Tribunal Session Summary

Session: conclave-20260604-012007-provide-a-best-in-ff0e

Depth: T6

Advocates: 6 | Judges: 5 | Cost: $6.5707 | Time: 59m 42s

THE PROMPT: Provide a best in class estimate of the EBITDA and EBITDA margins for Hebbia, decomposed by top line revenue, operating expenses with technology expense, headcount, depreciation, and other expenses broken out.

THE RULING: Hebbia is operating at a negative EBITDA margin of approximately **-25%** on **$47 million** in revenue (Mid-2026 TTM), resulting in an EBITDA loss of roughly **$11.7 million**. This conclusion accepts Claude Opus 4.7's position as the most rigorous and evidence-based. The company, while profitable at a smaller scale ($13M ARR), has deliberately traded profitability for aggressive growth post-Series B, hiring elite talent (driving headcount costs to ~$36M) and incurring high inference costs (~25% of revenue). The consensus across the tribunal is that EBITDA is negative, ranging from -10% to -30%, driven by the structural costs of AI-native operations.

The Prompt

Provide a best in class estimate of the EBITDA and EBITDA margins for Hebbia, decomposed by top line revenue, operating expenses with technology expense, headcount, depreciation, and other expenses broken out.

The Ruling

Hebbia is operating at a negative EBITDA margin of approximately **-25%** on **$47 million** in revenue (Mid-2026 TTM), resulting in an EBITDA loss of roughly **$11.7 million**. This conclusion accepts Claude Opus 4.7's position as the most rigorous and evidence-based. The company, while profitable at a smaller scale ($13M ARR), has deliberately traded profitability for aggressive growth post-Series B, hiring elite talent (driving headcount costs to ~$36M) and incurring high inference costs (~25% of revenue). The consensus across the tribunal is that EBITDA is negative, ranging from -10% to -30%, driven by the structural costs of AI-native operations.

The winning argument, presented by Claude Opus 4.7, successfully triangulated revenue using headcount and Revenue Per Employee (RPE) benchmarks, acknowledged the critical distinction between FY2025 and Mid-2026 TTM, and applied a realistic, function-weighted Fully-Loaded Cost (FLC) for NYC/SF talent. It correctly identified that the "profitability" achieved at $13M ARR was a small-scale artifact that cannot survive the step-function increase in costs associated with tripling headcount and scaling compute infrastructure.

Core Analysis

Thesis

Hebbia is currently in a deliberate investment phase, burning cash to capture market share in the AI-native financial and legal workflow space. The "profitability" achieved at $13M ARR (mid-2024) was a function of a small headcount (~55 FTEs) and does not scale linearly. Post-Series B, the company has aggressively hired elite talent and expanded infrastructure, causing operating expenses—specifically headcount and inference costs—to outpace revenue recognition. This results in a negative EBITDA margin of approximately -25% on ~$47M in revenue for the Mid-2026 TTM period.

Revenue Estimate: ~$47 Million (Mid-2026 TTM)

The revenue estimate is derived from a **headcount triangulation** method rather than speculative valuation multiples.

- **Mechanism:** Average headcount of ~135 Full-Time Equivalents (FTEs) multiplied by a Revenue Per Employee (RPE) benchmark of ~$350,000.

- **Evidence:** This aligns with the verified $13M ARR baseline from mid-2024 and accounts for the aggressive hiring ramp post-Series B (verified headcount growth from ~95 FTEs in Nov 2024 to ~149 FTEs in early 2026).
- **Why it matters:** This approach is superior to valuation-based caps (which assume specific revenue hurdles to justify entry pricing) or generic SaaS benchmarks, as it ties revenue directly to the company's observable operational capacity.

Operating Expenses Decomposition

1. Headcount Expense: ~$35.6 Million (75.7% of Revenue)

- **Mechanism:** Calculated as 135 average FTEs × ~$264,000 Fully-Loaded Cost (FLC).
- **Evidence:** The FLC figure is a standard corporate finance metric applied to the specific NYC/SF AI talent market (verified by industry benchmarks like Levels.fyi). This figure corrects the arithmetic impossibilities found in other models (e.g., Magistral Medium's implausibly low $172k/FTE and Advocate-A's $75k/FTE error).
- **Why it matters:** Headcount is the single largest expense category and the primary driver of the negative EBITDA outcome.

