Rubric, LLM Judge	18,711	✓	✓	✓

Table 3: Comparison of PRBench with select professional-domain benchmarks. GT stands for Ground-Truth based evaluation. *Partial* indicates a non-major subset of the dataset involved Open-Ended QA.

teria, PRBench enables interpretable, rubric-based evaluation of models on open-ended, economically consequential problems. Our analysis shows that while both proprietary and open-source models demonstrate steady progress, substantial gaps remain in process transparency and domain-specific diligence. Models frequently reach conclusions through incomplete or opaque reasoning, limiting their trustworthiness in professional settings. PRBench provides a framework for objective, fine-grained evaluation of model reasoning. By making this benchmark publicly available, we aim to advance research toward transparent, reliable, and economically valuable AI systems capable of assisting in real-world decision-making.

Limitations

The evaluation with PRBench relies on LLM-based judges and therefore inherits known limitations of automated judging (Szymanski et al., 2025). Although inter-rater agreement is comparable to that between human experts, manual inspection reveals systematic failure modes. In particular, LLM judges can be misled by surface-level or partially correct responses that mention required concepts without satisfying the full intent or granularity specified by the rubric (as exemplified in Section 4 contrasting GPT 5 with GPT 5.2). This may result in over-crediting answers that appear compliant but omit critical details or misapply analysis, introducing subtle biases into model scores.

Moreover PRBench evaluates reasoning in isolation, without requiring models to interact with external systems (e.g., databases, spreadsheets, or proprietary data sources) that are used in real professional workflows. Consequently, scores may not capture performance where tooling is essential.

Finally, although we recruited a geographically diverse set of annotators, the benchmark reflects

their specific areas of expertise and professional workflows. All prompts are written in English, which may limit the extent to which the evaluation captures true model performance in non-English settings or jurisdictions where legal and financial reasoning is typically conducted in other languages.

Ethical Considerations

We have worked with human experts in creating our prompts and rubrics. All annotators are independent contractors and were compensated at rates consistent with fair labor practices and designed to align with applicable local laws. Participation to this study was entirely voluntary, with the option to decline tasks at any time. Importantly, we do not collect any personally identifiable information or sensitive data. We used AI assistance for editing and improving the clarity of the paper and all technical content and conclusions are the authors’ own. We have also used AI to summarize the prompts in Table 9 through Table 12.

Acknowledgments

We thank Karmini Sampath, Emily Chan, Neel Guha and Jerry Xu for helpful feedback and discussions during the development of this work. We also thank Amir Fekrazad, Gabriel Mathews, Shannon Blakeney, Jermaine Ogwuda, Valerie Muigai, Karen Knighton, Diana Bonilla, and Hayden Morse for their contributions to data validation and quality control.

References

Ruth Appel, Peter McCrory, Alex Tamkin, Michael Stern, Miles McCain, and Tyler Neylon. 2025. Anthropropic economic index report: Uneven geographic and enterprise ai adoption. *Anthropic Research*.

- Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. 2025. [Healthbench: Evaluating large language models towards improved human health](#). *Preprint*, arXiv:2505.08775.
- Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use chatgpt. Technical report, National Bureau of Economic Research.
- Zhiyu Chen, SHIYANG LI, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. [Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Odysseas S. Chlapanis, Dimitrios Galanis, Nikolaos Aletras, and Ion Androutsopoulos. 2025. [GreekBar-Bench: A challenging benchmark for free-text legal reasoning and citations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25099–25119, Suzhou, China. Association for Computational Linguistics.
- Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. 2025. Arc-agi-2: A new challenge for frontier ai reasoning systems. *arXiv preprint arXiv:2505.11831*.
- Eva Eigner and Thorsten Händler. 2024. [Determinants of llm-assisted decision-making](#). *arXiv preprint arXiv:2402.17385*.
- Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstruwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, Daniel Brunner, Markus Leippold, Mrinmaya Sachan, Alexander Stremitzer, Christoph Engel, Elliott Ash, and Joel Niklaus. 2025. [Lexam: Benchmarking legal reasoning on 340 law exams](#). *ArXiv*, abs/2505.12864.
- Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Re, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, and 21 others. 2023. [Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. 2025. [Rubrics as rewards: Reinforcement learning beyond verifiable domains](#). *Preprint*, arXiv:2507.17746.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Luis Hillebrand, Sebastian Raisch, and Jonathan Schad. 2025. [Managing with artificial intelligence: An integrative framework](#). *Academy of Management Annals*, 19(1):343–375.
- Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, Xijun Gu, Peiyi Tu, Jiabin Liu, Wenyu Chen, Yuzhuo Fu, Zhiting Fan, Yanmei Gu, Yuanyuan Wang, Zhengkai Yang, and 2 others. 2025. [Reinforcement learning with rubric anchors](#). *Preprint*, arXiv:2508.12790.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. [Financebench: A new benchmark for financial question answering](#). *ArXiv*, abs/2311.11944.
- Mohsen Khosravi, Zahra Zare, Seyyed M. Mojtabaiean, and Reyhane Izadi. Artificial intelligence and decision-making in healthcare: A thematic analysis of a systematic review of reviews. *Health Services Research and Managerial Epidemiology*, 11.
- Sihyun Kim, Sangyoon Yi, and Sung-Pil Park. 2025. [Prioritizing challenges in ai adoption for the legal domain: A systematic review and expert-driven AHP analysis](#). *PLOS ONE*, 20(6):e0326028.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.
- LMarena. 2025. [Arena expert and occupational categories](#). *LMarena News*. Blog post.
- LMarena Team. 2025. [Arena expert and occupational categories](#). <https://news.lmarena.ai/arena-expert/>. Accessed: 2025-11-11.
- Mantas Mazeika, Alice Gatti, Cristina Menghini, Udari Madhushani Sehwal, Shivam Singhal, Yury Orlovskiy, Steven Basart, Manasi Sharma, Denis Peskoff, Elaine Lau, Jaehyuk Lim, Lachlan Carroll, Alice Blair, Vinaya Sivakumar, Sumana Basu, Brad Kenstler, Yuntao Ma, Julian Michael, Xiaoke Li, and 28 others. 2025. [Remote labor index: Measuring ai automation of remote work](#). *Preprint*, arXiv:2510.26787.
- Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. 2025. [Swe-lancer: Can frontier llms earn \\$1 million from real-world freelance software engineering?](#) *Preprint*, arXiv:2502.12115.
- nof1.ai. 2025. Alpha arena. <https://nof1.ai/>. Accessed: 2025-11-11.

- Melissa Pan, Negar Arabzadeh, Riccardo Cogo, Yuxuan Zhu, Alexander Xiong, Lakshya Agrawal, Huanzhi Mao, Emma Shen, Sid Pallerla, Liana Patel, Shu Liu, Tianneng Shi, Xiaoyuan Liu, Jared Davis, Emmanuele Lacavalla, Alessandro Basile, Shuyi Yang, Paul Castro, Daniel Kang, and 6 others. 2025. Measuring agents in production. *arXiv preprint arXiv:2512.04123*.
- Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Simón Posada Fishman, Marwan Aljubei, Phoebe Thacker, Lorraine Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. 2025. *Gdpval: Evaluating ai model performance on real-world economically valuable tasks*. Preprint, arXiv:2510.04374.
- Julio Pereyra, Elizabeth Lebens, Matthew Guillod, Laura Toulme, Cameron MacGregor, David Muder, Karl de la Roche, Emilie McConnachie, Jeremy Pushkin, Rina Kim, Aaron Chan, Jenny Pan, Boling Yang, Nan Wu, Niko Grupen, Lauren Oh, Aatish Nayak, and Gabriel Pereyra. 2024. Introducing biglaw bench. <https://www.harvey.ai/blog/introducing-biglaw-bench>. Accessed: 2025-11-11.
- Long Phan, Alice Gatti, Nathaniel Li, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dan Hendrycks, Ziwen Han, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, and 1081 others. 2026. *A benchmark of expert-level academic questions to assess ai capabilities*. *Nature*, 649(8099):1139–1146.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. *GPQA: A graduate-level google-proof q&a benchmark*. In *First Conference on Language Modeling*.
- MohammadHossein Rezaei, Robert Vacareanu, Zihao Wang, Clinton Wang, Bing Liu, Yunzhong He, and Afra Feyza Akyürek. 2025. *Online rubrics elicitation from pairwise comparisons*. Preprint, arXiv:2510.07284.
- Nikki Rojas. 2024. *Does ai help humans make better decisions?* Harvard Gazette.
- Scale AI. 2025. *Seal showdown: Technical report*. Technical report, Scale AI. Preliminary results and methodology for the SEAL Showdown leaderboard.
- Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. 2025. *Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms*. Preprint, arXiv:2501.17399.
- Mark Steyvers and Aakriti Kumar. 2024. *Three challenges for ai-assisted decision-making*. *Perspectives on Psychological Science*, 19(5):722–734.
- Annalisa Szymanski, Noah Ziemis, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. 2025. *Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks*. In *Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI '25*, page 952–966, New York, NY, USA. Association for Computing Machinery.
- Zichen Tang, Haihong E, Ziyang Ma, Haoyang He, Jiacheng Liu, Zhongjun Yang, Zihua Rong, Rongjin Li, Kun Ji, Qing Huang, Xinyang Hu, Yang Liu, and Qianhe Zheng. 2025. *FinanceReasoning: Benchmarking financial numerical reasoning more credible, comprehensive and challenging*. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15721–15749, Vienna, Austria. Association for Computational Linguistics.
- Inc. Vals AI. 2025. *Corpfin (v2): A private benchmark evaluating understanding of long-context credit agreements*. Online benchmark. Accessed on April 20, 2026.
- Bertie Vidgen, Abby Fennelly, Evan Pinnix, Chirag Mahapatra, Zach Richards, Austin Bridges, Calix Huang, Ben Hunsberger, Fez Zafar, Brendan Foody, Dominic Barton, Cass R. Sunstein, Eric Topol, and Osvald Nitski. 2025. *The ai productivity index (apex)*. Preprint, arXiv:2509.25721.
- Darko B. Vuković, Senanu Dekpo-Adza, and Stefana Matović. 2025. *AI integration in financial services: A systematic review of trends and regulatory challenges*. *Humanities and Social Sciences Communications*, 12(1):562.
- Zhilin Wang, Jaehun Jung, Ximing Lu, Shizhe Diao, Elie Evans, Jiaqi Zeng, Pavlo Molchanov, Yejin Choi, Jan Kautz, and Yi Dong. 2025. *Profbench: Multi-domain rubrics requiring professional knowledge to answer and judge*. Preprint, arXiv:2510.18941.
- Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. *Browsecomp: A simple yet challenging benchmark for browsing agents*. Preprint, arXiv:2504.12516.
- Jannik Zeiser. 2024. *Owning decisions: Ai decision-support and the attributability-gap*. *Science and Engineering Ethics*, 30(4):1–19.
- Junkai Zhang, Zihao Wang, Lin Gui, Swarnashree Mysore Sathyendra, Jaehwan Jeong, Victor Veitch, Wei Wang, Yunzhong He, Bing Liu, and Lifeng Jin. 2025. *Chasing the tail: Effective rubric-based reward modeling for large language model post-training*. Preprint, arXiv:2509.21500.
- Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024. *Financemath: Knowledge-intensive math reasoning in finance domains*. In *Proceedings of the 62nd Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 12841–12858.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *Preprint*, arXiv:2311.07911.

