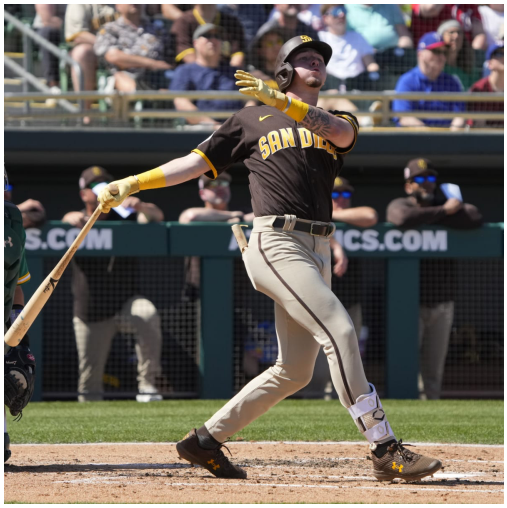
Rookie Excellence:
Using Data to Determine MLB's Top Freshmen

Peter D. DePaul III

09-15-2024



Contents

1 Introduction

The Rookie of the Year race has been somewhat of a huge debate throughout the year in the MLB. In the NL, between Paul Skenes, Jackson Chourio, and Jackson Merrill. In the AL, between Colton Cowser, Austin Wells, and Luis Gil. This has been a stellar production year for rookie batters as they have the 6th highest FanGraphs WAR (fWAR) since 1974 and the highest since 2002. This speaks for the quality of talent in this draft class and offers some explanation for why the Rookie of the Year races have been so close at this point in the season.

While looking at rookie stats recently I found myself questioning who should be Rookie of the Year in both leagues. I took it upon myself to give my best effort at determining who should be Rookie of the Year. I will make my best effort to provide objective statistical analysis, and eliminate any personal biases I have.

When I began this project on September 2nd, 2024 the odds were as seen below:

Click Here to View Rookie of the Year Odds Source

1.1 Rookie of the Year Odds (as of Sep. 2, 2024)

These odds were sourced from **FanDuel** according to the source.

Table 1: **American League Rookie of the Year Odds**

Player	Odds	Implied Probability
Colton Cowser	-230	69.70%
Austin Wells	+170	37.04%
Colton Keith	+1000	9.09%
Wilyer Abreu	+2800	3.45%

In the American League, Colton Cowser has been the odds favorite all year, but Austin Wells is not far behind. Cowser and Wells have both played well, but Cowser is an outfielder and there's some positional bias for them to win Rookie of the Year. Cowser shouldn't be number one, Wells has been better and the second problem is that Mason Miller is nowhere in sight. Despite this there has been some positive odds shift.

Table 2: **National League Rookie of the Year Odds**

Player	Odds	Implied Probability
Jackson Merrill	-900	90.00%
Paul Skenes	+550	15.38%
Jackson Chourio	+2800	3.45%
Shota Imanaga	+20000	0.50%
Masyn Winn	+20000	0.50%

1.2 Rookie of the Year Odds (as of Sep. 12, 2024)

Table 3: **American League Rookie of the Year Odds**

Player_Name	MGM	ESPN BET	Caesars	Draft Kings	FanDuel	Average Odds
Austin Wells	+150	+125	-250	+140	+170	+67
Luis Gil	+115	+165	+500	+200	+125	+221
Colton Cowser	+250	+225	+225	+230	+220	+230

Luckily for the AL the odds have shifted towards the better players. Austin Wells is now readily the favorite with Luis Gil surging up the odds due to his resounding performance in his last 2 starts. Cowser has been slumping this past month as a whole, and especially hard since August 31st. Wells has also been slumping a bit since August 31st but overall he's significantly outperformed Cowser in every way since August 1st.

Table 4: **National League Rookie of the Year Odds**

Player_Name	MGM	ESPN BET	Caesars	Draft Kings	FanDuel	Average_Odds
Jackson Merrill	-400	-350	-350	-360	-430	-378
Paul Skenes	+250	+250	+260	+240	+270	+254
Jackson Chourio	+2000	+1500	+1600	+2000	+3000	+2020

The NL odds have not shifted significantly at this point. While there has been some movement in favor of Skenes, Jackson Merrill remains the heavy favorite with an 80% implied probability to win.

2 Understanding the Data

The data utilized in the models was sourced from FanGraphs using their custom report function

2.1 Cleaning/Pre-processing

To not overkill this report with information I will skip over discussion of cleaning and pre-processing of the data. If you are interested please look no further than the Data Cleaning Python Notebook linked here.

