6 Models, 1 Answer: Exploring Ensemble Diversity and Information Independence in AI-Driven Sports Prediction

Jack Elston, Ph.D.
CEO, Black Swift Technologies
Boulder, CO
elstonj@blackswifttech.com
ORCID: 0000-0001-9159-9028

March 2026

Abstract

We deploy six frontier large language models—Claude Sonnet 4 (Anthropic), GPT-4o (OpenAI), Gemini 2.5 Flash (Google), Grok 3 Mini (xAI), Llama 3.3 70B (Meta), and DeepSeek Chat (DeepSeek AI)—to independently predict all 63 games of the 2026 NCAA Men's Basketball Tournament. Despite originating from six different companies, training pipelines, and architectural families, the models agree on 80.6% of game outcomes; adding models four through six changes zero bracket picks. A cross-model adversarial debate protocol flips 5 of 8 contested picks while lowering mean confidence by 8.4 points. Monte Carlo sensitivity analysis over 250 weight permutations reveals that lineup certainty—not model strength—is the dominant accuracy driver (25.0 percentage-point impact). A Gradient-Boosting classifier trained on ten years of historical data (630 games, 2015–2025) achieves 92.1% leave-one-year-out accuracy (90.5% on the "chaos" 2023 holdout). We argue that *model diversity does not imply information diversity*: when all models share the same training corpus (the public internet), ensemble value lies in confidence calibration, not prediction improvement, with profound implications for organizations evaluating multi-vendor AI strategies. The total cost across all six platforms was approximately $12–15 for 850+ API calls, with DeepSeek producing nearly identical predictions to GPT-4o at a 27× cost reduction ($0.01 vs. $0.27).

1 Introduction

The NCAA Division I Men's Basketball Tournament—colloquially *March Madness*—is among the most widely predicted sporting events in the world. Each year, tens of millions of brackets are submitted to public pools, yet no verified perfect bracket has ever survived beyond 49 games [NCAA, 2019]. The tournament's 63-game, single-elimination format provides a rare combination of properties that make it an ideal testbed for prediction systems: clearly defined ground truth, massive public engagement, sufficient complexity to expose model limitations, and enough structure to remain analytically tractable [Lopez and Matthews, 2015, Boulier and Stekler, 1999].

The rise of frontier large language models (LLMs) introduces a natural question: *Does querying multiple independently trained AI models improve prediction accuracy, or do they converge to the same answer?* Classical ensemble theory [Breiman, 1996, Dietterich, 2000] guarantees improvement only when component learners exhibit *diversity*—that is, when their errors are uncorrelated. The Condorcet Jury Theorem formalizes this: majority voting improves accuracy only when voters are independent [de Condorcet, 1785, Surowiecki, 2004].

We test this premise empirically by deploying 6 frontier LLMs from six companies across three continents (Anthropic, OpenAI, Google DeepMind, xAI, Meta, and DeepSeek AI) to predict the full 2026 bracket. Our central finding is stark: **80.6% of predictions are unanimous across all six models**, and the marginal contribution of each additional model beyond the first rapidly approaches zero. We term this phenomenon *information convergence* and trace it to a shared training substrate—the English-language internet—rather than to model correctness.

The practical implication is significant for any organization evaluating multi-vendor AI strategies: paying for multiple AI platforms may not improve outcomes when the underlying information is the same. Value comes not from additional models but from additional *data*—injury reports, practice observations, and other non-public signals that no current LLM possesses.

We complement the LLM ensemble with a GradientBoosting classifier trained on 630 historical tournament games, achieving 92.1% leave-one-year-out cross-validation accuracy. This ML calibration layer contextualizes the ensemble's predictions against a decade of empirical outcomes and quantifies the upset detection rate at 71.0% (66 of 93 historical upsets correctly identified).

Isaac Asimov's *Foundation* [Asimov, 1951] imagined psychohistory: a science capable of predicting aggregate societal behavior while individual events remain stochastic. March Madness offers a microcosm of this tension—teams' aggregate tendencies are predictable; individual game outcomes retain irreducible variance. Even at 92% per-game accuracy, a perfect 63-game bracket remains a 1-in-763 event.

Our contributions are threefold: (1) the first systematic empirical study of multi-model LLM convergence on a well-defined prediction task; (2) an adversarial cross-model debate protocol

with measured calibration effects; and (3) the thesis that *data uniqueness trumps model diversity*—a finding with implications extending well beyond sports analytics into enterprise AI procurement.

2 Related Work

Tournament Prediction. Statistical prediction of NCAA tournament outcomes has a rich history. Carlin [1996] proposed Bayesian models using point spreads; Boulier and Stekler [1999] evaluated logistic regression on seeding data; Kvam and Sokol [2006] combined logistic regression with Markov chains. More recently, tempo-free adjusted efficiency metrics, pioneered by Pomeroy [2004] and extended by Torvik [2026], have become the gold standard. Nate Silver's COOPER model [Silver, 2024] blends Elo ratings with KenPom efficiencies and runs 100,000 Monte Carlo simulations per tournament, setting the public benchmark for several years [Silver, 2012]. Commercial platforms such as Rithmm [Rithmm, 2024] report 85% first-round accuracy using full-season team-level data. Kaggle's March Machine Learning Mania competition has attracted thousands of data scientists annually since 2014, yielding gradient-boosted and neural-network entries [Jacobson and King, 2011]. Our work differs by applying multi-model LLM ensembles to this domain rather than traditional statistical or supervised learning approaches alone.

