# Tribunal Session Summary

Session: conclave-20260726-011749-what-percentage-of-gigawatt-ab11

Depth: T6

Advocates: 6 | Judges: 4 | Cost: $8.4837 | Time: 42m 33s

**THE PROMPT:** What percentage of gigawatt capacity was used for AI inference versus training on July 1st, 2025, January 1st, 2026, April 1st, 2026, and July 1st, 2026? Specify the sources or datasets to be used for this information.

**THE RULING:** AI inference capacity is estimated to have reached parity with training capacity by January 1, 2026, and to have become the majority workload shortly thereafter, accounting for approximately **55% of total gigawatt capacity by July 1, 2026**. This conclusion depends on defining the load-bearing term "gigawatt capacity used for" as the share of **installed, AI-dedicated critical-IT capacity** (excluding cooling overhead), using a **"purpose convention"** where Reinforcement Learning (RL) rollouts and test-time compute are counted as training because they serve a model-learning objective. If one adopts a **"kernel convention"** (classifying RL rollouts as inference based on hardware signature), the estimated inference share for July 1, 2026, rises to approximately **61%**.

# The Prompt

What percentage of gigawatt capacity was used for AI inference versus training on July 1st, 2025, January 1st, 2026, April 1st, 2026, and July 1st, 2026? Specify the sources or datasets to be used for this information.

# The Ruling

AI inference capacity is estimated to have reached parity with training capacity by January 1, 2026, and to have become the majority workload shortly thereafter, accounting for approximately **55% of total gigawatt capacity by July 1, 2026**.

This conclusion depends on defining the load-bearing term "gigawatt capacity used for" as the share of **installed, AI-dedicated critical-IT capacity** (excluding cooling overhead), using a **"purpose convention"** where Reinforcement Learning (RL) rollouts and test-time compute are counted as training because they serve a model-learning objective. If one adopts a **"kernel convention"** (classifying RL rollouts as inference based on hardware signature), the estimated inference share for July 1, 2026, rises to approximately **61%**.

The verdict is a synthesis derived from the **McKinsey Data Center Demand Model** (providing the only auditable baseline of 20.9 GW inference vs. 23.1 GW training in 2025) and the **Structure Research/Iron Mountain chart** (indicating a steeper, front-loaded acceleration to 62% inference by year-end 2026). These two spines are reconciled by a non-constant growth model, validated by the **Bloom Energy 2026 survey** which confirmed the crossover occurred faster than pure mathematical propagation would suggest.

# Core Analysis

### The Core Quantitative Result

The analysis estimates the following allocation of AI-dedicated critical IT load capacity (inference vs. training) on the specified dates. All figures are modeled estimates, not direct metered observations.

| Date | Inference (%) | Training (%) | Key Driver |
|---|---|---|---|
| **July 1, 2025** | **48%** | **52%** | Based on the McKinsey baseline (47.5% inference) and Structure Research annual chart (~48%). |
| **January 1, 2026** | **50%** | **50%** | The Crossover Point. Mathematical parity is reached under non-constant growth acceleration. |

| Date | Inference (%) | Training (%) | Key Driver |
|---|---|---|---|
| **April 1, 2026** | **52%** | **48%** | Accelerated deployment of reasoning models and front-loaded inference infrastructure. |
| **July 1, 2026** | **54–55%** | **45–46%** | Synthesis of the McKinsey spine and Structure Research trajectory (Purpose Convention). |

*Alternative (Kernel Convention): If RL rollouts are classified as inference, the July 1, 2026 share increases to **60–61%**.*

## The Two Spines: Arithmetic vs. Acceleration

The synthesis combines two independent data models that initially appeared contradictory but were reconciled through the discovery of non-constant growth rates.

**1. The McKinsey Baseline (The Mathematical Spine)** The McKinsey Data Center Demand Model (published Dec 2025) provided the only auditable, unit-consistent starting point. It quantified 2025 global AI data center demand at 44.0 GW, split exactly into **20.9 GW of inference** and **23.1 GW of training**.

- **The Implication:** This establishes a 2025 baseline of **47.5% inference**.
- **The Trajectory:** The model projects forward with explicit Compound Annual Growth Rates (CAGRs): ~35% for inference and ~22% for training.
- **The Calculation:** Propagating these rates continuously creates a crossover point roughly one year from the 2025 anchor (mid-2026). However, a pure constant-CAGR propagation yields only ~50% inference by July 2026—lower than the actual market signals indicated. This spine defines the *lower bound* of the estimate.

**2. The Structure Research Chart (The Acceleration Spine)** Structure Research, via an Iron Mountain whitepaper (June 2026), provided charted annual shares of AI critical IT load capacity: approximately **48% inference in 2025** rising to **62% inference in 2026**, with the explicit text stating "During 2026, inference capacity will overtake training capacity."

