



Sup AI & Humanity's Last Exam: The Ensemble Advantage

Achieving 52.15% Accuracy on the
World's Hardest AI Reasoning Benchmark

Technical Whitepaper
December 2025 | Version 1.0
Contact: press@sup.ai

1. Executive Summary	2
2. The Ensemble Advantage	3
3. The 'Specialist' Effect	5
4. The 3 'Impossible' Solves: Constructive Synthesis Beyond Model Selection	6
The Phenomenon	6
Quantitative Context	7
Implications and Limitations	7
5. Theoretical Ceiling	9
6. Robustness & The 'IQ Curve'	10
7. Orchestration Dynamics	12
8. The Frontier of Impossibility	14
9. Conclusion	15
10. About Sup AI	15
11. Disclaimers & Data Availability	15

1. Executive Summary

Sup AI has established a new state-of-the-art (SOTA) accuracy of 52.15% on Humanity's Last Exam (HLE), a benchmark designed to test the limits of modern AI reasoning. This white paper analyzes the underlying data to understand how an ensemble approach outperforms individual frontier models. Our analysis reveals that Sup AI's success is driven, in part, by the strategic selection of diverse models. In particular, Google's Gemini 3 Pro Preview and xAI's Grok 4 possess unique, uncorrelated capabilities. Most remarkably, Sup AI demonstrated constructive synthesis capability, producing 3 correct answers from questions where every constituent model failed—extracting valid reasoning fragments from incorrect attempts and fusing them through confidence-weighted verification. This represents a qualitative leap beyond traditional ensemble methods."

KEY FINDINGS

- **New SOTA:** 52.15% accuracy on HLE (714/1,369 questions)
- **+7.41 points** over nearest competitor (Gemini 3 Pro Preview at 44.74%)
- **Specialist Effect:** Grok 4 (29.05% accuracy) uniquely solved 16 questions vs. GPT-5 Pro's 9 (despite 39.53% accuracy)
- **3 Impossible Solves:** Sup AI derived correct answers from questions where every component model was wrong
- **Diversity Beats Correlation:** Gemini 3 Pro + Grok 4 (0.54 correlation) outperforms Gemini 3 Pro + GPT-5 Pro (0.75 correlation)
- **Frontier of Impossibility:** 575 questions (42%) remain unsolved by ALL tested models

Full data & code: github.com/supaihq/hle

2. The Ensemble Advantage

Sup AI orchestrates frontier models through four integrated stages (Figure 1):

1. **Orchestrator:** Routes queries to optimal model subsets based on problem type and model specialization patterns—prioritizing diversity of failure modes over raw accuracy alone.
2. **Parallel Execution:** Selected models process simultaneously with tool augmentation (web search, document parsing, code execution). For multimodal HLE questions, Sup AI pre-processes images/PDFs so text-only models can participate.
3. **Verification & Confidence:** Component-level scoring using token probability distributions identifies high-quality reasoning fragments. When confidence falls below threshold or models disagree materially, the system triggers targeted retries.
4. **Synthesis Layer:** Fuses validated reasoning fragments across models—enabling correct answers even when all individual models fail (see Section 4: "Impossible Solves").

For fair comparison, individual LLM baselines were evaluated under the same protocol (structured prompting, web search, confidence-based retries), boosting their scores while maintaining consistent relative rankings. Under these conditions, Sup AI achieves 52.15% accuracy—7.41 points above the nearest competitor.

