Tribunal Session Summary

Session: conclave-20260604-005912-is-10x-compute-required-7531

Depth: T6

Advocates: 6 | Judges: 5 | Cost: $5.8349 | Time: 36m 14s

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

THE RULING: No. "10x compute" is not a required, law-like relationship for doubling model intelligence. The relationship between compute and capability is governed by power laws with shallow exponents, meaning that large capability gains are routinely achieved through algorithmic improvements, data efficiency, and compute-optimal scaling—often with significantly less than a 10x increase in physical compute.

The Prompt

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

The Ruling

No. "10x compute" is not a required, law-like relationship for doubling model intelligence. The relationship between compute and capability is governed by power laws with shallow exponents, meaning that large capability gains are routinely achieved through algorithmic improvements, data efficiency, and compute-optimal scaling—often with significantly less than a 10x increase in physical compute.

The strongest evidence comes from the scaling laws themselves. Under the canonical Chinchilla scaling laws, a 10x increase in compute yields only a ~1.5x reduction in *reducible* loss, not a doubling of capability. Furthermore, empirical counterexamples like Chinchilla-70B outperforming Gopher-280B at roughly equal compute demonstrate that better allocation and data quality can substitute for raw scale. While the "10x rule" serves as a useful heuristic for budgeting perceived generational leaps, it fails as a strict requirement because it ignores the compounding impact of algorithmic efficiency (estimated at ~3x per year) and the non-linear mapping between loss reduction and practical intelligence.

The Tribunal accepts GPT-5's position as the most comprehensive and rigorously defended submission, synthesizing the mathematical reality of power laws with concrete empirical counterexamples.

Core Analysis

The Mathematical Reality: Power Laws, Not Doubling Steps

The core of the argument against a "10x requirement" is mathematical. Neural scaling laws establish that training loss (L) scales as a power law of compute (C), expressed as $L \propto C^{-\alpha}$.

- **The Exponents:** The exponent α is small. Kaplan et al. (2020) estimate $\alpha \approx 0.05$, while Hoffmann et al. (2022) estimate the exponent for *reducible* loss under compute-optimal conditions at $\alpha \approx 0.178$.
- **The Implication:** Because the exponent is small, returns to compute diminish rapidly. A 10x increase in compute reduces loss by only ~11–21% under Kaplan's formulation, or reduces reducible loss by a factor of ~1.51x under Chinchilla.
- **Why this refutes "10x per doubling":** If "intelligence" were strictly defined as the inverse of loss, achieving a literal doubling (halving the loss) would require thousands of times more compute, not 10x. Conversely, if "intelligence" is defined by downstream benchmarks, the relationship is even looser. The math simply does not support a fixed 10x multiplier for a 2x gain.

The Chinchilla Counterexample: Allocation Beats Scale

A single, robust counterexample falsifies the claim that 10x compute is *required* for major gains.

- **The Evidence:** Hoffmann et al. (2022) demonstrated that Chinchilla-70B (70 billion parameters) outperformed Gopher-280B (280 billion parameters) across a wide range of benchmarks, including MMLU (67.5% vs. 60.0%).
- **The Mechanism:** Both models were trained on roughly the same total compute budget (~5.76 × 10^23 FLOPs). Chinchilla achieved its superior performance by reallocating that compute toward more data tokens (1.4T vs. 300B) rather than more parameters.
- **The Takeaway:** This proves that "intelligence" gains can be unlocked by *how* you spend compute (allocation and data quality) rather than *how much* you spend. A 10x increase was not necessary to achieve a significant capability jump.

Algorithmic Efficiency: The Moving Target

The "10x rule" assumes a static relationship between hardware and capability, ignoring the rapid pace of software improvements.

- **The Trend:** Research by Epoch AI indicates that algorithmic efficiency—the compute required to reach a fixed performance level—improves by a factor of ~3x per year (or halves every 8–9 months).
- **The Impact:** Over a standard 2-year model development cycle, algorithmic progress alone provides a ~9x effective compute multiplier. This means a lab can achieve a capability doubling that *would* have required 10x physical compute in the past using only ~1.1x physical compute today, simply by waiting for better algorithms.
- **The Conclusion:** The physical compute requirement is not a fixed constant; it is decaying over time due to algorithmic gains.

