

# Large Language Models on Race Commentary: Towards Granular Data in Cycling Analytics

Bram Janssens<sup>1,2,3</sup>[0000-0003-3078-7919], Matthias Bogaert<sup>1,2</sup>[0000-0002-4502-0764] and Steven Verstockt<sup>4</sup>[0000-0003-1094-2184]

<sup>1</sup> Ghent University, Department of Marketing, Innovation, and Organization, Tweekerkenstr. 2, 9000 Ghent, Belgium

<sup>2</sup> FlandersMake@UGent-corelab CVAMO

<sup>3</sup> Research Foundation Flanders

<sup>4</sup> Ghent University - imec, IDLab; Technologiepark-Zwijnaarde 122, 9052 Gent, Belgium

**Abstract.** Current cycling analytics studies are limited to data about the eventual race results. This study searches how online commentary can be used to capture information about in-race dynamics by harnessing the power of large language models. The results show that the direct application of these models is already promising but not accurate enough to base end-to-end machine learning applications on the generated data. Our results show the tendency of these models to use information from previous queries in its generation step, which indicates data leakage and might hamper the scientific validation of approaches comparing various techniques. To capture overall rider behavior we suggest using graph representation learning. Our results indicate that this method is capable of identifying similar rider behavior, which to date was not yet feasible.

**Keywords:** Cycling Analytics, Text Mining, Data Collection, Large Language Models, Graph Representation Learning

## 1 Introduction

Cycling analytics has seen an increase in popularity in the last couple of years, with applications such as race outcome estimation (Kholkina et al., 2021), talent identification (Van Bulck, Vande Weghe & Goossens, 2023; Janssens, Bogaert & Maton, 2023), performance evaluation (de Leeuw et al., 2023; Janssens & Bogaert, 2023), and team roster optimization (Sagi et al., 2024).

Despite this surge in interest, one large issue is the fact that currently no data is available about what happens during the races. This is unfortunate, as this information should ideally be included in all the mentioned applications to get an objective race and rider evaluation. For instance, in talent identification a rider who is attacking-oriented but met a lot of malfortune due to crashes and did not finish the race will not be identified by the current models. However, this rider can still be an interesting prospect for scouting purposes.

Other sports such as soccer saw a similar revolution occurring recently. Historically, soccer analytics was heavily focused on match outcome (i.e., goals), which resulted in the majority of research being directed towards the valuation of goal attempts (Decroos et al., 2019). This limited the reach of the studies, after which researchers started

collecting and sharing more fine-grained data sets about what happened during the match (Pappalardo et al., 2019). This new information allowed for totally new applications such as action evaluation (Decroos et al., 2019), player vectorizations (Decroos & Davis, 2019), pressing evaluation (Robberechts, 2019) and player chemistry evaluation (Bransen & Van Haaren, 2020).

Unfortunately, the situation is a bit more complicated in road cycling. First, computer vision-based approaches (Okuma et al., 2004) popular in team sports on a fixed playing field (e.g., hockey field, soccer field) with fixed cameras are much more difficult to succeed in road cycling races, where up to 200 participants ride very close to each other, in a landscape and camera angle which changes the entire time. This explains that attempts to develop a tracking technique (Verstockt et al., 2020) which was developed for both road cycling and cyclocross only matured into a methodology to create fine-grained data for cyclocross segments (De Bock & Verstockt, 2021). A second option, the usage of tracking spatio-temporal data using wearable devices, currently popular in soccer (Pappalardo et al., 2019), also did not result into a useful publicly available data source to date. Knowing the exact physical abilities of your opponents can result into large competitive gains in endurance sports such as cycling, which hinders many competitors from being willing to share the data from these devices.

Rather, current in-race data gathering approaches focus on the timing of when riders' devices pass by some hardware which is installed next to the race course (Kolaja & Ehlerova, 2019; Decorte et al., 2024). To install this hardware across all races, a serious investment would be required, while it also does not bear the potential to retrieve information about past races, which limits the retroactive abilities of this technique. Accordingly, this study sets out to explore an alternative in-race data collection methodology. A lot of online commentary on professional road cycling races is provided on the online platform X, formerly known as Twitter. This textual information bears a lot of potential to capture in-race dynamics and events.

We deploy popular large language models (i.e., GPT-3.5-Turbo and GPT-4o) and demonstrate that these models can be used to retrieve structured information about who is positioned where during the race after the usage of few-shot learning. The results are promising, but only after some manual cleaning. GPT-4.o is demonstrated to outperform its predecessor on our task, but there is an additional cost associated to this when deployed on a large sample of races. Finally, we show that representing the uncleaned large language model output as a graph combined with node representation learning is demonstrated as an interesting avenue to automate in-race rider representation which might be useful for downstream tasks.

