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

Sélection de résumés par aspect pour le déploiement industriel : agrégation de Borda multi-objectif sans référence

Carl Hatoum[1,2] Catherine Combes[1] Virginie Burnaz-Fresse[1]
Christophe Gravier[1] Mathieu Orzalesi[2]
(1) Université Jean Monnet, Laboratoire Hubert Curien, UMR CNRS 5516, Saint-Étienne, France
(2) SEGULA Technologies, France
{carl.hatoum, catherine.combes, virginie.fresse,
christophe.gravier}@univ-st-etienne.fr
mathieu.orzalesi@segula.fr

RÉSUMÉ

Le résumé par aspect (*Aspect-Based Summarization*, ABS) cible un aspect thématique explicitement spécifié. Lorsqu'un LLM génère plusieurs candidats pour un même couple document–aspect, sélectionner le meilleur est difficile : la qualité est multi-dimensionnelle (couverture, adhérence à l'aspect, cohérence factuelle) et les résumés de référence sont rarement disponibles en déploiement industriel. Nous formalisons cette sélection comme une décision sans référence sur un espace multi-objectif, évaluée par deux règles d'agrégation comparées (Z-Moyenne sur les scores calibrés, Borda sur les rangs) sur ACLSum et LexAbSumm. Les résultats montrent que les trois dimensions capturent des axes indépendants, que Borda est plus robuste aux signaux bruités pour des taux de corruption intermédiaires (les deux règles convergeant à forte corruption), et que les classements s'alignent modérément avec les préférences humaines. L'entropie sémantique des distributions de rang constitue par ailleurs un outil de diagnostic opérationnel pour le choix des paramètres de génération.

ABSTRACT

Reference-Free Selection of Aspect-Based Summaries via Multi-Objective Borda Aggregation

Aspect-Based Summarization (ABS) targets an explicitly specified thematic aspect. When a large language model generates multiple candidates for the same document–aspect pair, selecting the best is challenging : quality is multi-dimensional (content coverage, aspect adherence, factual consistency) and reference summaries are rarely available in industrial deployments. We formalize this selection as a reference-free decision over a multi-objective space, evaluated through two aggregation rules (z-mean over calibrated scores, Borda over ranks) on ACLSum and LexAbSumm. Results show that the three quality dimensions capture independent axes, that Borda offers better robustness to noisy signals at intermediate corruption rates (with both rules converging at high corruption), and that rankings moderately align with human preferences. The semantic entropy of rank distributions further provides an operational diagnostic for selecting generation parameters.

MOTS-CLÉS : Résumé par aspect, Grands modèles de langue, Décision multi-objectif, Agrégation de rangs.

KEYWORDS: Aspect-Based Summarization, Large Language Models, Robust Decision-Making, Rank Aggregation.

Actes de l'atelier sur l'évaluation des modèles génératifs (LLM) et challenges (EvalLLM 2026), pages 179–190.

1 Introduction

Les grands modèles de langue (LLM) peuvent aujourd'hui produire des résumés de haute qualité, y compris dans le cadre du **résumé par aspect** (*Aspect-Based Summarization*, ABS) (Takeshita *et al.*, 2024; Huang *et al.*, 2024). Dans ce paradigme, la génération est conditionnée sur un **aspect** : un angle thématique ou une dimension d'analyse explicitement spécifiée par l'utilisateur (par exemple, « méthodologie » ou « résultats expérimentaux » dans un article scientifique, ou « motifs du jugement » dans une décision juridique). Contrairement au résumé générique, l'ABS n'admet pas de cible canonique unique : plusieurs résumés peuvent légitimement satisfaire les exigences de la tâche tout en mettant l'accent sur des contenus différents.

Dans la pratique, les LLM (Touvron *et al.*, 2023; OpenAI, 2023) produisent souvent plusieurs résumés candidats pour un même couple document–aspect, via l'échantillonnage, des variations de *prompt* ou des stratégies de décodage. La question centrale est alors : comment sélectionner automatiquement le meilleur candidat parmi cet ensemble ?

Complexité multi-dimensionnelle de la qualité. Un bon résumé par aspect doit simultanément satisfaire plusieurs **dimensions de qualité** distinctes : il doit couvrir le contenu sémantiquement pertinent du document source, se concentrer sur l'aspect cible, et rester cohérent avec les faits du document sans introduire d'informations non étayées. Ces propriétés sont partiellement indépendantes : un résumé peut être très adhérent à l'aspect mais omettre du contenu important, ou être très complet mais factuellement imprécis. Réduire la qualité à un score scalaire unique occulte ces compromis et entrave la transparence et la diagnosticabilité (Fabbri *et al.*, 2021; Deutsch *et al.*, 2022; Wang *et al.*, 2023).

