Improving NHL Draft Outcome Predictions using Scouting Reports

Hubert Luo. [1]

Abstract

We leverage Large Language Models (LLMs) to extract information from scouting report texts and improve predictions of National Hockey League (NHL) draft outcomes. In parallel, we derive statistical features based on a player's on-ice performance leading up to the draft. These two datasets are then combined using ensemble machine learning models. We find that both on-ice statistics and scouting reports have predictive value, however combining them leads to the strongest results.

1. Introduction

Hockey has long emphasized the importance of the eye test: evaluating players using visual observations of hockey games. However, the eye test is prone to biases (Deaner et al., 2013), anecdotal evidence, and the reality that machines can incorporate information from a much larger sample size of games than a human can.

On the other hand, analytical models lack situational context, nuance, and information that only humans can collect from discussions with other humans (Wolfson et al., 2011). Therefore, this paper's objective is to leverage the strengths of both approaches by extracting information gleaned from the eye test and combining it with quantitative analyses.

Leading up to the NHL draft, scouts attend games worldwide to identify top draft-eligible players, write scouting reports for them, and compare prospects with their peers through rankings. These scouting report texts are a rich source of qualitative information about players, encompassing information about their playing style, strengths/weaknesses, and a scout's feedback on how well a prospect's ability will translate to the professional game.

For example, scout Scott Wheeler writes about Connor Bedard, who ended up being drafted first overall in the 2023 NHL Draft:

> "[Bedard]'s got unbelievably quick hands and the loose grip that all great handlers have. He's got high-end speed with his galloping crossover strides and strong acceleration from a standstill, which help him carry the puck up the ice, create one-on-one off the rush, or join in transition as the trailer whenever he has to play catchup." (Wheeler, 2023a)

Scouts spend a vast amount of time canvassing thousands of potential prospects, providing unique information and learning important context about a player's on-ice performance that is only observable by a person or by interviewing coaches, prospects, and their teammates. This off-ice context is an important factor in a player's performance that has typically only been available to humans.

Incorporating these scouting reports into existing hockey analytics processes has been particularly challenging given the need to translate these texts into usable and comprehensive quantitative features. This is an area where LLMs have demonstrated strong potential over the last year, for example in engineering features for financial stock predictions (Lopez-Lira and Tang, 2023). This paper will be one of the first to apply LLMs to hockey, allowing us to integrate off-ice context into analytical models using scouting reports.

We end up with two different datasets, one featuring a player's on-ice statistics prior to the draft, and another with features engineered from a player's scouting reports. Base models are trained independently on each dataset in parallel, and then combined using a stacking ensemble learning methodology. Stacking was chosen to mitigate spurious decision-making and facilitate better comparison between base models as outlined in section 3.10.

In this paper, we will demonstrate that LLMs are useful in extracting information from scouting reports and that combining it with on-ice statistics will improve our predictions of post-draft outcomes.

2. Previous Work

Recent years have seen an uptick in the adaptation of sports analytics, ranging from spin rates in baseball to fourth-down in-game strategies for football. Quantitative applications in hockey has also likewise advanced in the last few years, going from puck possession and shot-share metrics like Corsi

[1]Department of Computer Science, University of Texas at Austin and Data Analytics Group, Lazard.

to more recent focuses on expected goals, biomechanics, and player tracking (Nandakumar and Jensen, 2018).

Draft-eligible players are spread across multiple leagues, countries, and continents. These leagues vary tremendously in strength, for example a point-per-game player in the Ontario Hockey League (OHL) is more impressive than a point-per-game player in the German junior hockey league given the OHL has a much higher quality of competition in comparison.

Therefore, it is important to adjust player performance relative to which league they are playing in - this was the motivation and findings from one of the first approaches to applying a quantitative perspective to the NHL draft (Desjardins, 2005). These NHL equivalences (NHLe) translate a player's performance from one league to what they would have reasonably achieved in the NHL, allowing a straightforward comparison between player performances.

Network graphs can then be leveraged to translate player transitions between leagues graphically with nodes representing each league and edges representing moving between leagues (Turtoro, 2020). When creating an equivalency factor from the American Hockey League (AHL) to NHL, NHLe looks solely at the edge between these two nodes. On the other hand, Network NHLe (NNHLe) considers additional paths to traverse the graph between these nodes, i.e., AHL to Kontinental Hockey League (KHL) to NHL. This provides a more robust estimate of inter-league comparison.

Different metrics have been used to define a successful draft pick. Previous work used a Poisson general additive model to predict the number of NHL games a player picked in their first seven seasons after they were drafted (Schuckers, 2016). Tree-based approaches have also been leveraged successfully to look at the probability a NHL draftee plays a game in the NHL (Liu et al., 2019) and AHL even-strength point production (Seppa et al., 2017).

Seppa et al showed that scouting report texts can create useful signals for predicting NHL draft outcomes. They focused on players in the Canadian major junior leagues (CHL) and engineered features using n-grams to classify players by their attributes. Our work builds on this previous work by broadening the scope to include all drafted players, applying LLMs to engineer features, and using deep learning architectures instead of tree-based methods. See sections 3.3 and 3.10 for more details.

Similar work has also previously been done in other sports. For example, in the National Football League (NFL), non-parametric regression modelling was used to measure the value of draft picks based on their post-draft playing performance in the NFL (Schuckers, 2011a). In the National Basketball Association (NBA), important factors for predicting draft outcomes varied by the dependent variable involved. Specifically, variables which were important for predicting whether a player would make it to the NBA varied compared to those which were important for predicting their performance once in the league (Berri et al., 2011).

This large collection of previous research in not only hockey but also other sports demonstrates the potential of quantitative analytical approaches to draft picks and evaluating post-draft player performance. On the other hand, individual scouts and the eye test have also proven valuable, for example in a study looking at NFL decisions drafting quarterbacks (Wolfson et al., 2011). Therefore, it is important to combine lessons learned from both a visual and analytical perspective.

3. Method

3.1. Dependent Variable

The motivation behind the dependent variable y was to capture the outcome of being a regular NHL player. We used a cumulative sum threshold to gauge whether a player was on track to play at least 200 NHL games within the first eight seasons after they were drafted:

$$t = \text{Draft Year} \in \mathbf{Z}_{\leq 2020} \tag{1}$$

$$x = \text{\# Games Played in NHL (GP)} \tag{2}$$

$$z = \text{Cumulative GP threshold} \tag{3}$$

$$= \begin{cases} \frac{82*(2022-t)}{3} & \text{if } t \in [2016, 2020] \\ 200 & \text{if } t \leq 2015 \end{cases} \tag{4}$$

$$y = \text{exceeds cumulative GP threshold} \tag{5}$$

$$= \mathbb{1}(x \geq z) \tag{6}$$

This threshold decreases the later a player was drafted: 200 games for those drafted in 2015, 164 for 2016, 136 for 2017, and so on. We used 200 NHL games within eight years as the starting point as teams typically have no control over players they drafted after this timeline and it is rare for players to make it to the NHL afterwards.

To evaluate the validity of the rolling threshold in our dependent variable, we first defined a true NHL regular outcome as someone who played over 200 NHL games before the end of their eighth season. Precision is then defined as the percentage of those on track after delta years who ended up actually being NHL regulars. Recall is the percentage of eventual NHL regulars who looked on track after delta years.

Table 1 shows precision and recall for players drafted in 2015. Our rolling threshold is highly precise as few players initially on track to become NHL regulars are not by their eighth season. False positives include players like Oliver

Table 1. Recall and precision for players drafted in 2015. Precision is % of those on track after delta years who ended up actually being NHL regulars. Recall is % eventual NHL regulars who looked on track after delta years. Note delta of 3 means "3 years after", i.e., by end of 2018-19 season.

DELTA	PRECISION	RECALL
3	0.949	0.787
4	0.935	0.796
5	0.956	0.800
6	0.944	0.909
7	0.929	0.981
8	1.000	1.000

Kylington, who faced significant injury troubles however now looks on pace to reach the 200-game threshold.