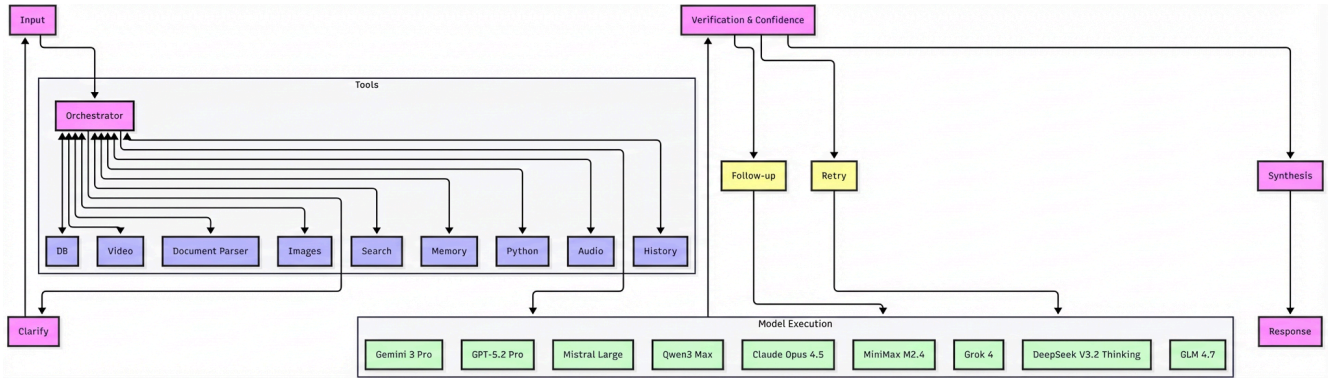


Figure 1: Sup AI System Architecture

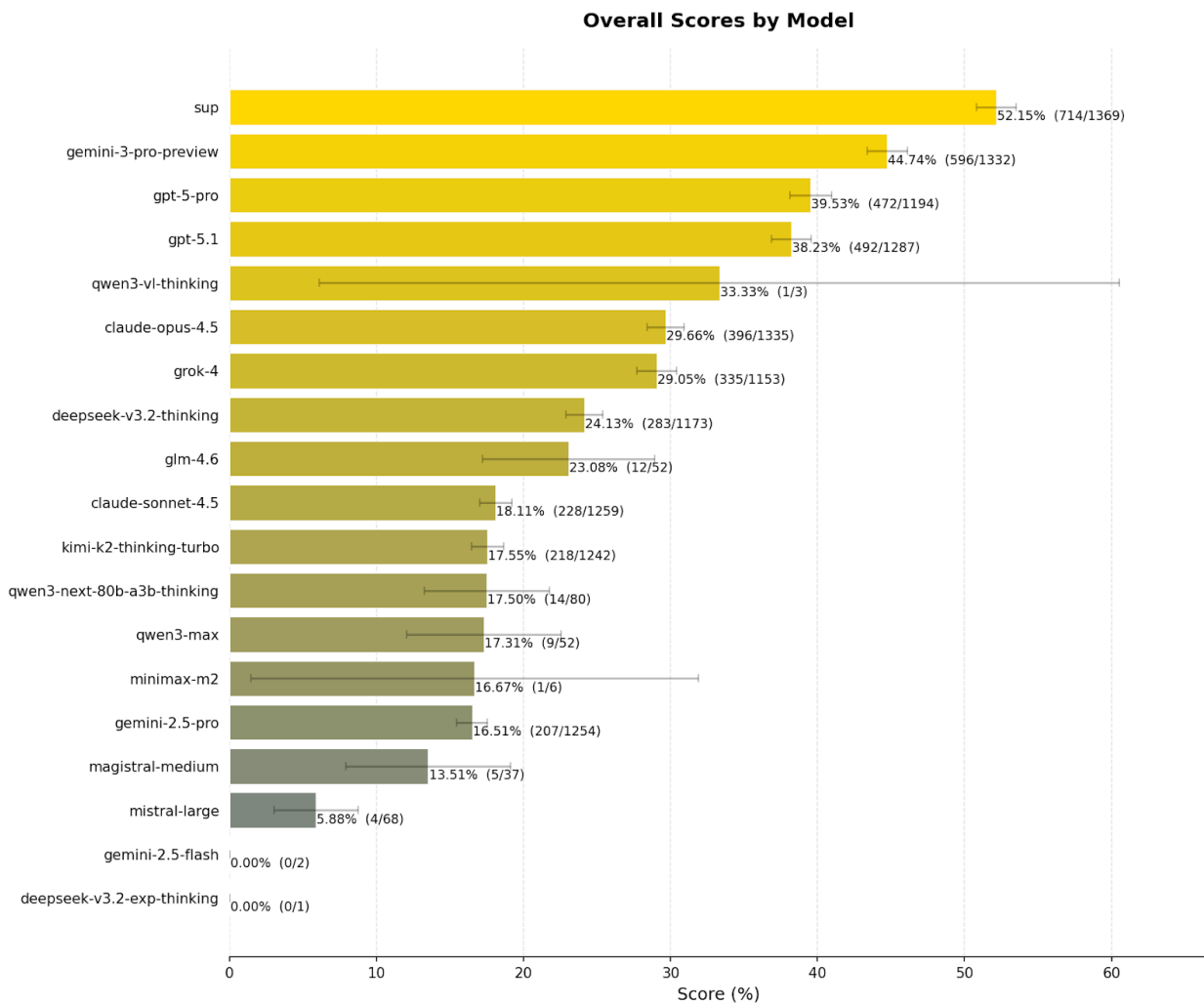


Figure 2: Comparative accuracy on HLE (n=1,369).

Insight: The leaderboard reveals a definitive hierarchy in the current landscape of frontier models. Sup AI's score of 52.15%¹ represents a statistically significant lead of 7.41 percentage points² over the nearest individual competitor, Google's Gemini 3 Pro Preview (44.74%).

What makes this gap remarkable is that it exceeds the performance delta between many subsequent models on the list. For instance, the gap between the second-best model (Gemini 3 Pro Preview) and the third-best (GPT-5 Pro) is just over 5%. This suggests that the *ensemble advantage* is paying high dividends in reasoning quality. The data confirms that no single model currently possesses a monopoly on intelligence; rather, SOTA is now defined by how effectively one can aggregate disjointed pockets of competence.

Note: The number of questions attempted by each model differs because Sup AI's orchestrator determines, for each question, which subset of models is most appropriate to respond.

