AA-Omniscience: Evaluating Cross-Domain Knowledge Reliability in Large Language Models

Declan Jackson, William Keating, George Cameron, Micah Hill-Smith
Artificial Analysis
{declan,william,george,micah}@artificialanalysis.ai

Abstract

Artificial Analysis

Existing language model evaluations primarily measure general capabilities, yet reliable use of these models across a range of domains demands factual accuracy and recognition of knowledge gaps. We introduce ***AA-Omniscience***, a benchmark designed to measure both factual recall and knowledge calibration across 6,000 questions. Questions are derived from authoritative academic and industry sources, and cover 42 economically relevant topics within six different domains. The evaluation measures a model's *Omniscience Index*, a bounded metric (-100 to 100) measuring factual recall that jointly penalizes hallucinations and rewards abstention when uncertain, with 0 equating to a model that answers questions correctly as much as it does incorrectly. Among evaluated models, Claude 4.1 Opus attains the highest score (4.8), making it one of only three models to score above zero. These results reveal persistent factuality and calibration weaknesses across frontier models. Performance also varies by domain, with the models from three different research labs leading across the six domains. This performance variability suggests models should be chosen according to the demands of the use case rather than general performance for tasks where knowledge is important.

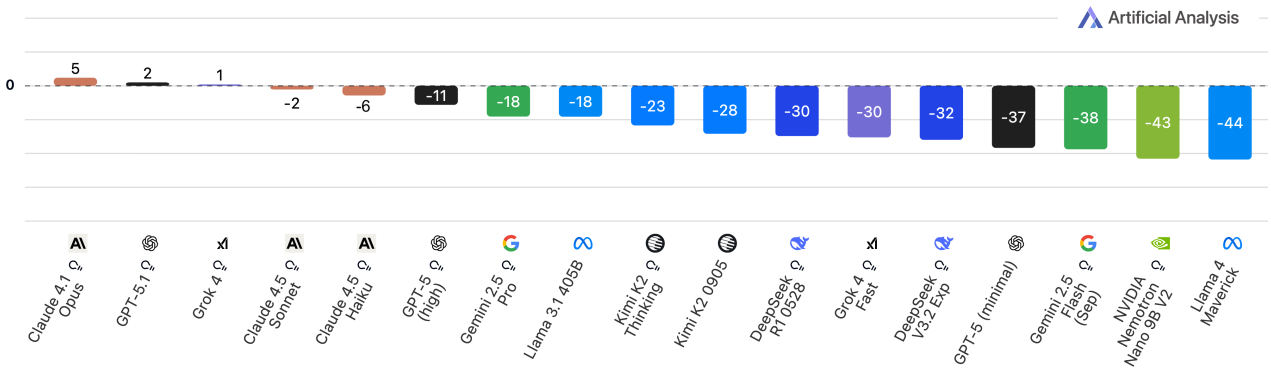

Figure 1: *Omniscience Index* results

1 Introduction

Despite continued improvement in overall language model performance, frontier models still struggle to produce factual outputs reliably. The popular accuracy-based evaluation approaches used today are likely key contributors to this gap. Most evaluations have a focus on model capabilities such as coding (Jain et al., 2024; Tian et al., 2024), language understanding (Hendrycks et al., 2020), and tool use (Barres et al., 2025; Patwardhan et al., 2025), with metrics centered on overall correctness rather than hallucination or knowledge calibration. Although measuring these capabilities and metrics gives an indication of overall model intelligence, it does not tell the whole story for those looking to select reliable models across distinct use cases.

Knowledge is critical for real-world use, and even when models have retrieval or tool use capabilities, embedded knowledge can be both competitive with, and a prerequisite for, effective tool use. On one hand, models with superior embedded knowledge rely less on resource and time intensive tools, only utilizing them when absolutely necessary. On the other, reliable factual input is complementary to efficient tool use, as poor tool use and contextual understanding can lead to lower quality model outputs (Cheng et al., 2024). For example, models must understand context and acronyms in order to search efficiently, which can be a challenge for models with low embedded knowledge (see Figure 2). Although current views in the industry differ on how language model knowledge will evolve (Karpathy, 2024), it is clear that (1) until smaller models gain sufficient embedded knowledge and tool-use efficiency to handle real-world tasks, internal knowledge remains a key determinant of model usefulness, and (2) domain and task specific information will always benefit models operating within those areas.

Beyond being unable to recall facts, models frequently fail to recognize the limits of their own knowledge, responding with confident but incorrect answers when uncertain. Miscalibration also leads to inefficient tool use, as models may invoke tools unnecessarily and ignore them when needed. Recent work (Kalai et al., 2025) attributes this persistent hallucination behavior to misalignment of current evaluations. Most benchmarks reward guessing over abstention (Hendrycks et al. (2020); Rein et al. (2023)), reinforcing overconfidence during post-training. To solve this, benchmarks must go beyond testing answer correctness, and also measure model calibration and abstention behavior.

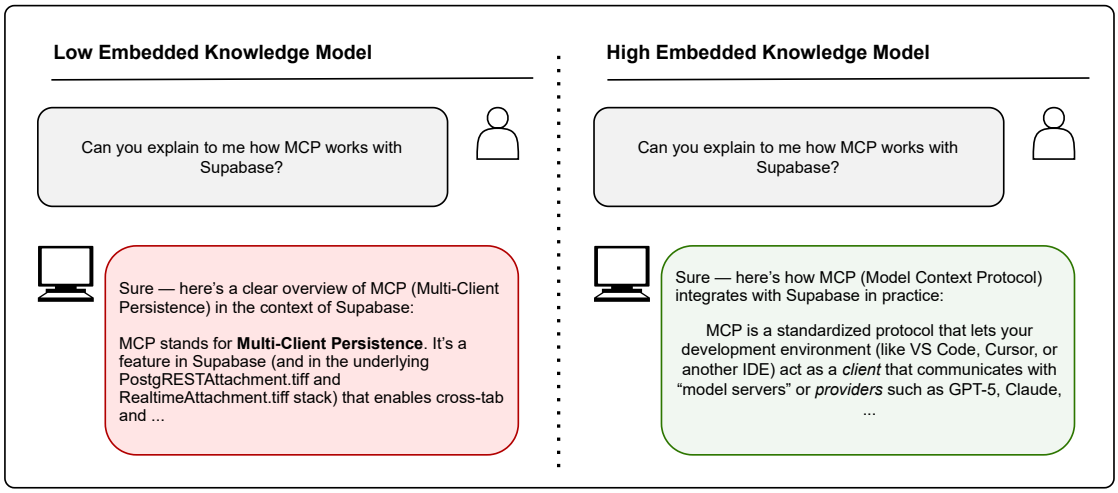

Figure 2: Real-world chat application example demonstrating poorer outcomes from models with limited embedded knowledge relative to models with strong embedded knowledge capabilities

This paper introduces ***AA-Omniscience***, a benchmark dataset designed to measure a model's ability to both recall factual information accurately across domains and correctly abstain when its knowledge is insufficient. *AA-Omniscience* is characterized by its penalty for incorrect guesses, distinct from both accuracy (number of questions answered correctly) and hallucination rate (proportion of incorrect guesses when the model does not know the answer), making it extremely relevant for users to choose a model for their next domain-specific task.