## 2 Literature Review

The application of data mining and machine learning in road cycling has remarkably increased in recent years. While the range of applications has grown, they have been focused on a relatively narrow list of data sources, as depicted in Table 1. Historically, two types of data have been used: race results databases and sensor data. Sensor data is typically utilized when an application is developed for a single team, in collaboration with a team. Recent examples are the automation of rider selection per race at Israel –

Premier Tech (Sagi et al., 2024), and the pre-race estimation of a single rider's race form at Team Jumbo-Visma (de Leeuw et al., 2023). These applications work very well when a single team is involved, but are difficult to deploy across multiple teams as many teams are not willing to share their sensor data to other teams, which renders applications such as talent identification or race outcome estimation infeasible with sensor data, as one would need the sensor data for practically every participating rider. Rather, these studies use race results databases. These databases, such as ProCyclingStats<sup>1</sup> or CQ ranking<sup>2</sup>, typically contain detailed information about the race results, as well as some limited structured information about the course (e.g., total distance, total elevation gain), with the recent addition (Janssens & Bogaert, 2023; De Bock & Verstockt, 2023) of detailed course information into the applications. However, just looking at the end result in a sport with races taking up to 6-7 hours, while race tactics heavily influence race outcome, entails an unavoidable loss of information. Current applications give no value to valuable team mate contributions, dangerous attacks, or rider misfortune (e.g., crashes or flat tires). Up until today, only one study used in-race information. When building a methodology to score how dangerous different road segments are to improve rider safety, De Bock and Verstockt (2023) use a database about in-race crashes, which is created through the manual input by stakeholders in the sport.

**Table 1.** Previous Studies in Cycling Analytics. Lack of in-race data has resulted in ignorance of this aspect in previous research. RRD = Race Results Databases; SD = Sensor Data; DCI = Detailed Course Information; IRI = In-Race Information.

Study	Application	RRD	SD	DCI	IRI
Hilmkil et al. (2018)	Heart Rate Prediction		x		
Kataoka & Gray (2018)	Power Output Prediction		x		
De Spiegeleer (2019)	Race Outcome Estimation	x			
Karetnikov (2019)	Power Output Prediction	x	x		
Kholkine et al. (2021)	Race Outcome Estimation	x			
Van Bulck et al. (2023)	Talent Identification	x			
Janssens, Bogaert & Maton (2023)	Talent Identification	x			
Baron, Janssens & Bogaert (2023)	Rider Representation	x			
de Leeuw et al. (2023)	Form Estimation		x		
Kholkine et al. (2023)	Peak Age Estimation	x			
De Bock & Verstockt (2023)	Crash Prevention			x	x
Janssens & Bogaert (2023)	Performance Evaluation	x		x	
Sagi et al. (2024)	Team Roster Optimization	x	x		
de Leeuw & Kholkine (2024)	Peak Age Estimation	x			

Such a manual database is a necessity, as currently no public data set is available on in-race information. Current approaches (Kolaja & Ehlerova, 2019; Decorte et al., 2024) to capture in-race information need the deployment of hardware installation near the race course, and then register when each rider passes next to these installations to have

<sup>1</sup> <https://www.procyclingstats.com/>

<sup>2</sup> <https://cqranking.com/men/asp/gen/start.asp>

some in-race information. Such an approach cannot gather information about past historic races, while optical tracking (Verstockt et al., 2020) did not yield sufficient accuracy. Therefore, this study explores another avenue: the processing of unstructured textual race commentary.

### 3 Methodology

#### 3.1 Data

The goal of this study is to investigate whether it is feasible to structure Tweets (i.e., messages on X) into structured information which can aid current cycling analytics solutions. As a case study, we have selected the Dutch WorldTour Amstel Gold Race 2024. This race was selected as its course lends itself towards various attacks, which results into a dynamic race progression, while it is also as one of the more important one-day races on the calendar. This ensures sufficient coverage by the accounts discussing professional road cycling races on X.

A list of X accounts was created who frequently discuss road cycling races on the platform. From these accounts, all Tweets posted between the race starting (10:45:00 CET) and ending (16:43:17 CET) times were collected, resulting in a sample of 336 Tweets. While this sample would still be humanly annotatable, this would no longer be feasible if the process was re-iterated for the thousands of races reported in other studies (Janssens & Bogaert, 2023). Figure 1 gives an example of such a Tweet which clearly contains valuable information about the positioning during the race. Unfortunately, the information is in an unstructured format (i.e., text) which often differs across tweets. Moreover, this is even further complicated by the fact that that riders can be addressed through family names, given names, full names, and nicknames. Accordingly, a method needs to be formulated to re-structure this information into machine-readable formats and link the correct riders.

#### 3.2 Prompt Engineering

Large Language Models (LLMs) are quickly becoming the new standard in Natural Language Processing. These models are trained on billions of data points and parameters and are capable of capturing complex contextual information hidden in natural language and text. This pre-training on vast amounts of data makes them effective in various tasks without a lot of training data if the model is provided with some contextual intermediate reasoning steps (Wei et al., 2022), also called prompt engineering. This can be especially useful in our case, given the situation we encounter. For example, in Figure 1 the tweet mentions 11 unique riders from the starting list of 175 riders. This translates to a 175-class problem with only 336 observations, with the class options (i.e., participating riders) also changing from race to race. Moreover, the task is even further complicated by the fact that besides rider, we also need to retrieve the mentioned

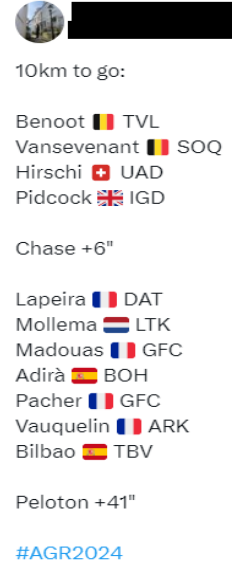


Fig. 1: Example Tweet

time gaps per rider, making it a multi-label task. Using more traditional fine-tuning machine learning approaches would require vast amounts of training data. Accordingly, we use a few-shot learning approach where we ask the model to act as a Tweet annotator (i.e. system role) for a cycling observation task: “You are a helpful assistant specialized in interpreting Tweets, which has to go through a large chunk of text on updates about a bicycling race.”. The final prompt (i.e. user role) is:

-----  
 “You are a helpful assistant specialized in interpreting Tweets. You are a huge cycling fan and want to get information about what happens during a cycling race, as currently you only have information about the results at the end of the race. To do so, you have to analyze a concatenated heap of text, and restructure this into information about who is positioned where. If no explicit information is given, you must return 'No Information'. Otherwise you should return a list with the mentioned rider(s) and their reported time gaps. Note that the time gaps can be both negative (chasing groups) as well as positive (leading group) and a distinction should be made. The Tweets are written in multiple languages, but your response should always follow the same English structure, regardless of the Tweet. If informative, ONLY return the list, do not explain your reasoning.