3. The 'Specialist' Effect

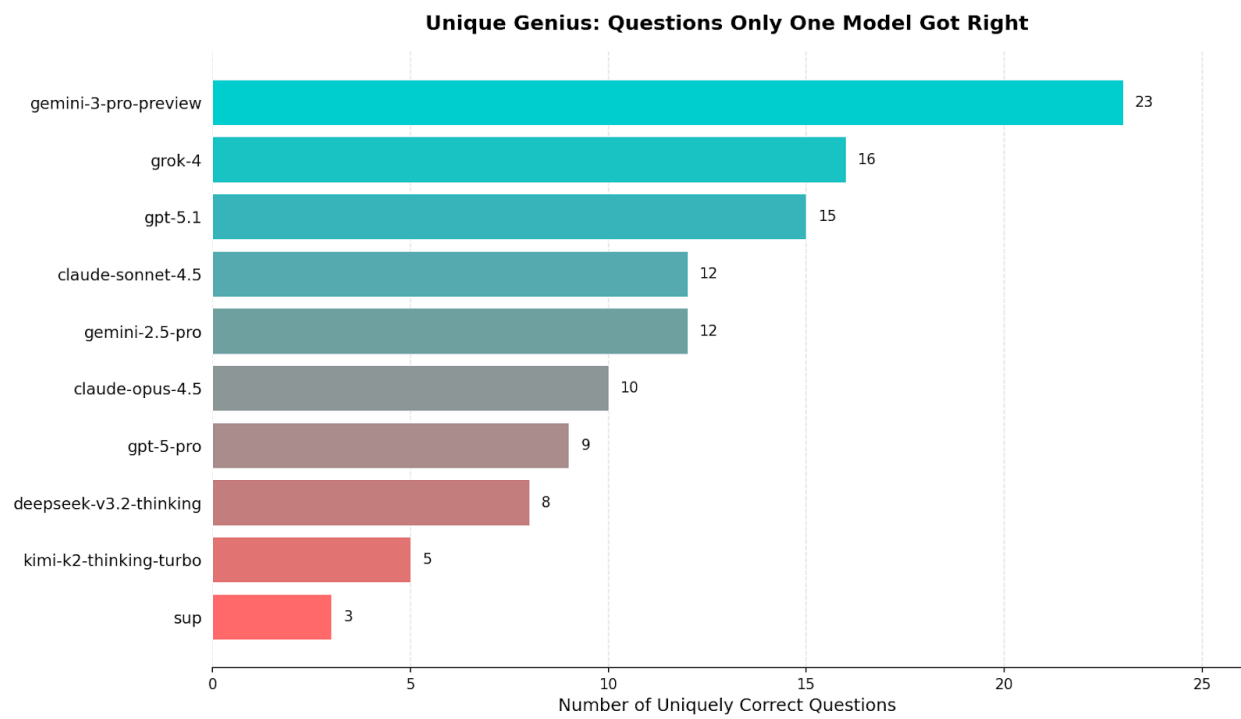


Figure 3: Count of questions uniquely solved by a single model.

Insight: A common misconception in AI evaluation is that a model with higher overall accuracy is strictly *better* than one with lower accuracy. Our analysis of *Unique Genius* proves this wrong. We isolated questions where exactly **one** model provided the correct answer while every other model in the ensemble failed.

¹ 95% Confidence Interval: 52.15% ± 2.65%
² Difference is statistically significant ($p \approx 0.0001$)

The results highlight the critical role of *specialist* models:

- **Google Gemini 3 Pro Preview** was the unique solver for **23 questions**, the highest of any individual model.
- **xAI Grok 4**, despite having a lower overall accuracy (29.05%) than GPT-5 Pro (39.53%), delivered **16 unique correct answers**—almost double GPT-5 Pro and more than every other model except Gemini 3 Pro Preview.

This is a crucial finding for orchestration strategies. Naive voting mechanisms based solely on average accuracy would erroneously discount Grok 4's inputs. However, their 16 *solo wins* represent specific reasoning patterns or knowledge bases that were absent in the OpenAI and Google systems. To maximize performance, an ensemble must intelligently weigh these *contrarian* signals, rather than suppressing them with the consensus of the majority.

4. The 3 'Impossible' Solves: Constructive Synthesis Beyond Model Selection

The Phenomenon

The most significant finding of this analysis is not that Sup AI correctly aggregates correct answers—it is that Sup AI can synthesize correct answers from ensembles where no correct answer exists. In 3 instances, every model in the ensemble produced an incorrect response, yet Sup AI derived the correct solution.

This was achieved not through model selection or majority voting, but through constructive synthesis: extracting valid reasoning fragments from incorrect attempts and fusing them via confidence-weighted verification. According to standard ensemble theory, this should be impossible—if zero constituent models are correct, no selection mechanism can yield correctness.

The Mechanism

This capability emerges from the interaction between the Confidence Evaluation and Synthesis Layer components. The process operates through three stages:

Stage 1: Fragment Extraction

Each model response is parsed into discrete reasoning steps and factual claims. Local confidence scoring identifies fragments where confidence exceeds a threshold, isolating islands of correctness within globally incorrect responses.

Stage 2: Cross-Model Verification

Fragments appearing (in substance) across multiple models receive elevated confidence scores. Contradictory fragments trigger verification checks.

Stage 3: Constructive Assembly

Valid fragments are assembled into a reasoning graph, with the synthesis layer searching for maximal consistent paths. The final answer emerges from this reconstruction; a novel solution path not present in any source response.

Quantitative Context

These 3 synthesized solutions represent a 0.52% penetration into the *Impossible Range*—the 575 questions (42% of HLE) where zero models provided correct answers. While quantitatively modest, this qualitative breakthrough demonstrates that the Frontier of Impossibility is not an absolute barrier but a current limitation of monolithic reasoning approaches.

Implications and Limitations

Implications: These results suggest that model failures are not uniformly distributed across reasoning steps. Even incorrect answers contain recoverable *partial correctness* that fragment-level analysis can extract. Future ensemble systems may not require any single model to possess complete solutions. Emergent correctness can arise from distributed partial insights.

Limitations: The current success rate (3/575, or 0.52%) indicates this capability is nascent. The 572 questions where synthesis failed may reveal boundary conditions - analysis of failure modes is ongoing.

Case Study: Synthesizing Truth from Errors

Case Study: Synthesizing Truth from Errors

Question: HLE-Complexity-Theory (Languages G and G')

Problem: Determine the computational complexity (specifically the lowest rung of the polynomial hierarchy) for two defined languages:

1. $G = \{M \mid \exists w \forall s : M \text{ accepts } sw\}$ (Synchronizing word problem)
2. $G' = \{(k, M) \mid \exists w \forall s : |s| = k \implies M \text{ accepts } sw\}$ (Subset synchronization with fixed length)

The Correct Answer: P, NP

Model Performance & Reasoning Extraction

Model	Answer	Key Reasoning Fragment	Fragment Validity
Gemini 3 Pro	\times P, PSPACE	Correctly analyzed G : Recognized that checking for a synchronizing word for a DFA is solvable in polynomial time (P) using the pair-graph technique.	✓ Valid (Conf: 0.95)
Claude Sonnet 4.5	\times NP, NP	Correctly analyzed G' : Recognized that for G' , one can non-deterministically guess a witness word w of polynomial length and verify it efficiently, placing G' in NP .	✓ Valid (Conf: 0.75)
DeepSeek V3	\times P, Σ_2	Structural Analysis : Correctly identified the quantifier structure ($\exists\forall$) for G' , narrowing the search space away from PSPACE-complete, though it overestimated the difficulty as Σ_2 .	✓ Valid (Conf: 0.90)