2.2 Training/Testing Data

- From 1974-2024
  - Excluded the 1994, and 2020 seasons (since they were shortened)
  - Split by Season
- Included both MLB specific and Rookie specific data (separated)
- Split into
  - Relievers
  - Starters
  - Batters
- Data Conditions
  - Relievers: minimum of 40 Innings Pitched (IP)
  - Starters: minimum of 100 Innings Pitched (IP)
  - Batters: minimum of 300 Plate Appearances (PA)

2.3 Prediction Data

- From 2024 Season
- Same splits and conditions as above

2.4 Other Data

This analysis incorporates data sourced from Stathead, specifically focusing on spans of pitchers during the early stages of their careers and further discussion about their general performance.

3 Let's talk about the Past

3.1 Proportions of Each Position

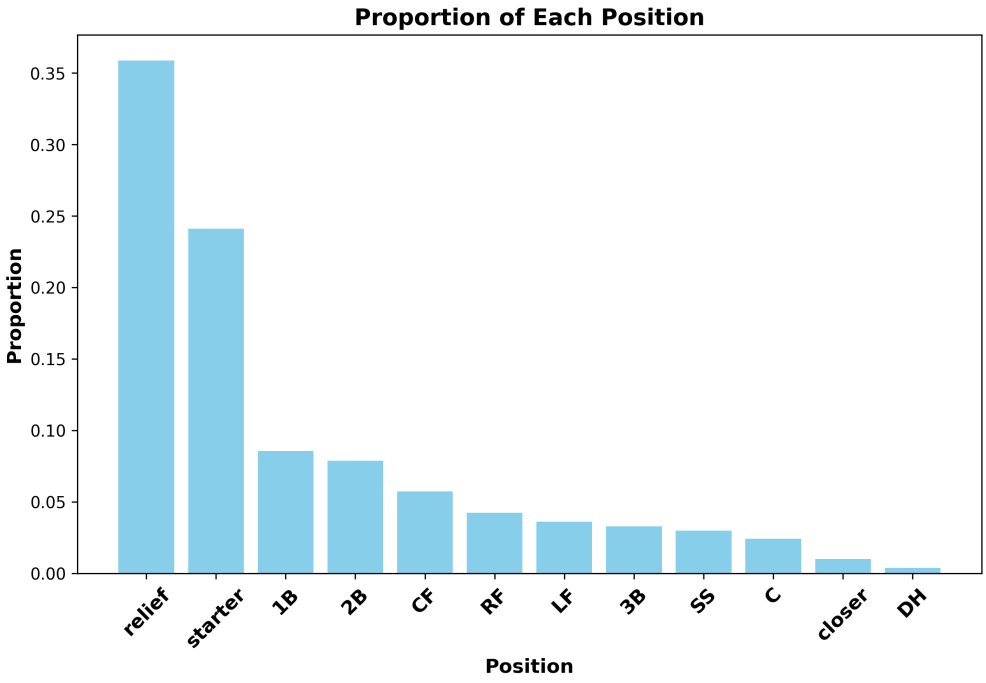


Figure 1: Proportions of Each Position

There are a ton of relievers and starters, and there are barely any Catchers, closers, or designated hitters (DH) within the data. This is important to keep in mind for further interpretation of the data presented after this point.