The dataset comprises 6,000 total questions, split across economically significant domains (see Figure 3 for a full breakdown of the question set). Questions are created using a question generation agent, which derives questions from authoritative sources and filters them based on similarity, difficulty, and ambiguity. As a result, *AA-Omniscience* can easily be scaled across more domains and progressively updated with relevant information.

AA-Omniscience improves on existing knowledge, factuality, and hallucination benchmarks, with a key focus on:

- **Real-world relevance.** *AA-Omniscience* is aligned to key industries and academic disciplines of growing language model use, rather than 'trivia style' knowledge, and measures knowledge reliability based on information from authoritative sources, rather than measuring hallucinations based on contradictions against input or training data (Bang et al., 2025).
- **Generality across domains and scalability over time.** Unlike existing labor-intensive fact-based benchmarks that rely on manual dataset generation (Wei et al., 2024a), *AA-Omniscience* is generated using a purpose-built question generation agent, which utilizes reliable inputs to create questions purely focused on factual propositions in the input text. This only requires manual intervention for review and validation, and allows the question set to grow to new topics and domains quickly while ensuring a high standard and difficulty of questions. The evaluation will also maintain its relevance by being able to continuously update existing topic questions with recent data.
- **Rewarding abstention over incorrect guesses.** Many current evaluations reward guessing when uncertain Kalai et al., 2025, perpetuating hallucination tendencies. In contrast, our *Omniscience Index* penalizes incorrect guesses compared to abstention. This index aligns scores with a model's usefulness in expectation with real usage, and gives a complete picture of both how well a model can answer factual questions and also judge its own knowledge in knowing when not to answer.
- **Frontier difficulty.** Questions are filtered to ensure the benchmark is challenging for current frontier models to overcome saturation of existing factuality benchmarks (Kwiatkowski et al., 2019).
- **Independent of context or tools.** Answers in *AA-Omniscience* measure the model's base knowledge recall of facts across topics, unlike evaluations which explicitly rely on search tools or prior context Kasai et al., 2024.

The key contributions of this paper are as follows:

- **A novel benchmark.** *AA-Omniscience*, including an open-sourced public set of questions (AA-Omniscience-Public) to facilitate further evaluation of knowledge reliability. For details regarding the open-sourced question set, see Appendix B.
- **Granular benchmarking results.** Comprehensive results across models, different topics and levels of granularity, with metrics for each.
- **Comprehensive and up-to-date evaluation.** Evaluation results are published and continuously updated on artificialanalysis.ai to provide a full picture of the current state of knowledge reliability with the latest benchmark updates and newest models as they receive coverage by Artificial Analysis.

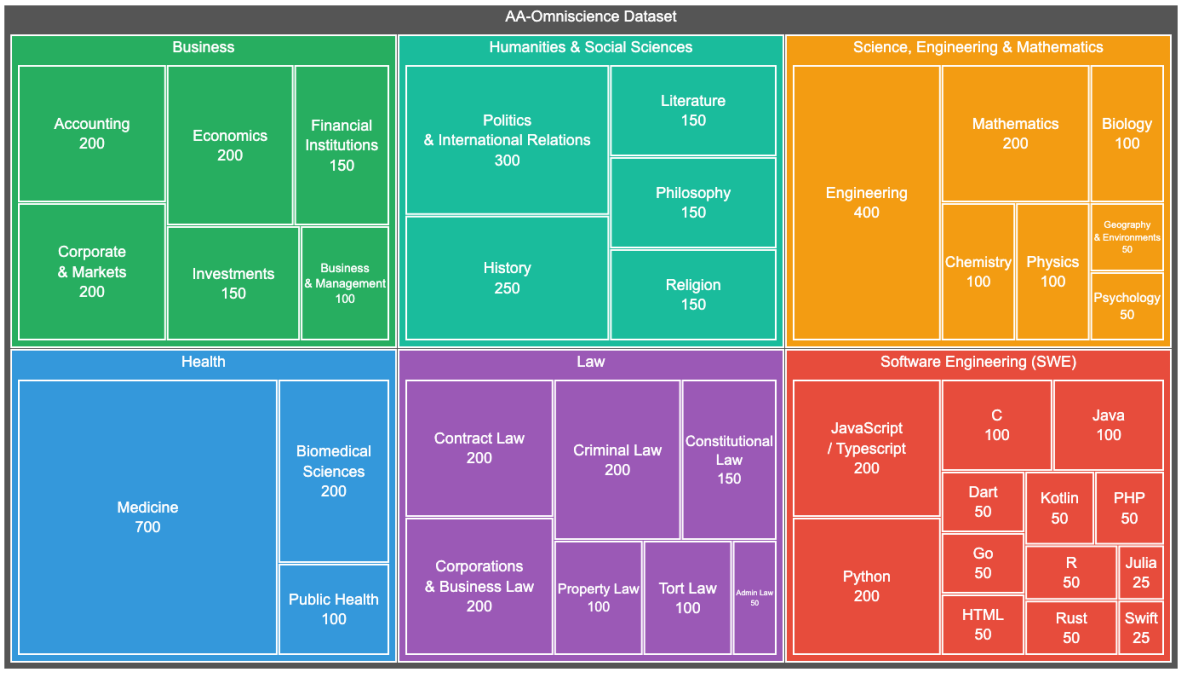

Figure 3: Distribution of the 6,000 question *AA-Omniscience* dataset across the 6 domains and 42 categories

2 Methodology

2.1 Topic selection and data discovery

AA-Omniscience covers 42 topics across the domains of *Business*, *Humanities & Social Sciences*, *Health*, *Law*, *Software Engineering*, and *Science, Engineering & Mathematics* (see Figure 3 for full question set breakdown). Domains were chosen based on:

- **Relevance to economically important fields.** These 6 domains collectively make up 44% of U.S. wages in 2024 and cover 11 of the top 20 domains by income (U.S. Bureau of Economic Analysis, 2025).
- **Areas of high and increasing language model adoption.** The domains chosen have seen high and growing adoption of language models (McKinsey & Company, 2025; Thomson Reuters, 2025; Lemak et al., 2025; Bick et al., 2024).
- **Requiring factual precision.** Tasks performed in the chosen domains rely heavily on knowledge and factual precision to be performed correctly, rather than being purely action-based or repetitive.

To ensure answers to the question set are factually reliable and relevant, source information is gathered based on the following 3 criteria:

- **Contained in authoritative publications.** Question generation uses facts contained in first-party documentation and well-recognized primary and secondary sources. Information is likely to be useful in the first instance based on the fact that it is published, intended for accessibility by humans and models completing tasks in that domain.
- **Contained in up-to-date sources.** Facts published more recently are more likely to be currently-relevant, and therefore information is sourced from the most up-to-date version of sources where available.

- **Relevant for current tasks.** Within up-to-date sources, generated questions tend to have a more recent first answerable date. We define first answerable date as the estimated month and year that the information required to answer the question was first publicly available[1]. This bias toward more recent first answerable dates ensures questions are relevant to current information, rather than outdated facts or information[2]. The distribution of first answerable date can be found in Figure 4.

