Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Résumé automatique de commentaires de football en direct avec grands modèles de langage

Aurélien Bossard[1] Christophe Rodrigues[2]
(1) Laboratoire d'Intelligence Artificielle et Sémantique des Données, Université Paris 8 (EA4383), 93200 Saint-Denis, France
(2) Léonard De Vinci Pôle Universitaire, Research Center, 92916 Paris La Défense, France
aurelien.bossard@iut.univ-paris8.fr, christophe.rodrigues@devinci.fr

RÉSUMÉ

Cet article traite du résumé de commentaires sportifs en direct. Il décrit un corpus de commentaires de matchs de football associés à un résumé manuel rédigé par un expert, tiré du site L'Équipe. Il propose différentes méthodes, certaines fondées sur les aspects afin de tester les capacités des grands modèles de langage à générer des résumés dans le style des experts. Les résultats montrent que les modèles récents arrivent à résumer efficacement les commentaires en direct. En revanche, ces résultats mettent en lumière la relative inadaptation des métriques d'évaluation au résumé génératif.

ABSTRACT

Automatic Summarization of Live Football Commentaries with Large Language Models

In this article, we address the task of summarizing live sports commentaries. We introduce a corpus of live comments from football matches along with their corresponding manual summaries collected from the L'Équipe website. We propose several methods, some of which are aspect-based, to evaluate the ability of recent LLMs to generate summaries in the same style as those written by experts. Results indicate that recent models can effectively summarize live comments, but also reveal the limitations of current evaluation metrics for generative summaries.

MOTS-CLÉS : Résumé automatique, grands modèles de langage, résumé fondé sur les aspects.

KEYWORDS: Automatic summarization, large language models, aspect based summarization.

1 Introduction

Certains sites sportifs proposent des commentaires en direct de matchs de football. Ces mêmes sites proposent parfois un résumé manuel du match, rédigé par le commentateur. Ces commentaires et leur résumé constituent un cas d'étude intéressant. D'une part, le besoin est réel : une aide au résumé voire une génération entièrement automatique de qualité libérerait au commentateur un temps précieux, qui pourrait être mieux investi dans des tâches analytiques. D'autre part, ce type de ressources – des textes associés à leur résumé écrit à la main – est coûteux et rare.

Dans cet article, nous nous intéressons donc au résumé automatique de commentaires en direct de matchs de football, tirés du site du journal L'Équipe [1]. Notre objectif est de résumer les commentaires en direct en conservant le style habituel de la rédaction du journal et en y incluant les informations

1. http://www.lequipe.fr

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 500–511.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

habituellement présentes dans les résumés.

Dans un premier temps, nous présentons les travaux connexes, puis les spécificités du corpus de commentaires sportifs en direct, avant de présenter les différents modèles, notre méthode d'évaluation puis nos résultats. Dans une dernière partie, nous abordons les limites des modèles utilisés.

2 Travaux connexes

Le résumé automatique de commentaires sportifs en direct est, à notre connaissance, traité de manière assez limitée dans la littérature. On peut citer l'approche de Bouayad-Agha *et al.* (2012) qui proposent un système de génération de résumé à base de règles qui s'appuient sur une ontologie. Le principal inconvénient d'un tel système est la nécessité de disposer ou de construire une ontologie peuplée et l'aspect stéréotypé des résumés générés par des règles ad hoc.

Corney *et al.* (2014) adoptent une approche radicalement différente, fondée non pas sur des commentaires en direct rédigés par un commentateur, mais par les tweets publiés en direct sur un match. On se situe alors dans le cas du résumé multidocument dans lequel la redondance joue un rôle important dans la sélection des informations à extraire dans le résumé. Les auteurs utilisent, après avoir détecté automatiquement les matchs auxquels se rapportent les tweets, une variante du tfidf afin d'extraire les tweets les plus représentatifs du match.

(Zhang *et al.*, 2016) résument des commentaires en ligne en direct en utilisant un système en trois étapes : la modélisation des phrases d'après les fréquences des mots et la présence d'indices de surface définis empiriquement et montrant l'importance d'une phrase dans le déroulé d'un match (but, expulsion, penalty accordé...), l'apprentissage de la prédiction du score ROUGE des phrases des commentaires en direct d'après cette modélisation, puis l'extraction des meilleures phrases dans le résumé. Ce système souffre de plusieurs limitations : la définition empirique d'indices de surface, la pénalisation de phrases courtes même fortement informatives, et la génération d'un résumé extractif qui, dans le cas des commentaires en direct, rend un texte qui diffère fortement de ce qui est attendu.