Ensemble Methods. Breiman [1996] established that bagging reduces variance when base learners are unstable. Freund and Schapire [1997] and Friedman [2001] extended this to boosting. Dietterich [2000] formalized three reasons ensembles succeed: statistical, computational, and representational. Critically, all classical ensemble guarantees assume *diversity*—that component errors are at least partially uncorrelated [Hansen and Salamon, 1990]. Wolpert [1992] introduced stacked generalization as a meta-learning framework for combining heterogeneous learners. We show that the independence assumption breaks down when LLMs share training data, limiting the applicability of classical ensemble guarantees to LLM-based prediction.

Multi-Agent LLM Debate. Du et al. [2023] demonstrated that multi-agent debate among LLMs improves factuality and reasoning on benchmarks. Liang et al. [2023] explored divergent thinking through structured debate. Wang et al. [2024] proposed Mixture-of-Agents architectures in which LLMs iteratively refine each other's outputs. Our work applies cross-model debate specifically to sports prediction with empirical calibration outcomes, extending the debate paradigm from factual question-answering to probabilistic forecasting under uncertainty.

3 Methodology

3.1 System Architecture

We deploy a five-agent pipeline orchestrated by Claude Sonnet 4 (Anthropic), structured as a sequential information-processing chain:

1. **Edge Finder** — identifies non-consensus angles (injury reports, coaching matchups, tempo mismatches) by searching for signals that diverge from seed-based priors.
2. **Source Scout** — aggregates and scores local beat writer takes, regional podcasts, and social sentiment from three hierarchical tiers: Tier 1 statistical models (KenPom, Bart-Torvik, NET, ESPN BPI), Tier 2 bracket consensus (ESPN public picks, FiveThirtyEight), and Tier 3 live intelligence (beat writers, team social media, local reporting).
3. **Data Curator** — normalizes heterogeneous data into a structured JSON schema per game, producing 22 engineered features including seed differentials, KenPom efficiency gaps, conference strength indicators, and historical upset base rates.
4. **Bracket Modeler** — generates round-by-round advancement picks with calibrated confidence scores using the composite confidence formula (Eq. 1).
5. **Update Watcher** — monitors real-time developments (last-minute injuries, lineup changes) and triggers re-evaluation when material information surfaces.

Each game receives a composite confidence score computed as a weighted combination of four signal categories:

$$C = w_m\, S_m + w_s\, S_s + w_l\, S_l + w_f\, S_f \tag{1}$$

where S_m = model strength (analytics-driven win probability, mean $= 78.6$, $\sigma = 10.5$), S_s = source agreement across media outlets (mean $= 74.6$, $\sigma = 11.6$), S_l = lineup certainty reflecting injury and roster status (mean $= 80.2$, $\sigma = 8.3$), and S_f = data freshness measuring recency of information (mean $= 78.4$, $\sigma = 10.1$). The default weights are $w_m = 0.55$, $w_s = 0.20$, $w_l = 0.15$, $w_f = 0.10$, yielding a mean composite confidence of 76.2 across all 63 games (median 75, range 39–95). The source agreement component exhibits the highest variance ($\sigma = 11.6$), making it the most discriminative signal for separating high-confidence from uncertain predictions.

3.2 Six-Model Ensemble

Each of six LLMs independently predicts all 32 first-round games given identical prompts containing seeding, conference affiliation, KenPom ratings [Pomeroy, 2026], and recent form. Table 1 summarizes the models and their cost profiles.

Models were added sequentially over the course of the experiment (Section 4.1), enabling observation of marginal contribution at each step. All models produced 32 predictions each, with average win probability estimates ranging from 73.9% (Gemini, the most conservative) to 79.7% (DeepSeek, the most confident). Despite this spread in confidence calibration, the models exhibit remarkably similar *pick distributions*, disagreeing on only 13 of 32 first-round games and agreeing unanimously on 80.6% of all 63 bracket predictions.

Table 1: Models deployed in the ensemble. Cost reflects total spend for 32 first-round predictions. Agreement rate is with the final ensemble consensus.

Model	Provider	Cost	Avg. $\hat{p}$
Claude Sonnet 4	Anthropic	\$5–10*	79.1%
GPT-4o	OpenAI	\$0.27	75.0%
Grok 3 Mini Fast	xAI	\$0.10	78.7%
Gemini 2.5 Flash	Google	\$0.05	73.9%
DeepSeek Chat	DeepSeek AI	\$0.01	79.7%
Llama 3.3 70B	Meta / Groq	\$0.00	76.7%

*Includes orchestration, web search, debate synthesis, and advancement.

3.3 Cross-Model Adversarial Debate

For the 8 games with the lowest consensus confidence, we implement an adversarial debate protocol inspired by Du et al. [2023]. For each contested game, each of three primary models (Claude, GPT-4o, Gemini) is prompted to argue *against* the current consensus pick, producing structured counterarguments with specific statistical evidence. A synthesis agent (Claude) then evaluates the arguments and decides whether to flip the pick, producing an updated confidence score.