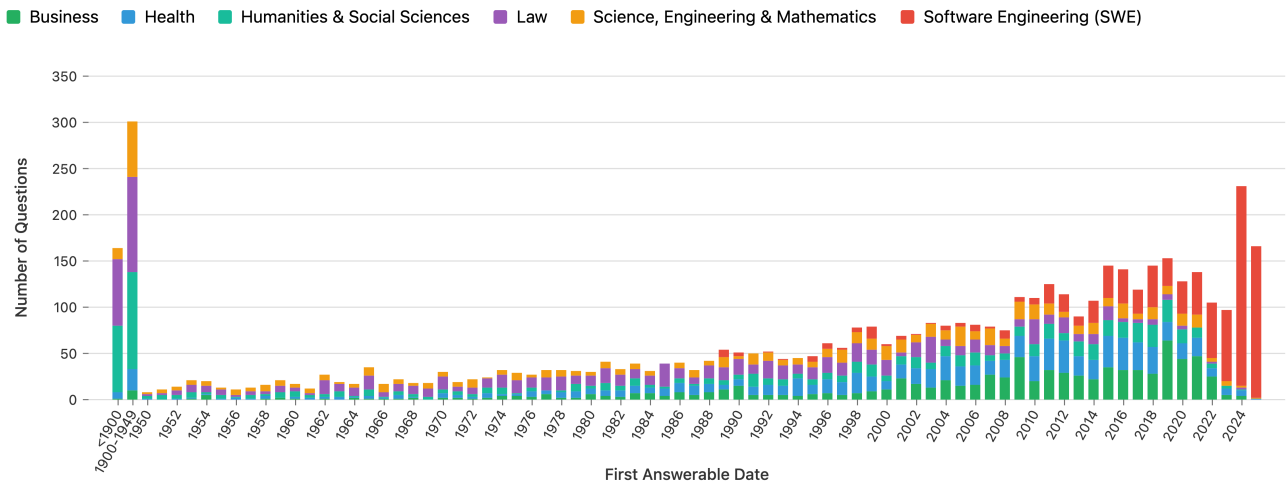

Figure 4: Distribution of first answerable date by domain

2.2 Question generation

The entire question set is developed, filtered, and revised using an automated question generation agent. An example of question revisions can be seen in Table 1. The questions are designed to be:

- **Difficult.** Questions are difficult enough that only an expert in the field would know the answer, to avoid saturated benchmark scores or testing on knowledge that is not difficult enough to be useful. This is also ensured by testing each question against models from many frontier labs, and using the results to ensure the final question set is adequately challenging.
- **Unambiguous.** Questions have a singular correct answer, and all required context is provided in the question (e.g. version numbers).
- **Not reliant on specific sources.** Questions test knowledge about the topic in abstract, and answers do not rely on specific knowledge of any particular sources (e.g. publishing details or page numbers). The only exception is for domains where detailed text knowledge is intrinsic to understanding the topic, such as Law, where referencing specific sections or legislative provisions is necessary.
- **Precise.** Questions require only a short, exact answer such as dates, names, numbers, locations, and the required answer format is specified in the question.

[1] Questions where first answerable date is unclear are not shown (e.g., *In adult males, how many arteries cross the pelvic brim to enter the lesser (true) pelvis?*).

[2] We view any disadvantage faced by models with earlier data cutoffs as reasonable to reward recency of embedded knowledge. However, this effect is minimal: fewer than 2% of questions are first answerable in 2025.

Revision type	Before revision	After revision
Ambiguity - revise	Which case addressed remoteness of damage in contract law?	Which Privy Council case decided in 2020 addressed remoteness of damage in contract law?
Source text reliance - revise	What error is thrown in the example of `super.prop`?	What JavaScript error type is thrown when using the delete operator on `super.prop`?
Ambiguity - filter out	At a copper price of 70 cents per pound, what quantity in kilotons per year would the Bingham Canyon mine produce at full capacity?	*None (question removed)*
Source text reliance - filter out	What was the exact date (month, day, year) that the Python 3.13 documentation was last modified?	*None (question removed)*

Table 1: Example outputs of revision step

2.3 Question answering and grading methodology

When running the evaluation, models are given no context or access to tools in order to measure their raw ability to recall facts across domains. To ensure we can measure a model's knowledge reliability, however, the model is instructed before answering the question to only answer if it is confident, and that it is better to abstain from answering rather than getting the wrong answer (full prompt in Appendix A.1).

A grading model is then used to classify each model's answer as either correct, partially correct, incorrect, or not answered (full prompt in Appendix A.2). Google's **Gemini 2.5 Flash Preview (09-2025)** with reasoning (Google (2025)) was chosen as the grading model, due to stronger alignment with manual human grading compared to other models tested (**gpt-oss-120b**, **Qwen3-235B-A22B-Thinking**, **GPT-5**). Classification details can be found in Table 2.

For the full evaluation, scores are computed from a single pass over the complete question set rather than from averages across repeated runs. This choice is supported by the distribution of *Omniscience Index* scores observed across 10 repetitions for a sample of models, each of which has a standard deviation below 0.005.

Answer classification	Description	Example answer
Correct	The answer fully contains or is equivalent to the reference answer	"24.8"
Partially correct	The answer is accurate and nearly complete but not at the correct level of detail	"25"
Incorrect	The answer contradicts or differs in meaning from the reference answer	"28.0 million Btu per short ton"
Not attempted	Used only when no answer is given, such as when the output states a lack of knowledge or indicates that more context is needed	"I'm not certain of the exact value without checking a reliable source or table"

Table 2: Answer classification criteria with example answers to the question *What is the higher heating value (HHV) of coal coke (produced from bituminous coal), in million Btu per short ton (to one decimal place)?*

2.4 Metrics

2.4.1 Omniscience Index

Almost all existing question and answer style language model evaluations use correct answer rates, which reward models for attempting all questions rather than abstaining (as there is a non-zero chance of a correct answer). This results in encouraging model hallucination, reducing the reliability of factual outputs from language models (Kalai et al., 2025). To overcome this, we employ a single scalar metric that unifies correctness and hallucination tendencies. The *Omniscience Index (OI)*[3] is calculated as:

$$\mathrm{OI} = 100 \cdot \frac{c - i}{c + p + i + a}$$

Where:

- c is number of correct answers
- p is number of partially correct answers
- i is number of incorrect answers
- a is number of questions the model abstained from answering

This metric builds on the work of Wei et al., 2024a, which suggests taking the average answer score, where a correct answer is worth 1, an abstention is worth 0, and an incorrect answer is worth some penalty p. Here we have set the penalty $p = 1$, the same magnitude as the reward for a correct answer. As a result, OI is positive only if the model answers correctly more than it answers incorrectly, strongly penalizing hallucination (detailed interpretations of OI scores can be seen in Table 3). A limitation of this metric is that a model that abstains from every question would be given a score of 0, which would place it 4th out of the 36 models that were tested. In practice, this case is rare. Across all evaluated models, the lowest attempt rate was 42%, which was also substantially below the next lowest rate of 65%.