Some examples:

Input: 'The lead has an advantage of 1 minute 30.', rider list: *example\_rider\_list*.

Response: '['Unknown', '1:30'];

Input: 'Schmid fighting for his 15 second lead.', rider list: *example\_rider\_list*. Response: '['SCHMID Mauro', '0:15'];

Input: 'Oomen and Gregoire 30 seconds behind, while the peloton is also closing in.', rider list: *example\_rider\_list*. Response: '['GRÉGOIRE Romain', '-0:30'], ['OOMEN Sam', '-0:30'];

Input: 'The gap to the peloton has been reduced to just over 2 minutes now', rider list: *example\_rider\_list*. Response: '['Peloton', '-2:00'];

Input: 'Attack Oomen', rider list: *example\_rider\_list*. Response: 'No Information';

Input: 'Attack Oomen', rider list: *example\_rider\_list*. Response: 'No Information';

Now do the same for following input: *actual\_input* and rider list: *rider\_list*."

-----  
 With *actual\_input* being the actual Tweet text (i.e., 336 unique values), *rider\_list* being a textual list representing the starting list of riders (and ‘Peloton’) at the Amstel Gold Race (known upfront, retrieved from ProCyclingStats.com), and *example\_rider\_list* being a random list of 10 riders (and ‘Peloton’), which includes the riders included in the responses (e.g., Sam Oomen). The few-shot approach allows the model to identify different structures of giving information about the race update, how to perform named entity recognition from a list of starting riders (which differs from the new task), what to do when no riders can be identified, and what to do when no race situation information is provided. This approach proved to be superior over other approaches during preliminary testing which aggregated tweets per time bin, used zero shot learning, and other prompts. Two popular OpenAI models (i.e., GPT-3.5-turbo and GPT-4o)

are evaluated against each other. As no ground truth is available, we will compare the stability and known hallucinations of both approaches.

## 4 Results

An important first observation to make is the large difference in sample size generated by both methods. GPT-3.5-turbo generates 723 rider-timing tuples, while GPT-4o only creates 279 of these tuples. This is even more surprising as the number of Tweets which did not return ‘No Information’ or ‘Unknown’ is much more similar (i.e., 85 and 82, respectively). This means that GPT-3.5-turbo is either capable of retrieving more information per Tweet, or that it hallucinates more. Moreover, only 50 of these Tweets overlap for both methods (i.e., only about 60% of each set identified as informative). This already indicates large differences between both method outcomes. To identify differences, we inspect the first five overlapping Tweets (Tweets 1-5), the first five only identified by GPT-3.5-turbo (Tweets 6-10), and the first five only identified GPT-4o (Tweets 11-15) in Table 2. The table already shows clear issues with the results. A first issue is to also include women’s race results (which happens around the same time) despite the fact that these riders are not included in the *rider\_list* variable. Moreover, both models create false information. GPT-3.5-turbo has the tendency to ‘hallucinate’ large rider lists when no exact information is provided, explaining the larger list of tuples. Interestingly, this list often contains information about riders who are actually involved in race dynamics but not mentioned in the original tweet. For example, the twelve rider list from Tweet 4 (i.e., ['PIDCOCK Thomas', 'HIRSCHI Marc', 'BENOOT Tiesj', 'VANSEVENANT Mauri', 'LAPEIRA Paul', 'MADOUAS Valentin', 'MOLLEMA Bauke', 'PACHER Quentin', 'BILBAO Pello', 'MATTHEWS Michael', 'VAN DEN BERG Marijn', 'ADRIÀ Roger']) actually did contain the riders who were riding at the front, despite the fact that the Tweet itself does not contain this information. Rather, the model probably infers this from earlier calls by the model. For example, Tweet 7 generates a random list of 155 rider (+ peloton), and Tweet 8 generates the actual male top-10 (which was not known at time of posting while Tweets are fed chronologically), while the Tweet actually included an embedded picture of the women’s race top-10. This has two main implications. First, too much noise is induced in this application to use the generated output on the single observation level. Second, it highlights an understudied problem in current studies on zero- and few-shot learning studies, that the model learns from earlier calls, which may lead to data leakage when different prompts are compared. The output for GPT-4o is of higher quality, but several issues remain, such as the inclusion of women’s race results, instability due the indication of time differences both relative to the front group (e.g., ['PIDCOCK Thomas', '0:00'] in Tweet 4 and ['Peloton', '-3:00'] in Tweet 2) as well as relative to peloton (e.g., ['BENOOT Tiesj', '0:30'] in Tweet 1, and the lack of adhering to the required structure at times (e.g., Tweets 5 and 10). While these might be processed after extensive data cleaning, we are less confident that this process can be done in a high quality way when deployed on hundreds or thousands of races, as other issues might be surfacing. This is especially worrisome given the large costs associated with using GPT-4o. At time of writing, GPT-4o is billed at a price which is ten times higher than the price of GPT-3.5-

turbo, which in our study resulted in a cost of around \$ 4 for a call deployed to all 336 individual Tweets for GPT-4o. This could amount to a very high cost when deployed on a large set of races while still being of questionable quality.