Aires (2016) propose un système de résumé « données vers texte » fondé sur les données structurées des matchs et des *templates* construits à partir d'articles rédigés par des journalistes. Des connaissances du domaine sont incorporées afin d'ajouter de la variabilité dans les résumés générés, mais ceux-ci restent tout de même stéréotypés. Les auteurs insistent sur le fait que si la qualité des résumés est globalement bonne, elle ne l'est pas assez pour qu'ils soient publiés sans intervention humaine a posteriori.

Huang *et al.* (2020) présentent SPORTSSUM, un corpus de 5 428 matchs de football contenant des commentaires en direct en chinois et les articles de presse correspondants pour étudier la génération automatique de résumés sportifs. Les auteurs proposent un modèle en deux étapes : la sélection des commentaires importants (résultat, score, buts) et un générateur séquence-à-séquence qui les transforme en texte journalistique. Contrairement à notre corpus, les textes à générer sont longs (24 phrases en moyenne), et les auteurs posent le problème des métriques d'évaluation traditionnelles (ROUGE et BERTScore par exemple), qui ne permettent pas de rendre compte de la factualité des textes générés. Wang *et al.* (2021) ont amélioré ce travail en nettoyant le corpus d'origine, et en proposant un algorithme de sélection des phrases, qui prend en compte la fluidité et la redondance

des résumés générés.

Wang *et al.* (2022) bâtissent sur les travaux de ces derniers en augmentant la taille du corpus, qui passe à 7854 matchs, et comblent le manque de connaissances extractibles depuis les commentaires et nécessaires à la génération d'un résumé grâce à un corpus de connaissances externe, et améliorent ainsi la qualité des résumés générés par les systèmes précédents.

Belemkoabga *et al.* (2021) introduisent un corpus de résumé automatique de commentaires en direct de matchs de football extrait du site L'Équipe. Après une analyse manuelle des aspects présents dans les résumés, ils extraient les commentaires importants à l'aide d'un SVM et génèrent le résumé à l'aide d'un *Pointer Generator* (See *et al.*, 2017). Cependant, les résumés générés manquent de qualité linguistique, et sont globalement extractifs, en raison du fonctionnement des systèmes à base de *Pointer Generator* alors que le corpus de résumé de commentaires en direct est plus adapté aux systèmes génératifs.

Proches du résumé de commentaires textuels en direct, Gautam *et al.* (2022) proposent un système de résumé de commentaires audio, et Sarkhoosh *et al.* (2024) du résumé multimodal.

Toutes ces études montrent la difficulté de la sélection d'informations propre au type de données étudié. Les méthodes fondées sur la fréquence sont inopérantes, et les auteurs se fondent donc sur des méthodes d'apprentissage automatique pour sélectionner les informations, en y apportant ou non des indices de surface définis empiriquement. L'autre difficulté inhérente aux commentaires sportifs en direct est que l'information présente dans les résumés est souvent absente des commentaires d'origine ou doit en être déduite.

Nous voulons tester la capacité des grands modèles de langage à sélectionner les informations pertinentes et à les inférer des textes d'origine quand elles sont disponibles, puis à générer un résumé concis dans le style de ceux du corpus. Pour cela, nous proposons plusieurs types de modèles, entièrement non guidés, guidés par une analyse préalable des données, ajustés sur le corpus et entièrement guidés par une analyse réalisée par un grand modèle de langage.

3 Corpus

Le corpus étudié dans cet article est composé des commentaires en direct des matchs de Ligue 1 de football de 2013 à 2025 récupérés depuis le site du journal L'Équipe. Celui-ci est la version du corpus introduit par Belemkoabga *et al.* (2021) et étendue avec les saisons de 2020 à 2025.

Le corpus compte 7200 matchs de football commentés en direct, chacun accompagné du résumé rédigé par le commentateur à l'issue du match. Cela constitue un véritable point fort de la ressource. En effet, le fait que les résumés aient été rédigé par un humain, qui plus est expert du domaine, garantit leur qualité. De plus, de telles ressources sont habituellement coûteuses à produire en grande quantité, et donc rares. Chaque match est composé d'un ensemble de commentaires ordonnés du plus récent au plus ancien, et accompagnés de la minute de l'action qu'ils décrivent.