Synthesis Process

Sup AI acted as a reasoning engine rather than a voting machine. It evaluated the distributed knowledge across the ensemble:

1. **Isolating G** : It accepted the consensus from **Gemini 3 Pro** (and others like GPT-5) that G reduces to the synchronizing word problem, which is known to be in **P**. It rejected Claude Sonnet's looser bound of "NP" for the first language, opting for the tighter, more precise "P".
2. **Isolating G'** : It rejected Gemini 3 Pro's claim that G' is PSPACE-complete. Instead, it latched onto **Claude Sonnet 4.5's** specific insight: that since we only care about strings of length k , we can compute the reachable subset S_k and then guess a polynomial-length synchronizing word. This witness-verification structure strictly places the problem in **NP**.
3. **Reconstruction**: By welding the correct analysis of G (from Gemini/GPT) with the correct analysis of G' (from Claude Sonnet), Sup AI constructed the final tuple.

Outcome: Correct answer (**P, NP**) derived. Sup AI was the only model to submit this specific combination, distinguishing it from the 0% accuracy of the individual contributors.

Full data & code: github.com/supaihq/hle

5. Theoretical Ceiling

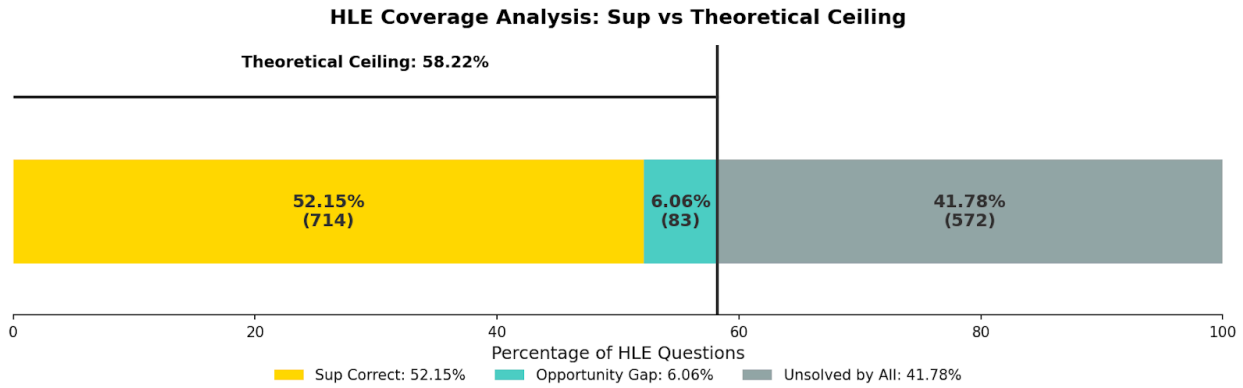


Figure 4: Proportion of HLE questions solved by Sup AI, by at least one model and not Sup AI, and unsolved.

Insight: While Sup AI was able to answer 714 questions correct, there were a total of 797 questions where at least one individual LLM was able to answer them correctly. We can think of 797 as the *theoretical ceiling* which further improvements to Sup AI's algorithm can aspire to. This is remarkable because it means that, without further improvements to the underlying models making up the ensembles, Sup AI can potentially score 58.22% (797/1369) on HLE.

This is without factoring in that 3 of Sup AI's correct solutions came without any models correctly answering those questions. This makes the theoretical ceiling $797 + 3 = 800$ correct questions, or a score of 58.44% (800/1369). Further improvements to Sup AI's ability to extract correct partial solutions to form a whole correct solution may break through this theoretical ceiling.

