Tribunal Session Summary

Session: conclave-20260603-231334-is-10x-compute-required-b910

Depth: T6

Advocates: 6 | Judges: 5 | Cost: $5.8145 | Time: 39m 16s

THE PROMPT: Is "10x Compute" required for each doubling of model intelligence?

THE RULING: No. The "10x compute per doubling of intelligence" heuristic is not a physical law, nor is it a strict requirement for advancing model capability. It is a historical budgeting pattern observed during a specific era of dense pretraining that has been mathematically refuted by scaling-law exponents and empirically falsified by recent advances in algorithmic efficiency, data curation, and inference-time compute.

The Prompt

Is "10x Compute" required for each doubling of model intelligence?

The Ruling

No. The "10x compute per doubling of intelligence" heuristic is not a physical law, nor is it a strict requirement for advancing model capability. It is a historical budgeting pattern observed during a specific era of dense pretraining that has been mathematically refuted by scaling-law exponents and empirically falsified by recent advances in algorithmic efficiency, data curation, and inference-time compute.

The strongest evidence against the rule comes from the mathematics of the scaling laws themselves. Published research (Kaplan et al., 2020; Hoffmann et al., 2022) establishes that cross-entropy loss scales as a power law with compute ($L \propto C^{-\alpha}$) using exponents (α) between approximately 0.05 and 0.16. Under these exponents, a 10x increase in compute reduces the reducible loss gap by only 11% to 31%. To actually halve the loss—a proxy often conflated with capability gains—requires roughly 90x to 1,000,000x more compute, not 10x. Furthermore, capability doublings (operationalized as error-halving on benchmarks) are routinely achieved through mechanisms other than raw pretraining scale, such as algorithmic efficiency improvements (which yield ~2-3x effective compute gains annually) and inference-time scaling (e.g., Chain-of-Thought prompting), which deliver substantial performance jumps without additional pretraining FLOPs. While frontier labs have historically spent roughly 10x more compute per generation, this reflects path-dependent industrial choices and risk management, not a fundamental constraint on intelligence scaling.

Core Analysis

The claim that "10x compute is required for each doubling of model intelligence" fails under mathematical scrutiny, empirical observation, and operational definition. The winning position demonstrates that intelligence scaling is multi-dimensional, and the 10x heuristic is an artifact of a narrow, historical paradigm rather than a universal law.

The Mathematical Refutation: Scaling Law Exponents

The foundational evidence against the 10x rule lies in the power-law equations governing neural language model training. The relationship between compute (C) and loss (L) is defined as $L(Gemini 3.1 Pro) = L_{\infty} + k \cdot C^{-\alpha}$.

- **The Exponent Mismatch:** Peer-reviewed literature establishes the compute exponent (α) for language models in the range of ~0.05 (Kaplan et al., 2020) to ~0.154 (Hoffmann et al., 2022, Chinchilla).
- **The Compute Cost of Loss Halving:** To halve the reducible loss component ($L - L_{\infty}$), the required compute multiplier is $2^{1/\alpha}$.

  - At $\alpha = 0.154$ (Chinchilla), halving loss requires $2^{1/0.154} \approx 90\times$ compute.
  - At $\alpha = 0.05$ (Kaplan), halving loss requires $2^{1/0.05} \approx 1,000,000\times$ compute.
- **The 10x Reality:** A 10x increase in compute yields only a marginal reduction in the reducible loss gap—specifically, a multiplier of $10^{-\alpha}$, which results in an 11% to 31% improvement depending on the exponent.
- **Conclusion:** If "doubling intelligence" is mapped to "halving loss," the 10x rule is mathematically insufficient by one to five orders of magnitude. If "doubling intelligence" is *not* mapped to loss halving, the rule lacks a derived mathematical foundation altogether.

Algorithmic Efficiency as a Compute Substitute

Algorithmic progress acts as a force multiplier that decouples capability gains from raw physical FLOPs.