Les matchs sont généralement structurés de la manière suivante :

1. Résumé du match ;
2. Commentaires de la seconde mi-temps ;
3. Compte-rendu de la première mi-temps ;
4. Commentaires de la première mi-temps ;

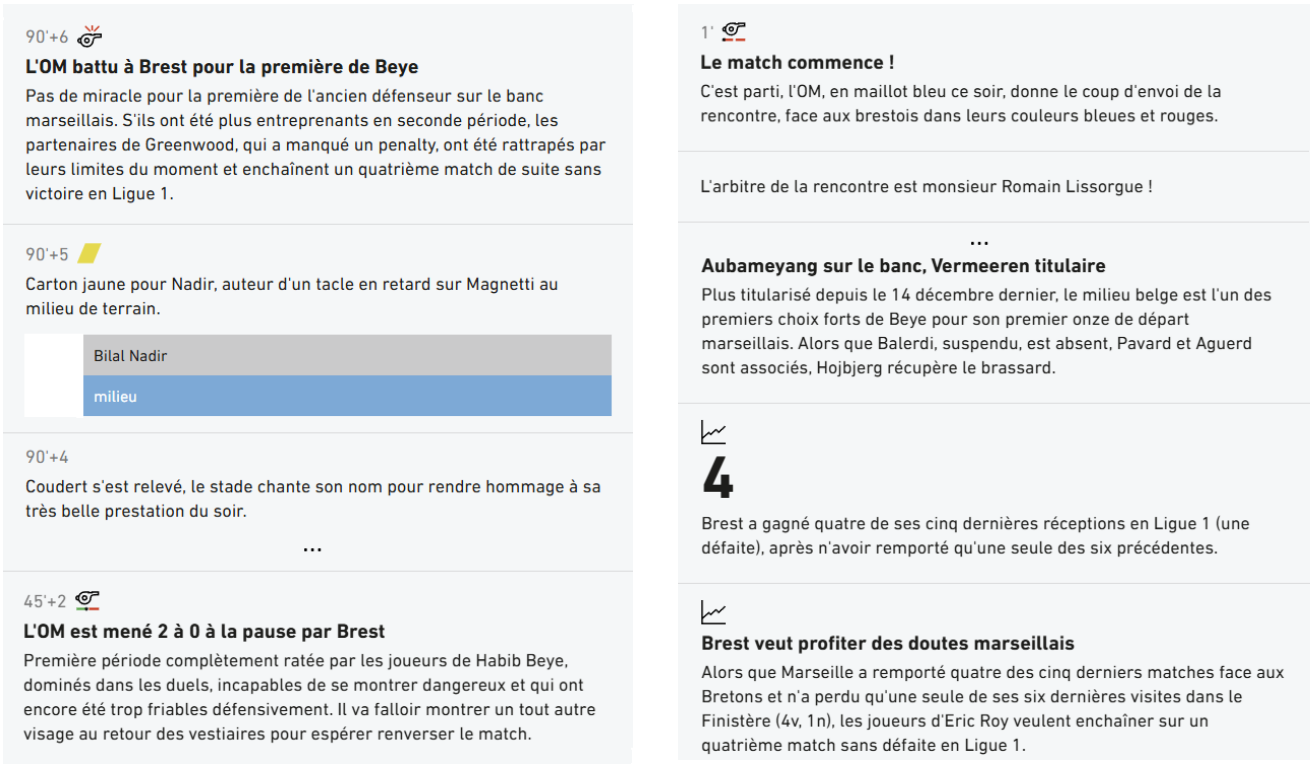

FIGURE 1 – Exemple de commentaires en direct tronqués associés à leur résumé. À gauche, le résumé suivi des derniers commentaires du match et du résumé à la mi-temps. À droite, le commentaire de début de match suivi des commentaires pré-match.
Source : https://www.lequipe.fr/Football/match-direct/ligue-1/2025-2026/brest-om-live/675995

5. Commentaires pré-match : météo, informations sur les deux équipes, informations marquantes sur des joueurs : absence, retour, point sur le classement et les conséquences du futur résultat ;

La figure 1 montre un exemple de commentaires de match où l'on voit clairement la structuration en cinq parties. Les commentaires ont une longueur moyenne de 2900 mots et les résumés 59 mots, soit un taux de compression moyen de 20.3%.

Les commentaires en direct de match présentent des difficultés inhérentes à leur forme quand il s'agit de les résumer. Premièrement, les systèmes de résumé s'appuient généralement sur des indices forts concernant l'emplacement des informations pertinentes des documents à synthétiser. Les deux types de documents les plus étudiés du domaine : les dépêches de presse et les articles scientifiques, ont une structuration de l'information très normée. La pyramide inversée pour le premier (les informations les plus importantes sont au début du document), une structure explicite et une structure rhétorique conventionnelle pour le second. Si les commentaires en direct suivent également une structure assez commune, les informations les plus essentielles du match, la victoire d'une équipe par exemple, doivent être inférées d'autres informations essentielles : les buts, qui peuvent apparaître à n'importe quelle position du document. De plus, les indices statistiques sont bien souvent non pertinents, comme l'ont montré Belemkoabga *et al.* (2021) sur ce même corpus (mais sur la période 2013-2020). Le tableau 1 présente les valeurs de nouveaux n-grammes (présents dans les résumés et absents des commentaires), de couverture et de densité sur la partie entraînement du corpus. Il montre bien son aspect abstractif : le lexique est emprunté à plus de 80% par les résumés, mais les segments longs, eux, ne sont pas extraits.

