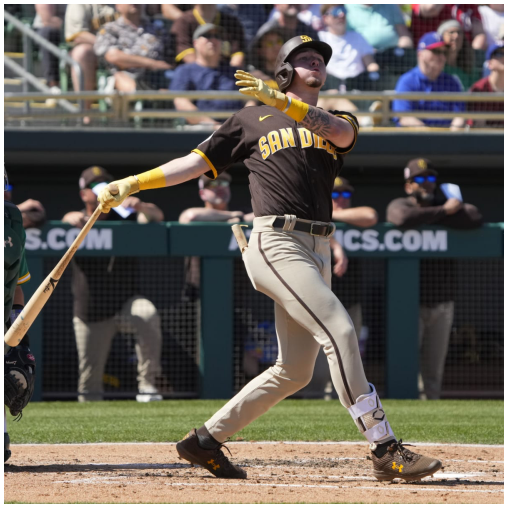
Rookie Excellence:
Using Data to Determine MLB's Top Freshmen

Peter D. DePaul III

09-17-2024



Contents

1 Introduction

The Rookie of the Year race has raised significant debate throughout the MLB, with Paul Skenes, Jackson Chourio, and Jackson Merrill leading in the NL, and Colton Cowser, Austin Wells, and Luis Gil in the AL. 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 aim to provide objective statistical analysis and eliminate any personal bias.

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. 17, 2024)

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

Player	MGM	ESPN BET	Caesars	Draft Kings	FanDuel	Average Odds
Luis Gil	+110	+135	+100	+110	+120	+115
Austin Wells	+150	+145	+125	+160	+175	+151
Colton Cowser	+225	+225	+220	+230	+210	+222

Luckily for the AL the odds have shifted towards the better players. Luis Gil is now readily the favorite 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	MGM	ESPN BET	Caesars	Draft Kings	FanDuel	Average_Odds
Jackson Merrill	-500	-425	-625	-600	-340	-498
Paul Skenes	+225	+260	+220	+280	+220	+241
Jackson Chourio	+3000	+3300	+2200	+3000	+4000	+3100

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 83.28% 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 avoid overwhelming the report with excessive detail, I have omitted the full data cleaning and pre-processing discussion. 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

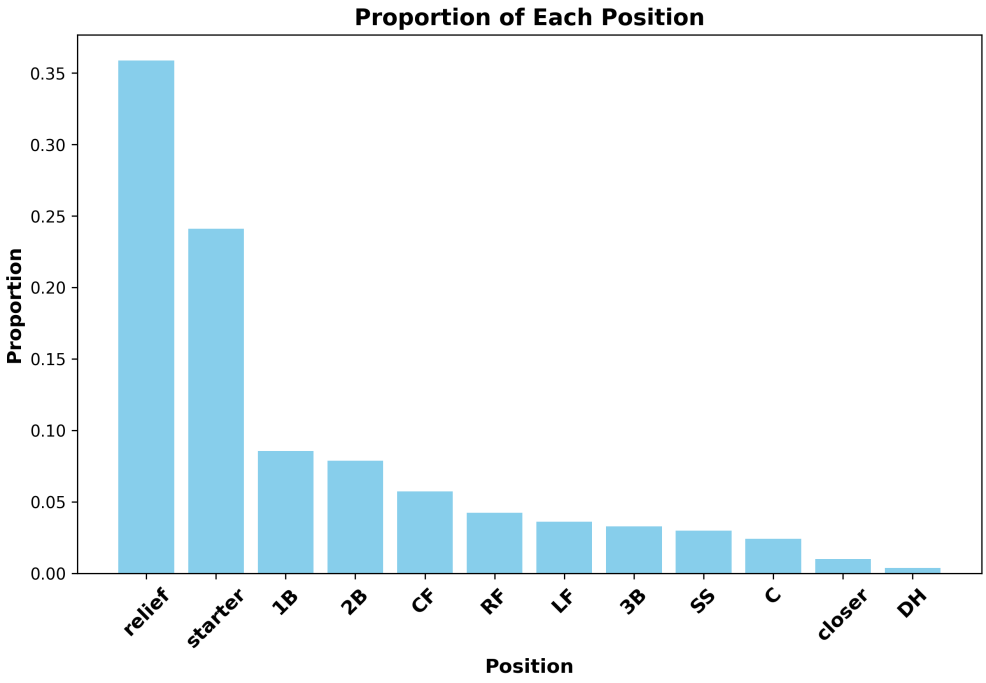


Figure 1: Proportions of Each Position

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. It is important to note the positions because it deeply affects the models ability to learn. While `lightgbm` is an ensemble method that, along with hypertuning, can perform well with limited data, having more observations generally improves performance by exposing the model to more variability. This implies these models might be sensitive to the less frequented positions.