A Evaluations with Web Search & Code Interpreter

In this section, we explore the performance of chat models when given access to web search and code interpreter tools. Agents are evaluated over this subset of hard tasks, and o4-mini is used as the judge to grade their responses. The Gemini models are allotted a budget of 32,768 thinking tokens. For Anthropic models ≤ 4.5 use a thinking budget of 16,384 tokens. All models (if configurable) use a reasoning effort of ‘high’. For Opus 4.5 we use the checkpoint 20251101.

Table 4 reveals that access to web search is helpful for legal tasks in more than half of the models tested. Our analysis reveals that when regression happens it is primarily due to over-reliance on external sources rather than providing a cohesive answer to the question. In Table 5, we enable both search and code interpreter tools; we observe that the code interpreter does not provide additional performance boosts over web search, except for Grok 4 and Gemini 3 Pro Preview on legal tasks.

B Assessing the Impact of Response Length on Rubric Scores

In Figure 7, we present average response lengths for all models. We observe that while Kimi K2 Thinking, Claude Sonnet 4.5, GPT OSS 120B, and Grok 4 Fast Reasoning achieve similar scores for Finance, Kimi K2 Thinking produces substantially shorter responses than the others. GPT-5 and O3 models strike a strong balance between performance and efficiency. In contrast, Claude Opus 4.6 produces about 40% longer sequences from the runner-up model. GPT-OSS 120B, Gemini 2.5 Flash, and Claude Sonnet 4.5 are the top three models that produce significantly longer responses, nearly twice the length of the comparable responses.

C Dataset Details

In Figure 8, we provide distribution of topics covered in PRBench. Figure 9 shows the distribution of weights assigned to each criterion and Figure 10

shows the frequencies of rubric categories. Negative weights are strictly reserved for penalizing undesired properties. In multi-turn conversations, we sample intermediate assistant turns from one of GPT OSS 20B, Mistral or Deepseek R1. User turns are provided by the human annotators.

C.1 Topic Distribution

PRBench covers a diverse range of topics as shown in Figure 8.

C.2 Rubric Category Definitions

We provide annotator-facing definitions of our rubric categories in Table 6 and 7.

C.3 Rubric Desiderata

- **Constructive:** Each criterion should be correct, precise, and free of internal errors or misconceptions.
- **Mutually Exclusive and Collectively Exhaustive:** No criterion is repeated or redundant, so that a model is not penalized twice for the same mistake. At the same time, the sum of all criteria should be thorough enough to cover all aspects of a perfect response.
- **Atomic:** Each rubric criterion evaluates exactly one distinct aspect and should contain no bundling of multiple criteria into a single criterion.
- **Objective:** Criteria should be binary (true or false) and objective, where a majority of readers should agree on whether a given model response satisfies the criteria.
- **Self-Contained:** All info needed to score a response must be included in the criterion.

C.4 Criteria Importance Scale

We allow annotators to choose from six categories associated with different importance levels as described in Table 8.

C.5 Jurisdiction Coverage of Legal Prompts

Figure 11 shows distribution of jurisdictions for Legal domain prompts across the US and globally.

D LLM Judge Prompt Template

Our LLM judge template follows a similar structure to [Arora et al. \(2025\)](#) and will be available via the evaluation harness upon acceptance.

Table 4: Average clipped performance scores over Hard subset of 300 finance and 250 legal tasks using o4-mini as the judge. We report scores with web search turned on (**on**) and off (**off**). Results are over a single trial and obtained using respective provider APIs when available to enable right tools hence might differ slightly from those presented in the main text.

	Finance		Legal	
	on	off	on	off
Gemini 2.5 Pro	0.207	0.266	0.255	0.297
Gemini 3 Pro Preview	0.265	0.270	0.289	0.286
Gemini 3.1 Pro Preview	0.286	0.291	0.322	0.317
o3 (High)	0.347	0.336	0.398	0.352
GPT-5 (High)	0.382	0.394	0.383	0.377
GPT-5.2 (High)	0.328	0.326	0.348	0.315
GPT-5.4 (High)	0.317	0.330	0.356	0.325
Sonnet 4.5	0.290	0.323	0.281	0.294
Opus-4.5	0.346	0.397	0.352	0.375
Opus-4.6	0.358	0.383	0.376	0.397
Grok 4 Fast Reasoning	0.333	0.314	0.374	0.325

Table 5: Average scores in the Hard subset of 300 finance with a o4-mini as the judge. We report scores with web search + code interpreter turned on and off. Results are over a single evaluation run. Grok 4 is Fast Reasoning.

	Finance		Legal	
	on	off	on	off
o3 (High)	0.342	0.336	0.400	0.352
GPT-5 (High)	0.381	0.394	0.383	0.377
GPT-5.2 (High)	0.318	0.326	0.338	0.315
GPT-5.4 (High)	0.328	0.330	0.365	0.325
Opus-4.5	0.271	0.351	0.293	0.315
Gemini 2.5 Pro	0.192	0.266	0.255	0.297
Gemini 3 Pro	0.262	0.270	0.292	0.286
Gemini 3.1 Pro	0.281	0.291	0.319	0.317
Grok 4	0.325	0.314	0.377	0.325

E Evaluation Details

We prepend reference texts to the prompt. All models are evaluated at reasoning mode set to High except for Claude Sonnet 4.5 where we set the thinking budget at 32,768 and for Claude Opus models to 16,384. Gemini 3 Pro Preview is tested with default thinking settings; Gemini 2.5 Pro and Flash models are evaluated at thinking_budget=-1 after observing no consistent improvements for setting a fixed thinking budget. Otherwise, we use the default sampling settings for each model. The code to evaluation harness will be released along with the dataset.

E.1 PRBench Scores

Following Arora et al. (2025), scores for each model M on PRBench are calculated as follows:

1. For the desired dataset, we evaluate prompts p_1, \dots, p_n . Each prompt p_j has k_j rubrics $r_{j,1}, \dots, r_{j,k_j}$ with weights $w_{j,i} \in [-10, 10]$, $w_{j,i} \neq 0$
2. The model M produces a response $m_j = M(p_j)$ given the prompt
3. An LLM judge grades m_j using each rubric individually, assigning a binary indicator $I_{j,i} \in \{0, 1\}$

Table 6: Rubric category definitions for Finance.

Dimensional Rating	Definition
Financial Accuracy	Maintains mathematical, factual, and financial accuracy, applying financial metrics and financial principles (e.g., time value, conservatism, materiality, etc.) correctly. Generally aligns with GAAP or IFRS standards and avoids contradictions.
Process Transparency & Auditability	Demonstrates correct work by providing formulas, reasoning steps, references, or supporting data so the answer can be reviewed, reproduced, or challenged by another professional.
Handling Uncertainty	Addresses incomplete or ambiguous information by highlighting assumptions, proposing clarifying questions, or presenting alternative scenarios.
Practical Utility	Provides concrete, actionable guidance such as next steps, strategies, checklists, examples, or references to external resources as needed, ensuring the response is directly useful rather than purely theoretical when appropriate.
Risk & Regulatory Disclosure	Describes associated financial or regulatory compliance-related risks or considerations connected to either the user request or the methods outlined in the response.
Supplemental Insight	Covers other relevant information, steps, or exceptions needed for a reliable answer beyond the primary objective of the question and answers.
Instruction Following	Follows auxiliary instructions in the prompt outside of answering the primary question, including tailoring for the finance function (such as corporate finance, advisory, investment banking, or investment management), geographic location, demographic, or personal situation, and ensuring the response matches the required role (expert vs. non-expert).

Table 7: Rubric category definitions for Law.

Dimensional Rating	Definition
Legal Accuracy	Identifies applicable law (jurisdiction) correctly and ensures statements of law are correct, legally valid, and consistent with authoritative, verifiable sources (such as statutes, case law, and regulations). The definition is materially complete and answers the question: “Is the statement of the law applicable, correct, and complete?”
Application of Law to the Facts	Correctly applies the law to the provided facts and answers the question: “Given these specific facts, what does the law require or permit – i.e., what rights, duties, remedies, or outcomes follow?”
Procedural Correctness	Conforms to legal processes and formal requirements, including deadlines, document structure, and jurisdiction-specific rules. It answers: “Does the response follow the official rules of how this is done?”
Handling Uncertainty	Addresses incomplete or ambiguous information by highlighting assumptions, asking follow-up questions to clarify the facts, or presenting alternative scenarios that explain how the law applies to different sets of facts.
Practical Utility	Provides concrete, actionable guidance as needed, such as next steps, strategies, checklists, examples, or references to external resources.
Risk & Ethical Disclosure	Flags limitations, includes disclaimers where necessary, avoids misleading or unsafe advice, and respects boundaries on unauthorized practice of law.
Supplemental Insight	Covers additional legally relevant principles, elements, steps, defenses, or exceptions that contribute to a reliable answer beyond the primary objective of the question.
Instruction Following	Follows auxiliary instructions in the prompt outside of answering the primary question, including implicit or explicit requirements, role-appropriate tailoring (lawyer vs. non-expert), and matches the jurisdiction, task fidelity or difficulty.

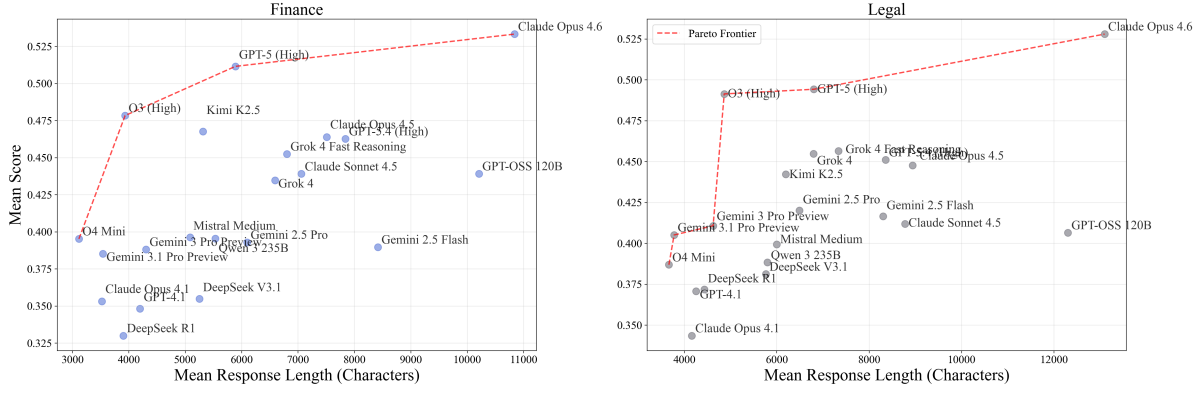


Figure 7: Scores vs. response lengths. While some models achieve similar performance (e.g., Kimi K2 Thinking and Claude Sonnet 4.5), conciseness appears as a differentiating factor. We only count alphanumeric characters, as most outputs contain Markdown formatting.