Ces deux phénomènes expliquent pourquoi les travaux sur le résumé automatique de commentaires

Novel-1	Novel-2	Novel-3	Novel-4	Couverture	Densité
0,189	0,628	0,829	0,917	0,811	2,026

TABLE 1 – Mesures de nouveaux n-grammes (Novel-1, Novel-2, Novel-3 et Novel-4), de couverture et de densité des résumés par rapport aux commentaires sur la partie entraînement du corpus L'Équipe.

de match se sont orientés vers l'extraction d'éléments clés, suivie d'une phase de génération. Ils ont guidé notre approche, que nous présentons dans la section suivante. Nous présentons en tableau 2 l'analyse manuelle réalisée par Belemkoabga *et al.* (2021), qui liste les aspects présents dans les résumés d'une saison entière de Ligue1 et leur pourcentage d'apparition. On constate que mis à part les aspects « Buteur », « Expulsion », « Penalty manqué », et « Penalty transformé », qui peuvent être extraits directement des commentaires, les autres informations nécessitent une inférence en plusieurs étapes. Par exemple, connaître le résultat requiert d'extraire les buts, pour quelle équipe ils ont été marqués, et de les additionner pour chaque équipe. Une telle tâche peut se révéler compliquée, y compris pour un humain.

Aspect	%
Résultat	80
Classement au championnat	55
Buteur	45
Domination de l'équipe	24
Série de victoires/défaites	22
Efficacité	19
Qualité 1re / 2e mi-temps	18
Qualité du match	18
Expulsion	16

Aspect	%
Fin d'une série de victoires/défaites	14
Penalty manqué	7
Match équilibré	5
Penalty transformé	4
Blessure	3
1er match depuis (retour)	3
Joueur absent	3
Coaching décisif	3

TABLE 2 – Pourcentage des résumés présentant un aspect particulier. Reprise de l'analyse présentée dans (Belemkoabga *et al.*, 2021).

4 Notre approche

Les nouveaux grands modèles de langage entraînés pour le raisonnement (OpenAI *et al.*, 2025; Grattafiori & Al, 2024; Liu & Al, 2026; Yang & Al, 2025) semblent être adaptés à la tâche de résumé de commentaires sportifs. En effet, l'étape de raisonnement pourrait permettre, même sans ingénierie de prompt, d'obtenir des résultats convaincants. Nous voulons tester ces modèles, avec un prompt naïf sans a priori sur les données, puis avec un prompt guidé par l'analyse des aspects des résumés.

Nous faisons l'hypothèse que les résultats de ces modèles pourraient être bons d'un point de vue informatif, mais ne pas correspondre au style de résumé habituellement utilisé par les commentateurs. Une étape de réglage fin pourrait remédier à ce problème.

Enfin, nous testons une approche entièrement fondée sur les LLM, dans laquelle nous réalisons l'analyse manuelle des aspects des résumés par grand modèle de langage.

La section suivante présente les différents modèles qui correspondent à chacune de ces hypothèses.

5 Modèles

Nous présentons ici les modèles mis en place pour tester les hypothèses présentées dans la section §4.

5.1 Modèle naïf

Ce modèle nous fournit à la fois une *baseline* et nous permet de tester les capacités des nouveaux modèles génératifs avec raisonnement à générer un résumé qui contient les informations attendues pour un résumé de commentaires sportifs en direct. Nous testons ici quatre grands modèles de langage libres avec le prompt naïf et minimaliste suivant, suivi de l'ensemble des commentaires en direct d'un match : "Produis un résumé de fin de match d'après le texte suivant : ".

Les quatre grands modèles de langage testés sont :
- Qwen3 14B (Yang & Al, 2025) ;
- GPT-OSS 20B (OpenAI *et al.*, 2025) ;
- Mistral3 8B Instruct 2512 (Liu & Al, 2026) ;
- Llama 3.2 3B (Grattafiori & Al, 2024).

5.2 Modèle guidé par une analyse manuelle des aspects

Ce modèle permet de tester l'hypothèse selon laquelle un prompt plus adapté aux données peut aider un LLM, même dôté d'une étape de raisonnement, à produire des résumés plus informatifs qu'un prompt naïf comme celui testé précédemment. Nous nous servons de l'analyse des aspects des résumés du corpus, réalisée par Belemkoabga *et al.* (2021) et dont les résultats sont présentés dans le tableau 2 afin d'écrire un nouveau prompt.