3.2 Proportions of Rookie of the Year (by Position)

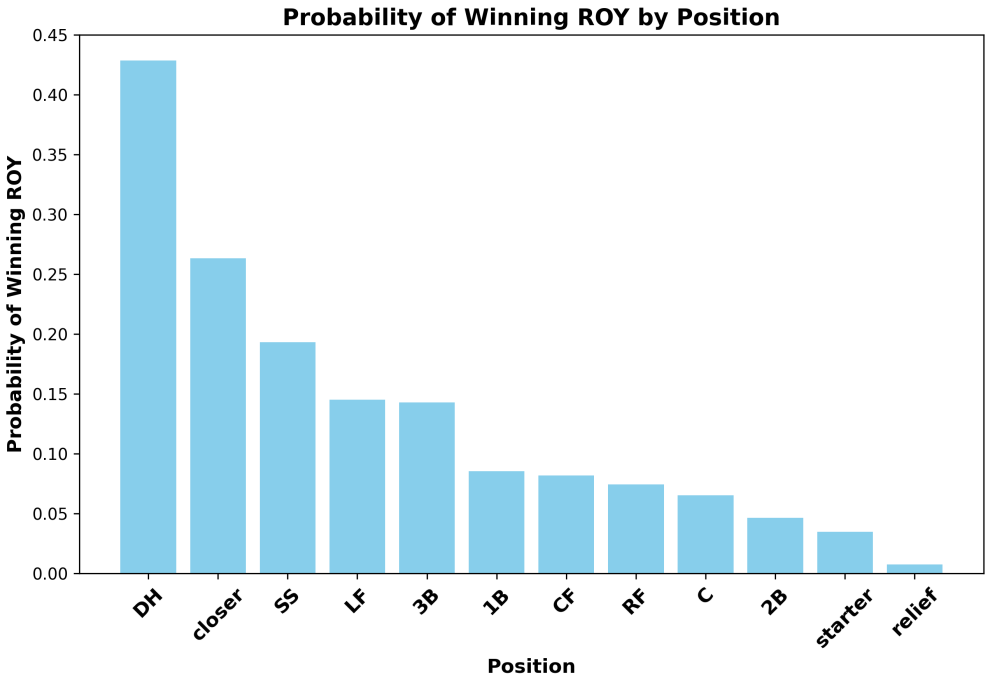


Figure 2: Proportions of Each Position

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. I developed separate models for each positional group: 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 used `hyperopt` for hyper-parameter tuning the models.

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	0.94	0.96	190
1	0.61	0.83	0.70	23
macro avg	0.80	0.88	0.83	213
weighted avg	0.94	0.92	0.93	213

Table 6: **Starters Vote Model Report**

category	precision	recall	F1_score	n
0	0.85	0.77	0.81	87
1	0.67	0.77	0.71	52
macro avg	0.76	0.77	0.76	139
weighted avg	0.78	0.77	0.77	139

Table 7: **Batters Vote Model Report**

category	precision	recall	F1_score	n
0	0.83	0.73	0.78	105
1	0.79	0.87	0.83	120
macro avg	0.81	0.80	0.80	225
weighted avg	0.81	0.80	0.80	225

We can see from the three tables above that the model for Relievers overall did the best (macro F1-Score = 0.83), then the model for Batters (macro F1-Score = 0.80), and then Starters (macro F1-Score = 0.76).

I want to remark that the Batters and Starters had much more consistent performance despite lower macro 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
Austin Wells	NYY	AL	C	1.00
Colton Cowser	BAL	AL	LF	0.99
Cade Smith	CLE	AL	relief	0.98
Mason Miller	OAK	AL	closer	0.98
Nolan Schanuel	LAA	AL	1B	0.97
Luis Gil	NYY	AL	starter	0.95
Wilyer Abreu	BOS	AL	RF	0.95
Colt Keith	DET	AL	2B	0.95

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

Name	Team	league	pos	vote_getter
Jackson Merrill	SDP	NL	CF	1.00
Michael Busch	CHC	NL	1B	1.00
Jackson Chourio	MIL	NL	RF	1.00
Joey Ortiz	MIL	NL	3B	0.98
Masyn Winn	STL	NL	SS	0.95
Shota Imanaga	CHC	NL	starter	0.95
Paul Skenes	PIT	NL	starter	0.95
Jacob Young	WSN	NL	CF	0.74

Most of these predictions are reasonable. 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	1.00	1.00	210
1	1.00	0.33	0.50	3
macro avg	1.00	0.67	0.75	213
weighted avg	0.99	0.99	0.99	213

Table 11: **Starters Vote Model Report**

category	precision	recall	F1_score	n
0	0.99	0.99	0.99	134
1	0.75	0.60	0.67	5
macro avg	0.87	0.80	0.83	139
weighted avg	0.98	0.98	0.98	139

Table 12: **Batters Vote Model Report**

category	precision	recall	F1_score	n
0	0.96	0.92	0.94	203
1	0.48	0.68	0.57	22
macro avg	0.72	0.80	0.75	225
weighted avg	0.92	0.90	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.

