The AI Productivity Index (APEX)

Bertie Vidgen[1] **Abby Fennelly**[1] **Evan Pinnix**[1] **Chirag Mahapatra**[1]
Zach Richards[1] **Austin Bridges**[1] **Calix Huang**[1] **Ben Hunsberger**[1]
Fez Zafar[1] **Brendan Foody**[1] **Dominic Barton**
Cass R. Sunstein[2] **Eric Topol**[3] **Osvald Nitski**[1*]
[1]Mercor [2]Harvard Law School [3]The Scripps Research Institute

Abstract

We introduce the first version of the **AI Productivity Index (APEX)**, a benchmark for assessing whether frontier AI models can perform knowledge work with high economic value. APEX addresses one of the largest inefficiencies in AI research: outside of coding, benchmarks often fail to test economically relevant capabilities. APEX-v1.0 contains 200 test cases and covers four domains: investment banking, management consulting, law, and primary medical care. It was built in three steps. First, we sourced experts with top-tier experience e.g., investment bankers from Goldman Sachs. Second, experts created prompts that reflect high-value tasks in their day-to-day work. Third, experts created rubrics for evaluating model responses. We evaluate 23 frontier models on APEX-v1.0 using an LM judge. GPT 5 (Thinking = High) achieves the highest mean score (64.2%), followed by Grok 4 (61.3%) and Gemini 2.5 Flash (Thinking = On) (60.4%). Qwen 3 235B is the best performing open-source model and seventh best overall. There is a large gap between the performance of even the best models and human experts, highlighting the need for better measurement of models' ability to produce economically valuable work.

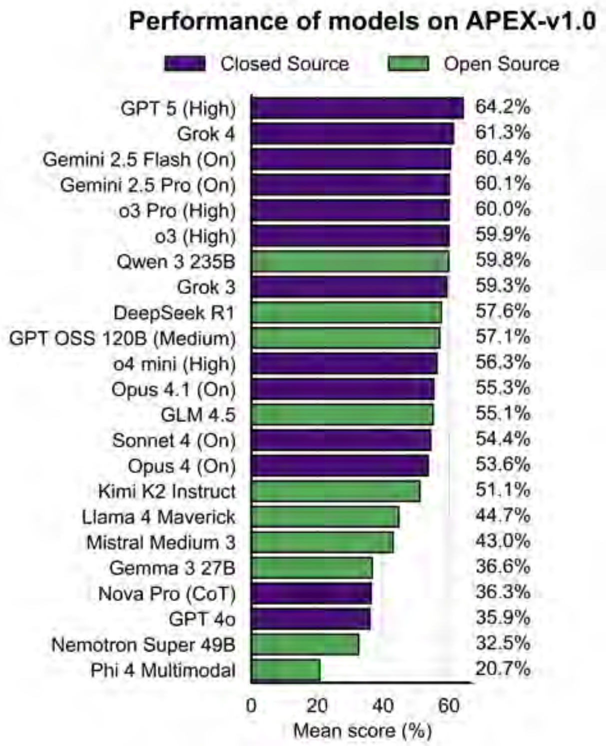


Figure 1: Models' mean score on APEX-v1.0. Models are ranked in descending order. The labels in parentheses indicate the "Thinking" settings used where a choice is available.

1 Introduction

Benchmarks help the AI community track progress, assess model capabilities, and hillclimb performance (Kiela et al., 2021; Schwartz et al., 2025; Weidinger et al., 2025). However, most benchmarks measure abstract model capabilities, rather than economically valuable outputs. The true economic impact of advances in frontier AI, and the potential of models to replace or augment human work, remains unquantified (Topol, 2019; Cahn, 2024; Bar-Gill and Sunstein, 2025). This gap between (1) what benchmarks evaluate and (2) what AI is actually being used for in production is one of the biggest obstacles to creating real-world value. To solve this problem, we have created the first version of the AI Productivity Index (APEX-v1.0), working with a team of industry experts. The purpose of APEX is to set a goal for AI progress that is aligned with economically useful tasks in the real-world.

APEX-v1.0 will remain a closed heldout dataset for rigorous evaluation of frontier models on their ability to execute tasks across four high-value knowledge jobs: investment banking associate, management consultant, big law associate, and primary

*Email: apex@mercor.com, osvald@mercor.com

First published: arXiv:2509.25721v2 2 Oct 2025

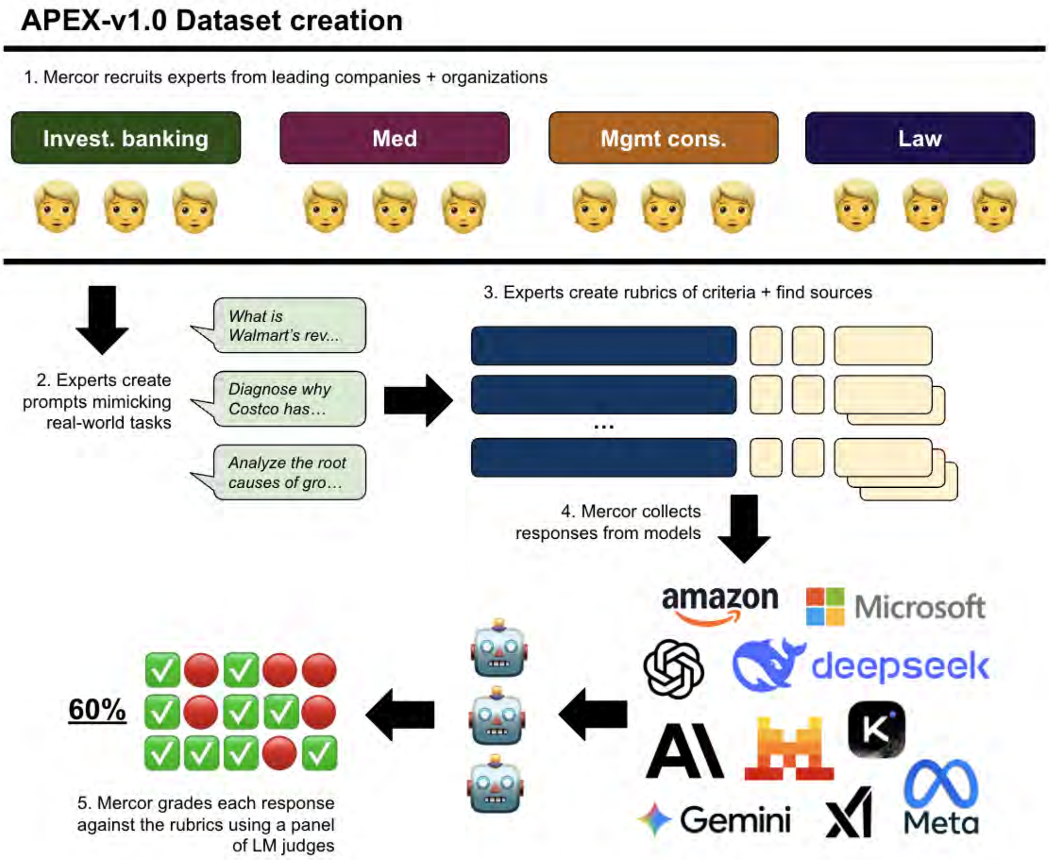


Figure 2: Workflow for creating the AI Productivity Index (APEX-v1.0). Quality control is applied at every step in production to ensure that prompts and rubrics are high-quality.

care physician (MD). Each prompt in APEX-v1.0 is a knowledge task that would take an expert between 1 and 8 hours to complete (with a mean of 3.5), and requires sophisticated reasoning. Prompts are provided with evidence sources, and a rubric of prompt-specific quality criteria. We use the rubrics to autograde responses using a panel of judge LMs, which is highly correlated with human grades (see Section 4). An example prompt and part of a rubric is given in Figure 3, and complete examples for each domain are given in Appendix D. We have produced a public-facing leaderboard for APEX-v1.0 with results for 23 models from 13 providers.[1] 13 models are closed source and 10 are open source. Contact us to submit your model for testing.

2 Dataset overview

APEX-v1.0 comprises $n = 200$ cases, split evenly across the four domains, as shown in Table 1. Every case was created and reviewed by experts with extensive industry experience.

[1] See mercor.com/apex.

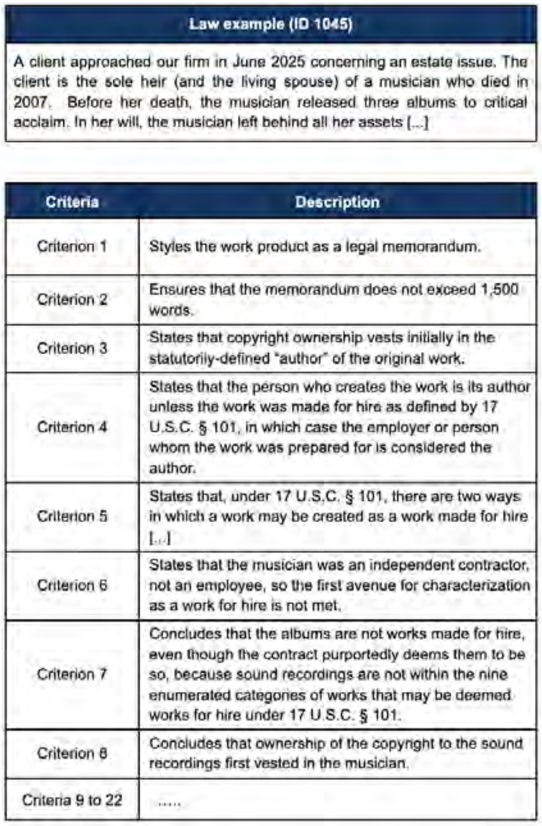
2.1 Expert sourcing and vetting

Experts were sourced through the Mercor platform. We targeted experts with appropriate experience, such as 3+ years at a top consulting firm, and prioritized experts with data labeling experience. Once sourced, experts completed a 30 to 45-minute interview. Experts who demonstrated excellent domain knowledge, strong communication and reasoning skills, and understood the potential role of AI in their industry, were paid to complete a 1-2 hour assessment that tested their ability to write prompts and rubrics. If they successfully completed the assessment, they were contracted to work on APEX-v1.0. Throughout the project we continually checked in with the experts, gave qualitative feedback, and as needed offboarded underperformers. In total, 76 experts, with a mean experience of 7.25 years, contributed at least one case to APEX-v1.0.

- **Investment Banking:** Twenty investment bankers with between 2 and 18 years of experience, and a mean of 8.7 years. They have held

Table 1: Overview of APEX-v1.0. Mean number of criteria and sources per case, and mean number of tokens in all the sources combined per case, and mean number of tokens in the prompt.

Domain	**No. of cases**	**No. of criteria**	**No. of sources**	**No. of tokens in all sources**	**No. of tokens in prompt**
Medicine	50	36.20	6.50	41,443	242
Law	50	27.74	7.02	31,029	494
Management Consulting	50	32.10	5.20	18,445	587
Investment Banking	50	20.32	4.62	15,787	398
Mean	**50**	**29.09**	**5.83**	**26,676**	**430**


Law example (ID 1045)

A client approached our firm in June 2025 concerning an estate issue. The client is the sole heir (and the living spouse) of a musician who died in 2007. Before her death, the musician released three albums to critical acclaim. In her will, the musician left behind all her assets [...]

Criteria	Description
Criterion 1	Styles the work product as a legal memorandum.
Criterion 2	Ensures that the memorandum does not exceed 1,500 words.
Criterion 3	States that copyright ownership vests initially in the statutorily-defined "author" of the original work.
Criterion 4	States that the person who creates the work is its author unless the work was made for hire as defined by 17 U.S.C. § 101, in which case the employer or person whom the work was prepared for is considered the author.
Criterion 5	States that, under 17 U.S.C. § 101, there are two ways in which a work may be created as a work made for hire [...]
Criterion 6	States that the musician was an independent contractor, not an employee, so the first avenue for characterization as a work for hire is not met.
Criterion 7	Concludes that the albums are not works made for hire, even though the contract purportedly deems them to be so, because sound recordings are not within the nine enumerated categories of works that may be deemed works for hire under 17 U.S.C. § 101.
Criterion 8	Concludes that ownership of the copyright to the sound recordings first vested in the musician.
Criteria 9 to 22	.....

Figure 3: Example rubric for **Law (ID 1045)** with the first 8 criteria out of 22 total. It is supported by 8 evidence sources, which total to under 100,000 tokens. ID 1045 is not part of the APEX-v1.0 heldout test set. It was created concurrently by the same group of experts.

positions at firms including Goldman Sachs, Evercore, and JPMorgan.

- **Management Consulting:** Eighteen management consultants with between 2 and 20 years of experience, and a mean of 6.9 years. They have held positions at firms including McKinsey, BCG and Bain.

