Generating Factually Consistent Sport Highlights Narrations

Noah Sarfati
Amazon
Tel-Aviv, Israel
nsarfati@amazon.com

Ido Yerushalmy
Amazon
Tel-Aviv, Israel
idoy@amazon.com

Michael Chertok
Amazon
Tel-Aviv, Israel
chertokm@amazon.com

Yosi Keller
Amazon
Tel-Aviv, Israel
yoskelleamazon.com

ABSTRACT

Sports highlights are an important form of media for fans worldwide, as they provide short videos that capture key moments from games, often accompanied by the original commentaries of the game's announcers. However, traditional forms of presenting sports highlights have limitations in conveying the complexity and nuance of the game. In recent years, the use of Large Language Models (LLMs) for natural language generation has emerged and is a promising approach for generating narratives that can provide a more compelling and accessible viewing experience. In this paper, we propose an end-to-end solution to enhance the experience of watching sports highlights by automatically generating factually consistent narrations using LLMs and crowd noise extraction. Our solution involves several steps, including extracting the source of information from the live broadcast using a transcription model, prompt engineering, and comparing out-of-the-box models for consistency evaluation. We also propose a new dataset annotated on generated narratives from 143 Premier League plays and fine-tune a Natural Language Inference (NLI) model on it, achieving 92% precision. Furthermore, we extract crowd noise from the original video to create a more immersive and realistic viewing experience for sports fans by adapting speech enhancement SOTA models on a brand new dataset created from 155 Ligue 1 games.

CCS CONCEPTS

• **Computing methodologies → Natural language generation**; **Speech recognition**; *Discourse, dialogue and pragmatics.*

KEYWORDS

Large Language Models (LLMs), Hallucinations, Factual consistency evaluation, Prompt engineering, Natural Language Inference, Speech enhancing

MMSports '23, October 29, 2023, Ottawa, ON, Canada
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0269-3/23/10…$15.00
https://doi.org/10.1145/3606038.3616157

ACM Reference Format:
Noah Sarfati, Ido Yerushalmy, Michael Chertok, and Yosi Keller. 2023. Generating Factually Consistent Sport Highlights Narrations. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports (MMSports '23), October 29, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3606038.3616157

1 INTRODUCTION

Sports highlights have been a popular form of media for fans all around the world. Whether it's a favorite player scoring a game-winning goal, a critical moment in a championship game, or a jaw-dropping feat of athleticism, the excitement of watching a great play unfolds transcends borders and cultures. Sports highlights are short videos that showcase key moments from a particular game, often accompanied by the original commentaries of the game's announcers. They do not only serve as a way for fans to relive the excitement of a particular match but also as a recap of the game's most significant events. However, traditional forms of presenting sports highlights have limitations in conveying the complexity and nuance of the game. The selected clips in the highlight may not contain the moment where the announcers explain what happened, leaving out important context and details that are necessary for a comprehensive understanding of the game. Another limitation of using the original commentaries is that the announcer's voice may be truncated during the production of the highlights, resulting in an incoherent clip.

To provide a more complete and accessible viewing experience, broadcasters like Ligue 1 have hired professional narrators to develop their own narratives and commentators to read them during the sports highlight[1]. While these narratives can be effective, they are often time-consuming and expensive to produce and may not be available for all games or every moment of a game. For example, Amazon Prime Video introduced Rapid-Recap, a module that allows a viewer who joined late a live game to watch a quick highlight of what happened before they joined. The above-mentioned solution would be impossible to use in this case, as Rapid-Recap is dynamic, a new version is produced every few minutes, hence human narration becomes infeasible.

As a result, there has been a growing interest in the development of techniques for automatically generating natural language narratives of sports events [23, 42, 49, 52]. This work explores the use of Large Language Models (LLMs) as a promising approach, leveraging

[1] Here is an example of such a narrated highlight: https://www.youtube.com/watch?v=a3DbpAUV__Y

their ability to learn patterns and relationships from vast amounts of text data to generate human-like language output. However, LLMs can generate inaccurate or misleading narratives that are not consistent with the source of information or the prompt that was given to it. This phenomenon is called "hallucinations" [13, 16, 36, 51]. Ensuring accuracy in sports is crucial, particularly when it comes to addressing hallucinations, as viewers heavily depend on precise and factual information. Another important aspect of sports media is the role of crowd noise in creating a sense of excitement and atmosphere for the viewers. Mixing it with the generated narrative is important because it helps to create a more immersive and realistic viewing experience for sports fans. By harnessing LLMs, factual consistency evaluation and crowd noise extraction, we aim to overcome the limitations of traditional sports highlights and enhance the viewer's experience by creating automated and engaging narratives.