**Table 2.** Differences between GPT-3.5-turbo and GPT-4o. \* Denotes a large list of names with the number of riders reported between brackets with no time differences reported unless stated otherwise.

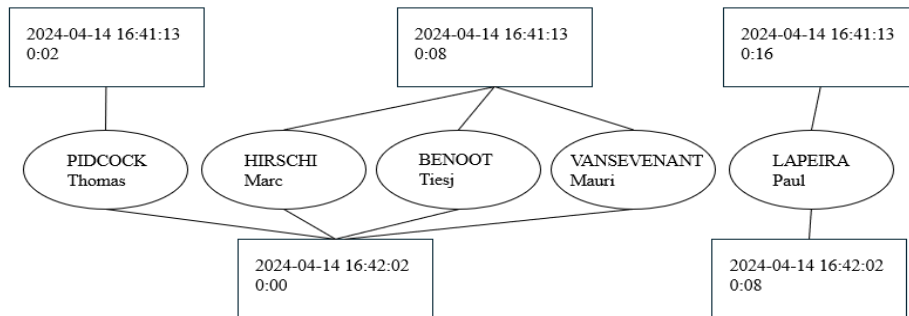
Id	Tweet	GPT-3.5-turbo	GPT-4o
1	NL #AGR2024 Tiesj is part of the 12-rider front group after Keutenberg. 🗣️ ⌚ 30" 🚩 27 km	['BENOOT Tiesj', '0:30']	['BENOOT Tiesj', '0:30']
2	NL #AGR2024   #TVLSlive 104KM Het peloton lijkt toch gas te geven. De laatste 20min hebben ze 0:48 van de voorsprong afgereden. De 4 hebben nog 3:00 over	['Peloton', '-0:48']	['Peloton', '-3:00']
3	NL #AGR2024 Tiesj is part of the 12-rider front group after Keutenberg. 🗣️ ⌚ 30" 🚩 27 km	['BENOOT Tiesj', '0:30']	[[['BENOOT Tiesj', '0:30']]]
4	Sigue en directo la Amstel Gold Race 2024: Tom Pidcock, Pello Bilbao y Roger Adriá, entre los 12 corredores al frente de la carrera a 26 km del final #Ciclismo	* (12)	[[['PIDCOCK Thomas', '0:00'], ['BILBAO Pello', '0:00'], ['ADRIÁ Roger', '0:00']]]
5	Amstel Gold Race donne: Vos beffa Wiebes	['VOS Marianne', 'WIEBES Lorena']	['VOS Marianne', 'WIEBES Lorena']
6	Auch die Gruppe mit Bilbao macht die Lücke zur Spitze wieder zu! Stark. #AGR2024	['BILBAO Pello', '0:00']	[[['BILBAO Pello', '0:00']]]
7	#AGR2024 - 27 km, ora in 12 in testa... 😊 🚴 #AGR #AGR24 #Ciclismo #Cyclisme #Cycling	* (156)	'No Information'
8	¡Así quedó el top 10 de una carrera que tuvo todos los condimentos! 😊 #AmstelGoldRace #StarPlusLA	* (10)	'No Information'
9	RT @Eurosport_FR: La cruauté de la photo-finish : Cosnefroy annoncé vainqueur en 2022... avant l'énorme désillusion et le triomphe de Kwiat...	['COSNEFROY Benoît', '0:00']	'No Information'
10	40 Sekunden – die Gruppe hat nun richtig gute Karten. #AGR2024	['Peloton', '0:40']	['Unknown', '0:40']
11	VIDEO: Dramatic finale to Amstel Gold Race Ladies 2024 as Marianne Vos punishes Lorena Wiebes' premature celebration #Cycling	['No Information']	['VOS Marianne', 'WIEBES Lorena']
12	Tiesj Benoot valt aan op de Keutenberg en Mathieu van der Poel moet in de achtervolging! Nog 25 kilometer, volg de finale van de Amstel Gold Race in ons liveblog!	['BENOOT Tiesj', 'Unknown']	[[['BENOOT Tiesj', '0:00'], ['VAN DER POEL Mathieu', '-0:00']]]
13	¡Total white para #MVDP en la #AmstelGoldRace! ❤️ 🚴 Sintoniza toda la acción por la pantalla de #StarPlusLA	['No Information']	['VAN DER POEL Mathieu', '0:00']
14	RT @UCI_WWT: @marianne_vos wins the 10th édition of @Amstelgoldrace ! 🏆 🚴 #UCIWWT #AmstelGoldRace	['MARIANNE VOS', 'No Information']	['MARIANNE Vos', '0:00']
15	#cyclisme 🚴 #AmstelGoldRace NL 12 à l'avant : Hirschi, Honoré, Pidcock, Benoot, Vansevenant, Adria, Mollema, Bilbao, Lapeira, Vauquelin, Madouas et Pacher #velo	* (11, including 'No Information')	* (12, indicated at 0:00 time gap)

Rather, we suggest an approach to aggregate these faulty observations in an automated manner, which can be used as an overall representation of rider behavior during

the race through graph representation learning. We create a bipartite graph based upon the structured data generated by GPT-3.5-turbo, with the two node types being ‘reported group timing’ and ‘rider’. Consider Table 3 and Figure 2 to make this clear. Table 3 visualizes our data after re-structuring it to a traditional table format, with for each rider the time gaps, as well as the original posting of the Tweet. The data is faulty, as the top-4 riders were actually all riding together with an 8 second gap on the group with Lapeira at the time. However, our graph in Figure 2 shows how this can be ‘smoothed’ using graph structures. The 16:41:13 observation in Table 3 does not link Pidcock to his three fellow attackers and it will be very hard to identify on a large scale when an LLM makes these types of mistakes. However, as long as there is any other time of posting / gap reported together (i.e., 16:42:02), the overall graph structure will link riders together who are often reported in the same group, which should theoretically give smoothed information about who exhibits similar riding behavior during the race.