- **Law:** Twenty lawyers, with experience at Big Law firms, with between 3 and 22 years of experience, and a mean of 5 years. They have held positions at firms including Latham & Watkins, Skadden, and Cravath, Swaine & Moore, and hold JDs from institutions including Harvard, Yale, Stanford, and other Top 14 US Law Schools.

- **Medicine:** Eighteen physicians with between 3 and 22 years of clinical experience in primary care, and a mean of 8.8 years. They have experience at hospitals including Brigham & Women's and Mount Sinai, and hold MDs from institutions including the University of Pennsylvania, Northwestern, Cornell, and other top US medical schools.

2.2 Expert creation of the prompts and quality-assessment rubrics

On joining the project, experts outlined the 3-5 most common tasks in their day-to-day work, describing the seniority and expertise required. An overview, with an estimate of their prevalence, is given in Appendix C. We used this information to steer the creation of prompts in APEX-v1.0, ensuring they have fidelity to the real distribution of economically-valuable tasks. For each case, experts found or created the evidence sources (if suitable sources were not freely available, such as patient records). We allowed both PDFs and CSVs, up to a maximum combined length of 100,000 tokens, checked against OpenAI's tokenizer for GPT 4o.[2] This ensures the sources fit within the context window of all tested models.

Experts created a rubric of quality criteria for each prompt, decomposing the hard-to-measure concept

[2]See openai.com/tokenizer.

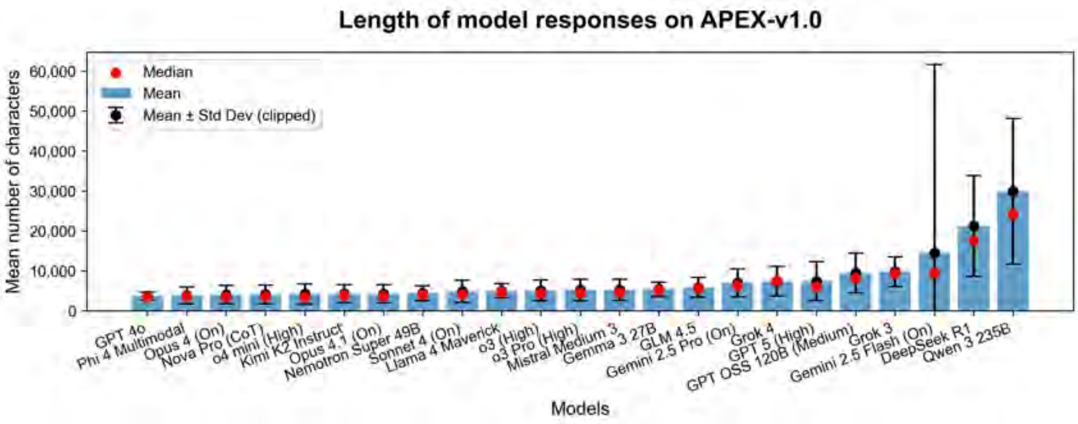


Figure 4: Median, mean and standard deviation of the length of model responses on APEX-v1.0, measured in characters and split by model. We clip mean minus standard deviation at 0 for readability.

of response "quality" into testable discrete components (Saad-Falcon et al., 2024; Arora et al., 2025; Starace et al., 2025). Each criterion is an objective, well-specified and self-contained statement about the response, phrased as a descriptive claim. They are analogous to unit tests for code, and can be assessed as Pass or Fail. For instance, if the prompt asks the model to analyze the growth opportunities for the five largest US Airlines in 2025, one criterion in the rubric could stipulate "The response identifies Delta airlines as one of the five largest airlines in the USA by market capitalization in 2025." We aim for parsimonious rubrics: more criteria create a more fine-grained signal to assess the quality of responses but can also encourage unnecessarily long answers if badly constructed.

2.3 Quality control

Experts created and submitted each prompt for review. They were then approved or rejected by reviewers. For rejections, the reviewers could either remove the prompt from product or request specific changes so it would met the quality bar. After the prompt was approved the contributor created the quality rubric, which was then reviewed. Multiple rounds of review help to ensure the data points meet the requirements of the task, are varied in terms of topic, style and model behavior, and are sufficiently challenging. Experts start as contributors and, only if they perform exceptionally well and demonstrate strong understanding of the task requirements, can be promoted to reviewers. We use in-house LM-powered reviewing tools that give experts immediate feedback on their work. Experts always remain responsible for the quality of their submissions. In total, 300 prompts were started by contributors, of which 200 were approved by reviewers and added to APEX-v1.0.

2.4 Dataset description

The mean number of tokens per prompt is 430, ranging from 75 to 2,806, and the mean number of characters is 2,029. The length of prompts differs across domains. The mean number of tokens for medicine is 242 and for management consulting is 587, over twice as many. Our qualitative review suggests this is because the medicine prompts rely more on the source evidence to specify the details of the prompt compared with the other domains. The mean number of criteria per rubric is 29.09, ranging from 7 to 54, with 5,818 criteria in total. There are large differences in the mean criterion count per rubric across domains, from 20.32 for investment banking to 36.20 for medicine. Our qualitative review suggests this is because rubrics in medicine tend to be very precise about outcomes, often referencing specific industry-standard guidelines and regulations. However, it is difficult to generalize as there is a large spread of values in every domain. The mean number of evidence sources for each case is 5.83, ranging from 1 to 18. The mean number of tokens for all sources associated with each case is 26,677, ranging from 89 to 93,667.

3 Experimental setup

3.1 Models selection and access

We tested 23 models on APEX-v1.0. All models were released in 2025 apart from Amazon's Nova Pro (released December 2024) and OpenAI's GPT 4o (initially released in May 2024 and last updated in November 2024). The most recent model

Table 2: Performance and rankings of models on APEX-v1.0 across four metrics: mean score (%), pairwise wins (%), times ranked first (%), and times ranked last (%). For the first three metrics, a higher score is better. For the % of times that a model is ranked last, a lower score is better. Models are grouped by whether they are closed source (top section) or open source (bottom section).

		Mean Score		**Pairwise Wins**		**Ranked 1st**		**Ranked Last**	
Model	**Provider**	%	Rank	%	Rank	%	Rank	%	Rank
Nova Pro (Thinking = CoT)	Amazon	36.3	20	20.5	20	0.0	19	7.0	4
Opus 4.1 (Thinking = On)	Anthropic	55.3	12	55.5	13	2.0	11	**0.0**	**17**
Opus 4 (Thinking = On)	Anthropic	53.6	15	50.9	15	0.5	17	1.0	8
Sonnet 4 (Thinking = On)	Anthropic	54.4	14	54.3	14	1.0	15	**0.0**	**17**
Gemini 2.5 Flash (Thinking = On)	Google	60.4	3	69.6	3	11.5	3	0.5	10
Gemini 2.5 Pro (Thinking = On)	Google	60.1	4	69.2	4	5.0	6	0.5	10
GPT 5 (Thinking = High)	OpenAI	**64.2**	**1**	**77.5**	**1**	**27.5**	**1**	**0.0**	**17**
o3 Pro (Thinking = High)	OpenAI	60.0	5	67.7	7	2.0	11	0.5	10
o3 (Thinking = High)	OpenAI	59.9	6	67.9	6	3.5	8	**0.0**	**17**
o4 mini (Thinking = High)	OpenAI	56.3	11	58.5	11	2.0	11	**0.0**	**17**
GPT 4o	OpenAI	35.9	21	18.2	21	0.5	17	14.0	2
Grok 4	xAI	61.3	2	72.5	2	9.5	4	**0.0**	**17**
Grok 3	xAI	59.3	8	66.3	8	3.5	8	0.5	10
DeepSeek R1	DeepSeek	57.6	9	63.1	9	13.5	2	1.0	8
Gemma 3 27B	Google	36.6	19	21.5	19	0.0	19	5.0	5
Llama 4 Maverick	Meta	44.7	17	32.5	17	1.0	15	2.0	6
Phi 4 Multimodal	Microsoft	20.7	23	4.3	23	0.0	19	50.5	1
Mistral Medium 3	Mistral	43.0	18	31.3	18	0.0	19	0.5	10
Kimi K2 Instruct	Moonshot	51.1	16	47.4	16	2.0	11	0.5	10
Nemotron Super v1 49B	Nvidia	32.5	22	14.9	22	0.0	19	14.0	2
GPT OSS 120B (Thinking = Medium)	OpenAI	57.1	10	61.5	10	7.5	5	0.5	10
Qwen 3 235B	Qwen	59.8	7	68.3	5	5.0	6	**0.0**	**17**
GLM 4.5	Z	55.1	13	56.5	12	2.5	10	2.0	6

is GPT 5 (Thinking = High), which was released in early August 2025. Model responses were collected at the start of August 2025. The 13 closed source models were accessed through their respective APIs and the 10 open source models were accessed through open source providers. We use the recommended temperature for each model. If a setting is not recommended, we set temperature to 0.7. Thinking is turned On if available and set to the recommended level (usually, High) if it can be configured. We do not explicitly set the system prompt, apart from Nova Pro where it is used to enable “thinking”.[3] Implementation details are given in Appendix A.

3.2 Model scoring

We collect responses from each model three times for each prompt and score them with an LM judge (see below). Because models are non-deterministic they can give different responses to the same prompt, resulting in different scores each run. Across all models, the mean range of the scores for the three runs is 11.9 percentage points. This is fairly consistent across models, from a mean range of 9.0 percentage points for Grok 3 to 16.2 percentage points for Gemini 2.5 Flash (Thinking = On). We use the median of the three scored responses for our leaderboard and analysis. We considered reporting the maximum score across runs (i.e. pass@3) as this would show the upper boundary of model performance with multiple attempts. However, this risks inflating scores for models that produce more variable outputs – a single lucky completion could obscure low performance on average.

The length of model responses varies considerably, as given in Figure 4, from a mean of 3,679 characters for GPT 4o to 29,914 for Qwen 3 235B. Gemini 2.5 Flash (Thinking = On) has the third longest mean responses at 14,522 characters. However, this is not representative due to the large standard deviation. It is primarily due to five cases that have an exceptionally high number of characters (821,912, 704,329, 364,202, 175,267 and 155,257), partly due to long and irrelevant sequences of consecutive spaces. Outside of these cases, Gemini 2.5 Flash

[3]See nova/prompting-chain-of-thought.html.

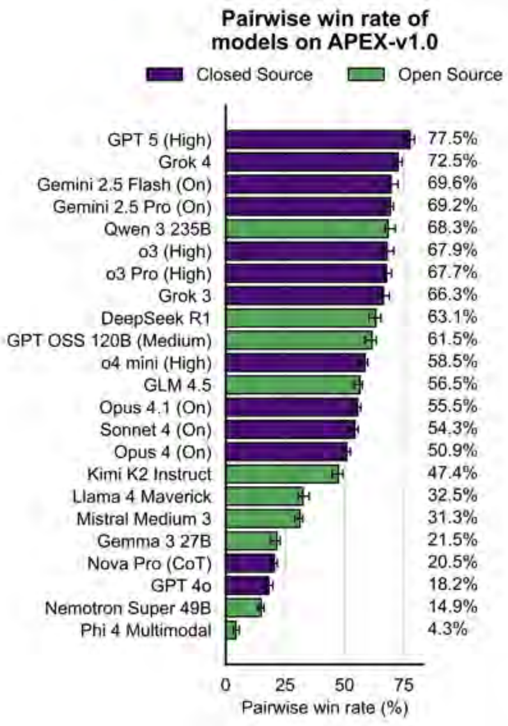


Figure 5: Models' pairwise win rate on all $n = 200$ cases in APEX-v1.0. Pairwise win rate is based on a head-to-head comparison of models' autograded scores on each task. The model that scores higher wins and the other model loses. Both models are awarded half a point for exact ties. The error bars show 95% Confidence Intervals, bootstrapped $n = 1,000$.

(Thinking = On) responses are not much longer than other models'.

4 Grading responses with a panel of LM judges

LM judges are widely used to scalably grade the quality of model responses (Baumann et al., 2025; Gu et al., 2025; Zhu et al., 2025). LMs can be biased judges and require careful prompting and adequate evaluation. LM judging is well-suited for grading rubrics as each criterion is a short self-contained statement. In principle, these statements are easier to judge than overall quality or abstract concepts like "usefulness" or "helpfulness". We used a panel of three LMs to grade model responses on APEX-v1.0: o3 (Thinking = Low), Gemini 2.5 Pro (Thinking = Off), and Sonnet 4 (Thinking = Off). The same prompt is passed to all three LM judges. It is given in Appendix B.

Each judge independently graded each criterion as Pass or Fail, and we took the majority vote (i.e., 3/3 or 2/3 agreement) as the final grade. The score for each response is the percentage of criteria in the rubric that it passes. For instance, if a response passed 16 out of the 20 criteria in a rubric we score it 80%. This approach lets us calculate a scalar grade for each response from the overall percentage of criteria passed. We assess the performance of the LM judges on four metrics: (1) judge consistency, (2) inter-judge agreement, (3) judge preference for itself, and (4) agreement between the judges and human labels. Overall, we find our LM judge panel is a consistent and high-quality grader.

