# PLAYING THE ODDS: AGENTIC LLMS FOR REAL-TIME NBA FORECASTING AND MARKET BETTING

Ethan Jeon[1*], Minseong (Leo) Sim[2], Woohyun Kim[3]
1. Yongsan International School of Seoul, 285 Itaewon-ro, Yongsan-gu, Seoul, 04347, South Korea
2. Chadwick International School, 45 Art Center-daero 97beon-gil, Yeonsu-gu, 21985, South Korea
3. Seoul Science High School, 63, Hyehwa-ro, Jongno-gu, Seoul, 03066, South Korea
* Corresponding author email: 10.ethan.jeon@gmail.com

## Abstract

Predicting the outcomes of professional basketball games is a challenging problem due to the intrinsic stochastic uncertainty of sports competitions and the heterogeneity of relevant information sources. While existing approaches in sports analytics primarily rely on structured historical statistics, such methods often struggle to incorporate timely and unstructured information. In this paper, we propose a unified framework that leverages large language models for probabilistic forecasting and decision making for NBA games and their prediction markets. Our approach combines specialized information retrieval agents with multiple role-based LLM predictors, whose forecasts are aggregated into the final forecasting probabilities. These probabilities are then operationalized through a fractional Kelly betting strategy in binary prediction markets. We evaluate the proposed system using both Brier Scores and simulated market returns, demonstrating that LLM-based forecasting can effectively complement traditional models and translate predictive improvements into economic values. Overall, our results demonstrate the potential of LLMs and in-context learning as flexible tools for decision-oriented sports analytics.

## Keywords

Sports analytics, large language models, Game prediction

## Introduction

Predicting the results of professional sports games has long been a central problem in sports analytics, with its applications ranging from coaching strategy to betting markets. However, as the volume and complexity of game-related data have increased, the limitations of traditional forecasting approaches have become evident. Early sports sciences and analytics relied heavily on the expertise of coaches and analysts, with competition data primarily being collected and interpreted manually (Ozkan 2020; Horvat et al. 2023). And while this was effective in small-scale settings, such approaches struggled to scale up as data volume grew, often resulting in incomplete utilization of the available information and limited predictive accuracy (Zhao, Du, and Tan 2023).

Advancements in computing power and data infrastructure initially enabled the adoption of machine learning (ML) techniques in sports analytics. In the NBA, researchers have explored a wide range of models, including naive bayes, artificial neural networks, and decision tree models (Thabtah, Zhang, and Abdelhamid 2019), achieving varying degrees of success in predicting the outcome of games (Tax and Joustra 2015; Pathak and Wadhwa 2016). Despite methodological diversity, state-of-the-art models typically report predictive accuracies in the range of approximately 67% to 75% (Gumm, Barrett, and Hu 2015; Kim, Magnusen, and Jeong 2023), suggesting both meaningful gains over heuristic baselines and persistent limits to purely data-driven performance in a highly stochastic environment.

More recently, progress in natural language processing has substantially expanded the scope of machine learning methods, particularly through the emergence of large language models (LLMs). Built on transformer architectures, LLMs are trained on massive corpora of text drawn from heterogeneous sources, enabling them to capture rich, semantic, contextual, and relational patterns in language. Rapid improvements in computational resources, model scale, and training data have led to dramatic gains in LLM performance across a wide range of different tasks (Wei et al. 2022). These advances have been most visible through interactive systems such as OpenAI's ChatGPT, which expose LLM capabilities through open-ended prompt–completion interfaces. In contrast to traditional machine learning pipelines that require carefully engineered numeric features, LLM-based approaches operate directly on unstructured text, therefore bypassing explicit feature construction and allowing flexible adaptation to diverse prediction problems (Liu et al. 2023).

An additional strength of LLMs lies in their ability to perform in-context learning. By conditioning on a small number of task-specific examples provided at inference time, LLMs can approximate task-specific behavior without parameter updates or retraining. Techniques such as few-shot prompting have been shown to generate robust performance across domains, even when labeled training data are scarce or costly to obtain (Brown et al. 2020). These properties suggest that LLMs provide a promising alternative to conventional numerical models, particularly in settings where relevant predictive information is embedded in textual, qualitative, or rapidly evolving sources.

Motivated by these developments, we explore whether large language models can be leveraged to improve probabilistic forecasting of NBA game outcomes by incorporating information sources that are largely inaccessible to traditional numerical models. Unlike prior basketball prediction studies that rely almost exclusively on structured historical statistics, our approach exploits timely and unstructured textual information, including injury reports, news coverage, and media content surrounding upcoming games. We design a unified LLM-based forecasting framework in which external information is first collected by specialized search agents and then synthesized by multiple role-specific predictors through prompt-based inference. By providing recent contextual evidence and, where appropriate, illus- trative examples of past games, the model produces calibrated win probabilities rather than point predictions. This design allows the forecasting system to adapt dynamically to rapidly evolving pre-game conditions while remaining compatible with downstream decision-making tasks, such as trading in binary prediction markets. More broadly, our framework demonstrates how LLMs and in-context learning can be integrated into sports analytics pipelines, offering a complementary alternative to purely numeric approaches in highly stochastic environments like professional basketball.