- **The Efficiency Trend:** Research by Epoch AI (Ho et al., 2024) indicates that the compute required to reach a fixed performance threshold halves approximately every 8 to 10 months. This corresponds to a ~2.5x to 3x annual increase in compute-equivalent efficiency derived solely from better algorithms, data curation, and optimization.
- **The Intercept Shift:** This efficiency gain shifts the *intercept* of the scaling curve rather than changing its exponent. A model trained today with 1x physical compute can achieve the performance that required 3x physical compute two years ago.
- **Implication:** The "10x requirement" assumes a static algorithmic baseline. In reality, the effective compute available to frontier labs grows significantly faster than their hardware purchases due to these compounding software improvements.

Inference-Time Compute: A New Scaling Axis

The most significant challenge to the pretraining-centric 10x rule is the emergence of inference-time scaling.

- **Chain-of-Thought (CoT) & Self-Consistency:** On the GSM8K benchmark, techniques like CoT and self-consistency improved accuracy from ~18% to ~58% (Wei et al., 2022; Wang et al., 2023) without any additional pretraining compute. This represents a >3x absolute gain (and a halving of error rate) achieved purely by allocating more compute during inference.
- **OpenAI o1:** The o1 model improved AIME math accuracy from ~13% (GPT-4o baseline) to ~83% without a 10x increase in pretraining compute. The gains were driven by Reinforcement Learning on chain-of-thought traces and test-time search.
- **Mechanism:** These methods trade cheap inference-time search for expensive pretraining FLOPs. They prove that "intelligence" (reasoning capability) can be doubled by spending compute at generation time, invalidating the assumption that pretraining compute is the sole bottleneck.

Recipe Optimization and Data Quality

The Chinchilla experiment (Hoffmann et al., 2022) serves as a definitive counterexample to the necessity of 10x scaling.

- **The Result:** Chinchilla (70B parameters, 1.4T tokens) outperformed Gopher (280B parameters, 300B tokens) on 56 out of 62 benchmarks.
- **The Compute Parity:** Both models utilized roughly similar total training FLOPs (on the order of $5-6 \times 10^{23}$).

- **The Mechanism:** The capability jump came from optimizing the parameter-to-token ratio (compute-optimal training), not from increasing the compute budget. This demonstrates that substantial "intelligence" gains can be extracted from a fixed compute envelope by changing the training recipe.

The Failure of the "YES" Position

The initial "YES" arguments collapsed under evidentiary pressure.

- **Mathematical Errors:** Magistral Medium relied on an exponent of 0.3 to derive the 10x rule. This exponent does not exist in the peer-reviewed literature for language model compute scaling (which cites 0.05–0.16), rendering the derivation invalid.
- **Anecdotal Evidence:** Grok 4 relied on industry statements from Elon Musk and Morgan Stanley reports to prove the rule. These are investment heuristics, not scientific measurements of a causal law.
- **Lack of Cardinal Metric:** No advocate produced a peer-reviewed, cardinal definition of "intelligence" or a validated function mapping loss reduction to "intelligence doubling." Without this, the 10x rule is unfalsifiable rhetoric.

Surviving Counterarguments

While the YES position failed, valid points were raised regarding the *frontier*.

- **Frontier Budgeting:** It is true that frontier labs (Anthropic, OpenAI) have historically spent ~10x more compute per generation. However, this is a description of *spending behavior*, not a proof of *necessity*. Labs choose 10x jumps to minimize risk and guarantee progress, not because 3x or 5x is impossible.
- **Data Exhaustion:** As high-quality text data runs out (projected ~2026–2028), the cost of further scaling may rise. However, this reinforces the need for algorithmic efficiency and synthetic data, rather than validating a fixed 10x multiplier.