Our recall is generally lower as there were more instances of players who ended up playing over 200 NHL games by their eighth year who did not make the NHL immediately especially within the first five years of being drafted. These include late bloomers, such as Jonas Siegenthaler, as well as players who spend their early years in other leagues. For example, Niko Mikkola stayed in Finland after being drafted and only moved to North America ahead of the 2018 season.

Overall, the high precision and recall observed on these players validate the rolling threshold used as our dependent variable. This resulted in a classification problem where players were assigned to two groups, those who were on track to become a NHL regular and those not on track.

3.2. Scouting Reports: Data Sources

Scouting reports were collected from two primary publicly-available sources: *The Athletic* and *ESPN*. These sources were selected because they have contributions from well-known and reputable scouts with established track records in the public sphere.

These reports present a differentiated perspective through viewings and interviewing their coaches, teammates, and players themselves. They encompass not only a player's on-ice impact, but also the context around these players. For example, a player may have a poor relationship with their coach, they may be going through a challenging living situation off-ice, or they may have recently moved to a new country and are learning a second language (Wheeler, 2023b). This context is hidden in traditional statistical approaches, however has significant on-ice impact.

As these reports are public-facing, they are typically released throughout the season in the form of articles featuring rankings of the top 32-100 prospects for the NHL draft. We focus on the final articles published prior to the draft to provide the most up-to-date information and ensure we compare players at similar points in time.

Table 2. Scouting reports by year. Note not all scouts publish lists of their top 100 prospects - in some cases, they publish lists of their top 32 or 64 prospects, corresponding to the first or first two rounds of the draft.

DRAFT YEAR	# SCOUTS	# UNIQUE PLAYERS	AVERAGE REPORT LENGTH
2015	1	99	718.5
2016	1	100	864.9
2017	1	100	871.3
2018	3	94	988.0
2019	3	124	839.7
2020	2	138	1,125.4
2021	2	164	872.8
2022	2	142	967.5
2023	2	151	1,047.4

We collected this data across eight years, starting with the 2015 draft class. This time threshold was chosen because the specific scouts who are writing reports for prospects today primarily started after 2015. Earlier reports were not easily accessible and/or written by others who differ significantly in evaluation style, report length, and how these reports were collected. Prior to 2015, few news outlets employed writers focused on scouting year-round so these scouts were primarily hired recently due to increased public demand for in-depth reporting.

Table 2 demonstrates the evolution of collected scouting reports by year, which shows that average report length has generally increased each year. There was a spike in average report length in 2020, likely due to the fact scouts had more time to write these reports without significant travel obligations due to the pandemic.

Another trend we can observe from table 2 is that there was generally less consensus among who scouts place in their top prospect lists over time. The pandemic's impact was again clear in 2021 as some prospects barely played that year due to many leagues shutting down in 2020 and 2021. Therefore, scouts had wider divergence in players they ranked in their respective lists of top prospects.

The number of prospects drafted has varied in length given NHL expansion over time, ranging from 210 to 224 prospects depending on the specific year. Therefore, approximately half of these drafted prospects have a scouting report in our database each year.

We also collected information on where scouts ranked these players among all available players that year. These rankings were an implicit summarization of the text and were especially meaningful given they were fixed and a direct

representation of a scout's feedback on a player relative to their peers.

However, only using these rankings would hinder our ability to compare across draft years as not all years are equal. For example, Owen Power and Connor Bedard may have both been consensus first overall in their respective draft rankings however there are clear differences in their playing abilities.

Rankings also assume that the difference between players is equidistant, i.e., the difference between players ranked first and tenth is the same as that between players ranked 31st and 40th. However, past research (Schuckers, 2011b) and empirical evidence has shown dropoff in abilities between ranks is much steeper at the beginning of the draft.

Furthermore, a player's rank may be subject to a scout's bias, i.e., a player may be ranked highly even if their report is not as positive about them due to that scout's biases about a player's playing style, league, or performance in certain games. This means the rank they assign a player may not be consistent with their true thoughts in the text.

Therefore, we need to supplement the ranks that scouts provided with additional features more apt at comparing between years, capturing more granular distinctions between players, and a more reflective representation of the written scouting report.

3.3. Scouting Reports: Feature Engineering

As motivated in section 3.2, we then engineered a few additional features based on scouting report texts. The goal of these features was to capture information and context missing from on-ice statistics.

We first processed scouting report texts to be more suitable for feature engineering by using unidecode to standardize text, which was especially important given high prevalence of surnames involving special characters. We removed casing by converting all text to lower-case.

Lemmatization is the process of only keeping a word's base form, i.e., "write" instead of "written" (Manning et al., 2008). We decided not to use lemmatization so the LLM has a better understanding of the text given the entire word rather than just its base form. We also explored removing stopwords like "and" or "the" however again decided to retain them to keep additional context for the LLM.

This processing resulted in 1,682 reports on 1,074 unique players over nine seasons. We then leveraged a LLM to extract information from these textual data sources to create three new features. These were a player's likelihood of making it to the NHL, represented by $\alpha_m \in \mathbb{Z}_{[0,100]}$, a sentence describing that player's strengths, and a sentence describing their weaknesses.

A likelihood score of zero in α_m means the model believes that player has no chance of making it to the NHL. On the other hand, a likelihood score of 100 mean the model has strong conviction that a prospect will be drafted into the NHL.

These features were derived using OpenAI's ChatGPT model, which uses a transformer architecture with self-attention layers to better capture contextual relevance (Vaswani et al., 2023). OpenAI then trained this model on vast amounts of web-scraped data in an unsupervised learning approach before fine-tuning it with human feedback (Stiennon et al., 2020).

Previous research has shown that prompts perform best when provided precise instructions to act in a specific role (Chen et al., 2023). This served as the basis for the following prompt:

```
nhl_template = """You are an ice hockey
    expert. You are given this player's
    scouting report: {report_text}.
Answer in this format:
"SCORE: integer between 0 and 100 for
    whether he will make it to the NHL (1
    is impossible, 99 is certain)
STRENGTHS: one sentence on his strengths,
    based only on the provided report
WEAKNESS: one sentence on his weaknesses,
    based only on the provided report
"
"""
```

We first used this prompt in a zero-shot approach on OpenAI's gpt-3.5-turbo-1106 model. In order to evaluate our prompt's performance, we selected 70 players at random without replacement from our training set.

Based on this subset of players, we examined those with a large discrepancy between the model's score α_m and their actual NHL regular outcome y. For example, Vasili Podkolzin had a high α_m due to a scouting report that was mostly stellar other than raising concerns about his inefficient skating technique. This flaw in his scouting report meant his α_m score was initially too high, an observation that would be applicable to other players with similar skating issues.

We then replicated this process with a few additional players, specifically Urho Vaakanainen, who the model was too optimistic about given his weak draft season, and Adam Fox, who the model was too pessimistic about given his superstar offensive upside. This moved us to a few-shot prompting approach. See appendix 8.2 for the exact adjustments made to these players.

Table 3 evaluates our prompting approaches by showing

Table 3. Comparing correlations between likelihood of making the NHL α_m generated by different prompting approaches vs. dependent variable of being a NHL regular y on training set sample.

APPROACH	CORRELATION
ZERO-SHOT	0.538
ONE-SHOT	0.547
FEW-SHOT	0.601

Table 4. LLM consistency checks applying same prompt to same sample of players in training dataset over 10 iterations. Strength and weaknesses consistencies measured by overlap in words generated between LLM responses.

METRIC	CONSISTENCY
LIKELIHOOD OF MAKING THE NHL α_m	0.994
STRENGTHS	0.946
WEAKNESSES	0.892

correlation between α_m and y for the random sample of players in our training dataset. We observed higher correlations between α_m and y in a few-shot learning approach compared to one-shot and zero-shot approaches.

Throughout this section, we used a temperature of zero to reduce variation in responses given the same prompt. We tested the model's consistency by applying the same few-shot prompt to our random training sample over 10 iterations. This resulted in the model being given the same player scouting report and prompt 10 times.