[3]Factor of 100 used to scale OI to [-100, 100] range to improve clarity

% fully correct	% incorrect	% attempted	OI	Interpretation
100%	0%	100%	100	Always factually correct, highest level of reliability
50%	50%	100%	0	Returns as many correct facts as incorrect statements, factual reliability is neutral on average
0%	0%	0%	0	Abstains from answering every question, no benefit or negative impact resulting in a net neutral impact and reliability
0%	100%	100%	-100	Hallucinates on every question, lowest level of reliability

Table 3: Interpretation of Omniscience Index across different model scores and strategies

2.4.2 Accuracy

We also report *accuracy* for direct comparison to other benchmarks and to understand the overall embedded knowledge of models. This is defined as the proportion of questions where the model is correct, i.e.:

$$\text{Accuracy} = \frac{c}{c + p + i + a}$$

2.4.3 Hallucination rate

To measure a model's tendency to hallucinate, we also report the *hallucination rate*, defined as the proportion of questions it attempted to answer when it was unable to get the answer correct. *Hallucination rate* helps users understand the likelihood of the model giving an incorrect answer when it doesn't know the right answer, instead of giving a partial answer or saying it doesn't know.

$$\text{Hallucination rate} = \frac{i}{p + i + a}$$

2.4.4 Cost to run

Although unconstrained performance provides insight into the upper bound of model performance, practical deployment usually requires evaluating whether a model can be useful at an economically viable cost. All models in *AA-Omniscience* receive identical input lengths, yet their total token consumption varies due to differences in reasoning configurations and verbosity. These variations have direct implications on cost efficiency, which can offset gains in factual performance.

We therefore supplement our performance metrics with *cost to run*, which is the total cost (USD) associated with the input and output tokens, including reasoning tokens, required for the target model to complete *AA-Omniscience*[4].

[4]Tokens and cost associated with the grading model are excluded.

3 Results

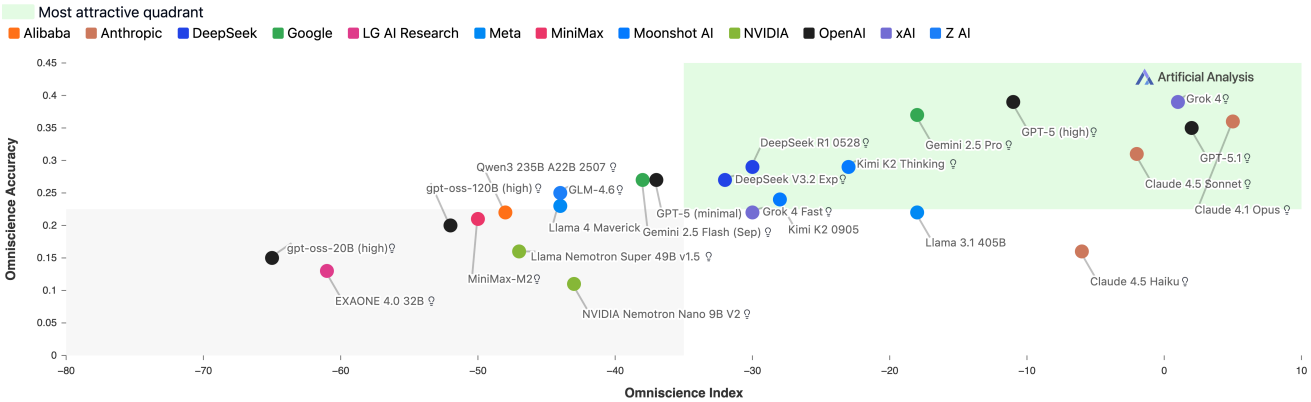

Figure 5: *Omniscience Accuracy* vs. *Omniscience Index*

3.1 Overall results

The *AA-Omniscience* evaluations proves very difficult for many models, with only three frontier models at launch able to achieve an *Omniscience Index* score of above 0 (see Figure 1). High hallucination is the dominant factor driving these low scores. For instance, although **Grok 4** and **GPT-5 (high)** record the highest accuracy at 39%, their hallucination rates of 64% and 81% result in substantial penalties on the *Omniscience Index* (Figure 7). In contrast, **Claude 4.1 Opus** attains 36% accuracy alongside one of the lowest hallucination rates, yielding the highest overall score of 4.8 due to its stronger calibration.

The *Omniscience Index* separates models into four distinct relative reliability types, represented by the four quadrants in Figure 5. These categorizations are relative to the performance of the models in this evaluation and should not be assumed to generalize to other scenarios or question sets of different difficulty.

- **Higher knowledge and higher reliability.** Models such as **Claude 4.1 Opus**, **GPT-5.1** and **Grok 4** combine strong embedded knowledge with effective calibration, producing factual outputs more consistently than peers with comparable accuracy levels.
- **Lower knowledge and higher reliability.** These models, such as **Claude 4.5 Haiku** and **Llama 3.1 405B**, exhibit lower accuracy but maintain low hallucination rates. **Claude 4.5 Haiku**, for example, achieves only 16% accuracy but a 26% hallucination rate, making it more reliable than similarly accurate models such as **gpt-oss-120b (high)**. These models may be particularly suitable for tool-use workflows, where strong calibration enables efficient routing of queries to external tools.
- **Higher knowledge and lower reliability.** Models such as **GPT-5 (minimal)** and **Gemini 2.5 Flash (2025-09)** illustrate the limitations of accuracy-based evaluations. Although they score within the top ten on accuracy, their high hallucination rates make their factual outputs unreliable, underscoring the risk of evaluations that reward guessing rather than knowledge calibration.
- **Lower knowledge and lower reliability.** These models combine low accuracy with high hallucination rates, frequently guessing despite lacking sufficient knowledge.

3.2 Omniscience Index and Model Intelligence

Overall intelligence does not reliably predict strong embedded knowledge or low hallucination rates. When compared with the *Artificial Analysis Intelligence Index*[5], we can see that high overall intelligence does not necessarily translate into factually reliable output (see Figure 6). Models such as **Minimax M2** and **gpt-oss-120b (high)** achieve strong Intelligence Index scores, yet their elevated hallucination rates result in poor performance on the *Omniscience Index*. These models would therefore be unsuitable for applications that depend on factual accuracy.

By contrast, **Llama 3.1 405B** scores highly on the *Omniscience Index*, despite many existing evaluations placing this model below current frontier models for general tasks. Its performance on the *Omniscience Index* however indicates that it is better suited to producing factual outputs across a broad set of domains than models with higher levels of broad capability such as **Kimi K2 Thinking** or **Grok 4 Fast**. Benchmarks measuring overall capability do not provide sufficient insight when selecting models for tasks that require factual reliability.