How We Got Here

Scorecard

Judge	Grok 4	Claude Opus 4.7	Gemini 3.1 Pro	Magistral Medium	DeepSeek V4 Pro	GPT-5	Winner
GPT-OSS 120B	REJECT	ADOPT	ADOPT	REJECT	ADOPT	ADOPT	Claude Opus 4.7
Zhipu GLM-5	REJECT	ADOPT	ADOPT	REJECT	ADOPT	ADOPT	GPT-5
Zhipu GLM-4.7	REJECT	ADOPT	ADOPT	REJECT	ADOPT	ADOPT	Synthesis

Judge	Grok 4	Claude Opus 4.7	Gemini 3.1 Pro	Magistral Medium	DeepSeek V4 Pro	GPT-5	Winner
DeepCogito v2.1 671B	REJECT	ADOPT	ADOPT	REJECT	ADOPT	ADOPT	Synthesis
Mistral Large 3	REJECT	ACCEPT	ADOPT	REJECT	ADOPT	ADOPT	Claude Opus 4.7

Convergence Assessment

The convergence was **epistemic**. The shift in position by Grok 4 and Magistral Medium was triggered by verifiable corrections: the exposure of the incorrect 0.3 exponent and the lack of peer-reviewed support for the 10x-to-2x mapping. There was no evidence of sycophantic drift; the NO advocates (Claude Opus 4.7, Gemini 3.1 Pro, DeepSeek V4 Pro, GPT-5) maintained their positions stably despite the eventual unanimity, and the YES advocates only moved after their evidentiary pillars were demolished.

Closing Statements

Grok 4: A conceded the original YES position, acknowledging that 10x compute is not a derived law or universal requirement. A refined the stance to "10x is a useful budgeting heuristic under fixed paradigms," noting that algorithmic efficiency and inference-time scaling break the rule for general capability doublings.

Claude Opus 4.7: B maintained a confident NO, arguing that the 10x rule is a folk heuristic conflating loss scaling with intelligence. B highlighted that published exponents (0.05–0.154) make 10x insufficient for loss halving, and that algorithmic efficiency and inference-time compute (o1) routinely deliver capability doublings at lower multipliers.

Gemini 3.1 Pro: C held to NO, focusing on the "intercept shift" from algorithmic efficiency and the "inference axis." C conceded the S-curve argument but emphasized that algorithmic progress (~2.8x/year) and data quality (Phi-1) mathematically substitute for physical FLOPs, making the 10x requirement an economic artifact, not a technical one.

Magistral Medium: D conceded the YES position, admitting the 0.3 exponent was incorrect and that the 10x figure is an approximation. D acknowledged that factors like model architecture and training methods alter the compute-to-performance relationship, concluding the rule is not a universal necessity.

DeepSeek V4 Pro: E held to NO, characterizing the 10x rule as an artifact of the dense transformer paradigm. E cited Chinchilla and MoE as proof that capability doublings occur at <10x compute, and argued that the rule fails under any rigorous operationalization of "intelligence."

GPT-5: F held to NO, providing the most rigorous mathematical critique. F operationalized "doubling" as error-halving on benchmarks and demonstrated that 10x compute yields only marginal loss reduction. F cited CoT/self-consistency and Chinchilla as definitive counterexamples to the necessity claim.

Reality Checks

Claim	Challenge	Result
Magistral Medium: "For a typical exponent of 0.3... doubling performance requires roughly 10x compute."	**DeepSeek V4 Pro/B:** The exponent for LLM loss scaling is ~0.05–0.16 (Kaplan/Hoffmann), not 0.3. 0.3 is invented.	**Full Concession.** D admitted the exponent was incorrect and the derivation invalid.
Claude Opus 4.7: "To halve loss... you need ~10^(1/0.154) ≈ 10^6.5x compute."	**DeepSeek V4 Pro:** Your arithmetic is wrong. $2^{1/0.154} \approx 90$, not $10^{6.5}$. You used base-10 instead of base-2.	**Full Concession.** B corrected the math to ~90x, noting this still refutes the 10x rule.
Grok 4: "Morgan Stanley (2026) and Elon Musk cite that applying 10x compute... doubles model intelligence."	**DeepSeek V4 Pro/F:** This is financial analysis, not peer-reviewed evidence. It is circular reasoning (bankers citing labs citing folklore).	**Full Concession.** A admitted this is an industry heuristic, not a scientific law.
Gemini 3.1 Pro: "Wei et al. (2022) demonstrated that downstream... capabilities exhibit 'emergent' S-curve phase transitions."	**DeepSeek V4 Pro:** Schaeffer et al. (2023) proved these are artifacts of nonlinear metrics, not real phenomena.	**Full Concession.** C abandoned the S-curve argument.