Consistency was measured using the following formula:

$$N = \text{\# players in training sample} \tag{7}$$

$$\alpha^*_{m,k} = \text{mode}(\alpha_m) \text{ for player } k \tag{8}$$

$$\alpha^t_{m,k} = \alpha_m \text{ for player } k \text{ in iteration } t \tag{9}$$

$$\text{Consistency}_{\alpha_m} = \frac{1}{10 * N} \sum_{k=1}^{N} \sum_{t=1}^{10} \mathbb{1}(\alpha^*_{m,k} == \alpha^t_{m,k}) \tag{10}$$

In table 4, we observed that in 99.4% of situations, the model returned the same α_m score as a player's most common α_m score. In all cases where the α_m was not the same, the difference was minor and ranged in value from 5-10.

We observed lower consistency for strengths and weaknesses in table 4, as expected since the model is generating an entire sentence rather than a numerical score. Differences were minor and did not impact overall meaning when sentences diverged. For example, the only difference in weaknesses generated from a report on Nikita Chibrikov was "inconsistent in his level of engagement and presence on the ice" instead of "inconsistent in his level of engagement and activity."

Non-determinism in model results may be due to floating point errors or model architecture (Chann, 2023), however our consistency checks demonstrate that differences were not only rare but also minor and thus unlikely to impact our results.

See appendix 8.2 for the code, which can be used to reproduce results.

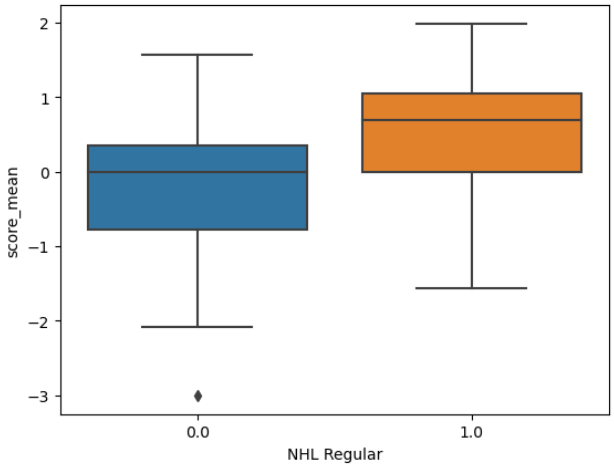

Figure 1. Likelihood of Making the NHL (Standardized α_m) vs NHL Regular Outcome in training dataset.

As part of our exploratory data analysis, we compared these average ranks to our dependent variable of being a NHL regular in our training set - see section 3.9 for more details about train-test split. In Figure 1, we see that players who ended up being NHL regulars in our training dataset had a median standardized α_m of 0.7 compared to median standardized α_m of zero for non-NHL regulars. See the following section for more details on data processing.

To generate summaries of player strengths and weaknesses, we asked OpenAI's gpt-3.5-turbo-1106 model to create lists of topics using the following prompt:

```
topics_template = """You are an ice hockey
    expert. You are given a set of player
    reports separated by periods:
    {report_text}.
 Return a list of 10-15 generalized traits
     mentioned in these reports in the
     following format:
"Name of Topic 1: Explanation of Topic 1
Name of Topic 2: Explanation of Topic 2
...
"
"""
```

We generated different topics for forwards and defencemen due to differences in skillsets, i.e., a defenceman's transition ability to move the puck up ice is important, however less important for forwards. We sampled without replacement from all forwards in the training dataset to obtain a random sample of 100 forwards. This was then repeated for defencemen. See appendix 8.3 for the code, which can be used to reproduce results.

These raw LLM-generated topics were refined using human intervention to be more representative of on-ice behaviour and to reduce overlap between topics. For example, the LLM generated separate topics for "competitiveness" and "work ethic", two topics which logically led to substantial overlap in their descriptions.

This resulted in the following human-adjusted forward strengths and descriptions:

- Skating: Strong skating ability with good speed, agility, and balance
- Playmaking: Able to create scoring chances, make great passes, and has strong vision
- Shooting: Impressive shot, quick release, and goal-scoring ability
- Puckhandling: Quick hands and puckhandling ability to beat opponents easily
- Hockey IQ: Has smart positioning, able to anticipate plays and make quick decisions on the ice
- Competitiveness: Able to win battles, competitive nature, and strong work ethic
- Physical Game: Strong and physical play on the ice
- Size: Large player who uses it effectively on the ice
- Versatility: Able to play a variety of roles and excel in all situations
- Defensive Abilities: Responsible defensive player and able to disrupt opponent plays
- Leadership: Good leadership qualities

This list encapsulated not only physical attributes like skating but also mental attributes like leadership. Likewise, we generated the following forward weaknesses:

- Skating: Concerns about speed, quickness, and stride technique
- Offensive Ability: Questioned in terms of playmaking, finishing, and overall skill level
- Hockey IQ: Poor decision-making, reads, and understanding of the game
- Defensive Play: Concerns about consistency, defensive engagement, and battles
- Consistency: Inconsistent effort and weak play away from the puck
- Puck Management: Tendency to force plays, make risky decisions, and have issues with turnovers
- Size: Undersized and lacks physicality
- Physical Game: Lack of strength and physical play on the ice
- Inexperience: Concerns about facing more experienced players at the next level
- Injury History: Significant injury history that might impact his play on the ice in the future

See Appendix 8.1 for corresponding strengths/weaknesses for defencemen. While some topics like skating were universal across both positions, others were more relevant for one position.

We then leveraged OpenAI's gpt-3.5-turbo-1106 model to classify each player into their corresponding strengths/weaknesses using the following prompt:

```
classification_template = """You are an
    ice hockey expert. You are given a
    player report: {report_text}.
 Which of the following {comment_type} are
     mentioned in this report?
 Only use {comment_type} from this list
     with their description (delimited with
     ":"): {topic_list}
 Return a list of relevant {comment_type}
     for this report. If no {comment_type}
     in that list are present, return an
     empty list: []
 """
```

In the above prompt, we replace `comment_type` with either "strengths" or "weaknesses" depending on which classifications we want to obtain. As mentioned previously in this section, we used temperature of zero to maximize determinism in model results. See appendix 8.4 for the code, which can be used to reproduce results.

Finally, these topics were one-hot encoded to integrate these topics into the model, i.e., a variable such as "strength: skating" was created and equal to 1 if skating was mentioned as a strength, 0 otherwise.

3.4. Scouting Reports: Data Processing

These scouting report features were processed to serve as input into gradient-boosted and neural net base models.

A player's rank and LLM-generated features were then concatenated and aggregated so each row was unique by player and season. For example, Connor Bedard was ranked first overall in his draft year by all scouts and thus had an average rank of 1. This was then further standardized using a modified z-score calculated with median instead of mean to be more robust to outliers.

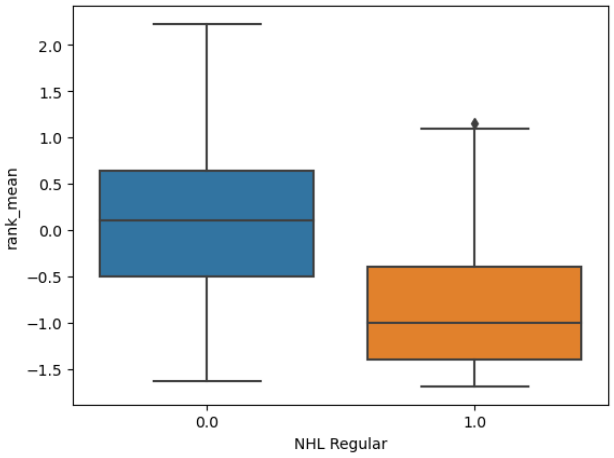

Figure 2. Standardized Average Rank vs NHL Regular Outcome in training dataset.

Figure 2 shows that on a median basis, players in our training dataset who ended up being NHL regulars had an average standardized rank of -1.0 compared to 0.1 for players who did not end up becoming NHL regulars. However, there were a number of players with a low rank who did not become a NHL regular.

A similar process was repeated for LLM-generated features, as discussed previously in section 3.3. If there were more than one scouting report for a player in a season, this aggregation would get average values for α_m. For one-hot-encoded topics, we also aggregated using the average to get the percentage of reports which mentioned that strength or weakness in that player's reports for a specific season.

We then further processed these features using standardization, winsorization, and imputation. Data was standardized using a modified z-score calculated with median instead of mean to be more robust to outliers in most variables. For columns derived from one-hot encoded topics, we kept original scores to be consistent with their original meaning indicating whether that topic was mentioned in the player's report.