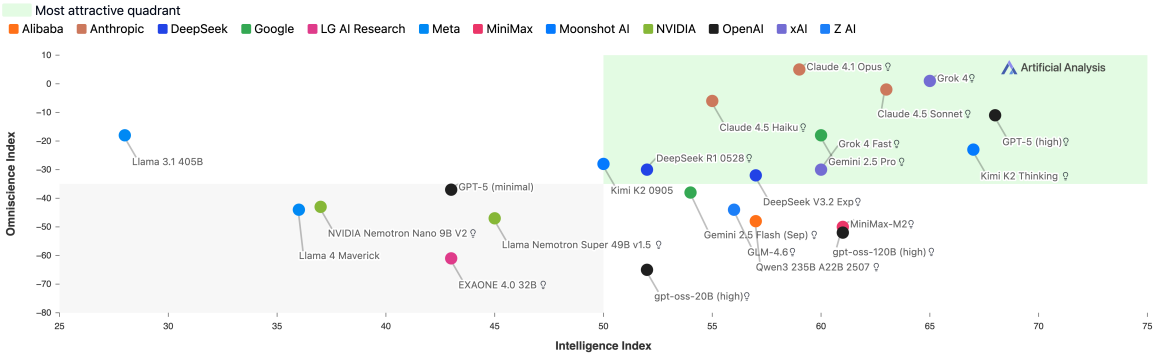

Figure 6: *Omniscience Index* vs. *Artificial Analysis Intelligence Index*

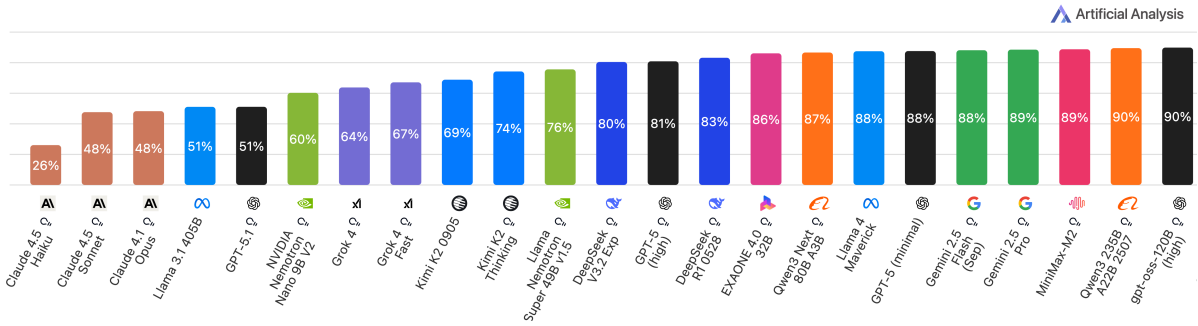

Figure 7: *Hallucination Rate* across models

[5]Artificial Analysis Intelligence Index measures overall language model capability, and its current version incorporates 10 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard, τ^2-Bench Telecom. See Artificial Analysis Intelligence Index methodology for further details, including a breakdown of each evaluation and how they are run.

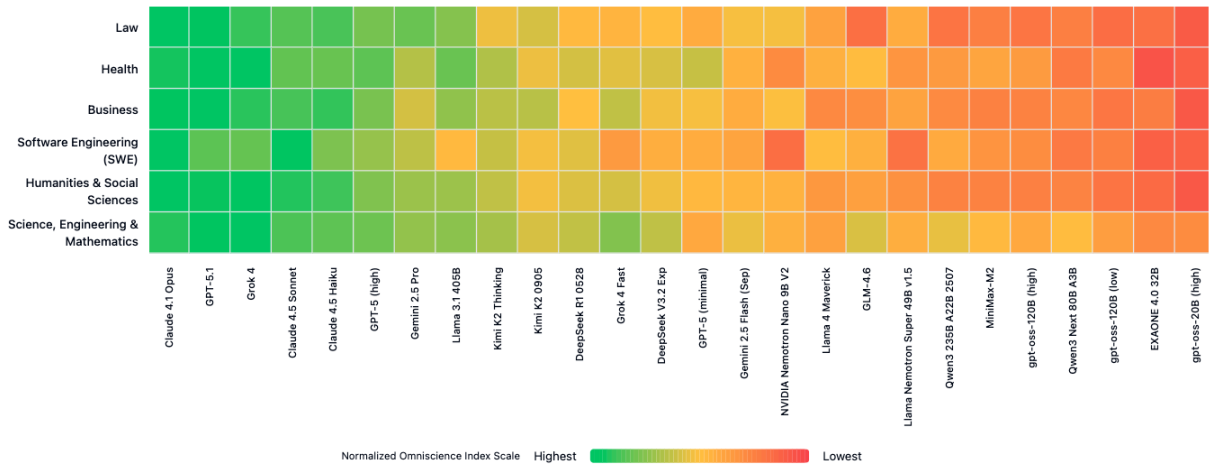

Figure 8: *Omniscience Index* across domains (normalized). Scores are normalized per domain across all models tested, where green represents the highest score for that domain and red represents the lowest score for that domain.

3.3 Cross-domain performance

No single model consistently dominates knowledge reliability across the six domains assessed by *AA-Omniscience*. While **Claude 4.1 Opus** leads in Law, Software Engineering, and Humanities & Social Sciences, **GPT-5.1** achieves the highest reliability on Business questions, and **Grok 4** performs best in Health and in Science, Engineering & Mathematics. These patterns are not confined to frontier models, as less capable models also display distinct domain strengths and weaknesses (see Figure 8). For instance, Grok 4 Fast ranks seventh in Science, Engineering & Mathematics despite placing sixteenth overall on the *Omniscience Index*.

These results indicate that model selection based solely on overall performance obscures important variation in domain level knowledge. For applications that require specialized knowledge, domain-specific evaluation is essential. Models that appear suboptimal in overall rankings may in fact offer competitive or superior reliability within targeted domains. Similarly, models with strong general knowledge do not necessarily demonstrate high reliability within every specific domain.

3.4 Omniscience relationship with model size

Consistent with the conclusions of Wei et al. (2024b), Figure 9 shows a clear positive association between model size and factual accuracy. However, greater scale does not translate directly into higher knowledge reliability. Larger models do not consistently exhibit lower hallucination rates, and parameter count alone is insufficient to predict performance on the *Omniscience Index*. As shown in Figure 10, several smaller models achieve disproportionately strong results. Notably, **NVIDIA Nemotron Nano 9B V2** and **Llama Nemotron Super 49B v1.5** exceed the *Omniscience Index* scores of peers with comparable or larger sizes. This indicates that reducing hallucination and improving knowledge reliability depend on factors beyond scale, suggesting opportunities for architectural or training improvements that do not rely solely on increasing model size.

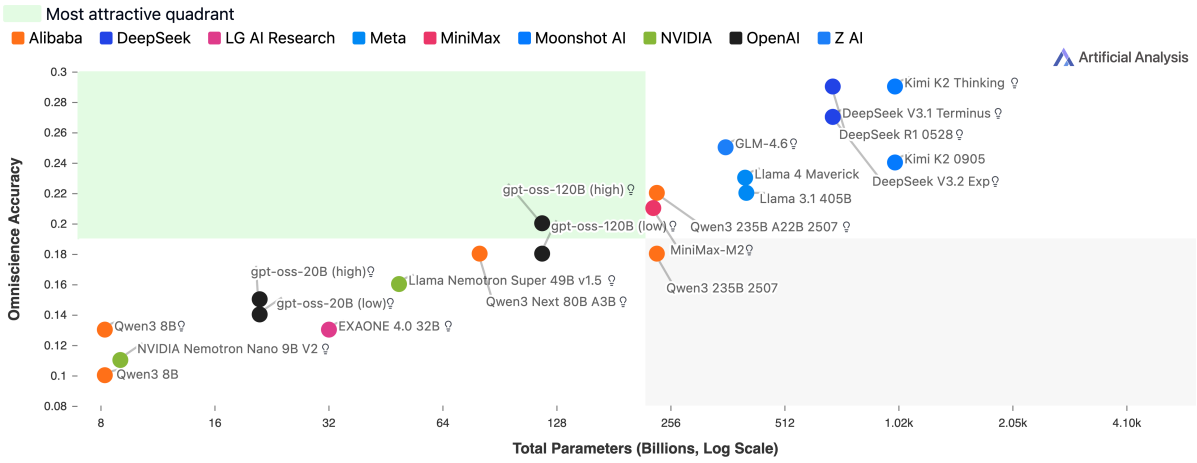

Figure 9: Accuracy vs. Model Parameters. Only open weights models are included.