3.2 Proportions of Rookie of the Year (by Position)

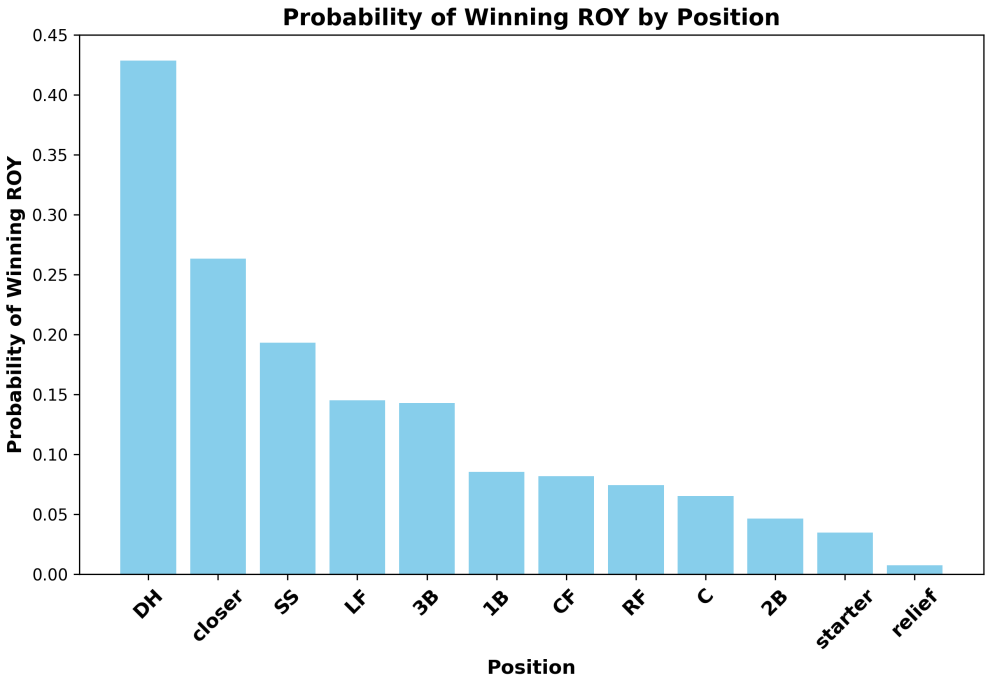


Figure 2: Proportions of Each Position

I want to focus on with the data in particular is the positions of the players. Batters are overwhelmingly favored every year over pitchers (they've won 73.5% of the awards since 1974). This is important to know because it likely suggests a positional bias, or perhaps hesitancy to vote for pitchers for Rookie of the Year. It's a weird bias, but it seems very real.

Keep in mind less than 1% of the training data was made up by DHs, and a lot of these guys are elite talents such as Yordan Alvarez, and Shohei Ohtani. I wanted to make this clear so people don't assume being a DH automatically wins Rookie of the Year, they're just really talented hitters. Who doesn't love to watch some great hitting?

4 Modeling Rookie of the Year

For both of my models I utilized the `lightgbm` interface in Python to create two binary classification models. For the models, I created individual models for each positional set of relievers, starters, and batters. Each of these positions had different predictors however the pitching models, did share significant overlap in columns overall.

In addition to this, I utilized `hyperopt` for hypertuning my models for the best parameters.

For my models I decided to create predictive models to model two distinct topics:

- Who will receive a vote this year? (`vote_getter`)
- Who will win Rookie of the Year? (`rookie_of_the_year`)

4.1 Who will receive votes?

Predicted Variable: `vote_getter`

Predicted Output Format: Probability (between 0 and 1)

Levels of `vote_getter`:

- 0 = Did NOT receive a vote
- 1 = Received at least 1 vote

4.1.1 Choosing Recipients across Positions

For the `vote_getter` model, which was trained using positional data sets, predictions were generated based on the predicted class probabilities. After grouping all players by league (AL or NL), the top 8 probabilities were selected from each league. While this method for selecting vote recipients may not be flawless, it seemed the best approach for comparing across positional groups.

4.1.2 Results

Table 5: **Relievers Vote Model Report**

category	precision	recall	F1_score	n
0	0.98	0.91	0.94	190
1	0.53	0.87	0.66	23
macro avg	0.75	0.89	0.80	213
weighted avg	0.93	0.90	0.91	213

Table 6: **Starters Vote Model Report**

category	precision	recall	F1_score	n
0	0.83	0.83	0.83	87
1	0.71	0.71	0.71	52
macro avg	0.77	0.77	0.77	139
weighted avg	0.78	0.78	0.78	139

Table 7: **Batters Vote Model Report**

category	precision	recall	F1_score	n
0	0.80	0.78	0.79	105
1	0.81	0.83	0.82	120
macro avg	0.81	0.81	0.81	225
weighted avg	0.81	0.81	0.81	225

We can see from the three tables above that the model for Relievers overall did the best (weighted F1-Score = 0.91), then the model for Batters (weighted F1-Score = 0.81), and then Starters (weighted F1-Score = 0.78).

I want to remark that the Batters and Starters had much more consistent performance despite lower weighted F1-Scores. Meanwhile the Relievers performance was heavily carried by the near perfect precision of the negative class (0s). In fact the precision for the positive class was rather abysmal, however I more concerned with overall performance.