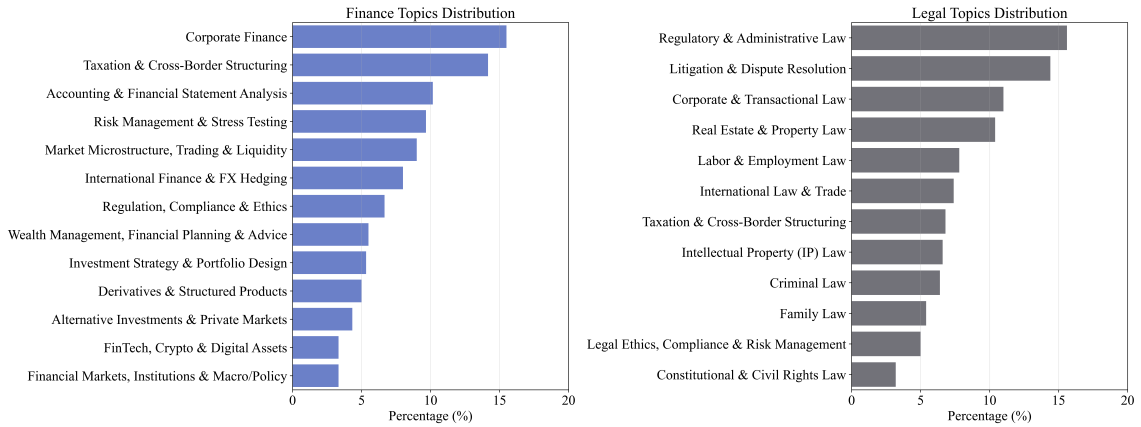


Figure 8: Topic distribution of prompts in PRBench across Finance and Legal domains.

The score for response m_j is

$$s_j = \frac{\sum_{i=1}^{k_j} w_{j,i} I_{j,i}}{\sum_{i: w_{j,i} > 0} w_{j,i}} \quad (1)$$

The denominator is always > 0 because each prompt has at least one positive-weight rubric. The overall score for model M is the mean over prompts

$$S(M) = \max\left(0, \frac{1}{n} \sum_{j=1}^n s_j\right). \quad (2)$$

E.2 Min-Normalized Scoring

We also propose a normalized score for rubric-based benchmarks, which adjusts each response score by the minimum possible score for its corresponding prompt’s rubrics. The normalized score

for response m_j is

$$\tilde{s}_j = \frac{\sum_{i=1}^{k_j} w_{j,i} I_{j,i} - \sum_{i: w_{j,i} < 0} w_{j,i}}{\sum_{i: w_{j,i} > 0} w_{j,i} - \sum_{i: w_{j,i} < 0} w_{j,i}} \quad (3)$$

The normalized model score is the mean over prompts:

$$\tilde{S}(M) = \frac{1}{n} \sum_{j=1}^n \tilde{s}_j \quad (4)$$

Normalization makes scores more comparable across prompts with different numbers and magnitudes of positive and negative rubrics, such as rubrics that are in different categories. For example, it is natural for “Supplemental Insight” to have lower absolute weighted negative rubrics than the “Legal / Financial Accuracy” category. Thus, we use score normalization to compute the per-rubric-category scores reported in Figure 5.

Table 8: Rubric criteria scoring levels.

Score Range	Description
[+8, +10]	Critically Important: These are essential criteria without which the response would fail to adequately address the prompt. They define the minimally viable rubric set and capture only the core, indispensable elements of a correct and sufficient answer.
[+4, +7]	Important: Criteria that meaningfully strengthen a response by adding depth, accuracy, or completeness. They materially shape the response’s quality but are not strictly required for it to be acceptable.
[+1, +3]	Slightly Important: Optional enhancements or “nice-to-have” details that improve clarity or precision but do not affect the core correctness of the response.
[−3, −1]	Slightly Detrimental: Minor issues or irrelevant tangents that slightly detract from quality or focus but do not undermine reasoning or factual integrity.
[−7, −4]	Detrimental: Significant errors or omissions that meaningfully weaken the response, such as misleading reasoning, incorrect facts, or major structural flaws, though the response remains generally valid.
[−10, −8]	Critically Detrimental: Severe errors that render the response fundamentally invalid, harmful, or unethical. These issues directly contradict the prompt or destroy the credibility of the reasoning.

Furthermore, normalization also reduces sensitivity to how a rubric is phrased (e.g., “presence of a problem” as a negative rubric versus “absence of a problem” as a positive rubric). This helps avoid situations where adding more negative rubrics, or reframing positives as negatives, would mechanically deflate raw averages. For PRBench, in line with our best practices described in Section 3.2, we ensure our rubrics are phrased to always check the presence (existence) of desired or undesired characteristics of a response, which further mitigates this issue.

F Additional Insights from Hierarchical Clustering on Rubrics

For a granular assessment of the rubrics, we perform hierarchical clustering on all criteria in PRBench. The objective is to move from high-level performance scores to a fine-grained understanding of specific model capability gaps. We first identify the capabilities required to score highly on each rubric and then construct a five-level hierarchical clustering tree by clubbing similar criteria together.

We perform this analysis separately for the Finance and Legal domains. Using the model performance of each criterion, we identify fine-grained clusters where the model demonstrates significant under-performance compared to other clusters. This information can be used to guide further analysis and fine-tuning of model capabilities.

Figure 12 illustrates the hierarchical clustering using a representative example from our dataset from the Finance domain. The hierarchy includes four levels of clustering. In this instance, we analyze the criteria from the “Risk and Regulatory

Disclosure” rubric category (Level 1 in Figure 12). We then dynamically derive the subsequent granular layers (Levels 2 to 4) across the entire dataset. Specifically, this example maps to “The Law Application: International and Domestic Compliance” at Level 2 and further refines to “Compliance and risk analysis under global tax anti-avoidance frameworks” at Level 3. This multi-level hierarchical clustering enables flexibility in deriving insights, allowing us to select a specific level of granularity depending on interested capabilities.

Applying this methodology to our rubrics shows significant performance differences between models. For instance, within the Legal domain, the cluster “Advanced corporate and *international tax law*, strategy, and compliance services.”, shows a significant performance difference: GPT-5 achieved 0.64 accuracy, whereas Grok-4 only achieved 0.16. A similar disparity was observed in the Finance domain for “Strategic planning, compliance, and optimization for international corporate tax.” cluster. Here, we found that Claude Opus 4.1 lags behind substantially, with only 0.34, where the top performing model achieved 0.76.

G Additional Details for Rubric Validation Assessment

To validate the reliability of the evaluation used in PRBench, we conducted a follow-up study, which we refer to as Rubric Validation Assessment. In this study, subject-matter experts (who are different from the authors of the original rubric) reviewed each rubric criterion and selected one of *Agree* and *Disagree* where agreement was based on verifying whether each criterion was *well-constructed*, *accu-*

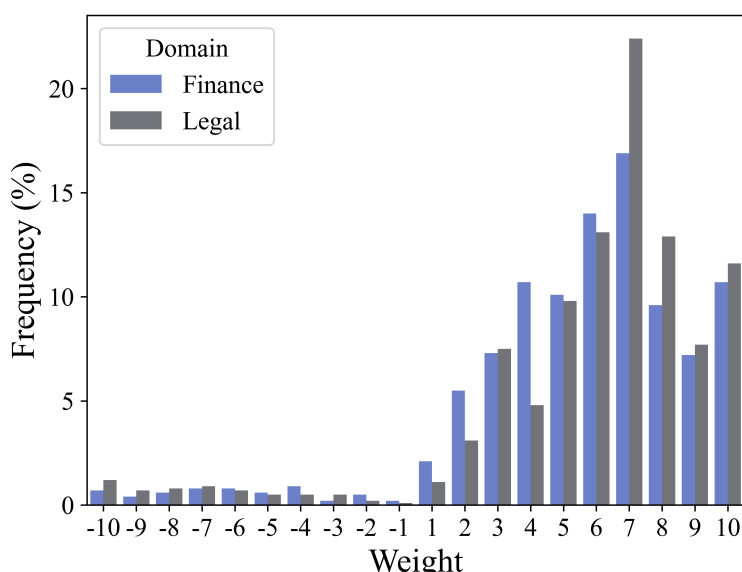


Figure 9: Distribution of weights for each rubric across Finance and Legal domains. We observe that the weights on the both ends of the spectrum are used more frequently in Legal than Finance.

rate, and relevant to the task. In either case, the annotators provided short written justifications to explain their rationale. As a result, the annotators agreed that the criteria were justified 93.9% of the time.

H Definitions for Decision Type and Economic Pathway

Table 9-12 include definitions for decision type and economic pathway annotations for both domains.

I Sample Tasks

We provide two sample tasks in Figure 13 and Figure 14 and the entire dataset which will be released publicly upon acceptance is submitted as supplementary material.

J Results Across Topics

Figure 15 and Figure 16 demonstrate scores for the top models across topics of PRBench.

K Analysis of Performance Across Generations of LLM Models

More recent models improved on PRBench: we observe gains from Kimi K2 Instruct to Kimi K2 Thinking (+0.04), Claude Opus 4.1 to 4.5 (+0.10), and GPT-4.1 to GPT-5 (+0.15). In contrast, the Gemini 3 suite performs largely on par with Gemini 2.5. A closer examination reveals that Gemini 3

Pro is substantially more token-efficient: as shown in Figure 7, Gemini 2.5 Pro achieves similar scores for both domains while using at least 50% more tokens.

We also find that GPT-5.2 Pro trails GPT-5 by approximately 0.05 points. To understand this gap, we manually analyzed 3,000 criteria where GPT-5.2 underperforms GPT-5, in collaboration with finance and legal experts. We identify that the primary difference lies in response style and coverage: GPT-5 more frequently delivers comprehensive answers that explore multiple scenarios and include simulated calculations for open-ended queries, whereas GPT-5.2 responses tend to be more outline-driven and geared toward setting up multi-turn interactions. This gap is amplified by the fact that PRBench rubrics explicitly reward comprehensive, fully worked responses over high-level or scaffolded answers.