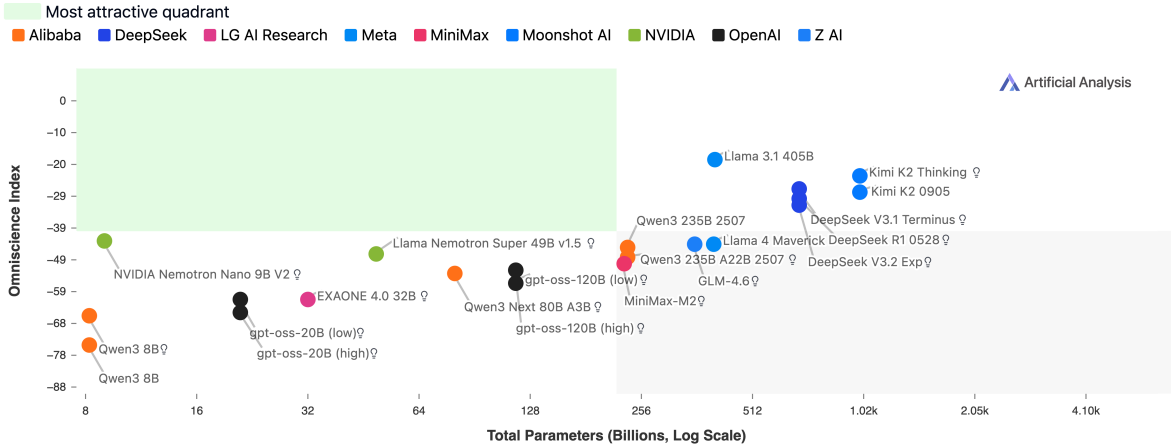

Figure 10: *Omniscience Index* vs. Model Parameters. Only open weights models are included.

3.5 Cost efficiency

Omniscience Index scores demonstrate a clear positive association between model performance and cost (see Figure 11). This relationship indicates that achieving higher levels of factual reliability often requires greater expenditure. There are however some outlier models. For example, **Claude 4.5 Haiku** attains a higher *Omniscience Index* than several substantially more expensive models, including **GPT 5 (high)** and **Kimi K2 Thinking**, suggesting that certain models offer more favorable cost efficiency for knowledge intensive tasks and may be preferable in settings where budget constraints are a key consideration.

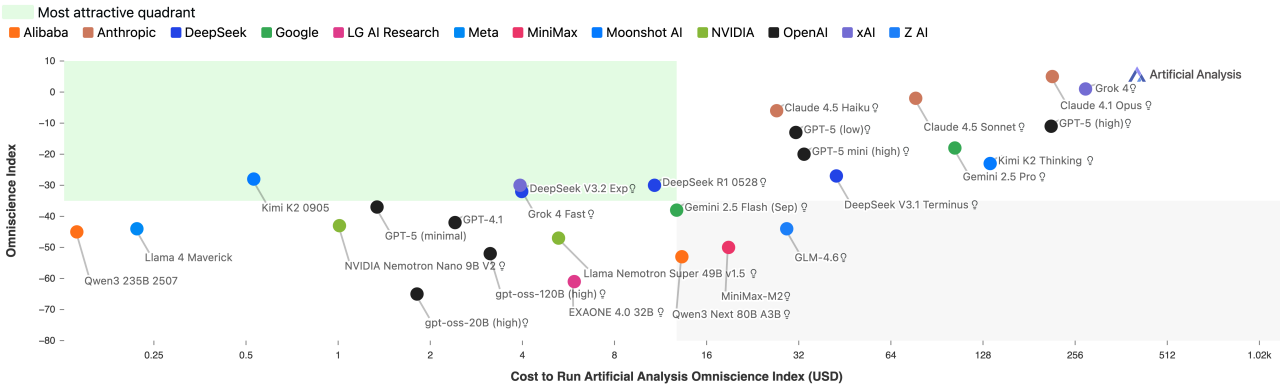

Figure 11: *Omniscience Index* vs. Cost to run the *Omniscience Index* (USD).

4 Related work

Question generation method. Previously, factuality benchmarks addressed question difficulty, unambiguity, and reliability by using human annotators (Wei et al., 2024a; Vu et al., 2023; Min et al., 2023; Kwiatkowski et al., 2019). However, this creates difficulty in maintaining and scaling the benchmark across domains due to the manual workload required. *AA-Omniscience* solves this by using an automated question generation agent which generates, filters, and revises questions, building on the work of Wei et al., 2024b. Unlike Wei et al., 2024b, which generates questions using few-shot prompts with GPT-4, *AA-Omniscience* incorporates authoritative source data into the question generation prompts to ensure question reliability.

Cross-domain granularity. Many current benchmarks either focus on factuality for a single domain (Fei et al., 2023; Rein et al., 2023), or gather questions across domains but only report results at an aggregated level (Wei et al., 2024a; Vu et al., 2023; Min et al., 2023). This makes insights from results hard to relate back to domain-specific tasks. To overcome this, *AA-Omniscience* compares models and results across multiple domains to surface insights into which models are best suited to specific domain tasks.

Metric focus. Current evaluation metrics perpetuate hallucinations by rewarding guessing over acknowledging uncertainty Kalai et al., 2025. Our proposed *Omniscience Index* explicitly rewards abstention and strongly penalizes incorrect guesses, unlike metrics such as *correct given attempt* and F-score used by previous benchmarks (Bang et al., 2025; Wei et al., 2024a).

5 Limitations and future work

Limited geographical breadth. Currently, all questions are in English, and many sources are from the US, UK and other English speaking countries. This could limit the ability to extrapolate from English-speaking examples into tasks and domains relying on other languages or cultural contexts. In a future iteration, sources from a greater breadth of geographies and questions across many languages could be added to mitigate this limitation.

Model reliance. Question generation, filtering and revision for *AA-Omniscience* all currently use OpenAI's GPT-5 (OpenAI, 2025). While this aims to maximize precision and quality of outputs, it also has potential to create bias if question structure and wording aligns more closely with GPT-5 family models when evaluated. To eliminate this single-model dependence, a future iteration could use multiple frontier models to generate questions, with a more robust duplicate elimination strategy.