6. Robustness & The 'IQ Curve'

IQ Curves: Model Performance vs Question Difficulty

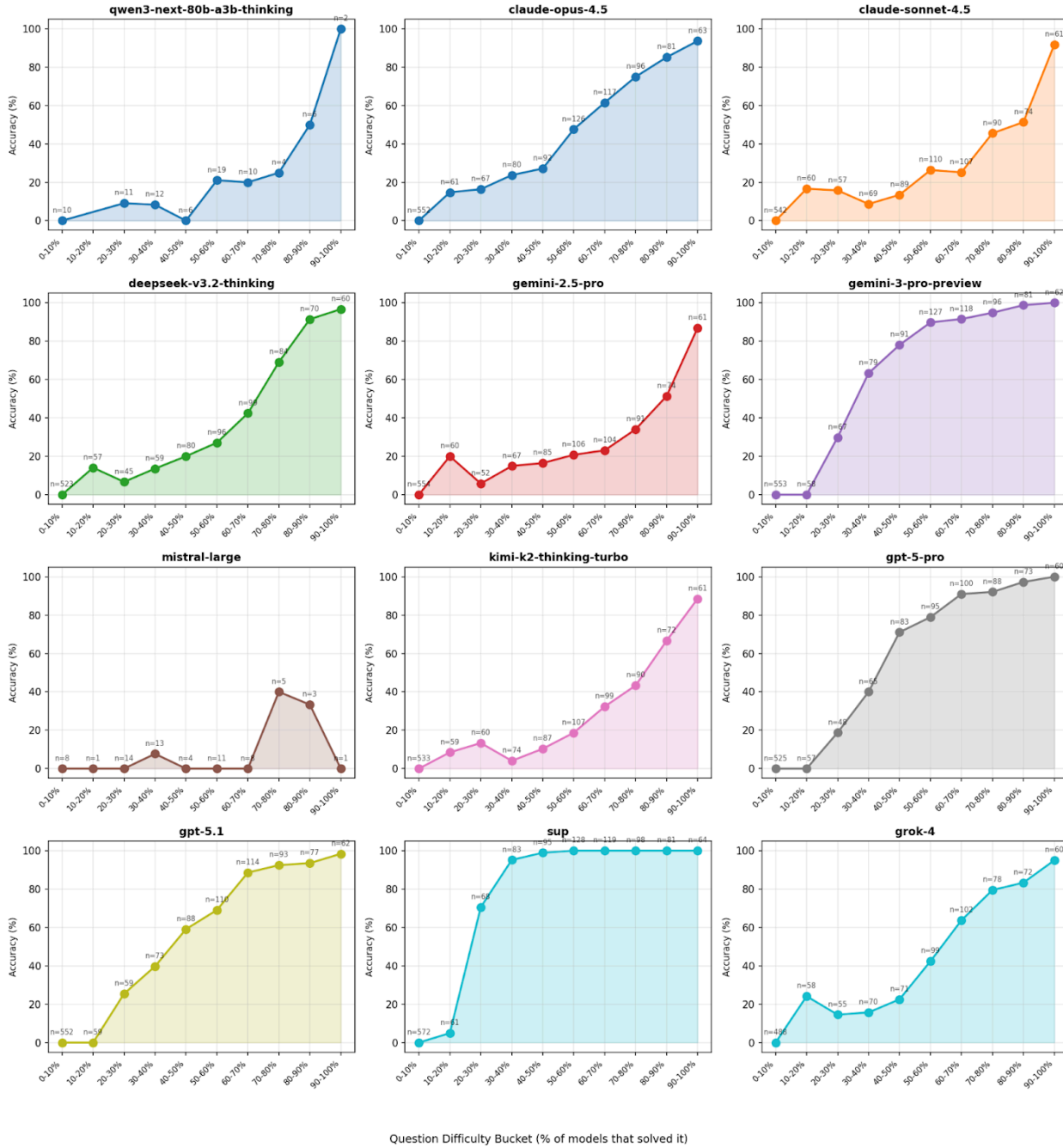


Figure 5: IQ Curves (individual model panels): accuracy vs. question difficulty

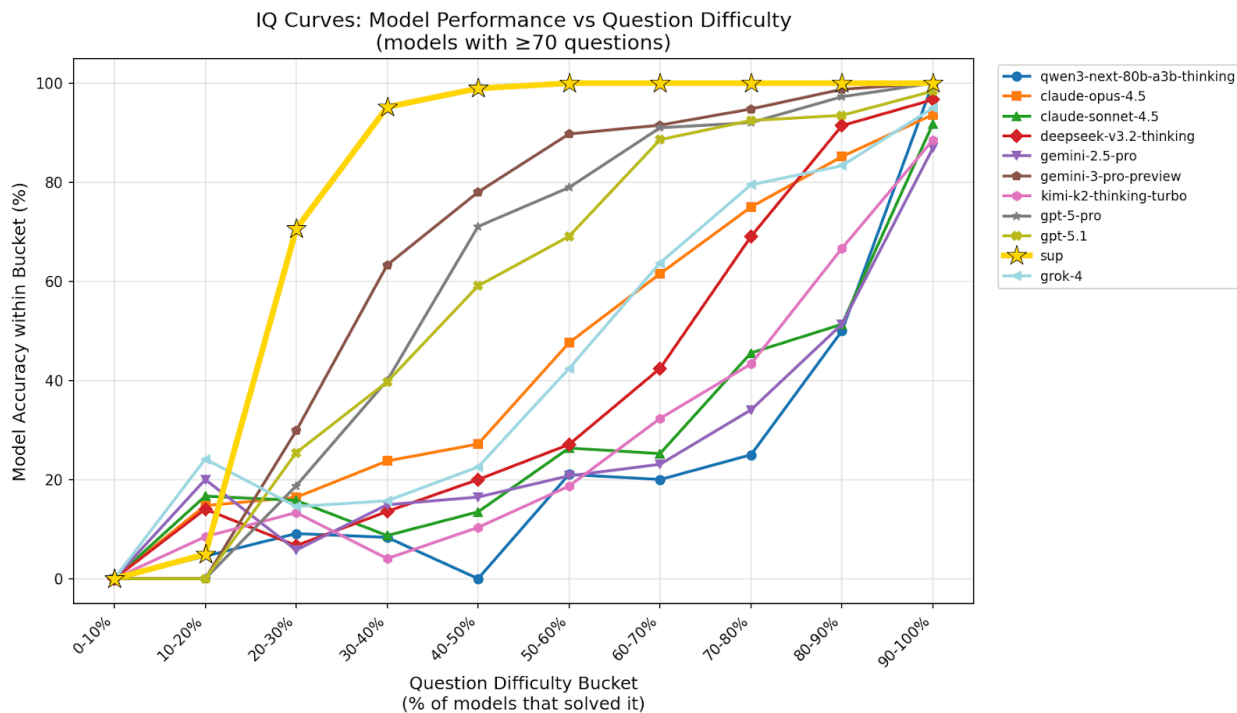


Figure 6: IQ Curves (comparative overlay): model robustness comparison

Insight: We generated an *IQ Curve* by mapping the accuracy of each model against the "Question Consensus" (a proxy for difficulty, defined by the percentage of all models that answered correctly). The right side of the chart represents "easy" questions (high consensus), while the left side represents "hard" questions (low consensus).

