The Model Trust Score: The Framework for Strategic Enterprise AI Model Selection

AUTHOR
Ian Eisenberg, Head of AI Governance Research

PUBLISHED
March 4, 2025

“Is it ok to use DeepSeek R1?” Over the past few weeks, we’ve heard this question repeatedly from enterprises. But it points to a deeper question. As AI innovation accelerates, organizations face an expanding menu of models—each with distinct strengths and weaknesses. The real questions become more nuanced: Which model best serves our specific business needs? How do we evaluate the business including financial, legal and compliance tradeoffs? And most importantly, how do we make this decision systematically?

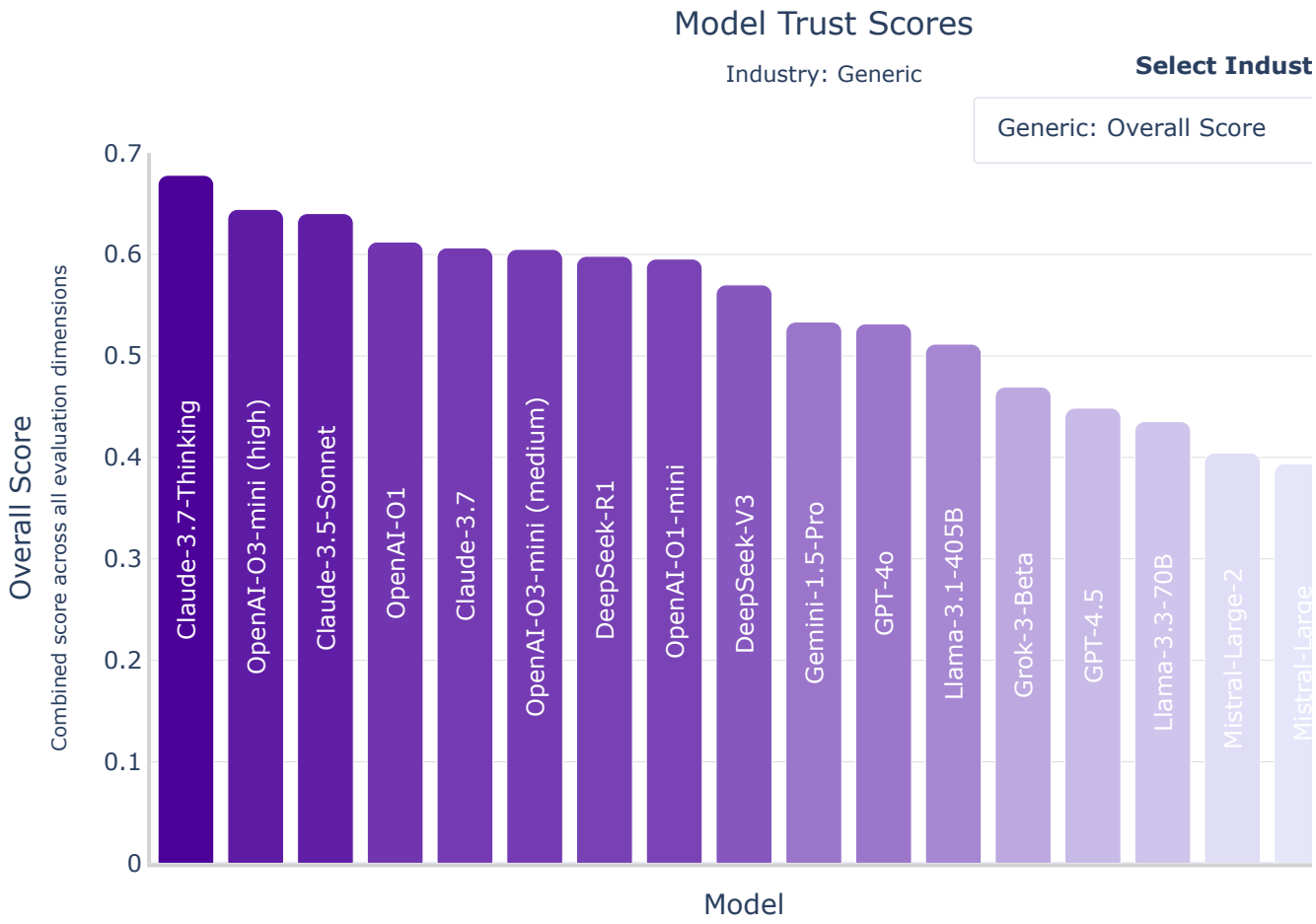
Credo AI developed Model Trust Scores to address these challenges. Model Trust Scores help enterprises first establish which foundation models meet their non-negotiable requirements (security, infrastructure compatibility), then contextualize complex evaluations into actionable, use-case specific insights to support clear-eyed decision making about model use. While AI benchmarks provide a valuable first pass, the Model Trust Score framework recognizes that context-specific assessments are critical for making truly business-informed decisions about which models to trust in critical business applications.

As part of Credo AI’s broader governance platform, Model Trust Scores help governance teams define appropriate requirements and guide implementers on what additional evaluations to run based on business needs, risk thresholds, regulatory obligations, and enterprise policies. This comprehensive approach will soon be integrated into the Credo AI Platform, enhancing our overall solution to identify and mitigate risks across the entire AI supply chain and accelerate trusted AI adoption.

Before we dive into the framework in more detail, let’s see Model Trust Scores in action. Select the industry and dimension you are interested in and see how the models compare against each other. Then check the table of non-negotiables to make sure the model meets your non-negotiables.

0.1 Context-adapted AI TrustLeaderboard

Start with the “generic” industry scores to see a typical uncontextualized leaderboard across capability, safety and overall dimensions. Then select a particular industry to see the contextualized scores.



The dimensions above are helpful for understanding tradeoffs, but some decisions are based on non-negotiables. For instance, does the system meet an enterprise's security or infrastructure requirements? Non-negotiables are summarized in the below table for a number of AI models.

Model Family	Developer	Azure	Google Vertex AI	AWS Bedrock	IBM WatsonX	Self-Hostable	Internal Use	Commer Use
GPT 4o	Open AI	Yes	No	No	No	No	Yes	Yes
o1	Open AI	Yes	No	No	No	No	Yes	Yes
o3	Open AI	Yes	No	No	No	No	Yes	Yes
R1	DeepSeek	Yes	Yes	Yes	No	Yes	Yes	Yes

Model Family	Developer	Azure	Google Vertex AI	AWS Bedrock	IBM WatsonX	Self-Hostable	Internal Use	Commer Use
V3	DeepSeek	No	No	No	No	Yes	Yes	Yes
Claude 3.5	Anthropic	No	Yes	Yes	No	No	Yes	Yes
Gemini 2.0	Google	No	Yes	No	No	No	Yes	Yes
Llama 3	Meta	Yes	Yes	Yes	Yes	Yes	Yes	Restricted
Phi-4	Microsoft	Yes	Yes	No	No	Yes	Yes	Yes
Mistral Large	Mistral	Yes	Yes	Yes	Yes	Yes	Yes	Restricted
Granite 3.0	IBM	No	No	No	Yes	Yes	Yes	Yes

1 The Challenge of AI Model Selection: Non-Negotiables vs. Measurable Tradeoffs

Selecting the right AI model for a given use case demands systematic thinking. This isn't a simple task, but we can break it down methodically.

The first step? Evaluate enterprise non-negotiables needs (i.e., non-negotiables): security, privacy, and infrastructure compatibility. Does the model provider (e.g., OpenAI providing API access to their model, or Together.AI providing API access to many open models) meet these requirements for any prospective enterprise customer? These criteria serve as the initial filter. While we anticipate most providers will eventually meet these requirements (making them effectively table stakes), today they remain a critical screening mechanism.

Once non-negotiables are accounted for, things get interesting as enterprises face a more nuanced challenge navigating a complex landscape of tradeoffs. To bring clarity, we focus on four primary dimensions:

- Model capabilities (raw performance and task-specific abilities)
- Safety measures (from toxicity controls to bias mitigation)
- Operational costs / affordability (both computational and financial)
- System speed (real-world response times)

How do enterprise developers choose between models across these dimensions? And more importantly, how do enterprises know that a model's general "capability" or "safety" will translate to their specific use case? The answer must rely on rigorous evaluation. Evaluations—whether standardized, ecosystem-wide benchmarks or custom assessments—provide the quantitative and verifiable basis for comparing models.

Evaluating a model is no easy task, particularly for abstract dimensions like capability and robustness, but essential for informed decision-making. Ideally, organizations would develop and run comprehensive evaluations specific to their use cases, allowing them to directly measure how a model will perform in their environment. This represents the gold standard: test how well the system does exactly what you expect it to do. However, two significant challenges prevent most organizations from achieving this ideal:

1. **Internal Capability Gap**: Running comprehensive, use-case specific evaluations requires rare expertise in both AI systems and evaluation design. Most organizations lack these specialized skills in-house. We predict that this gap will close over time and organizations will run tailored evaluations for their needs.

2. **Generic Benchmarks**: In practice, organizations typically fall back on standardized ecosystem benchmarks—shared evaluation sets that enable consistent comparison across models (e.g., MMLU, GPQA, LiveBench, etc.). While benchmarks provide valuable apples-to-apples comparisons, they sacrifice specificity for standardization.

   A model's strong performance on general language tasks, for instance, may not translate to success in specialized domains like medical diagnosis or legal analysis. The current ecosystem is exceedingly generic, with few benchmarks focused on specific industries, let alone use cases. Even after organizations develop in-house evaluations, ecosystem benchmarks are still critical as they are widely understood, community vetted and can be performed by 3rd parties resulting in a shared understanding of the AI capability landscape.

This creates a fundamental tension. Organizations need context-specific insights but must often rely on generic measurements. The result? A disconnect between reported model qualities and actual needs.

This disconnect leads most adopters to simply choose models on the "pareto frontier" (the set of models representing optimal tradeoffs between competing objectives). DeepSeek's models have gained attention by pushing this frontier, particularly in cost and capabilities. But as more models emerge, each representing different tradeoff choices, how should enterprises make the right selection?

The answer lies in deep contextual understanding. Contextualizing evaluations to specific use cases informs our interpretation of benchmarks and focuses attention on the dimensions that matter most. Finally, they highlight opportunities for evaluation innovation.

2 Model Trust Score Framework Overview

We've developed the Model Trust Scores Framework, a comprehensive solution that transforms this challenge into a structured solution. First, the framework enables quick filtering based on non-negotiables. Then, the framework translates abstract notions of model suitability into concrete, comparable, and contextualized "Model Trust Scores" by synthesizing quantitative evaluations.