# Methods

## *Problem Description: LLM-based Forecasting and Trading Agents*

We study the problem of probabilistic forecasting and decision-making for the professional basketball games under heterogeneous information and time constraints. Our setting combines real-time game outcome prediction with the downstream actions in binary prediction markets, such as Kalshi.

## *2.1 Modeling the Game Outcomes*

In every NBA game, there is always a winning team and a losing team. No games end in a tie, meaning that a given team for a given game will always be limited to just one of two results: a win or a loss. In our case, the games will always be in the perspective of the home team.

Let $j \in \{1, \ldots, N\}$ index NBA games.

Each game resolves to a binary outcome

$$o_j \in \{0,1\},$$

where $o_j = 1$ denotes a win by the home team and $o_j = 0$ otherwise.

## *2.2 Information Sets, Probabilistic Forecasting and Dynamism*

Unlike classic pattern recognition tasks, a key challenge of event forecasting is its dynamism – that is, both available information for forecasting and the predicted probabilities based on the information evolves over time (before the event's resolution). Hence the time stamp of the forecasting matters a lot. At any given moment before the start of game $j$, the forecaster observes and information set $Ij$.

Given $Ij$ , the forecaster outputs a predicted probability:

$$\rho_j \in [0,1]$$

It is interpreted as the probability that the home team wins game $j$ at that given moment.

We use probabilistic forecasting rather than just predicting the results since probabilities are more informative and directly usable in downstream decision-making. Forecasting quality is evaluated using scoring rules such as the Brier Scores.

## *Dynamism.*

The information is heterogeneous and time-varying, consisting of: (i) Historical Statistics; (ii) Official injury reports and lineup announcement; (iii) news articles and social media posts.

## *2.3 Prediction Markets and Contracts*

Prediction markets are markets that allow people to trade contracts, where the value of these contracts depend on the outcome of a future event. The idea is that the trading price reflects what the market collectively believes about the likelihood of that outcome. For example, the more likely that the market believes that an event will happen, the more expensive the Yes contract for that event will become, and vice versa.

Each game $j$ is associated with a binary prediction market.

• A Yes contract pays \$1 if $o_j = 1$ and \$0 otherwise, and is traded at price $q_j \in [0,1]$.

• A No contract pays \$1 if $o_j = 0$ and costs $1 - q_j$.

These prices are often interpreted as the market-implied probabilities.

## *2.4 Sporting Betting and the Kelly Model*

To formalize the betting decision, we use a fractional Kelly criterion.

Given a predicted win probability $p_j$ and contract price $q_j$ , the Kelly-optimal fraction of capital allocated to the favorable side is:

$$f_j \frac{(1-q_j)p_j - q_j(1-p_j)}{1-q_j}$$

When betting on the Yes contract, and analogously for No. In practice, we use a fractional Kelly strategy,

$$f_j = \alpha max(0, f_j^{Kelly})$$

where $\alpha \in 0, 1$ controls risk aversion.

## *3 The Design of an LLM-based Sport Betting Agent*

We propose an end-to-end framework that integrates probabilistic forecasting with market-based decision-making for NBA games under real-time constraints. The approach consists of several steps.

### *3.1 Phase 1: Retrieval of Event-relevant Data Sources via a Searcher Agent*

We designed a unified probabilistic forecasting framework on large language models, consisting of two core components: *searchers* and *predictors*. The searchers are responsible for collecting the relevant external information, while the predictors synthesize this information to generate probability forecasts for future NBA games outcomes.

For each NBA game $j$, we retrieve an associated binary prediction market from NBA website and Kalshi. NBA website provides the schedule of the incoming games, and Kalshi provides a *Yes* contract price $q_j$ and a *No* contract price $1 - q_j$.

We restrict the sample to markets that are still open, ensuring that all decisions are made before the game outcome is realized.

#### *Searchers*

On the search side, we design a set of specialized search agents which their roles are stated by their prompt. They are each responsible for a distinct dimension of NBA game outcomes, and this design enforces a clear separation of informational roles by decomposing the search process into multiple agents instead of aggregating all external data into a single unstructured input. Each agent is dedicated to a specific factor that is known to influence basketball performance. We instruct each agent to collect reliable, up-to-date information relevant to its assigned dimension. The following six search agents(searchers) are used:

- **Injury Searcher**: focuses on collecting information on player injuries across all teams, including individual injury reports and expected absences or playing restrictions
- **Star Injury Searcher:** specializes in tracking player injuries with particular emphasis on each team's star player, closely monitoring injury reports, expected absences, and restrictions that may significantly affect team performance
- **Extended Injury Searcher:** provides a comprehensive injury overview for each NBA team by analyzing official injury reports, latest updates, return timelines, and minutes restrictions, while explicitly evaluating how star player injuries influence team performance and win probability
- **Home Advantage Searcher:** analyzes home-court advantage by examining the home team's performance, the away team's road performance, and historical home-court advantage statistics
- **Rest Searcher:** evaluates scheduling-related factors by tracking days of rest and travel distance for each team to assess fatigue and recovery effects
- **Team Momentum Searcher:** assesses team momentum by monitoring winning and losing streaks as well as current league rankings

Each searcher independently queries the web using a role-specific prompt and retrieves 5 sources. The union of these sources is then passed to the prediction stage.