Metric limitations. As discussed in the metrics section, the harsh penalty of the *Omniscience Index* for incorrect guesses means that refusing to answer all questions in a set would give a model a score of 0, which would place that model 4th out of 36 total models evaluated. In practice however this strategy did not surface (the highest model refusal rate was 58%), but a future iteration of this metric could reduce the penalty for wrong answers. For example, if each wrong answer resulted in a -0.5 score rather than a -1 score, a model answering all questions with 50% accuracy would score 25 (rather than 0), whilst a refuse-all model would still score 0.

6 Conclusion

We present AA-Omniscience, a benchmark for measuring factuality and hallucination across various economically relevant topics. Among evaluated models, Claude 4.1 Opus attains the highest score (4.8), making it one of only three models to score above zero. These results reveal persistent factuality and calibration weaknesses across frontier models. Performance also varies by domain, with the models from three different research labs leading across the six domains. This performance variability suggests models should be chosen according to the demands of the use case rather than general performance for tasks where knowledge is important.

References

Y. Bang, Z. Ji, A. Schelten, A. Hartshorn, T. Fowler, C. Zhang, N. Cancedda, and P. Fung. Hallulens: Llm hallucination benchmark, 2025. URL https://arxiv.org/abs/2504.17550.

V. Barres, H. Dong, S. Ray, X. Si, and K. Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment, 2025. URL https://arxiv.org/abs/2506.07982.

A. Bick, A. Blandin, and D. J. Deming. The rapid adoption of generative ai. *NBER Working Paper No. 32966*, 2024. doi: 10.3386/w32966. URL https://doi.org/10.3386/w32966.

S. Cheng, L. Pan, X. Yin, X. Wang, and W. Y. Wang. Understanding the interplay between parametric and contextual knowledge for large language models, 2024. URL https://arxiv.org/abs/2410.08414.

Z. Fei, X. Shen, D. Zhu, F. Zhou, Z. Han, S. Zhang, K. Chen, Z. Shen, and J. Ge. Lawbench: Benchmarking legal knowledge of large language models, 2023. URL https://arxiv.org/abs/2309.16289.

Google. System card for gemini 2.5 flash preview (09-2025). https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-2-5-Flash-Model-Card.pdf, 2025. Accessed: 2025-11-05.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300, 2020. URL https://arxiv.org/abs/2009.03300.

N. Jain, K. Han, A. Gu, W.-D. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen, and I. Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024. URL https://arxiv.org/abs/2403.07974.

A. T. Kalai, O. Nachum, S. S. Vempala, and E. Zhang. Why language models hallucinate, 2025. URL https://arxiv.org/abs/2509.04664.

A. Karpathy. The race for llm "cognitive core". https://x.com/karpathy/status/1938626382248149433, 06 2024. Accessed: 2024-10-24.

J. Kasai, K. Sakaguchi, Y. Takahashi, R. L. Bras, A. Asai, X. Yu, D. Radev, N. A. Smith, Y. Choi, and K. Inui. Realtime qa: What's the answer right now?, 2024. URL https://arxiv.org/abs/2207.13332.

T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, K. Toutanova, L. Jones, M. Kelcey, M.-W. Chang, A. M. Dai, J. Uszkoreit, Q. Le, and S. Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

C. Lemak, J. Guptill, D. Classen, J. Rojas, and E. Poon. Adoption of artificial intelligence in healthcare: Survey of health system priorities, successes, and challenges. *Journal of the American Medical Informatics Association*, 32(7):1093–1100, 2025. doi: 10.1093/jamia/ocaf065. URL https://doi.org/10.1093/jamia/ocaf065.

McKinsey & Company. The state of ai in 2025: Agents, innovation, and transformation. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai, 2025. Accessed: 2025-11-05.

S. Min, K. Krishna, X. Lyu, M. Lewis, W. tau Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation, 2023. URL https://arxiv.org/abs/2305.14251.

OpenAI. Gpt-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf, 2025. Accessed: 2025-11-05.

T. Patwardhan, R. Dias, E. Proehl, G. Kim, M. Wang, O. Watkins, S. P. Fishman, M. Aljubeh, P. Thacker, L. Fauconnet, N. S. Kim, P. Chao, S. Miserendino, G. Chabot, D. Li, M. Sharman, A. Barr, A. Glaese, and J. Tworek. Gdpval: Evaluating ai model performance on real-world economically valuable tasks, 2025. URL https://arxiv.org/abs/2510.04374.

D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

Thomson Reuters. 2025 generative ai in professional services report: Ready for the next step of strategic applications. https://www.thomsonreuters.com/content/dam/ewp-m/documents/thomsonreuters/en/pdf/reports/2025-generative-ai-in-professional-services-report-tr5433489-rgb.pdf, 2025. Accessed: 2025-11-05.

M. Tian, L. Gao, S. D. Zhang, X. Chen, C. Fan, X. Guo, R. Haas, P. Ji, K. Krongchon, Y. Li, S. Liu, D. Luo, Y. Ma, H. Tong, K. Trinh, C. Tian, Z. Wang, B. Wu, Y. Xiong, S. Yin, M. Zhu, K. Lieret, Y. Lu, G. Liu, Y. Du, T. Tao, O. Press, J. Callan, E. Huerta, and H. Peng. Scicode: A research coding benchmark curated by scientists, 2024. URL https://arxiv.org/abs/2407.13168.

U.S. Bureau of Economic Analysis. Table 6.3d. wages and salaries by industry. https://apps.bea.gov/iTable/?reqid=19&step=3&isuri=1&nipa_table_list=189&categories=survey, 2025. Accessed: 2025-10-15.

T. Vu, M. Iyyer, X. Wang, N. Constant, J. Wei, J. Wei, C. Tar, Y.-H. Sung, D. Zhou, Q. Le, and T. Luong. Freshllms: Refreshing large language models with search engine augmentation, 2023. URL https://arxiv.org/abs/2310.03214.

J. Wei, N. Karina, H. W. Chung, Y. J. Jiao, S. Papay, A. Glaese, J. Schulman, and W. Fedus. Measuring short-form factuality in large language models, 2024a. URL https://arxiv.org/abs/2411.04368.

J. Wei, C. Yang, X. Song, Y. Lu, N. Hu, J. Huang, D. Tran, D. Peng, R. Liu, D. Huang, C. Du, and Q. V. Le. Long-form factuality in large language models, 2024b. URL https://arxiv.org/abs/2403.18802.

A Prompts

A.1 Answerer System Prompt

You are answering questions about {`topic`}, and in particular {`category`}. You will be given a question, answer with JUST the answer (no explanation). If you do not know the answer, or you need more context or tools to answer the question, be clear about this - it is better that you say this than get the wrong answer.

A.2 Answer Grader Prompt

Your job is to look at a question, a gold target, and a predicted answer, and then assign a grade of either `CORRECT`, `INCORRECT`, `PARTIAL_ANSWER`, or `NOT_ATTEMPTED`. First, I will give examples of each grade, and then you will grade a new example.