4.1.3 2024 Rookie of the Year Votes

Table 8: **American League Rookie Vote Getter Predictions**

Name	Team	league	pos	vote_getter
Luis Gil	NYY	AL	starter	0.99
Mason Miller	OAK	AL	closer	0.98
Austin Wells	NYY	AL	C	0.96
Colton Cowser	BAL	AL	LF	0.96
Nolan Schanuel	LAA	AL	1B	0.94
Wilyer Abreu	BOS	AL	RF	0.92
Cade Smith	CLE	AL	relief	0.91
Spencer Horwitz	TOR	AL	2B	0.70

Table 9: **National League Rookie Vote Getter Predictions**

Name	Team	league	pos	vote_getter
Jackson Chourio	MIL	NL	RF	1.00
Paul Skenes	PIT	NL	starter	1.00
Jackson Merrill	SDP	NL	CF	1.00
Masyn Winn	STL	NL	SS	0.99
Shota Imanaga	CHC	NL	starter	0.98
Joey Ortiz	MIL	NL	3B	0.97
Michael Busch	CHC	NL	1B	0.97
Jacob Young	WSN	NL	CF	0.97

I don't have a lot to say about the predictions themselves. I believe most of these make sense, the NL Candidates as a whole have been stronger this year than the AL. Additionally all the important candidates (Merrill, Wells, Skenes, Cowser) were all predicted to receive votes, so I'm happy with the performance of the model

4.2 Who will win Rookie of the Year?

Predicted Variable: `rookie_of_the_year`

Predicted Output Format: Probability (between 0 and 1)

Levels of `rookie_of_the_year`:

- 0 = Did NOT win Rookie of the Year
- 1 = Did win Rookie of the Year

Utilized a similar selection process as the `vote_getter` model across positions, utilizing the probabilities.

4.2.1 Results

Table 10: **Relievers Vote Model Report**

category	precision	recall	F1_score	n
0	0.99	0.99	0.99	210
1	0.33	0.33	0.33	3
macro avg	0.66	0.66	0.66	213
weighted avg	0.98	0.98	0.98	213

Table 11: **Starters Vote Model Report**

category	precision	recall	F1_score	n
0	0.99	0.99	0.99	134
1	0.60	0.60	0.60	5
macro avg	0.79	0.79	0.79	139
weighted avg	0.97	0.97	0.97	139

Table 12: **Batters Vote Model Report**

category	precision	recall	F1_score	n
0	0.94	0.97	0.95	203
1	0.60	0.41	0.49	22
macro avg	0.77	0.69	0.72	225
weighted avg	0.91	0.92	0.91	225

Overall, none of the Rookie of the Year models performed exceptionally well, which was expected. It's relatively straightforward to identify who is unlikely to win the award; the real challenge lies in predicting the winner. The objective of this model was not to definitively select the Rookie of the Year, but rather to explore which factors the model considers most relevant for making such a prediction and to highlight the information that may be most important.

4.2.2 2024 Rookie of the Year Predictions

Table 13: **American League Rookie of the Year Predictions**

Name	Team	league	pos	rookie_of_the_year
Mason Miller	OAK	AL	closer	0.88
Cade Smith	CLE	AL	relief	0.83
Colton Cowser	BAL	AL	LF	0.36
Wilyer Abreu	BOS	AL	RF	0.04
Luis Gil	NYY	AL	starter	0.03
Spencer Horwitz	TOR	AL	2B	0.01
Ceddanne Rafaela	BOS	AL	SS	0.01
Brayan Rocchio	CLE	AL	SS	0.00

Mason Miller and Cade Smith are undoubtedly exceptional this season, which is likely the reason why they had the top 2 odds in the American League. In comparison top AL batters like Colton Cowser, have been rather mediocre in comparison to the pitchers.

Table 14: **National League Rookie of the Year Predictions**

Name	Team	league	pos	rookie_of_the_year
Jackson Merrill	SDP	NL	CF	0.99
Jackson Chourio	MIL	NL	RF	0.97
Masyn Winn	STL	NL	SS	0.15
Jacob Young	WSN	NL	CF	0.03
Bryan Hudson	MIL	NL	relief	0.01
Michael Busch	CHC	NL	1B	0.01
Shota Imanaga	CHC	NL	starter	0.00
Paul Skenes	PIT	NL	starter	0.00

There were no major surprises in the National League predictions, with the top three candidates being as expected. Merrill and Chourio have undoubtedly been the two best batters. However, the absence of Paul Skenes is intriguing, and it's unclear what factors are causing the model to downplay his odds. It could be due to the lower win rate for starting pitchers, but the reasoning isn't entirely clear.