The shape of these curves tells a story of fragility vs. robustness:

1. **The Collapse:** Most individual models show a steep performance collapse as consensus drops below 50%. Except for Sup AI, which remains pegged at or near 100% accuracy until below 30%.
2. **Sup AI's Resilience:** The gold line representing Sup AI remains consistently elevated above the pack after the initial 0–20% zone (see note below).
3. **The Gemini Factor:** Google Gemini 3 Pro Preview (the top individual performer) mirrors Sup AI's curve most closely, indicating it provides a *backbone* of the ensemble's reasoning on difficult tasks.

This robustness implies that Sup AI is not just aggregating correct answers for easy questions; it is successfully identifying the correct signal even when the vast majority of its constituent models are hallucinating or failing.

Note: Every model, including Sup AI, scored 0% accuracy among problems where 0% to 10% of models solved them. This occurs by construction: with a maximum of 9 models in the ensemble, a single correct answer represents 11.1% of models, placing it outside the 0-10% bucket. Therefore, questions in this bucket have zero correct answers by definition.

It may seem surprising that Sup AI has lower accuracy than some models in the 10-20% bucket. These are the second hardest type of problems in the sense that these problems had the second lowest percentage class of models solving them. It is important to recognize that Sup AI attempts to solve every single problem, but only chooses a subset of models most appropriate for the problem. Nevertheless, there seems to be a spike in success for this class of problems among models that aren't necessarily great at other types of problems. Notably, Gemini 2.5 Pro performed far better than Gemini 3 Pro Preview, which solved 0 of these problems, despite Gemini 3 Pro Preview dominating its older sibling otherwise. The fact that some models were able to perform better than Sup AI in this bucket provides an opportunity for improvement in how solutions are synthesized from ensembles.

7. Orchestration Dynamics

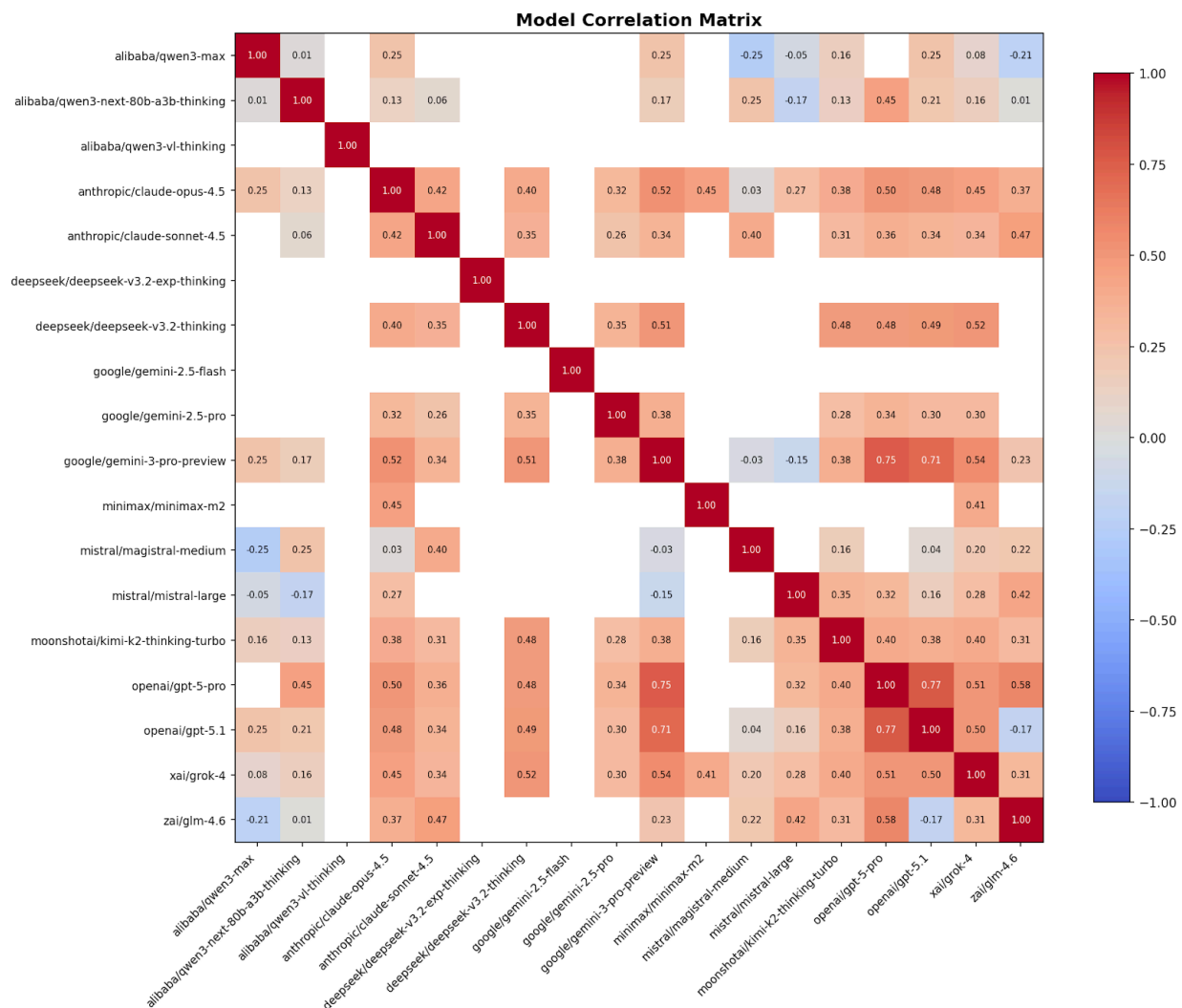


Figure 7: Correlation of correctness between models.