### *3.2 Phase 2: Probabilistic Forecasting with LLMs via a Predictor Agent*

On the prediction side, we use multiple predictors, each representing a different forecasting perspective. This ensemble design is motivated by the idea that forecasting benefits from combining heterogeneous beliefs rather than relying on a single persona. We employ six predictors:

- **Analyst Predictor:** emphasizes quantitative evidence such as recent performance trends and historical statistics, and produces more conservative predictions
- **Fan Predictor:** is confident of the team, and captures team chemistry and intuitions
- **Sportsbook Trader Predictor:** mimics a market-orientated perspective, accounting for implied odds and risk management considerations
- **Kalshi Trader Predictor:** mimics a market-orientated perspective, but using specifically Kalshi to match public sentiment and beat the market with sharper predictions
- **Home Team Coach Predictor:** views the matchup with a basketball-oriented mind through the home team's strategy and personnel, weighing factors like player matchups, team morale, and on-court chemistry with a natural bias for their own team
- **Away Team Coach Predictor:** views the matchup with a basketball-oriented mind through the away (road) team's strategy and personnel, weighing factors like player matchups, team morale, and on-court chemistry with a natural bias for their own team

Each predictor receives the same aggregated set of sources but is guided by a role-specific prompt.

## *Probability Aggregation*

Let $s$ denote the Brier Score (for the definition of the Brier Score, see Section 4.1) of by predictor $k \in \{1, \ldots, K\}$. For each Brier score $s$, we convert the score into $1 - s$, because a lower Brier Score is better. This makes the lower values correspond to better performance and a higher weight.

This is the inverse sigmoid function that we will be using:

$$\sigma^{-1}(x) = ln(\frac{x}{1-x})$$

We plug this value $1 - s$ into the inverse sigmoid (logit) function to find the raw weight of each predictor.

$$w_k = \sigma^{-1}(1-s) = ln(\frac{1-s}{1-(1-s)}) = ln(\frac{1-s}{s})$$

We then normalize the weights by taking a sum of the raw weights of our K predictors and dividing each raw weight by the sum:

$$Normalized\ Weight = \frac{w_k}{\sum_{l=1}^{K} w_l}$$

Then the aggregated the predicted probability for game $j$ is computed as:

$$p_j{}^{agg} = \sum_{k=1}^{K} w_k p_{j,k}$$

## *3.3 Phase 3: Betting Optimization and Execution via a Trading Agent*

To translate probabilistic forecasts into trading decision, the understanding employs a decision rule based on expected (EV) with fractional Kelly betting. This framework shows how rational agents should act in prediction markets when probabilities are estimated with uncertainty.

## *The Fractional Kelly Betting*

After selecting the trade direction, the position size is determined using the Kelly Criterion (Wysocki 2025) which specifies the optimal fraction of capital to bet in order to maximize long-run benefit. For a contract purchased at price c with win probability $p_{adj}$, the Kelly fraction $f^*$ is given by:

$$f^* = \frac{(1-c)p_{adj} - (1-p_{adj})c}{c}$$

In practice, full Kelly betting assumes accurate probability estimates and can therefore lead to excessive volatility when probabilities are not specified. To reduce the risk, the strategy employs *fractional Kelly betting*, scaling the optimal fraction by a constant $\alpha \in (0,1)$: $i.e., f = \alpha f^*$.

In our study, we set α = 0.20, meaning that only 20% of the full Kelly criterion is used. This conservative adjustment acknowledges in $p_{adj}$ and reduces the impact of probability estimation errors on position sizing.

Additionally, a hard cap is imposed on the amount of betting per game. The final stake size is therefore computed as:

$$Stake = min(Bankroll \times f, MaxBet)$$

This cap prevents the strategy from allocating excessive capital to a single trade, even when the Kelly criterion suggests a large position. Such position limits are standard in real-world trading markets and serve as an additional layer for risk management. Algorithm 1 presents the full algorithmic procedure as a concise, rule-based framework.