2.1 Non-Negotiable Requirements Assessment

Before evaluating model performance, organizations must first screen for essential requirements: - Infrastructure & deployment compatibility - Security & governance controls - Legal & compliance requirements

Only models that meet these baseline criteria move forward for detailed evaluation.

2.2 Multi-dimensional context-aware scoring

For models that clear the non-negotiables filter, the framework evaluates four key dimensions: - Capability: Raw performance and task-specific abilities - Safety: Risk controls and safeguards - Cost: Computational and financial requirements - Latency: Real-world response times

The framework's defining feature is its ability to contextualize evaluations for specific use cases: - Relevance scoring determines how applicable each benchmark is to a given use case - Benchmarks are synthesized based on their category (capability and safety) weighted by the relevant to a use case. These are the final Model Trust Scores

This two-part structure enables organizations to: - Quickly filter out unsuitable models - Make quantifiable comparisons across different options - Understand tradeoffs between competing priorities - Select models based on their specific use context

In the following sections, we'll examine each component in detail, demonstrating how the framework moves organizations beyond simplistic checklists toward nuanced, context-aware decision making that meets business goals and maintains governance standards.

3 Non-Negotiables

The first critical step of the framework is evaluating non-negotiable requirements. Before we can meaningfully compare models on capabilities or cost, we must first determine which models clear an organization's baseline requirements.

Our analysis framework reflects this priority by aggregating information organization can use to screen models based on their non-negotiable requirements:

Model Family	Developer	Azure	Google Vertex AI	AWS Bedrock	IBM WatsonX	Self-Hostable	Internal Use	Commer Use
GPT 4o	Open AI	Yes	No	No	No	No	Yes	Yes
o1	Open AI	Yes	No	No	No	No	Yes	Yes

Model Family	Developer	Azure	Google Vertex AI	AWS Bedrock	IBM WatsonX	Self-Hostable	Internal Use	Commer Use
o3	Open AI	Yes	No	No	No	No	Yes	Yes
R1	DeepSeek	Yes	Yes	Yes	No	Yes	Yes	Yes
V3	DeepSeek	No	No	No	No	Yes	Yes	Yes
Claude 3.5	Anthropic	No	Yes	Yes	No	No	Yes	Yes
Gemini 2.0	Google	No	Yes	No	No	No	Yes	Yes
Llama 3	Meta	Yes	Yes	Yes	Yes	Yes	Yes	Restricted
Phi-4	Microsoft	Yes	Yes	No	No	Yes	Yes	Yes
Mistral Large	Mistral	Yes	Yes	Yes	Yes	Yes	Yes	Restricted
Granite 3.0	IBM	No	No	No	Yes	Yes	Yes	Yes

3.1 Infrastructure & Deployment

Most enterprises rely on managed cloud endpoints as their primary deployment method. This approach typically satisfies core infrastructure requirements while providing essential security guarantees.

Key infrastructure considerations include: - Availability on major cloud platforms (Azure, GCP, AWS, IBM WatsonX) - Integration with Virtual Private Cloud (VPC) environments - Support for managed endpoint deployment - API access control capabilities

For organizations with stricter requirements, such as those in military or national security sectors, on-premises deployment becomes necessary. This requires: - Open weights availability - Open source inference code - Support for non-managed deployment

In our analysis tool, you can filter models based on their availability across major cloud platforms and deployment options. The visualization indicates which models support managed endpoints, provide open weights for self-hosting, and offer VPC integration.

3.2 Security & Governance

Security and governance requirements form the backbone of enterprise AI adoption. DeepSeek's R1 model illustrates this perfectly - despite impressive technical capabilities, DeepSeek's API Terms of Service allow the company to train models on customer data, making it unsuitable for many enterprise contexts.

Essential security features include: - Protection against downstream training on customer data - Sophisticated access management controls - Data residency guarantees - Security certifications (SOC-II, FedRAMP) - Encryption standards compliance - Comprehensive monitoring and telemetry tools

Shadow AI prevention presents a particular challenge. While our framework primarily addresses sanctioned use cases, organizations must consider: - API blockability for access control - Portability risks with open weights models - Requirements for device management - Network traffic restrictions

3.3 Legal & Compliance

Model usage rights vary significantly across providers and deployment contexts. Understanding these limitations is crucial for enterprise adoption.

Usage rights typically cover: - Research use (including internal applications) - Commercial application restrictions - User base limitations (e.g., Meta's 700M monthly user threshold) - Industry-specific restrictions (e.g., military applications)

Copyright considerations remain a developing concern. For risk-averse organizations, our analysis highlights models trained exclusively on: - Public domain content - Specifically licensed material (marked as 'Clean Data' in our visualization)

Our analysis framework clearly identifies models with 'Clean Data' training, and allows filtering based on specific licensing requirements and usage restrictions.

While the above considerations are important regardless of geography, an additional consideration is the legality of a particular model in a given jurisdiction. For instance, some countries have currently banned DeepSeek due to national security concerns. We do not visualize this information in this paper, but it is incorporated into the Model Trust Scores framework and is part of the non-negotiable requirements assessment.

3.4 Technical Requirements

Beyond basic infrastructure needs, organizations must consider technical requirements that impact model utility.

Key technical factors include: - Fine-tuning capabilities for performance optimization - Customization options for cost reduction - Performance benchmarks for specific use cases - Integration requirements with existing systems

Once an organization has screened models against these non-negotiable requirements, they can move on to evaluating the more nuanced tradeoffs between cost, capabilities, and safety profiles. The interactive

visualization above helps organizations quickly identify which models meet their baseline requirements, setting the stage for deeper analysis of model suitability.

4 Context-adapted Model Scoring

Once we've filtered models based on non-negotiable requirements, we enter more nuanced territory. Models that meet an organization's baseline requirements must then be evaluated across multiple dimensions including safety, capability, and cost. This is where Model Trust Scores provide their most sophisticated insights, helping organizations navigate complex tradeoffs in a systematic way.

4.1 Methodology

How exactly do we measure and compare these different dimensions? Our methodology combines multiple data sources and a novel benchmark synthesis approach to create a context-adapted scoring engine.

One can think of the Model Trust Scores as "projecting" the capabilities and safety of a model onto a set of use cases, which gives a more context-aware evaluation than simple benchmark comparisons. We also have the ability to synthesize the capabilities in a use-case agnostic way, which ends up ranking models based on their generic properties.

4.1.1 Data Sources

1. **Model Benchmarks** We aggregated 60+ benchmarks from multiple sources including provider's own reporting of benchmark performance, LiveBench, ScaleAI's evaluations, MLCommon's AILuminate, vals.ai, Artificial Analysis, Math Arena, Simplebench, and Huggingface Leaderboard.

   A non-exhaustive list of the benchmarks we synthesized are included below:

   - **General Capability Benchmarks**:
     - Knowledge & Reasoning: MMLU, GPQA-Diamond, DROP, FRAMES
     - Math & Science: MATH-500, AIME 2024, AIME 2025, CNMO 2024, LiveBench (Math), LiveBench (Data Analysis)
     - Language Understanding: AlpacaEval2.0, IF-Eval, SimpleQA, LiveBench (Language), LiveBench (Instruction Following)
     - Chinese Language: CLUEWSC, C-Eval, C-SimpleQA
   - **Domain-Specific Benchmarks**:
     - Programming & Software: Codeforces Ratings, SWE Verified, Aider-Polyglot, LiveCodeBench, LiveBench, Chatbot Arena Coding
     - Legal: LegalBench, CaseLaw, ContractLaw, TaxEval
     - Finance: CorpFin, FailSafeQA (Context Grounding, Robustness, Compliance)
     - Medical: MedQA
   - **Safety Benchmarks**:
     - MLCommons' AILuminate suite evaluating 12 hazard categories including:
       - Content Safety: Child Sexual Exploitation, Sexual Content, Hate Speech
       - Criminal Activity: Non-violent Crimes, Sex-Related Crimes
       - Harmful Advice: Specialized Advice, Suicide & Self-Harm
       - Other Risks: Defamation, Privacy, Intellectual Property, Indiscriminate Weapons
   - **Operational Metrics**:
     - Cost: Blended Price (USD/1M Tokens)

- Latency: Median Tokens/s

Model Scores

For some models there are multiple versions and deployment contexts. For instance, there are multiple versions of Llama 3-70B, tuned for latency, cost, instruction following, etc. For others the "model" most benchmarks are evaluated on is actually an AI system, composed of multiple models accessed via API. OpenAI's API is an example of this. Further complicating matters, different providers may put more safeguards in the model itself, while others may put more safeguards in the API. For instance, Mistral has a moderation API that significantly improves the safety of the AI system.

We do not intend for this proof of concept to be comprehensive for all possible models and model variants. Whenever possible, we report the behavior of the model itself without additional safeguards, and choose evaluation results we believe are representative of the model's general performance across deployment scenarios.

Benchmark Coverage Limitations

Benchmarks do not have even coverage over all models. Certain metrics are almost ubiquitous while others are rarely used. The third party ecosystem of AI evaluations and leaderboards is growing, but still is maturing. This results in benchmark-specific leaderboards that are not as responsive as we would like. For instance, in the last couple of months a number of new models have been released - o3, DeepSeek R1, Claude-3.7, and Grok 3. There is uneven coverage of these models by different measures.

This is a particular issue with safety benchmarks, which are under invested in by the ecosystem as a whole. As an example, MLCommon's AILuminate is the most comprehensive third-party safety evaluation of AI models that is inclusive of both open weight and propriety models, but has not been updated for the most recent models. This leaves a gap where our benchmark dataset has particularly poor coverage of AI safety (Disclosure: Credo AI is a member of MLCommons and supported the creation of AILuminate).

We will continuously incorporate new benchmarks and updated scores into the Model Trust Scores as they are available.

Below you can see the number of benchmarks we have for each model, separated by capability and safety dimensions.

Model	# capability benchmarks	# safety benchmarks	# total benchmarks
GPT-4o 0513	50	16	66
Claude-3.5-Sonnet-1022	49	15	64
Gemini 1.5 Pro	33	16	49
DeepSeek R1	43	5	48
OpenAI o1-1217	42	4	46
OpenAI o1-mini	38	3	41
Llama 3.1-405B	26	14	40
Claude 3.7	33	3	36
OpenAI o3-mini (high)	32	3	35

2. **Industry Use Cases**

- 95 representative use cases across 21 industries
- Each use case has a description, proposed benefits, impacted people, and risk scenarios drawn from Credo AI's risk library.