Après une analyse plus approfondie des aspects du tableau 2, nous avons constaté que le deuxième aspect le plus présent dans les résumés : le changement de classement en championnat induit par le résultat du match n'est quasiment jamais déductible des commentaires. En effet, s'il est parfois fait mention du classement des équipes avant le match, le classement après le match dépend à la fois du résultat du match, des concurrents et de l'écart avec ceux-ci avant le match. Sauf exception (un match avancé qui causera un changement provisoire au classement ou retardé et que les résultats des concurrents sont alors déjà connus), demander à un LLM une information non inférable du texte risquerait de provoquer des hallucinations préjudiciables à la qualité des résumés générés. Nous avons donc exclu cet aspect du prompt, et nous sommes concentrés arbitrairement sur les aspects présents dans plus de 10% des résumés.

La prompt qui en résulte est présenté en figure 2, suivi de l'ensemble des commentaires du match. Il a été testé sur les mêmes grands modèles de langage libres que le prompt naïf, listés en section §5.1.

5.3 Modèle ajusté sur les commentaires complets

Ce modèle consiste en l'ajustement de grands modèles de langage par entraînement à partir du corpus de commentaires sportifs. Plutôt que de mettre à jour entièrement les modèles, ce qui aurait un coût de calcul conséquent, nous adaptons les modèles de langage suivants à l'aide de LoRa (Hu *et al.*, 2022) dans leur version instructions :

Produis un résumé de fin de match en français en incluant prioritairement et sans structurer explicitement l'information l'issue du match, les buteurs, si le match est dominé par une équipe ou équilibré, la série de victoires ou de défaites en cours si elle existe, l'efficacité des équipes, la qualité du match et des mi-temps, les éventuelles expulsions d'après le texte suivant :

FIGURE 2 – Prompt rédigé d'après les analyses en aspect des résumés manuels

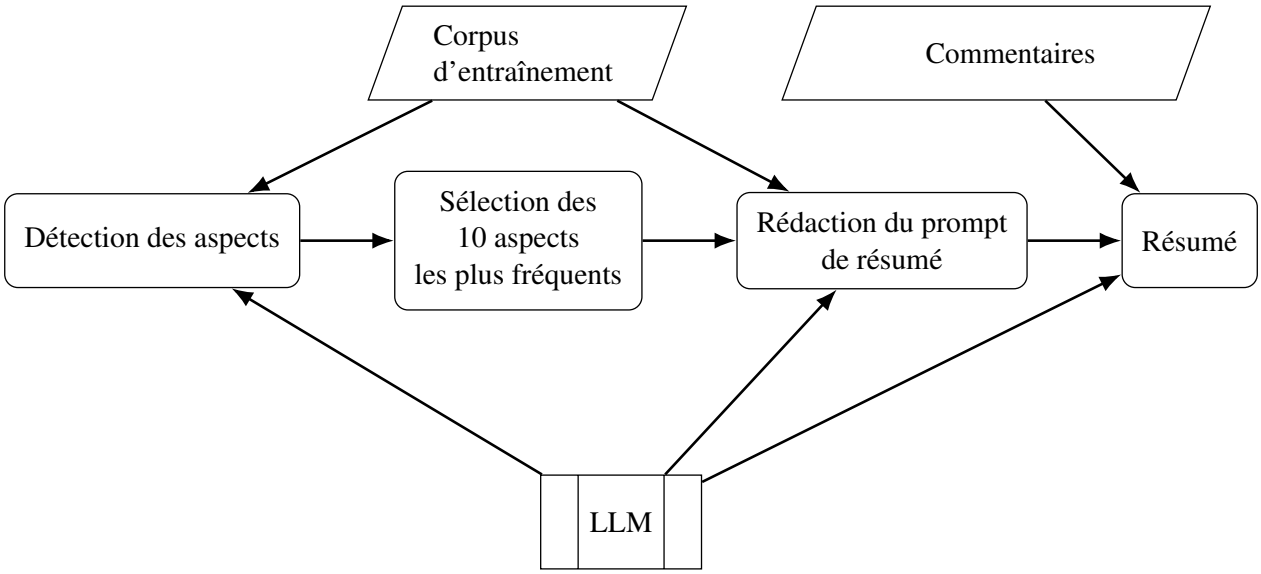

FIGURE 3 – Pipeline de génération de résumés guidé par les aspects entièrement guidée par un LLM

— Qwen3-14B ;
— GPT-OSS-20B ;
— Mistral3-8B (modèle multimodal, ajustement uniquement sur les couches du LLM) ;
— Llama32-3B ;

et sans LoRa :

— T5 (Raffel *et al.*, 2020) : notre *baseline*, l'ajustement d'une version déjà ajustée pour le résumé automatique en français [2].