L Results Across Economic Pathway and Decision Types

For all evaluated models, results per decision type are shown in Figure 18.

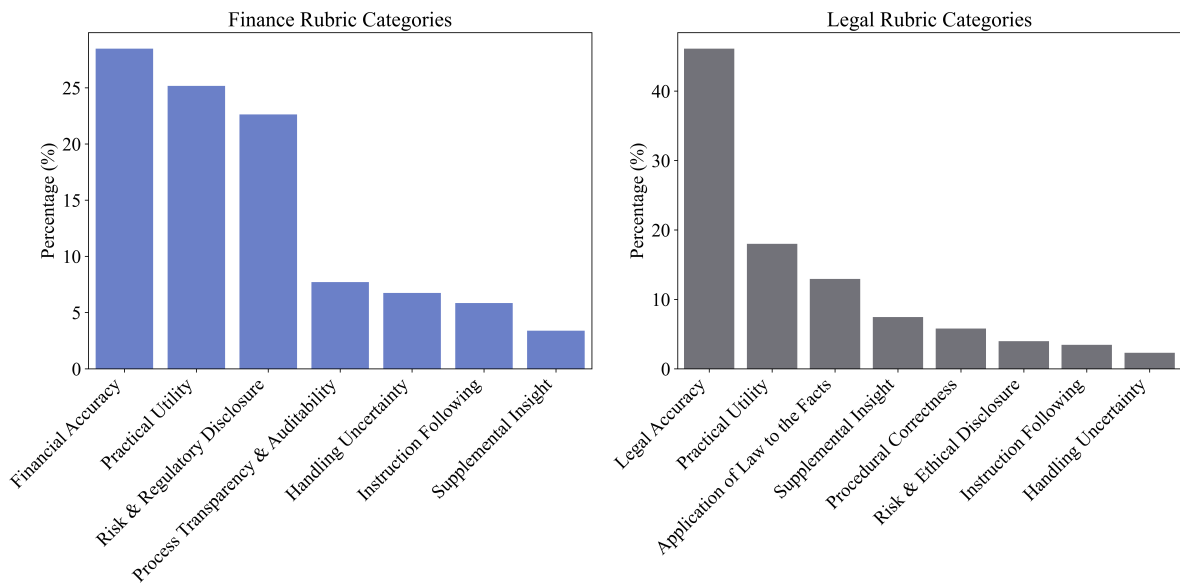


Figure 10: Distribution of rubric categories

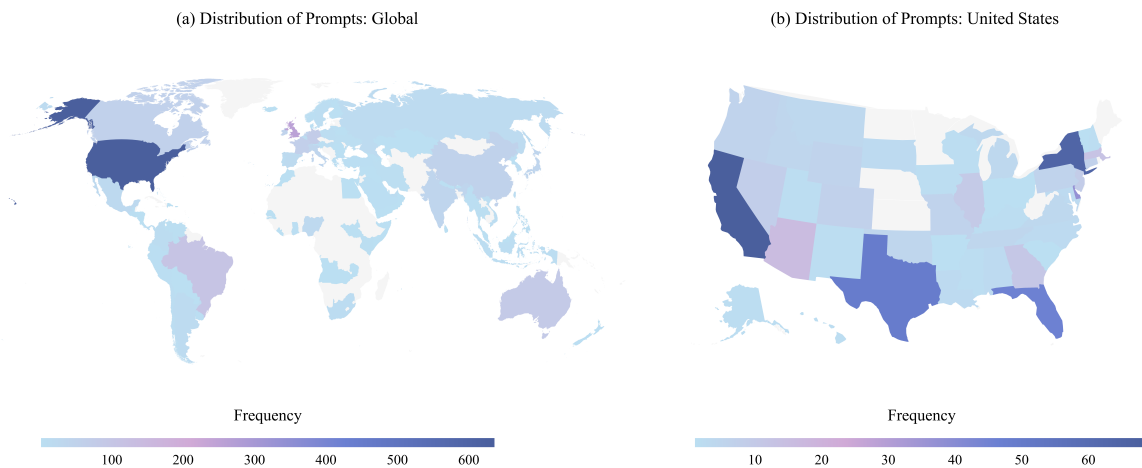


Figure 11: (a) A global frequency map showing the 114 countries and dependencies, and (b) A localized frequency map showing the 47 US states and territories covered by PRBench prompts across both Finance and Legal domains.

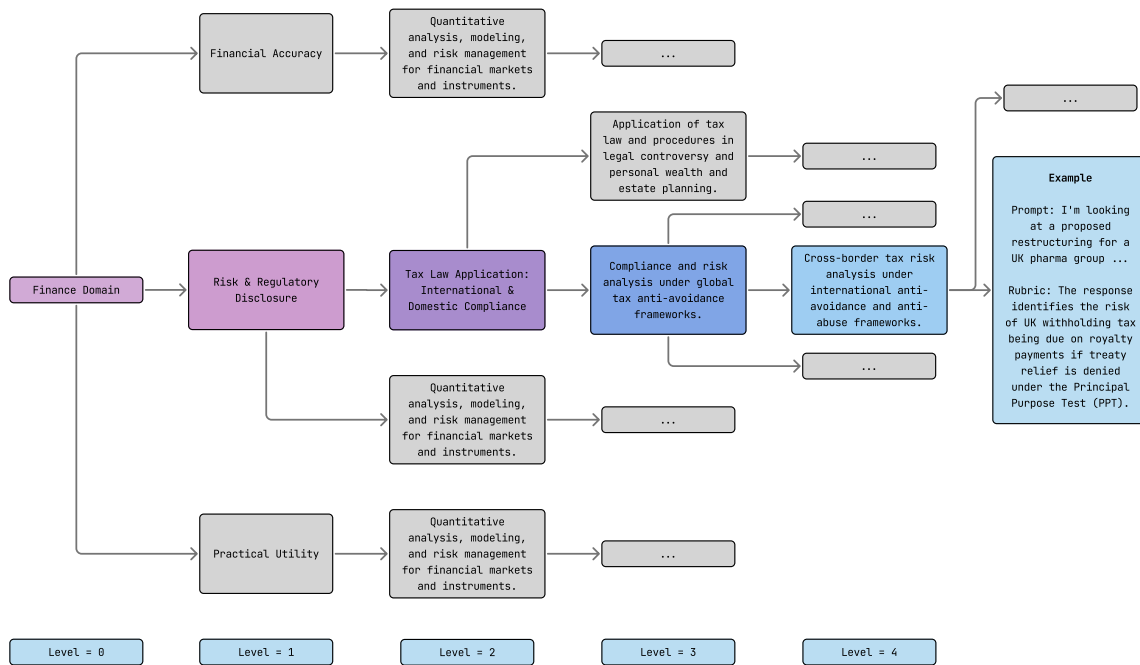


Figure 12: Automatically clustering the available rubric criteria enables fine-grained analysis of failure modes.

Table 9: Legal – Decision Types

Name	Description	Examples
Governing Law & Rule	Determines whether a law, regulation, clause, or doctrine applies to the facts.	Does FDA regulation apply here? Do NY overtime laws govern remote staff? Does GDPR cover this dataset? Is this contract clause enforceable? Does constitutional protection extend to corporations?
Duty and Obligation	Defines what parties must do — statutory, contractual, or regulatory requirements.	Must we provide paid parental leave? Are directors required to disclose conflicts? Do we owe continuing care duties? Is notification to regulator mandatory? When must tax be remitted?
Rights / Entitlement / Exemptions	Identifies what parties may claim, enjoy, or be exempt from — rights, permissions, privileges.	Can employee demand severance pay? Do we have exclusive patent rights? Is tenant entitled to early termination? Can a parent relocate a child abroad? Do shareholders have inspection rights?
Compliance	How to operationalize laws or structure transactions to stay compliant or implement policies.	How to structure merger to avoid liability? What HR policy updates are required? How to comply with PBS prescribing rules? Which filings needed for EU expansion? How to implement anti-bribery controls?
Procedure, Forum & Jurisdiction	Where and how a matter proceeds — forum choice, motion sequence, appellate route.	Which court has jurisdiction? Should we file in federal court? Will the appellate court affirm? Can dispute be sent to arbitration? When is the appeal deadline?
Claims & Litigation Strategy	What claims or defenses to assert, and how to frame them procedurally and doctrinally.	Should we move to dismiss? Can negligence rely on criminal statute? What precedent supports our motion? Should we plead estoppel or waiver? Is summary judgment strategically sound?
Risk & Outcome Forecasting	Predicts likely results, exposure, penalties, or success probabilities.	What's our exposure under wage law? How likely is appellate reversal? What damages could be awarded? What's the fine range for violation? What's litigation success probability?
Negotiation & Deal Strategy	How to bargain, structure, or trade concessions in business, regulatory, or settlement contexts.	How to negotiate stock-for-tax swap? What's best anchor in settlement talks? How to balance indemnity vs. price? Which terms are fallback vs. walk-away? How to sequence multi-party negotiation?
Other	Decision requests that don't fit the above in this lean scheme; use sparingly.	–
Non-decision / Informational	General explanation, commentary, or background.	–

Table 10: Legal – Economic Pathways

Name	Description	Examples
Penalty and Damages Avoidance	Decisions that prevent fines, lawsuits, or sanctions by ensuring lawful conduct and reducing liability exposure.	<ul style="list-style-type: none"> Will failing to notify regulators trigger penalties? How do we avoid wage-and-hour violations? Does our ad campaign risk consumer-protection fines? Should we update safety policies to reduce liability? What steps prevent data-breach penalties?
Transaction Economics	Structuring deals or tax arrangements to maximize value, efficiency, and post-transaction outcomes.	<ul style="list-style-type: none"> How should we structure the merger for tax efficiency? Does an asset purchase reduce future liabilities? Should we use a holdco structure to improve economics? Which deal terms minimize post-closing disputes? Would a licensing model generate better economics?
Compliance Efficiency	Designing cost-effective systems and controls to meet regulatory requirements and minimize compliance burden.	<ul style="list-style-type: none"> How do we streamline AML checks without overspending? Should we centralize compliance reviews to cut costs? What's the least burdensome way to meet new reporting rules? Can we automate disclosures to reduce manual workload? How do we simplify our governance policies efficiently?
Market Access	Securing or maintaining licenses, approvals, or conditions needed to operate and expand legally in target markets.	<ul style="list-style-type: none"> Do we need new licenses to enter the EU market? How do we maintain eligibility for Medicaid contracts? What requirements must we meet to sell in California? Will our product updates trigger new certifications? How do we retain export authorization after expansion?
Rights and Asset Protection	Safeguarding ownership, IP, and contractual rights to preserve or recover economic value.	<ul style="list-style-type: none"> Should we file a trademark to protect brand value? Can we enforce our patent against new entrants? How do we prevent a partner from misusing our data? Should we pursue damages for IP infringement? How do we secure title before selling the asset?
Contractual Risk Allocation	Managing risk through contract terms such as indemnities, liability caps, and dispute clauses.	<ul style="list-style-type: none"> Should we negotiate a higher liability cap? Does the draft indemnity expose us to excess risk? Which dispute clause minimizes future cost? Should we require reps and warranties insurance? How do we allocate compliance obligations in the contract?
Other	Legal-economic effects that do not clearly fit in the main pathways.	–
Informational / Educational Only	Purely explanatory or conceptual content with no direct economic consequence.	–