The Operationalization Problem: Loss vs. Intelligence

A critical flaw in the "10x" heuristic is the ambiguity of "intelligence."

- **Loss vs. Capability:** Cross-entropy loss is a proxy for next-token prediction accuracy, not a direct measure of reasoning or problem-solving. A 10% reduction in loss might correspond to a massive jump in a specific benchmark (like coding or math) or no perceptible change at all.
- **Perception vs. Measurement:** The heuristic persists because of log-scale perception (Weber-Fechner law); humans perceive an order-of-magnitude increase in compute as a linear step in capability. However, this is a psychological effect, not a physical law.
- **Test-Time Compute:** Recent advances in inference-time scaling (e.g., OpenAI's o1 models) show that "intelligence" can be increased by spending compute at *inference* time rather than *training* time, further breaking the link between training compute and capability.

Summary of the Winning Position

GPT-5's position is accepted because it integrates the mathematical proof (power laws), the empirical disproof (Chinchilla), and the contextual nuance (algorithmic efficiency) to demonstrate that 10x compute is a contingent heuristic, not a requirement. The other "No" advocates provided valuable supporting evidence—such as the ~4x compute needed to double parameters under Chinchilla scaling—but GPT-5 provided the most comprehensive and falsifiable framework.

How We Got Here

Scorecard

Judge	GPT-5	Claude Opus 4.7	Magistral Medium	DeepSeek V4 Pro	Gemini 3.1 Pro	Grok 4	Winner
GPT-OSS 120B	ACCEPT	REJECT	REJECT	REJECT	REJECT	REJECT	GPT-5
DeepCogito v2.1 671B	ADOPT (partial)	ADOPT (partial)	REJECT	ADOPT (partial)	REJECT	REJECT	Synthesis
Qwen 3.5 397B	ACCEPT	REJECT	REJECT	REJECT	REJECT	REJECT	GPT-5
Zhipu GLM-5.1	ACCEPT	REJECT	REJECT	REJECT	REJECT	REJECT	GPT-5

Convergence Assessment

The convergence was **epistemic** (evidence-driven). The shift from the initial "Yes" stance by Grok 4 to a final "No" consensus was triggered by verifiable corrections: the mathematical derivation of power-law exponents, the verified results of the Chinchilla paper, and the failure to verify the Merali citation. The position stability was generally high among the "No" advocates, with refinements occurring only to narrow the scope of claims (e.g., conceding that RLHF is preference, not capability) rather than flipping the core thesis.

Closing Statements

GPT-5: Maintained that "10x compute" is not a requirement, arguing that capability growth depends on compute, data, and algorithms. They emphasized that while 10x steps sometimes coincide with capability leaps, this is contingent, not necessary. Their strongest evidence remained the Chinchilla counterexample and the shallow scaling exponents.

Claude Opus 4.7: Argued the "10x rule" is a folk heuristic, not a scaling law. They highlighted that the answer depends entirely on the operationalization of "intelligence" (loss vs. benchmarks vs. productivity). They provided the key quantitative synthesis that per-doubling physical compute is actually lower (~2–3.5x) when accounting for efficiency trends.

Magistral Medium: Initially argued that halving loss requires ~8000x compute, implying 10x is insufficient. They later conceded that "halving loss" is not the correct definition of "doubling intelligence." Their final position rested on the complexity of the relationship and the lack of a consistent 10x pattern.

DeepSeek V4 Pro: Focused on the Chinchilla parameter scaling law ($N \propto C^{0.5}$), arguing that doubling parameters requires only ~4x compute. They retracted speculative claims about GPT-4's size and compute, refining their argument to focus on the lack of a universal multiplier.

Gemini 3.1 Pro: Proposed that algorithmic progress and test-time compute drastically reduce physical needs. They retracted their "Compute Allocation Matrix" after challenges regarding variable independence. Their final position rested on the ~3x/year algorithmic efficiency trend (Epoch AI) as the primary disproof of the 10x rule.

Grok 4: Initially argued "Yes," claiming 10x compute is an empirical rule of thumb. They conceded every load-bearing claim (exponent math, GPT-4 compute estimates, loss-to-intelligence mapping) and ultimately flipped to "No," acknowledging the rule is a heuristic, not a requirement.