In this paper, we propose an end-to-end solution to enhance the experience of watching sports highlights by automatically generating natural language narratives using LLMs. Our solution involves several steps. First, one extracts the source of information from the sports live broadcast using a transcription model and extracts metadata messages describing what happened in each play. With this, we experienced with prompt engineering and propose one to capture the essence of the play. To ensure the consistency of the generated narratives, we propose a new dataset annotated on generated narratives and compare out-of-the-box models and a fine-tuned an NLI model on this dataset. Finally, to further enhance the viewer's experience, one extracts crowd noise from the original video using a dataset we created from Ligue 1, which includes both mixed audio and crowd noise only. The major contributions of the paper are:

- Propose a novel way of creating narratives for highlights by leveraging LLMs and prompt engineering.
- Analyzed the out-of-the-box consistency of the OpenAI GPT-3 model and propose a model trained on a new dataset for factual consistency evaluation.
- Applied state-of-the-art speech enhancing models to enhance noise on a dataset of crowd noise that we created.
- Designed and develop an ML-based pipeline to produce an end-to-end highlights narrated video, see Fig 1.

2 RELATED WORK

To address the limitations of traditional sports highlights, broadcasters have explored various techniques and approaches. One approach involves the use of professional narrators to develop narratives that accompany the highlights. Ligue 1, for instance, has employed this method to enhance the viewing experience of their sports highlights. However, this approach is resource-intensive and may not be feasible for all games or real-time highlights.

One promising approach that has emerged in recent years is the use of Large Language Models (LLMs) for natural language generation. LLMs are large neural networks that learn from vast amounts of text data patterns and relationships, and generate human-like language output. They have shown impressive performance in various natural language processing tasks, such as translation [14, 41], question-answering [30, 31], and text summarization [50]. In the context of sports highlights, LLMs have been trained to enhance sports narratives that are used during the game. In [37], the authors propose an approach to produce more natural-sounding narratives, overcoming the problem of rigidity of statistics that is hard to understand. Other methods use live commentaries, metadata, and players' information to generate sports news articles [44, 45].

However, LLMs can generate inaccurate or misleading narratives that are not consistent with the source of information, the prompt, that was given to it. This phenomenon is called "hallucinations" [13, 16, 36, 51]. This issue can be addressed by Factual Consistency Evaluation (FCE) [11], to detect the hallucinations automatically by predicting whether a generated text is factually consistent with respect to a grounding text. Several methods have been proposed [11] but we will focus on the task of Textual Entailment (ANLI) [6] or Natural Language Inference (NLI) [1] and Question Generation-Question Answering (QG-QA) [8, 12, 43] as these methods proved to be the best for text summarization [11]. NLI methods takes an hypothesis and a premise and determine whether the hypothesis is entailed by the premise, contradicted by it or is neutral w.r.t to it and one use the entailment probability as a factual consistency score. In QG-QA methods, one use a Name-Entity-Recognition model (NER) to generate spans on the premise which will led to generated questions that will be asked to the hypothesis, then the answer is compared to the span.

Another important aspect of sports media is the role of crowd noise in creating a sense of excitement and atmosphere for the viewers. Mixing it with the generated narrative is important because it helps to create a more immersive and realistic viewing experience for sports fans. In the literature, the classical approach is the inverse of this task, aiming to enhance speech signals captured under noisy or degraded conditions [5, 7, 24]. Here we would like to enhance the background noise.