Judge consistency. Each judge graded the responses from one arbitrarily selected model (Grok 4) three times. Their grades are internally consistent, with very high levels of agreement: 99.5% for o3 (Thinking = Low), 99.4% for Gemini 2.5 Pro (Thinking = Off), and 99.6% for Sonnet 4 (Thinking = Off).

Inter-judge agreement. We checked whether judges agree on the Pass/Fail grades given to each criterion in the dataset (comprising over 400,000 data points). The three judges agree on criterion grades 3/3 in 81.16% of cases. For cases where there is not 3/3 agreement, there is a fairly even spread of 2/3 majorities: 42.1% majority agreement between Sonnet 4 (Thinking = Off) and Gemini 2.5 Pro (Thinking = Off), 33.1% majority agreement between Sonnet 4 (Thinking = Off) and o3 (Thinking = Low), and 24.8% majority agreement between Gemini 2.5 Pro (Thinking = Off) and o3 (Thinking = Low). The three models are also moderately correlated: Gemini 2.5 Pro (Thinking = Off) and Sonnet 4 (Thinking = Off) have a correlation of 0.80, o3 (Thinking = Low) and Sonnet 4 (Thinking = Off) have a correlation of 0.75, and o3 (Thinking = Low) and Gemini 2.5 Pro (Thinking = Off) have a correlation of 0.72.

Judge preference for itself. As all three LM judges are models evaluated on APEX-v1.0 (although with thinking turned off for Gemini 2.5 Pro and Sonnet 4 and set to Low for o3) we check judge self-preference by calculating the gap between the grade given by the LM judge to itself and the mean grade from the other two judges. Sonnet 4 (Thinking = Off) is the harshest judge, scoring itself 50.8% while Gemini 2.5 Pro (Thinking = Off) scores it 59.4% and o3 (Thinking = Low) scores it 54.2%. This is a gap of -6 percentage points. o3 (Thinking = Low) is the best calibrated, scoring itself 59.9% while Gemini 2.5 Pro (Thinking = Off) scores it 64.7% and Sonnet 4 (Thinking = Off) scores it 56.2%. This is a gap of -0.6 percentage points. Gemini 2.5 Pro (Thinking = Off) is the most generous judge, scoring itself 63.5% while Sonnet 4 (Thinking = Off) scores it 55.8% and o3 (Thinking = Low) scores it 60.4%. This is a gap of $+5.4$ percentage points. These differences align very closely with judges' overall biases when grading models. Compared with the other two judges, Sonnet 4 (Thinking = Off) typically scores criterion -5.6 percentage points, o3 (Thinking = High) scores -0.5 percentage points, and Gemini 2.5 Pro (Thinking = Off) scores $+6.1$ percentage points. Therefore, we do not find that judges prefer themselves but instead a general bias, which is partly mitigated by using a panel.

Agreement between judges and humans. For the responses from one model (Gemini 2.5 Pro (Thinking = On)) we collected Pass/Fail grades from the expert annotators (n=5,818 total). Note that these ratings were not validated by reviewers, creating a greater risk of errors than with our other annotations. Using majority vote, the LM judge panel is in 89% agreement with the human grades. This is a small lift compared with the individual judges – Gemini 2.5 Pro (Thinking = Off) is in 87.7% agreement with the human grades, Sonnet 4 (Thinking = Off) is 88.1% and o3 (Thinking = Low) is 88.4%. There are minor differences in judge-human agreement across the four domains: 84.3% in law, 85.7% in medicine, 88.0% in investment banking and 94.5% in management consulting. These differences could be a product of model's different reasoning capabilities across domains, stylistic differences in how the criteria are worded creating errors, confounding factors like different complexity of criteria, or mistakes from the expert annotations. We aim to investigate these differences in the future and improve LM judge performance.

5 Results

GPT 5 (Thinking = High) has the highest mean score on APEX-v1.0 at 64.2%, followed by Grok 4 (61.3%) and Gemini 2.5 Flash (Thinking = On) (60.4%). Just two percentage points separate the 2nd to 7th best models, from 59.3% to 61.3%. Performance differences are much larger at the bottom of the leaderboard, where seven models score under 50%. The lowest performing models are Phi 4 (20.7%), Nemotron Super v1 49B (32.5%) and GPT 4o (35.9%). Using a Kruskal-Wallis significance test, the differences between models are statistically significant at $\alpha = 0.00001$. We fit an Ordinary Least Squares regression model with score as the outcome variable, and can explain 22.8% of variance using just the model name and no other coefficients. Models' mean score on APEX-v1.0 is shown in Table 3.

Domains differ in difficulty. Across all models the mean score for medicine is 47.5%, investment banking is 47.6%, management consulting is 52.6%, and law is 56.9%. This is a range of 9.4 percentage points. The highest model scores in each domain are similarly varied (62.0%, 59.7%, 64.8%, 70.5% respectively). Yet, despite differing difficulty and the different knowledge and reasoning skills they require, models' rank positions are fairly consistent across domains. GPT 5 (Thinking = High) is the best performing model in every domain and Phi 4 Multimodal is the worst. Grok 4 and Gemini 2.5 Pro (Thinking = On) are tied second in management consulting, and DeepSeek R1 and Gemini 2.5 Pro (Thinking = On) are tied second in investment banking. Interestingly, Gemini 2.5 Pro (Thinking = On) performs much worse in medicine (ranked 9th) and law (ranked 7th). Gemini 2.5 Flash (Thinking = On), in contrast, performs well on medicine (tied 2nd with GPT OSS 120B (Thinking = Medium) and medicine (tied 3rd with o3 Pro (Thinking = High), after o3 (Thinking = High) in second) but is worse on management consulting (ranked 4th) and investment banking (ranked 9th).

To better understand how models compare we calculate (1) the percentage of times each model performs best on a given task, (2) the percentage of times each model performs worst, and (3) the percentage of pairwise head-to-head battles that each model wins. Results are given in Table 2. There is a large spread in the pairwise win rate, from GPT 5 (Thinking = High) winning 77.5% of all head-to-heads, and Phi 4 winning just 4.3%. GPT 5 (Thinking = High) is also clearly differentiated from the second best-performing model, Grok 4 at 72.5%, and the third-best performing, Gemini 2.5 Flash (Thinking = On) at 69.6%. This shows how even fairly small differences can translate to larger differences in ranked position. We calculate 95% confidence intervals for the pairwise win rate

Table 3: Model performance across knowledge domains in APEX-v1.0 (investment banking, law, management consulting, and medicine). Scores are the mean percentage of criteria passed for tasks in each domain. In all four domains, GPT 5 (Thinking = High) is the highest performing model.

Model	Provider	Invest. Banking	Law	Mgmt. Consulting	Medicine
Nova Pro (Thinking = CoT)	Amazon	31.4%	44.2%	36.2%	33.5%
Opus 4.1 (Thinking = On)	Anthropic	55.2%	59.7%	56.6%	49.8%
Opus 4 (Thinking = On)	Anthropic	53.8%	57.4%	54.6%	48.5%
Sonnet 4 (Thinking = On)	Anthropic	53.7%	57.8%	55.3%	50.7%
Gemini 2.5 Flash (Thinking = On)	Google	55.0%	67.2%	60.6%	58.9%
Gemini 2.5 Pro (Thinking = On)	Google	57.8%	65.2%	62.7%	54.9%
GPT 4o	OpenAI	32.6%	41.6%	38.9%	30.5%
GPT 5 (Thinking = High)	OpenAI	**59.7%**	**70.5%**	**64.8%**	**62.0%**
o3 (Thinking = High)	OpenAI	56.9%	68.0%	58.4%	56.2%
o3 Pro (Thinking = High)	OpenAI	57.1%	66.7%	58.6%	57.7%
o4 mini (Thinking = High)	OpenAI	54.8%	63.6%	56.8%	50.1%
Grok 3	xAI	55.9%	65.2%	59.1%	57.1%
Grok 4	xAI	56.6%	66.2%	63.1%	59.3%
DeepSeek R1	DeepSeek	58.2%	59.4%	60.4%	52.4%
Gemma 3 27B	Google	25.5%	45.5%	40.6%	34.9%
Llama 4 Maverick (I)	Meta	46.5%	45.7%	51.8%	34.7%
Phi 4 Multimodal (I)	Microsoft	10.4%	30.7%	21.5%	20.1%
Mistral Medium 3	Mistral	37.6%	50.3%	45.8%	38.3%
Kimi K2 Instruct	Moonshot	47.3%	58.2%	50.4%	48.4%
Nemotron Super v1 49B	Nvidia	29.0%	37.9%	36.1%	26.9%
GPT OSS 120B (Thinking = Medium)	OpenAI	50.0%	62.2%	57.7%	58.6%
Qwen 3 235B	Qwen	56.2%	65.7%	61.3%	56.0%
GLM 4.5	Z.ai	52.7%	59.0%	56.6%	52.1%

by bootstrapping $n = 1,000$. Error bars are small for each model, as shown in Figure 5. The percentage of times that models are best performing and worst performing is mostly highly-associated with the other two metrics. Interestingly, Sonnet 4 (Thinking = On), which is a mid-performing model (ranked 14/23 for mean score), is not ranked last on any single task; and DeepSeek R1 (ranked 9/23 for mean score), is ranked 1st the second most of all models.

5.1 Do models find similar cases difficult in APEX-v1.0?

For a small number of cases in APEX-v1.0 nearly all models score 90%+ while, at the other extreme, for a small number of cases nearly all models score under 5%. This reflects our design choices as we wanted APEX to capture the variety of real-world tasks, which differ in complexity and time-to-complete. To understand whether models find the same cases difficult we computed models' pairwise correlations. This accounts for models' different means and standard deviations, allowing us to assess how *relatively* difficult they find cases rather than their absolute performance. Higher correlation indicates that models find the same cases similarly difficult / easy.

As expected, models from the same provider are often highly correlated. o3 (Thinking = High) and o3 Pro (Thinking = High) have correlation of 0.93 (the highest reported), Opus 4.1 (Thinking = On) and Opus 4 (Thinking = On) have correlation of 0.91, Opus 4 (Thinking = On) and Sonnet 4 (Thinking = On) have correlation of 0.88, Gemini 2.5 Flash (Thinking = On) and Gemini 2.5 Pro (Thinking = On) have correlation of 0.82. Interestingly, mean correlation between closed-source models (0.73) is higher than mean correlation between open-source models (0.68). This is somewhat surprising given the latter often share architectures and training methods (Eiras et al., 2024). Phi 4, the lowest-performing model, has lowest mean pairwise correlation at 0.50, indicating orthogonal performance. Surprisingly, the best performing model, GPT 5 (Thinking = High), has a mean pairwise correlation of just 0.65, making it only the 20th most correlated model out of 23.

5.2 Does performance of open source and closed source models differ on APEX-v1.0?

There is a moderate performance gap between the open source and closed source models. The mean score of all closed source models is 55.2% whereas

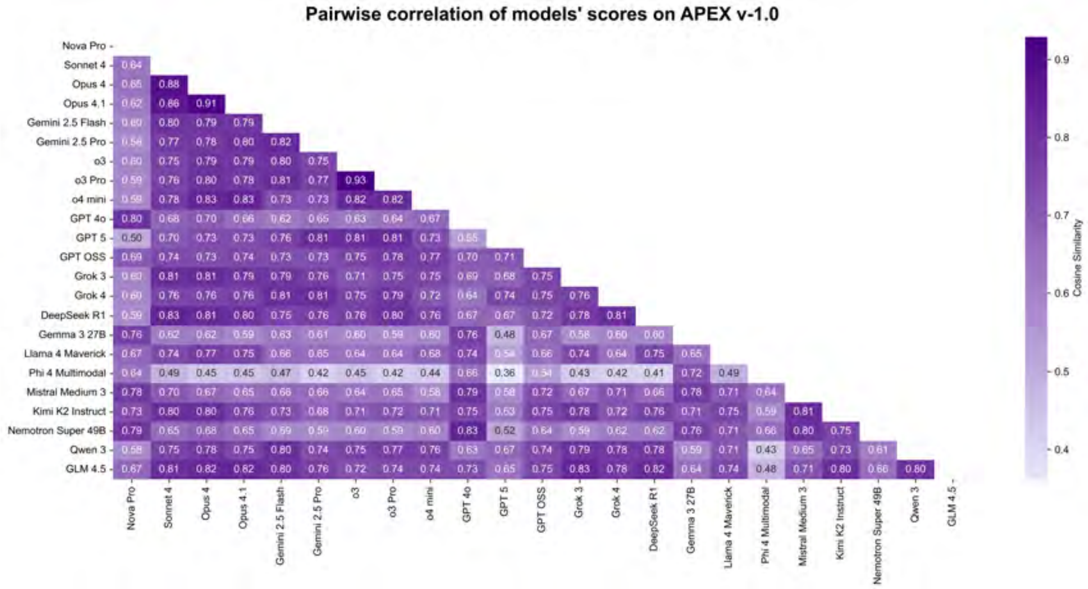


Figure 6: Correlation between pairs of models on APEX-v1.0. Higher values indicate greater similarity. Across the 23 models there are 253 unique pairwise combinations. Values range from 0.39 (for Phi 4 and GPT 5 (Thinking = High)) to 0.93 (for o3 (Thinking = High) and o3 Pro (Thinking = High)).

for open source models it is 45.8%, a drop of 9.4 percentage points. The gap is even larger in terms of pairwise win rates at 57.6% and 40.2%, which is a drop of over 15 percentage points. However, there are some notable exceptions where open source models perform well. Qwen 3 235B is ranked seventh on mean score and DeepSeek R1 is ninth.