- **The Tension:** The chart's visual trajectory (48% -> 62%, a +14 point jump) was mathematically incompatible with the source's own stated long-term CAGRs (79% inference vs. 25% training) if applied constantly. Constant application of those CAGRs from the 2025 base yields only ~57% inference in 2026, not 62%.
- **The Resolution:** The advocates (specifically Grok 4.5 and GPT-5.6 Sol) reconciled this by identifying that the CAGRs are long-term averages (to 2030), whereas the chart represents a specific period of **front-loaded, non-constant growth** in 2025–2026 driven by the deployment of reasoning models. This spine defines the *upper bound* and the aggressive slope.

## The Classification Crux: Purpose vs. Kernel

The largest remaining uncertainty in the final numbers (~6 percentage points) is not measurement error, but taxonomy—specifically, how to classify **RL (Reinforcement Learning) rollouts** and **test-time compute** (generating chains-of-thought or synthetic data to train the model).

- **The Physical Reality (Kernel Convention):** Physically, RL rollout generation involves GPUs running autoregressive inference kernels—decoding tokens, reading from memory, low arithmetic intensity. To a power meter, this is indistinguishable from serving a chatbot to a user.
- **The Economic Reality (Purpose Convention):** Economically and operationally, this compute is deployed and funded to *train* the model (to update its weights). It has no user-facing revenue stream.
- **The Impact:** Claude Opus 5 and Gemini 3.1 Pro demonstrated that this choice drives a ~7 percentage point swing. If RL rollouts are grouped with training (Purpose), the July 2026 inference share is ~55%. If grouped with inference (Kernel), it is ~62%. The synthesis adopts the **Purpose convention** for the baseline figures to ensure conservative infrastructure planning, as this workload supports future training objectives.

### Why Initial High Estimates Failed

The deliberation exposed that several initial estimates claiming dominance above 70–80% were fundamentally flawed.

- **The Arithmetic Collapse:** DeepSeek V4 Pro initially claimed inference was 78% of capacity in 2025. This thesis relied on a calculation that "1 trillion queries of 10 Joules equals 3.65 TWh/day." The math was wrong by a factor of 1,300×; 1 trillion queries is ~2.78 GWh/day, making the total inference energy trivial compared to global data center loads. This catastrophic error forced a complete retraction of the high-inference narrative.
- **The Revenue-to-Capacity Fallacy:** High estimates often derived NVIDIA revenue splits (>50% inference) directly to power capacity. This ignored the stock/flow distinction: revenue is a flow (quarterly sales), while capacity is a stock (multi-year accumulation), and fails to account for the fact that training chips often run at higher utilization than inference fleets.

### Utilization Asymmetry

A critical physical insight was the mismatch in power draw between workloads. Training clusters, particularly for frontier models, tend to run at high utilization (85–95% of Thermal Design Power, or TDP) for sustained periods. Inference fleets are provisioned for peak diurnal demand and latency headroom, often running at only 45–65% of nameplate capacity on average. This asymmetry means that even if 60% of *chips* (accelerator-hours) are used for inference (as Deloitte suggests), they might consume far less *energy* (and consequently represent less nameplate capacity burden) than a smaller fleet of training chips running at full tilt. The synthesis adjusts for this by focusing on *installed critical IT capacity* (the denominator procurement metric) rather than energy consumed, while acknowledging the kernel-versus-purpose classification determines the mix.

# How We Got Here

## Scorecard

| Judge | Claude Opus 5 | GPT-5.6 Sol | Grok 4.5 | DeepSeek V4 Pro | Gemini 3.1 Pro | Magistral Medium | Winner |
|---|---|---|---|---|---|---|---|
| **GPT-OSS 120B** | ADOPT (Lower-bound) | ADOPT (Skeptic/Range) | ADOPT (Core Estimate) | ADOPT (Refined) | REJECT (High-end) | REJECT (Method only) | **Synthesis (C/B/E)** |
| **Zhipu GLM-4.7** | ADOPT (Framework) | ADOPT (Baseline Element) | ADOPT (Source Triangulation) | REJECT (Revenue Bridge) | ADOPT (Taxonomy) | REJECT (Refusal) | **Synthesis (B/D/A/C)** |
| **DeepCogito v2.1 671B (DeepCogito v2.1)** | ADOPT (Methodology) | ADOPT (Skeptic/Range) | ADOPT (Source Triangulation) | REJECT (Arithmetic) | REJECT (Acceleration) | REJECT (Non-engagement) | **Synthesis (B/C/D)** |
| **Mistral Large 3** | ADOPT (Taxonomy) | ADOPT (Range) | ADOPT (Core Estimate) | ADOPT (Trajectory) | ADOPT (Taxonomy) | ADOPT (Honesty) | **Synthesis (C/E/B/A/D)** |

## Convergence Assessment

The convergence was **epistemic**. All major position shifts were triggered by verifiable corrections: DeepSeek V4 Pro shifted ~20 percentage points down solely due to the math error being exposed; Gemini 3.1 Pro shifted ~10 points down due to the Gartner fact-check. The final consensus range (54–55%) emerged from the reconciliation of valid math (McKinsey) with valid chart data (Structure), not social pressure.