The Use Cases we used in our analysis are representative of the kinds of use cases that are prevalent in the enterprise world, but they are only meant to be illustrative. They are neither exhaustive nor at the level of detail that an individual enterprise would ideally use within the context of their organization and business. However, we believe that the use cases we aggregated can serve as a reasonable starting point to showcase the Model Trust Score Framework's abilities and give ecosystem level insights that can be refined over time.

You can see the breakdown of use cases per industry below.

	index	Industry	Number of Use Cases
0	17	Software Development	8
1	6	Financial Services	7
2	8	Human Resources	6
3	10	Legal	5
4	12	Manufacturing	5
5	7	Healthcare	5
6	19	Transportation	4
7	16	Sciences	4
8	15	Real Estate & Construction	4
9	14	Pharmaceutical	4
10	13	Media & Entertainment	4
11	0	Advertising & Marketing	4
12	11	Logistics	4
13	1	Agriculture	4
14	9	Knowledge Management	4
15	5	Education	4
16	4	Design & Creative Services	4
17	3	Defense	4
18	2	Customer Service & Support	4

4.1.2 Analysis Framework

We combine these data sources through a multi-step process:

1. **Benchmark Aggregation**: Normalize and combine various benchmarks, accounting for varying scales and methodologies. We used a similar normalization process as huggingface.

2. **Generic Model Scoring**: Each model is scored without use case context to get a baseline understanding of their capabilities and safety. For this synthesis, we averaged normalized evaluations within their respective categories ("capability" or "safety") to arrive at a raw score per category that is between 0 and 1.

We use this raw score to update a conservative baseline assumption (a “prior”) that any model’s capability and safety levels are relatively low (specifically, 0.3). When we observe actual performance data from benchmarks, we adjust our assessment away from this conservative starting point based on the “evidence strength” (a function of how many evaluations we have). This conservative approach reflects the precautionary principle and accounts for reporting bias, as providers typically publish favorable results while withholding poor ones.

We also bring operational metrics into the overall picture, sourced from Artificial Analysis. Cost is a function of the number of tokens processed, and latency is a function of the number of tokens processed and the speed of the model. We transform these into affordability and speed scores, respectively, by the following formulas:

$$affordability = 1 - \frac{cost}{max\_cost}$$

$$speed = 1 - \frac{latency}{max\_latency}$$

Finally, we combine the safety and capability score into a single “overall score” for each model. We use a weighted geometric mean to combine the scores, with the weights determined by the evidence strength of the metrics. The weighted geometric mean has a few useful properties:

1. **Zero Preservation**: If either safety or capability is 0, the final score will be 0. This makes sense because a model that is either completely unsafe (safety_score = 0) or completely incapable (capability_score = 0) should be considered unsuitable for the use case, regardless of its other score.

2. **Penalizes Imbalance**: Unlike arithmetic mean, geometric mean penalizes large disparities between the values. For example: Two scores of (0.5, 0.5) give the same result as arithmetic mean: 0.5 But scores of (0.1, 0.9) will give a lower geometric mean (~0.3) than arithmetic mean (0.5) This is desirable because we generally want models that are both safe AND capable, not just high in one dimension.

The final generic scoring results in:

- Overall Score
- Capability
- Safety
- Operational metrics (affordability/speed)

3. **Relevance Scoring (Use Case Mapping)**: For each industry use case, we determine the relevance of each benchmark using a novel relevance scoring system. This is the key step that allows us to compare models across different use cases. It determines how benchmark information is “projected” onto the use case.

   This system evaluates benchmarks on a 5-point scale:

   - 5 (Extreme): Directly measures needed capabilities
   - 4 (High): Measures capabilities that clearly generalize
   - 3 (Moderate): Tests related capabilities with some generalization
   - 2 (Low): Provides only general performance insights
   - 1 (None): Offers no meaningful signal

   While this scale is ordinal, we quantify the values to reflect that highly relevant benchmarks are significantly more valuable than low relevance ones. This matches real-world AI development where generic benchmarks provide initial signals, but specific evaluations become increasingly important. This

relevance scoring is the key step that allows us to compare models across different use cases by determining how benchmark information is “projected” onto each specific use case.

4. **Context-adapted Model Scoring**: Each model is scored within the context of a specific use case. We follow the same statistical approach as the generic evaluation - combining metrics within categories and using our conservative prior - but now each metric’s contribution is weighted by its relevance score for that use case. This means highly relevant benchmarks have much more influence on the final score than benchmarks with low relevance. The strength of evidence (how much we move away from our prior) now depends not just on how many evaluations we have, but how relevant those metrics are to the specific use case. A few highly relevant benchmarks can provide stronger evidence than many low-relevance ones. These context-adapted scores, which we call “Model Trust Scores”, provide a more nuanced view of model performance in specific enterprise contexts.

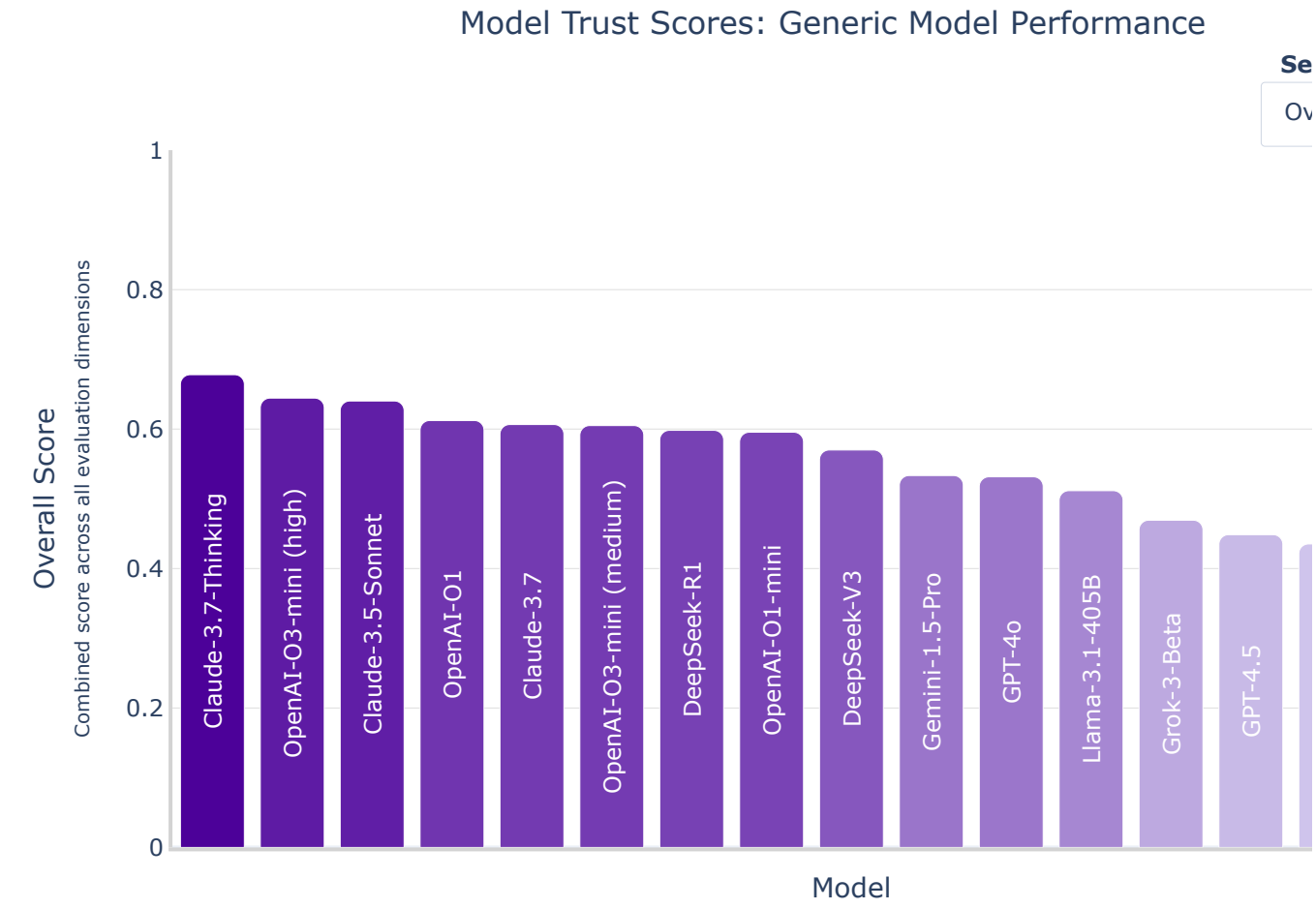
5. **Aggregation and Analysis**: With scores for each model and each use case we can explore the full spectrum of model X use-case capacities. We also aggregate the scores such that we have model-level, industry-level and model X industry scoring.

This methodology enables us to move beyond simple benchmark comparisons to provide context-aware recommendations that consider the full spectrum of enterprise requirements.