**Algorithm 1** Trading Algorithm with EV Threshold and Fractional Kelly

1. Inputs:
   - Adjusted probability ($p_{adj}$)
   - Market YES price ($mktYES$)
   - Market NO price ($mktNO$)
   - Current Bankroll ($B$)
   - EV threshold ($\tau$)
   - Fractional Kelly parameter ($\alpha$)
   - Maximum stake per game ($C$)

2. Step 1: Compute expected values
   - $EV_{YES} = p_{adj} - mktYES$
   - $EV_{NO} = (1 - p_{adj}) - mktNO$

3. Step 2: Apply EV threshold
   - If $max(EV_{YES}, EV_{NO}) < \tau$, set decision = HOLD and stop.

4. Step 3: Select trade side
   - If $EV_{YES} \geq EV_{NO}$:
     –side=YES, $p_{win} = p_{adj}, c = mktYES$
   - Else:
     – side = NO, $p_{win} = 1 - p_{adj}, c = mktNO$

5. Output:
    - Trade decision (YES / NO / HOLD)
    - Number of contracts
    - Execution Price

### *Implications of Using Adjusted Probabilities*

The use of $p_{adj}$ highlights that trading decisions are made under uncertainty rather than perfect foresight. The combined use of adjusted probabilities expected value thresholds, fractional Kelly sizing, and position caps leads to three key outcomes. First, many markets are intentionally skipped,

reflecting disciplined selectivity. Second, higher EV thresholds results in fewer but more profitable trades which indicates that filtering weak signals is more important than trade frequency. Lastly, the results, demonstrate that the Kelly criterion is effective when it is used with conservative adjustments that reduce errors in probability estimation.

Overall, this framework shows that profitability in prediction market depends not only on estimating probabilities, but on recognizing and managing uncertainty in those predictions. These considerations are incorporated into rule-based trading algorithm that uses adjusted probability estimates, expected value filtering, and fractional Kelly sizing to determine whether and how to trade.

## Results and Discussion

We evaluate the proposed system from both a predictive and decision-making perspectives. Our primary focus is on the probabilistic calibration and downstream economic performance in the prediction markets.

### *4.1 Evaluation Metrics.*

### *4.1.1 Brier Score*

Since our model outputs probabilistic forecasts, we evaluate the predictive quality using the Brier Score. For a set of N games, the Brier Score is defined as:

$$BS = \frac{1}{N}\sum_{j=1}^{N}(p_j - o_j)^2$$

where $p_j \in [0,1]$ is the predicted probability that the home team wins game $j$, and $o_j \in \{0,1\}$ is the realized outcome of game $j$.

### *Proposition 1*

*Suppose $p, q$ are two different probability forecast for an event $E$,then the Brier score for $\lambda p + (1-\lambda)q$ is a quadratic function in $\lambda \in [0,1]$.*

*Proof.* Suppose $p, q \in [0,1]$ are two distinct probabilistic forecasts for a binary event $E$, and let:

$$\hat{p}(\lambda) = \lambda p + (1-\lambda)q, \qquad \lambda \in [0,1]$$

Let the $p^* \in [0,1]$ denote the true probability of E, and $O \in \{0,1\}$ denote the realized outcome of the event, with:

$$P(O=1) = p^*, \qquad P(O=0) = 1 - p^*$$

The Brier score for a forecast r is:

$$\begin{aligned} BS(\lambda) &= E[(\hat{p}(\lambda) - O)^2] \\ &= E[\hat{p}(\lambda)^2 - 2\hat{p}(\lambda)O + O^2] \\ &= \hat{p}(\lambda)^2 - 2\hat{p}(\lambda)E[O] + E[O^2]. \end{aligned}$$

Since $O \in \{0,1\}$, we have $O^2 = O$, and therefore,

$$E[O] = E[O^2] = p^{*\,*}$$

which gives:

$$BS(\lambda) = \hat{p}(\lambda)^2 - 2p^*\hat{p}(\lambda) + p^*$$

Since

$$\hat{p}(\lambda) = q + \lambda\,(p - q)$$

is a linear function of $\lambda$. Hence the $\hat{p}(\lambda)^2$ is a quadratic function of $\lambda$, therefore, the brier score $BS(\lambda)$is a quadratic function of $\lambda$ as well.

More explicitly, letting $\Delta = p - q$, we obtain:

$$BS(\lambda) = \Delta^2\lambda^2 + 2\Delta(q - p^*)\lambda + (q - p^*)^2 + p^*(1 - p^*)$$

## *Proposition 2*

*The function $BS(\lambda)$ is convex in $\lambda$ and is monotone on $[0,1]$ if and only if p∉ $[min(p,q), max(p,q)]$.*

*Proof.* From the quadratic expansion in the previous claim, the coefficient of $\lambda^2$ is $(p-q)^2 \geq 0$, so $(BS)(\lambda)$ is convex.

Take the derivative over $\lambda$:

$$BS'(\lambda) = 2(p - q)\hat{p}(\lambda) - p^*$$

Setting the derivative to zero gives the minimizer:

$$\lambda^* = \frac{p^* - q}{p - q}$$

if $\lambda^* \in (0,1)$, then the $BS(\lambda)$ decreases for $\lambda < \lambda^*$ and increases for $\lambda > \lambda^*$, and hence not monotone on $[0,1]$. This occurs if and only if $p^* \in (min(p,q), max(p,q))$

If instead $\lambda^* \leq 0$ or $\lambda^* \geq 1$, the minimum is attained at an endpoint, and the function is monotone on [0, 1].