**Table 3.** Example of Structured Information.

Time of Posting	Rider Name	Gap
2024-04-14 16:41:13	PIDCOCK Thomas	0:02
2024-04-14 16:41:13	HIRSCHI Marc	0:08
2024-04-14 16:41:13	BENOOT Tiesj	0:08
2024-04-14 16:41:13	VANSEVENANT Mauri	0:08
2024-04-14 16:41:13	LAPEIRA Paul	0:16
2024-04-14 16:42:02	PIDCOCK Thomas	0:00
2024-04-14 16:42:02	HIRSCHI Marc	0:00
2024-04-14 16:42:02	BENOOT Tiesj	0:00
2024-04-14 16:42:02	VANSEVENANT Mauri	0:00
2024-04-14 16:42:02	LAPEIRA Paul	0:08



**Fig. 2:** Graph derived from Table 3. If any other timing is reported together, then Pidcock will still be placed together with the other attacking riders.

To capture this information in a machine-readable format which can easily be used in other tasks, we deploy Node2Vec (Grover & Leskovec, 2016). This algorithm creates random sequences of nodes by ‘walking’ along the edges after which the skip-gram Word2Vec algorithm is deployed to learn the contextual embedding of each node in the network. We create 20 sequences (i.e., ‘random walks’) per rider in the network, each

of length 10, window size is set to 10, and the dimensionality of the resulting vectors (i.e., the hidden layer) is set to 32. This results into a 32-dimensional representation of each participating rider. To see whether each embedding is a good representation of in-race behavior, we check for each rider node what nodes have the most similar embeddings based on cosine similarity. Table 4 compares (1) the eventual winner Tom Pidcock, which placed a successful attack on the Keutenberg, (2) the pre-race favorite Mathieu van der Poel, (3) Paul Lapeira, who anticipated with an attack before the Keutenberg, and (4) Tosh Van der Sande, who was in the early breakaway. Each rider gets matched to the riders which displayed the most similar racing behavior. Tom Pidcock is matched to his three fellow attackers who rode away on the Keutenberg climb, but also to the random value >30. This is a consequence of the data quality issues stated above. Luckily, this is much easier to filter out based on the rider starting list. Similarly, Paul Lapeira is matched to four of the riders who also anticipated before the Keutenberg and ended up in the chasing group after the Keutenberg, while Tosh Van der Sande is matched to the three other riders who were also in the early break away, but also to Tomáš Kopecky, who was not actively involved in the breakaway, but interestingly joined the breakaway in many race leading up to this race, and finally race favorite Mathieu van der Poel is matched to four apparently random riders. This nicely reflects the fact that van der Poel had a rather anonymous race despite him ending in the second finishing group. These are all dynamics which would be difficult to capture based purely on eventual race results as currently is typically done.

**Table 4.** Most Similar Riders after Node2Vec

Base Rider	Most Similar Riders	
PIDCOCK Thomas	>30	HIRSCHI Marc
	VANSEVENANT Mauri	BENOOT Tiesj
VAN DER POEL Mathieu	SHAW James	OOMEN Sam
	DE POOTER Dries	CHZHAN Igor
LAPEIRA Paul	PACHER Quentin	MOLLEMA Bauke
	MADOUAS Valentin	MATTHEWS Michael
VAN DER SANDE Tosh	LEIJNSE Enzo	KYFFIN Zeb
	HAJEK Alexander	KOPECKÝ Tomáš

## 5 Conclusion

This study assesses whether it is feasible to use large language models to structure textual commentary on  $X$  into a useful and machine-readable structure. Results are disappointing with regard to the direct application of these models. GPT-4o outperforms GPT-3.5-turbo but the performance is too low to validate the currently large cost associated with using the model. Rather, we suggest a graph representation learning approach on the generated data which seems to capture complex in-race dynamics which are currently not captured in the used data sources, despite the errors made by the LLMs. This seems promising for future applications which want to acknowledge rider performances which are not observable in the end results. Future work should focus on how this information should be encapsulated in these applications and how information across races should ideally be captured. Moreover, we will expand upon additional

aces, which have some sort of in-race data being captured, which would allow for a detailed benchmarking of these methods. We will also investigate more deeply exactly what confuses the LLMs, whether some textual structures are more compatible than others, and whether this can be leveraged in a multi-stage methodology.