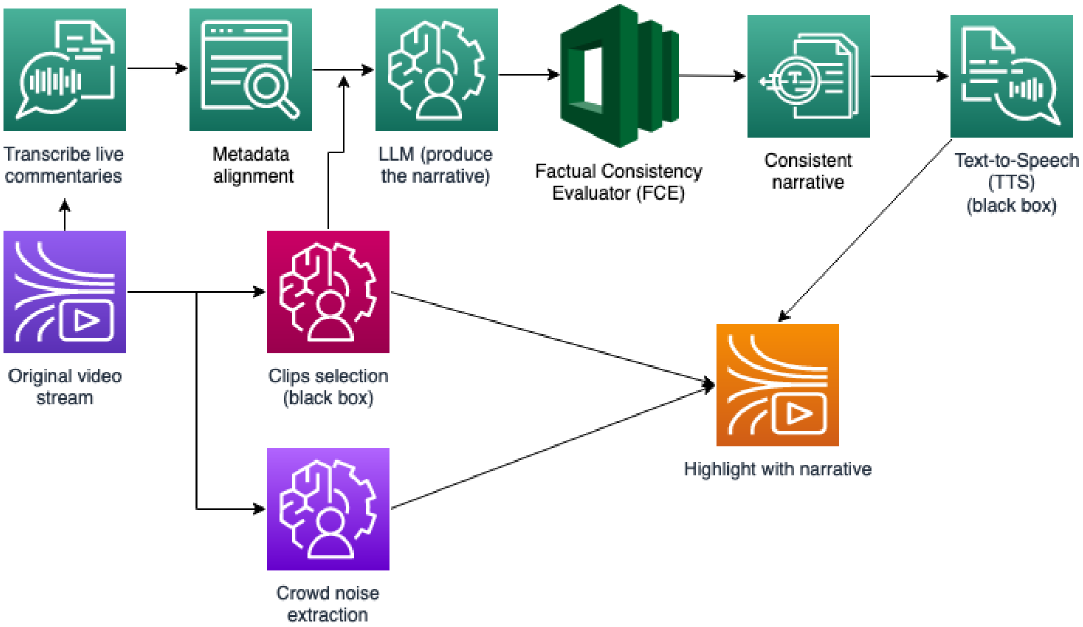
3 GENERATING THE NARRATIVE

In this section we describe how we generate the narrative, i.e how we create the prompt for the LLM and which model we chose. We make the assumption that we are getting for a given game an already created highlight composed of several clips. We are using metadata from a third party provider, each clip is coming with its associated metadata ID. The metadata fields we are interested in are the exact moment when the play occurred, the type of play and a message describing it.

3.1 The Source of Information

To generate the narrative via a LLM, we need to feed it with some sources of information. We mentioned earlier that given a clip in a highlight we can't be sure the announcers will be describing what is happening in it specifically, but we can make a robust assumption that they will do so at some point during or after the moment it occurs. For each play that occurred at a precise moment t_p, we capture the commentaries that happened around it $[t_p - \delta_b; t_p + \delta_a]$. To do so we transcribe the game using Whisper [28] which yields state-of-the-art results on Word-Error-Rate (WER) with Zero-shot Evaluation. t_p is coming from the metadata we obtained, we also get access to a message describing the play, we take it as our second source of information.

Figure 1: Pipeline to produce a highlight with narratives. The clips selections and Text-to-Speech parts are treated as black box, we make the assumption that they are provided.


Transcribe live commentaries
Metadata alignment
LLM (produce the narrative)
Factual Consistency Evaluator (FCE)
Consistent narrative
Text-to-Speech (TTS) (black box)
Original video stream
Clips selection (black box)
Highlight with narrative
Crowd noise extraction

3.2 LLM choice

Given a prompt (see subsection 3.3) we compared several state of the art LLM models. We report the results in Appendix A.1. Only GPT-3 [2] gave acceptable results. Models like T5 [29] were trained with given tasks such as summarization and translation and won't be good with open prompts, nevertheless they are very good for transfer learning. Hence we decided to pursue with the pre-trained GPT-3 model from OpenAI.

3.3 Prompt engineering

3.3.1 Definition. Prompt engineering [20] is the process of designing and optimizing prompts for language models such as GPT-3 [2], with the goal of improving their ability to generate high-quality and relevant responses to user inputs. Indeed, despite the impressive performances of LLMs, these models still suffer from various limitations, including poor output coherence, lack of context awareness, and the tendency to generate biased or offensive content.

To address these issues, prompt engineering has emerged as a promising approach to fine-tune language models for specific applications or domains. It involves carefully crafting prompts that provide relevant context and constraints for the model, such as input-output examples [2] or chain of thoughts [46], natural language descriptions, or structured data formats. By providing more explicit guidance to the model, prompt engineering can help it learn more effective representations of language and reduce the likelihood of generating irrelevant or harmful content.

In addition, prompt engineering can also help improve the interpretability and trustworthiness of language models. By designing prompts that require the model to provide explanations or justifications for its outputs, or by including human feedback in the training process [38], prompt engineering can enable more transparent and accountable AI systems. This is particularly important in applications such as healthcare, finance legal domains or sports where the consequences of incorrect or biased predictions can be severe.