2. Technology/Inference Expense: ~$11.8 Million (25.0% of Revenue)

- **Mechanism:** Captures the high cost of inference (API costs to OpenAI/Anthropic) and cloud hosting.
- **Evidence:** While exact token volumes are unknown, this percentage is supported by the founder's statement regarding high volume (2% of OpenAI daily volume at $13M ARR) and the nature of Hebbia's ISD (Iterative Source Decomposition) architecture, which fans out parallel LLM calls per query. This is significantly higher than traditional SaaS COGS (10-15%).
- **Why it matters:** This "structural COGS floor" limits gross margin expansion, distinguishing AI-native companies from legacy software.

3. Other Operating Expenses: ~$11.3 Million (24.0% of Revenue)

- **Mechanism:** Includes office space (NYC/SF), legal, insurance, and sales/marketing not directly tied to headcount or compute.
- **Evidence:** Derived as a residual in the model, this figure is consistent with a company scaling to ~150 employees in high-cost metro areas.

EBITDA and Margin Calculation

- **EBITDA:** Revenue ($47.0M) - Tech COGS ($11.8M) - Headcount ($35.6M) - Other OpEx ($11.3M) = **-$11.7 Million**.
- **EBITDA Margin: -24.9%**.
- **Depreciation:** ~$1.0M (Asset-light software model; minimal capitalized hardware or leaseholds).

Key Drivers of Performance

- **The "Profitability Trap":** The transition from $13M ARR (profitable) to $47M revenue (negative margin) is driven by the non-linear scaling of costs. Elite AI talent commands high fixed costs, and inference costs scale with usage, not just seats.

- **Growth vs. Efficiency:** The negative margin is a strategic result of the $130M Series B raise, which was explicitly deployed to fund growth (hiring, acquisition of FlashDocs, infrastructure) rather than preserve margins.

How We Got Here

Scorecard

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

Convergence Assessment

Convergence was **epistemic** (evidence-driven). The shift toward negative margins was triggered by verifiable corrections: the exposure of fabricated data points (Gemini 3.1 Pro), the demonstration of arithmetic impossibilities regarding compensation (A/B), and the correction of valuation multiples (DeepSeek V4 Pro). The consensus that headcount costs (~$35M) exceed gross profit at current revenue levels ($40-50M) is a mathematical necessity derived from public headcount data and market compensation benchmarks, not social pressure.

Closing Statements

Claude Opus 4.7: Maintained the negative margin thesis (-25%) as the most defensible position. Highlighted the use of headcount triangulation and Revenue Per Employee (RPE) to anchor revenue, arguing that valuation-based revenue caps were flawed. Emphasized that the company is executing a deliberate post-Series B playbook: growth ahead of profitability, with sufficient runway to sustain -25% margins for years.

GPT-5: Converged to a -30% margin estimate. Refined the model to use function-weighted Fully-Loaded Costs (FLC), distinguishing between R&D, S&M, and G&A roles. Dropped generic SaaS comps (C3.ai,

Palantir) in favor of mechanism-first COGS drivers (agentic fan-out, token price cuts).

Gemini 3.1 Pro: Underwent a dramatic revision, swinging from a -47% margin to -11.5%. Abandoned fabricated data points (GPT-5, Opus 4.5) and adopted the headcount triangulation logic proposed by Claude Opus 4.7. Argued that revenue *must* be $55M to justify the valuation and hiring velocity, though this was challenged as a category error by others.

Grok 4: Conceded the initial thesis of sustained profitability (8-12% margin). Acknowledged that the arithmetic of headcount costs ($33M+ payroll) makes positive EBITDA impossible at $40-50M revenue. Revised the outlook to -5% to -12% margin, citing the "investment phase" narrative.

Magistral Medium: Revised from a positive 20% margin to a negative 18% margin. Conceded that generic SaaS benchmarks were inappropriate for an AI-native company and that initial headcount cost assumptions were too low.

DeepSeek V4 Pro: Revised the revenue anchor from $20M to $40M after conceding the valuation multiple error. Maintained a negative margin thesis (-31%), emphasizing the "investment-phase narrative" and the heavy burden of AI talent and inference costs.