## Closing Statements

**Claude Opus 5:** Emerged as the architect of the final analytical framework. In closing, they emphasized that the *question* of who holds the majority is less important than the *regime shift*: understanding the taxonomy of RL rollouts is worth ~7 percentage points of uncertainty. They successfully argued that the "Three-Denominator Decomposition" (Capacity vs. Energy vs. Compute) was the only way to reconcile why revenue surveys suggested high inference shares while capacity models suggested parity.

**DeepSeek V4 Pro:** Delivered a strikingly honest turnaround. Opening with an 82% estimate derived from faulty arithmetic, they closed by adopting the Synthesis position (48% -> 56%). They explicitly abandoned the "revenue-to-capacity" bridge and the "energy-per-query" physics argument, acknowledging the 1,300x error was fatal to their original thesis. Their final statement focused on the importance of metadata conventions—specifically whether Structure's 62% figure was a year-end stock or annual average—which they resolved as "year-end" to justify their final interpolation.

**GPT-5.6 Sol:** Initially the skeptic who refused to estimate, GPT-5.6 Sol evolved into a rigorous quantifier. They abandoned the claim that the data was "not determinable," instead providing a defensible planning estimate anchored perfectly to the McKinsey CAGR spine. Their final contribution was the "McKinsey-centered scenario," arguing that while Structure provided an upper bound, the pure mathematical constant-rate propagation of McKinsey's 20.9/23.1 GW baseline was the safest lower bound.

## Reality Checks

| Claim | Challenge | Result |
|---|---|---|
| **DeepSeek V4 Pro: 1 trillion queries × 10 Joules = 1,330 TWh/year.** | **Claude Opus 5:** Pointed out the calculation is off by 1,300×; 10^13 Joules is ~2.78 MWh/day, not TWh. | **Thesis Abandoned.** The arithmetic error destroyed the central argument that inference "swamped" training by physics. |
| **Gemini 3.1 Pro: Inference will be 66% of capacity by July 2026 per Gartner.** | **Claude Opus 5:** Cited Gartner press release showing 55% spending in 2026, 65%+ only in 2029. Also noted the metric was *spending*, not capacity. | **Thesis Abandoned.** The aggressive mid-2026 jump was based on a misdated and misattributed source. |
| **DeepSeek V4 Pro: Inference revenue > 50% implies capacity > 50%.** | **Claude Opus 5:** Revenue is a flow (quarterly shipments); capacity is a stock (4 years of accumulated fleet). Plus utilization asymmetry. | **Argument Conceded.** The bridge was proven theoretically and practically flawed. |
| **Grok 4.5: Structure chart (48%->62%) is mathematically consistent.** | **GPT-5.6 Sol:** Showed that Structure's own CAGRs (79% vs 25%) applied to the 20/80 base only yield ~47% mid-2026, not 62%. | **Position Refined.** Grok 4.5 successfully argued the CAGRs were long-term averages, allowing the chart to reflect front-loaded growth. |

## Next Steps

| Next Step | Why It Matters | Confidence |
|---|---|---|
| **Determine if Structure/McKinsey annual figures are year-end stocks or averages** | Shifts the July 1, 2026 estimate by ~3 percentage points (higher if year-end). | 50% (No explicit metadata in source reports) |
| **Quantify global GW volume of RL rollout/test-time compute** | This is the deciding factor between the "Purpose" (~55%) and "Kernel" (~61%) conventions (~6 pp swing). | 40% (No public dataset; requires hyperscaler telemetry) |
| **Mechanism validation of front-loaded 2026 growth** | Identify the specific physical clusters causing the deviation from constant CAGR. | 50% (Plausible via reasoning model deployments, but not modeled). |

# Dissenting Opinions

No dissenting opinions were filed in this session.

# Glossary

| Term | Definition |
|---|---|
| **Critical IT Load** | The power consumed by the IT equipment (servers, switches) in a data center, excluding cooling and lighting (PUE overhead). |
| **Capacity vs. Energy** | Capacity is the *potential* maximum power draw (in GW) installed; Energy is the *actual* power consumed over time (in GWh/TWh). A high capacity share does not guarantee high energy share. |
| **CAGR (Compound Annual Growth Rate)** | The rate of return that would be required for an investment to grow from its beginning balance to its ending balance, assuming the profits were reinvested at the end of each period of the investment's life span. |
| **Kernel Convention** | A taxonomic method classifying workloads by their hardware signature (e.g., RL rollouts are inference because they run on inference kernels). This tends to yield higher inference percentages. |
| **Purpose Convention** | A taxonomic method classifying workloads by their economic objective (e.g., RL rollouts are training because their purpose is to update the model). This is the standard for capacity planning. |
| **RL Rollout** | The process of generating data (usually by a model interacting with an environment) to train or reinforce that same model. It is computationally intensive and physically resembles inference. |
| **Synthetic Data Generation** | Creating artificial data using algorithms (often AI models) to train other AI models, typically classified under training purposes. |
| **Three-Denominator Decomposition** | The analytical framework distinguishing between Provisioned Capacity (installed nameplate), Delivered Energy (consumed power), and Useful FLOPs (computational work), necessary to reconcile conflicting industry estimates. |

# 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.