3.3.2 Methodology. In this work we are interested in creating a narrative and not a summary of a given play. The nuance is very important as if we provide the user with a summary they may receive raw information, producing a sub-optimal result, see table 1 for a comparison. When an LLM is prompted to provide a narrative, it may lead to hallucinations as by definition it would need to narrate a story, potentially including non factual-consistent information, see section 4. Another issue coming from taking commentaries as a source of information is they can include unrelated elements to the play we are focusing on. To alleviate this, we can guide the LLM so that it knows what to speak about. In [18], the authors proposed a new framework called *Directional Stimulus Prompting* to learn a hint and proved that helping the LLM by providing it helped increased metrics for summarization. As we don't have labeled data, we take inspiration from this work and use the metadata we have access to, to provide the stimulus. In table 2 we show an example why this context is important and lead to better empirical results. The final prompt is of the form: *Knowing this context:\n <metadata message> \n Write a two-sentence narrative for the following information:\n<information>\n*

4 FACTUAL CONSISTENCY EVALUATION

As mentioned in Section 1, LLMs are prone to hallucinations. Honovich Et al. [11] introduced a comprehensive survey and assessment of factual consistency metrics on a standardized collection

Table 1: Comparison of GPT-3 output when asking to narrate or summarize. As information we used the commentaries with $\delta_b = 5$s and $\delta_a = 60$s.

prompt	Output
Write a two-sentence narrative for the following information:\n <information> \n	Willian took a stunning free kick, curling it around the wall and into the back of the net off the post. With this goal, Chelsea have equalized and now look to go on and win it with plenty of time left in the game.
Write a two-sentence summary for the following information:\n <information> \n	Willian scored an amazing free kick for Chelsea to level the score at 2-2. Both teams now have a chance to win, but Chelsea must focus on keeping their defense strong.

Table 2: Comparison of GPT-3 output when adding a context to the information. As information we used the commentaries with $\delta_b = 5$s and $\delta_a = 60$s. The information is: ***Here's Alonso. And Kante's going to have a go. Deflection, 3-1. Big slice of luck about the goal [...] Now, Willian is going to be able to bounce on this, and he was fouled by Maia Yoshida. Clumsy defending from both teams today.***
And the metadata message is:
Goal! Southampton 1, Chelsea 3. N'Golo Kanté (Chelsea) right footed shot from outside the box to the bottom right corner. Assisted by Marcos Alonso.

prompt	Output
Write a two-sentence narrative for the following information:\n <information> \n	Kante's strike from the edge of the box took a lucky deflection, restoring Chelsea's two-goal buffer and giving Kante his second goal of the season. Willian was then fouled by Maia Yoshida as both teams displayed clumsy defending throughout the match.
Knowing this context:\n <metadata message> \n Write a two-sentence narrative for the following information:\n <information> \n	N'Golo Kante unleashed a powerful right-footed shot from outside the box that found its way into the bottom corner with a deflection, restoring Chelsea's two-goal lead and giving them an unassailable 3-1 advantage. Marcos Alonso provided the assist for Kante's goal, which was met with luck as well as skill to give him his second of the season.

of existing texts from diverse tasks, manually annotated for factual consistency. We keep the same definition, a text is said to be factually consistent w.r.t its source of information if all the factual information it conveys is consistent with the factual information conveyed by the source of information. But for our task the only new informations it can add are players names extension or abreviation and clubs nicknames. To allow this external knowledge, we replace all players name and clubs names or nicknames by their full name using the Levenstein distance with a roster of players, in both the source of information and the narrative at the moment of comparison. The Levenshtein distance between two strings a, b is given by $\mathrm{lev}_{a,b}(|a|, |b|)$ where:

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j)+1 \\ \mathrm{lev}_{a,b}(i,j-1)+1 \\ \mathrm{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases} \tag{1}$$

For our task we are mainly interested in abstractive summary results and [11] showed that large scale NLI and QG-QA based approaches perform well across multiple tasks and datasets, leading us to focus on theses methods.

From the England Premiere League we took 66 games, selected tier one events (goal, own goal, red card, second yellow card, penalty saved, penalty goal, penalty miss), generated 15 narratives for each of them, deleted the ones with more than 100 tokens and annotated their factual consistency. This led to a dataset of 143 plays with 13.7 narratives in average for each of them.

4.1 NLI for textual entailment

The task of Textual Entailment or NLI is to determine, given two sentences, a hypothesis and a premise, whether the hypothesis in entailed by the premise, contradicts it or is neutral w.r.t to it. The resemblance of NLI to factual consistency evaluation has led to utilizing NLI models for measuring factual consistency. We use DeBERTa-v3 [9] large[2] fine-tuned on the MultiNLI [48], Fever-NLI [25], Adversarial-NLI (ANLI) [26] , LingNLI [27] and WANLI [19] datasets, which comprise 885 242 NLI hypothesis-premise pairs. We use this model in a zero-shot classification format, abbreviated NLI OOD. We also fine-tune it on our annotated data for soccer. To get the consistency score, we apply a softmax function to the entailment and contradictions probabilities and consider the entailment final probability. We train it using a binary cross entropy loss. For the testing, we apply a 5-Fold cross validation strategy.