5.3 Do more powerful models perform better on APEX-v1.0?

Providers often do not share all details of model training, resulting in a patchwork of information about compute, the number of parameters, and model architecture (Eiras et al., 2024). Given the lack of information, we compare different models from the same family where performance is clearly differentiated by the provider (e.g., o3 pro is pitched by OpenAI as a better version of o3). There are some surprising gaps. Opus 4 (Thinking = On) performs worse than Sonnet 4 (Thinking = On) on all four performance metrics, as shown in Table 3, and beats Sonnet 4 (Thinking = On) in only one domain based on mean score (investment banking). o3 Pro (Thinking = High) outperforms o3 (Thinking = High) by just 0.1% in terms of mean score, and performs worse on the other three metrics. Gemini 2.5 Flash (Thinking = On) is 0.3% better than Gemini 2.5 Pro (Thinking = On) on mean score. It also outperforms in terms of pairwise win rate and the percentage of times ranked first. These results suggest that the more powerful versions of models, which are often more expensive, are not necessarily better at performing the real-world high economic value tasks that comprise APEX-v1.0. There is a small performance improvement across different generations of models within the same family. Grok 4 outperforms Grok 3 on all metrics, Opus 4.1 (Thinking = On) outperforms Opus 4 (Thinking = On), and GPT 5 (Thinking = High) is substantially better than all other OpenAI models, and GPT 4o is substantially worse.

5.4 Do thinking models perform better on APEX-v1.0?

16 of the 23 models offer “thinking” tokens. They perform better than non-thinking models, with mean score of 55.8% (versus 40.4%) and pairwise win rate of 58.9% (versus 29.6%). However, models with thinking are more likely to be closed source and released more recently, which is associated with higher performance and could confound this finding. To reliably assess whether thinking delivers a performance boost on APEX-v1.0 we would need to run ablations with thinking turned on/off or set to high/low (where options are available). This is non-trivial to implement given that “thinking” functionality varies across model providers, as shown in Table 4. Similarly, due to the different information available about models’ architecture

we cannot make reliable claims about parameter scaling.

5.5 Does response length impact scores on APEX-v1.0?

We were concerned that models could achieve high scores on APEX-v1.0 by "scattergunning" – providing very long responses that are hard to read but have enough information to pass many of the criteria. There is some evidence to support this. For instance, qualitative inspection of responses shows that on several tasks both Qwen 3 235B and DeepSeek R1 provide a lot of detail about their thinking process, and are highly repetitive and in places off-topic. Yet both have high mean scores as we do not penalize length, and they provide enough of the correct information to pass many criteria. At the other end of the spectrum, several lower performing models (e.g., GPT 4o, Phi 4 Multimodal, and Nova Pro (Thinking = CoT)) have the shortest mean response lengths (1st, 2nd and 4th shortest, respectively). However, overall, we do not find evidence this is a substantial problem. We fit an Ordinary Least Squares regression model with mean response score as the outcome variable and response length as the only coefficient. The R^2 is only 0.02 (significant at $\alpha = 0.001$), indicating almost zero relationship.

5.6 How does APEX-v1.0 compare with other benchmarks?

The primary benefit of APEX-v1.0 as a benchmark is in evaluating models for well they perform at economically valuable tasks, which remains under-addressed in AI benchmarking. To understand how APEX-v1.0 differs from existing benchmarks, we correlated models' scores with their reported values on four existing benchmarks: Humanity's Last Exam (HLE) (Phan et al., 2025), MMLU (Pro) (Hendrycks et al., 2021), MMMU (Val) (Yue et al., 2024), and GPQA (Rein et al., 2023). Of the 23 models we test on APEX-v1.0, 12 have been publicly graded on HLE, 12 on MMLU Pro, 13 on MMMU (Val) and 22 on GPQA. Mean correlation is 0.79 between models' score on APEX-v1.0 and their score on the other four benchmarks. This compares to 0.58 for MMLU Pro versus the other four benchmarks, 0.75 for HLE, 0.79 for MMMU (Val), and 0.84 for GPQA. This suggests that APEX-v1.0 is approximately as different as the other benchmarks are. Note that the missing model scores limit the robustness of these results.

6 Discussion

6.1 Investment banking

Models' mean scores are second lowest for investment banking (47.6%). It also has the lowest top scoring model of the four domains, with GPT-5 (Thinking = High) at 59.7%. APEX-v1.0 was designed to capture the primary groups of products and sectors that bankers typically work on, with coverage concentrated on buy-side mergers and acquisitions advisory, equity capital markets, valuation methodologies, sponsor-led transactions, and FP&A. However, given the breadth of specialties in banking, certain niche transaction types are not substantially covered, such as sell-side mergers and acquisitions advisory, debt capital markets, project finance, and restructuring or distressed investing. There are particularly significant gaps in tasks requiring non-public data, including debt issuances, refinancings, and private placements. Sector coverage is weighted toward consumer and retail, pharmaceuticals, and technology, media, and telecommunications, with areas such as industrials, aviation, and financial institutions less represented. Future iterations of APEX will expand coverage across products and sectors, with a particular focus on live deal execution tasks—such as using precedent transactions as a valuation cross-check and drafting core transaction materials.

6.2 Management consulting

Models' mean scores are second highest for management consulting (52.5%) and it has the second highest top score model of the four domains, with GPT 5 (Thinking = High) at 64.8%. APEX-v1.0 reflects a broad set of core work in management consulting, covering major practice areas, industries, and analytical approaches, including strategy, operations, finance, and organizational effectiveness. Coverage was strongest in corporate and commercial strategy, financial performance evaluation, operational improvement, and market and competitive intelligence. Specialty areas—such as social impact, ESG and sustainability strategy, and change management—were less represented. Industry coverage was strongest in consumer, technology, finance, and healthcare services, with less representation in utilities, materials, communication services, and energy. Future iterations of APEX will aim for more granular measurement across industries and consulting practices, with deeper representation of implementation-focused and peoplecentric engagements. This will include a richer set of prompts for transformation programs, PMO, and organizational change, as well as more nuanced prompts for innovation, product strategy, and digital enablement. Further development will focus on simulating the full arc of real-world engagements—from diagnostic to recommendation to execution—requiring more judgment, prioritization, and stakeholder alignment alongside technical analysis.

6.3 Law

Models' mean scores are highest for Law (56.9%) and it has the highest top score of the four domains, with GPT 5 (Thinking = High) at 70.5%. APEX-v1.0 covers a diverse set of tasks and legal areas that are of primary importance to practicing lawyers. But since legal tasks and practice areas are so varied, certain tasks and areas were not substantially covered, such as contract redlining, writing regulation drafts, writing a will or a trust, responding to freedom-of-information requests, and assessing whether regulations are lawful. There was also little coverage of specialty areas, such as handling trusts and estates. Feedback from experts suggest the need to distinguish clearly between corporate and civil law, and the work of in-house versus outside counsel, as well as the differing roles of parties and counter-parties. Future iterations of APEX will provide more insight by providing breakdowns for scores by task types and areas of practice, aiming for more systematic and comprehensive coverage. Additionally, we see particular value in providing qualitative assessments of the significance of model scores as our legal partners suspect there are thresholds at which the value of LMs in live work settings will change.

6.4 Medicine

Models' mean scores are lowest for medicine (47.5%) and it has the second lowest top score of the four domains, with GPT 5 (Thinking = High) at 62.0%. The majority of model failures stem from a lack of depth in responses, reflecting an inability to account for the nuances of real-world scenarios. Our medical collaborators note that there is significant room for improvement both in LLM performance and in measuring their true clinical value and efficacy. Tasks in APEX-v1.0 were designed by physicians, focusing on high-impact situations from clinical practice. As a result, the dataset emphasizes emergency department presentations, but also includes broader challenges such as navigating CMS coverage requirements, triaging under resource constraints, and addressing ethical dilemmas involving multiple patients. Future iterations will expand coverage to include more areas such as outpatient care, address "gray-area" decisions (e.g., off-label use), and incorporate a wider range of specialties. Medicine proved to be the most challenging domain for prompt and rubric creation, with the highest levels of internal disagreement between contributors and reviewers. Clinical reviewers also had the lowest tolerance for ambiguity or gaps in reasoning. We caution that clinical value depends on local standards of care, and that greater rigor is required to achieve certainty in this complex domain. Despite strong performance on structured diagnostic benchmarks, our findings show that frontier LLMs still struggle with open-ended clinical reasoning tasks as judged by practicing physicians—and remain far from expert performance.

7 Limitations

Measurement error. There is a risk of measurement error given the challenges of creating rubrics and calibrating annotators, especially as we prioritized hiring annotators with domain-specific industry expertise. We collected model responses three times on each prompt, and saw inter-run variance (see Section 3.2). In future work, we could reduce biases due to inter-run variance by collecting responses more times. Our use of LM judges also introduces another potential measurement error. They are likely to make more errors on complex prompts and criteria in technically-advanced domains, such as medicine. Our analyses and quality control processes suggest this risk is unlikely to substantially change our results. It is worth noting that our testing setup is relatively easy for models to reason over as we pass them the sources they need in context. In a live production setting finding the right information would be much harder, likely requiring a document searching system or other tool usage, and scores would be lower.

Penalizing bad responses. Rubrics in APEX-v1.0 do not have negative criterion that penalize responses for containing incorrect or irrelevant claims. In theory, a response could pass many criteria but also hallucinate, i.e., create factually incorrect claims. This is related to the problem of scattergunning, where responses are far longer than needed. These responses are less valuable to end users than ones that are equally insightful and correct, but far shorter. Although it is not guaranteed that higher scored responses contain fewer errors, our qualitative reviews do not show systematic evidence of wrong claims appearing in otherwise highly scored responses.

Matching real-world value. The scores that models achieve on APEX-v1.0 are not necessarily correlated with real-world value. Value is often stage-gated, which means that a model scoring 60% does not have 60% of the value that one scoring 100% delivers. Instead a response that scores 60% might be effectively useless. On the other hand, if a strong *product* is built on top of a relatively weak model, possibly one that clearly communicates what it does not know, it can still be incrementally valuable to users. Ultimately, value depends on how AI models are used and in what context. In the future, as models improve, we aim to integrate assessing cost and latency as well as performance.

Saturation and contamination Long-term, models could saturate performance on APEX-v1.0, making it less useful as a benchmark to guide training efforts. We see this as an exciting development and is actually our main ambition with APEX – providing they are not overfitting, it would mean that models are delivering far more economic value. At the same time, we note that the best performing model, GPT 5 (Thinking = High), only achieves 64.2% on APEX-v1.0 despite scoring very highly on other benchmarks such as AIME 2025 without tools (94.6%), SWE-bench Verified (74.9%) and MMMU (84.2%).[4] This indicates substantial headroom to improve due to the difficulty of APEX-v1.0. A related concern is that models saturate performance because the data becomes contaminated rather than because they have improved their capabilities. This happens when model providers pre- or post-train on a dataset, possibly unintentionally. We have kept APEX-v1.0 as a heldout hidden set to minimize this risk but we acknowledge that it cannot be fully mitigated, especially as we have publicly shared details on its design and creation.

[4] https://openai.com/index/introducing-gpt-5/

8 Future expansions of APEX

Scope and coverage. APEX-v1.0 contains evals for four important knowledge-intensive jobs. Future iterations of APEX should include a wider range of roles, as well as more granular measurement of performance withing common workflows and specialties. We see software engineering, teaching, insurance, and graphic design as promising new roles to benchmark.

Tool use and data rooms. The evals in APEX-v1.0 can be extended to more closely mimic the day-to-day activities of knowledge workers. This includes adding software and tooling, knowledge stores for large volumes of documents, and multi-turn interactions with an LM live. In particular, we see benefits in creating data rooms that contain all of the relevant files and resources for a set of prompts. A single data room could support multiple tasks that a professional might undertake.