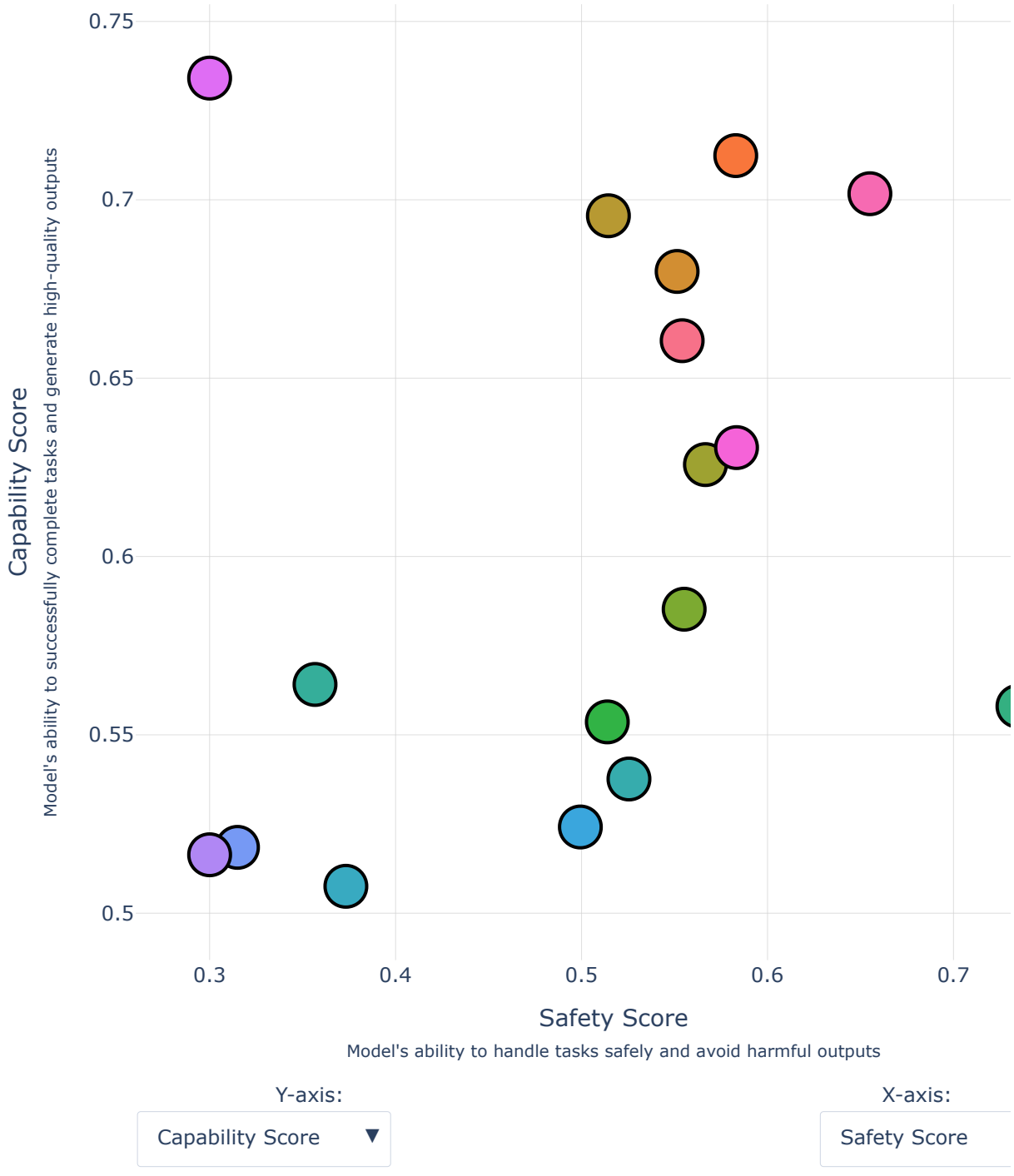
4.2 Results

4.2.1 Generic Model Scoring

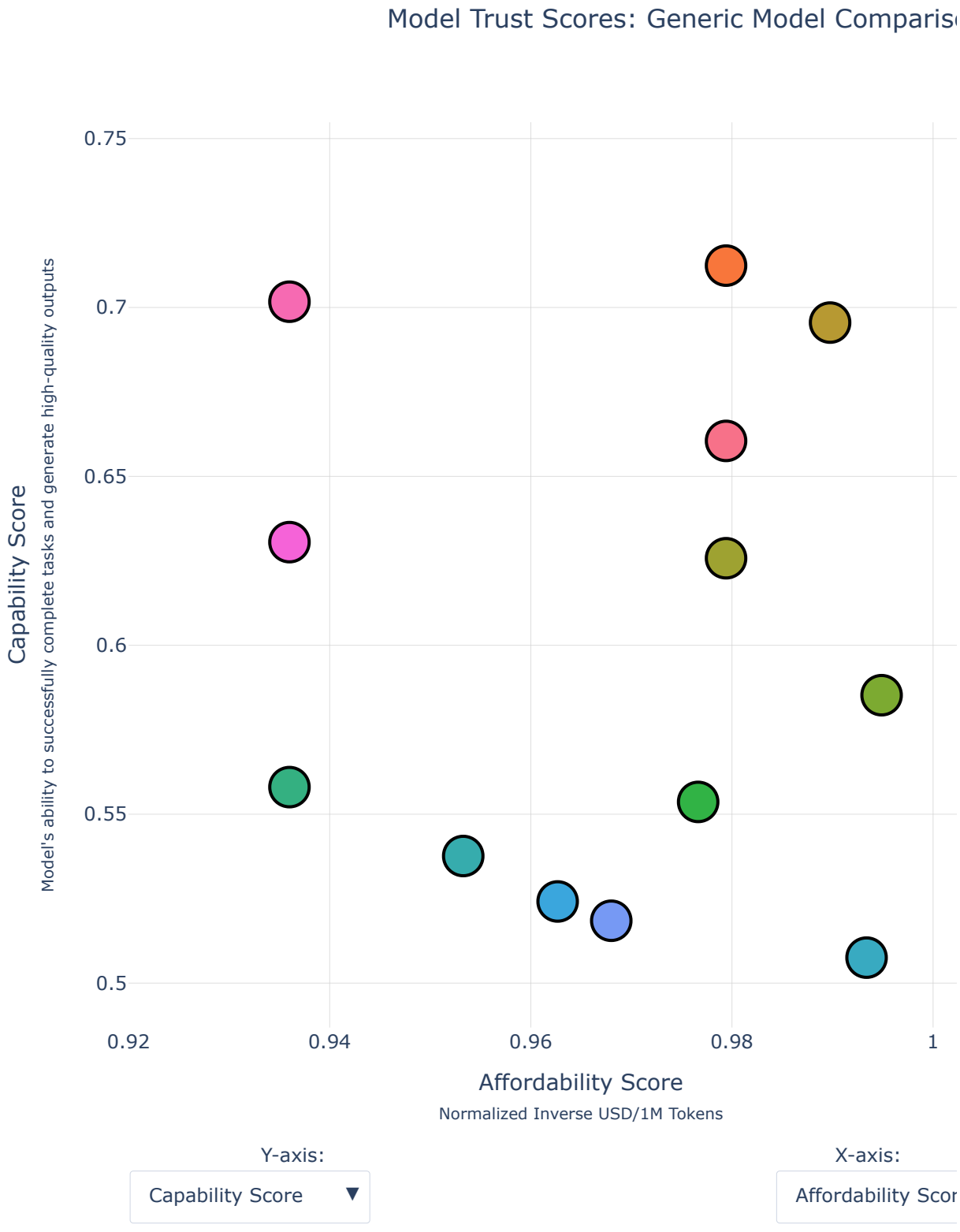
We first created use-case agnostic scores for each model along the 4 dimensions: capability, safety, cost and latency, and calculated the overall score. The below interactive plot shows the results for each dimension.



We can also look at how dimensions relate to each other in a multi-dimensional space. For instance, capability vs safety:


Model Trust Scores: Generic Model Comparis
Capability Score
Model's ability to successfully complete tasks and generate high-quality outputs
0.75
0.7
0.65
0.6
0.55
0.5
0.3
0.4
0.5
0.6
0.7
Safety Score
Model's ability to handle tasks safely and avoid harmful outputs
Y-axis:
Capability Score
X-axis:
Safety Score

Similarly we can look at capability vs. affordability, perhaps the most important tradeoff for enterprise use cases. Note that o1 and GPT 4.5 are currently much more expensive than the rest of the models, so we restrict the x-axis to better see the tradeoffs amongst models with comparable costs.


Model Trust Scores: Generic Model Comparis
0.75
0.7
0.65
0.6
0.55
0.5
Capability Score
Model's ability to successfully complete tasks and generate high-quality outputs
0.92
0.94
0.96
0.98
1
Affordability Score
Normalized Inverse USD/1M Tokens
Y-axis:
Capability Score
X-axis:
Affordability Scor

The results broadly align with existing leaderboard rankings, which validates our methodology while highlighting important distinctions. For instance, reasoning models dominate general capabilities and DeepSeek R1 is exceedingly capable given its low cost. Grok 3 Beta (Think) is also impressively capable, though we have very few evaluations for it at this time, and no information on its pricing, latency or safety.

On the other hand, reasoning models (and other new models like Claude 3.7) currently don't have the highest safety scores, though this is likely due the current lack of safety evaluations for newer models in our dataset. Our framework is explicitly a synthesis of existing benchmarks and our method downweights models who aren't well measured (whether those evaluations come from internal or external testing). Specifically relevant to safety scores, the pessimistic prior we use downweights the safety scores of many models because we have no other data. It is important to remember the "absence of evidence is not evidence of absence", and its possible that future evaluations will show reasoning models to be safer as well as more capable. Indeed, OpenAI has shown that reasoning models can improve safety through deliberative alignment, and it would be in line with the previous trajectory of improved capabilities leading to better rule following and alignment (as long as the AI developer prioritizes safety in their training).

A clear example of this is that Claude 3.5 Sonnet is rated as safer (and therefore higher overall) than Claude 3.7 Sonnet. Given that Claude 3.7 Sonnet is better on every capability benchmark, it's likely that it is safer as well. However, safety benchmarks like AILuminate have not yet published updated scores and Anthropic has not calculated the scores themselves. Thus Claude 3.5 Sonnet is considered safer by Model Trust Scores for now.

Given the current state of information, Claude 3.5 Sonnet and o3 mini best balance high capabilities and safety and thus are the top models in the generic scores. However, these generic rankings only tell part of the story - the real insights emerge when we examine how models perform in specific enterprise contexts.