4.2 QG-QA approach

Factual consistency evaluation via Question Generation-Question Answer was proposed in [8, 43]. The process is as follow, first from the generated text (the narrative in our case), spans are generated from a name entity recognition model, we use spacy [10] library for this. These spans correspond to answers of potential questions that are generated by the QG model. Then we present the question to the grounding text or source of information using a QA model. Finally the spans and the answers are compared to get a score which will represent the consistency evaluation.

We reproduce the Q^2 metric introduced in [12] and adapt it as described in [11]. It is a QG-QA method that employs an NLI model to compare the two answers for each question. In the original Q^2 paper [12], the authors used a pre-trained T5-Base [29] trained on the SQuAD Dataset [31] for the QG model. Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage. Hence, originally this dataset was not intended for the QG task but more for the QA task, this is why there are no official benchmarks. In

[2] https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli

the mentioned model the authors tweaked the dataset giving to the model as input: *"answer: <answer> context: <context>"*. It leads good results but we find that when processing lengthy context the generated question is often with lower quality. Second, when an answer phase appears multiple times in the context, there is ambiguity when selecting which one to generate questions. This why [3] introduced the HLSQG method. For a given context paragraph $C = [c_1, .., c_{|C|}]$ and an answer phase $A \subset C$, $A = [a_1, ..., a_{|A|}]$, one creates a new text $C' = \left[c_1, .., [HL], a_1, ..., a_{|A|}, [HL], ..., c_{|C|}\right]$. In C', one designs and inserts a new token $[HL]$ to indicate the answer phase in the context. The authors observed that the design of [HL] token helped avoiding possible ambiguities.

5 CROWD NOISE EXTRACTION

An important part for the user experience is to have the crowd noise as a background audio.

Given a discrete input waveform $x(t) \in \mathbb{R}^T$, the speech separation tasks is the estimation of C sources $s_1(t), ..., s_c(t) \in \mathbb{R}^T$, where:

$$x(t) = \sum_{i=1}^{C} s_i(t) \tag{2}$$

We would like to estimate the C different sources directly from the waveform $x(t)$. In this work we want to separate the input waveform into two sources, the commentators and the background crowd/stadium noise:

$$x(t) = s(t) + n(t) \tag{3}$$

With $s(t) \in \mathbb{R}^T$ the waveform associated to the commentators and $n(t) \in \mathbb{R}^T$ the one associated to the noise. This task is not new, but to the best of our knowledge most of the proposed solutions focus on the quality of the estimation of s and don't pay attention to n, it is called speech enhancement. Several challenges/dataset have been proposed [5, 7, 24].

Deep learning enabled a large improvement in end-to-end audio source separation in time-domain [4, 21, 22, 39, 40]. Like in natural language tasks and computer vision, most of the recent SOTA methods use increasingly complex architectures to achieve better speech enhancement [4, 39]. In [17], the authors proposed an efficient attention-based architecture TDANet achieving SOTA results, comparable to SepFormer [39], on Libri2Mix [5] and WHAM! [47], public datasets for speech separation in noisy environment. Thus, we decided to compare results from TDANet and SepFormer on our task.

6 EXPERIMENTAL RESULTS

6.1 Factual Consistency Evaluation

We report the results of several QG models in table 3 tested on the dev set of Squad, the suffix *hl* means that the the HLSQG method was used otherwise the method of Q^2 was. Based on these results we select the model T5-large-hl. For the QA model we use a pretrained Deberta-v3[3] on SQuAD v2 [30], an extension of SQuAD but questions without answer were added so that the model knows not to answer when he doesn't know. It helps if an information was invented by the generated narrative, indeed a question will be asked about it to the source of information and the QA model will not answer. For the NLI part we use the same model as described in section 4.1. We call the final model Q^2HL.

For the consistency evaluation task we evaluate the different models with two different metrics:

(1) For a given play, we rank all the generated narratives according to their factual consistency score, select the best and check whether or not it is consistent. We call this metric *rankedPrecision*
(2) For a given recall r we measure the precision. We denote it @Prec$_r$

Out of the 143 annotated plays, 2 of them didn't have at all consistent narratives among their generated narratives (13.7 in average). We report the results in table 4. As a baseline, we take the outputs of GPT-3 as they are. Among all the generated narratives 68.11% are consistent, and if we look on the play scale, then 68.30% are. As in theory we can generate as much narrative as we wish, we focus on the precision. We evaluate all the models with three different recalls, 0.25, 0.5, 0.75. As mentionned in [11] ensembling NLI OOD and Q^2HL improves the benchmark. For the rankedPrecision we simply sum the ranks, and for the @Prec we used a min-max normalization of score and average them.