## References

1. Baron, E., Janssens, B., & Bogaert, M. (2023). Bike2Vec: Vector Embedding Representations of Road Cycling Riders and Races. *arXiv preprint arXiv:2305.10471*.
2. Bransen, L., & Van Haaren, J. (2020). Player chemistry: Striving for a perfectly balanced soccer team. *arXiv preprint arXiv:2003.01712*
3. De Bock, J., & Verstockt, S. (2021). Video-based analysis and reporting of riding behavior in cyclocross segments. *Sensors*, 21(22), 7619.
4. De Bock, J., & Verstockt, S. (2023). Road cycling safety scoring based on geospatial analysis, computer vision and machine learning. *Multimedia Tools and Applications*, 82(6), 8359-8380.
5. Decorte, R., De Bock, J., Taelman, J., Slembrouck, M., & Verstockt, S. (2024). Fully Automatic Camera for Personalized Highlight Generation in Sporting Events. *Sensors*, 24(3), 736.
6. Decroos, T., Bransen, L., Van Haaren, J., & Davis, J. (2019, July). Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 1851-1861).
7. Decroos, T., & Davis, J. (2019). Player vectors: Characterizing soccer players' playing style from match event streams. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 569-584). Springer, Cham.
8. de Leeuw, A. W., Heijboer, M., Verdonck, T., Knobbe, A., & Latré, S. (2023). Exploiting sensor data in professional road cycling: personalized data-driven approach for frequent fitness monitoring. *Data Mining and Knowledge Discovery*, 37(3), 1125-1153.
9. de Leeuw, A. W., & Kholkina, L. (2024). The relationships between age and race performance in women's road cycling. *International Journal of Performance Analysis in Sport*, 1-13.
10. De Spiegeleer, E. (2019). *Predicting cycling results using machine learning*. MA thesis. Ghent: Ghent University
11. Grover, A., & Leskovec, J. (2016, August). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 855-864).
12. Hilmkil, A., Ivarsson, O., Johansson, M., Kuylentierna, D., & van Erp, T. (2018). Towards machine learning on data from professional cyclists. *CoRR abs/1808.00198*.
13. Janssens, B., & Bogaert, M. (2022, April). Imputation of non-participated race results. In *Machine Learning and Data Mining for Sports Analytics: 8th International Workshop, MLSA 2021, Virtual Event, September 13, 2021, Revised Selected Papers* (pp. 155-166). Cham: Springer International Publishing.
14. Janssens, B., & Bogaert, M. (2023, September). Performance Measurement 2.0: Towards a Data-Driven Cyclist Specialization Evaluation. In *International Workshop on Machine Learning and Data Mining for Sports Analytics* (pp. 179-190). Cham: Springer Nature Switzerland.

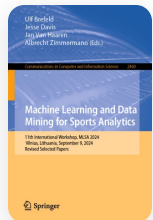
15. Janssens, B., Bogaert, M., & Maton, M. (2023). Predicting the next Pogačar: a data analytical approach to detect young professional cycling talents. *Annals of Operations Research*, 325(1), 557-588.
16. Kolaja, J., & Ehlerova, J. K. (2019, September). Effectivity of sports timing RFID system, field study. In *2019 IEEE International Conference on RFID Technology and Applications (RFID-TA)* (pp. 220-223). IEEE.
17. Kholkina, L., Latré, S., Verdonck, T., & de Leeuw, A. W. (2023). Age of peak performance in professional road cycling. *Journal of Sports Sciences*, 41(3), 298-306.
18. Kholkina, L., Schepper, T. D., Verdonck, T., & Latré, S. (2020, September). A machine learning approach for road cycling race performance prediction. In *International workshop on machine learning and data mining for sports analytics* (pp. 103-112). Springer, Cham.
19. Kholkina, L., Servotte, T., De Leeuw, A. W., De Schepper, T., Hellinckx, P., Verdonck, T., & Latré, S. (2021). A Learn-to-Rank Approach for Predicting Road Cycling Race Outcomes. *Frontiers in sports and active living*, 3.
20. Okuma, K., Taleghani, A., De Freitas, N., Little, J. J., & Lowe, D. G. (2004). A boosted particle filter: Multitarget detection and tracking. In *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I 8* (pp. 28-39). Springer Berlin Heidelberg.
21. Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., & Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. *Scientific Data*, 6(1), 236.
22. Robberechts, P. (2019). Valuing the art of pressing. In *Proceedings of the StatsBomb Innovation In Football Conference* (pp. 1-11). StatsBomb.
23. Sagi, M., Saldanha, P., Shani, G., & Moskovitch, R. (2024). Pro-cycling team cyclist assignment for an upcoming race. *PLoS one*, 19(3), e0297270.
24. Steyaert, M., De Bock, J., & Verstockt, S. (2022). Sensor-Based Performance Monitoring in Track Cycling. In *International Workshop on Machine Learning and Data Mining for Sports Analytics* (pp. 167-177). Springer, Cham.
25. Van Bulck, D., Vande Weghe, A., & Goossens, D. (2023). Result-based talent identification in road cycling: discovering the next Eddy Merckx. *Annals of operations research*, 1-18.
26. Verstockt, S., Van Vooren, B., De Smul, S., & De Bock, J. (2020). Data-driven summarization of broadcasted cycling races by automatic team and rider recognition. In *icSPORTS 2020, the 8th International Conference on Sport Sciences Research and Technology Support* (pp. 13-21). SCITEPRESS.
27. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35, 24824-24837.

Home > [Machine Learning and Data Mining for Sports Analytics](#) > Conference paper


# Large Language Models on Race Commentary: Towards Granular Data in Cycling Analytics

| Conference paper | First Online: 26 March 2025


| pp 14–25 | [Cite this conference paper](#)



**Machine Learning and Data Mining**  
**for Sports Analytics**  
(MLSA 2024)

[Bram Janssens](#) , [Matthias Bogaert](#) & [Steven Verstockt](#)


 Part of the book series: [Communications in Computer and Information Science](#)  
(CCIS, volume 2460)

 Included in the following conference series:  
[International Workshop on Machine Learning and Data Mining for Sports Analytics](#)

 69 Accesses

## Abstract

Current cycling analytics studies are limited to data about the eventual race results. This study searches how online commentary can be used to capture information about in-race dynamics by harnessing the power of large language models. The results show that the direct application of these models is already promising but not accurate enough to base end-to-end machine learning applications on the generated data. Our results show the tendency of these models to use information from previous queries in its generation step, which indicates data leakage and might hamper the scientific validation of approaches comparing various techniques. To capture overall rider behavior we suggest using graph representation learning. Our results indicate that this method is capable of identifying similar rider behavior, which to date was not yet feasible.