Outliers were defined as z-scores three standard deviations above or below the median and were handled using winsorization to maximize information preservation while making model results more robust. Missing values were also imputed using a regression-based approach to fill in missing data using non-missing data from that scouting report.

Finally, we dropped highly correlated columns over an 80% threshold to mitigate multicollinearity in the data. This processing on scouting reports led data on 1,029 unique players. This processed scouting report data was the input into the first part of our ensemble machine learning approach, creating a model only on scouting reports to represent the eye test.

3.5. Player Statistics: Data Sources

We obtained on-ice statistics for players before and after their draft year from a publicly-available online database called EliteProspects (EP).

Data was collected for 2,072 players who were drafted in the NHL in the range $[2015, 2023]$ or were undrafted and had a scouting report written about them. This resulted in 372,000 records unique by EP player URL, season, and which team/league they were on. For each record, we had data on numerous attributes listed below. All on-ice statistics were available for regular-season, playoffs, or relegation game types:

- Metadata: URL, name, date/place of birth, and position
- Physical Attributes: Height, weight, and handedness
- Team: Team/league name, season
- Draft: Which year they were drafted and which which pick
- On-Ice (All): # Games Played
- On-Ice (Forwards/Defencemen): Goals/Assists scored, penalty minutes (PIM), and +/-
- On-Ice (Goalies): Goals-Against Average (GAA) and save percentage
- Captaincy: Whether a player was a captain or assistant captain on their team

3.6. Player Statistics: Feature Engineering

The overall goal of feature engineering was to create a robust, consistent, and comprehensive set of features with contextual knowledge such as number of games played and level of difficulty.

Aggregated positions were first created so each player was classified as a forward, defender, or goalie. These were

created to make our analysis more robust given EP often used granular classifications that have limited impact on draft outcomes. Skaters, i.e., forwards and defencemen, included 90% of all unique players so we decided to focus on skaters and exclude goalies for the rest of this section. This decision was also made due to large differences in the type of on-ice statistics tracked for goalies, a difficulty in translating goalie performance between different leagues, and the simple fact that goalie performance is highly volatile even among top NHL goalies.

Top players are selected by their national hockey federations to participate in international competitions each year. The most important tournament for draft outcomes is the World Juniors (WJC), featuring the best players in the world under age 20. This is an especially important tournament as scouts have the opportunity to compare top prospects who may play in different leagues directly against each other for the first time. The World Juniors' high-pressure environment, even playing field, and elevated level of competition is also an important environmental and contextual factor for evaluating players who usually participate in less-competitive leagues.

A player's performance in the World Juniors has a significant impact on how scouts evaluate them. For example, Jesse Puljujärvi's domination of the World Juniors as a draft-eligible player in 2016 was a significant factor for why scouts thought highly enough of him to take him third overall in that year's draft.

Furthermore, as outlined in section 2, Network NHLe (NNHLe) translates a player's performance from one league into what they would have obtained in the NHL using graph traversal aggregation to increase robustness (Turtoro, 2020). These NNHLe ratios were also added to a player's on-ice statistics and leveraged to translate a player's total points, the sum of their goals and assists. This resulted in an Adjusted Total Points (TP) metric.

As each player may be on more than one team during a season, we then aggregated the dataset so each year was a unique combination of player and season. For example, a player's Adjusted TP for a season was the sum of their Adjusted TP for each team they played at that season.

Players participate in a different number of games, so a further adjustment needs to be made to calculate an Adjusted Total Points per Games Played (Adjusted TP/GP). Similar per-game adjustments were made to PIM and +/-.

Top draft-eligible players by Adjusted TP/GP are shown in table 5, demonstrating that top performers by this metric are generally selected with the first three picks of the draft.

Another important factor to account for is not all players are the same age at the draft. In general, players age 18-21 inclusive as of September 15 on the year of the draft are eligible, provided they have not been picked previously. As a result, players who were undrafted the previous year remaing eligible to be picked, however, expectations are higher as they have had another year to develop compared to first-time draft-eligible players. Thus, we calculated player age as of the time they were drafted as an important contextual datapoint for our downstream model.

Table 5. Selected on-ice statistics of top draft-eligible players by Adjusted Total Points/Games Played.

NAME	DRAFT YEAR	DRAFT RANK	ADJUSTED TP/GP	TP/GP WJC
AUSTON MATTHEWS	2016	1	0.479	1.571
ADAM FANTILLI	2023	3	0.417	0.715
CONNOR BEDARD	2023	1	0.406	3.286
CONNOR MCDAVID	2015	1	0.396	1.571
JACK EICHEL	2015	2	0.362	0.800

3.7. Player Statistics: Data Processing

These on-ice player statistics were processed to serve as input into gradient-boosted and neural net base models.

First, the dataset was aggregated so each row represents a single player. Given this dataset was used to predict draft outcomes, we focused on a player's performance during the season leading up to their draft (D-1) and the two seasons prior (D-2, D-3). There were also three distinct game situations, specifically regular-season, post-season, and world juniors. This resulted in nine possible combinations of season relative to draft and game situation.

Additional features were engineered to examine player growth from season to season, i.e., percentage change in Adjusted TP/GP from their D-2 to D-1 seasons.

A snapshot of some of these raw variables is shown in table 6, which demonstrates that 25.6% of players were wingers and only 4.3% of them served as a captain of a team at some point during the season leading up to their draft. Table 6 also shows that the median player had 54 games played in their D-1 season, with a median Adjusted TP/GP metric of 0.092. On a median basis, this was an 83.5% improvement from their Adjusted TP/GP during their D-2 season.

Differences between the median and mean suggested that some variables have a skewed distribution. An example of this skewness is seen in the distribution plot shown in Figure 3. The presence of outliers leads to a long right tail,

Table 6. Snapshot of raw selected variables based on player statistics.

VARIABLE	MEAN	MEDIAN
POSITION - WINGER	0.256	0.000
D-1 CAPTAIN	0.043	0.000
D-1 # TEAMS	1.372	1.000
D-1 # GAMES PLAYED	49.407	54.000
D-1 ADJUSTED TP/GP	0.101	0.092
D-2 TO D-1 ADJUSTED TP/GP % CHANGE	1.344	0.835

thus demonstrating a right-skewed distribution. This can be explained by the presence of certain prospects in each draft class who have dominated lower levels of competition to a degree unmatched by their peers. For example, a generational player like Connor McDavid is clearly an outlier compared to the typical player drafted.

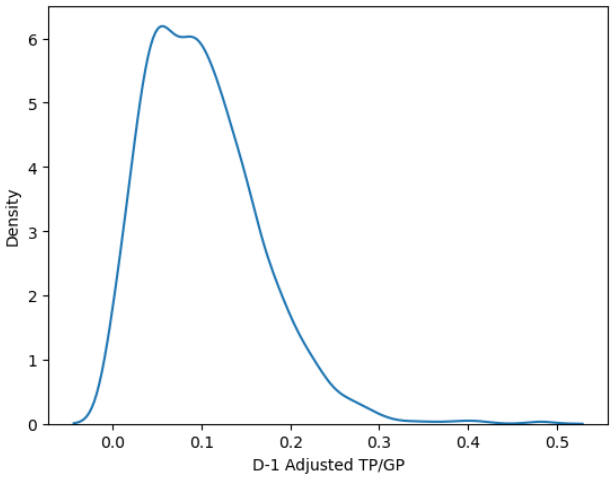

Figure 3. Distribution of Adjusted Total Points/Games Played during the season preceding their draft

Skewness and the presence of outliers meant that further processing needed to be done. Data was then standardized using a modified z-score calculated with median instead of mean to be more robust to outliers for most variables. For boolean variables, such as whether a player was a winger, the z-score was still calculated using the mean since the median for these variables was just zero or one.

Outliers were defined to be z-scores three standard deviations above or below the median/mean, depending on the variable as outlined in the above paragraph. Winsorization was used to handle these outliers, to maximize information preservation while making model results more robust.