Key Moments

- **Fabrication Exposed (Gemini 3.1 Pro):** In Round 3, Gemini 3.1 Pro claimed Hebbia used "GPT-5" and "Opus 4.5" and processed "1 million queries per day." Claude Opus 4.7 challenged these specific model names as non-existent and the query volume as unverified. Gemini 3.1 Pro was forced to concede that these were fabricated and retracted the entire claim, revising their model significantly.
- **Arithmetic Impossibility (Grok 4):** Grok 4 initially claimed profitability with $7.5M in headcount expenses for 100 employees, implying a Fully-Loaded Cost of $75,000 per person in NYC. Gemini 3.1 Pro exposed this as mathematically impossible given market rates. Grok 4 conceded the error and revised their position from positive margins to negative.
- **Stale Data Correction (DeepSeek V4 Pro):** DeepSeek V4 Pro anchored revenue to $20M using a flawed valuation multiple calculation (35x vs. the correct 54x). Advocate-D corrected the math, forcing Advocate-E to abandon the $20M figure and revise revenue upward to $40M+, aligning with the growth trajectory implied by the Series B raise.
- **Convergence on Negative Margins:** By Round 7, the debate converged on a consensus that Hebbia is burning cash. Claude Opus 4.7 and GPT-5, and E arrived at nearly identical EBITDA figures (~-$12M) on ~$40M revenue. Claude Opus 4.7 refined the timeframe to Mid-2026 TTM, raising the revenue estimate to $47M to reflect the current state of the company.

Reality Checks

Claim	Challenge	Result
Gemini 3.1 Pro: Hebbia uses "GPT-5" and "Opus 4.5" and processes "1M queries/day."	**Claude Opus 4.7:** These models do not exist in production and the query volume is unverified.	**Full Concession.** Gemini 3.1 Pro admitted the data was fabricated and retracted the claim.

Claim	Challenge	Result
Grok 4: Headcount expense is $7.5M for 100 employees (implying $75k/FTE).	**Gemini 3.1 Pro:** This is mathematically impossible in NYC given your own $220-280k benchmark.	**Full Concession.** Grok 4 admitted the arithmetic error and revised the cost model.
DeepSeek V4 Pro: Revenue is $20M based on a 35x valuation multiple.	**Claude Opus 4.7:** The correct multiple is 54x ($700M / $13M ARR). Your math is wrong.	**Full Concession.** DeepSeek V4 Pro corrected the multiple and revised revenue to $40M+.
Magistral Medium: EBITDA is +$10M (20% margin) on $50M revenue.	**Gemini 3.1 Pro:** Your headcount cost ($25M) implies $172k/FTE for 145 employees, which is below market base salary.	**Revision.** Magistral Medium revised to negative EBITDA (-18%) to align with realistic headcount costs.

Next Steps

Next Step	Why It Matters	Confidence
Verify FY2024 Year-End ARR	The $13M mid-2024 figure is verified, but the YE-2024 figure (critical for 2025 projections) is unverified and disputed (Gitnux claim).	Medium (DeepCogito v2.1 671B)
Audit Inference Unit Economics	Technology COGS (25% of revenue) is a major swing factor. Actual token volumes and negotiated Anthropic/OpenAI rates are unknown.	Low (Zhipu GLM-5.1)
Confirm Headcount Timing	Average FTE vs. snapshot FTE significantly impacts expense calculations. Exact ramp dates are approximated.	Medium (DeepCogito v2.1 671B)

Dissenting Opinions

No dissenting opinions were filed in this session.

Glossary

Term	Definition
ARR (Annual Recurring Revenue)	A metric used by subscription-based businesses to measure the revenue they expect to receive annually from their customers.
EBITDA (Earnings Before Interest, Taxes, Depreciation, and Amortization)	A measure of a company's operating performance, calculated as revenue minus operating expenses (excluding interest, taxes, depreciation, and amortization).
EBITDA Margin	EBITDA expressed as a percentage of total revenue.
FLC (Fully-Loaded Cost)	The total cost of an employee to the company, including base salary, bonuses, benefits, payroll taxes, and overhead (e.g., equipment, office space).
RPE (Revenue Per Employee)	A productivity metric calculated by dividing total revenue by the number of employees.
ISD (Iterative Source Decomposition)	Hebbia's proprietary architecture that decomposes queries into parallel sub-tasks across multiple model calls, increasing inference intensity.
COGS (Cost of Goods Sold)	The direct costs of producing the goods or services sold by a company. In Hebbia's case, this primarily refers to inference and cloud computing costs.
OpEx (Operating Expenses)	The ongoing costs for running a company's core business operations, excluding COGS (e.g., R&D, S&M, G&A).
Rule of 40	A benchmarking rule of thumb for SaaS companies stating that the sum of their revenue growth rate and profit margin should exceed 40%.

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.