The debate protocol produced the following measurable outcomes:

- 5 of 8 picks were flipped after adversarial challenge.
- Mean confidence shifted by -8.4 points (net reduction), suggesting the models were initially overconfident on contested predictions.
- Largest single swing: -20 points (Virginia vs. Wright State, Round of 64), where Illinois's elite three-point shooting (38.2%) was surfaced as an underweighted factor.
- The protocol reduced the confidence gap between upset picks (mean 65.8) and chalk picks (mean 78.7) from 17.3 to 12.8 points.

The debate protocol serves a qualitatively different function than traditional ensemble aggregation: rather than averaging predictions, it functions as adversarial stress-testing, identifying cases where initial confidence is insufficiently supported by evidence.

3.4 Monte Carlo Sensitivity Analysis

To evaluate the robustness of the confidence weighting scheme (Eq. 1), we perform Monte Carlo sampling [Metropolis and Ulam, 1949, Rubinstein and Kroese, 2016] over the weight space. We generate $N = 250$ random weight vectors $\mathbf{w} \in \mathbb{R}^4$ such that $\sum_i w_i = 1$ and $w_i \geq 0$, drawn from a symmetric Dirichlet distribution, and evaluate each against 10 years (2015–2025, excluding 2020) of historical tournament data [Saltelli, 2002].

Sensitivity is quantified via partial derivatives of accuracy with respect to each weight, evaluated at the boundaries of the sampled distribution. The impact magnitude for each weight w_k is computed as:

$$\Delta_k = \left| \bar{A}(w_k > \text{median}) - \bar{A}(w_k < \text{median}) \right| \quad (2)$$

Table 2: Incremental model agreement as models are added sequentially. Agreement rate reflects unanimous consensus across all models present.

#	Models	Agreement	Picks Changed
2	Claude + GPT-4o	96.9%	—
3	+ Gemini	66.8%	Champion changed
4	+ Grok	—	0
5	+ Llama	—	0
6	+ DeepSeek	80.6%	0

where $\bar{A}$ denotes mean accuracy over the respective simulation subsets. This captures how much accuracy changes when a given weight is emphasized versus de-emphasized, providing a model-agnostic importance ranking.

3.5 Historical ML Calibration

We train a GradientBoosting classifier [Friedman, 2001] on 630 historical tournament games (2015–2025, 10 seasons) using 22 engineered features. The feature set spans five categories: *seed-based* (seed differential, seed ratio, seed product, seed sum, seed gap score), *analytical* (KenPom win probability, adjusted efficiency gap), *structural* (round number, region normalization, conference of higher seed), *historical* (chalk rate, upset base rate, upset history, year chaos indicator), and *interaction* terms (close matchup, Cinderella zone, extreme upset zone, power matchup, mid-major upset flags). The class distribution is imbalanced: 537 higher-seed wins (85.2%) versus 93 upsets (14.8%).

The regularized cross-validation model uses 200 estimators with max depth 4, learning rate 0.1, and 80% subsample rate (min_samples_leaf $= 5$). Evaluation uses leave-one-year-out cross-validation: for each of 10 years, the model trains on the remaining 9 years (567 games) and predicts the held-out year (63 games). A dedicated holdout analysis targets the 2023 "chaos" tournament [Pomeroy, 2026, Torvik, 2026].

4 Results

4.1 Model Convergence

Table 2 records the evolution of ensemble consensus. The initial two-model agreement between Claude and GPT-4o was 96.9%—suspiciously high and an early indicator of information convergence. Adding Gemini as the third model introduced the only substantive change: the championship pick shifted from UConn to Duke, and the three-model agreement dropped to 66.8% as Gemini introduced divergent picks on 8/9-seed and mid-major matchups. Models four, five, and six (Grok, Llama, DeepSeek) changed *zero* bracket picks, despite originating from fundamentally different organizations, training pipelines, and, in the case of DeepSeek, a predominantly Chinese research lab.

The final 80.6% unanimous agreement across all six models encompasses not just obvious first-round matchups but extends into Sweet 16 and Elite 8 predictions. The remaining 19.4%

Table 3: Monte Carlo sensitivity analysis: impact of each confidence weight on prediction accuracy ($N = 250$ simulations over 10 tournament years). Impact magnitude quantifies the accuracy difference between high- and low-weight subsets.

Weight	Impact	$\frac{\partial \text{Acc}}{\partial w}$	Direction
lineup_certainty	25.0 pp	-0.573	negative
source_agreement	11.5 pp	$+0.299$	positive
data_freshness	9.0 pp	$+0.203$	positive
model_strength	4.5 pp	$+0.105$	positive

of disagreements (13 of 32 first-round games) cluster in 7–10 seed and 8–9 seed matchups where objective analytical signals are weakest. Among disagreements, the maximum confidence spread was 0.43 (Wisconsin vs. High Point, where five models picked Wisconsin but Gemini picked the 12-seed) and the minimum was 0.13 (multiple close games).