Various strategies were used to impute missing variables. Number of games played, number of teams/leagues they were on that season, and whether they were a captain of a team were imputed with zeroes since a missing value was equivalent to a zero. For example, if a player did not participate in the World Juniors during their D-1 season, the games played category would be N/A which was equivalent to a zero.

On the other hand, missing values in other columns such as Adjusted TP/GP needed to be imputed more carefully. For example, a missing value in Adjusted TP/GP at the World Juniors during their D-1 season may mean that player was insufficiently good enough to participate or that their nation did not qualify for the World Juniors. For example, a German player may not have any data since their country did not qualify however they would have been selected if their country qualified for the tournament. Therefore, a regression-based approach was necessary to use non-missing data for each player to fill in their missing data.

Finally, we dropped highly correlated columns over an 80% threshold to mitigate multicollinearity in the data while optimizing information retention. This processing on player statistics leading up to the draft resulted in data on 1,713 unique players. This processed player statistics data was the input into the second part of our ensemble machine learning approach, creating a model only on player statistics.

3.8. Mapping

To create a mapping between scouting reports and player statistics, we used a full and fuzzy string match strategy based on player names. This automated methodology resulted in 89.7% of scouting reports having a single assigned player name, which was then augmented using a manual process. This resolved edge cases such as Sebastian Aho when different players had the same name.

3.9. Modelling

We trained different base models for forwards and defencemen separately on each dataset, i.e., scouting reports or player statistics. Therefore, there were four separate base models required. We decided to train separate models for each position group due to differences in scouting report features as outlined in section 3.3 and large differences in player scoring between position groups, i.e., forwards score more than defencemen.

For each modelling task, we first split each dataset into training, validation, and testing sets. Training data contained all players with a draft year of 2019 or earlier while validation data contained players with a draft year in 2020. It was too early to judge players drafted in 2021 or later and thus we decided to denote these players as the testing set. Data was split by year to avoid temporal data leakage between years.

We then found the best model for two different types of models, a gradient-boosted approach and a deep learning

approach, which will be discussed separately.

First, gradient-boosting was selected as an approach to improve the model's performance by giving more weight to weak learners trained on datapoints that were previously incorrect. We then used extensive hyperparameter tuning over 100 trials and adjusting for parameters such as learning rate, number of estimators, and maximum tree depth. A Tree-structured Parzen Estimator was used to navigate our search space with an objective to minimize validation RMSE (Akiba et al., 2019).

Note that the number of defencemen in our training set was smaller than that of the number of forwards. Therefore, to avoid overfitting, we defined the search space differently for defencemen than we did for forwards, specifically setting smaller max values for the number of estimators and the learning rate.

A similar process was done for our second modelling approach, using a fully-connected neural network. Our network architecture contained of two or more hidden layers, each consisting of a linear layer, a Leaky ReLU activation function, and then a dropout layer. We used dropout to reduce overfitting given the relatively small dataset, and used leaky ReLU to reduce the risk of vanishing gradients as the gradients are less likely to be saturated around zero or the endpoints with this activation function.

After all hidden layers, we applied a final activation function on the output, differing based on the dependent variable. Whether a player would be a NHL regular was a classification problem so a sigmoid activation function was applied to the final output to convert model results into probability space.

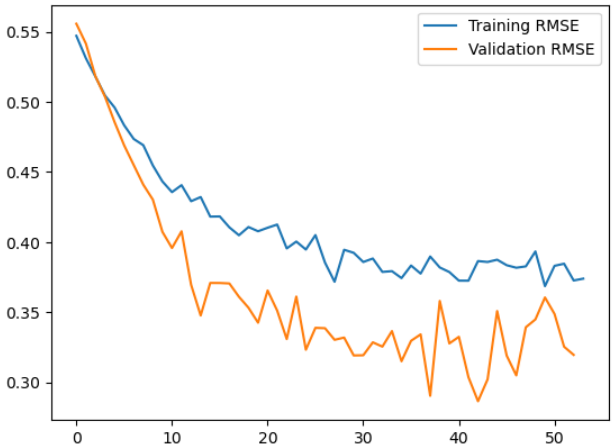

Figure 4. RMSE using Neural Net on scouting report dataset (Hyperparameter Trial # 70)

We underwent extensive hyperparameter tuning on neural net models across 100 trials adjusting hyperparameters such as the number of hidden layers, learning rate, and dropout percentage. For each trial, we trained on training dataset with an Adam optimizer for stochastic gradient descent and a learning rate scheduler to decrease the learning rate on plateau. A Tree-structured Parzen Estimator was used to navigate our search space with an objective to minimize validation RMSE (Akiba et al., 2019).

We also implemented early stopping criteria so the model stopped training if the validation RMSE had not decreased after ten epochs to reduce runtime and prevent overfitting. Figure 4 shows results from a hyperparameter tuning trial displaying RMSE over the number of epochs for a neural net when predicting NHL regular outcomes for forwards with an early stop around epoch 55 due to early stopping criteria. We can see a clear downwards trend on training RMSE, while validation RMSE is more volatile however still trends downwards overall.

These two modelling approaches were then repeated for each position group and dataset.

3.10. Modelling: Ensemble Learning

After creating base models independently on different datasets, we created an overall prediction using a meta model as part of a stacking ensemble approach. For example, a forward's combined probability of being a NHL regular would be based on predictions from base models trained on each dataset. We repeated the model-training procedure outlined in section 3.9 to train gradient-boosted and neural-net models on top of base model predictions for each player.

Model stacking was was used due to large differences between scouting report and player statistics which were the inputs into our base models. These differences would make it less robust to process, impute, or model on a combined dataset. Therefore, we decided to train separate base models on each dataset in parallel to avoid spurious conclusions. Note that if a player had no data on one of the base datasets, we imputed that missing dataset's prediction using the non-missing dataset's prediction.

Another benefit of a stacking framework was we could easily compare model results and identify players for whom one model outperforms the other. This is further discussed in section 4.3.

4. Results

4.1. Results for Forwards

In order to compare our results against a benchmark, we created a trivial model which randomly assigned players to NHL regular outcomes based on the proportion of NHL regulars in the training dataset. We used a random seed to

Table 7. Summary of all results for forwards.

Dataset	Model	Position	Train RMSE	Val RMSE
N/A	Trivial	F	0.587	0.521
Stats	GB	F	0.241	0.343
Stats	NN	F	0.325	0.311
Reports	GB	F	0.162	0.356
Reports	NN	F	0.332	0.319
Ensemble	NN	F	0.197	0.268

Table 8. Selected important features for forwards on each respective base model. Deltas are average Z-scores of NHL regulars in training set subtracted by average Z-scores of non-NHL regulars in training set.

Feature	Delta	Importance
AVG Likelihood of making NHL (α_m)	1.237	0.425
AVG Rank	-0.929	0.164
AVG Report Length	0.707	0.079
STDEV α_m	-0.021	0.072
STDEV Rank	-0.160	0.070
AVG # Strengths	0.399	0.022
D-1 Adjusted TP/GP	0.737	0.119
D-1 TP/GP WJC	0.401	0.101
D-1 Draft Age	-0.251	0.079
D-2 Adjusted TP/GP	0.605	0.051
D-1 GP WJC	0.547	0.048

ensure deterministic and reproducible results.

We then used the methodology described previously to obtain results for forwards in table 7, separated by dataset and model type. We observe that base models on player statistics generally outperform base models on scouting reports alone. However, a combined model outperformed each individual model across the board with lower validation RMSE scores than base models trained on only one of the two datasets.

Table 8 shows selected Gini feature importance scores for a gradient-boosted model predicting whether a forward will be a NHL regular. This feature importance demonstrated that α_m, the LLM-generated likelihood score on whether a player will make it to the NHL was the most important feature by a large margin. NHL regulars have substantially higher average values of α_m compared to non-NHL regulars as seen in the deltas between these two groups' z-scores.

Other important features were average rank, average report length, and number of strengths. These findings again aligned with our prior that forwards who were on average ranked earlier by scouts, i.e., first overall instead of tenth overall, were more likely to make the NHL. Likewise, table 8 shows that scouts not only ended up writing longer reports for forwards who ended being more likely to make the NHL, but also included more strengths in their reports.