 This is a preview of subscription content, [log in via an institution](#)  to check access.

### Access this chapter

[Log in via an institution](#)

### Subscribe and save

Springer+ Basic

€32.70 /Month

Get 10 units per month

Download Article/Chapter or eBook

1 Unit = 1 Article or 1 Chapter

Cancel anytime

[Subscribe now](#) →

**Buy Now**

 [Chapter](#)

EUR 29.95

Price includes VAT (Belgium)

Available as PDF  
Read on any device  
Instant download  
Own it forever

Buy Chapter

^ eBook

EUR 50.28  
Price includes VAT (Belgium)

Available as EPUB and PDF  
Read on any device  
Instant download  
Own it forever

Buy eBook

^ Softcover Book

EUR 63.59  
Price includes VAT (Belgium)

Compact, lightweight edition  
Dispatched in 3 to 5 business days  
Free shipping worldwide - [see info](#)

Buy Softcover Book

Tax calculation will be finalised at checkout

Purchases are for personal use only

[Institutional subscriptions](#) →

## Notes

---

1. <https://www.procyclingstats.com/>.
2. <https://cqranking.com/men/asp/gen/start.asp>.

## References

---

Baron, E., Janssens, B., Bogaert, M.: Bike2Vec: Vector Embedding Representations of Road Cycling Riders and Races (2023). arXiv preprint [arXiv:2305.10471](https://arxiv.org/abs/2305.10471)

Bransen, L., Van Haaren, J.: Player chemistry: Striving for a perfectly balanced soccer team (2020). arXiv preprint [arXiv:2003.01712](https://arxiv.org/abs/2003.01712)

De Bock, J., Verstockt, S.: Video-based analysis and reporting of riding behavior in cyclocross segments. *Sensors* 21(22), 7619 (2021)

[ADS](#) [PubMed](#) [PubMed Central](#) [MATH](#) [Google Scholar](#)

De Bock, J., Verstockt, S.: Road cycling safety scoring based on geospatial analysis, computer vision and machine learning. *Multimedia Tools Appl.* 82(6), 8359–8380 (2023)

[MATH](#) [Google Scholar](#)

Decorte, R., De Bock, J., Taelman, J., Slembrouck, M., Verstockt, S.: Fully automatic camera for personalized highlight generation in sporting events. *Sensors* 24(3), 736 (2024)

[ADS](#) [PubMed](#) [PubMed Central](#) [Google Scholar](#)

Decroos, T., Bransen, L., Van Haaren, J., Davis, J.: Actions speak louder than goals: valuing player actions in soccer. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 1851–1861 (2019)

[Google Scholar](#)

Decroos, T., Davis, J.: Player vectors: Characterizing soccer players' playing style from match event streams. In: Joint European conference on machine learning and knowledge discovery in databases, pp. 569–584. Springer, Cham (2019)

[Google Scholar](#)

de Leeuw, A.W., Heijboer, M., Verdonck, T., Knobbe, A., Latré, S.: Exploiting sensor data in professional road cycling: personalized data-driven approach for frequent fitness monitoring. *Data Min. Knowl. Disc.* 37(3), 1125–1153 (2023)

[Google Scholar](#)

de Leeuw, A.W., Kholkina, L.: The relationships between age and race performance in women's road cycling. *Int. J. Perform. Anal. Sport* 1–13 (2024)

[Google Scholar](#)

De Spiegeleer, E.: Predicting cycling results using machine learning. MA thesis. Ghent University, Ghent (2019)

[MATH](#) [Google Scholar](#)

Grover, A., Leskovec, J.: node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864 (2016)

[Google Scholar](#)

Hilmkil, A., Ivarsson, O., Johansson, M., Kuylenstierna, D., van Erp, T.: Towards machine learning on data from professional cyclists. CoRR abs/1808.00198 (2018)

[Google Scholar](#)

Janssens, B., Bogaert, M.: Imputation of non-participated race results. In: Machine Learning and Data Mining for Sports Analytics: 8th International Workshop, MLSA 2021, Virtual Event, September 13, 2021, Revised Selected Papers, pp. 155–166. Cham: Springer International Publishing (2022)

[Google Scholar](#)

Janssens, B., Bogaert, M.: Performance measurement 2.0: towards a data-driven cyclist specialization evaluation. In: International Workshop on Machine Learning and Data Mining for Sports Analytics, pp. 179–190. Cham: Springer Nature Switzerland (2023)

[Google Scholar](#)

Janssens, B., Bogaert, M., Maton, M.: Predicting the next Pogačar: a data analytical approach to detect young professional cycling talents. Ann. Oper. Res. 325(1), 557–588 (2023)

[MathSciNet](#) [PubMed](#) [MATH](#) [Google Scholar](#)

Kataoka, Y., Gray, P.: Real-time power performance prediction in Tour de France. In: Brefeld, U., Davis, J., Van Haaren, J., Zimmermann, A. (eds) Machine Learning and Data

Mining for Sports Analytics. MLSA 2018. LNCS, vol 11330. Springer, Cham (2019).