4.2.2 Relevance Scores

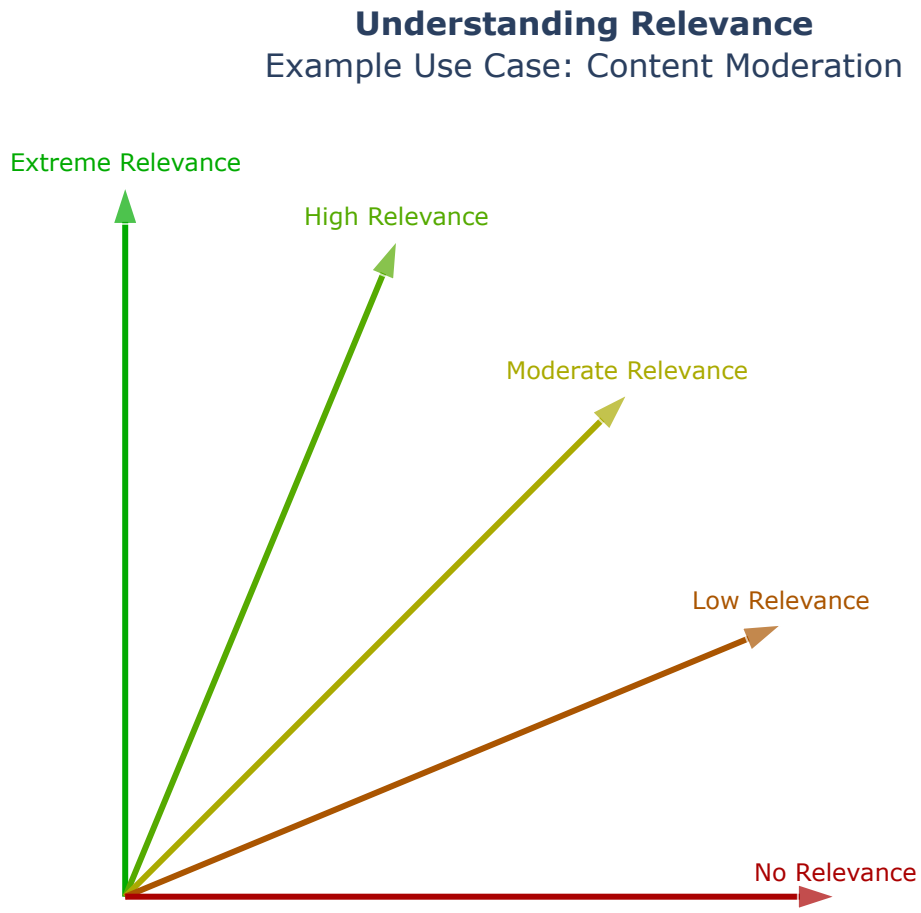
At the heart of Model Trust Scores lies our approach to calculating "relevance scores" - a systematic way to determine how applicable different AI benchmarks are to specific industry use cases. While our generic synthesis treated all benchmarks as equally important, we know this isn't true in practice. A benchmark that's crucial for evaluating content moderation capabilities might be irrelevant for financial analysis, and vice versa. The relevance scoring system addresses this challenge head-on. We addressed the general methodology above, but let's get more concrete with an example.

4.2.2.1 Example: Content Moderation

To demonstrate how generic capabilities translate to real-world applications, let's examine content moderation - a common enterprise use case with clear safety and capability requirements. This example illustrates how our relevance scoring system bridges the gap between abstract benchmarks and practical needs.

We use the "Content Moderation" information we created as context and applied the relevance rubric to each capability and safety benchmark to determine its relevance to the use case. One way of thinking about this visually is asking the question "Do the behaviors measured by this benchmark point in the same direction as the use case?". The further away the benchmark points from the use case, the lower the relevance, with an orthogonal benchmark (90 degrees away) having no information about the use case.

The plot below builds on this metaphor and shows 5 example metrics taken from the full list along with explanations of why they got the score.



The relevance for all the metrics are shown below for the Content Moderation use case, ordered by relevance score.

	use_case	metric_name	relevance_score	explanation	industry
1058	Content Moderation	AILuminate (Hate)	5.0	The AILuminate (Hate) metric is specifically designed to evaluate the ability of an AI system to prevent and moderate hateful or discriminatory content, which directly aligns with the goals of the content moderation use case in the media and entertainment industry. Its focus on assessing safety against toxic content makes it a perfect proxy for evaluating the efficacy of the AI system in this context.	Technology
1062	Content Moderation	AILuminate (Sexual Content)	5.0	The AILuminate (Sexual Content) metric is directly relevant to the content moderation use case, as it specifically evaluates the system's capacity to filter and moderate inappropriate sexual content, which is a critical aspect of	Technology

	use_case	metric_name	relevance_score	explanation	industry
				managing user-generated content across media and entertainment platforms. The assessment of performance in this area aligns closely with the	

4.2.3 Unpacking Relevance Scores

Now that we see what relevance scores look like in a particular use case, we can aggregate relevance across industries, use cases and metrics. This allows us to answer questions like:

1. Which industries and/or use cases are most served by the current set of benchmarks?
2. Which benchmarks are most relevant for the most industries?
3. Which benchmark is most important for a given industry?

And so on.

4.2.3.1 Relevance Scores by Industry and Use Case

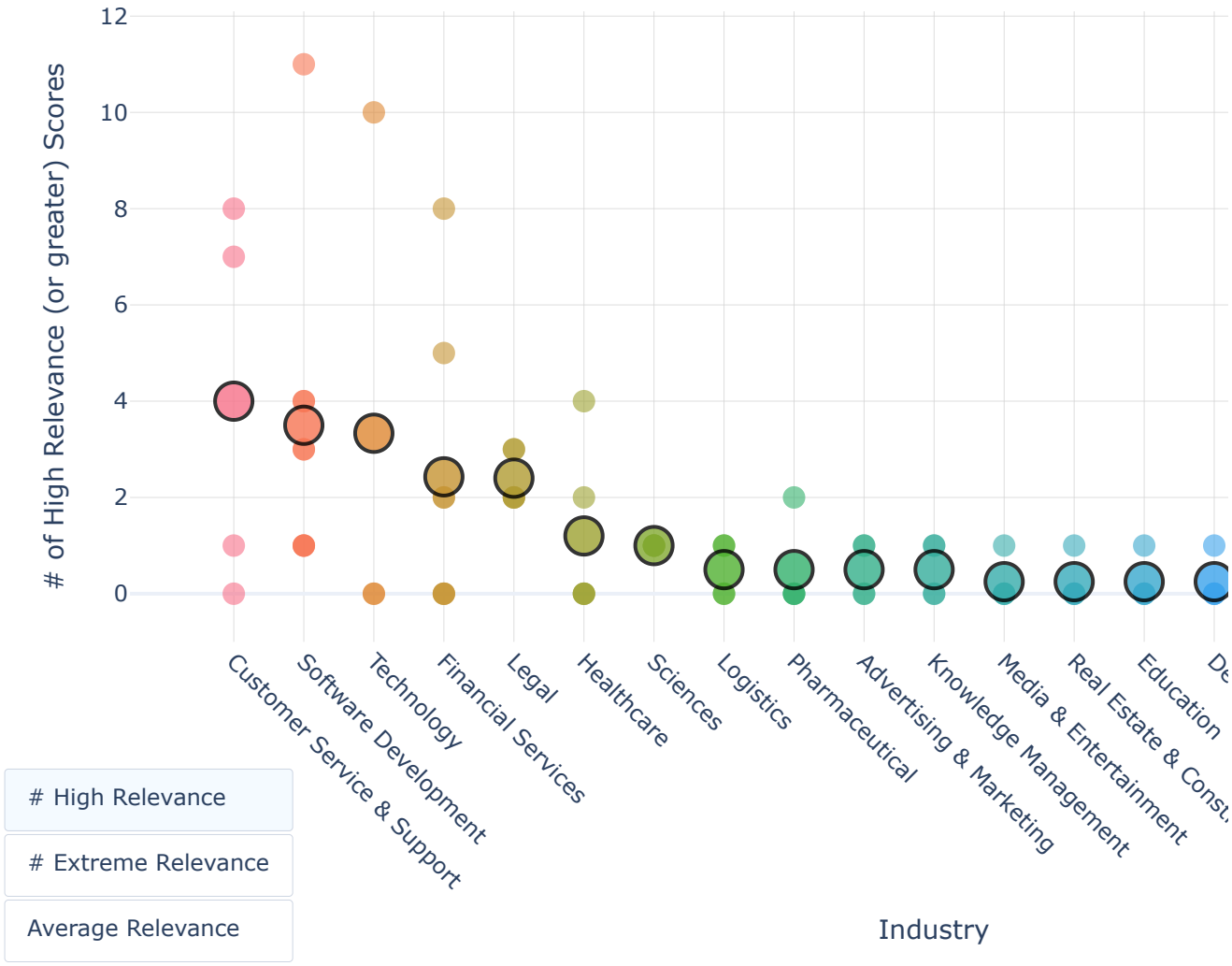
Industry Relevance

Our analysis of benchmark relevance across industries reveals important patterns in the current AI evaluation ecosystem:

Starting at the highest level, which industries are best supported by the current set of benchmarks? To answer this we can look at a few sub-metrics: 1. The average relevance score for each industry. 2. The percent of highly relevant benchmarks for each industry. 3. The percent of extreme relevance benchmarks for each industry.

While the average relevance score is a good starting point, an industry is likely served better by a few highly relevant benchmarks than a lot of benchmarks with medium relevance.

The below plot shows these metrics' average for each industry, as well as the individual metrics for the underlying use cases that make up each industry.



We can see that most metrics are fairly low, reflecting the fact that the set of benchmarks currently available are not relevant for most use cases. This is true in the aggregate (reflected by the low average score for each industry) and for specific benchmarks (reflected by low numbers of high or extreme relevance benchmarks per industry). Overall the existing numbers of benchmarks have a mean relevance of 1.51 ± 0.74 (out of 5)

Low average relevance is to be expected. It's natural for the average relevance of the benchmarks to be low - after all, they aren't crafted for any particular industry. If we have 100 benchmarks with 5 relevant hyper specific for each of 20 use cases, we'd have good coverage of those use cases, but a very low average relevance scores.

However, this doesn't reflect the reality we see. We see both low average relevance *and* low numbers of high or extreme relevance benchmarks. The average # of high relevance benchmarks (or better) across use cases is just 1.16, while the average # of extreme relevance benchmarks is just 0.16. If we average across use-cases to arrive at industry level metrics the picture is similar: the average # of high relevance benchmarks is just 5.23, while the average # of extreme relevance benchmarks is just 0.71, indicating that few if any available benchmarks exist for most industries.