Next Steps

Next Step	Why It Matters	Confidence
Define a cardinal metric for "intelligence"	The debate stalled on the definition of "doubling." A standardized, economic, or task-based metric is needed to make scaling claims falsifiable.	Medium (Mistral Large 3)
Validate inference-time scaling laws	Snell et al. (2024) suggests inference compute substitutes for training compute, but this is preprint. Peer review is needed to confirm this new axis.	High (Mistral Large 3)
Monitor data exhaustion impacts	Villalobos et al. (2024) projects high-quality text exhaustion by 2026–2028. This may force a shift back to compute-heavy synthetic data or alter scaling dynamics.	High (Mistral Large 3)

Dissenting Opinions

No dissenting opinions were filed in this session.

Glossary

Term	Definition
Chinchilla Scaling Laws	A set of scaling laws (Hoffmann et al., 2022) establishing the compute-optimal trade-off between model size and training data tokens for language models.
Compute-Optimal Training	Training a model at the point on the scaling curve where it achieves the lowest possible loss for a given fixed compute budget.
Cross-Entropy Loss	A loss function measuring the difference between two probability distributions (the model's predictions and the true labels). In LLMs, it is the primary metric minimized during pretraining.
Emergent Abilities	Capabilities that are not present in small models but arise unpredictably in large models, often associated with "phase transitions" in performance.
Inference-Time Compute	Computational resources expended by a model when generating a response (processing tokens), distinct from the compute used during training.
Mixture-of-Experts (MoE)	A sparse architecture where only a subset of parameters (experts) are activated for each input token, decoupling total parameter count from active compute cost.
Scaling Laws	Empirical power-law relationships describing how model performance improves with increases in compute, data, and parameters.
Self-Consistency	An inference technique where multiple reasoning paths are sampled and the most frequent answer is selected, improving performance on reasoning tasks.

How The Conclave Works

The Conclave is a structured multi-model deliberation system. Instead of asking one AI model a question and trusting its answer, the Conclave convenes a panel of independent models, forces them to argue, and subjects their conclusions to judicial review. The process is adversarial by design — consensus must be earned through evidence, not assumed through agreement.

Advocates are independent AI models (e.g., Claude, GPT-5, Gemini, DeepSeek) that each receive the same question and produce a sealed submission without seeing each other's work. Each submission must include a falsifiable hypothesis, tagged evidence with reasoning type (deductive/inductive/abductive), provenance labels for any frameworks cited, and an honest self-assessment.

Challenges follow submissions. Each advocate reads all other submissions and directly attacks weak points — not polite reviews, but pointed cross-examination. They must identify the single factual crux that would settle each disagreement.

Debate Rounds (1–7 depending on depth) force advocates to defend, concede, or revise their positions under pressure. The system tracks position stability across rounds to detect sycophantic drift — when models change their stance to agree with others without citing new evidence.

Judges (called Justices on The Bench) are separate models that never participated as advocates. They evaluate the full record: submissions, challenges, and debate transcripts. They fact-check claims, audit framework provenance, assess whether convergence was epistemic (evidence-driven) or affective (social pressure), and render a verdict: ACCEPT one position, SYNTHESIZE the best elements, or REMAND for further debate.

Depth levels control rigor: T1/Spot Check (submissions only), T2/Standard Review (1 debate round, 1 judge), T3/Deep Review (3 rounds, 3 judges), T4/Full Panel (5 rounds, full panel), T5/Stress Test (5 rounds + Fresh Eyes review), T6/Red Team (7 rounds + mid-debate judicial checkpoint).

The system is deterministic code — it cannot be sycophantic. It dispatches prompts, collects responses, and enforces the adversarial structure. The models argue; the code referees.