Loss analysis with criteria and prompt tags. We see a lot of potential in applying finegrained tags to the data in APEX. Each criterion could be tagged for secondary information, such as importance, cross-criteria dependencies, the type of model behavior it relates to (e.g., reasoning, stating information, generation and instruction following), and domain-specific topics. Similarly, prompts can be tagged for their complexity, impact, workflow, and the type of tools they require. These tags can be used to run loss analyses, providing granular insight into models' strengths and weaknesses. They could also be used to weight rewards during training.

9 Related work

Numerous benchmarks have been developed to test the reasoning, tool use, instruction following, and generative capabilities of LMs. GAIA introduces a benchmark of 466 real-world multi-step tasks that require advanced capabilities such as reasoning, tool use (e.g. web browsing, coding), and multi-modal comprehension to be solved (Mialon et al., 2023). Answers are generally unambiguous, enabling high-quality scoring. MMLU tests models on general knowledge and reasoning across 57 academic and professional subjects using multiple-choice questions (Hendrycks et al., 2021). It is widely used to measure general knowledge and reasoning, but has been criticized for data errors

(Gema et al., 2025) and saturation (Wang et al., 2024), especially as the data is open-source. GPQA is a dataset of 448 expert-crafted, graduate-level multiple-choice questions in biology, physics, and chemistry that even PhD-level experts achieve only approximately 65% (Rein et al., 2023). This high-quality resource only has multiple choice questions, and is focused on academic performance rather than real-world value. Similarly, Humanity's Last Exam (HLE) is a benchmark of expert-level, high-stakes questions from professional and academic domains, created by PhDs in each respective field (Phan et al., 2025). It also lacks focus on economically-valuable outputs. GSM8K is a less challenging but widely used dataset. It contains high-quality, grade-school math word problems that need to be solved step-by-step (Cobbe et al., 2021), and is designed to evaluate models' arithmetic reasoning. Many other question answering datasets evaluate model's knowledge and reasoning capabilities in specific domains, like FinanceBench (Islam et al., 2023), PubMedQA (Jin et al., 2019) and ScienceQA (Lu et al., 2022). There are also abstract reasoning benchmarks such as ARC-1 (Clark et al., 2018) and ARC-2 (Chollet et al., 2025), which are part of the AI2 Reasoning Challenge. ARC contains visual reasoning tasks, assessing models' ability to perform complex reasoning without explicit instructions.

AI models' ability to improve workers' productivity is a growing area of interest in research. OpenAI's GDPval is a benchmark for testing AI capabilities at performing real world economically valuable tasks (OpenAI, 2025). It covers 44 occupations across the top 9 sectors contributing to U.S. Gross Domestic Product. Industry professionals created the prompts and rubrics, and GPDval is one of the first benchmarks to reflect real-world high-value tasks. The authors open source 220 gold cases and show that for some tasks frontier models are approaching industry experts' quality of work. Anthropic's Economic Index (Handa et al., 2025) aims to understand the impact of AI on the economy over time. An open source project, it uses real data from Claude to track where and how AI is used, focusing on broad groupings like "Computer & Mathematical" and "Business & Financial". It shows that partial automation, and use of AI to augment human workers, is the dominant paradigm of AI use. Evaluations of AI models' ability to perform knowledge is mixed. VendingBench (Backlund and Petersson, 2025) tests LLM-based agents' ability to operate a simulated vending machine. Agents have access to tools and must execute simple tasks (e.g., placing orders), managing workload over a long time horizon. Models are capable of returning a profit in many runs, but sometimes derail by losing track of the task or making poor decisions and losing money. TheAgentCompany is a benchmark for evaluating agents' performance at the tasks of a digital worker, where they can browse the Web, write code, run programs, and communicate with other coworkers. They find that the best agent can complete 30% of tasks autonomously (Xu et al., 2025). Hendrix et al. (2022) look at use of AI among clinicians and argue that while it could increase productivity, it will increase workloads if badly implemented, and needs to be used for suitable tasks. Becker et al. (2025) run a controlled study on 16 developers with moderate AI experience. Participants complete 246 tasks in mature projects on which they have a mean of 5 years of prior experience. Although participants anticipate that AI will reduce completion time by 24%, the authors found it increased completion time by 19%. This finding challenges other research that shows AI coding assistants can generate a significant uplift in output and speed (Peng et al., 2023; Paradis et al., 2024). Miserendino et al. (2025) introduce SWE-Lancer, a benchmark of over 1,400 freelance software engineering tasks from Upwork, valued at 1 million dollars. Tasks vary from simple bug fixes to complex tasks involving advanced reasoning and management. Three frontier models earn between 30% (o3) and 40% (Claude 3.5 Sonnet) of the money available.

10 Acknowledgments

We are deeply grateful to all of the expert annotators who took contributed to APEX-v1.0. We thank Nitin Nohria, Larry Summers, Paul Röttger, Scott Hale, Jonathan Frankle, Nathan Lambert, Akul Datta, Asher Spector, Ryan Hu, Nolan Brown, Apara Sivaraman, Supriya Shah, Kailash Sundaram, Shaun VanWeelden, and Tara Razeai for their feedback and help on the design, implementation and analysis.

References

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-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.

Axel Backlund and Lukas Petersson. 2025. Vending-bench: A benchmark for long-term coherence of autonomous agents.

Oren Bar-Gill and Cass R. Sunstein. 2025. Algorithmic harm: Protecting people in the age of artificial intelligence. Working Paper 25-21, Harvard Public Law Working Paper Series. Available at SSRN.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using llms for text annotation.

Joel Becker, Nate Rush, Elizabeth Barnes, and David Rein. 2025. Measuring the impact of early-2025 ai on experienced open-source developer productivity.

David Cahn. 2024. Ai's $600b question: The growing gap between investment and revenue.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. 2025. Arc-agi-2: A new challenge for frontier ai reasoning systems.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems.

Francisco Eiras, Aleksandar Petrov, Bertie Vidgen, Christian Schroeder, Fabio Pizzati, Katherine Elkins, Supratik Mukhopadhyay, Adel Bibi, Aaron Purewal, Csaba Botos, Fabro Steibel, Fazel Keshtkar, Fazl Barez, Genevieve Smith, Gianluca Guadagni, Jon Chun, Jordi Cabot, Joseph Imperial, Juan Arturo Nolazco, Lori Landay, Matthew Jackson, Phillip H. S. Torr, Trevor Darrell, Yong Lee, and Jakob Foerster. 2024. Risks and opportunities of open-source generative ai.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile van Krieken, and Pasquale Minervini. 2025. Are we done with mmlu?

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A survey on llm-as-a-judge.

Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. 2025. Which economic tasks are performed with ai? evidence from millions of claude conversations.

Nathaniel Hendrix, David L. Veenstra, Mindy Cheng, Nicholas C. Anderson, and Stéphane Verguet. 2022. Assessing the economic value of clinical artificial intelligence: Challenges and opportunities. *Value in Health*, 25(3):331–339.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants.

Samuel Miserendino, Michele Wang, Tejal Patwardhan, and Johannes Heidecke. 2025. Swe-lancer: Can frontier llms earn 1 million from real-world freelance software engineering?

OpenAI. 2025. Gdpval: Evaluating ai's economic value. Accessed: 2025-09-29.

Elise Paradis, Kate Grey, Quinn Madison, Daye Nam, Andrew Macvean, Vahid Meimand, Nan Zhang, Ben Ferrari-Church, and Satish Chandra. 2024. How much does ai impact development speed? an enterprise-based randomized controlled trial.

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The impact of ai on developer productivity: Evidence from github copilot.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala, Noah Burns, Gashaw M. Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G. Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B. Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G. Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod, Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad Hogg, Carlo Bosio, Brian P Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Kelsey Van den Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever, Alexei Kopylov, Benjamin Myklebust, Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong Yang, John Maar, Julian Wykowski, Martí Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Ariel Ghislain Kemogne Kamdoum, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Gongbo Sun, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Joseph M Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan, Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes, Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser, Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheswaran, Adam Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, Josef Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T. Ryan Rogers, Tom Goertzen, Declan Grabb, Abhishek Shukla, Alan Givré, John Arnold Ambay, Archan Sen, Muhammad Fayez Aziz, Mark H Inlow, Hao He, Ling Zhang, Younesse Kaddar, Ivar Ängquist, Yanxu Chen, Harrison K Wang, Kalyan Ramakrishnan, Elliott Thornley, Antonio Terpin, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Martin Stehberger, Peter Bradshaw, JP Heimonen, Kaustubh Sridhar, Ido Akov, Jennifer Sandlin, Yury Makarychev, Joanna Tam, Hieu Hoang, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu, Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K. Cohen, Orr Paradise, Jan Hendrik Kirchner, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Michael Wang, Yuzhou Nie, Anna Sztyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani, Shreyas Verma, Prashant Joshi, Eli Meril, Ziqiao Ma, Jérémy Andréoletti, Raghav Singhal, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Marco Piccardo, Hamid Mostaghimi, Qijia Chen, Virendra Singh, Tran Quoc Khánh, Paul Rosu, Hannah Szlyk, Zachary Brown, Himanshu Narayan, Aline Menezes, Jonathan Roberts, William Alley, Kunyang Sun, Arkil Patel, Max Lamparth, Anka Reuel, Linwei Xin, Hanmeng Xu, Jacob Loader, Freddie Martin, Zixuan Wang, Andrea Achilleos, Thomas Preu, Tomek Korbak, Ida Bosio, Fereshteh Kazemi, Ziye Chen, Biró Bálint, Eve J. Y. Lo, Jiaqi Wang, Maria Inês S. Nunes, Jeremiah Milbauer, M Saiful Bari, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Hossam Elgnainy, Guillaume Douville, Daniel Tordera, George Balabanian, Hew Wolff, Lynna Kvistad, Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Andrew Favre D. O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Sherwin Abdoli, Tim Santens, Shaul Barkan, Allison Tee, Robin Zhang, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Jiayi Pan, Emma Rodman, Jacob Drori, Carl J Fossum, Niklas Muennighoff, Milind Jagota, Ronak Pradeep, Honglu Fan, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Moritz Firsching, Carter Harris, Stefan Ciobâcă, Jason Gross, Rohan Pandey, Ilya Gusev, Adam Jones, Shashank Agnihotri, Pavel Zhelnov, Mohammadreza Mofayezi, Alexander Piperski, David K. Zhang, Kostiantyn Dobarskyi, Roman Leventov, Ignat Soroko, Joshua Duersch, Vage Taamazyan, Andrew Ho, Wenjie Ma, William Held, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska, Claudio Di Fratta, Edson Oliveira, Joseph W. Jackson, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Bita Golshani, David Stap, Egor Kretov, Mikalai