Aside from the difficulty in predicting the best rookie, another reason for the lackluster performance could be the lack of data. We have a total of 98 Rookie of the Year winners in our data. Only 10 of those were relievers, 16 who were starters, and 72 who were batters. This means there's a relative lack of positive data for the pitchers overall.

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
Colton Cowser	BAL	AL	LF	0.67
Austin Wells	NYY	AL	C	0.64
Mason Miller	OAK	AL	closer	0.38
Wyatt Langford	TEX	AL	LF	0.19
Colt Keith	DET	AL	2B	0.15
Nolan Schanuel	LAA	AL	1B	0.04
Wilyer Abreu	BOS	AL	RF	0.04
Cade Smith	CLE	AL	relief	0.02

Unsurprisingly, Austin Wells and Colton Cowser are the top 2 AL Candidates for Rookie of the Year. To my surprise Mason Miller, who's been exceptional this year, is the third highest candidate. His role as a closer likely helps his case for Rookie of the Year.

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

Name	Team	league	pos	rookie_of_the_year
Jackson Merrill	SDP	NL	CF	0.91
Jackson Chourio	MIL	NL	RF	0.72
Joey Ortiz	MIL	NL	3B	0.59
Masyn Winn	STL	NL	SS	0.49
Paul Skenes	PIT	NL	starter	0.48
Michael Busch	CHC	NL	1B	0.31
Jacob Young	WSN	NL	CF	0.07
Pete Crow-Armstrong	CHC	NL	CF	0.01

There were no major surprises in the National League predictions, with the top two candidates being as expected. Merrill and Chourio have undoubtedly been the two best batters. However, the low ranking 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. He excels in key pitching metrics such as `ERA-` (0.91), `WHIP+` (0.86), and `WAR` (1.00 - rookie leader), showing a well-rounded ability to dominate on the mound. His high `IP` rank (1.00 - rookie leader) 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 - rookie leader), `HR/9+` (1.00 - rookie leader), 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, also shows promise, particularly with his `WAR/IP` (1.00 - rookie leader), `K%+` (1.00 - rookie leader), and rookie leading SD (shutdowns), 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.

What's intriguing is that the model predicts Mason Miller to have the highest probability among relievers for winning Rookie of the Year, and it's not quite clear why.

Table 15: **Best Rookie Relievers (Closer and Relief)**

Name	ERA-	FIP-	WHIP+	SD-int	WAR/IP	Average Rank
Jonathan Papelbon	2	15	2	10	1	6.0
Dellin Betances	13	6	8	7	14	9.6
Takashi Saito	42	5	16	15	4	16.4
Troy Percival	20	30	5	14	15	16.8
Mark Eichhorn	21	28	32	23	17	24.2
Ken Giles	11	1	11	101	5	25.8
Cade Smith	57	2	58	20	9	29.2
Joey Devine	1	11	7	117	13	29.8
Emmanuel Clase	9	17	60	45	20	30.2
Greg Holland	29	27	41	33	25	31.0

The data in the table above came from the training/testing data from FanGraphs. I combined all relievers from 1974 to 2024, and focused on `ERA-`, `FIP-`, `WHIP+`, `WAR/IP`, and `SD-int`. `SD-int` was calculated by

$$\text{SD-int} = \text{Number of Shutdowns} * \frac{\text{Number of Shutdowns}}{\text{Number of Shutdowns} + \text{Number of Meltdowns}}$$

The `Average Rank` was calculated by taking the average of the aforementioned 5 statistics. We can see that Cade Smith is having a historically dominant seasons among all relievers. He has the 7th best season among Rookie Relievers according to this ranking. This is incredible, and puts him in competition with the most dominant rookie relievers such as 2006 Jonathan Papelbon, 2014 Dellin Betances, and 2006 Takashi Saito. He even outpaces fellow Guardians Reliever Emmanuel Clase who has been the best Closer in the MLB since 2019.

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 better 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 0.86) further solidify his case as the top rookie, making him one of the most well-rounded pitchers in his rookie class.