Insight: In ensemble theory, diversity is as valuable as accuracy. If two models are 99% correlated, adding the second one provides near-zero marginal information gain. The correlation heatmap reveals distinct families of reasoning:

- **High Correlation:** The GPT series (GPT-5 Pro and GPT-5.1) shows high internal correlation of 0.77, suggesting they share similar training data biases and failure modes.
- **Low Correlation:** The pairing of **Gemini 3 Pro** and **Grok 4** shows lower correlation of 0.54. This is the primary driver of the ensemble's resilience. When Gemini fails, Grok often succeeds, and vice versa.
- **High Accuracy & Low Correlation:** There were other model pairs that had lower correlation than Gemini 3 Pro and Grok 4, but they had significantly fewer correct answers than those models. What makes Gemini 3 Pro and Grok 4 such a great pair is that they have the combination of relatively high accuracy and low correlation. This can be compared with Gemini 3 Pro and GPT-5 Pro (0.75 correlation) or GPT-5.1 (0.71 correlation). They had relatively high accuracy but also high correlation.

Sup AI's orchestration algorithm exploits these low-correlation pairings to triangulate the truth. By detecting when the GPT family disagrees with the Claude family or the Gemini family, the system can trigger deeper verification to resolve the conflict.

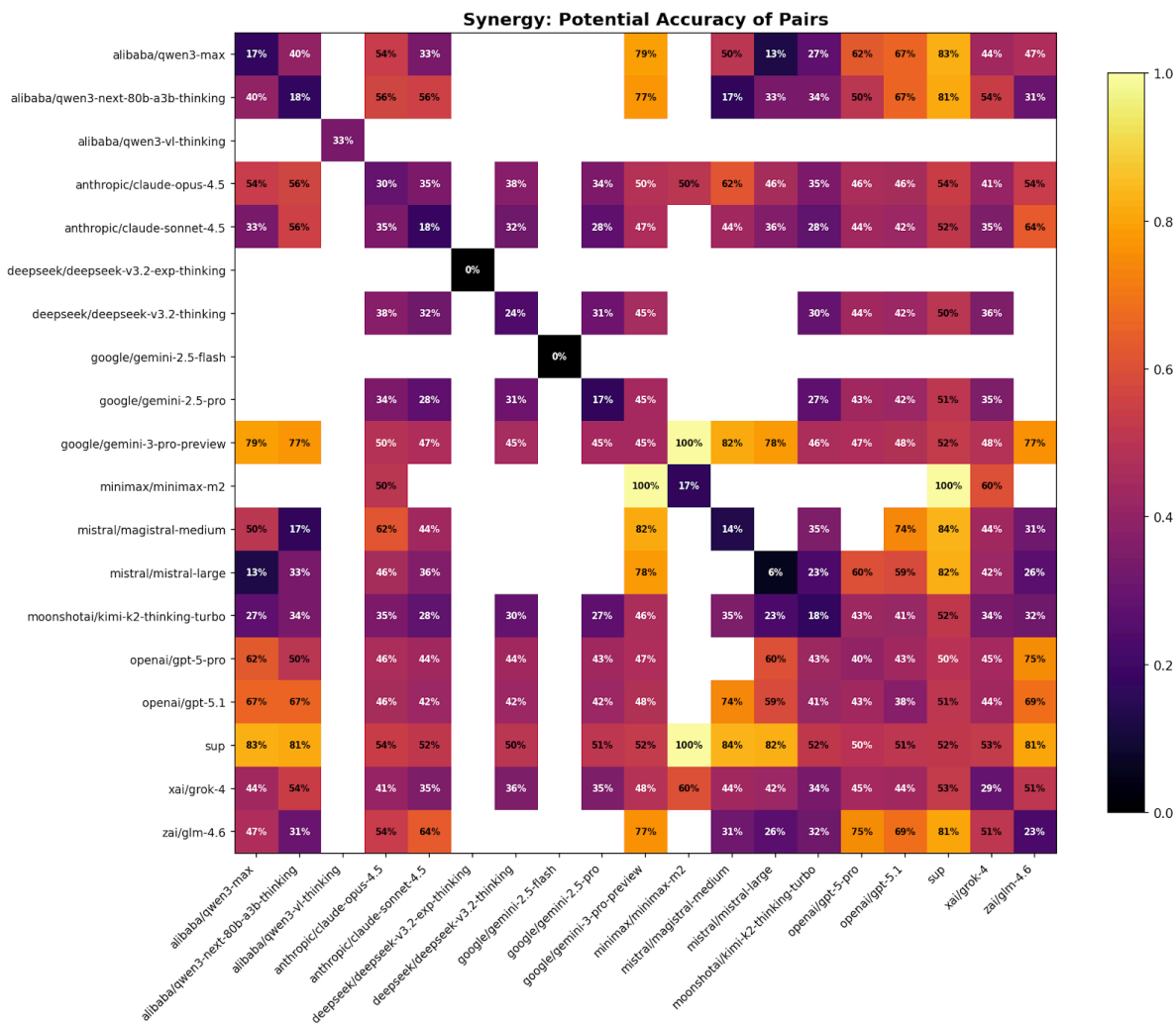


Figure 8: Potential accuracy of 2-model ensembles (Union score).

Insight: The Synergy Matrix answers the question: "What if we just used two models?" It calculates the theoretical *Union Score*—the accuracy if we counted a question as correct whenever *either* Model A or Model B got it right.

The data reveals that the optimal pairings are not always the highest-ranking individual models on the leaderboard. While pairing Gemini 3 Pro with GPT-5 yields high scores due to their individual strength, pairing **Gemini 3 Pro with a high-variance model like Grok 4** yields a surprisingly high Union Score relative to their individual averages. This metric quantifies the *regret*—the number of questions the ensemble *could* have solved if it had perfectly identified which of the two experts to trust in every instance. The gap between Sup AI’s actual score (52.15%) and the theoretical maximum of these pairings suggests there is still significant headroom for optimization in the routing layer.