The condition $p^* \notin [min(p,q), max(p,q)]$ implies either $p, q$ are both smaller than $p^*$, or are both larger – that is, they make the same forecasting mistake of being overly optimistic or overly pessimistic. Claim 2 formally show that in such situation, the mixed Brier score is monotone, hence the best choice is to choose $\lambda = 0$ or 1, i.e., to not mix. On the other hand, if $p^* \in [min(p,q), max(p,q)]$, this implies one of $p, q$ overestimates $p^*$ whereas the other underestimates $p^*$. In this situation, mixed forecaster can do better. This echoes the fundamental insight of boosting (Schapire 1990) that boosting is helpful when different weak learner/forecasters have strengths over different subsets of data points or tasks.

### *4.1.2 Expected Calibration Error*

While proper scoring rules such as the Brier Score evaluate the overall accuracy of probabilistic forecasts, they do not directly assess whether the predictions are well calibrated. Calibration measures whether events predicted to occur with probability p indeed occur approximately p fraction of the time.

We evaluate calibration using the Expected Calibration Error (ECE). The ECE quantifies the discrepancy between predicted probabilities and empirical outcome frequencies by partitioning predictions into bins.

Let $\{(p_j, o_j)\}_{j=1}^{N}$denote the predicted probabilities and realized outcomes for $N$ NBA games, where $p_j \in [0,1]$ is the predicted probability that home team wins game $j$, and $o_j \in \{0,1\}$ is the realized outcomes.

We divide the unit interval $[0,1]$ into $M$ disjoint bins $\{B_m\}_{m=1}^{M}$ For each bin $B_m$, define:

- $|B_m|$: the number of games whose predicted probabilities fall into bin $B_m$,
- $conf(B_m) = \frac{1}{|B_m|}\sum_{j \epsilon B_m} p_j$ : the average predicted probability in the bin,
- $acc(B_m) = \frac{1}{|B_m|}\sum_{j \epsilon B_m} o_j$ : the empirical frequency of home-team wins in the bin.

The Expected Calibration Error is then defined as:

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} |acc(B_m) - conf(B_m)|$$

Intuitively, ECE measures the weighted average absolute deviation between predicted probabilities and observed outcome frequencies across bins. A perfectly calibrated forecasting system attains ECE = 0, while larger values indicate increasing miscalibration.

### *4.1.3 Simulated Market Returns*

To evaluate the system, we simulate the market settlement. Given the realized outcomes $o_j$, the profit-and-loss (PnL) for each bet is computed as the contract payoff minus its cost.

## *4.2 Evaluation of the Forecasting Agent*

## *Experimental Setup*

To evaluate the performance of the forecasting agents, a total of 39 NBA games across 5 days were evaluated. Three initial predictor agents (Analyst, Fan, Sportsbook Trader) were used for the first 29 games across 4 days, and for the final day with 10 games, an additional three predictors (Kalshi Trader, Home Team Coach, Away Team Coach) were added on top of the existing three that we originally had. Games from December 29th, 2025 to January 2nd, 2026 were used.

To set up the testing, a prompt was fed into the forecasting agent, and this prompt asked the agent for a probability between 0 and 1 for the home team to win for this specific NBA game. The two teams involved in the game, the home team, the date, and the location of the game were involved in this prompt, making it detailed and clear in what we were asking for. The close times for the LLM to obtain sources and conduct research were set to prior to the start of the given game.

After each run, the predicted win probabilities for each of the different predictor agents involved in forecasting that NBA game were recorded on a spreadsheet. The descriptions of each of these agents and their specific roles can be found in Section 3.2. In addition, to compare our predictor agents' results with the actual market and what the general population thought, we also recorded the home team win probabilities for all of the games from Kalshi, an online prediction market. These win probabilities were retrieved from Kalshi at the point 24 hours before the start of each NBA game. This was done to ensure that a stable and accurate prediction of the prediction market was obtained, as fluctuations in the win probabilities would be too volatile close to the start of the game or especially during the game itself. It also wouldn't be a fair prediction, as these fluctuations would likely be influenced by information that wouldn't have been available to our forecasting agent at the close time that we set.

Finally, in addition to recording the win probabilities for each of the predictor agents and the Kalshi market, we also recorded the binary realized outcome for the home team after the games had taken place. A win for the home team results in 1, and a loss for the home team results in 0.

## *Experimental Results*

Table 1 reports the predicted home team win probabilities produced by different forecasting personas, or agents, together with the realized game outcomes and the corresponding market prices. This is data from NBA games on Janurary 2nd, 2026, and these were the 10 games in which all 6 predictor agents were used.

For each NBA game, we collect the home-team winning probabilities from multiple predictors. These forecasts are compared with the realized game outcome, where a home-team winning coded as 1. We also included the Kalshi market prices as a market-implied benchmark probability.