5.4 Modèle guidé de bout en bout par un LLM

Ce modèle est entièrement guidé par un grand modèle de langage. Dans un premier temps, le prompt suivant est appliqué aux résumés du corpus d'entraînement : "Quels sont les différents aspects généraux abordés dans le résumé de match de football suivant : ". Les aspects sont automatiquement regroupés et comptés, et seuls sont gardés les 10 premiers. Un deuxième prompt sert alors à générer le prompt qui servira à l'entrée du modèle de résumé, d'après les 10 premiers aspects et trois exemples guides. Le prompt résultant permet d'interroger un grand modèle de langage afin de générer les résumés. La figure 3 illustre le pipeline de ce modèle.

2. https://huggingface.co/plguillou/t5-base-fr-sum-cnndm

Aspect	%	Rang
Résultat du match	47	1
Buts marqués	10	–
Le résultat du match	5	1
Positionnement des équipes	4	2
Performances individuelles	4	3
Résultat	3	1
Qualité du match	2	8
Classement	2	2
Qualité des joueurs	2	–

Aspect	%	Rang
Le score	2	1
Qualité de l'équipe	2	–
Évolution du match	2	–
Résultat final	2	1
Les performances individuelles	2	3
Le score et la victoire	2	1
Analyse du match	2	–
L'atmosphère du match	2	–
Les buteurs	3	3

TABLE 3 – Aspects détectés automatiquement par LLM sur le corpus d'entraînement et leur rang dans l'analyse manuelle. Dans la même couleur : les aspects redondants

Le tableau 3 présente les 20 aspects les plus fréquents selon cette analyse, et leur rang d'après l'analyse manuelle s'ils y avaient été détectés. On voit que le LLM a capté les trois informations les plus fréquentes selon l'analyse manuelle du corpus de (Belemkoabga *et al.*, 2021), mais ne capte pas les informations les plus fines. De plus, la majorité des aspects sont redondants. Nous avons colorié les lignes qui présentent des aspects redondants de la même couleur pour mieux les identifier.

6 Évaluation

Nous avons évalué les différents modèles selon deux métriques, sur un sous-ensemble de 3% du corpus tiré aléatoirement, soit 227 paires de commentaires / résumé. Les deux métriques utilisées sont :

— ROUGE (Lin, 2004)
— BERTScore (Zhang *et al.*, 2020).

Le tableau 4 présente les résultats obtenus avec les modèles naïfs, guidés par un prompt rédigé d'après une analyse manuelle des aspects, ajustés et guidé de bout en bout par un grand modèle de langage et présentés en section §5.

7 Limites des modèles

Nous observons que guider les grands modèles de langage en leur indiquant les aspects à extraire prioritairement n'améliore pas sensiblement les performances en termes de score ROUGE ou BERT. En revanche, nous observons un gain de performance considérable en ROUGE2 entre modèle naïf et modèle ajusté. Cela laisse penser que l'ajustement sur les données permet de capter et mieux reproduire le style des résumés de référence.

Afin de vérifier ces résultats et l'intuition concernant le style des résumés générés, nous avons mené une étude qualitative sur 15 exemples du corpus d'évaluation afin de déterminer si chacun des modèles générait des résumés fidèles au déroulé du match. Nous avons donc évalué manuellement la factualité du résumé sur les aspects suivants : issue du match, score du match, buteurs, domination, série de victoires/défaites. Nous avons profité de cette évaluation manuelle pour évaluer la proximité stylistique avec les résumés de référence et la qualité linguistique, notées sur 5 ainsi que les hallucinations (on

Modèle	R1	R2	RL	BERT P	BERT R	BERT F1
Naïfs						
qwen3_14B	0,216	**0,039**	**0,118**	**0,652**	**0,741**	**0,694**
GPT_OSS_20B	0,200	0,026	0,108	0,622	0,695	0,656
Mistral3_8B_Instr	**0,231**	0,033	0,120	0,601	0,701	0,647
Llama32_3B	0,187	0,032	0,113	0,641	0,712	0,675
Aspects manuels						
Qwen3_14B	**0,204**	**0,043**	**0,114**	**0,655**	**0,748**	**0,698**
GPT_OSS_20B	0,163	0,027	0,090	0,612	0,721	0,670
Mistral3_8B_Instr	0,174	0,034	0,097	0,622	0,741	0,676
Llama32_3B	0,164	0,028	0,099	0,627	0,719	0,667
Ajustés						
T5	0,226	0,039	0,137	**0,696**	0,653	0,673
qwen3_14B	0,289	0,062	0,161	0,668	0,706	0,686
GPT_OSS_20B	0,255	0,034	0,138	0,687	0,680	0,683
Mistral3_8B_Instr	0,214	0,038	0,129	0,667	0,681	0,674
Llama32_3B	**0,308**	**0,063**	**0,173**	0,692	**0,710**	**0,700**
LLM bout en bout						
Qwen3_14B	0,276	0,043	0,143	0,676	0,704	0,689
GPT_OSS_20B	0,220	0,028	0,112	0,636	0,682	0,637
Mistral3_8B_Instr	0,205	0,036	0,109	0,633	0,751	0,687
Llama32_3B	0,103	0,017	0,070	0,594	0,708	0,646