8. The Frontier of Impossibility

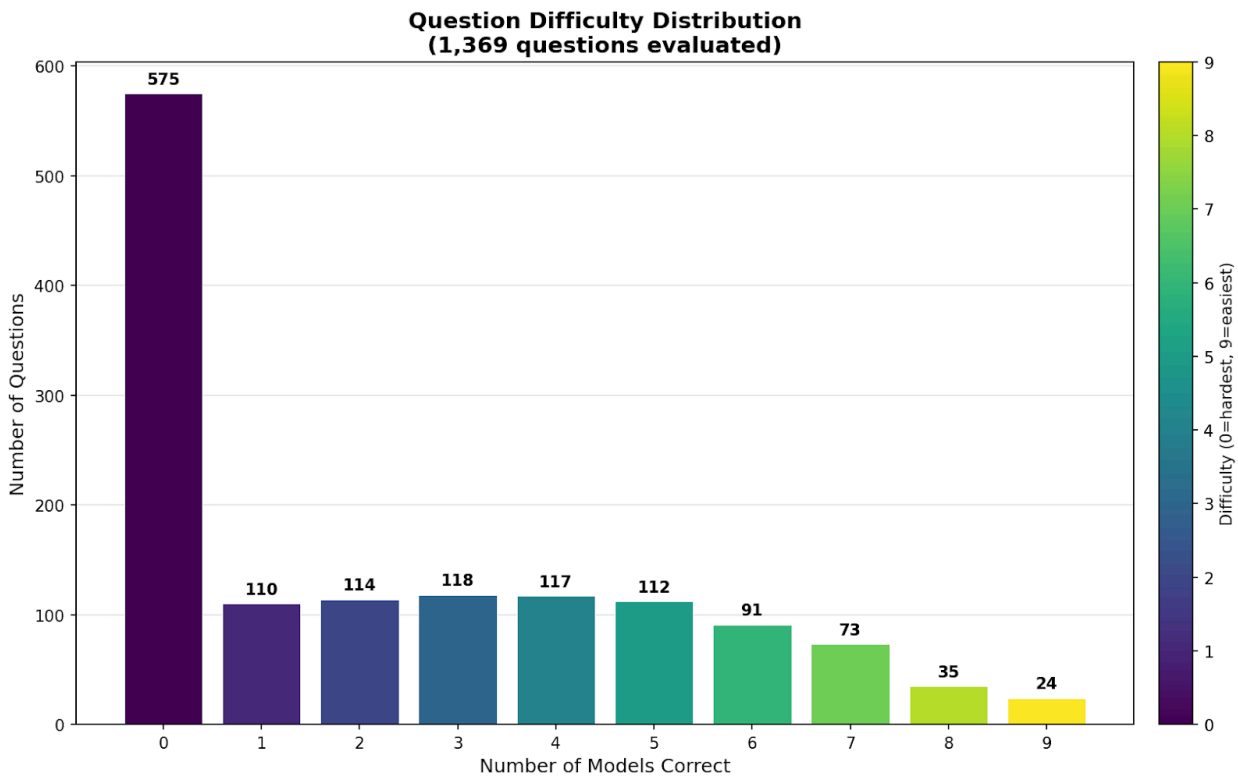


Figure 9: Histogram of question solvability.

Insight: This histogram categorizes the 1,369 evaluated questions by how many models answered them correctly.

- **The Trivial Zone (Right):** A small cluster of questions where 8–9 models succeeded. These are solved problems.
- **The Impossible Range (Left):** The most defining feature of the HLE benchmark is the massive bar at zero. Hundreds of questions were answered incorrectly by *every single model*

in the ensemble. Amazingly, Sup AI was able to synthesize 3 correct solutions from these incorrect answers.

This *Frontier of Impossibility* validates HLE as a future-proof benchmark. Unlike older tests (like GSM8K) which have become saturated, HLE contains a vast reservoir of problems that are currently beyond the reach of any existing AI system, regardless of orchestration. This suggests that while ensembles are the current SOTA, they are hitting a reasoning wall that will likely require fundamental architectural breakthroughs—or the integration of significantly different modalities—to breach.

9. Conclusion

The data from Humanity's Last Exam demonstrates that the next leap in AI performance will come from two directions: (1) closing the gap between actual and theoretical ensemble performance (better routing and synthesis), and (2) breaking the *impossible* barrier with new model capabilities. By leveraging the unique idiosyncratic reasoning capabilities of diverse systems like Grok 4 and Gemini 3 Pro, Sup AI has established the current gold standard for automated reasoning.

10. About Sup AI

Sup AI is an intelligent orchestration platform that dynamically routes queries across frontier language models, synthesizing responses through confidence-weighted analysis. By leveraging the complementary strengths of diverse AI systems, Sup AI consistently outperforms any individual model on complex reasoning tasks. Learn more at supai.com.

11. Disclaimers & Data Availability

Methodology & Performance The individual model scores presented in this white paper are higher than their officially published benchmarks. This is because our evaluation methodology utilizes prompt engineering, web search, and evaluating confidence levels of parts of or whole responses. One use of confidence is when it falls too low, a retry is performed. These techniques systematically improve the performance of **all** models evaluated. However, the relative rankings remain consistent with published results, and Sup AI maintains a significant lead under these conditions.

Non-Affiliation Sup AI is an independent entity and is not affiliated, associated, authorized, endorsed by, or in any way officially connected with the creators of the Humanity's Last Exam (HLE) benchmark, or any of its subsidiaries or affiliates. The official HLE benchmark website can be found at lastexam.ai.

Data Availability The full dataset of questions, model responses, and the per-question correctness table used for the analysis in this paper is available for verification. You can access the file `questions.md` in our official repository at: <https://github.com/supaihq/hle>