Table 1: Predicted Home Win Probabilities and Game Outcomes

| Date | Game | Home Team | Analyst Prob | Fan Prob | Sportsbook Trader Prob | Kalshi Trader Prob | Home Coach Prob | Away Coach Prob | Result (Win=1) | Market Prob (Kalshi) |
|---|---|---|---|---|---|---|---|---|---|---|
| Jan 2, 2026 | 1: IND vs SAS | IND | 0.30 | 0.33 | 0.33 | 0.33 | 0.40 | 0.34 | 0 | 0.33 |
| Jan 2, 2026 | 2: WAS vs BKN | WAS | 0.64 | 0.70 | 0.62 | 0.60 | 0.64 | 0.59 | 1 | 0.59 |
| Jan 2, 2026 | 3: CLE vs DEN | CLE | 0.74 | 0.82 | 0.78 | 0.74 | 0.74 | 0.75 | 1 | 0.86 |
| Jan 2, 2026 | 4: NYK vs ATL | NYK | 0.71 | 0.76 | 0.66 | 0.67 | 0.71 | 0.58 | 0 | 0.72 |
| Jan 2, 2026 | 5: CHI vs ORL | CHI | 0.37 | 0.43 | 0.35 | 0.36 | 0.42 | 0.35 | 1 | 0.35 |
| Jan 2, 2026 | 6: MIL vs CHA | MIL | 0.73 | 0.75 | 0.73 | 0.74 | 0.74 | 0.68 | 1 | 0.66 |
| Jan 2, 2026 | 7: NOP vs POR | NOP | 0.66 | 0.77 | 0.68 | 0.65 | 0.68 | 0.66 | 0 | 0.48 |
| Jan 2, 2026 | 8: PHX vs SAC | PHX | 0.76 | 0.81 | 0.77 | 0.81 | 0.76 | 0.69 | 1 | 0.84 |
| Jan 2, 2026 | 9: GSW vs OKC | GSW | 0.22 | 0.21 | 0.18 | 0.22 | 0.24 | 0.24 | 0 | 0.16 |
| Jan 2, 2026 | 10:LAL vs MEM | LAL | 0.64 | 0.69 | 0.64 | 0.65 | 0.65 | 0.58 | 1 | 0.60 |

Table 2 below shows the Brier scores for each of the 6 different personas, or predictor agents, along with the Brier score of the Kalshi prediction market. As mentioned in the experimental setup, all 6 agents were only tested for 10 games, while the first 3 were tested for all 39 games that were evaluated as the sample size.

Table 2: Brier Score for Each Predictor Agent

| | Analyst | Fan | Sportsbook Trader | Kalshi Trader | Home Coach | Away Coach | Market (Kalshi) |
|---|---|---|---|---|---|---|---|
| Brier Score (10 games) | 0.1932 | 0.1966 | 0.1910 | 0.1892 | 0.1965 | 0.1973 | 0.1795 |
| Brier Score (39 games) | 0.1982 | 0.2031 | 0.1990 | N/A | N/A | N/A | 0.1876 |

## *Key Takeaways*

There are multiple key takeaways that can be formed from the evaluation and results of the forecasting agent. First, Table 1 shows substantial heterogeneity across forecasting personas. This means that each persona, as it should, gives its own unique opinion on the game at hand in according to their personal role and what they look at and are more knowledgeable in. For example, it can be seen that fans and coaches tend to produce more extreme probabilities, while the analyst and trader predictors are generally more conservative. That is why a combination of different predictors is often used, as it can be seen as the most beneficial and accurate, as it incorporates all different unique views, taking each agent's strengths to make the best possible prediction. And even in situations where a combination of different agents don't result directly in a lower Brier score for a given sample size of games, this method can still be worth integrating. Oftentimes, especially in smaller sample sizes, results may be unstable as one or two games can largely impact the final score as a whole. One decently wrong prediction or upset in an NBA game could sway a Brier score up or down within a small set of NBA games. This leads to certain agents showing to be more accurate in some sets of

games, while others showing to be more accurate in another set of games, which is why a blend can still be helpful in the grand scheme and looking towards a more consistent forecasting agent.

Next, as seen in Table 2, it is an extremely difficult task to beat the market in accuracy, especially over a big range of games. A prediction market like Kalshi is built off of real people betting their money as they see events and information come up in real time. It is also therefore an aggregation of a wide range of opinions in the population, and this inevitably makes a prediction market like Kalshi immensely consistent and a naturally excellent forecaster. Beating the market consistently would mean that money could be earned at a high rate in the real world.

Finally, one trend that was noticed for these NBA games was that the predictor agents and the LLM as a whole tend to not be as confident in their predictions for lopsided games. When one team is clearly better than another team, the prediction market and the general population tend to favor the better team heavily, regardless of home/away games, injuries, rest days, travel distance, or other factors. However, the predictor agents seem to overestimate the impact of these secondary factors, oftentimes giving a bad team a higher chance of winning than what they are capable of. And while this strategy can prove successful sometimes through upsets, these upsets are unexpected and are upsets for a reason. They are rare and one underdog win here or there doesn't make up for the constant, consistent predictions that the prediction market makes game by game. This phenomenon also proves to be truer as the sample size gets larger, as outliers are increasingly overshadowed by the majority, making the more consistent and bolder predictor more accurate regarding these lopsided games.