Contraintes des déploiements industriels. La plupart des évaluateurs multi-dimensionnels existants (Zhong *et al.*, 2022; Koto *et al.*, 2022) reposent sur des **résumés de référence** : des exemples annotés manuellement servant d'étalon. Or, en déploiement industriel, de telles références sont rarement disponibles : leur construction est coûteuse, les annotations non-expertes manquent de nuance dans le domaine, et les annotations expertes peuvent introduire des connaissances propres à l'annotateur qui sont distinctes du contenu factuel du document (Chawla *et al.*, 2026).

Notre proposition. Nous formulons la sélection de résumés par aspect comme une décision définie sur un espace multi-objectif induit par la tâche, sans résumé de référence ni annotation humaine. Chaque candidat est caractérisé selon des dimensions de qualité explicites, estimées par des modèles *proxy* légers déployables sur site. Les règles d'agrégation (par scores ou par rangs) sont traitées comme des règles de décision sur cet espace, permettant transparence et diagnosticabilité.

Ce travail apporte trois contributions : (1) une formalisation de la sélection de résumés par aspect comme décision multi-objectif sans résumé de référence, ancrée dans la sémantique du document et de l'aspect ; (2) une comparaison des règles d'agrégation par scores (Z-Moyenne) et par rangs (Borda) comme règles de décision sur cet espace, et l'analyse de leurs propriétés de robustesse ; (3) une étude empirique sur deux corpus montrant que la décomposition multi-dimensionnelle révèle des compromis de qualité que les critères scalaires occultent.

2 Travaux connexes

Les travaux antérieurs ont établi que la qualité des résumés est intrinsèquement multi-dimensionnelle et ne peut pas être fidèlement représentée par un score scalaire unique. En conséquence, plusieurs cadres d'évaluation décomposent la qualité en dimensions interprétables afin de mettre en évidence les compromis que les métriques scalaires occultent (Zhong *et al.*, 2022; Koto *et al.*, 2022; Jain *et al.*, 2023). Cependant, la plupart de ces évaluateurs multi-dimensionnels sont basés sur des résumés de référence, ce qui les rend inapplicables dans des contextes industriels où ces références sont indisponibles. Les approches plus récentes fondées sur le jugement de modèles de langue (Wang *et al.*, 2024) assouplissent l'exigence de références explicites, mais elles réduisent les évaluations multi-dimensionnelles à des scores scalaires ou des préférences par paires, et requièrent un accès à un modèle externe.

Ces limitations sont amplifiées dans l'ABS, où la génération est conditionnée par un aspect et où plusieurs sorties peuvent légitimement satisfaire les exigences de la tâche à travers différents équilibres de dimensions de qualité.

La théorie du choix social (Brandt *et al.*, 2016) offre une perspective alternative en fournissant des règles d'agrégation opérant sur des préférences ordinales sans supposer de comparabilité cardinale entre critères. Les méthodes basées sur les rangs, comme l'agrégation de Borda, traitent chaque dimension de qualité comme un votant et favorisent les candidats qui atteignent un large consensus entre les critères. Nous adoptons cette perspective pour cadrer la sélection de résumés par aspect sans résumé de référence comme un problème d'agrégation sur plusieurs dimensions de qualité.

À notre connaissance, la sélection de résumés par aspect sans résumé de référence n'a pas encore été abordée comme une décision multi-objectif avec les outils de la théorie du choix social.

3 Méthodologie

Nous formulons la sélection de résumés par aspect sans résumé de référence comme un problème de décision sur des résumés candidats générés par LLM. Étant donné un document, un aspect cible et un ensemble de résumés échantillonnés sous un conditionnement LLM contrôlé, le but est de sélectionner un résumé optimal unique sans annotations de référence. La procédure est résumée dans la Figure 1.

3.1 Définition de la tâche

Soit $\mathcal{D}$ un corpus de documents et $\mathcal{A}$ un ensemble d'aspects cibles. Pour chaque $(\delta, \alpha) \in \mathcal{D} \times \mathcal{A}$, un LLM génère un ensemble fini de résumés candidats

$$S_{\delta,\alpha} = \{s_1, \ldots, s_N\},$$

en utilisant des variables de conditionnement $C = \{c_1, \ldots, c_N\}$, où chaque c peut être une variante de prompt, une stratégie de décodage, etc. Tous les résumés d'un même ensemble partagent la même entrée (δ, α) mais, en raison de conditionnements différents et de la stochasticité du modèle, ils peuvent différer par leur angle thématique, leur style et leur réalisation. L'ensemble est traité comme