We managed to improved considerably all the metrics, with a 91.72% rankedPrecision and 86.78% @Prec$_{0.25}$. But it is far from the results one we would like to reach for real life application. We only trained on 143 plays, based on the results we expect that augmenting the dataset to a decent size such as 1000 plays coming from a variety of games will help us reach better results. At the end, the real focus would be on @Prec$_{0.25}$, indeed knowing that GPT-3 gives us a consistent narrative with 68.11% chance, then if we had a 99% @Prec$_{0.25}$ we would need approximately 6 draws to obtain a desired output.

Table 3: Results on the dev set of Squad for question generation. The suffix *hl* means that the the HLSQG method was used otherwise the method of Q^2 was.

model name	BLEU-4 ↑	METEOR ↑	ROUGE-L ↑
T5-large-hl	**23.37**	**25.99**	**49.04**
T5-large	22.76	25.49	48.83
T5-base-hl	22.19	25.12	47.38
T5-base	21.64	24.69	46.93
Bart-base-hl	21.45	24.71	46.42

6.2 Crowd Noise Extraction

In the literature, the main metric for the task of source seperation or speech enhancement is the SI-SDR improvement [15] or SI-SDR$_i$, hence we pursue with it for our task. The SI-SNR for the estimated noise $\hat{n}$ and the original noise n is defined as:

$$\text{SI-SNR}(\hat{n}, n) = 10 \log_{10} \frac{||\mathbf{A}_{target}(\hat{n}, n)||}{\mathbf{e}_{noise}(\hat{n}, n)} \tag{4}$$

where:

$$\mathbf{A}_{target}(\hat{n}, n) = \frac{\langle \hat{n}, n \rangle n}{||n||_2^2}, \quad \mathbf{e}_{noise}(\hat{n}, n) = \hat{n} - \mathbf{A}_{target}(\hat{n}, n) \tag{5}$$

[3] https://huggingface.co/deepset/deberta-v3-large-squad2

Table 4: Results on factual consistency evaluation of our annotated soccer narratives.

model name	rankedPrecision ↑	@Prec$_{0.25}$ ↑	@Prec$_{0.5}$ ↑	@Prec$_{0.75}$ ↑
Baseline (random choice)	0.6830	0.6811	0.6811	0.6811
Q^2HL	0.7342	0.7620	0.7335	0.7131
NLI OOD	0.7972	0.8182	0.7603	0.7292
Ensemble OOD	0.8180	0.7744	0.7441	0.7223
NLI fine-tuned	**0.9172**	**0.8678**	**0.8483**	**0.8014**

Given the original audio x containing both the commentators and the background crowd/stadium noise, the SI-SDR$_i$ is defined as:

$$\text{SI-SNR}_i(\hat{n}, x, n) = \text{SI-SNR}(\hat{n}, n) - \text{SI-SNR}(x, n) \qquad (6)$$

In this work we adapt TDANet [17] and SepFormer [39] SOTA models on speech enhancements to focus on the noise extraction and provide a new baseline on the Libri2Mix train-360 16k [5], see results in table 6. The TDANet architecture, showed to be parameters efficient for the task of speech enhancement, reveals to also be for noise extraction even though in this case SepFormer achieves better results.

We work on a new Dataset called CrowdSoccer, composed of 155 games from the Ligue 1 to which we cut the audios into segments of 3 seconds and downscale the frequency to 16 kHZ. Each game has two audio channels, the mixture ($x(t)$) and the crowd noise only ($n(t)$) i.e the groundtruth. We finetuned the two models we trained on Libri2Mix. In this dataset TDANet achieves better results. The dataset is large in term of raw audios, approximately 280000, but they come from a very small set of games and then are correlated to each other which makes the domain very small. It can explain why SepFormer has trouble generalizing and why the SI-SDR$_i$ is much smaller than the one on the Libri2Mix dataset.

Automatically evaluating the quality of output audios, noises or speeches is not trivial, the best way remains subjective evaluation [34]. For speech enhancement quality evaluation, models were trained [35] to give metrics on audio quality based on frameworks such as P.808 [33] or P.835 [32]. To the best of our knowledge, such trained quality evaluators on speech removal or noise enhancement don't exist.

Table 5: Results of crowd extraction on Libri2Mix.

Model name	SI-SDR$_i$ ↑	Params (M) ↓
TDANet Large	11.37	**2.20**
Sepformer	**12.23**	25.60

Table 6: Results of crowd extraction on CrowdSoccer.