5 Who has been the Best?

5.1 Relievers

Cade Smith (Cleveland Guardians)

Cade Smith remains the strongest candidate for Rookie of the Year among relief pitchers. Smith demonstrates incredible overall performance, with an impressive rookie rank of 1 and an MLB rank of 3, highlighting his excellence both among rookies and across the league. He excels in key pitching metrics such as ERA- (0.91), WHIP+ (0.86), and WAR (1.00 - best in the league), showing a well-rounded ability to dominate on the mound. His high IP rank (1.00 - most in the league) indicates that he consistently performs across many innings, a crucial factor for maintaining reliability throughout the season. His consistency across important

categories, including FIP- (1.00 - best in the league), HR/9+ (0.95), and K/BB+ (0.95), suggests that he effectively limits home runs, strikes out more batters than he walks, and maintains control over the game while on the mound. Cade Smith has truly shown the makings of a standout rookie pitcher with hall-of-fame potential, even though Cooperstown and Rookie of the Year voters tend to overlook relief pitchers.

Mason Miller (Oakland Athletics)

Mason Miller, with a rookie rank of 2 and an MLB rank of 5, also shows promise, particularly with his WAR/IP (1.00 - best in league) and K%+ (1.00 - best in league), but he trails Smith in several other key metrics, such as ERA- (0.77) and K/BB+ (0.91), keeping him behind Smith in overall performance. Miller's slightly lower IP rank (0.59 - due to him being a closer) indicates that he hasn't logged as many innings as Smith, which may impact his candidacy for Rookie of the Year. However one can never expect a closer to really log the same amount of innings as an ordinary reliever.

Interestingly, the three pitchers with legitimate chances at winning Rookie of the Year are all from the American League, and two of them—Smith and Gaddis—are from the same team, the Cleveland Guardians. Gaddis, with a rookie rank of 3 and an MLB rank of 9, also boasts strong numbers, including a reliever leading ERA- (1.00), but overall he's a step behind both Smith and Miller.

What's intriguing is that the model predicts Mason Miller to have the highest probability among relievers for winning Rookie of the Year, despite Smith being the top performer in most categories. This could be due to a potential bias in the model toward closing pitchers, with Miller benefiting from his positional role in the model's predictions. Additionally, despite Smith's superior overall performance, he has the lowest odds of winning Rookie of the Year among the three, which could be a reflection of this positional bias.

The Reliever Rookie of the Year

Overall, Cade Smith's consistent excellence throughout the season, particularly his performance across multiple key metrics, makes him the clear choice for Rookie of the Year among these pitchers.

5.2 Starters

Paul Skenes (Pittsburgh Pirates)

Paul Skenes is the clear standout candidate for Rookie of the Year among the four starting pitchers (predicted to receive votes). He ranks 1st in the rookie category and 4th in the MLB, showcasing his dominance among both rookies and across the league. His performance metrics are outstanding, with a perfect 1.00 rank in key categories such as ERA-, FIP-, HR/9+, WHIP+, WAR, K%+, and WPA. This demonstrates his all-around effectiveness in run prevention, strikeouts, and overall contribution to team success. His high WAR/IP rank (1.00) and strong control (K/BB+ rank of 1.00) further solidify his case as the top rookie, making him one of the most well-rounded pitchers in his rookie class.

Imanaga's overall performance is great. However, Skenes' rapid ascent to the MLB and dominance in nearly every category makes his rookie season even more impressive. At just 22 years old and within a year of being drafted, Skenes has already established himself as a dominant pitcher in the league. Meanwhile Imanaga has had 8 year of NPB experience prior to his MLB rookie season..

Overall, Paul Skenes' consistency and excellence across all major categories make him the clear choice for Rookie of the Year among these starting pitchers. His youth, rapid development, and nearly flawless performance metrics place him far ahead of his peers.

It's important to emphasize just how dominant Paul Skenes has been, even more so than commonly perceived. Using Stathead, I analyzed all rookie starting pitchers' first 20 games, focusing specifically on those with at least 100 innings pitched — about the same amount as Skenes' innings at the time of data collection.