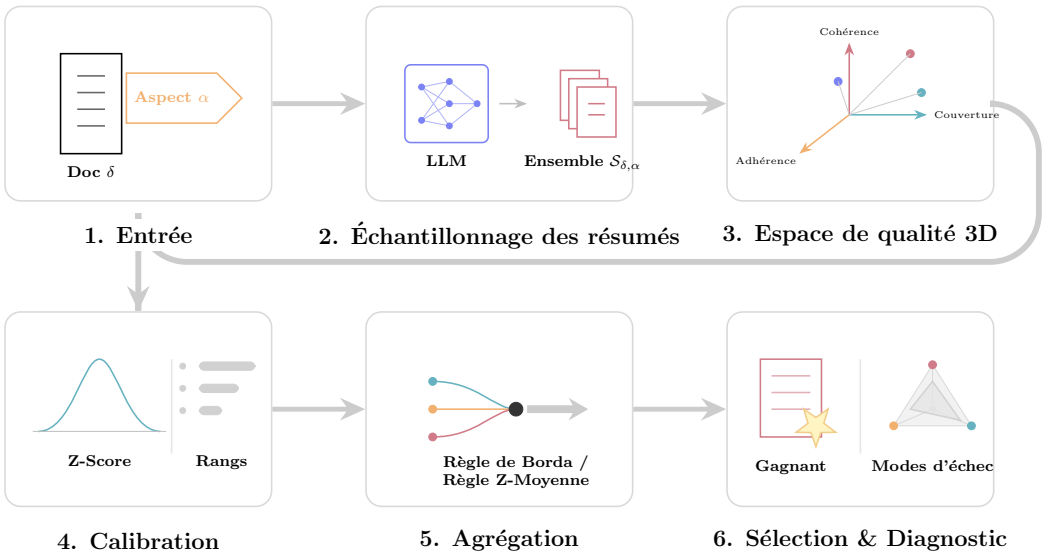

FIGURE 1 – Vue d'ensemble de la méthode de sélection multi-objectif sans résumé de référence.

un ensemble de décision fermé (les candidats ne sont pas modifiés après génération) : toutes les opérations de normalisation, de calibration et de classement sont définies relativement à $S_{\delta,\alpha}$. La tâche consiste à sélectionner un résumé

$$s^*_{\delta,\alpha} \in S_{\delta,\alpha}$$

en utilisant uniquement des dimensions de qualité relatives à (δ, α).

3.2 Espace des dimensions de qualité

La qualité ABS est intrinsèquement multi-dimensionnelle. Dans ce travail, nous nous concentrons sur trois dimensions, choisies pour couvrir les deux ancres sémantiques de la tâche (le document source et l'aspect cible) selon des axes complémentaires :

— **Couverture** (m_{cov}) : le résumé doit capturer le contenu documentaire sémantiquement pertinent. Elle est mesurée par similarité sémantique entre le résumé et le document source.
— **Adhérence** (m_{adh}) : le résumé doit se concentrer sur l'aspect cible. Elle est mesurée par la pertinence du résumé par rapport à la requête formée par l'aspect dans le contexte du document.
— **Cohérence** (m_{coh}) : le résumé ne doit pas introduire d'informations non étayées par le document (hallucinations). Elle est mesurée par inférence de langage naturel (NLI) au niveau des phrases entre le résumé et le document.

Ces trois dimensions couvrent les axes qualitatifs centraux de l'ABS, ancrés dans le document et dans l'aspect. D'autres dimensions (lisibilité, longueur, fluidité) sont pertinentes en général mais moins spécifiques à ce type de tâche.

Chaque résumé est représenté par un vecteur de qualité

$$m(s_i) = \big(m_{\text{cov}}(s_i),\ m_{\text{adh}}(s_i),\ m_{\text{coh}}(s_i)\big) \in \mathbb{R}^3, \tag{1}$$

conditionné implicitement sur (δ, α).

3.3 Calibration des dimensions

Les dimensions brutes différant par leur échelle, une calibration est effectuée au sein de chaque ensemble. Pour chaque dimension $d \in \{\text{cov}, \text{adh}, \text{coh}\}$:

Calibration par z-score (centrage réduit).

$$z_d(s_i) = \frac{m_d(s_i) - \mu_d}{\sigma_d}, \tag{2}$$

où μ_d et σ_d sont la moyenne et l'écart-type de m_d sur $S_{\delta,\alpha}$. Cette transformation ramène chaque dimension à une moyenne nulle et un écart-type unitaire, rendant les scores comparables entre dimensions indépendamment de leur échelle brute.

Calibration par rang. Chaque dimension induit également un classement

$$r_d(s_i) \in \{1, \ldots, |S_{\delta,\alpha}|\},$$

un rang plus faible indiquant de meilleures performances.