Key Moments

- **The Chinchilla Opening (GPT-5):** GPT-5 opened with the Chinchilla vs. Gopher counterexample, establishing immediately that equal compute could yield superior performance through better allocation. This set the factual baseline that "more compute" is not synonymous with "more intelligence."
- **The Power Law Breakdown (Claude Opus 4.7):** Claude Opus 4.7 provided the crucial mathematical analysis, distinguishing between the Kaplan (~0.05) and Chinchilla (~0.178) exponents. They demonstrated that 10x compute yields only fractional loss reductions, directly attacking the mathematical plausibility of the heuristic.
- **The Matrix Collapse (Gemini 3.1 Pro vs. Claude Opus 4.7):** Gemini 3.1 Pro proposed a "Compute Allocation Matrix" claiming 2.5x physical compute could yield >15x effective compute through multipliers. Claude Opus 4.7 challenged the independence of these variables. Gemini 3.1 Pro eventually conceded and retracted the matrix, a pivotal moment where a complex theoretical claim failed under scrutiny.
- **The "Yes" Position Collapse (Grok 4):** Grok 4 was the sole advocate arguing "Yes." Their position relied heavily on industry anecdotes (Musk quotes) and a misinterpretation of scaling exponents. Under cross-examination regarding the sign of the exponent and the lack of primary data, Grok 4 conceded their position and flipped to "No," leaving the panel with unanimous opposition to the "10x requirement."
- **The Citation Purge (Merali 2024):** Multiple advocates (Claude Opus 4.7, Magistral Medium, DeepSeek V4 Pro) initially cited a "Merali 2024" study claiming 12% productivity gains. When challenged, none could verify the source. The withdrawal of this citation across multiple rounds strengthened the integrity of the final consensus, removing weak evidence from the record.

Reality Checks

Claim	Challenge	Result
Gemini 3.1 Pro: "Compute Allocation Matrix" (2.5x physical × 2x data × 3x inference = >15x effective).	**Claude Opus 4.7:** The variables (data, inference, algorithms) are not independent; they often target the same loss and cannot be multiplied.	**Full Retraction.** Gemini 3.1 Pro conceded the matrix was mathematically suspect and abandoned the >15x figure.

Claim	Challenge	Result
Grok 4: "10x compute doubles intelligence" based on Musk/Morgan Stanley quotes.	**GPT-5/B/D:** These are appeals to authority, not empirical evidence. The scaling law exponents mathematically contradict the claim.	**Thesis Abandoned.** Grok 4 conceded the "Yes" position and flipped to "No," acknowledging the lack of primary data.
Claude Opus 4.7/C/D: "Merali 2024" study shows 12% productivity gain per 10x compute.	**GPT-5:** The citation is unverifiable; the arXiv ID does not exist.	**Claim Withdrawn.** Claude Opus 4.7 and Magistral Medium, and D all conceded they could not verify the source and removed it from their evidence.
Grok 4: GPT-4 required ~10x compute over GPT-3.	**DeepSeek V4 Pro:** Public estimates place the increase at ~60x, not 10x.	**Correction.** Grok 4 conceded the 10x figure was incorrect and retracted the specific claim.

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) determining the compute-optimal ratio of model parameters to training data tokens for large language models.
Compute-Optimal Training	Training a model in a way that minimizes loss for a given fixed compute budget, typically by using smaller models and more data.
Cross-Entropy Loss	A metric measuring the difference between two probability distributions (the model's predictions and the true labels), commonly used to train language models.
Effective Compute	A measure of capability that accounts for algorithmic improvements and efficiency gains, distinct from raw physical FLOPs.
Emergent Capabilities	Abilities in large language models (e.g., arithmetic, chain-of-thought) that are not present in smaller models and appear unpredictably at scale.
Inference-Time Compute	Computational power expended during the model's use (generating a response) rather than during its training.
Power Law	A mathematical relationship where one quantity varies as a power of another (e.g., $L \propto C^{-\alpha}$), often used to describe scaling in AI.
Process Reward Models (PRMs)	Models trained to evaluate intermediate steps in a reasoning process, used to guide search algorithms during inference.
Scaling Laws	Empirical rules describing how model performance improves with increases in compute, data, and parameters.
Weber-Fechner Law	A psychological principle stating that the perceived change in a stimulus is proportional to the logarithm of the stimulus intensity, explaining why OOM compute gains feel linear.

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.