Uzhou, Alina Borisovna Zhidkovskaya, Nick Winter, Miguel Orbegozo Rodriguez, Robert Lauff, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Fortuna Samuele, Fredrik Ekström, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Rayner Hernandez Perez, Daniel Pyda, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Kenchi Okutsu, Mike Battaglia, Mohammad Maghsoudimehrabani, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Handoko, Anton Peristyy, Stephen Malina, Mustafa Mehkary, Rami Aly, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Mukhwinder Singh, Hassan Shapourian, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Harsh Kumar, Chiara Ceconello, Chao Zhuang, Haon Park, Micah Carroll, Andrew R. Tawfeek, Stefan Steinerberger, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Jainam Shah, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Abram Jackson, Paolo Giordano, Philipp Petersen, Adrian Cosma, Jesus Colino, Colin White, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Petr Spelda, Vit Stritecky, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Koen Sponselee, Renas Bacho, Zheng-Xin Yong, Florencia de la Rosa, Nathan Cho, Xiuyu Li, Guillaume Malod, Orion Weller, Guglielmo Albani, Leon Lang, Julien Laurendeau, Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Julien Degorre, Yiğit Yalın, Gbenga Daniel Obikoya, Rai, Filippo Bigi, M. C. Boscá, Oleg Shumar, Kaniuar Bacho, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Thomas C. H. Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu, Liu, Stefano Cavalleri, Olle Häggström, Emil Verkama, Joshua Newbould, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Ting Wang, Yosi Kratish, Wen-Ding Li, Sivakanth Gopi, Andrea Caciolai, Christian Schroeder de Witt, Pablo Hernández-Cámara, Emanuele Rodolà, Jules Robins, Dominic Williamson, Vincent Cheng, Brad Raynor, Hao Qi, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Christoph Demian, Peyman Kassani, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk Bodur, D. P. Shinde, Yan Carlos Leyva Labrador, Zahra Adoul, Mohamed Zekry, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Earth Anderson, Rodrigo De Oliveira Pena, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Ross Finocchio, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Isaac C. McAlister, Alejandro José Moyano, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Yana Malysheva, Daphiny Pottmaier, Omid Taheri, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M. R. Minissi, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Ronald Clark, Josh Ducey, Matheus Piza, Maja Somrak, Eric Vergo, Juehang Qin, Benjámin Borbás, Eric Chu, Jack Lindsey, Antoine Jallon, I. M. J. McInnis, Evan Chen, Avi Semler, Luk Gloor, Tej Shah, Marc Carauleanu, Pascal Lauer, Tran Đuc Huy, Hossein Shahrtash, Emilien Duc, Lukas Lewark, Assaf Brown, Samuel Albanie, Brian Weber, Warren S. Vaz, Pierre Clavier, Yiyang Fan, Gabriel Poesia Reis e Silva, Long, Lian, Marcus Abramovitch, Xi Jiang, Sandra Mendoza, Murat Islam, Juan Gonzalez, Vasilios Mavroudis, Justin Xu, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong, Thorben Jansen, Antonella Pinto, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh, Tong Jiang, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Gang Zhang, Zhehang Du, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D. Hartman, Tim Tarver, Stephen Mensah, Gautier Abou Loume, Wiktor Morak, Farzad Habibi, Sarah Hoback, Will Cai, Javier Gimenez, Roselynn Grace Montecillo, Jakub Łucki, Russell Campbell, Asankhaya Sharma, Khalida Meer, Shreen Gul, Daniel Espinosa Gonzalez, Xavier Alapont, Alex Hoover, Gunjan Chhablani, Freddie Vargus, Arunim Agarwal, Yibo Jiang, Deepakkumar Patil, David Outevsky, Kevin Joseph Scaria, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Ashley Cartwright, Sergei Bogdanov, Niels Mündler, Sören Möller, Luca Arnaboldi, Kunvar Thaman, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Tony Fruhauff, Glen Sherman, Mátyás Vincze, Siranut Usawasutsakorn, Dylan Ler, Anil Radhakrishnan, Innocent Enyekwe, Sk Md Salauddin, Jiang Muzhen, Aleksandr Maksapetyan, Vivien Rossbach, Chris Harjadi, Mohsen Bahaloohoreh, Claire Sparrow, Jasdeep Sidhu, Sam Ali, Song Bian, John Lai, Eric Singer, Justine Leon Uro, Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Dario Bezzi, Yashaswini Jain, Ashley Aaron, Murat Tiryakioglu, Sheeshram Siddh, Keith Krenek, Imad Ali Shah, Jun Jin, Scott Creighton, Denis Peskoff, Zienab EL-Wasif, Ragavendran P V, Michael Richmond, Joseph McGowan, Tejal Patwardhan, Hao-Yu Sun, Ting Sun, Nikola Zubić, Samuele Sala, Stephen Ebert, Jean Kaddour, Manuel Schottdorf, Dianzhuo Wang, Gerol Petruzella, Alex Meiburg, Tilen Medved, Ali ElSheikh, S Ashwin Hebbar, Lorenzo Vaquero, Xianjun Yang, Jason Poulos, Vilém Zouhar, Sergey Bogdanik, Mingfang Zhang, Jorge Sanz-Ros, David Anugraha, Yinwei Dai, Anh N. Nhu, Xue Wang, Ali Anil Demircali, Zhibai Jia, Yuyin Zhou, Juncheng Wu, Mike He, Nitin Chandok, Aarush Sinha, Gaoxiang Luo, Long Le, Mickaël Noyé, Michał Perełkiewicz, Ioannis Pantidis, Tianbo Qi, Soham Sachin Purohit, Letitia Parcalabescu, Thai-Hoa Nguyen, Genta Indra Winata, Edoardo M. Ponti, Hanchen Li, Kaustubh Dhole, Jongee Park, Dario Abbondanza, Yuanli Wang, Anupam Nayak, Diogo M. Caetano, Antonio A. W. L. Wong, Maria del Rio-Chanona, Dániel Kondor, Pieter Francois, Ed Chalstrey, Jakob Zsambok, Dan Hoyer, Jenny Reddish, Jakob Hauser, Francisco-Javier Rodrigo-Ginés, Suchandra Datta, Maxwell Shepherd, Thom Kamphuis, Qizheng Zhang, Hyunjun Kim, Ruiji Sun, Jianzhu Yao, Franck Dernoncourt, Satyapriya Krishna, Sina Rismanchian,

Bonan Pu, Francesco Pinto, Yingheng Wang, Kumar Shridhar, Kalon J. Overholt, Glib Briia, Hieu Nguyen, David, Soler Bartomeu, Tony CY Pang, Adam Wecker, Yifan Xiong, Fanfei Li, Lukas S. Huber, Joshua Jaeger, Romano De Maddalena, Xing Han Lù, Yuhui Zhang, Claas Beger, Patrick Tser Jern Kon, Sean Li, Vivek Sanker, Ming Yin, Yihao Liang, Xinlu Zhang, Ankit Agrawal, Li S. Yifei, Zechen Zhang, Mu Cai, Yasin Sonmez, Costin Cozianu, Changhao Li, Alex Slen, Shoubin Yu, Hyun Kyu Park, Gabriele Sarti, Marcin Briański, Alessandro Stolfo, Truong An Nguyen, Mike Zhang, Yotam Perlitz, Jose Hernandez-Orallo, Runjia Li, Amin Shabani, Felix Juefei-Xu, Shikhar Dhingra, Orr Zohar, My Chiffon Nguyen, Alexander Pondaven, Abdurrahim Yilmaz, Xuandong Zhao, Chuanyang Jin, Muyan Jiang, Stefan Todoran, Xinyao Han, Jules Kreuer, Brian Rabern, Anna Plassart, Martino Maggetti, Luther Yap, Robert Geirhos, Jonathon Kean, Dingsu Wang, Sina Mollaei, Chenkai Sun, Yifan Yin, Shiqi Wang, Rui Li, Yaowen Chang, Anjiang Wei, Alice Bizeul, Xiaohan Wang, Alexandre Oliveira Arrais, Kushin Mukherjee, Jorge Chamorro-Padial, Jiachen Liu, Xingyu Qu, Junyi Guan, Adam Bouyamourn, Shuyu Wu, Martyna Plomecka, Junda Chen, Mengze Tang, Jiaqi Deng, Shreyas Subramanian, Haocheng Xi, Haoxuan Chen, Weizhi Zhang, Yinuo Ren, Haoqin Tu, Sejong Kim, Yushun Chen, Sara Vera Marjanović, Junwoo Ha, Grzegorz Luczyna, Jeff J. Ma, Zewen Shen, Dawn Song, Cedegao E. Zhang, Zhun Wang, Gaël Gendron, Yunze Xiao, Leo Smucker, Erica Weng, Kwok Hao Lee, Zhe Ye, Stefano Ermon, Ignacio D. Lopez-Miguel, Theo Knights, Anthony Gitter, Namkyu Park, Boyi Wei, Hongzheng Chen, Kunal Pai, Ahmed Elkhanany, Han Lin, Philipp D. Siedler, Jichao Fang, Ritwik Mishra, Károly Zsolnai-Fehér, Xilin Jiang, Shadab Khan, Jun Yuan, Rishab Kumar Jain, Xi Lin, Mike Peterson, Zhe Wang, Aditya Malusare, Maosen Tang, Isha Gupta, Ivan Fosin, Timothy Kang, Barbara Dworakowska, Kazuki Matsumoto, Guangyao Zheng, Gerben Sewuster, Jorge Pretel Villanueva, Ivan Rannev, Igor Chernyavsky, Jiale Chen, Deepayan Banik, Ben Racz, Wenchao Dong, Jianxin Wang, Laila Bashmal, Duarte V. Gonçalves, Wei Hu, Kaushik Bar, Ondrej Bohdal, Atharv Singh Patlan, Shehzaad Dhuliawala, Caroline Geirhos, Julien Wist, Yuval Kansal, Bingsen Chen, Kutay Tire, Atak Talay Yücel, Brandon Christof, Veerupaksh Singla, Zijian Song, Sanxing Chen, Jiaxin Ge, Kaustubh Ponkshe, Isaac Park, Tianneng Shi, Martin Q. Ma, Joshua Mak, Sherwin Lai, Antoine Moulin, Zhuo Cheng, Zhanda Zhu, Ziyi Zhang, Vaidehi Patil, Ketan Jha, Qiutong Men, Jiaxuan Wu, Tianchi Zhang, Bruno Hebling Vieira, Alham Fikri Aji, Jae-Won Chung, Mohammed Mahfoud, Ha Thi Hoang, Marc Sperzel, Wei Hao, Kristof Meding, Sihan Xu, Vassilis Kostakos, Davide Manini, Yueying Liu, Christopher Toukmaji, Jay Paek, Eunmi Yu, Arif Engin Demircali, Zhiyi Sun, Ivan Dewerpe, Hongsen Qin, Roman Pflugfelder, James Bailey, Johnathan Morris, Ville Heilala, Sybille Rosset, Zishun Yu, Peter E. Chen, Woongyeong Yeo, Eeshaan Jain, Ryan Yang, Sreekar Chigurupati, Julia Chernyavsky, Sai Prajwal Reddy, Subhashini Venugopalan, Hunar Batra, Core Francisco Park, Hieu Tran, Guilherme Maximiano, Genghan Zhang, Yizhuo Liang, Hu Shiyu, Rongwu Xu, Rui Pan, Siddharth Suresh, Ziqi Liu, Samaksh Gulati, Songyang Zhang, Peter Turchin, Christopher W. Bartlett, Christopher R. Scotese, Phuong M. Cao, Aakaash Nattanmai, Gordon McKellips, Anish Cheraku, Asim Suhail, Ethan Luo, Marvin Deng, Jason Luo, Ashley Zhang, Kavin Jindel, Jay Paek, Kasper Halevy, Allen Baranov, Michael Liu, Advaith Avadhanam, David Zhang, Vincent Cheng, Brad Ma, Evan Fu, Liam Do, Joshua Lass, Hubert Yang, Surya Sunkari, Vishruth Bharath, Violet Ai, James Leung, Rishit Agrawal, Alan Zhou, Kevin Chen, Tejas Kalpathi, Ziqi Xu, Gavin Wang, Tyler Xiao, Erik Maung, Sam Lee, Ryan Yang, Roy Yue, Ben Zhao, Julia Yoon, Sunny Sun, Aryan Singh, Ethan Luo, Clark Peng, Tyler Osbey, Taozhi Wang, Daryl Echeazu, Hubert Yang, Timothy Wu, Spandan Patel, Vidhi Kulkarni, Vijaykaarti Sundarapandiyan, Ashley Zhang, Andrew Le, Zafir Nasim, Srikar Yalam, Ritesh Kasamsetty, Soham Samal, Hubert Yang, David Sun, Nihar Shah, Abhijeet Saha, Alex Zhang, Leon Nguyen, Laasya Nagumalli, Kaixin Wang, Alan Zhou, Aidan Wu, Jason Luo, Anwith Telluri, Summer Yue, Alexandr Wang, and Dan Hendrycks. 2025. Humanity's last exam.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark.

Jon Saad-Falcon, Rajan Vivek, William Berrios, Nandita Shankar Naik, Matija Franklin, Bertie Vidgen, Amanpreet Singh, Douwe Kiela, and Shikib Mehri. 2024. Lmunit: Fine-grained evaluation with natural language unit tests.

Reva Schwartz, Rumman Chowdhury, Akash Kundu, Heather Frase, Marzieh Fadaee, Tom David, Gabriella Waters, Afaf Taik, Morgan Briggs, Patrick Hall, Shomik Jain, Kyra Yee, Spencer Thomas, Sundeep Bhandari, Paul Duncan, Andrew Thompson, Maya Carlyle, Qinghua Lu, Matthew Holmes, and Theodora Skeadas. 2025. Reality check: A new evaluation ecosystem is necessary to understand ai's real world effects.

Giulio Starace, Oliver Jaffe, Dane Sherburn, James Aung, Jun Shern Chan, Leon Maksin, Rachel Dias, Evan Mays, Benjamin Kinsella, Wyatt Thompson, Johannes Heidecke, Amelia Glaese, and Tejal Patwardhan. 2025. Paperbench: Evaluating ai's ability to replicate ai research.

Eric Topol. 2019. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*, 1st edition. Basic Books, Inc., USA.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark.

Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. 2025. Toward an evaluation science for generative ai systems.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. 2025. Theagentcompany: Benchmarking llm agents on consequential real world tasks.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. Judgelm: Fine-tuned large language models are scalable judges.

A Models Under Test

We parse the PDFs and CSVs of the evidence sources using Reducto and append to the prompt in the context window.[5] We did not use the native file handling functionalities in closed source models' respective APIs as early testing shows it can massively increase the total number of tokens. By using a single file parsing strategy we also create a fair evaluation setting for all models. Without the evidence sources it is very difficult for a model to give a high-quality response. As such, we believe that models can only be fairly graded against APEX-v1.0 if passed the full set of documents, and will otherwise have deflated and uninformative scores.