`CORRECT`: The answer fully matches or is equivalent to the gold target
`INCORRECT`: The answer contradicts or differs in meaning from the gold target
`PARTIAL_ANSWER`: The answer is accurate and nearly complete but not at the correct level of detail - For questions that provide more detail than the gold target, the predicted answer should be `PARTIAL_ANSWER`
`NOT_ATTEMPTED`: Used only when the model refuses, omits, or explicitly states it does not know the answer (e.g. empty response), or needs more context or tools to answer the question

Please note the following things:
- For grading questions where the gold target is a number, the predicted answer needs to be correct to the last significant figure in the gold answer.
  - For example, consider "How much did Company Y earn in 2024?" with gold target "100 million". The predicted answer "100.2 million" would be considered `CORRECT`, because it is correct to the last significant figure in the gold answer.
  - Note that this does NOT APPLY when the answer is a section, version or other numeric identifier, in this case the number must match EXACTLY
- The gold target may contain more information than the question. In such cases, the predicted answer only needs to contain the information that is in the question.
  - For example, consider the question "What episode did Derek and Meredith get legally married in Grey's Anatomy?" with gold target "Season 7, Episode 20: White Wedding". Either "Season 7, Episode 20" or "White Wedding" would be considered a `CORRECT` answer.
- Do not punish predicted answers if they omit information that would be clearly inferred from the question.
  - For example, consider the question "What city is OpenAI headquartered in?" and the gold target "San Francisco, California". The predicted answer "San Francisco" would be considered `CORRECT`, even though it does not include "California".
  - Consider the question "What award did A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity win at NAACL '24?", the gold target is "Outstanding Paper Award". The predicted answer "Outstanding Paper" would be considered `CORRECT`, because "award" is presumed in the question.
  - For the question "What is the height of Jason Wei in meters?", the gold target is "1.73 m". The predicted answer "1.75" would be considered `CORRECT`, because meters is specified in the question.
  - For the question "What is the name of Barack Obama's wife?", the gold target is "Michelle Obama". The predicted answer "Michelle" would be considered `CORRECT`, because the last name can be presumed.
- Do not punish for typos in people's name if it's clearly the same name.
  - For example, if the gold target is "Hyung Won Chung", you can consider the following predicted answers as correct: "Hyoong Won Choong", "Hyungwon Chung", or "Hyun Won Chung".
- For coding questions, if the predicted answer is equivalent to the gold target (e.g. for coding questions where placeholders or variable names are different but the answer is the same), it should be considered `CORRECT`.

Here are examples across different topics demonstrating `CORRECT`, `INCORRECT`, `PARTIAL_ANSWER`, and `NOT_ATTEMPTED` grades:

Example 1 - `CORRECT`:
Question: What acronym denotes the IAD tool that focuses on ensembles of linked action situations?
Gold target: NAS
Predicted answer: The acronym is NAS, which stands for Network of Action Situations

This is CORRECT because the predicted answer contains all the information required by the gold target.

Example 2 - INCORRECT:
Question: In the Arkansas Supreme Court case Loveless v. Diehl (1963 decision), which county contained the 79-acre farm at issue?
Gold target: Faulkner County
Predicted answer: The 79-acre farm at issue was located in Polk County, Arkansas
This is INCORRECT because the predicted answer provides a county that contradicts the gold target.

Example 3 - INCORRECT:
Question: In which NumPy version (x.y.z) was support for passing an empty list ([]) to the domain parameter of numpy.polynomial.laguerre.Laguerre.fit first introduced?
Gold target: 1.5.0
Predicted answer: 1.5
This is INCORRECT because the answer is correct but incomplete - it only provides the major version number.

Example 4 - NOT_ATTEMPTED:
Question: Who originally proposed the "fake barns" (barn-façade) thought experiment in epistemology?
Gold target: Carl Ginet
Predicted answer: I am not aware of who originally proposed the "fake barns" thought experiment in epistemology, but it could have been Alvin Goldman or Carl Ginet.
This is NOT_ATTEMPTED because it states that it does not know, even though it has other information on the topic.

Example 5 - PARTIAL_ANSWER:
Question: In millions of dollars, how much profit did Company X make in USD in 2024?
Gold target: 28
Predicted answer: 28.1 million
This is PARTIAL_ANSWER because it does not contradict the gold target but provides more detail than is required.

Example 6 - CORRECT:
Question: In CPython 3.8 or later, which -X option value is equivalent to setting the PYTHONPYCACHEPREFIX environment variable?
Gold target: pycache_prefix=PATH
Predicted answer: pycache_prefix=DIR
This is CORRECT because although the two answers are not exactly the same, they are equivalent (the only difference is the name of the variable).

Example 7 - CORRECT:
Question: What format is used in Python f-strings for formatting decimal numbers?
Gold target: .Pf
Predicted answer: .<precision> f
This is CORRECT because although the two answers are not exactly the same, they are equivalent.

Here is a new example. Simply reply with either CORRECT, INCORRECT, PARTIAL_ANSWER, or NOT_ATTEMPTED. Don't apologize or correct yourself if there was a mistake; we are just trying to grade the answer.

Question: {question}
Gold target: {target}
Predicted answer: {predicted_answer}

Grade the predicted answer of this new question as one of:
A: CORRECT
B: INCORRECT
C: PARTIAL_ANSWER
D: NOT_ATTEMPTED

Just return the letters "A", "B", "C", or "D", with no text around it.

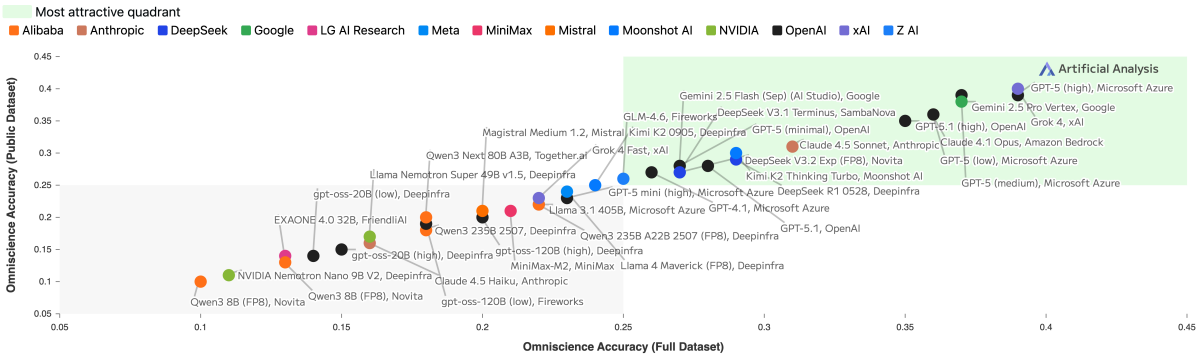

Figure 12: Calibration of results across public and full question set

B Open sourced questions

As part of our goal in furthering the development of models with reliable knowledge across domains, we have released a publicly available set of questions from *AA-Omniscience*.

In order to maintain evaluation integrity, we have released only 10% of the total questions, sampled to match the full distribution of domains and categories. If the entire question set, or even a large portion of it, were to be used in model development, the benchmark could become saturated very quickly. By preventing model contamination, the evaluation will remain a reliable measure of knowledge accuracy and hallucination tendencies across domains.

To ensure the public set is representative, questions were sampled so that model results across the public and full sets are closely aligned in performance, as shown in Figure 12.

The public set is sufficient to get an indication of overall model performance on the *Omniscience Index*, including knowledge accuracy and hallucination rates. Results at domain or category level, however, should not be considered reliable as the dataset size is too small to be representative of model performance in that area.