Table 11: Finance – Decision Types

Name	Description	Examples
Governance & Policy	Set enduring rules or postures such as accounting/tax elections, risk appetite, or disclosure stance.	Should we elect LIFO or FIFO for tax reporting? Do we raise our risk appetite for credit exposure? Should dividends be fixed or discretionary? Do we disclose climate risks in MD&A this year?
Modeling & Measurement	Define how value, exposure, or performance is measured, modeled, and interpreted.	How should we measure portfolio VaR across currencies? What's the right discount rate for project valuation? Do we model beta using weekly or monthly returns? How to estimate expected credit loss under IFRS 9?
Capital & Funding	Choose balance-sheet structure, financing mix, and capital allocation priorities.	Should we issue new equity or refinance debt? How much leverage can we take without breaching covenants? Do we fund expansion from retained earnings or external capital? Is it optimal to repurchase shares at current valuation?
Markets & Transactions	Decide how, when, and at what price to transact in markets or strategic deals.	When's the best time to execute the bond buyback? Should we hedge FX now or wait for better liquidity? At what price do we enter the secondary offering? Which trading venue minimizes slippage for this order?
Operations, Processes & Controls	Set repeatable cash, control, and process steps to meet operational and financial obligations.	How do we automate vendor payment approvals? Should we shorten the monthly close cycle? What's the best control for petty cash discrepancies? How can we speed up receivables collection safely?
Planning & Forecasts	Set budgets, targets, scenarios, and rolling forecasts.	Should we raise our revenue target for next quarter? How much buffer to build into cash forecasts? Do we base next year's budget on trend or zero-based planning? What's the scenario if rates rise by 100 bps?
Compliance & Reporting	Ensure financial actions, records, and disclosures align with regulatory, accounting, and internal standards.	Do we meet IFRS 16 lease disclosure requirements? Are we compliant with new AML reporting thresholds? What filings are due after our debt restructuring? Do we need auditor sign-off before publishing results?
Other	Decision requests that don't fit the above in this lean scheme; use sparingly.	–
Non-decision / Informational	General explanation or background without a decision component.	What's the difference between EBITDA and operating income? How do interest rate swaps work? What is free cash flow conversion? How is goodwill impairment tested?

Table 12: Finance – Economic Pathways

Name	Description	Examples
Value Creation	Decisions that increase profitability, valuation, or investment performance through higher earnings, NPV, IRR, or ROE.	Should we invest in automation to boost ROI? Does expanding into Asia improve our NPV? Will share buybacks lift EPS more than dividends? How much value does the new product add to EBITDA?
Operating Efficiency	Actions that improve cost structure, productivity, or capital utilization.	Can we cut logistics costs without hurting service? Should we consolidate warehouses to free up capital? Will outsourcing payroll improve margin efficiency? How do we reduce idle capacity in production?
Risk & Resilience	Strategies that reduce exposure to market, credit, liquidity, or operational risks.	Should we hedge commodity exposure at current prices? What's the best mix of fixed vs. floating debt now? How do we diversify revenue to cushion downturns? Can we add liquidity buffers to handle a credit crunch?
Funding Optimization	Financing, treasury, or strategic choices that improve funding cost, stability, or flexibility.	Should we issue longer-term bonds at today's rates? Do we refinance now or wait for better spreads? How can we improve our interest coverage ratio? Is a revolving credit facility better than short-term loans?
Compliance and Reporting Integrity	Ensuring regulatory, accounting, and disclosure accuracy to maintain transparency and trust.	Are our revenue disclosures aligned with IFRS 15? Do we need to restate last year's tax provision? How do we ensure audit trails meet SOX standards? What steps prevent misstatement of fair values?
Other	Economic outcomes not clearly aligned with the main pathways.	–
Informational / Educational Only	Purely explanatory or conceptual content with no direct economic consequence.	What's the difference between NPV and IRR? How does leverage amplify returns? What is Basel III capital adequacy? How do rating agencies assess liquidity risk?