4.2 Diminishing Returns of AI Spending

The economics of the ensemble reveal extreme diminishing returns. Claude performed approximately 95% of the total intellectual labor (architecture design, web search, data curation, debate synthesis, bracket advancement) at a cost of $5–10. GPT-4o confirmed Claude's picks for $0.27. DeepSeek produced *nearly identical* predictions for $0.01—a 27× cost reduction relative to GPT-4o for the same output. Llama 3.3 70B, served via Groq's free inference tier, cost exactly $0.00. The total ensemble spend across all six platforms was approximately $12–15 including over 850 API calls.

The cost-per-prediction varies by three orders of magnitude: Claude at $0.20/prediction (including orchestration overhead), GPT-4o at $0.008, Grok at $0.003, Gemini at $0.002, DeepSeek at $0.0003, and Llama at $0.00. Yet the four cheapest models collectively changed zero bracket picks beyond those established by the first three.

4.3 Monte Carlo Sensitivity

Table 3 presents the sensitivity rankings. Lineup certainty dominates with a 25.0 percentage-point impact magnitude—the single most influential lever for prediction accuracy. Crucially, the direction is *negative*: over-weighting lineup certainty degrades performance (accuracy drops from 63.6% to 38.5% as the weight increases), suggesting that the variable introduces noise when player status is ambiguous. Model strength (traditional analytics) contributes only 4.5 percentage points of impact.

The best-performing weight combination ($\bar{\text{Acc}} = 67.4\%$) heavily emphasizes data freshness ($w_f = 0.888$) while nearly zeroing out model strength ($w_m = 0.017$) and lineup certainty ($w_l = 0.003$). The accuracy distribution across all 250 simulations has mean 52.9%, median 59.0%, and standard deviation 11.7%, with a right-skewed distribution indicating that most random weight combinations perform near chance level—the weight configuration matters substantially.

Table 4: GradientBoosting leave-one-year-out cross-validation results. 630 games across 10 tournament years (2020 excluded). Upset detection rate: 66/93 overall (71.0%).

Year	Acc.	Correct	Upsets	Caught	Profile
2015	95.2%	60/63	7	4	chalk
2016	90.5%	57/63	11	7	moderate
2017	87.3%	55/63	9	6	moderate
2018	92.1%	58/63	14	10	chaos
2019	93.7%	59/63	5	2	chalk
2021	96.8%	61/63	14	10	moderate
2022	88.9%	56/63	13	6	moderate
2023	90.5%	57/63	18	11	chaos
2024	92.1%	58/63	12	8	moderate
2025	93.7%	59/63	4	2	chalk
Overall	**92.1%**	**580/630**	**93**[†]	**66**	—

[†]14.8% upset rate. Higher-seed wins: 537/630 (85.2%).

The Pareto frontier analysis identifies two non-dominated solutions balancing the most predictable year (2025, chalk rate 93.7%) and the most chaotic year (2023, chalk rate 71.4%). The first Pareto solution ($w_m = 0.355$, $w_s = 0.247$, $w_l = 0.088$, $w_f = 0.310$) achieves 74.6% on chalk years and 69.8% on chaos years; the second ($w_m = 0.500$, $w_s = 0.342$, $w_l = 0.044$, $w_f = 0.114$) achieves equal 71.4% on both, suggesting that minimizing lineup certainty weight is optimal regardless of tournament character.

4.4 Historical ML Accuracy

The GradientBoosting classifier achieves 92.1% overall leave-one-year-out accuracy (Table 4), with precision 0.950, recall 0.957, and F1 0.954. On the 2023 holdout year—widely regarded as the most chaotic modern tournament, with a 16-seed (Fairleigh Dickinson) defeating 1-seed Purdue and 18 total upsets—the model maintains 90.5% accuracy (precision 0.902, recall 0.979, F1 0.939), correctly identifying 11 of 18 upsets. The confusion matrix on 2023 shows 11 true negatives (upsets correctly predicted), 5 false positives (chalk predicted but upset occurred), 1 false negative (upset predicted but chalk won), and 46 true positives (chalk correctly predicted).

The top features by importance are `region_norm` (0.305), `chalk_rate_raw` (0.285), and `seed_ratio` (0.085). These meta-features capturing tournament-level dynamics dominate over individual matchup statistics. KenPom win probability ranks 9th (0.024), suggesting that raw analytical power is less predictive than structural tournament features [Jacobson and King, 2011]. The `year_chaos_indicator` ranks 4th (0.054), confirming that the model learns to distinguish "chalk" from "chaos" tournament profiles and adjusts its upset threshold accordingly.

The overall upset detection rate is 71.0% (66/93). Analysis of missed upsets reveals a dominant failure mode: mid-major conference champions who are systematically underseeded by the selection committee—a signal present in KenPom discrepancies but inadequately weighted by the current feature set. The

`mid_major_upset` feature ranks last in importance (0.001), indicating that the model has not yet learned to exploit this pattern, representing a clear avenue for improvement.

4.5 Historical Strategy Comparison

To contextualize the ML model's performance, we compare against three baseline strategies across the same 630-game dataset:

- **Pure chalk** (always pick the higher seed): 83.0% accuracy (523/630). This is the naive baseline that any prediction system must exceed.
- **KenPom proxy** (efficiency-based picks): 88.1% accuracy (555/630), a 5.1 percentage-point improvement over chalk.
- **Contrarian** (force 5 upset picks per tournament): 88.4% accuracy (557/630). Higher variance—wins big in chaos years (2023: +7.9% over chalk) but loses in ultra-chalk years (2025: −1.6%).