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.

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 21 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 16: **Best Rookie Starting Pitcher (in first 20 games)**

Player	Team	IP	WPA/IP	ERA	BF/IP	BR/IP	Average Rank
Paul Skenes	PIT	126.0	1	2	1	1	1.25
Mark Fidrych	DET	168.2	6	1	2	3	3.00
Jose Fernandez	MIA	127.2	10	4	3	2	4.75
Hideo Nomo	LAD	153.0	4	8	23	9	11.00
Bill Laskey	SFG	133.2	11	5	15	13	11.00

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/IP`, `BF/IP`, and `BR/IP` while ranking second in ERA. 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.

Table 17: **Best Rookie Starting Pitcher Seasons**

Name	ERA-	FIP-	WHIP+	WAR/IP	Average Rank
Paul Skenes	2	5	9	6	5.50
Roy Oswalt	10	6	6	5	6.75
Cal Eldred	1	16	2	15	8.50
Dave Righetti	5	3	24	3	8.75
Spencer Strider	23	2	12	1	9.50

The data in the table above came from the training/testing data from FanGraphs. I combined all starters from 1974 to 2024, and focused on `ERA-`, `FIP-`, `WHIP+`, `WAR/IP` then ranked all the starters. The `Average Rank` is the mean of these 5 variables. From this data we can see Skenes has been dominant. Since 1974 he has the second best `ERA-`, the fifth best `FIP-`, and is top 10 in both `WHIP+` and `WAR/IP`. Only one other pitcher in this data has done the same, 2001 NL Rookie of the Year runner-up (to Albert Pujols), Roy Oswalt. All baseball fans should recognize that we are currently witnessing history among rookies with this performance.

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+` (1.00 - rookie leader), and `ISO+` (0.95). While his `OBP+` (0.59) suggests room for improvement in getting on base, his performance across multiple offensive metrics and his mediocre `Def` (0.55) make him the most complete candidate in the NL.

AL Candidate: Austin Wells (NYY)

Austin Wells has been the top rookie batter in the AL, with strong ranks in `WAR/G` (1.00 - rookie leader), and `WAR` (0.91). Aside from this his high `OBP+` (0.91) highlights his consistency to reach base offensively. His stellar `Def` (1.00 - rookie leader) 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 Austin Wells (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 performed well in predicting the betting favorite. Colton Cowser, Austin Wells, Mason Miller and Wyatt Langford were the top candidates according to the model; Miller, and Langford are not among the top four favorites on any betting platforms.

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. Luis Gil has surged from a distant third place to become the overall betting favorite.

In my opinion, Cade Smith deserves to win the AL Rookie of the Year over Cowser. Cade Smith is having one of the best seasons since 1974 by a rookie reliever. He has been dominant in `FIP-` suggesting he can

control the game without need of his fielders behind him, and he has been excellent at shutting down the other teams hitters in his appearances.

It's also worth noting that Austin Wells was not featured in the model's Rookie of the Year predictions. Catchers have rarely won the Rookie of the Year award (only three times since 1974), which may explain why Wells is not favored (since Catchers are usually defensive specialists and not much more).

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. At this point it's likely that Merrill will win the award in real life, but he shouldn't.

Paul Skenes is having a top 5 rookie pitching campaigns since 1974. 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.

7 Conclusion

The Rookie of the Year race in 2024 has proven to be one of the most competitive and debated in recent memory, with standout performances across both leagues. We have identified key contenders for this prestigious award. While players like Jackson Merrill and Colton Cowser have dominated the conversation, our statistical approach has revealed important nuances that might not be reflected in public perception or betting odds.

In the American League, the competition has tightened considerably, with Luis Gil surging late in the season. However, based on both performance metrics and positional impact, Cade Smith has consistently outperformed his rivals and is a deserving candidate for the award. Wells' ability to contribute as a reliever, a position historically underrepresented in Rookie of the Year voting, sets him apart as the strongest contender.

In the National League, Jackson Merrill remains the betting favorite, but our analysis suggests that Paul Skenes should be the frontrunner. Skenes' performance as a rookie pitcher has been historically dominant, placing him among the top rookie campaigns since 1974. Despite the typical positional bias that favors batters in Rookie of the Year voting, Skenes' all-around excellence and ability to impact the game make a compelling case for his selection.

Ultimately, the results of this analysis underscore the importance of objective data in evaluating rookie performances. While public opinion and positional bias may influence the final outcome, it is clear from the numbers that both Smith and Skenes have demonstrated the highest level of excellence in their respective leagues and are most deserving of the 2024 Rookie of the Year awards.