Model name	SI-SDR$_i$ ↑	Params (M) ↓
TDANet Large	**8.48**	**2.20**
Sepformer	8.05	25.60

7 CONCLUSION AND FUTURE WORK

We have presented an end-to-end solution to enhance the experience of watching sports highlights by automatically generating natural language narratives using Large Language Models (LLMs). Our approach involves extracting the source of information from the live broadcast, prompt engineering, and using a fine-tuned Natural Language Inference (NLI) model to ensure factual consistency. Additionally, we extracted crowd noise from the original video to further enhance the viewer's experience. Our proposed solution has several potential applications in the sports media industry, including the production of more comprehensive and accessible sports highlights, pre-game analyses, and post-game recaps. It can also be extended to other forms of media, such as news articles and podcasts. However, to make it to real life applications we need to have more a robust factual consistency evaluator and a better quality crowd noise extractor. To attain these goals we would need more annotated data and start a broader scale annotation. Additionally, we can explore the use of multi-modal data, such as incorporating video analysis to capture important visual cues during the game. Finally, we can evaluate the effectiveness of our approach through user studies and feedback to improve the overall user experience.

REFERENCES

[1] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[3] Ying-Hong Chan and Yao-Chung Fan. 2019. A recurrent BERT-based model for question generation. In *Proceedings of the 2nd workshop on machine reading for question answering*. 154–162.

[4] Jingjing Chen, Qirong Mao, and Dong Liu. 2020. Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation. *arXiv preprint arXiv:2007.13975* (2020).

[5] Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. 2020. Librimix: An open-source dataset for generalizable speech separation. *arXiv preprint arXiv:2005.11262* (2020).

[6] Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment: First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, April 11-13, 2005, Revised Selected Papers*. Springer, 177–190.

[7] Harishchandra Dubey, Vishak Gopal, Ross Cutler, Ashkan Aazami, Sergiy Matusevych, Sebastian Braun, Sefik Emre Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, et al. 2022. Icassp 2022 deep noise suppression challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 9271–9275.

[8] Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *arXiv preprint arXiv:2005.03754* (2020).

[9] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543* (2021).

[10] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[11] Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991* (2022).

[12] Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. *Q*{2}: Evaluating Factual Consistency in Knowledge-Grounded Dialogues via Question Generation and Question Answering. *arXiv preprint arXiv:2104.08202* (2021).

[13] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[14] Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

[15] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. 2019. SDR–half-baked or well done?. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 626–630.

[16] Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. (2018).

[17] Kai Li, Runxuan Yang, and Xiaolin Hu. 2022. An efficient encoder-decoder architecture with top-down attention for speech separation. *arXiv preprint arXiv:2209.15200* (2022).

[18] Zekun Li, Baolin Peng, Pengcheng He, Michel Galley, Jianfeng Gao, and Xifeng Yan. 2023. Guiding Large Language Models via Directional Stimulus Prompting. *arXiv preprint arXiv:2302.11520* (2023).

[19] Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. *arXiv preprint arXiv:2201.05955* (2022).

[20] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *Comput. Surveys* 55, 9 (2023), 1–35.

[21] Yi Luo, Zhuo Chen, and Takuya Yoshioka. 2020. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 46–50.

[22] Yi Luo and Nima Mesgarani. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing* 27, 8 (2019), 1256–1266.

[23] Xue-Qiang Lv, Xin-Dong You, Wen-Chao Wang, and Jian-She Zhou. 2020. Generate Football News from Live Webcast Scripts Based on Character-CNN with Five Strokes. *Journal of Computers* 31, 1 (2020), 232–241.

[24] Matthew Maciejewski, Gordon Wichern, Emmett McQuinn, and Jonathan Le Roux. 2020. WHAMR!: Noisy and reverberant single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 696–700.

[25] Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

[26] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial NLI: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599* (2019).

[27] Alicia Parrish, William Huang, Omar Agha, Soo-Hwan Lee, Nikita Nangia, Alex Warstadt, Karmanya Aggarwal, Emily Allaway, Tal Linzen, and Samuel R Bowman. 2021. Does Putting a Linguist in the Loop Improve NLU Data Collection? *arXiv preprint arXiv:2104.07179* (2021).

[28] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. https://doi.org/10.48550/ARXIV.2212.04356

[29] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[30] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822* (2018).

[31] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250* (2016).

[32] ITUT Rec. 2003. P. 835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. *International Telecommunication Union, Geneva* (2003).

[33] ITUT Rec. 2018. P. 808, Subjective evaluation of speech quality with a crowdsourcing approach. *ITU-T, Geneva* (2018).

[34] Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke. 2019. A scalable noisy speech dataset and online subjective test framework. *arXiv preprint arXiv:1909.08050* (2019).

[35] Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2022. DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 886–890.

[36] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156* (2018).

[37] Saman Sarraf, ML Amazon, and Arbi Tamrazian. [n. d.]. Sports Narrative Enhancement with Natural Language Generation. ([n. d.]).