TABLE 4 – Moyennes ROUGE et BERTScore des modèles interrogés avec un prompt naïf, un prompt adapté d'après les analyses manuelles des aspects, les modèles ajustés et les modèles à base de LLM bout en bout.

compte les hallucinations majeures : buteurs, séries, classement etc). Nous présentons les résultats de cette évaluation dans le tableau 5.

On peut constater que Qwen3-14B arrive à inférer le score exact dans la majorité des cas, quand il est interrogé avec un prompt, qu'il soit naïf, ajusté manuellement pour contenir les aspects les plus fréquents, ou construit de bout en bout, comme décrit en Section §5.4. En revanche, il semble que l'ajustement du modèle lui fasse perdre sa capacité à inférer sur le corpus, puisqu'il n'infère le score exact que dans 16% des cas. L'ajustement permet néanmoins de mieux capter le style des résumés, puisque le modèle Qwen3-14B obtient la note maximale de 5 en style. La solution bout en bout semble être un bon compromis, puisqu'elle conserve une très bonne note en style tout en restant capable d'inférer correctement le score exact dans la majorité des cas. Cette analyse manuelle est cependant à considérer avec précaution étant donné le relatif faible nombre d'exemples évalués (15). Il est intéressant de noter que T5 ajusté, pourtant classé d'après les métriques ROUGE et BERTScore au même niveau voire au-dessus de la majorité des modèles, n'extrait aucune information essentielle selon les évaluations manuelles.

8 Conclusion et perspectives

Nous avons présenté un corpus de résumé automatique de commentaires sportifs tiré du site l'Équipe, introduit dans Belemkoabga *et al.* (2021) et que nous avons mis à jour avec les données les plus récentes. Nous avons présenté ses particularités et les défis qu'elles induisent pour le résumé automa-

Système	Issue	Score	Buteurs	Domination	Série	Halluc.↓	Style	Qual. Ling.
Naïf								
Qwen3-14B	**1,000**	**0,500**	0,580	0,333	0,333	0,667	3,167	**4,833**
GPT-oss-20B	0,333	0,333	0,333	0,000	0,000	0,833	2,167	0,833
Mistral3-8B-Instr	0,833	0,333	**0,692**	0,500	0,333	2,333	1,000	3,167
Llama-3.2-3B-Instr	0,833	0,333	0,583	0,250	0,000	2,167	1,000	2,667
Aspects manuels								
Qwen3-14B	**1,000**	0,333	0,605	**1,000**	**0,667**	0,667	3,333	4,667
GPT-oss-20B	0,833	0,333	0,498	0,833	0,167	3,667	1,667	1,667
Mistral3-8B-Instr	0,833	**0,500**	0,580	0,667	0,333	2,167	2,167	4,167
Llama32-3B	0,833	0,167	0,575	0,750	0,500	2,000	1,500	3,500
Ajustés								
Qwen3-14B	0,667	0,167	0,443	0,333	0,000	1,833	**5,000**	4,667
GPT-oss-20B	0,167	0,000	0,055	0,000	0,000	2,667	3,000	0,333
Mistral3-8B-Instr	0,500	0,167	0,413	0,167	0,000	2,000	2,167	1,500
Llama-3.2-3B-Instr	0,667	0,167	0,635	0,167	0,000	2,500	4,667	4,000
T5	0,000	0,000	0,000	0,000	0,000	**0,500**	0,833	3,500
LLM bout en bout								
Qwen3-14B	**1,000**	**0,500**	0,358	0,333	0,333	0,667	4,500	**4,833**
Mistral3-8B-Instr	**1,000**	0,333	0,408	0,833	0,250	3,000	1,667	4,000

TABLE 5 – Résultats moyens par système de l'évaluation manuelle. Les métriques Issue, Score, Buteurs, Domination, Série, Style et Qualité Linguistique sont à maximiser, tandis qu'Hallucinations est à minimiser.

tique. D'après celles-ci, nous avons mis au point et testé plusieurs méthodes de résumé automatique, fondées sur des grands modèles de langage, et montré que les modèles les plus récents étaient capables d'inférer depuis de longs commentaires non structurés certains éléments clés du résumé d'un match. Nous avons aussi montré que si l'ajustement des modèles à la tâche permet de mieux copier le style des résumés manuels, ce dernier semble faire perdre des capacités d'inférence.