There was also less disagreement among scouts, i.e., lower standard deviation of α_m and rank, among players in our training set who ended up being NHL regulars.

Table 8 also shows Gini feature importance and deltas between NHL regulars and non-NHL regulars for a gradient-boosted base model trained on player statistics data. We see a more balanced set of important features when predicting whether a forward is on track to be a NHL regular.

The more points a player has, adjusted for level of competition and games played, the more likely they are to be a NHL regular. Their point production at the world juniors and the number of games they played at the tournament were also important factors. Their point production in in the season leading up to their draft (D-1) was more important than that of the season prior (D-2).

Younger players at the time of their draft are slightly more likely to be a NHL regular, with a negative delta in table 8. This may be because younger players have more biological time to develop compared to older players after they are drafted, thus increasing the chance of them taking a regular shift at the NHL level.

Table 9. Summary of all results for defencemen.

Dataset	Model	Position	Train RMSE	Val RMSE
N/A	Trivial	D	0.551	0.398
Stats	GB	D	0.239	0.297
Stats	NN	D	0.291	0.246
Reports	GB	D	0.248	0.317
Reports	NN	D	0.259	0.306
Ensemble	NN	D	0.199	0.220

4.2. Results for Defencemen

Similar to our results on forwards in section 4.1, Table 9 shows ensemble models combining both player statistics and player reports outperform either base model on a single dataset alone across the board.

Base models on defencemen statistics again generally outperform base models on their scouting reports alone. Note that for the trivial model, we used the average percentage of defencemen to become NHL regulars in a random assignment as outlined previously in section 4.1.

Table 10 shows important scouting report features. This was also based on Gini feature importance scores from a gradient-boosted model when predicting NHL regular

Table 10. Selected important features for defencemen on each respective base model. Deltas are average Z-scores of NHL regulars in training set subtracted by average Z-scores of non-NHL regulars in training set.

FEATURE	DELTA	IMPORTANCE
AVG LIKELIHOOD OF MAKING NHL (α_m)	1.019	0.480
STDEV RANK	-0.146	0.077
STDEV α_m	-0.170	0.070
AVG REPORT LENGTH	0.467	0.067
NET STRENGTHS/WEAKNESSES	0.289	0.052
AVG RANK	-0.755	0.045
D-1 ADJUSTED TP/GP	0.701	0.253
D-1 +/- PER GP WJC	0.357	0.066
D-1 +/- PER GP	0.395	0.065

outcomes for defencemen. Findings are broadly consistent with results for forwards in table 8, with NHL regulars generally having higher LLM-generated likelihood scores of being NHL regulars, longer reports, and lower average ranks.

An interesting area where forwards and defencemen differ is the standard deviation of rank is a more important factor than the average rank for defencemen. Therefore, this suggests that having consensus among scouts, i.e., lower standard deviations in their ranks, may be related to a player becoming more likely to being a NHL regular.

Another difference is the net number of strengths and weaknesses was more important for defencemen, while the average number of strengths was more important for forwards. This suggests having fewer weaknesses is more important for defencemen in becoming NHL regulars than it is for forwards.

In table 10, we also see that for player statistics, a player's point production remains the most important feature in predicting whether a defenceman will become a NHL regular like it was for forwards in table 8.

However, +/- becomes a more important stat for defencemen than it was for forwards. Although +/- is a flawed stat as it does not account for quality of competition and which teammates are on the ice with a player, it seems to have some value as a measurement for a player's capabilities especially on the defensive end.

4.3. Comparing Scouting Report Predictions vs Player Statistics Predictions

There were numerous cases where scouting report models outperformed player statistics models, and vice versa. We defined a model's outperformance over another model as a difference in an absolute error greater than 0.1. This margin was empirically selected to focus on relatively large differences in model conclusion rather than minor differences in model errors.

Table 11. Average attributes where one model outperforms another with absolute error greater than 0.1 on training dataset.

FEATURE	REPORTS OUTPERFORM	STATS OUTPERFORM
# FORWARDS	100/622	36/622
% NHL REGULAR (F)	0.610	0.306
# DEFENCEMEN	38/362	17/362
% NHL REGULAR (D)	0.816	0.235

Table 11 shows that when predicting whether a forward would become a NHL regular, there were 100/622 forwards in our training set where a scouting report base model performed better and 36/622 forwards where a player statistics base model performed better. Among forwards for whom the scouting report base model performed better, a higher percentage ended up being NHL regulars than among forwards for whom the player statistics base model performed better.

A similar pattern is also seen among defencemen in table 11. Note that all sample sizes are relatively small, especially among defencemen.

4.4. Example Player Results

We will look at a few players to illustrate our model's results, starting with Dylan Holloway. Scouts were quite positive about him as he scored 1.66 standard deviations above the median for his likelihood of making it to the NHL (α_m). While scouts praised his competitiveness, physical game, and shooting, they raised concerns about his offensive ability. Overall, however, scouts were quite positive compared to a statistical model which saw respectable but not amazing point production in his draft year at just 0.308 standard deviations above the median adjusted TP/GP. He also did not produce a lot of points at the world juniors, ending up 0.371 standard deviations below the median. This mixed bag ended up resulting in the scouting report base model producing more optimistic results about him being on track to becoming a NHL regular. Holloway has currently played over a full season of games in the NHL for the Edmonton Oilers.

Moritz Seider is another player who scouts were positive about as he scored an average 1.295 standard deviations above the median for α_m. Scouts praised his defensive abilities, poise, and skating while raising concerns about

his offensive upside. His point production was weak as he scored 0.511 standard deviations below the median Adjusted TP/GP among defencemen. In this case, the statistical model lacks context that he held his own against as a 17-year-old against grown men while winning a championship in Germany's top pro league. However, this context was discussed in scouting reports and thus the scouting report base model performed much better than the player statistics model. Seider has evolved into one of the best young defencemen in the NHL today having already played over 200 games for the Detroit Red Wings.

On the other hand, Timo Meier had an extremely strong statistical profile, posting an adjusted TP/GP that was 2.20 standard deviations above the median and performed well at the world juniors in his draft year. However, scouts were relatively less positive about him as he only had a standardized α_m Z-score of 0.555. This resulted in the player statistics base model producing more optimistic results about him being on track to becoming a NHL regular. Meier has now played over 500 NHL games for the San Jose Sharks and New Jersey Devils.

Ridley Greig caused substantial disagreement among scouts in his ranking, with a rank standard deviation Z-score of 0.928. Specifically, Scott Wheeler ranked him 64th overall while Cory Pronman had him ranked 37th overall in their respective 2020 draft rankings. His point production, however, was quite clear as he scored 1.256 standard deviations above the median Adjusted TP/GP production in his draft year. The player statistics base model ended up being more optimistic about his chances of being a NHL regular than the scouting reports base model was. Greig looks on track to be a NHL regular and has currently played just under a full season of games in the NHL for the Ottawa Senators.

5. Future Work

The mapped dataset used in this paper combining player attributes and scouting reports is a starting point for numerous potential future applications. For example, we could take a permutation testing approach to identify whether scouts are more likely to say players of a certain category have particular strengths or weaknesses. This may help identify blind spots where scouts are generally over or under-valuing players.

In addition, it may be interesting in a future work to evaluate scout tendencies and criteria. For example, when a scout says a player is a good skater, that statement may have different meanings depending on which scout said it.

Another interesting future area of work would be adding video or images from hockey games into future models. This would leverage video understanding (Tang et al., 2024) and reinforcement learning to further incorporate the eye test into analytical models.

The results we found in this paper can be broadly expanded to other metrics, sports, and industries. A hockey player's games played and points are only one measurement of a player's impact and undervalues players with strong defensive results. Therefore, other metrics such as GSVA (Luszczyszyn, 2023), which is separated by offensive and defensive on-ice impact, would be a useful addition to the model based on quantitative metrics.

In addition, other sports such as baseball have extensive precedents in scouting and quantitative analysis. For example, baseball scouts evaluate players using a standardized rubric on a 20-80 scale and those metrics would be a valuable datapoint in a baseball application.

Finally, this general principle of creating features based on text documents and combining it with structured data has utility for a wide variety of industries outside of sports, as recently demonstrated in healthcare (Tu et al., 2022) and finance (Lopez-Lira and Tang, 2023).