CONVERSATION	RUBRICS																																																																								
<p>USER</p> <p>If we encounter a 150 bps sovereign spread shock and 10% increase in repo overcollateralization, how to maintain Liquidity Coverage Ratio above the regulatory minimum without forced asset sales?</p> <p>ASSISTANT</p> <p>Of course. This is an excellent and highly relevant question that gets to the heart of modern treasury and liquidity risk management. Avoiding forced asset sales during a stress scenario is critical to prevent realizing losses and entering a destructive feedback loop.</p> <p>Here is a comprehensive framework for maintaining the LCR above the regulatory minimum under the given shock without resorting to forced asset sales.</p> <p>### Understanding the Shock</p> <p>First, let's break down what the shock implies:</p> <p>1. **150 bps Sovereign Spread Shock:** This means the credit spreads on sovereign bonds (and by extension, other high-quality liquid assets - HQLA) have widened by 1.5%. This has two primary effects:</p> <ul style="list-style-type: none"> * **Market Value Decline:** The market value of your existing HQLA portfolio falls. This is a *capital* issue, not directly a *liquidity* issue, but it erodes the buffer. * **Higher Haircuts:** For assets pledged as collateral (e.g., in repo transactions), counterparties will demand higher haircuts to account for the increased risk and price volatility. This reduces the amount of cash you can raise from a given pool of collateral. <p>2. **10% Increase in Repo Overcollateralization:** This is a direct outcome of the spread shock. If you were required to post 102% collateral for a repo loan, you might now need to post 112%. This effectively reduces the efficiency of your collateral and...</p> <p>USER</p> <p>Now working backward from the point where our liquidity ratio would fail. What early warning triggers would you set? And what exact steps would you have taken, like changing to higher quality collateral, using pre arranged secured funding, or adding internal fees to keep it steady during payment deadlines?</p>	<table border="1"> <thead> <tr> <th>Criterion</th> <th>Axis</th> <th>Importance</th> <th>Weight</th> </tr> </thead> <tbody> <tr> <td>The response correctly defines the liquidity ratio by stating the formula: liquidity ratio = high-quality liquid assets / 30-day stressed net cash outflows.</td> <td>Financial Accuracy</td> <td>Critically Important</td> <td>#10</td> </tr> <tr> <td>The response explicitly defines a single point of failure to be used everywhere (e.g. liquidity coverage ratio falls below 100% within 5 days under stress, cash forecast turns negative on day 4).</td> <td>Instruction Following</td> <td>Critically Important</td> <td>#10</td> </tr> <tr> <td>The response provides at least 3 numerical early warning triggers (e.g., liquidity ratio within 10 percentage points of min requirement; intraday line use > 70%; Level 1 share (top tier assets) < 75%; pledged high quality liquid assets (HQLA) > 30%; payment backlog > 30%; expected collateral calls > 75th percentile).</td> <td>Practical Utility</td> <td>Critically Important</td> <td>#10</td> </tr> <tr> <td>The response states all 4 of the following steps to take following the early warnings: payment timing and netting; same-day secured borrowing; collateral upgrade; portfolio rotation/asset sales.</td> <td>Practical Utility</td> <td>Critically Important</td> <td>#10</td> </tr> <tr> <td>The response states at least 2 of the following payment timing tactics: advance small routine payments; hold non-urgent outflows; net offsetting flows; use payment-vs-payment FX settlement; re-queue by closing times; align inflows before settlement deadlines.</td> <td>Practical Utility</td> <td>Important</td> <td>#7</td> </tr> <tr> <td>The response explicitly states that you should use prearranged secured borrowing (e.g. committed repo, secured facilities).</td> <td>Practical Utility</td> <td>Important</td> <td>#7</td> </tr> <tr> <td>The response states the correct order of use for pre-arranged secured borrowing: draw same-day/intraday first; use term secured borrowing (overnight/1 week) only if still needed.</td> <td>Practical Utility</td> <td>Important</td> <td>#7</td> </tr> <tr> <td>The response explains the concept of a collateral upgrade (e.g., swapping into top-tier government bonds/cash, exchanging corporate bonds for short-term treasury bills).</td> <td>Practical Utility</td> <td>Important</td> <td>#7</td> </tr> <tr> <td>The response explicitly states whether each step raises the available liquid assets (numerator) or reduces the 30-day net outflows (denominator).</td> <td>Process Transparency & Auditability</td> <td>Important</td> <td>#7</td> </tr> <tr> <td>The response addresses cutoff times by mentioning at least 2 of the following: central bank payment system closing time; FX settlement deadline; clearing house margin deadline; custodian delivery deadline.</td> <td>Risk & Regulatory Disclosure</td> <td>Important</td> <td>#6</td> </tr> <tr> <td>The response states at least 5 of the following key assumptions: size of secured lines; number of counterparties; mix of liquid assets; where assets are held; currencies needed; eligibility rules; limit on already pledged assets.</td> <td>Handling Uncertainty</td> <td>Important</td> <td>#7</td> </tr> <tr> <td>The response states a timeframe for how each step changes the liquidity coverage ratio (e.g., execute a collateral swap within 24 hours, deploy repo funding within 72 hours).</td> <td>Practical Utility</td> <td>Important</td> <td>#5</td> </tr> <tr> <td>The response assigns duties of each step to one of these roles: treasury/liquidity desk; collateral team; risk team; operations; approver.</td> <td>Practical Utility</td> <td>Important</td> <td>#7</td> </tr> <tr> <td>The response explicitly states that warning triggers are chosen by working back from the point of failure (e.g., elevated LCR in excess of 100%, available repo capacity dropping to sub 70%).</td> <td>Process Transparency & Auditability</td> <td>Important</td> <td>#9</td> </tr> <tr> <td>The response specifies at least 3 of the following live metrics to monitor: projected end-of-day LCR; intraday credit use; payment queue status; expected margin/variation calls; free top-tier assets (Level 1) by currency; encumbrance ratio; issuer/venue concentration.</td> <td>Practical Utility</td> <td>Important</td> <td>#8</td> </tr> <tr> <td>The response gives at least a backup option if secured lines shrink (e.g. add counterparties; shorten tenors; use a central bank facility; pledge other eligible assets; use same-day FX swaps).</td> <td>Practical Utility</td> <td>Important</td> <td>#4</td> </tr> <tr> <td>The response states that new trades that worsen today's liquidity face extra charges automatically (e.g., due to trade penalties, liquidity surcharges).</td> <td>Practical Utility</td> <td>Slightly Important</td> <td>#3</td> </tr> </tbody> </table>	Criterion	Axis	Importance	Weight	The response correctly defines the liquidity ratio by stating the formula: liquidity ratio = high-quality liquid assets / 30-day stressed net cash outflows.	Financial Accuracy	Critically Important	#10	The response explicitly defines a single point of failure to be used everywhere (e.g. liquidity coverage ratio falls below 100% within 5 days under stress, cash forecast turns negative on day 4).	Instruction Following	Critically Important	#10	The response provides at least 3 numerical early warning triggers (e.g., liquidity ratio within 10 percentage points of min requirement; intraday line use > 70%; Level 1 share (top tier assets) < 75%; pledged high quality liquid assets (HQLA) > 30%; payment backlog > 30%; expected collateral calls > 75th percentile).	Practical Utility	Critically Important	#10	The response states all 4 of the following steps to take following the early warnings: payment timing and netting; same-day secured borrowing; collateral upgrade; portfolio rotation/asset sales.	Practical Utility	Critically Important	#10	The response states at least 2 of the following payment timing tactics: advance small routine payments; hold non-urgent outflows; net offsetting flows; use payment-vs-payment FX settlement; re-queue by closing times; align inflows before settlement deadlines.	Practical Utility	Important	#7	The response explicitly states that you should use prearranged secured borrowing (e.g. committed repo, secured facilities).	Practical Utility	Important	#7	The response states the correct order of use for pre-arranged secured borrowing: draw same-day/intraday first; use term secured borrowing (overnight/1 week) only if still needed.	Practical Utility	Important	#7	The response explains the concept of a collateral upgrade (e.g., swapping into top-tier government bonds/cash, exchanging corporate bonds for short-term treasury bills).	Practical Utility	Important	#7	The response explicitly states whether each step raises the available liquid assets (numerator) or reduces the 30-day net outflows (denominator).	Process Transparency & Auditability	Important	#7	The response addresses cutoff times by mentioning at least 2 of the following: central bank payment system closing time; FX settlement deadline; clearing house margin deadline; custodian delivery deadline.	Risk & Regulatory Disclosure	Important	#6	The response states at least 5 of the following key assumptions: size of secured lines; number of counterparties; mix of liquid assets; where assets are held; currencies needed; eligibility rules; limit on already pledged assets.	Handling Uncertainty	Important	#7	The response states a timeframe for how each step changes the liquidity coverage ratio (e.g., execute a collateral swap within 24 hours, deploy repo funding within 72 hours).	Practical Utility	Important	#5	The response assigns duties of each step to one of these roles: treasury/liquidity desk; collateral team; risk team; operations; approver.	Practical Utility	Important	#7	The response explicitly states that warning triggers are chosen by working back from the point of failure (e.g., elevated LCR in excess of 100%, available repo capacity dropping to sub 70%).	Process Transparency & Auditability	Important	#9	The response specifies at least 3 of the following live metrics to monitor: projected end-of-day LCR; intraday credit use; payment queue status; expected margin/variation calls; free top-tier assets (Level 1) by currency; encumbrance ratio; issuer/venue concentration.	Practical Utility	Important	#8	The response gives at least a backup option if secured lines shrink (e.g. add counterparties; shorten tenors; use a central bank facility; pledge other eligible assets; use same-day FX swaps).	Practical Utility	Important	#4	The response states that new trades that worsen today's liquidity face extra charges automatically (e.g., due to trade penalties, liquidity surcharges).	Practical Utility	Slightly Important	#3
Criterion	Axis	Importance	Weight																																																																						
The response correctly defines the liquidity ratio by stating the formula: liquidity ratio = high-quality liquid assets / 30-day stressed net cash outflows.	Financial Accuracy	Critically Important	#10																																																																						
The response explicitly defines a single point of failure to be used everywhere (e.g. liquidity coverage ratio falls below 100% within 5 days under stress, cash forecast turns negative on day 4).	Instruction Following	Critically Important	#10																																																																						
The response provides at least 3 numerical early warning triggers (e.g., liquidity ratio within 10 percentage points of min requirement; intraday line use > 70%; Level 1 share (top tier assets) < 75%; pledged high quality liquid assets (HQLA) > 30%; payment backlog > 30%; expected collateral calls > 75th percentile).	Practical Utility	Critically Important	#10																																																																						
The response states all 4 of the following steps to take following the early warnings: payment timing and netting; same-day secured borrowing; collateral upgrade; portfolio rotation/asset sales.	Practical Utility	Critically Important	#10																																																																						
The response states at least 2 of the following payment timing tactics: advance small routine payments; hold non-urgent outflows; net offsetting flows; use payment-vs-payment FX settlement; re-queue by closing times; align inflows before settlement deadlines.	Practical Utility	Important	#7																																																																						
The response explicitly states that you should use prearranged secured borrowing (e.g. committed repo, secured facilities).	Practical Utility	Important	#7																																																																						
The response states the correct order of use for pre-arranged secured borrowing: draw same-day/intraday first; use term secured borrowing (overnight/1 week) only if still needed.	Practical Utility	Important	#7																																																																						
The response explains the concept of a collateral upgrade (e.g., swapping into top-tier government bonds/cash, exchanging corporate bonds for short-term treasury bills).	Practical Utility	Important	#7																																																																						
The response explicitly states whether each step raises the available liquid assets (numerator) or reduces the 30-day net outflows (denominator).	Process Transparency & Auditability	Important	#7																																																																						
The response addresses cutoff times by mentioning at least 2 of the following: central bank payment system closing time; FX settlement deadline; clearing house margin deadline; custodian delivery deadline.	Risk & Regulatory Disclosure	Important	#6																																																																						
The response states at least 5 of the following key assumptions: size of secured lines; number of counterparties; mix of liquid assets; where assets are held; currencies needed; eligibility rules; limit on already pledged assets.	Handling Uncertainty	Important	#7																																																																						
The response states a timeframe for how each step changes the liquidity coverage ratio (e.g., execute a collateral swap within 24 hours, deploy repo funding within 72 hours).	Practical Utility	Important	#5																																																																						
The response assigns duties of each step to one of these roles: treasury/liquidity desk; collateral team; risk team; operations; approver.	Practical Utility	Important	#7																																																																						
The response explicitly states that warning triggers are chosen by working back from the point of failure (e.g., elevated LCR in excess of 100%, available repo capacity dropping to sub 70%).	Process Transparency & Auditability	Important	#9																																																																						
The response specifies at least 3 of the following live metrics to monitor: projected end-of-day LCR; intraday credit use; payment queue status; expected margin/variation calls; free top-tier assets (Level 1) by currency; encumbrance ratio; issuer/venue concentration.	Practical Utility	Important	#8																																																																						
The response gives at least a backup option if secured lines shrink (e.g. add counterparties; shorten tenors; use a central bank facility; pledge other eligible assets; use same-day FX swaps).	Practical Utility	Important	#4																																																																						
The response states that new trades that worsen today's liquidity face extra charges automatically (e.g., due to trade penalties, liquidity surcharges).	Practical Utility	Slightly Important	#3																																																																						

Figure 13: Sample task for the Finance field.

		RUBRICS		
	Criterion	Axis	Importance	Weight
<p style="text-align: center;">PROMPT</p> <p>In New Hampshire, a police officer searched my client's house without a warrant. My client used his smartphone to record the police officer's behavior. While my client was standing in the driveway of his home and video-recording the officer, the police officer arrested my client claiming he did not have permission to record the officer. My client did not tell the officer that he was being recorded. My client was indicted for felony wiretapping because the video contains an audio component. What is the best strategy to defend against this claim before a trial?</p>	The response cites Glik v. Cunniffe, 655 F.3d 78 (2011), as the prevailing law in the First Circuit on the constitutionality of wiretapping charges.	Legal Accuracy	Critically Important	#10
	The response states that people have a constitutional right to record police officers performing their duties in public places subject to reasonable time, place, and manner restrictions.	Legal Accuracy	Critically Important	#10
	The response infers that a driveway is often a location that is publicly visible and where anyone could see or overhear a conversation.	Application of Law to the Facts	Important	#7
	The response argues that police officers do not have a reasonable expectation of privacy while performing their public duties in a public place, such as the client's driveway.	Application of Law to the Facts	Critically Important	#8
	The response recommends dismissing the indictment because it violates the First Amendment speech protections of the U.S. Constitution as applied under the Fourteenth Amendment.	Application of Law to the Facts	Critically Important	#10
	The response recommends dismissing the indictment because it violates the free speech protections of the New Hampshire Constitution.	Application of Law to the Facts	Critically Important	#8
	The response notes that the New Hampshire wiretapping statute does not match the Massachusetts wiretapping statute, such as the scope of the statute and required mental states.	Legal Accuracy	Important	#5
	The response recognizes that the New Hampshire Supreme Court has held that the intent element of a wiretapping charge under RSA 570-A is a knowing intent to break the wiretapping statute. State v. Mueller.	Legal Accuracy	Critically Important	#8
	The response argues that an indictment for wiretapping should be dismissed for insufficiency if it does not plead a knowing mental state.	Procedural Correctness	Critically Important	#9
	The response states that any civil rights damages will likely increase if the case proceeds towards trial.	Supplemental Insight	Slightly Important	#3
	The response states that the police officer's search of the client's house was potentially illegal without a warrant.	Application of Law to the Facts	Important	#6
	The response mentions pretrial motions or defenses such as suppression of evidence, right to a speedy trial, statutes of limitations defenses.	Practical Utility	Critically Detrimental	#8
	The response mentions the distinction between one-party or two-party consent states.	Legal Accuracy	Important	#6
	The response mentions State v. Gerring.	Legal Accuracy	Critically Detrimental	#8
	The response states that the lawyer may wish to present his arguments to the prosecutor before filing a motion to dismiss.	Practical Utility	Slightly Important	#3
	The response states that under the New Hampshire wiretapping law, to protect an oral communication there must be a reasonable belief that it would not be intercepted.	Legal Accuracy	Critically Important	#8
	The response states that in the user's case, the officer had no reasonable belief that the communication would not be intercepted in the public driveway.	Application of Law to the Facts	Critically Important	#8
	The response includes that in the user's case, the officer had no reasonable belief that the communication would not be intercepted even when searching the user's home (since the user has the right to privacy in their own home, but the officer clearly would not have).	Application of Law to the Facts	Critically Important	#8
	The response includes that New Hampshire is a two-party consent state, meaning both parties have to consent to a communication being recorded.	Legal Accuracy	Important	#7
	The response states that consent to recording of a communication is only required when there is a reasonable expectation of privacy.	Legal Accuracy	Important	#7
	The response includes that the officer did not have any reasonable expectation of privacy in their communications either in the home or in the driveway for purposes of arguing two-party consent.	Application of Law to the Facts	Critically Important	#8
	The response states that an additional defense the user could use is that the intent requirement of the wiretapping statute was not met, since the user's intent was to record the illegal search (not to violate the officer's privacy).	Application of Law to the Facts	Critically Important	#8
	The response states that under the fact pattern provided by the user, the criminal wiretapping claim will likely be dismissed.	Application of Law to the Facts	Critically Important	#8

Figure 14: Sample task for the Legal field.