[https://doi.org/10.1007/978-3-030-17274-9\\_10](https://doi.org/10.1007/978-3-030-17274-9_10)

Karetnikov, A.: Application of data-driven analytics on sport data from a professional bicycle racing team. Eindhoven University of Technology, The Netherlands (2019)

[Google Scholar](#)

Kolaja, J., Ehlerova, J.K.: Effectivity of sports timing RFID system, field study. In: 2019 IEEE International Conference on RFID Technology and Applications (RFID-TA), pp. 220–223. IEEE (2019)

[Google Scholar](#)

Kholkine, L., Latré, S., Verdonck, T., de Leeuw, A.W.: Age of peak performance in professional road cycling. *J. Sports Sci.* **41**(3), 298–306 (2023)

[PubMed](#) [MATH](#) [Google Scholar](#)

Kholkine, L., Schepper, T.D., Verdonck, T., Latré, S.: A machine learning approach for road cycling race performance prediction. In: International Workshop on Machine Learning and Data Mining for Sports Analytics, pp. 103–112. Springer, Cham (2020)

[Google Scholar](#)

Kholkine, L., et al.: A learn-to-rank approach for predicting road cycling race outcomes. *Front. Sports Act. Living* **3** (2021)

[Google Scholar](#)

Okuma, K., Taleghani, A., De Freitas, N., Little, J.J., Lowe, D.G.: A boosted particle filter: multitarget detection and tracking. In: Computer Vision–ECCV 2004: 8th European

Conference on Computer Vision, Prague, Czech Republic, May 11–14, 2004. Proceedings, Part I 8, pp. 28–39. Springer Berlin Heidelberg (2004)

[Google Scholar](#)

Pappalardo, L., et al.: A public data set of spatio-temporal match events in soccer competitions. *Sci. Data* 6(1), 236 (2019)

[PubMed](#) [PubMed Central](#) [MATH](#) [Google Scholar](#)

Robberechts, P.: Valuing the art of pressing. In: Proceedings of the StatsBomb Innovation in Football Conference, pp. 1–11. StatsBomb (2019)

[Google Scholar](#)

Sagi, M., Saldanha, P., Shani, G., Moskovitch, R.: Pro-cycling team cyclist assignment for an upcoming race. *PLoS ONE* 19(3), e0297270 (2024)

[CAS](#) [PubMed](#) [PubMed Central](#) [Google Scholar](#)

Steyaert, M., De Bock, J., Verstockt, S.: Sensor-based performance monitoring in track cycling. In: International Workshop on Machine Learning and Data Mining for Sports Analytics, pp. 167–177. Springer, Cham (2022)

[Google Scholar](#)

Van Bulck, D., Vande Weghe, A., Goossens, D.: Result-based talent identification in road cycling: discovering the next Eddy Merckx. *Ann. Oper. Res.* 1–18 (2023)

[Google Scholar](#)

Verstockt, S., Van Vooren, B., De Smul, S., De Bock, J.: Data-driven summarization of broadcasted cycling races by automatic team and rider recognition. In: icSPORTS 2020, the 8th International Conference on Sport Sciences Research and Technology Support, pp. 13–21. SCITEPRESS (2020)

[Google Scholar](#)

Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Process. Syst. **35**, 24824–24837 (2022)

[Google Scholar](#)

## Author information

---

### Authors and Affiliations

Department of Marketing, Innovation, and Organization, Ghent University,  
Tweakerkenstr. 2, 9000, Ghent, Belgium

Bram Janssens & Matthias Bogaert

FlandersMake@UGent-corelab CVAMO, Ghent, Belgium

Bram Janssens & Matthias Bogaert

Research Foundation Flanders, Brussel, Belgium

Bram Janssens

Ghent University - Imec, IDLab; Technologiepark-Zwijnaarde 122, 9052, Gent, Belgium

Steven Verstockt

### Corresponding author

Correspondence to [Bram Janssens](#) .

## Editor information

---

### Editors and Affiliations

Leuphana University, Lüneburg, Germany

Ulf Brefeld

Katholieke Universiteit Leuven, Leuven, Belgium

Jesse Davis

Katholieke Universiteit Leuven, Leuven, Belgium

Jan Van Haaren

Université de Caen Normandie, Caen Cedex 5, France

Albrecht Zimmermann

## Rights and permissions

---

[Reprints and permissions](#)

## Copyright information

---

© 2025 The Author(s), under exclusive license to Springer Nature Switzerland AG

## About this paper

---

### Cite this paper

Janssens, B., Bogaert, M., Verstockt, S. (2025). Large Language Models on Race Commentary: Towards Granular Data in Cycling Analytics. In: Brefeld, U., Davis, J., Van Haaren, J., Zimmermann, A. (eds) Machine Learning and Data Mining for Sports Analytics. MLSA 2024. Communications in Computer and Information Science, vol 2460. Springer, Cham. [https://doi.org/10.1007/978-3-031-86692-0\\_2](https://doi.org/10.1007/978-3-031-86692-0_2)

[.RIS](#) [.ENW](#) [.BIB](#)

DOI

[https://doi.org/10.1007/978-3-031-86692-0\\_2](https://doi.org/10.1007/978-3-031-86692-0_2)

Published

26 March 2025

Publisher Name

Springer, Cham

Print ISBN

978-3-031-86691-3

Online ISBN

978-3-031-86692-0

eBook Packages

Artificial Intelligence (R0)

## Publish with us

---

[Policies and ethics](#) 