[38] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems* 33 (2020), 3008–3021.

[39] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. 2021. Attention is all you need in speech separation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 21–25.

[40] Efthymios Tzinis, Zhepei Wang, and Paris Smaragdis. 2020. Sudo rm-rf: Efficient networks for universal audio source separation. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 1–6.

[41] David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting PaLM for Translation: Assessing Strategies and Performance. *arXiv preprint arXiv:2211.09102* (2022).

[42] Xiaojun Wan, Jianmin Zhang, Jin-ge Yao, and Tianming Wang. 2016. Overview of the NLPCC-ICCPOL 2016 shared task: sports news generation from live webcast scripts. In *Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24*. Springer, 870–875.

[43] Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228* (2020).

[44] Jiaan Wang, Zhixu Li, Tingyi Zhang, Duo Zheng, Jianfeng Qu, An Liu, Lei Zhao, and Zhigang Chen. 2022. Knowledge enhanced sports game summarization. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1045–1053.

[45] Jiaan Wang, Tingyi Zhang, and Haoxiang Shi. 2022. GOAL: Towards Benchmarking Few-Shot Sports Game Summarization. *arXiv preprint arXiv:2207.08635* (2022).

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903* (2022).

[47] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. 2019. Wham!: Extending speech separation to noisy environments. *arXiv preprint arXiv:1907.01160* (2019).

[48] Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426* (2017).

[49] Jin-ge Yao, Jianmin Zhang, Xiaojun Wan, and Jianguo Xiao. 2017. Content selection for real-time sports news construction from commentary texts. In *Proceedings of the 10th International Conference on Natural Language Generation*. 31–40.

[50] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848* (2023).

[51] Zheng Zhao, Shay B Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. *arXiv preprint arXiv:2009.13312* (2020).

[52] Liya Zhu, Wenchao Wang, Yujing Chen, Xueqiang Lv, and Jianshe Zhou. 2016. Research on summary sentences extraction oriented to live sports text. In *Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24*. Springer, 798–807.

A APPENDICES

A.1 LLM Comparisons on narratives

Model name	Output
prompt	Knowing this context: Goal! West Ham United 2, Chelsea 2. Willian (Chelsea) from a free kick with a right footed shot to the centre of the goal. Write a two-sentence narrative for the following information: Round the wall and off the post and in. Brilliant. What a stunning free kick. Can't stop scoring at the moment. Chelsea's number 10. It's absolutely brilliant, isn't it? I mean, he saw the tested Fabianski to say he was going nearside and Fabianski slow getting across, but it's perfectly executed in off the post. In the end, no chance for the goalkeeper. Chelsea back on level terms with, I don't want to say, plenty of time to get another. What a free kick that was. Beauty from Willian. Chelsea back at 2-2 will look to go on and win it, but must keep the back door shut. I don't intend to feel at this moment it's anybody's game now, isn't it? Most of them are going to fancy their chances because they've scored twice, but Chelsea must build on what they've done in the last five or six minutes, really.
OpenAI:GPT-3	Willian stepped up to take the free kick and curled a beautiful right-footed shot around the wall and off the post, leaving Fabianski with no chance. It was an absolutely stunning strike from Chelsea's number 10 that brought them back on level terms with plenty of time left to get another goal.
cohere:command-xlarge-nightly	Willian scored a stunning free kick to bring Chelsea back on level terms with West Ham United.
forefront:EleutherAI/GPT-J	So far we've had nothing. Redknapp: \n Perfectly alleged and a free kick a minute later and the referee has given West Ham the free kick. A minute after that Vydra had a headed effort off the goal line but it was really bl**dy u.ss blessed if the referee didn't wave my team back ontoside for me. West Ham back on level terms, although I didn't believe they were going to get back into the game at all. I'm shocked now. We were playing Wimbledon and Chelsea was really bad that day so it's not that relevant, really, but you can't deny they haven't looked good since that day. Chelsea don't look good. \n Hodgson: Dalglish:
forefront:pythia-20b	Most people want there to be five more minutes of action at the moment and at the end of the game so they can see the conclusion of the match. \n Didn't that kind of result happen in the conclusion of the second half against the Gunners. Osiek and his goal-scorer Samir Nasri will be pleased at the DM position. [...]
huggingface:bigscience/bloomz	Goal! West Ham United 2, Chelsea 2. Willian (Chelsea) from a free kick with
huggingface:google/flan-t5-xxl	Willian's free kick is absolutely stunning.
cohere:xlarge	The commentator was clearly very impressed with the quality of the free kick, but what was the goal-scorer's name? What is the position of the player in the team?
replicate:llama-13b	write