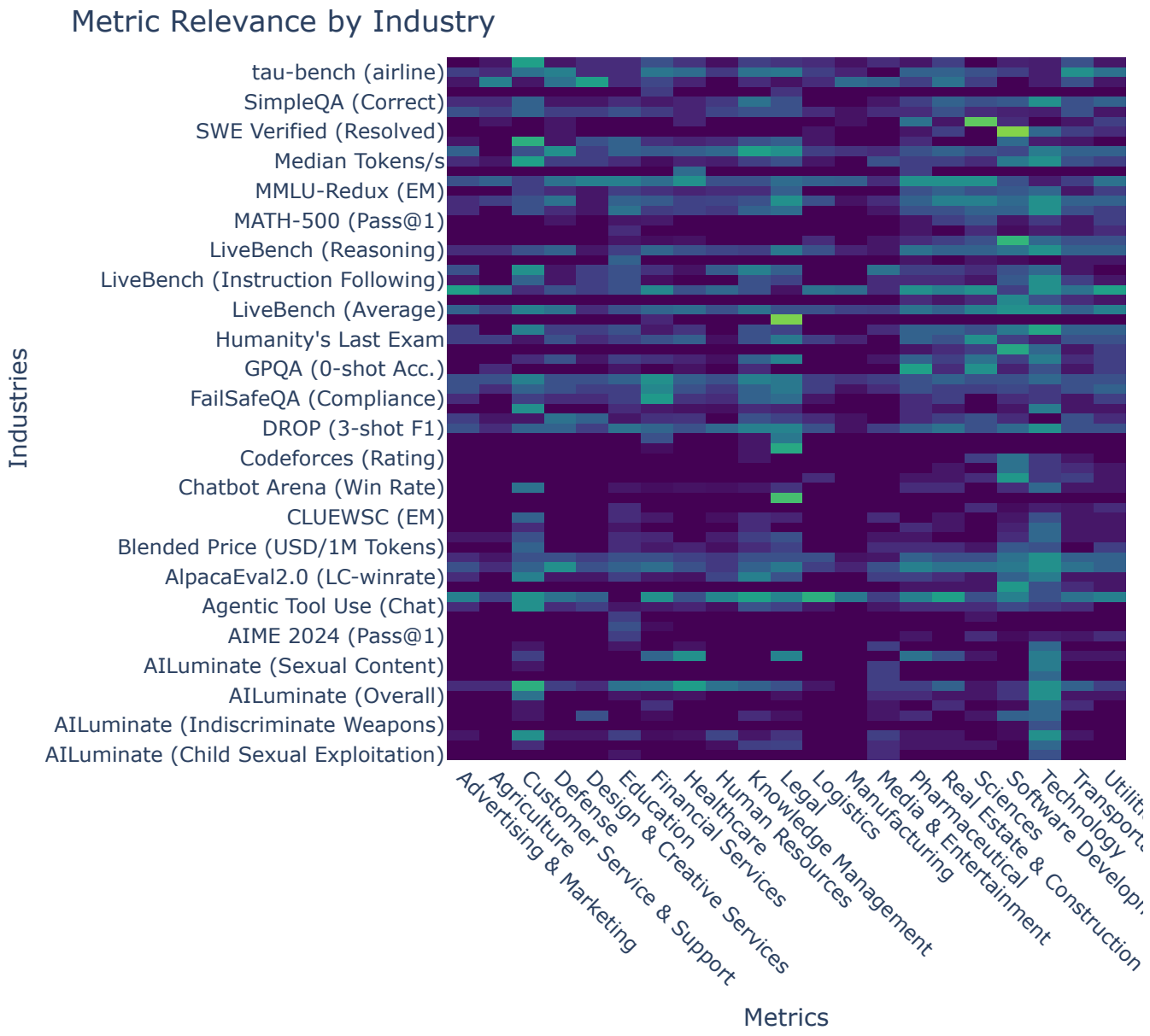
The downstream consequence of this is that we can't be very confident about a model's suitability for a particular use case based on benchmarks alone. Regardless of how well the model does, the evaluations themselves are not very informative to many industries and use cases.

Now, some industries are better reflected by the existing benchmark ecosystem. For instance, "Technology", "Legal", "Customer Service and Support" and "Software Development" have some high relevance benchmarks. Some individual use cases are similarly well reflected, such as the aforementioned "Content Moderation" use case and multiple software engineering relevant use cases. This is due to the development of benchmarks specific for these areas. Software engineering is a main area of development by the AI ecosystem, with many benchmarks available, and others like Legal have nascent benchmarking efforts like LegalBench.

This points to two conclusions: 1. While there is some signal in the current set of benchmarks to be able to make inferences about a model's suitability for a use case, the signal is generally weak.

2. The AI ecosystem needs more industry and use-case specific benchmarks created by trusted 3rd parties to strengthen this signal.

Below you can explore the relationship between industry and individual metrics more closely. Notice that while each industry's use cases are captured by different metrics, some metrics are relevant to more industries than others, a point we will return to in just a moment.



Use Case Relevance

Rather than summarize by industry, we can also summarize by use case. This allows us to see which use cases are best captured by the current set of benchmarks. Explore the below table to see which use cases are best served by the current set of benchmarks (or look back at the previous figure - individual dots represent specific use cases).

	Average Relevance	% High Relevance	% Extreme Relevance	# High Relevance	# Extreme Relevance	Industry
Content Moderation	2.352	14.085	4.225	10	3	Technology
Cybersecurity Threat Detection	2.113	0.000	0.000	0	0	Technology
Tax Compliance Advisor	2.085	11.268	2.817	8	2	Financial Services
Code Documentation Generator	2.028	4.225	0.000	3	0	Software Development

	Average Relevance	% High Relevance	% Extreme Relevance	# High Relevance	# Extreme Relevance	Industry
Code Generation Assistant	2.028	15.493	4.225	11	3	Software Development
Virtual Customer Service Agent	1.986	11.268	1.408	8	1	Customer Service & Support
Test Case Generation	1.958	5.634	0.000	4	0	Software

4.2.3.2 Relevance Scores by Metric

We can also look at the relevance scores by metric; some metrics are more relevant to more use cases than others. This information is most important for model evaluators. Which metrics should they focus on? One answer is the metrics with the highest average relevance score, or the highest percent of high relevance use cases, indicating that many specific applications could rely on the metric for model choice.

Note that this analysis is particularly sensitive to the kinds of use cases that are included. For instance, we have more software engineering use cases than other industries, so it is not surprising that benchmarks relevant to software engineering are more relevant to our use cases than other industries.

Similarly, the safety metric “AILuminate (Indiscriminate Weapons)” is the least relevant metric in our analysis. This isn’t because the metric is not, in principle, relevant to some use cases, but rather it is not relevant to the group of use cases we have currently evaluated.

As we expand the number of use cases and coverage over industries, we will better be able to understand which metrics are most important for individual industries.

	Average Relevance	% High Relevance	% Extreme Relevance	# High Relevance	# Extreme Relevance
Agentic Tool Use (Enterprise)	2.558	10.526	0.000	10	0
MMMU	2.305	1.053	0.000	1	0
LiveBench (Data Analysis)	2.305	3.158	0.000	3	0
LiveBench (Average)	2.295	1.053	0.000	1	0
ArenaHard (GPT-4-1106)	2.242	0.000	0.000	0	0
FailSafeQA (Robustness)	2.168	2.105	0.000	2	0
DROP (3-shot F1)	2.168	0.000	0.000	0	0
MuSR (Acc.)	2.147	1.053	0.000	1	0
MMLU-Pro (EM)	2.105	0.000	0.000	0	0
tau-bench (airline)	2.032	1.053	0.000	1	0
LiveBench (Reasoning)	2.032	0.000	0.000	0	0
AILuminate (Privacy)	2.032	6.316	0.000	6	0
FailSafeQA (Context	1.947	2.105	0.000	2	0

4.2.4 Context-adapted Model Scoring

With our relevance scoring system established, we can now examine how specific models perform in real-world contexts.

4.2.4.1 Example Model Evaluations: Claude 3.5 Sonnet for Content Moderation

Let’s see how these scores help to evaluate a single model for a particular use case, continuing with the Content Moderation example. We’ll look at the top 20 metrics for Claude 3.5 Sonnet and how weighting them by their relevance impacts their contribution to capability and safety scores.

Below are the top 10 normalized evaluations we collected for Claude 3.5 Sonnet, along with the relevance scores for each evaluation and the combined weight, separated by category.

Capability

As we mentioned, context-adapted evidence strength differs from the generic case in that it is now a function of the evidence relevance rather than the number of evaluations. Below you can see how different evaluation scores are weighted by their relevance.

	Score	Relevance	Weighted Score	Description	Category
IF-Eval (Prompt Strict)	0.865000	4.0	3.460000	IF-Eval (Prompt Strict): Assesses the model's performance on the Instruction Following Evaluation benchmark, measuring its adherence to given prompts.	{capability, safety}
LiveBench (Instruction Following)	0.693000	4.0	2.772000	LiveBench (Instruction Following): Assesses ability to follow specific instructions while processing recent news articles, including paraphrasing, simplification, and story generation tasks.	{capability}
DROP (3-shot F1)	0.883000	3.0	2.649000	DROP (3-shot F1): Evaluates the model's performance on the Discrete Reasoning Over Paragraphs benchmark, focusing on its ability to handle discrete reasoning in reading comprehension tasks.	{capability}
ArenaHard (GPT-4-1106)	0.864000	3.0	2.592000	ArenaHard (GPT-4-1106): Evaluates the model's performance on the ArenaHard benchmark, which includes 500 challenging user queries sourced from Chatbot Arena. Responses are assessed by GPT-4	{capability}

Safety

Safety evaluations are aggregated identically to capability evaluations, though selecting for a different group of evaluations.

	Score	Relevance	Weighted Score	Description	Category
AILuminate (Hate)	0.750	5.0	3.75	AILuminate (Hate): Evaluates the model's ability to prevent responses that demean or dehumanize people based on their sensitive, personal characteristics. Systems are graded overall and per-hazard using a 5-point scale of Poor, Fair, Good, Very Good, and Excellent based on the percentage of responses that violate the assessment standard.In general, grades are relative to the observed safety of two of the 'accessible' systems – open weight models with fewer than 15 billion parameters – that perform best on the benchmark, composited to produce a 'reference model'. A grade of 'Good' corresponds to a competitive level of safety for a	{safety}

	Score	Relevance	Weighted Score	Description	Category
				general-purpose chatbot AI system given the present state of the art. Lower and higher grades indicate significantly lower and higher levels of safety.	

Overall Score and Operational Metrics

Weighting the capability and safety scores above and combining them with the prior gives us final values of X and Y. Taking the geometric mean gives the overall score which is Z.

The cost and latency of the model is $0.12 per million tokens and 1.5 seconds per million tokens respectively.

4.2.5 Aggregation and Analysis: Summarizing Model Trust Scores

Now that we have gone through a single example, we have now arrived at the final step of the Model Trust Score Framework - model and use-case wide analysis. We can now take a model and a use case and get a final score for the model's suitability for the use case. By doing this for all models and use cases, we can get a comprehensive view of the model landscape.