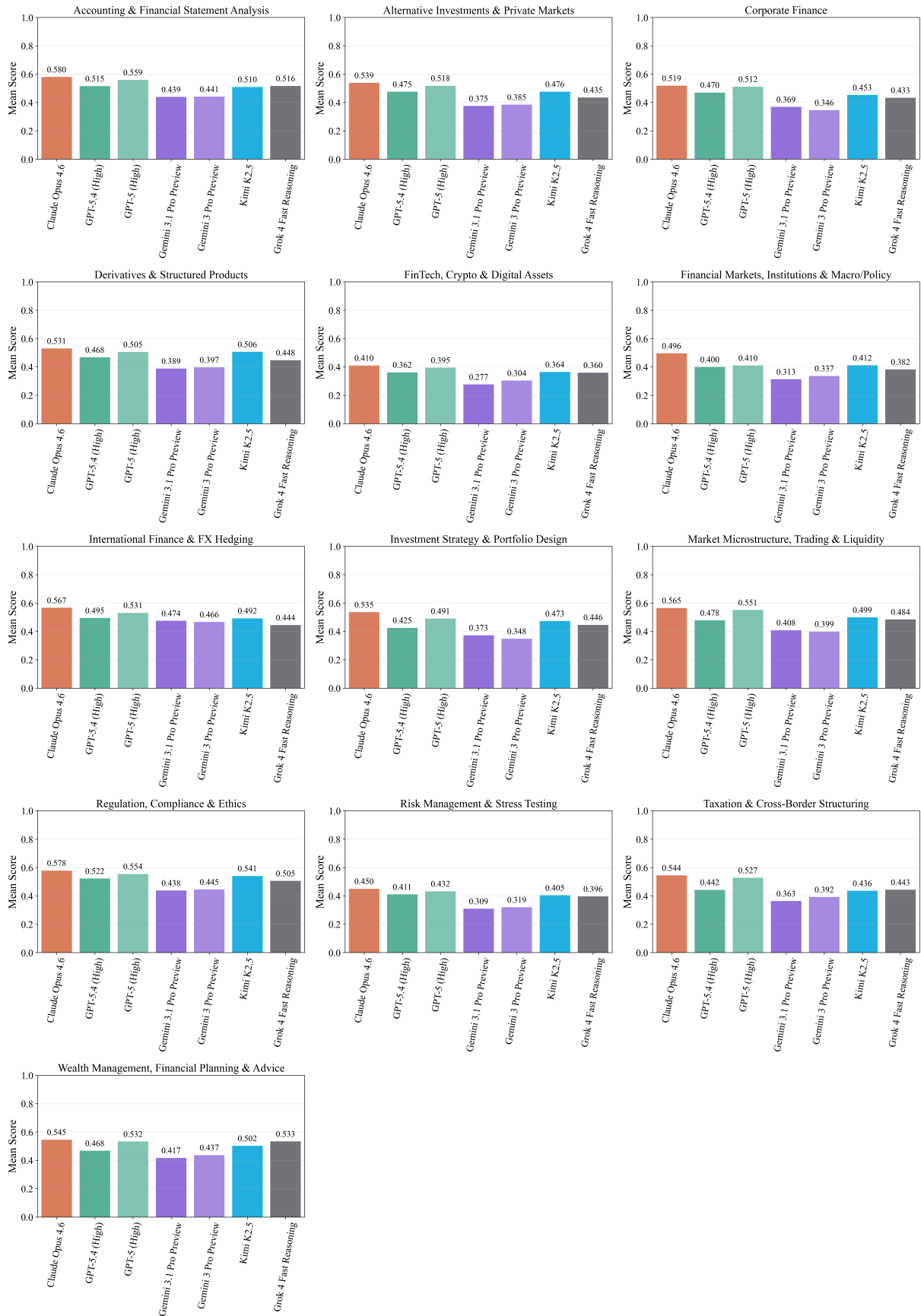


Figure 15: Results across Finance topics.

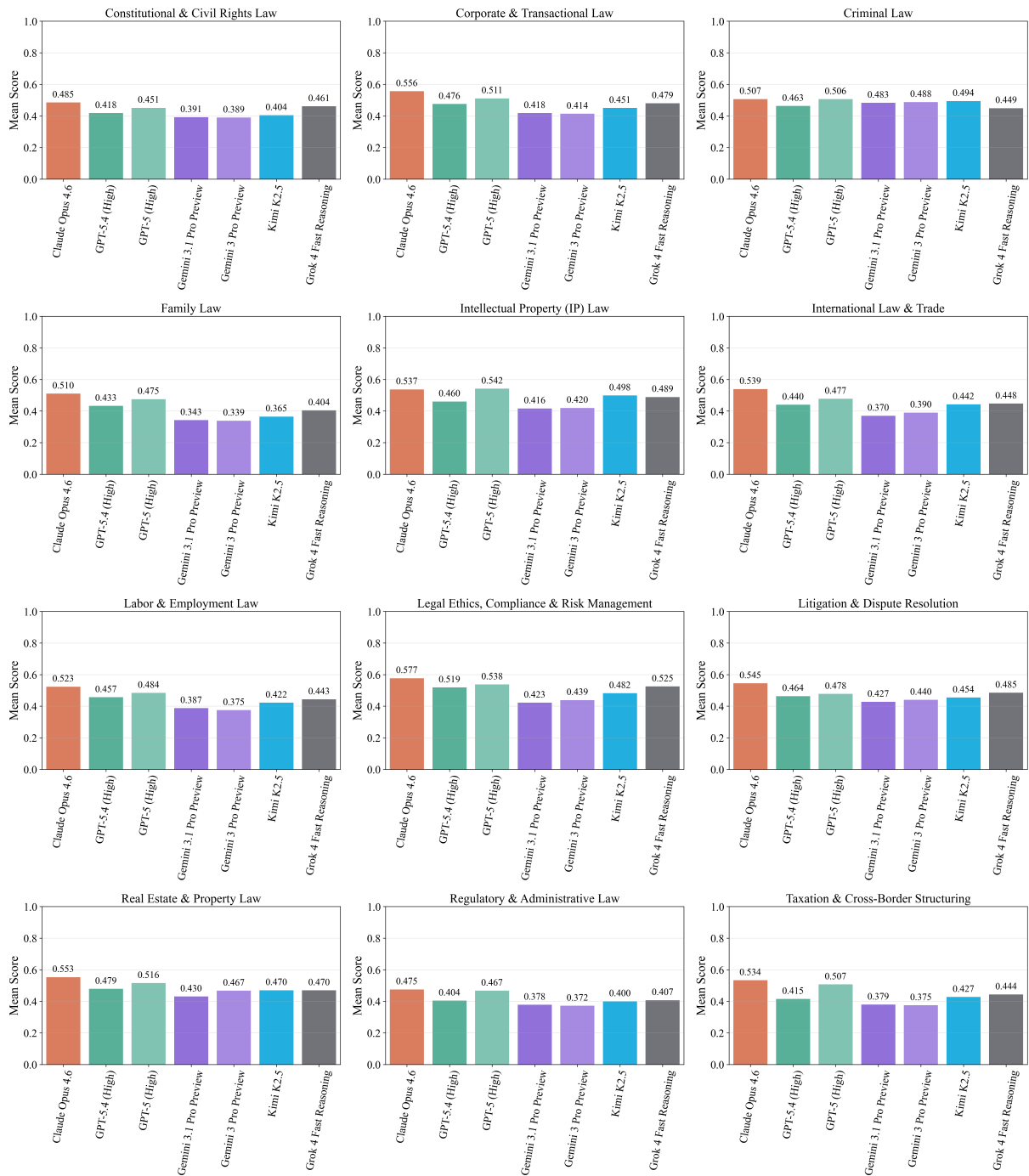


Figure 16: Results across Law topics.