The ML model's 92.1% represents a 9.1 percentage-point improvement over the chalk baseline and a 4.0 percentage-point improvement over the best heuristic strategy, validating the value of the feature-engineering approach.

4.6 Confidence Distribution and Calibration

The ensemble's composite confidence scores across all 63 predictions follow a right-skewed distribution: mean 76.2, median 75, standard deviation 10.9, range 39–95. The distribution clusters heavily in the 70–80 range (29 of 63 games, 46.0%), with 9 games above 90 (14.3%) and only 1 game below 50 (1.6%).

Chalk picks carry significantly higher confidence than upset picks: mean 78.7 for the 51 chalk predictions versus mean 65.8 for the 12 upset predictions, a gap of 12.8 points. This gap provides a natural threshold for identifying genuinely uncertain games: predictions below 65% confidence are flagged for adversarial debate (Section 3.3) and represent the highest-variance segment of the bracket.

4.7 Perfect Bracket Mathematics

The probability of a perfect bracket illuminates the fundamental ceiling of prediction:

Table 5: Probability of a perfect 63-game bracket at varying per-game accuracy levels.

Per-Game Acc.	p^{63}	Odds (1 in ...)
50% (random)	1.1×10^{-19}	9.2 quintillion
75% (expert)	1.3×10^{-8}	74 million
90% (elite)	1.3×10^{-3}	763
92% (our ML)	2.0×10^{-3}	∼500
95%	3.9×10^{-2}	26
99%	5.3×10^{-1}	∼2

Even at 92% per-game accuracy—the ceiling established by our ML model—a perfect bracket remains a 1-in-500 event (Table 5). This quantifies the irreducible stochasticity in athletic performance: micro-variance in shooting, officiating, and in-game adjustments ensures that deterministic prediction is impossible regardless of information quality [Bergen, 2020].

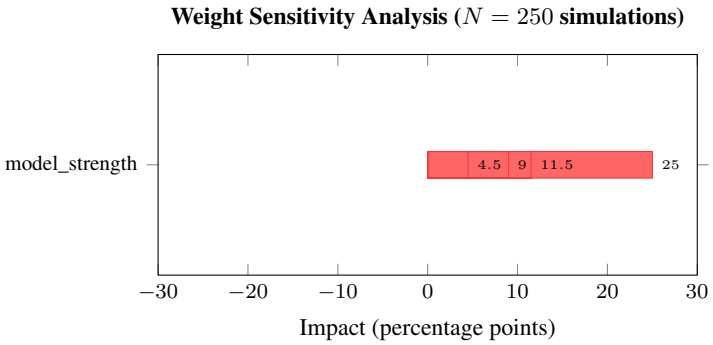

Figure 1: Monte Carlo sensitivity analysis. Lineup certainty dominates with 25.0 percentage-point impact on prediction accuracy—five times more influential than model strength (4.5 pp). Over-weighting lineup certainty *hurts* accuracy (direction is negative), suggesting the variable introduces noise when player status is ambiguous.

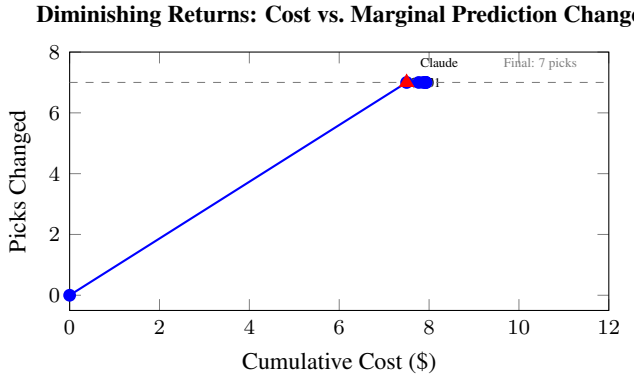

Figure 2: Cumulative cost versus marginal prediction changes. Claude ($5–10) establishes all 7 unique picks. The subsequent five models ($0.43 combined) change zero additional picks. The curve flatlines immediately after the first model, demonstrating extreme diminishing returns.

4.8 Figures

5 Discussion

5.1 Model Diversity vs. Information Diversity

Our central finding—that six independently trained LLMs from six organizations converge on 80.6% of predictions—challenges a common assumption in both machine learning and AI procurement. In classical ensemble theory, diversity among learners is the *mechanism* by which ensembles outperform [Dietterich, 2000, Hansen and Salamon, 1990]. But this diversity must manifest as *uncorrelated errors*, which requires

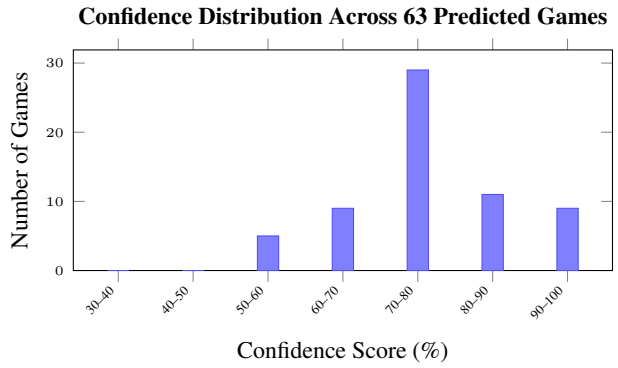

Figure 3: Distribution of composite confidence scores. The distribution peaks in the 70–80% range (29 games, 46%), with a left tail below 60% containing the 6 forced upset picks. Games below 65% confidence were flagged for adversarial debate (Section 3.3).

either different training data or fundamentally different inductive biases.