6. Conclusion

This paper has demonstrated that LLMs are useful in extracting information from textual data like scouting reports in hockey. In particular, these reports have information and context missing from a player's on-ice statistics. We have also demonstrated that the best results come from combining these datasets in an ensemble learning framework that integrates the eye test with analytics.

7. Acknowledgments

Thanks to Amanda Glazer for advice throughout this process, and thanks to Yayu Xu, Fresa Luo, Steve Cao, Kevin Lin, Arvind Kumar, Junsheng (Allen) Shi, Anisha Jahagirdar, Jack Han, and David Wang for helpful discussions and feedback.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Berri, D. J., Brook, S. L., and Fenn, A. J. (2011). From college to the pros: predicting the nba amateur player draft. *Journal of Productivity Analysis*, 35(1):25–35. DOI:10.1007/s11123-010-0187-x.

Chann, S. (2023). Non-determinism in gpt-4 is caused by sparse moe.

Chen, B., Zhang, Z., Langrené, N., and Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: a comprehensive review.

Deaner, R. O., Lowen, A., and Cobley, S. (2013). Historical perspectives and current directions in hockey analytics. *Public Library of Science ONE*, 8(1):1–7. DOI:10.1371/journal.pone.0057753.

Desjardins, G. (2005). Projecting junior hockey players and translating performance to the nhl. *Behind the Net*.

Liu, Y., Schulte, O., and Li, C. (2019). Model trees for identifying exceptional players in the nhl and nba drafts. In *Machine Learning and Data Mining for Sports Analytics*, pages 93–105. Springer International Publishing.

Lopez-Lira, A. and Tang, Y. (2023). Can chatgpt forecast stock price movements? return predictability and large language models. https://arxiv.org/abs/2304.07619.

Luszczyszyn, D. (2023). Introducing the 'new' nhl stats fans should know: Offensive and defensive rating. *The Athletic*.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). Stemming and lemmatization. In *Introduction to Information Retrieval*.

Nandakumar, N. and Jensen, S. T. (2018). Historical perspectives and current directions in hockey analytics. *Annual Review of Statistics and Its Application*, 6(1):19–36. DOI:10.1146/annurev-statistics-030718-105202.

Schuckers, M. (2011a). An alternative to the nfl draft pick value chart based upon player performance. *Journal of Quantitative Analysis in Sports*, 7(2):10–10. DOI:10.2202/1559-0410.1329.

Schuckers, M. (2011b). What's an nhl draft pick worth? a value pick chart for the national hockey league.

Schuckers, M. (2016). Draft by numbers: Using data and analytics to improve national hockey league player selection. *MIT Sloan Sports Analytics Conference*.

Seppa, T., Schuckers, M. E., and Rovito, M. (2017). Text mining of scouting reports as a novel data source for improving nhl draft analytics. *Ottawa Hockey Analytics Conference*.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020). Learning to summarize with human feedback. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Tang, Y., Bi, J., Xu, S., Song, L., Liang, S., Wang, T., Zhang, D., An, J., Lin, J., Zhu, R., Vosoughi, A., Huang, C., Zhang, Z., Zheng, F., Zhang, J., Luo, P., Luo, J., and Xu, C. (2024). Video understanding with large language models: A survey.

Tu, T., Loreaux, E., Chesley, E., Lelkes, A. D., Gamble, P., Bellaiche, M., Seneviratne, M., and Chen, M.-J. (2022). Automated loinc standardization using pre-trained large language models. In Parziale, A., Agrawal, M., Joshi, S., Chen, I. Y., Tang, S., Oala, L., and Subbaswamy, A., editors, *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193 of *Proceedings of Machine Learning Research*, pages 343–355. PMLR.

Turtoro, C. (2020). Network nhl equivalences (nnhle).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

Wheeler, S. (2023a). 2023 nhl draft ranking. *The Athletic*.

Wheeler, S. (2023b). What is the scouting process for nhl draft prospects? everything you need to know in 2023. *The Athletic*.

Wolfson, J., Addona, V., and Schmicker, R. H. (2011). The quarterback prediction problem: Forecasting the performance of college quarterbacks selected in the nfl draft. *Journal of Quantitative Analysis in Sports*, 7(3). DOI:10.2202/1559-0410.1302.

8. Appendix

8.1. Forward Strengths/Weaknesses

The following were the lists of strengths and weaknesses for players, after a human adjustment to the original classes generated by a LLM.

Forward Strengths

- Skating: Strong skating ability with good speed, agility, and balance
- Playmaking: Able to create scoring chances, make great passes, and has strong vision
- Shooting: Impressive shot, quick release, and goal-scoring ability
- Puckhandling: Quick hands and puckhandling ability to beat opponents easily
- Hockey IQ: Has smart positioning, able to anticipate plays and make quick decisions on the ice
- Competitiveness: Able to win battles, competitive nature, and strong work ethic
- Physical Game: Strong and physical play on the ice
- Size: Large player who uses it effectively on the ice
- Versatility: Able to play a variety of roles and excel in all situations
- Defensive Abilities: Responsible defensive player and able to disrupt opponent plays
- Leadership: Good leadership qualities

Defenceman Strengths

- Skating: Strong skating ability with good speed, agility, and balance
- Defensive Abilities: Strong defensive play and able to disrupt opponent plays
- Transition Game: Able to transition the puck up ice effectively, quickly, and cleanly
- Physical Game: Strong and physical play on the ice
- Size: Large player who uses it effectively on the ice
- Competitiveness: Able to win battles, competitive nature, and strong work ethic
- Hockey IQ: Has smart positioning, able to anticipate plays and make quick decisions on the ice
- Poise and Patience: Poised under pressure and patient in making plays
- Playmaking: Able to create scoring chances, make great passes, and has strong vision
- Puckhandling: Quick hands and puckhandling ability to beat opponents easily
- PowerPlay Quarterbacking: Able to quarterback the power play effectively
- Leadership: Good leadership qualities

Forward Weaknesses

- Skating: Concerns about speed, quickness, and stride technique
- Offensive Ability: Questioned in terms of playmaking, finishing, and overall skill level
- Hockey IQ: Poor decision-making, reads, and understanding of the game
- Defensive Play: Concerns about consistency, defensive engagement, and battles
- Consistency: Inconsistent effort and weak play away from the puck
- Puck Management: Tendency to force plays, make risky decisions, and have issues with turnovers
- Size: Undersized and lacks physicality
- Physical Game: Lack of strength and physical play on the ice
- Inexperience: Concerns about facing more experienced players at the next level
- Injury History: Significant injury history that might impact his play on the ice in the future

Defenceman Weaknesses

- Skating: Concerns about speed, quickness, and stride technique
- Defensive Play: Issues with positioning, decision-making, and battles
- Offensive Upside: Lack of creativity, puck skills, and scoring production
- Size: Undersized and lacks physicality
- Hockey IQ: Poor decision-making, reads, and understanding of the game

- Consistency: Inconsistent effort and weak play away from the puck
- Transition: Unable to move the puck up the ice
- Puck Management: Tendency to force plays, make risky decisions, and have issues with turnovers
- Physical Game: Lack of strength and physical play on the ice
- Inexperience: Concerns about facing more experienced players at the next level
- Injury History: Significant injury history that might impact his play on the ice in the future