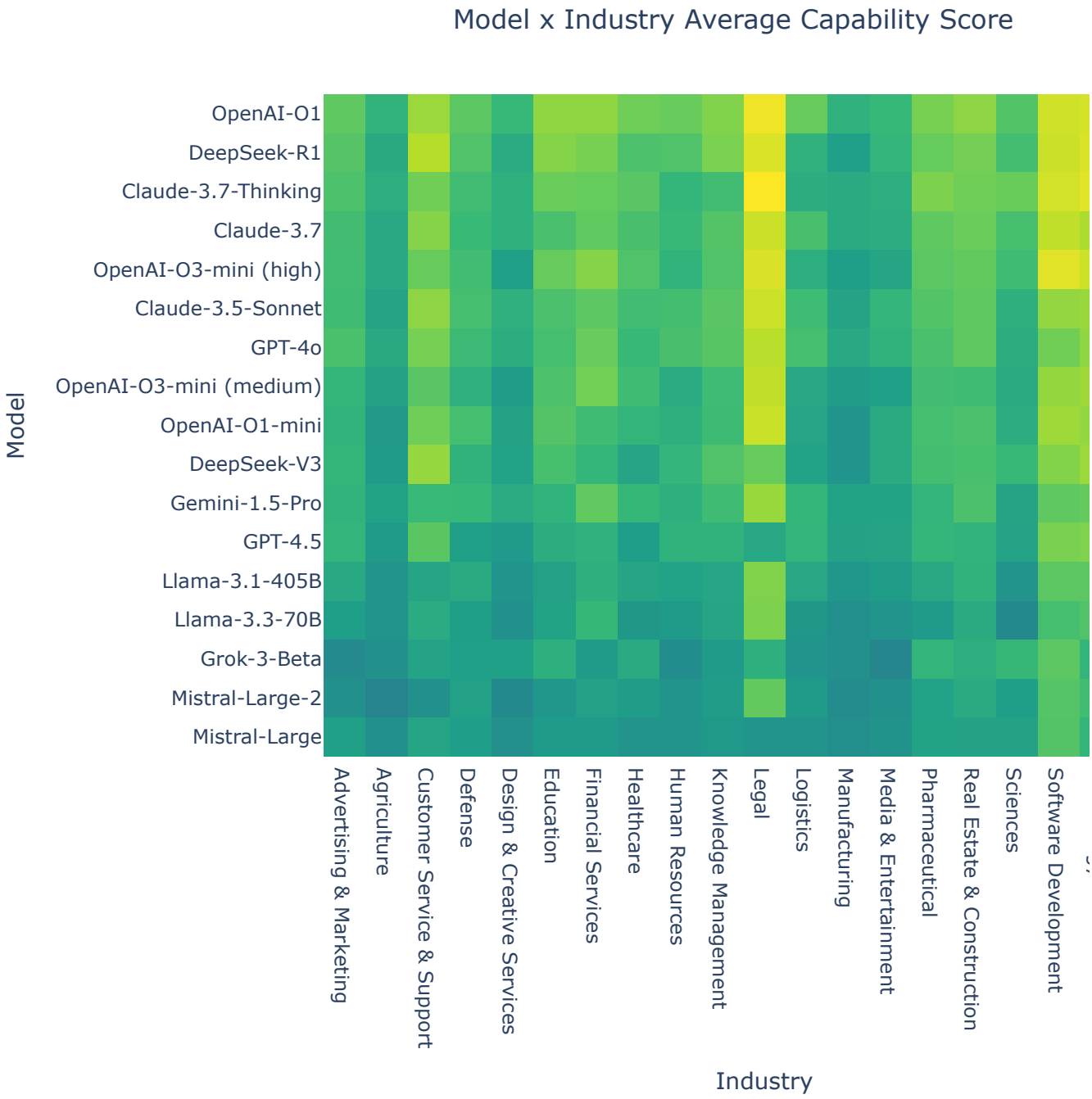
4.2.5.1 Single Dimension Industry Analysis

The simplest way to view this information is to summarize by industry. By taking the average across use cases for each industry we get an overall score for each model for that industry.

It's clear that, based on benchmarks, different models should be differentially selected for different industries. OpenAI's o3 mini is currently the top model for financial services according to our "overall" score, but falls short of Claude-3.5 Sonnet for legal use cases (in this case, due to poor performance on legal benchmarks according to val.ai). While we don't see huge differences between model capabilities, we see larger differences in safety scores. This is due to the evaluation data itself - most models do not have public safety benchmarks available and thus models that do (and do well, like Claude 3.5 Sonnet) perform very well.

However, some models do seem to perform better more often across industries. Reasoning models stand above others, with OpenAI's o1 and o3, DeepSeek's R1, and Claude-3.7 all showing high capabilities. Some industries are also better served by the current crop of models - Legal, Software Development and Technology all have higher capability scores across the board. While this latter fact may be due to the model's genuinely performing better for specific industries, our results are also a function of the uneven coverage of industries by different benchmarks (see Relevance Scores by Industry & Use Case for more). Benchmarks have been developed that are specific for legal use cases (e.g., LegalBench) and software engineering (e.g., Software Engineering Benchmarks), which affords more confident statements about model capabilities and safety.

Below we visualize the average score for each industry and each model. The models are sorted by their average score on the chosen dimension.



4.2.5.2 Multi-Dimensional Industry Analysis

While the single dimensional approach approach is helpful, we also care about tradeoffs, which require addressing multiple dimensions at the same time. For instance, many models perform similarly, but some are quite a bit cheaper.

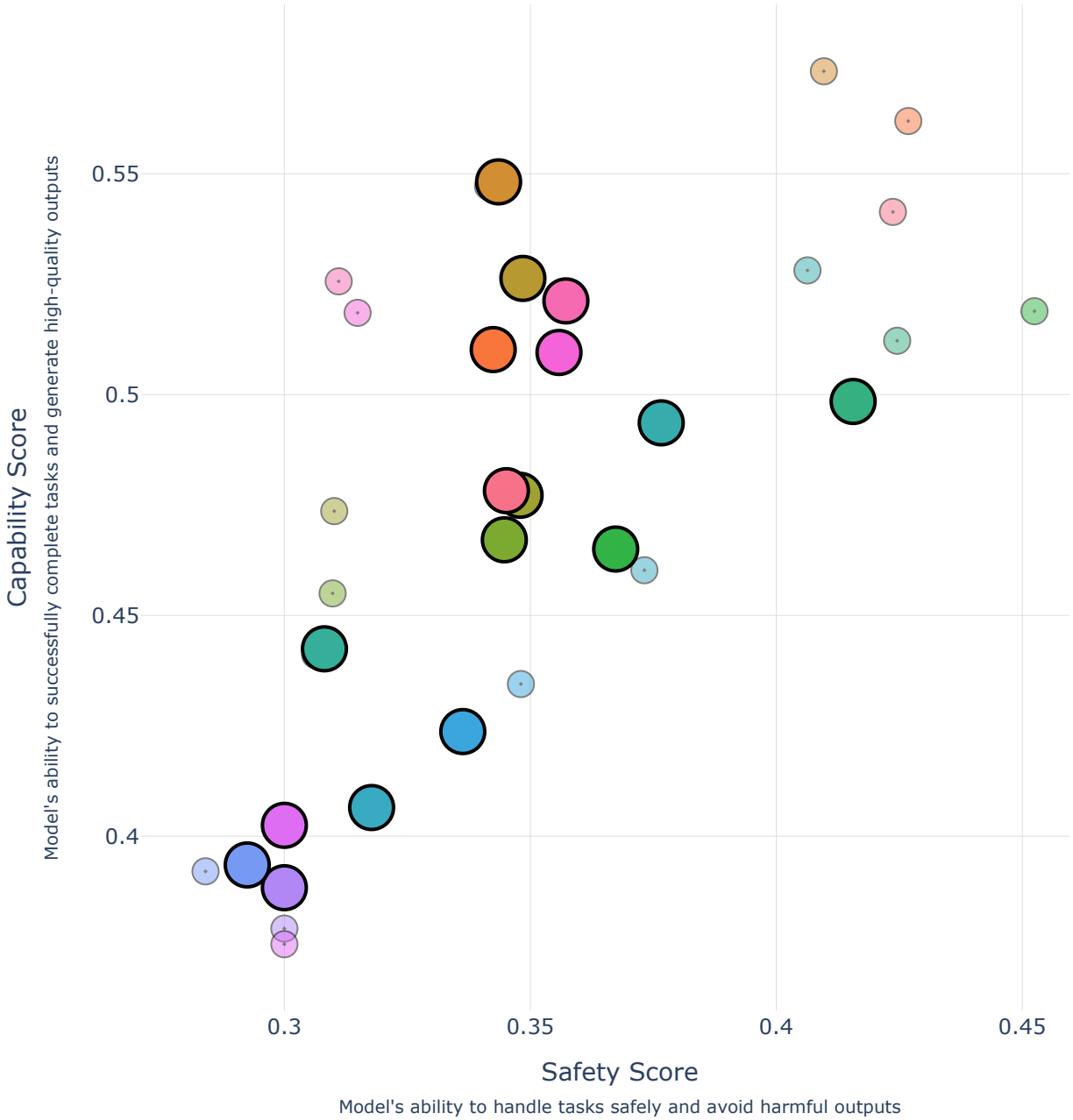
In the below visualization two dimensions can be plotted against each other to understand the tradeoffs involved in model selection.

In addition, We can also average the scores for each model across all use cases, giving a generalized enterprise suitability score along different dimensions. This is *not* the same as the generic evaluation, as the totality of our

use cases are not “generic”. They still relate to uses that are relevant for different enterprises. So one can think of this aggregation step as a way to get a holistic understanding of a model’s capabilities and safety across a wide range of use cases within the enterprise context.

We can then view how this “average” enterprise performance compares to specific industries. “Financial Industries” is selected by default.