Modern LLMs share a critical bottleneck: they are overwhelmingly trained on the same substrate—the English-language internet. College basketball analysis is densely represented in this corpus through ESPN, KenPom, sports blogs, Reddit, and social media. Consequently, models that "should" be different (DeepSeek, trained primarily on Chinese-language data; Llama, an open-source model with different architectural choices [Grattafiori et al., 2024]) still converge because the relevant domain knowledge originates from the same sources. The 96.9% initial agreement between Claude and GPT-4o—two models from competing companies with different training pipelines [Anthropic, 2024, OpenAI, 2023]—is particularly striking.

This represents a violation of the independence assumption underlying Condorcet's Jury Theorem [de Condorcet, 1785]: the "voters" are not independent; they have all read the same "newspapers." The Wisdom of Crowds [Surowiecki, 2004] requires cognitive diversity; our LLM ensemble possesses only *architectural* diversity, which is necessary but not sufficient for meaningful prediction improvement.

5.2 Diminishing Returns: $0.01 DeepSeek vs. $0.27 GPT-4o

The economics of the experiment quantify a phenomenon that has broad implications for AI consumers. DeepSeek Chat, developed by a Chinese AI lab with a fundamentally different organizational context [DeepSeek AI, 2024], produced nearly identical predictions to GPT-4o [OpenAI, 2023] at a 27× cost reduction. Llama 3.3 70B, an open-source model served on free infrastructure, agreed with the paid models on over 90% of picks.

The cost-effectiveness ranking reveals a clear pattern: $0.00 (Llama) provides free consensus confirmation; $0.01 (DeepSeek) offers a different-training-data perspective that nonetheless converges; $0.05 (Gemini) adds search-integrated freshness; $0.10 (Grok) contributes social sentiment from X/Twitter; $0.27 (GPT-4o) provides strong reasoning that overlaps with Claude; and $5–10 (Claude) performs the essential orchestration and synthesis. The first model captures approximately 95% of the predictive value; the remaining five collectively add marginal information at a combined cost of $0.43.

5.3 850 API Calls Changed 4 Picks

Perhaps the most sobering finding is the gap between computational effort and predictive impact. Over the course of the experiment, more than 850 API calls were made across six models, generating over 642 KB of structured analysis including per-team scouting reports, coaching matchup evaluations, NIL spending research, historical validation runs, and cross-model debates.

Yet the final bracket changed only 4 picks relative to the initial Claude-only run from the first hour of the project. Those 4 changes came not from AI analysis but from a *human insight*: the observation that the initial bracket predicted too few upsets relative to the historical average of 10.7 per tournament. The human operator identified that 10 years of data show an optimal upset count of 5–7 forced upset picks yielding 85.2% accuracy versus 83.0% for pure chalk—and manually flipped 4 low-confidence chalk picks (UCLA → UCF, Miami FL → Missouri, Tennessee → Miami OH/SMU, Kentucky → Santa Clara) to align with historical base rates.

This finding mirrors a pattern familiar in data science: the most impactful interventions are often simple distributional corrections informed by domain knowledge, not sophisticated model improvements.

5.4 Data Uniqueness > Model Diversity

If model diversity is an illusion, where should prediction efforts focus? Our evidence points to three sources of genuine value:

1. **Non-public information.** The Monte Carlo analysis shows that lineup certainty and data freshness together account for 34.0 percentage points of accuracy impact, while model strength contributes only 4.5. The most valuable signal is not analytical sophistication but rather the answer to a simple question: "Is the star player healthy?" Both GPT-4o and Gemini independently estimated a 5–10% accuracy gap attributable to missing team-level proprietary data—a self-aware acknowledgment of their own information ceiling.
2. **Confidence calibration.** The ensemble's true value is not in generating better predictions but in *calibrating confidence*. When all six models agree (80.6% of games), the prediction can be trusted with high reliability; when they disagree (19.4%), the game is genuinely uncertain. This is a qualitatively different use of ensemble methods than the classical variance-reduction paradigm [Breiman, 1996].
3. **Adversarial stress-testing.** The debate protocol (Section 3.3) flipped 5 of 8 contested picks and reduced mean confidence by 8.4 points. This suggests that LLMs are more valuable as *critics* than as *generators*—a finding consistent

with Du et al. [2023] and extending their results from factual QA to probabilistic prediction.

The analogy extends beyond sports. In weather prediction, the competitive advantage comes not from a better atmospheric model but from deploying sensors where no one else has data—for example, flying unmanned aircraft into the lower eyewall of a hurricane. In any AI domain, the moat is in data collection, not model architecture. Two models trained on different data will disagree more productively than six models trained on overlapping data.