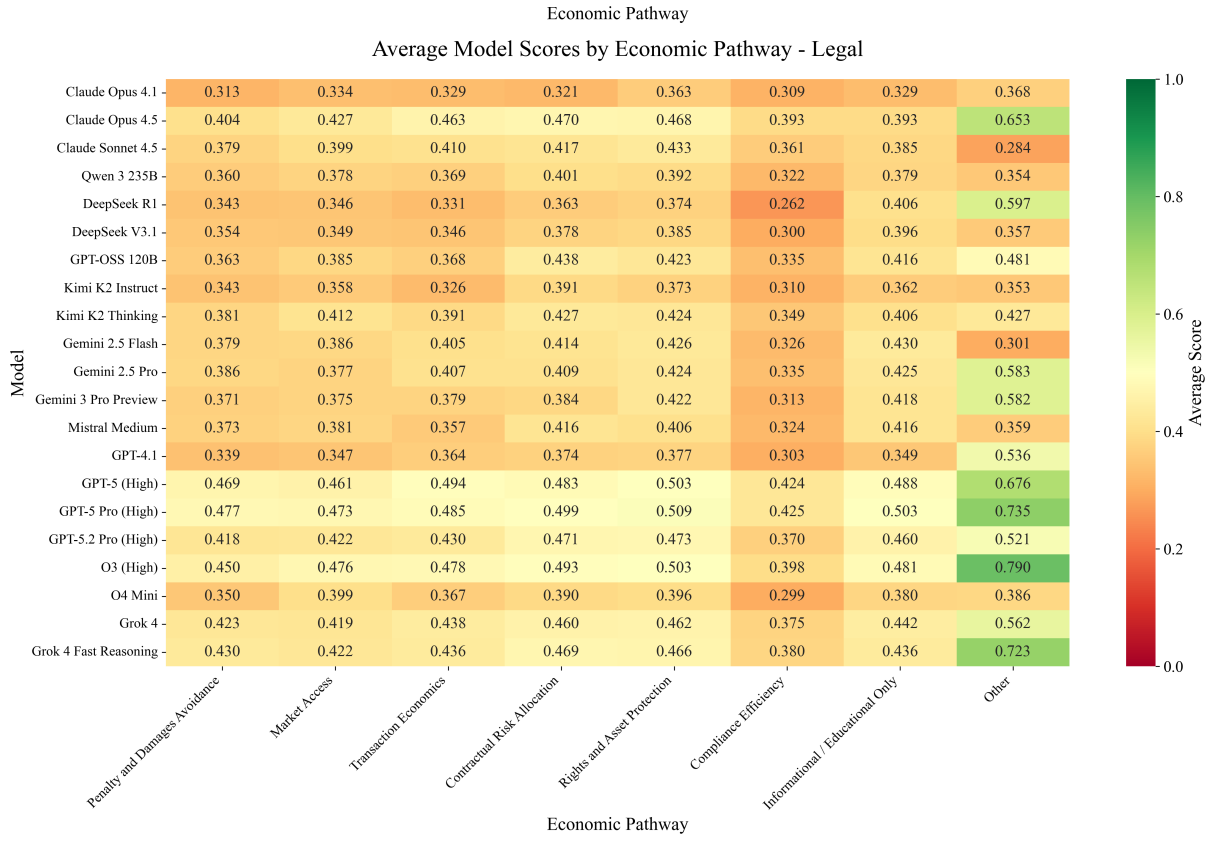
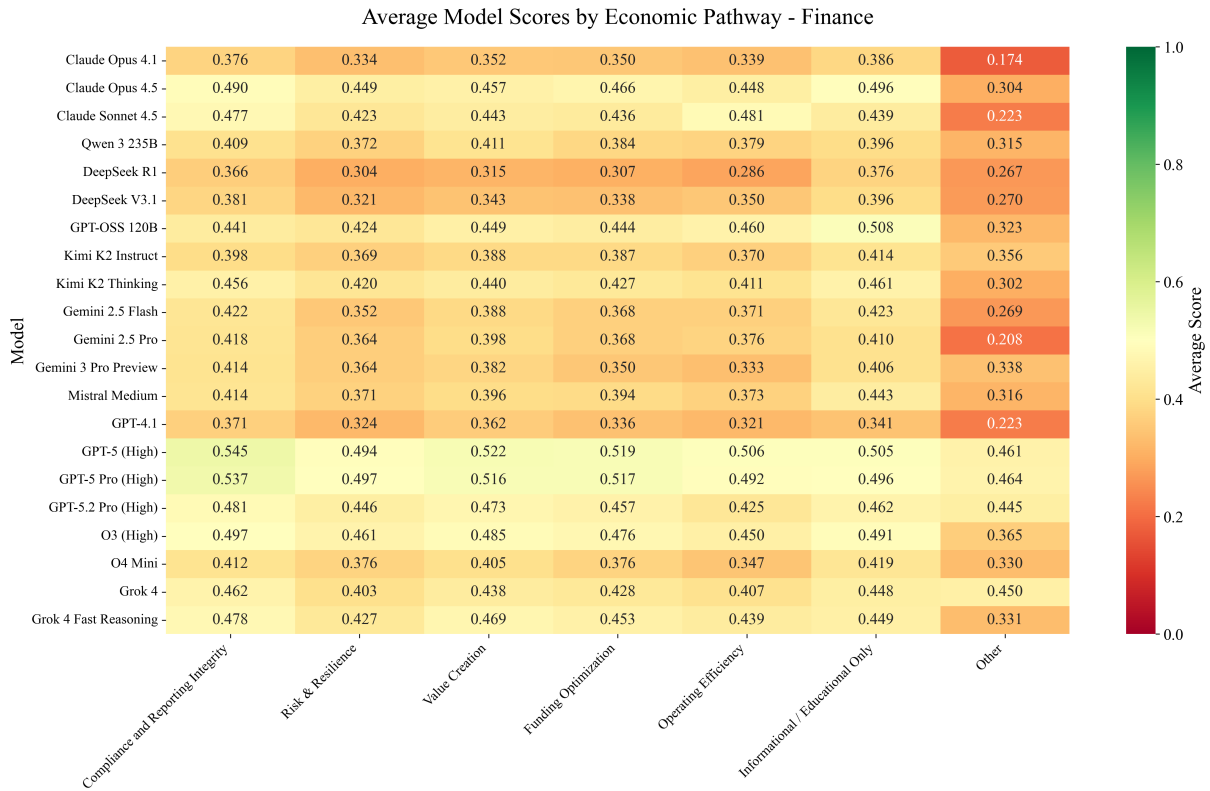


Figure 17: Results across economic pathways, defined as the mechanisms through which economic impact is realized, for all evaluated models.

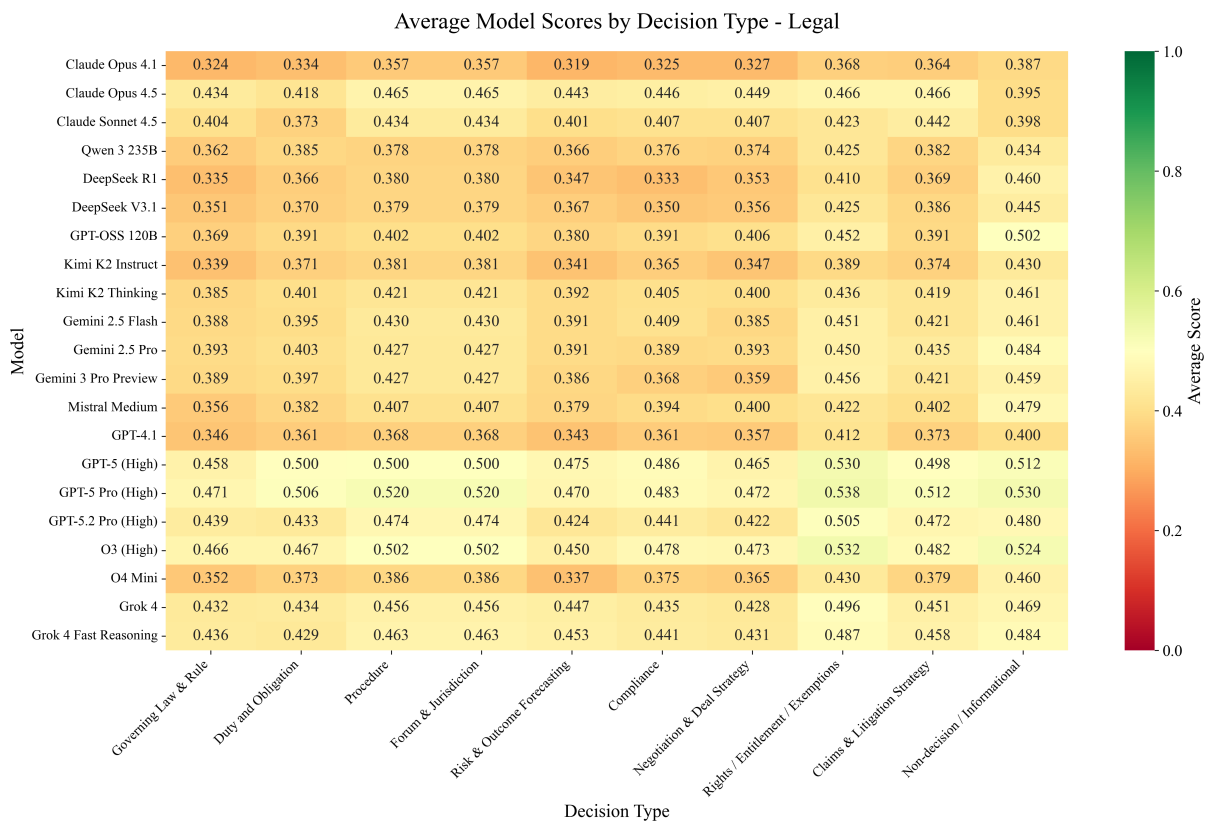
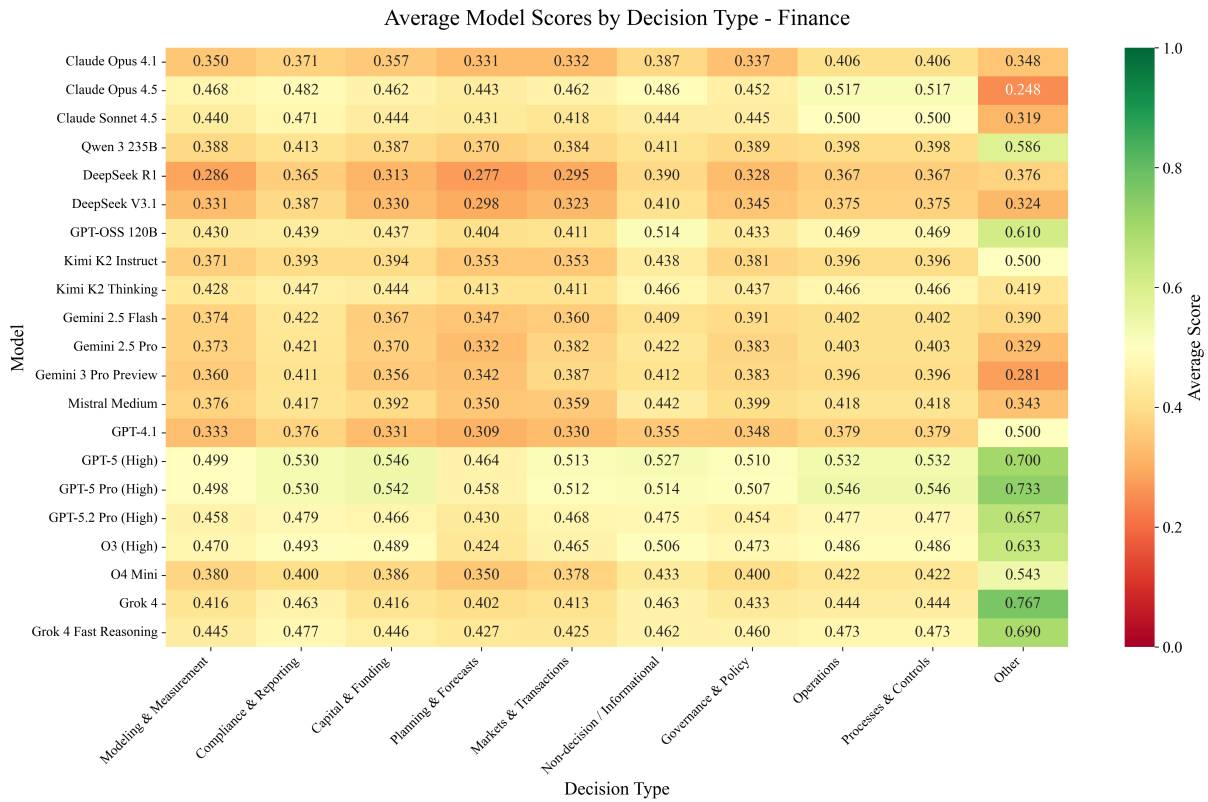


Figure 18: Results across decision types for all evaluated models.