Table 15: **Best Rookie Starting Pitcher (in first 20 games)**

Player	Team	IP	WPA/IP	ERA	BF/IP	BR/IP	Mean Rank
Paul Skenes	PIT	120.0	1	1	1	6	2.25
Fernando Valenzuela	LAD	156.0	3	5	3	3	3.50
Walker Buehler	LAD	117.2	6	7	5	1	4.75
Chris Sale	CHW	137.2	2	13	2	5	5.50
Mark Fidrych	DET	177.2	12	3	4	12	7.75

For the Stathead span data, the focus was on identifying the statistics that best capture pitcher dominance. The key metrics chosen for this analysis were `'WPA (Win Probability Added)'`, `'ERA'`, `'BF (batters faced)'`, and `'BR (baserunners allowed)'`. For cumulative stats, such as WPA, BF, and BR, these were normalized to a per innings pitched rate by dividing each by `IP`, given that many of the pitchers in the comparison, such as Hideo Nomo and Fernando Valenzuela, have significantly more innings pitched than Skenes.

As shown in the table above, Skenes has been exceptional in these categories. He ranks the highest in WPA, ERA, and BF, and is 6th in BR. This dominance across the board results in Skenes having the lowest average rank in these key metrics, outperforming Hall of Famer Fernando Valenzuela and future Hall of Famer Chris Sale. Skenes may very well have the most dominant start to a starter's career since 1974.

5.3 Batters

The standout candidate for Rookie of the Year in the NL would be Jackson Merrill (NL).

NL Candidate: Jackson Merrill (SDP)

Jackson Merrill stands out in the NL as the highest-rated rookie batter in the MLB. His WAR rank (1.00 - rookie leader) and WPA (1.00 - rookie leader) show that he consistently delivers both in terms of overall value and in contributing to team wins. Merrill also demonstrates a dominant offensive game, with high marks in wRC+ (0.95), SLG+ (0.95), and ISO+ (0.95). While his OBP+ (0.48) suggests room for improvement in getting on base, his performance across multiple offensive metrics and his mediocre defense (0.57) make him the most complete candidate in the NL.

AL Candidate: Wilyer Abreu (BOS)

Wilyer Abreu has been the top rookie batter in the AL, with strong ranks in SLG+ (1.00 - rookie leader), ISO+ (1.00 - rookie leader), and WAR (0.81). His balanced performance across key offensive categories highlights his ability to generate runs and power. Abreu's solid defense (0.62) further strengthens his case as the top rookie in the AL, making him a worthy candidate for the award.

In summary, Jackson Merrill (NL) and Wilyer Abreu (AL) would be the top choices for Rookie of the Year based on their overall performance and consistency.

6 Who Should be Rookie of the Year?

6.1 American League

Interestingly, the AL Rookie of the Year model underperformed in predicting the betting favorite. Mason Miller, Cade Smith, Colton Cowser, and Wilyer Abreu were the top candidates according to the model, yet Miller, Smith, and Abreu are not among the top four favorites on any betting platforms. This discrepancy supports the idea of positional bias existing, particularly given that Miller and Smith are a closer and a relief pitcher, respectively. However, to maintain relevance, the focus will remain on the odds previously discussed.

The AL Rookie of the Year race has become increasingly intriguing. Until earlier this week, Colton Cowser had been the frontrunner. However, Austin Wells has been on an impressive run since the All-Star break, gradually climbing the ranks and ultimately overtaking Cowser. As of yesterday, Luis Gil has surged from a distant third place to become the favorite on some betting platforms.

In my opinion, Austin Wells deserves to win the AL Rookie of the Year over Cowser. Wells has consistently outperformed Cowser in almost every aspect, but Cowser's advantage of having played 35-40 more games might make him a strong contender. His ability to perform well over a larger sample of games may play a significant role in the voting.

It's also worth noting that Austin Wells was not featured in the model's Rookie of the Year predictions. This omission again points toward potential positional bias. Catchers have rarely won the Rookie of the Year award (only three times since 1974), which may explain why Wells is not favored by the model.

6.2 National League

The model for predicting vote recipients performed exceptionally well. It accurately identified the top three candidates who are expected to receive votes, all of whom were given the highest probabilities of being selected.

The model for predicting the Rookie of the Year also did a solid job, correctly predicting that either Merrill or Chourio will win the NL Rookie of the Year award. I'm gonna imagine at this point that Merrill will win the award in real life, but he shouldn't.

Paul Skenes is having one of the top 10 rookie pitching campaigns since 1974, while Jackson Merrill's rookie hitting season might rank in the top 50. The odds in favor of Merrill likely stem from a positional bias that exists in Rookie of the Year voting, where hitters are often favored over pitchers.

Paul Skenes, based on his performance, should be the NL Rookie of the Year.