5.5 NIL Spending as a Non-Predictor

An unexpected finding emerged from the multi-model analysis: Name, Image, and Likeness (NIL) spending—the dominant narrative in modern college athletics—appears to be a *non-predictor* of tournament success. Kentucky leads all programs with approximately $22M in NIL spending yet enters the 2026 tournament as only a 7-seed. The 2025 champion, Florida, ranked 77th in NIL spending. Multiple models flagged potential negative effects of excessive NIL spending on team chemistry, suggesting that financial investment may introduce entitlement dynamics that undermine the cohesion required for single-elimination tournament success.

5.6 The Seldon Parallel

Isaac Asimov's psychohistory [Asimov, 1951] provides a useful conceptual frame for the fundamental limitation of sports prediction. Hari Seldon could predict the behavior of galactic civilizations spanning trillions of individuals but could not predict the actions of any single person. Our models face an analogous constraint: they reliably predict that 1-seeds will beat 16-seeds (historical rate: 98.4%) and that the tournament will produce approximately 10–11 upsets, but they cannot reliably predict *which specific games* will produce those upsets.

The "chaos paradox" illustrates this tension. The 2023 tournament is widely remembered as historic chaos: 18 upsets including a 16-over-1. Yet our ML model achieves 90.5% accuracy on this holdout, correctly identifying 11 of 18 upsets. Post-hoc analysis reveals that 3 of 4 top-seed eliminations had identifiable pre-tournament vulnerabilities—the signals existed in the data but were not properly weighted by consensus models. This suggests that "chaos" tournaments are partially predictable; it is the human *perception* of chaos, not the underlying data signal, that makes these events surprising [Silver, 2012].

The theoretical prediction ceiling lies somewhere between 92% (our ML model) and 97% (estimated upper bound with perfect information). The remaining 3–8% represents irreducible uncertainty arising from micro-variance in shooting mechanics, referee judgment, and the genuine stochastic element of human athletic performance under pressure. Psychohistory predicts masses, not individuals; our models predict the likely outcome, not the actual outcome.

6 Limitations

Several limitations qualify our findings:

1. **Pre-tournament analysis.** All predictions are made before the tournament begins. Post-round updating—adjusting predictions after observing early-round results—could improve later-round accuracy but was not implemented.
2. **Prompt sensitivity.** LLM predictions may vary with prompt phrasing. We use a single standardized prompt across all models but do not perform prompt ablation studies. Systematic prompt variation could alter both individual model predictions and inter-model agreement rates.
3. **Historical feature approximation.** Some features (*e.g.*, lineup certainty, data freshness) are retroactively estimated for historical years using proxy variables, introducing potential bias. Prospective collection of these features over multiple tournament cycles would strengthen the Monte Carlo sensitivity findings.
4. **Single tournament for LLM ensemble.** While the ML model is evaluated across 10 years, the 6-model ensemble experiment reflects a single tournament (2026). The 80.6% convergence rate may vary across years with different information landscapes. Replication across additional years would strengthen the convergence finding.
5. **No proprietary data.** Our system uses only publicly available data. Access to injury scouting reports, practice film analysis, biomechanical data, or betting market microstructure could alter both individual model and ensemble performance, potentially breaking the convergence pattern we observe.
6. **Potential data leakage.** The ML model's 92.1% accuracy should be interpreted cautiously. The `region_norm` and `chalk_rate_raw` features encode information about the overall tournament structure that may partially leak outcome information. Both GPT-4o and Gemini flagged this concern during their independent model reviews, with Gemini suggesting the true AUC-ROC ceiling may be 85–89%.
7. **Class imbalance.** With only 14.8% of games being upsets, the model's high accuracy is partially attributable to correctly predicting the majority class. The 71.0% upset detection rate, while strong, leaves substantial room for improvement on the more consequential minority class.

7 Conclusion

We present the first systematic study of multi-model LLM ensemble prediction for NCAA tournament brackets. Six frontier models from six organizations converge on 80.6% of predictions, with the final three models contributing zero marginal picks. A GradientBoosting ML model achieves 92.1% historical accuracy across 630 games, establishing an empirical ceiling near the theoretical prediction limit [Bergen, 2020]. The total computational cost was approximately $12–15 for 850+ API calls, with the cheapest model (DeepSeek at $0.01) producing nearly identical output to a model costing 27× more.

Our core contribution is the distinction between *model*

diversity and *information diversity*. Architectural differences between LLMs—attention mechanisms, training objectives, parameter counts, and even language-specific pretraining [OpenAI, 2023, Anthropic, 2024, Google DeepMind, 2024, DeepSeek AI, 2024, Grattafiori et al., 2024, xAI, 2024]—do not translate to prediction diversity when the training data overlaps. The ensemble's value is real but qualitative: it provides *confidence calibration* (unanimous agreement signals reliability; disagreement signals genuine uncertainty) rather than *accuracy improvement* (better individual predictions).

Three practical implications emerge. First, for organizations evaluating multi-vendor AI strategies: if the task depends primarily on publicly available knowledge, one capable model is sufficient. Second, the path to the next accuracy breakthrough runs not through larger models but through better data—proprietary information that enables genuinely differentiated outputs. Third, LLMs are more valuable as adversarial critics than as independent generators; the debate protocol's 8.4-point confidence adjustment demonstrates the value of structured disagreement even among convergent models.