Ces résultats laissent la porte ouverte à de nombreuses améliorations. Nous pensons notamment qu'une étape d'extraction des commentaires clés préalable à l'ajustement ou à l'inférence pourrait permettre de réduire la complexité du problème et donc d'améliorer les résumés générés. Nous pensons également qu'injecter dans le processus des bases de connaissances externes pourrait combler les lacunes des systèmes liées à l'absence de certaines informations dans les commentaires, comme le changement au classement, le transfert récent d'un joueur, le surnom de certaines équipes, ou encore le poste ou même l'équipe d'un joueur.

Nos résultats montrent également la difficulté à évaluer des systèmes génératifs : la différence entre systèmes est faible du point de vue de ROUGE ou BERTScore, tandis qu'une évaluation manuelle fait ressortir des différences prononcées, à tel point que certains systèmes bien classés selon des scores automatiques ne véhiculent aucune information intéressante selon un humain. D'autres métriques d'évaluation automatique adaptées à la tâche sont donc nécessaires.

Références

AIRES J. (2016). Automatic generation of sports news. Mémoire de master.

BELEMKOABGA D. S., BOSSARD A., ESSA A., RODRIGUES C. & SYLLA K. (2021). Neural network-based generation of sport summaries : A preliminary study. In R. MITKOV & G. ANGELOVA, Éds., *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, p. 147–154, Held Online : INCOMA Ltd.

BOUAYAD-AGHA N., CASAMAYOR G., MILLE S. & WANNER L. (2012). Perspective-oriented generation of football match summaries : Old tasks, new challenges. *ACM Trans. Speech Lang. Process.*, **9**(2), 3 :1–3 :31. DOI : 10.1145/2287710.2287711.

CORNEY D., MARTÍN DANCAUSA C. & GOKER A. (2014). Two sides to every story : Subjective event summarization of sports events using twitter. In *CEUR Workshop Proceedings*, volume 1198.

GAUTAM S., MIDOGLU C., SHAFIEE SABET S., KSHATRI D. B. & HALVORSEN P. (2022). Soccer game summarization using audio commentary, metadata, and captions. In *Proceedings of the 1st Workshop on User-Centric Narrative Summarization of Long Videos*, NarSUM '22, p. 13–22, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3552463.3557019.

GRATTAFIORI A. & AL (2024). The llama 3 herd of models.

HU E. J., YELONG SHEN, WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2022). LoRA : Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

HUANG K.-H., LI C. & CHANG K.-W. (2020). Generating sports news from live commentary : A Chinese dataset for sports game summarization. In K.-F. WONG, K. KNIGHT & H. WU, Éds., *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, p. 609–615, Suzhou, China : Association for Computational Linguistics. DOI : 10.18653/v1/2020.aacl-main.61.

LIN C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

LIU A. H. & AL (2026). Ministral3.

OPENAI, :, AGARWAL S. & AL (2025). gpt-oss-120b & gpt-oss-20b model card.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, **21**(140), 1–67.

SARKHOOSH M. H., GAUTAM S., MIDOGLU C., SABET S. S. & HALVORSEN P. (2024). Multimodal ai-based summarization and storytelling for soccer on social media. In *Proceedings of the 15th ACM Multimedia Systems Conference*, MMSys '24, p. 485–491, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3625468.3652197.

SEE A., LIU P. J. & MANNING C. D. (2017). Get to the point : Summarization with pointer-generator networks. In R. BARZILAY & M.-Y. KAN, Éds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1073–1083, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1099.

WANG J., LI Z., YANG Q., QU J., CHEN Z., LIU Q. & HU G. (2021). Sportssum2.0 : Generating high-quality sports news from live text commentary. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, p. 3463–3467, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3459637.3482188.

WANG J., LI Z., ZHANG T., ZHENG D., QU J., LIU A., ZHAO L. & CHEN Z. (2022). Knowledge enhanced sports game summarization. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, p. 1045–1053, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3488560.3498405.

YANG A. & AL (2025). Qwen3 technical report.

ZHANG J., YAO J.-G. & WAN X. (2016). Towards constructing sports news from live text commentary. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1361–1371, Berlin, Germany : Association for Computational Linguistics. DOI : 10.18653/v1/P16-1129.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). Bertscore : Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020* : OpenReview.net.