## *4.3 Evaluation of the Trading Agent*

Experimental Setup To evaluate the performance of the trading agent, a series of controlled simulation experiments were conducted using real-world NBA prediction market prices. Market prices for YES and NO contracts were retrieved from an online prediction market platform, Kalshi, where each contract pays $1 if the corresponding event occurs and $0 otherwise. These prices reflect the aggregated beliefs of market participants and serve as a benchmark for evaluating the agent's decisions.

The trading agent operates in two stages. First, it produces a forecasted probability for each game outcome. These probabilities are treated as *adjusted probabilities*, acknowledging uncertainty and potential estimation error rather than assuming perfect foresight. Second, the agent compares these adjusted probabilities against market prices using an expected value (EV) framework.

$$EV_{YES} = p_{adj} - yes\_price, \qquad EV_{NO} = (1 - p_{adj}) - no\_price.$$

Trades are executed only when the estimated edge exceeds a predefined EV threshold $\tau$ , such that

$$max(EV_{YES}, EV_{NO}) \geq \tau$$

When this condition is satisfied, the agent selects the contract (YES or NO) with the higher expected value; otherwise, it chooses to HOLD and does not place a trade. Position sizes are then determined using fractional Kelly betting with a fixed cap on the maximum stake per game.

The EV threshold is introduced to control the trade-off between selectivity and aggressiveness, limiting trades based on marginal edges that may arise from estimation noise rather than genuine market mispricing. To assess the sensitivity of the agent to this decision criterion, experiments are conducted under multiple threshold settings.

All experiments are performed in a backtesting environment that simulates contract settlement using realized game outcomes. The agent begins with a fixed initial bankroll and trades sequentially across games, allowing profits and losses to accumulate over time.

### *Experimental Results*

To assess the sensitivity of the trading agent to its decision criteria, multiple experiments were conducted using different EV thresholds. The EV threshold determines how selective the agent is when deciding whether to place a trade. Results in Table 3 reveal a clear trade-off between trade frequency and profitability. When a low EV threshold of 0.5% is used, the agent places a large number of trades but overall gets a negative return. This suggests that small apparent edges are often dominated by probability estimation error and market noise. Increasing the threshold to 1% yields modest positive returns, while a higher threshold of 2% results in fewer trades but substantially higher overall profitability.

Table 5: Performance of the Trading Agent under different EV Thresholds

| EV Threshold | Trades Placed | Total PnL($) | Final Bankroll($) |
|---|---|---|---|
| 0.5% | 19/29 | -166.36 | 2733.64 |
| 1.0% | 17/29 | +205.39 | 3105.39 |
| 2.0% | 14/29 | +408.22 | 3308.22 |

These findings indicate that filtering weak signals is more important than increasing trading frequency. By restricting trades to situations where the model's confidence exceeds market prices, the agent is better to use genuine mis-pricings rather than fluctuations caused by noises.

### *Key Takeaways*

Several key lessons emerge from the evaluation of the trading agent. First, probability accuracy alone is insufficiently for profitable trading. Even reasonably calibrated probability estimates can lead to losses if trades happen too aggressively or without filtering. Second, the Kelly criterion must be applied conservatively in practical settings. Fractional Kelly sizing and position caps are essential to mitigate risk when probability estimates are uncertain. Third, disciplined selectivity -choosing not to trade in many markets- is a critical component of long-term profitability.

## Conclusion

This paper presents a unified framework that integrates large language models with prediction-market-based decision making for probabilistic forecasting of NBA game outcomes. By leveraging timely unstructured textual information and aggregating heterogeneous LLM-based predictors, our approach moves beyond traditional data-driven models and produces calibrated win probabilities that are actionable in binary markets. The empirical results highlight the value of the probabilistic aggregation and disciplined betting strategies, such as fractional Kelly, in linking forecast quality to economic performance. Overall, our findings suggest that LLMs offer a promising direction for sports analytics in uncertain environments.

For the search agent, several extensions can be considered for future work. First, during the source aggregation process, the number of sources retrieved by each search agent can be varied. This would allow us to examine whether adjusting the number of sources from each agent leads to improved performance and to identify the best set of numbers. Through this analysis, we can also assess

whether the number of sources for each agent meaningfully reflects the importance of each factor represented by the agents. Second, we can investigate the relationship between the performance and the number of search agents used. In this study, we evaluated cases in which one or two agents were employed. Extending this to three or more agents would allow us to determine whether increasing the number of agents consistently improves performance, or whether there exists a saturation point which additional agents provide limited benefit. Finally, we can analyze whether the set of agents that performed best individually also forms the optimal joint set. For example, although the 'Extended Injury' agent had a lower Brier score than the 'Star Injury' agent, the set of 'Extended Injury' agent and 'Team Momentum' agent, which we used in this study, may have a higher Brier score than the set of 'Star Injury' agent and 'Team Momentum' agent. Since the complementary relationship between agents may affect the result of the sets of search agents, to systematically evaluate such interactions. In the future work, we can explore all possible combinations of agents and assess their joint effectiveness.