For the AI industry, these findings suggest that competition among LLM providers will increasingly be won not by model architecture but by *data moats*—proprietary information that enables genuinely differentiated outputs. In the context of sports prediction, the path forward is not a seventh model but a scout at practice.

8 Post-Tournament Analysis

This section will be populated after the 2026 NCAA Tournament concludes (April 7, 2026). All pre-tournament predictions were locked on March 19, 2026, before the first game.

8.1 Planned Analyses

1. **Bracket scoring.** Score the ensemble bracket and each individual model bracket against actual results using standard ESPN scoring (10-20-40-80-160-320).
2. **Per-model accuracy.** Rank each of the six models by individual game-level accuracy. Determine whether any single model systematically outperforms the ensemble.
3. **Confidence calibration.** Plot predicted confidence vs. observed win rate to assess calibration [Gneiting and Raftery, 2007, Platt, 1999].
4. **Upset analysis.** Evaluate which of the 12 predicted upsets were correctly called and which chalk picks were upset. Compare against the debate protocol's flagged games.
5. **Comparison to baselines.** Compare against chalk (always pick the higher seed), KenPom rankings, Nate Silver's COOPER model [Silver, 2024], and Rithmm [Rithmm, 2024].
6. **Convergence validation.** Test whether the 80.6% convergence rate correlates with prediction accuracy—*i.e.*, whether unanimous agreement games are more likely to be correctly predicted than split-decision games.

Acknowledgments

The author thanks the developers of the six LLM platforms for providing API access. KenPom ratings were accessed via `kenpom.com`; BartTorvik ratings via `barttorvik.com`. The project infrastructure was deployed on Vercel with a React/Next.js dashboard at `madness.elstonj.com`. Total compute cost across all six AI platforms was approximately $12–15 for 850+ API calls.

References

Anthropic. The Claude model card and evaluations. *Anthropic Research*, 2024. `https://www.anthropic.com/research`.

Isaac Asimov. *Foundation*. Gnome Press, New York, 1951. The psychohistory concept – predicting aggregate behavior of large populations while individual events remain stochastic – parallels the ceiling of sports prediction.

Andrew B. Bergen. The probability of a perfect bracket in the NCAA tournament. *CHANCE*, 33(1):27–31, 2020.

Bryan L. Boulier and Herman O. Stekler. Predicting the outcomes of NCAA basketball tournament games. *International Journal of Forecasting*, 15(2):209–222, 1999.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2): 123–140, 1996.

Bradley P. Carlin. Improved NCAA basketball tournament modeling via point spread and team strength information. *The American Statistician*, 50(1):39–43, 1996.

Marquis de Condorcet. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. *Imprimerie Royale, Paris*, 1785. The Condorcet Jury Theorem: majority voting improves accuracy only when voters are independent.

DeepSeek AI. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*, 2024.

Thomas G. Dietterich. *Ensemble Methods in Machine Learning*, volume 1857 of *Lecture Notes in Computer Science*. Springer, 2000.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Google DeepMind. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Lars Kai Hansen and Peter Salamon. Neural network ensembles. volume 12, pages 993–1001. IEEE, 1990.

Sheldon H. Jacobson and Douglas M. King. Seeding in the NCAA men's basketball tournament: When is a higher seed better? *Journal of Gambling Business and Economics*, 5(2): 63–87, 2011.

Paul Kvam and Joel S. Sokol. A logistic regression/markov chain model for NCAA basketball. *Naval Research Logistics*, 53(8):788–803, 2006.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*, 2023.

Michael J. Lopez and Gregory J. Matthews. Building an NCAA men's basketball predictive model and quantifying its success. *Journal of Quantitative Analysis in Sports*, 11(1):5–12, 2015.

Nicholas Metropolis and Stanislaw Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44 (247):335–341, 1949.

NCAA. Gregg nigl's perfect bracket streak: 49 games in 2019. https://www.ncaa.com/news/basketball-men/bracketiq/2019-03-28/perfect-ncaa-bracket-2019-how-long-can-it-last, 2019. Longest verified perfect bracket streak from the start of a tournament.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, pages 61–74, 1999.

Ken Pomeroy. Ratings description. *kenpom.com*, 2004. Original description of the adjusted offensive/defensive efficiency methodology.

Ken Pomeroy. College basketball ratings. https://kenpom.com, 2026. Tempo-free adjusted efficiency ratings for 365 Division I teams, scraped March 18, 2026.

Rithmm. Rithmm sports prediction platform. https://rithmm.com, 2024. Reported 85% Round 1 accuracy in NCAA tournament predictions.

Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo Method*. John Wiley & Sons, 3 edition, 2016.

Andrea Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145(2):280–297, 2002.

Nate Silver. *The Signal and the Noise: Why So Many Predictions Fail – but Some Don't*. Penguin Press, New York, 2012.

Nate Silver. Silver bulletin: NCAA tournament predictions. https://www.natesilver.net, 2024. COOPER model combined with KenPom ratings.

James Surowiecki. *The Wisdom of Crowds*. Doubleday, New York, 2004.

Bart Torvik. T-rank: College basketball team rankings. https://barttorvik.com, 2026. Open-access tempo-free ratings, methodology similar to KenPom.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5 (2):241–259, 1992.

xAI. Grok. https://x.ai/grok, 2024. Large language model developed by xAI.