8.2. LLM Code: Likelihood Scores + Generating Player Strengths/Weaknesses

```
llm = ChatOpenAI(
   openai_api_key=openai.api_key,
   model_name='gpt-3.5-turbo-1106',
   temperature=0,
   )

nhl_template = """You are an ice hockey
    expert. You are given this player's
    scouting report: {report_text}.
Answer in this format:
"SCORE: integer between 0 and 100 for
    whether he will make it to the NHL (1
    is impossible, 99 is certain)
STRENGTHS: one sentence on his strengths,
    based only on the provided report
WEAKNESS: one sentence on his weaknesses,
    based only on the provided report
"
"""

examples = [
   {
      "report_text": """podkolzin played a
          lot of hockey this season
          between multiple levels of
          junior, pro and international
          hockey, and impressed almost
          every single time. he almost
          always seems to have an impact
          on a game. he's super talented
          but also an elite competitor.
          podkolzin can make the flashy
          plays to deke defenders, but he
          rarely does that off a
          standstill or along the walls.
          he has hard skill. podkolzin is
          typically full speed ahead to
          the net; and if he needs to go
          around or through you, he will.
          he's also a very good playmaker
          and finisher who can take
          advantage of space if defenders
          make off him by making a pass or
          sniping from a distance. quite
          often he made passes this season
          that were elite, but he didn't
          rack up that many assists. the
          one thing that bugs me about him
          is his skating. his stride is a
          little awkward and inefficient,
          he's hunched over, kicks his
          heels out, but he generates
          decent speed and hustles so hard
          that any technical flaw isn't
          that exposed. he has two years
          left on his khl contract with
          ska and told the athletic he
          intends to see that contract
          out.russian u18 coach vladimir
          filatov on podkolzin: "he's the
          heart of a team. he always wants
          to set an example on and off the
          ice. he's maybe not the most
          elite skill player or an elite
          sniper, but he's a leader, he
          runs the game. his game is
          always about controlling the
          puck, pushing the play forward
          and putting everything on the
          net.""",
      "answer": """
      SCORE: 85
      STRENGTHS: podkolzin is a highly
          talented and competitive player
          who consistently makes an impact
          on the game with his hard skill,
          playmaking ability, and strong
          work ethic
      WEAKNESS: his skating technique is
          inefficient and will hinder his
          ability to be an impact player
          at the NHL level
      """,
},
{
      "report_text": """vaakanainen has
          been on the prospect radar for
          many years. while he didn't have
          the draft season he may have
          hoped for, he still showed well
          at various points and remains a
          coveted player. vaakanainen has
          excellent two-way hockey sense.
          he's a smooth, creative puck
          mover who can dictate tempo and
          qb a power play due to his
          vision and a good slap shot. his
          skating isn't explosive, but he
```

```
        has an easy stride, with the
            ability to evade pressure and
            get around the ice. defensively,
            he's solid. he can use his body
            to win battles and play a sound
            positional game, closing his
            gaps effectively. if he gets
            back to the development track he
            was on when he was 15 and 16
            years old, he could be a great
            pro.""",
        "answer": """
        SCORE: 80
        STRENGTHS: vaakanainen is a skilled
            and creative puck mover with
            excellent two-way hockey sense
            and solid defensive abilities
        WEAKNESS: he did not have a strong
            draft season and his skating is
            not explosive
        """,
    },
    {
        "report_text": """the harvard commit
            was a player who consistently
            impressed me all season, and he
            was a top player for the usntdp.
            he has big-time offensive upside
            and some of the best offensive
            tools among the draft-eligible
            defensemen. fox can control the
            play very well in all three
            zones, shows great patience,
            creativity and vision, and
            creates space well with his puck
            skills. although he isn't an
            elite skater, he has
            above-average speed and agility
            and is able to make plays that
            require evasion. fox has shown
            some improvement defensively,
            but that remains a big issue in
            his game. a small defender is
            never going to dominate in that
            area, and he has been a little
            inconsistent in terms of
            positioning on that end. he is a
            high-risk player at times,
            trying to do too much, and he
            can pass the puck to the other
            team more than you'd like.""",
        "answer": """
        SCORE: 90
        STRENGTHS: fox has superstar
            offensive upside and fantastic
            offensive tools, with the
            ability to control the play in
            all three zones and create space
            with his puck skills.
        WEAKNESS: his defensive game and
            positioning are inconsistent,
            and he can be a high-risk player
            at times, however he has shown
            improvement defensively
        """,
    },
]

# template for examples where answer is
    already known
example_prompt = PromptTemplate(
    input_variables=["report_text",
        "answer"],
    template=nhl_template + "\nAnswer:
        {answer}",
    )

nhl_prompt = FewShotPromptTemplate(
    examples=examples,
    example_prompt=example_prompt,
    input_variables=["input"],
    suffix=nhl_template
)


class ScoutGPTParser(
    BaseOutputParser[list[str]]):
    def parse(self, t_output: str) ->
        pd.DataFrame:
        """
        Parse LLM output for nhl_template
        :param t_output: str
        :return: pd.DataFrame
        """
        t_dict = {'output_raw': t_output}
        for t in t_output.split("\n"):
            t_split = t.split(":")
            t_key = t_split[0].lower().strip()
            if len(t_split) <= 1:
                t_dict[t_key] = [np.nan]
            else:
                if len(t_split) > 2:
                    utils.logger.warning(
                        f"Expected only one
                            comma, instead got:
                            {t} - now taking
                            first entry")
                t_dict[t_key] = [t_split[1]]

        df_text =
            pd.DataFrame.from_dict(t_dict)

        for col in ['score']:
            if col in df_text.columns:
                df_text[col] = (df_text[col]
                .str.replace(r"[^0-9]", "",
                    regex=True)
                .str.strip().astype(int)
```

```
            )

        df_text =
            utils.process_str_cols(df_text,
            verbose=False)

        return df_text


nhl_runnable = nhl_prompt | llm |
    ScoutGPTParser()
```

8.3. LLM Code: Generating Topics

```
llm = ChatOpenAI(
   openai_api_key=openai.api_key,
   model_name='gpt-3.5-turbo-1106',
   temperature=0,
   )

topics_template = """You are an ice hockey
    expert. You are given a set of player
    reports separated by periods:
    {report_text}.
 Return a list of 10-15 generalized traits
     mentioned in these reports in the
     following format:
"Name of Topic 1: Explanation of Topic 1
Name of Topic 2: Explanation of Topic 2
...
"
"""

topics_prompt = PromptTemplate(
   input_variables=["report_text"],
   template=topics_template,
)


class TopicGPTParser(
   BaseOutputParser[list[str]]):
   def parse(self, t_output: str) ->
       pd.DataFrame:
       """
       Parse LLM output for topics_template
       :param t_output: str
       :return: pd.DataFrame
       """
       topics = [i for i in
           t_output.split("\n") if len(i) >
           0]

       dict_topics = {}
       for t in topics:
          t_split = t.split(":")
          t_key = t_split[0]
          if len(t_split) <= 1:
             utils.logger.error(f"Only got
                 topic name without topic
                 explanation - now
                 skipping: {t}")
          else:
             if len(t_split) > 2:
                utils.logger.warning(
                   f"Expected only one
                       comma, instead got
                       {t} - now taking
                       first entry")
             dict_topics[t_key] =
                 [t_split[1]]

       df_topics = (pd.DataFrame.from_dict(
                dict_topics, orient='index')
                .reset_index(drop=False)
                )
       df_topics.columns = ['Topic Name',
           'Topic Description']
       df_topics['Topic Name'] =
           (df_topics['Topic Name']
       .str.replace(r"[^a-zA-Z\s]", "",
           regex=True)
       )
       df_topics =
           utils.process_str_cols(df_topics,
                   verbose=False)

       return df_topics


topic_runnable = topics_prompt | llm |
    TopicGPTParser()
```

8.4. LLM Code: Classifying Topics

```
llm = ChatOpenAI(
   openai_api_key=openai.api_key,
   model_name='gpt-3.5-turbo-1106',
   temperature=0,
   )

classification_template = """You are an
    ice hockey expert. You are given a
    player report: {report_text}.
 Which of the following {comment_type} are
     mentioned in this report?
 Only use {comment_type} from this list
     with their description (delimited with
     ":"): {topic_list}
 Return a list of relevant {comment_type}
     for this report. If no {comment_type}
     in that list are present, return an
     empty list: []
 """
```

```
classification_prompt = PromptTemplate(
input_variables=["report_text",
    "topic_list", "comment_type"],
template=classification_template,
)


class ClassificationGPTParser(
   BaseOutputParser[list[str]]):
   def parse(self, t_output: str) -> list:
      """
      Parse LLM output for
          classification_template
      :param t_output: str
      :return: list
      """
      list_topics =
          [re.sub(r"[^a-zA-Z\s]", "", t)
          for t in t_output.split(",")]

      return list_topics


classification_runnable =
    classification_prompt | llm |
    ClassificationGPTParser()
```