In terms of the forecasting agent, future work could integrate going the extra mile to add a larger variety of different personas, truly testing tens or hundreds of unique and creative roles to see which ones can breed the most accurate predictions and even have a chance at beating the market. In addition, future work could involve a bigger sample size, going from a couple of days' worth of games to months or even throughout an entire season to improve the validity of results. A larger sample size would allow clearer information on which agents are truly consistent and reliable over lots of data, and which agents show to have promising signs over smaller, favorable sets of games but fail to be accurate throughout the whole. Finally, future work could also explore a system of automated persona design, where techniques such as reinforcement learning would modify and create personas in groups so that the fit between these personas, when weighted and aggregated, would allow for the best possible prediction and lowest Brier score. These would automate and modify agent traits organically that also fit together well, in that each persona would be specialized and knowledgeable in different aspects of basketball and the NBA, so that the cumulative result after intentional weighting would breed favorable and possibly unprecedented results.

With respect to the trading agent, an important direction for future work would be the development of a language-model based trading agent that directly outputs both probability forecasts and stake recommendations given observed market prices and a fixed budget. This approach would integrate prediction and decision-making into a single agent. This will enable joint reasoning over uncertainty, expected value, and capital allocation. However, such a trading agent would also introduce additional risks, as protective measures would be required to prevent overconfidence and excessive risk-taking when probability estimates are not perfect.

## Acknowledgements

The author would like to express sincere gratitude to the professor and the teacher assistant for their guidance and support throughout the research process. Special thanks are to peers who provided valuable feedback and encouragement during the development of this paper.

## References

Brown, Tom et al. (2020). "Language models are few-shot learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Gumm, Jordan, Andrew Barrett, and Gongzhu Hu (2015). "A machine learning strategy for predicting march madness winners". In: IEEE, pp. 1–6. ISBN: 1-4799-8676-3.

Horvat, Tomislav et al. (2023). "A data-driven machine learning algorithm for predicting the outcomes of NBA games". In: *Symmetry* 15.4. Publisher: MDPI, p. 798. ISSN: 2073-8994.

Kim, Jun Woo, Mar Magnusen, and Seunghoon Jeong (2023). “March Madness prediction: Different machine learning approaches with non-box score statistics”. In: *Managerial and Decision Economics* 44.4. Publisher: Wiley Online Library, pp. 2223–2236. ISSN: 0143-6570.

Liu, Yiheng et al. (2023). “Summary of chatgpt-related research and perspective towards the future of large language models”. In: *Meta-radiology* 1.2. Publisher: Elsevier, p. 100017. ISSN: 2950-1628.

Ozkan, Ilker Ali (2020). “A novel basketball result prediction model using a concurrent neuro-fuzzy system”. In: *Applied Artificial Intelligence* 34.13. Publisher: Taylor & Francis, pp. 1038–1054. ISSN: 0883-9514.

Pathak, Neeraj and Hardik Wadhwa (2016). “Applications of modern classification techniques to predict the outcome of ODI cricket”. In: *Procedia Computer Science* 87. Publisher: Elsevier, pp. 55–60. ISSN: 1877-0509.

Schapire, Robert E (1990). “The strength of weak learnability”. In: *Machine learning* 5.2. Publisher: Springer, pp. 197– 227. ISSN: 0885-6125.

Tax, Niek and Yme Joustra (2015). “Predicting the Dutch football competition using public data: A machine learning approach”. In: *Transactions on knowledge and data engineering* 10.10, pp. 1–13.

Thabtah, Fadi, Li Zhang, and Neda Abdelhamid (2019). “NBA game result prediction using feature analysis and machine learning”. In: *Annals of Data Science* 6.1. Publisher: Springer, pp. 103–116. ISSN: 2198-5804.

Wei, Jason et al. (2022). “Emergent abilities of large language models”. In: arXiv preprint arXiv:2206.07682. Wysocki, Maciej (2025). “Sizing the Risk: Kelly, VIX, and Hybrid Approaches in Put-Writing on Index Options”. In: arXiv preprint arXiv:2508.16598.

Zhao, Kai, Chunjie Du, and Guangxin Tan (2023). “Enhancing basketball game outcome prediction through fused graph convolutional networks and random forest algorithm”. In: Entropy 25.5. Publisher: MDPI, p. 765. ISSN: 1099-4300.

## Authors

Ethan Jeon is a high school sophomore at Yongsan International School of Seoul in South Korea. He plans to apply to colleges around the New England area in the U.S. East Coast, and is considering majors in applied mathematics, data science, or finance.

Minseong (Leo) Sim is a student at Chadwick International School in South Korea. His academic interests include education, data analytics, economics, and applied artificial intelligence. He plans to pursue a university major related to business analytics, education, or kinesiology with a focus on sports.

Woohyun Kim is a high school student at Seoul Science High School. His research interests include applied statistics, machine learning, and large language models. He plans to further explore the intersection of data science and economic modeling at university.