B Judge LM

Implementation details: o3 has thinking turned on by default and temperature cannot be set. We set thinking to Low. For Gemini 2.5 Pro and Sonnet 4 we leave thinking turned off and set temperature to 0.01, minimizing the risk of intra-model inconsistency.

Prompt: You are evaluating a model-generated response against a specific criterion. Your task is to determine if the response satisfies this criterion and provide a concise explanation.

Criterion to evaluate: `evaluation_case`
Response to evaluate: `model_response`

Instructions:

1. First, analyze the response against the given criterion.
2. Determine if the response fully satisfies the criterion (`result = 1`) or not (`result = 0`).
3. Provide a concise explanation (maximum 2–3 sentences) that:
   - States whether the criterion is met or not,
   - Points to specific evidence from the response,
   - Avoids unnecessary details or repetition.

Return your evaluation in the following JSON format:

```
{
"result": <1 or 0>,
"reason": "<concise explanation>"
}
```

Keep your explanation brief and focus on the key points that justify your result.

[5] https://docs.reducto.ai/overview

Table 4: Overview of the Models Under Test based on their public documentation. Note that we do not set top_p because we set temperature (where possible).

Provider	Model	Release	Temperature	Thinking settings	Context window
Amazon	Nova Pro	Closed	0.7	On. Set in the System prompt	300,000
Anthropic	Opus 4.1	Closed	Disabled	On. Num tokens configurable	200,000
Anthropic	Opus 4	Closed	Disabled	On. Num tokens configurable	200,000
Anthropic	Sonnet 4	Closed	Disabled	On. Num tokens configurable	200,000
Google	Gemini 2.5 Flash	Closed	0.7	On. Num tokens configurable	1,048,576
Google	Gemini 2.5 Pro	Closed	0.7	On. Num tokens configurable	1,048,576
OpenAI	GPT 5	Closed	Disabled	On. Setting: High	400,000
OpenAI	GPT 4o	Closed	0.7	Not available	128,000
OpenAI	o3	Closed	Disabled	On. Setting: High	200,000
OpenAI	o3 Pro	Closed	Disabled	On. Setting: High	200,000
OpenAI	o4 mini	Closed	Disabled	On. Setting: High	200,000
xAI	Grok 3	Closed	0.7	Not available	131,072
xAI	Grok 4	Closed	0.7	On by default	256,000
DeepSeek	DeepSeek R1 0528	Open	0.7	On by default	163,840
Google	Gemma 3 27B	Open	0.7	Not available	131,072
Meta	Llama 4 Maverick (I)	Open	0.7	Not available	1,048,576
Microsoft	Phi 4 Multimodal (I)	Open	0.7	Not available	131,072
MoonshotAI	Kimi K2 Instruct	Open	0.6	Not available	128,000
Nvidia	Nemotron Super v1 49B	Open	0.6	On by default	131,072
OpenAI	GPT OSS 120B	Open	0.7	On. Setting: Medium	128,000
Qwen	Qwen 3 235B-A22B	Open	0.6	On by default	262,144
Mistral	Mistral Medium 3	Open	0.7	On by default	131,072
Z.ai	GLM 4.5	Open	0.7	On by default	131,072

C Workflows reported by experts in each domain

We surveyed all contributing experts before about their day-to-day tasks. The activities reported in Table 5, Table 6, Table 7, Table 8, and Table 9 are an estimate based on the experts' responses. We split law into litigation and corporate law given substantial differences in day-to-day work.

D Example prompts and rubrics for each domain

The four example prompts and rubrics given here are not in the APEX-v1.0 heldout set but they were created by the same group of experts with the same set of instructions. We provide one example from each domain.

Table 5: Law Workflows – Corporate

Workflow	Description of workflow	Time allocation
Contract Drafting & Negotiation	Draft, review, and negotiate transaction contracts	30%
Due Diligence & Document Review	Review corporate records to identify legal risks	20%
Regulatory Compliance & Legal Research	Research regulations and craft compliance advice	15%
Client Advising & Communication	Provide written advice and updates to clients	15%
Transaction Management & Filings	Coordinate closings and prepare official filings	10%
Administrative & Training Duties	Conduct meetings and workshops to collect inputs and present interim results	10%

Table 6: Law Workflows – Litigation

Workflow	Description of workflow	Time allocation
Legal Research & Memo Writing	Research precedent and draft legal memoranda	30%
Drafting Pleadings & Filings	Draft complaints, motions, and briefs for court	25%
Discovery & Evidence Review	Review and categorize documents during discovery	15%
Court Appearances & Trial Prep	Prepare for and attend hearings, depositions, and trials	15%
Client Communication & Strategy	Update clients and refine litigation strategy	10%
Administrative & Professional Duties	Manage billing, case files, and CLE requirements	5%

Table 7: Medicine Workflows

Workflow	Description of workflow	Time allocation
Documentation & Charting	Record encounters and complete required medical documentation	25%
Patient Consultation & Diagnosis	Conduct patient interviews and exams to reach diagnoses	20%
Treatment Planning & Management	Develop and adjust treatment plans and prescriptions	20%
Procedures & Acute Interventions	Perform clinical procedures and urgent care interventions	15%
Patient Communication & Coordination	Communicate results and coordinate with care team	10%
Administrative Tasks	Handle insurance forms, scheduling, and staff supervision	5%
Continuing Education & Research	Engage in CME, literature review, and clinical research	5%

Table 8: Investment Banking Workflows

Workflow	Description of workflow	Time allocation
Financial Modeling & Valuation	Create valuation analyses and financial projections supporting transactions	30%
Client Pitch Material Writing	Draft written client pitch materials describing transaction ideas and market context	25%
Industry Research & Due Diligence	Research sectors and targets; compile diligence notes from public sources	20%
Deal Execution & Documentation	Coordinate transactions and prepare deal documents	15%
Internal Memos & Committee Materials	Write internal approval memos summarizing opportunity and risk	5%
Client Communication & Meetings	Provide written updates and correspondence to clients	5%

Table 9: Consulting Workflows

Workflow	Description of workflow	Time allocation
Report & Memo Writing	Written reports and memos that convey findings and guidance to clients	25%
Data Analysis & Modeling	Evaluating data sets and building models to quantify business impacts	25%
Market & Competitive Research	Gather and synthesize market, competitor, and trend information	15%
Strategy Formulation & Business Case	Develop strategic options and business cases with projected outcomes	15%
Project Management & Coordination	Plan tasks, timelines, and coordination across team and client	10%
Client Meetings & Workshops	Conduct meetings and workshops to collect inputs and present interim results	5%
Proposal Development & Thought Leadership	Create proposals and thought-leadership content to win work and share insights	5%

Table 10: **Law (ID 1045)**. "A client approached our firm in June 2025 concerning an estate issue. The client is the sole heir (and the living spouse) of a musician who died in 2007. Before her death, the musician released three albums to critical acclaim. In her will, the musician left behind all her assets to the client. The musician's three albums were released in 1989, 1990, and 1995. Prior to recording these albums, on December 31, 1988, the musician entered into a single contract with Warner Music Group (the "Company" or "WMG"). The contract sets out the musician's and the Company's rights to the musician's three subsequently-released albums. Importantly for present purposes, the contract contains the following two clauses, both of which were heavily negotiated: Clause A (Independent Status): "For all purposes, including international tax and liability, the Artist* shall be considered an independent contractor. Nothing in this agreement shall be construed to create a partnership, joint venture, or employer/employee relationship." Clause B (Work for Hire): "The parties expressly agree that all sound recordings created hereunder shall be considered 'works made for hire' as defined in 17 U.S.C. § 101, with the Company (WMG) being deemed the sole author of the works in perpetuity. This stipulation is a material inducement for the Company entering into this agreement." As it is used in the contract, "Artist" refers to the musician. As set out above, the agreement stipulated that the musician would not be considered an employee of WMG and that the musician would assign the recording copyrights to all albums released between 1988 and 1999 to WMG. The agreement further stipulated that the albums would be considered "works made for hire." The client wants to know whether or not he owns the copyrights over the sound recordings of the musician's three albums. If not, the client would like to know if he can ever regain ownership over the copyrighted sound recordings and if so, how. Write a legal research memo of no more than 1,500 words that answers the client's questions. Assume that (1) the musician meets the definition of an independent contractor under the relevant agency test established in Community for Creative Non-Violence v. Reid, 490 U.S. 730 (1989); (2) any argument that the albums should be considered compilations is invalid; (3) the albums were made solely by the musician (i.e., were not joint works with another artist); and (4) the agreement does not cover the right of publication with respect to the albums. The analysis in your memo should reflect the state of the law as of Sunday, June 8, 2025. Confine your legal research to the uploaded sources. Do not rely on outside sources. Your memo must cite every authority and source on which it relies. Every citation should be in Bluebook format."

Criterion	Description
1	Styles the work product as a legal memorandum.
2	Ensures that the memorandum does not exceed 1,500 words.
3	States that copyright ownership vests initially in the statutorily-defined "author" of the original work.
4	States that the person who creates the work is its author unless the work was made for hire as defined by 17 U.S.C. § 101, in which case the employer or person whom the work was prepared for is considered the author.
5	States that, under 17 U.S.C. § 101, there are two ways in which a work may be created as a work made for hire: (1) if it is created by an employee acting within the scope of his or her employment; or (2) work-made-for-hire status may attach to works that are "specially ordered or commissioned" under a written work-made-for-hire agreement, but only if the works fall into one of nine exclusive categories of copyrightable works specified: a contribution to a collective work, as a part of a motion picture or other audiovisual work, as a translation, as a supplementary work, as a compilation, as an instructional text, as a test, as answer material for a test, or as an atlas.
6	States that the musician was an independent contractor, not an employee, so the first avenue for characterization as a work for hire is not met.
7	Concludes that the albums are not works made for hire, even though the contract purportedly deems them to be so, because sound recordings are not within the nine enumerated categories of works that may be deemed works for hire under 17 U.S.C. § 101.
8	Concludes that ownership of the copyright to the sound recordings first vested in the musician.
9	States that the musician, as original copyright holder, assigned the rights to the albums to Warner Music Group (WMG) pursuant to Clause B in their 1988 agreement.
10	Concludes that the client does not currently own the copyrights to the albums.
11	States that the Copyright Act grants the musician author or her heirs the right, subject to certain conditions, to terminate grants of copyright transfers or licenses that were executed on or after January 1, 1978.
12	States that the right to terminate grants of copyright transfers or licenses cannot be waived or alienated.
13	Concludes that the client, as the author's sole heir and living spouse, maintains transfer rights over the sound recordings.
14	States that for assignments executed on or after January 1, 1978, termination may be effected at any time during a five-year period beginning at the end of 35 years from the date of execution of the grant.

Law (ID 1045 cont.)

Criterion	Description
15	Concludes that the five-year termination window for all the works at issue opened on January 1, 2024.
16	Concludes that the five-year termination window for all the works at issue will close on December 31, 2028.
17	States that to give effect to the termination of rights, the client must serve a written notice upon WMG.
18	States that the written notice to terminate must state the effective date of termination.
19	States that the written notice to terminate must be served to WMG at least two years before, and at most ten years before, the stated effective date of termination.
20	States that a copy of the termination notice must be recorded with the Copyright Office before the effective date of termination.
21	Recommends that the client serve a termination notice as soon as possible, and before the end of 2028, with an effective termination date of two years from the date of service.
22	The model provides accurate Bluebook formatting for each citation.

Table 11: **Investment Banking (ID 810)**. "Imagine you're advising Medtronic, which is looking for ways to create shareholder value. One of the strategies they are considering is spinning off their diabetes segment. You are asked to create a valuation for a stand-alone public entity of the diabetes segment. The information you need to calculate is the present value of free cash flow for the next five years and the present value of the terminal value for the diabetes segment based on an EV/EBITDA multiple. Your comp set is Dexcom, Inc., and Insulet Corporation, and keep present value calculations as of April 26, 2024. Moreover, round your answers to two decimals and use the assumptions below. Assume that 10% of diabetes segment's assets, less goodwill and other intangible assets, is PP&E, and the rest is considered Other Operating Assets. Further assume that these Other Operating Assets are current. Assume the diabetes segment's intangible assets are 5% of Medtronic WholeCo's intangible assets, and that the segment's amortization expense each year is at the same percentage of WholeCo's. Additionally, assume a working capital ratio of 2, a WACC of 8%, and a tax rate of 23%. Lastly, assume that the previous years' NWC for the diabetes segment was 500 million dollars. Below are the following deliverables: 1) Please calculate the following information attributed to the diabetes segment for FY2024: the amount of PP&E, capex, other operating assets, net working capital, change in net working capital, the amortization attributed to the diabetes segment, EBITDA, and free cash flow. 2) Please calculate the following information as of April 26,2024: Median EV/EBITDA multiple for the comparables. 3) Assume for the next 5 years (projected period), that FCF grows at 4% y/y and EBITDA grows at 6% y/y. Calculate the present value of cash flows for the 5 years, as well as the present value of the terminal value. Use the EBITDA exit multiple method. 4) Finally, add the present value of cash flows during the projected period and terminal value to come up with a valuation for Medtronic's diabetes segment."