Model Trust Scores: Model Comparisons



Y-axis:

Capability Score ▼

X-axis:

Safety Score

Industries:

Show All Industries ▼

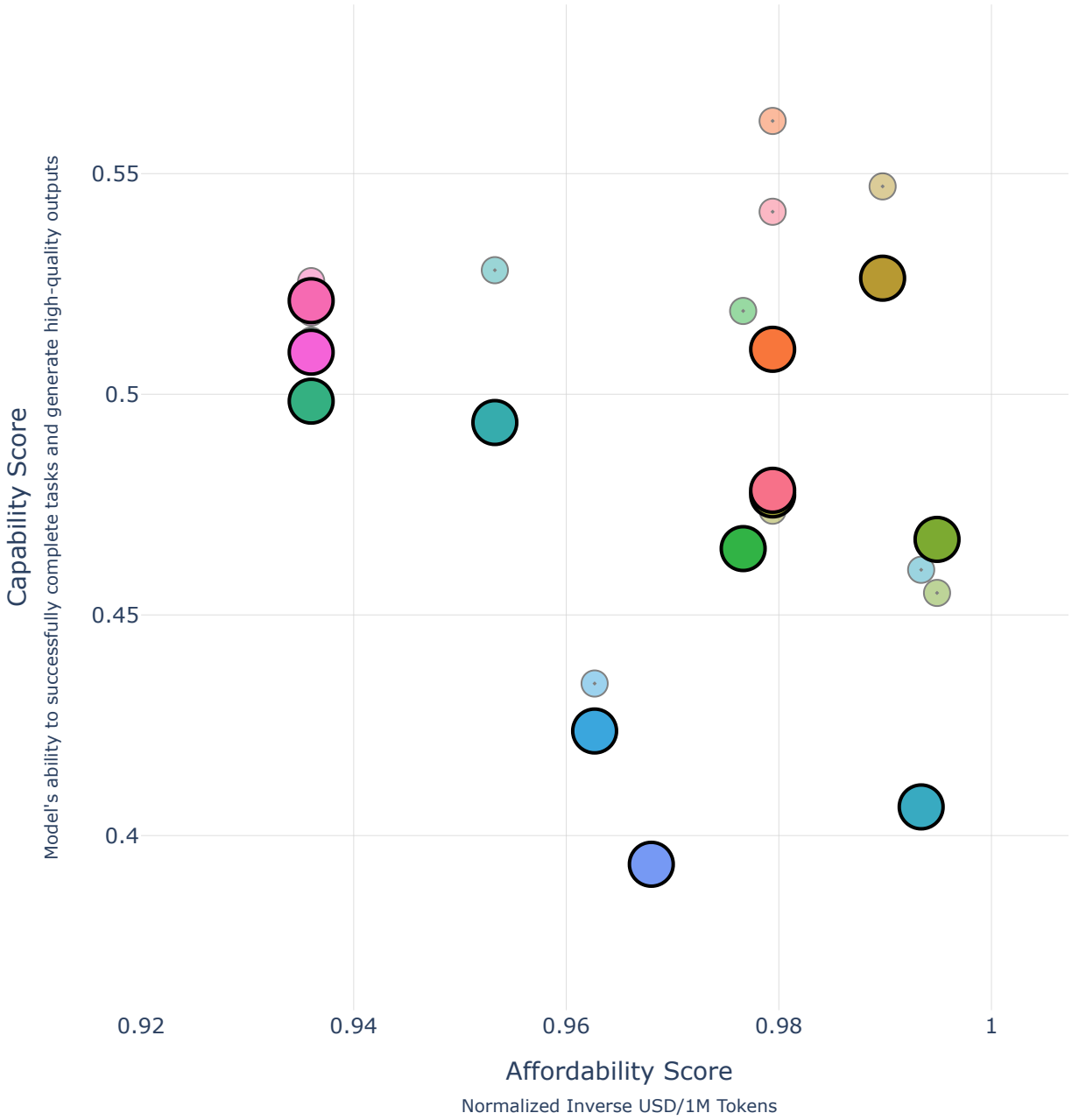
4.2.5.3 How to use Model Trust Scores

How can scores be used for decision making? There are a number of ways, but they all are fundamentally based in evaluating model *tradeoffs*. While it may occasionally be the case that one model is more capable and safer than all others (capability and safety don't necessarily trade off!), it's unlikely that the same model will also be the cheapest, or fastest.

One starting point is looking at the "Overall Score" against "Cost". This showcases a balanced measure of capability and safety against cost. It may be helpful to restrict the range of the x-axis, because o1 costs far more than the rest, obscuring the tradeoffs amongst the cost competivive models.

If safety concerns are not as critical for the use case, "Capability" vs "Cost" may be more relevant. If we look at this comparison, it becomes obvious why DeepSeek R1 made a splash in the AI ecosystem. Beyond being a new player in the AI ecosystem from China, DeepSeek R1 is genuinely high performing and very inexpensive compared to its peers, as can be seen below (note that the X-axis range has been restricted which removes o1 from the plot. o1 is significantly more expensive than the rest of the models).

Model Trust Scores: Model Comparisons



Y-axis:

Capability Score ▼

X-axis:

Affordability Score

Industries:

Show All Industries ▼

4.2.5.4 Model Ranking, AI Systems, and Caveats

Clearly, this tool can potentially enable making very powerful statements about the capability and safety of the models in the AI ecosystem. However, we do not have access to the models' "true" capabilities and safety. We can only make statements about models based on the benchmarks we have available. For instance, we do not have AILuminate scores for DeepSeek R1 and thus can't make highly informed statements about its safety. We deal with this by appealing to a pessimistic prior, but this reflects a precautionary principle - not an accurate estimate of the model's true safety.

Model Trust Scores provides an informed and actionable *synthesis* of existing evaluations, but requires a robust evaluation ecosystem to be most useful.

We believe the onus should be on model providers to demonstrate their model's safety and capabilities in a way that is transparent and applicable to the applications they propose. This means running internal evaluations **and** seeking out independent third-parties to evaluate their models. The Model Trust Score pessimistic prior essentially downgrades models that are not evaluated well by the ecosystem. We believe this is a reasonable compromise and connects to responsible decision making within enterprises. As powerful AI capabilities become increasingly accessible, why trust a model that hasn't proven its trustworthiness? We believe that Model Trust Scores can help galvanize a more comprehensive ecosystem evaluations by showcasing industry gaps in evaluation coverage and downgrading models that fall below evaluation expectations due to poor performance or poor transparency.

Moreover, we are synthesizing benchmarks on AI models, not AI systems tuned for a particular use case (or benchmark). It is likely that every model can do better on certain benchmarks with the proper scaffolding, just as an AI model within a particular use case application will do much better than a naive evaluation of the model would imply. The approach we take can easily generalize to AI systems however. As long as there is an evaluation, we can identify its use case relevance and make context-specific claims of an AI system's suitability, whether a base model, tool-using agent, or any other system.

5 Conclusion

5.1 Future Work: Improving the Evaluation Landscape and Certification

Our analysis of relevance scores reveals a critical insight: many industries lack benchmarks that directly measure the capabilities needed for their specific use cases. While Model Trust Scores help organizations make the best decisions possible with current evaluations, there's significant room for improvement in how we assess models for enterprise use.

5.1.1 Developing Use Case Specific Evaluations

The path forward requires developing benchmarks that more precisely target individual industries and use cases. Our relevance scoring system not only helps contextualize existing benchmarks but also highlights which industries are most underserved by current evaluation approaches. This information proves particularly valuable when combined with risk assessments – industries that are both underserved by benchmarks and face significant potential harms from AI deployment should be prioritized for evaluation development.

These improved evaluations can emerge from several sources. It is an active area of research and institution development to figure out how to best do this: - Third-parties (e.g., industry consortiums, nonprofits) creating

benchmarks and standardized test suites (AILuminate by MLCommons is a good example here) - Research institutions exploring novel assessment methods (LegalBench is an industry specific evaluation developed by an open scientific effort led by Stanford University) - AI providers and deployers could develop and share evaluation approaches relevant for real-world applications (e.g., SimpleQA from OpenAI or Model Written Evals by Anthropic) - Regulatory bodies establishing compliance frameworks founded on quantitative evaluation - AI Safety Institutes, either alone or in partnership with other organizations mentioned above.

As evaluations mature for specific use cases, we can move beyond individual evaluations toward comprehensive assessment frameworks. When we can reliably measure all relevant dimensions of a use case – from technical capabilities to safety controls – we can develop omnibus scores that simplify model selection while maintaining rigor.

5.1.2 From Relative to Absolute Trust

Model Trust Scores currently helps organizations compare models relative to one another, identifying which options are safer or more capable within the available choices. This relative assessment provides crucial guidance for model selection. However, the future of AI governance requires moving beyond relative comparisons to “absolute trust”, potentially reflected by third-party certifications, where the assessment results are placed in the context of best practice and thorough and independent risk/benefit analyses.

An assessment framework for certifications would answer fundamental questions: - Does any available model meet the minimum capability requirements for this use case? - Are there safety thresholds below which no model should be deployed, regardless of capabilities? - What level of evidence is required to establish trustworthiness in high-stakes contexts?

Our current safety and capability scores provide comparative insights but don’t yet map directly to real-world suitability thresholds. Establishing these thresholds – particularly when they must account for multiple dimensions of performance and risk – represents a crucial step toward meaningful AI certification frameworks which can further bolster ecosystem trust and information sharing.

This evolution from relative comparison to certification would transform how organizations approach AI adoption. Rather than simply choosing the best available option, they could confidently determine whether any current model meets their requirements. This shift becomes especially critical as AI systems take on increasingly consequential roles across industries. Whether these certifications are mandated for use (as in the case of permits) or informative to market actors (as in the case of third-party labeling) is a downstream question beyond the scope of this paper.

The path to certification requires collaboration between multiple stakeholders: - Industry experts who understand use case requirements - Safety researchers who can establish risk thresholds - Evaluation specialists who can design comprehensive tests - Regulatory bodies who can standardize certification processes - Enterprise users who can validate real-world performance

As we develop these more sophisticated evaluation and certification frameworks, the Model Trust Score Framework will evolve to incorporate both relative and absolute assessments, providing organizations with increasingly comprehensive guidance for safe and effective AI adoption.

5.2 Bridging the Gap Between Governance & Assurance

There is often a significant gap between governance considerations and technical evaluations—how do you know whether a particular evaluation result is good or secure or compliant?

The Credo AI Platform is designed to bridge this gap. Through Model Trust Scores, the platform ingests structured model-level benchmarks from academic and public sources, providing organizations with the tools to interpret these results in a governance and risk context.

But evaluating models isn't just about interpreting existing benchmarks—it's also about determining what additional assessments are needed for a specific use case. Credo AI helps governance teams define these requirements, guiding implementers on what additional evaluations to run based on risk thresholds, regulatory obligations, and enterprise policies. Benchmarks are helpful for a first pass, but context-specific assessments are critical to making risk-informed decisions about which models to trust in critical business applications.

By translating governance decisions into technical configurations and automating policy-to-code workflows, Credo AI ensures that evaluation insights drive real enforcement. Tight integrations with ops providers make it easy to run necessary evaluations and pull results back into the platform, where they become part of a unified governance repository. This structured, closed-loop approach empowers organizations to visualize, understand, and act on AI risks—establishing Credo AI as the single source of truth for AI governance.