Criterion	Description
1	Calculates PP&E related to the diabetes segment to be $107.98 million as of April 26, 2024 (acceptable range is between $106.90 to $109.05 million).
2	Calculates $108.70 million in capex for the diabetes segment in FY2024 (acceptable range is between $107.61 and $109.78 million).
3	Calculates other operating assets related to the diabetes segment to be $971.78 million as of April 26, 2024 (acceptable range is between $962.06 and $981.49 million).
4	Calculates net working capital related to the diabetes segment to be $485.89 as of April 26, 2024 (acceptable range is between $481.03 and $490.75 million).
5	Calculates the change in net working capital related to the diabetes segment to be $-$14.11 million in FY2024 (acceptable range is between $-$13.97 and $-$14.25 million).
6	States amortization related to the diabetes segment to be $84.65 million for FY2024 (acceptable range is between $83.80 and $85.50 million).
7	Calculates $572.65 million in EBITDA related to the diabetes segment for FY2024 (acceptable range is between $566.92 and $578.38 million).
8	Calculates $387.45 million in free cash flow related to the diabetes segment for FY2024 (acceptable range is $383.57 and $391.32 million).
9	Calculates the median EV/EBITDA multiple for the comparables to be 47.62x (acceptable range is between 47.57x and 47.67x).
10	Calculates the present value of free cash flow during the projected period to be $1,732.33 million as of April 26, 2024 (acceptable range is between $1,715.00 and $1,749.65 million).
11	Calculates the present value of terminal value to be $24,833.82 million as of April 26, 2024 (acceptable range is between $24,585.49 and $25,082.16 million).
12	Calculates the value of the diabetes segment as of April 26, 2024 to be $26,566.15 million (acceptable range is between $26,300.49 and $26,831.81 million).

Table 12: **Management Consulting (ID 828)**. "Your client recently started working on a business idea to reduce the number of plastic bottles that are not recycled. The idea is to commercialize a feedstock recycling solution for PET plastic, which is the most common plastic used in plastic bottles. Feedstock recycling is a way of breaking down PET back into its building blocks (monomers) using either chemical or biological methods. The big advantage is that new, high-quality PET can be created from these recovered building blocks. Your client would like an estimation of the world's consumption of PET plastic bottles from January 2024 through December 2030. Assume that the average weight of an empty PET bottle is 10 grams. Task Objectives: 1. Estimate the number of plastic bottles consumed from January 2024 through December 2030 using the following process: a) Using the files "828 - Important Plastic Water Bottle Stats.pdf" and "828 - Worldbank Population - Original.csv", determine the average consumption of plastic water bottles per person per year based on global daily consumption and the world population in 2023. Round to the nearest whole number. Assume the figure has remained constant since 2023, and will remain constant through 2030. b)Using the file "828 - Worldbank Population - Original.csv", determine the 2019 and 2023 World population and 2019-2023 compound annual growth rate (CAGR) rounded to two decimal places. Then forecast the World population in 2024-2030 using rounded CAGR, rounding the population to the nearest whole number. c) Using the calculations in 1a and 1b, determine the projected number of plastic bottles in 2024-2030, rounded to the nearest whole number 2. Global annual demand for PET plastic bottles: a) Using the file "828 - Percent of Plastic Water Bottles.pdf", determine the % of bottles that are made from PET, assuming the figure is for 2024. The % made from PET is expected to increase by 1 percent every year, compounded (i.e., each year's value is 1.01× the prior year), until 2030 (e.g., 50% would grow to 50.5% in one year). Round answer to one decimal place when displayed in % (e.g., 50.51% would round to 50.5%). b) Using the calculation in 2a, determine the global annual demand for PET plastic bottles from 2024-2030 in metric tons, rounded to the nearest whole number. This will serve as a proxy for the total volume of PET potentially available for recycling."

Criterion	Description
1	Identifies the World population in 2023 as 8,061,876,001.
2	Identifies the average consumption of bottles per person per year as 59.
3	Identifies the World population in 2019 as 7,776,892,015.
4	Calculates the 2019–2023 World population CAGR as 0.90%.
5	Forecasts the World population in 2024 as 8,134,432,885.
6	Forecasts the World population in 2025 as 8,207,642,781.
7	Forecasts the World population in 2026 as 8,281,511,566.
8	Forecasts the World population in 2027 as 8,356,045,170.
9	Forecasts the World population in 2028 as 8,431,249,577.
10	Forecasts the World population in 2029 as 8,507,130,823.
11	Forecasts the World population in 2030 as 8,583,695,000.
12	Calculates the number of plastic bottles in 2024 as 479,931,540,215.
13	Calculates the number of plastic bottles in 2025 as 484,250,924,079.
14	Calculates the number of plastic bottles in 2026 as 488,609,182,394.
15	Calculates the number of plastic bottles in 2027 as 493,006,665,030.
16	Calculates the number of plastic bottles in 2028 as 497,443,725,043.
17	Calculates the number of plastic bottles in 2029 as 501,920,718,557.
18	Calculates the number of plastic bottles in 2030 as 506,438,005,000.
19	Identifies the % of bottles made from PET in 2024 as 78.8%.
20	Calculates the % of bottles made from PET in 2025 as 79.6%.
21	Calculates the % of bottles made from PET in 2026 as 80.4%.
22	Calculates the % of bottles made from PET in 2027 as 81.2%.
23	Calculates the % of bottles made from PET in 2028 as 82.0%.
24	Calculates the % of bottles made from PET in 2029 as 82.8%.
25	Calculates the % of bottles made from PET in 2030 as 83.6%.
26	Calculates the global annual demand for PET plastic bottles in 2024 as 3,781,861 metric tons.
27	Calculates the global annual demand for PET plastic bottles in 2025 as 3,854,637 metric tons.
28	Calculates the global annual demand for PET plastic bottles in 2026 as 3,928,418 metric tons.
29	Calculates the global annual demand for PET plastic bottles in 2027 as 4,003,214 metric tons.
30	Calculates the global annual demand for PET plastic bottles in 2028 as 4,079,039 metric tons.
31	Calculates the global annual demand for PET plastic bottles in 2029 as 4,155,904 metric tons.
32	Calculates the global annual demand for PET plastic bottles in 2030 as 4,233,822 metric tons.

Table 13: **Medicine (ID 929)**. "A 6-year-old boy (new patient) presents to your clinic with wheezing. Per parents, he is on Fluticasone Propionate/Salmeterol (Advair® HFA), 2 puffs twice daily (total 460 mcg fluticasone/day) and montelukast. He also uses albuterol as needed. In the past 12 months, he has been to the ER 10 times for wheezing and given inhalers and oral medications. Three of the ER visits resulted in prolonged hospital stays for a "lung infection," and he is "always sick". Other than asthma, he has no medical problems that they are aware of. Parents also state that he is UTD on all his vaccines except for his 4-year vaccines. Parents have been hearing more about Tetanus and want to ensure that he is immune. On physical examination, he exhibits diffuse wheezing, which improves with albuterol and oral steroids in the office. He is sent home with an oral steroid burst, albuterol, and continuation of Advair. The rest of the ROS and physical exam are negative, and he has a normal height and weight. Initial labs and imaging are completed CBC with Diff: elevated eosinophils at 300 cells Mu. Reassuring CMP Immunoglobulins: normal IgE and IgM, lower IgG (<2 standard deviations below the norm), low IgA Tetanus non-immune Chest x ray: normal Two months later, the family returns and wants to discuss labs and additional medications for his symptoms. What labs should be drawn at this visit? Please include both visit specific labs and any follow-up labs needed from the last appointment. Based on the patient's history, presenting symptoms, and available lab work, what is the most likely diagnosis? What is the most appropriate medication to start immediately, and what other medications and treatments should be considered? Information should be based on the latest peer-reviewed medical literature as of May 2025."

Criterion	Description
1	Mentions that Common Variable Immunodeficiency (CVID) has lower IgG accompanied to reduced IgA and/or IgM concentrations.
2	Mentions that patients with Common Variable Immunodeficiency (CVID) have low IgG concentration (<2 standard deviation below age-appropriate references) for 2 measurements 3 weeks apart.
3	Mentions that the patient's medical history, recurrent severe sinopulmonary infection with 10 ER visits and 3 hospitalizations, is indicative of a possible Common Variable Immunodeficiency (CVID) diagnosis.
4	Mentions that the patient's non-immune Tetanus status is indicative of a possible Common Variable Immunodeficiency (CVID) diagnosis.
5	States the likely diagnosis is Common Variable Immunodeficiency (CVID).
6	States that the patient's age, greater than 4 years old, meets diagnostic criteria for Common Variable Immunodeficiency (CVID) diagnosis.
7	Recommends completing a repeat quantitative serum immunoglobulin panel to confirm low IgG concentration with a second measurement greater than 3 weeks apart.
8	Mentions that patients may be diagnosed with Common Variable Immunodeficiency (CVID) with only one serum study if the serum IgG level is very low (<100-300 mg/dL depending on the age).
9	Mentions that for a Common Variable Immunodeficiency (CVID) diagnosis, no secondary causes of hypogammaglobulinemia can be present.
10	Mentions memory B-cells may be reduced in Common Immunovariable Deficiency (CVID).
11	Recommends completing a B-cell subset analysis by flow cytometry for immunophenotyping of B-cells and to rule out X-linked Agammaglobulinemia.
12	Recommends completing a T-lymphocyte subset analysis by flow cytometry to evaluate for T cell deficiency.
13	Recommends testing an adequate immune response to previous vaccinations by measuring specific antibody titers to protein antigens (such as tetanus toxoid, diphtheria toxoid).
14	Mentions testing an adequate immune response to previous vaccinations by measuring specific antibody titers to polysaccharide antigens (such as pneumococcus).
15	Recommends infection avoidance (hand hygiene, drinking treated water, respiratory protection) as the first step in the treatment plan.
16	Recommends IV immunoglobulin replacement therapy as the first medical intervention to treat Common Immunovariable Deficiency (CVID).
17	Recognizes that the patient's immunoglobulin levels will need to be monitored every 6 months, with appropriate dose adjustment based on IgG production and weight.
18	Highlights that the patient will need to be screened for anti-IgA antibodies to prevent anaphylactic reactions to the treatment.
19	Recognizes that IV immunoglobulin therapy (IVIG) alleviates the state of chronic immune activation in CVID, helping improve cellular immunity.
20	Recognizes that IV immunoglobulin therapy should help to decrease the recurrence sinopulmonary in infections in this patient.
21	Considers antibiotics for infection prophylaxis given patient's notable history of recurrent sinopulmonary infections.
22	Recognizes that the patient is not up-to-date on 4-year vaccines.
23	Recognizes that the patient's upcoming vaccine schedule may need to be modified to delay or spread out the administration of live vaccines (MMR, varicella).

Medicine (ID 929 cont.)

Criterion	Description
24	Recommends educating the patient's parents regarding the risks and benefits of live-attentuated vaccinations and reaching a joint decision together.
25	Recognizes that autoimmunity, and specifically autoimmune cytopenias, are the most common noninfectious complication of Common Immunovariable Deficiency (CVID).
26	Recommends drawing a complete blood count with differential to screen for cytopenias (thrombocytopenia, anemia, neutropenia).
27	Mentions completing a repeat eosinophil count in lab work.
28	Mentions the patient's history of diffuse wheezing with elevated blood eosinophils (last visit) are indicative of eosinophilic asthma.
29	Classifies the patient's asthma as severe and poorly controlled given frequent exacerbations despite being prescribed the combination of Advair and montelukast.
30	Notes the anti-IL-5 monoclonal antibody, Mepolizumab, is indicated for severe eosinophilic asthma in this patient's age group.
31	Recognizes that of the three biologics (Dupilumab, Mepolizumab, Omalizumab) approved to treat severe asthma in patients 6 years and older, only Mepolizumab has an FDA label indicating eosinophilic asthma.
32	Recommends mepolizumab to treat the patient's severe eosinophilic asthma, recognizing that an eosinophil count 300 cells/ Mu L or greater in the past 12 months is sufficient to begin treatment without a repeat lab value.
33	Mentions the typical dose of mepolizumab for a 6 year old is 40 mg subcutaneously every 4 weeks.
34	Recognizes the need for parent education to administer an injectable biologic to a 6 year old and maintain medication adherence.
35	Mentions recent literature suggest that interleukin 5 (IL-5) receptor blockage can be safely co-administered with IVIG in patients who require both therapies.