3.4 Règles de décision

Les dimensions calibrées sont agrégées en un score de décision scalaire.

Z-Moyenne.

$$Q_Z(s_i) = \frac{1}{3} \sum_{d \in \{\text{cov,adh,coh}\}} z_d(s_i). \tag{3}$$

Cette règle suppose une comparabilité cardinale des dimensions après centrage réduit : elle favorise les résumés dont la somme des écarts à la moyenne est maximale. Elle est sensible aux valeurs aberrantes : un score extrêmement élevé sur une seule dimension peut dominer l'agrégation.

Borda.

$$Q_B(s_i) = \sum_{d \in \{\text{cov,adh,coh}\}} \big(|S_{\delta,\alpha}| - r_d(s_i)\big). \tag{4}$$

Cette règle s'appuie uniquement sur les classements relatifs. Chaque dimension agit comme un votant, et un candidat obtient un score élevé s'il est bien classé selon plusieurs dimensions simultanément. En projetant les scores sur des rangs, elle neutralise l'amplitude absolue des signaux, ce qui la rend robuste aux scores bruités ou mal calibrés.

La sélection est effectuée par

$$s^*_{\delta,\alpha} = \arg\max_{s_i \in S_{\delta,\alpha}} Q(s_i), \quad Q \in \{Q_Z, Q_B\}. \tag{5}$$

3.5 Entropie sémantique

Chaque règle de décision Q induit un ordre total sur $S_{\delta,\alpha}$, qui est discrétisé en R intervalles de rang $\mathcal{B} = \{B_1, \ldots, B_R\}$ ($R = 10$ déciles dans nos expériences). Pour mesurer la stabilité du classement selon le conditionnement, nous définissons l'entropie sémantique. Pour chaque conditionnement $c_i \in C$, soit $S_{\delta,\alpha}^{(c_i)} \subseteq S_{\delta,\alpha}$ les résumés générés sous c_i. La distribution empirique sur les intervalles de rang est :

$$p_k^{(c_i)} = \frac{1}{|S_{\delta,\alpha}^{(c_i)}|} \sum_{s \in S_{\delta,\alpha}^{(c_i)}} \mathbf{1}\{s \in B_k\}.$$

L'entropie sémantique est mesurée par l'entropie de Shannon discrète (Shannon, 1948) :

$$H_{\text{sem}}(c_i) = -\sum_{k=1}^{R} p_k^{(c_i)} \log p_k^{(c_i)}. \tag{6}$$

Une entropie sémantique faible indique que les résumés produits sous un conditionnement fixe se concentrent systématiquement dans une zone stable du classement, traduisant une cohérence entre le conditionnement et la qualité multi-dimensionnelle mesurée.

4 Analyse empirique de l'espace multi-objectif

Pour que la méthode soit utile en pratique, quatre propriétés doivent être vérifiées : les dimensions de qualité doivent capturer des axes indépendants ; l'agrégation doit résister à des signaux bruités ; les classements doivent distinguer les résumés de haute qualité des candidats générés ; et les classements obtenus doivent correspondre aux préférences humaines. Nous instancions et analysons empiriquement la méthode décrite en Section 3 sur deux corpus de résumé par aspect, en nous concentrant sur : (i) la complémentarité des dimensions (Section 4.2) ; (ii) la robustesse de l'agrégation (Section 4.3) ; (iii) le comportement sur les résumés de référence (Section 4.4) ; et (iv) l'alignement avec les préférences d'évaluateurs humains (Section 4.5).

4.1 Configuration expérimentale

Nous utilisons ACLSum (résumés de papiers scientifiques en anglais) (Takeshita *et al.*, 2024) et LexAbSumm (décisions juridiques en anglais) (Huang *et al.*, 2024), qui associent des documents à des résumés de référence par aspect, écrits par des humains. Ces deux corpus anglophones couvrent des registres différents (scientifique vs. juridique), permettant de tester la généralisation de la méthode.

Génération de résumés LLM. Pour chaque paire document–aspect (δ, α), nous générons des candidats avec Llama-3.1-8B-Instruct [1] selon deux axes de conditionnement : (i) un *prompting* à haute spécificité (HS) vs. faible spécificité (FS) (trois variantes de formulation du prompt chacune) ; et (ii) un échantillonnage à température moyenne (TM) de 0,7 vs. haute température (HT) de 1,4. Ces combinaisons produisent $N = 12$ configurations de conditionnement par couple (δ, α), constituant l'ensemble $S_{\delta,\alpha}$.

1. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

Instanciation des dimensions de qualité. Nous instancions les dimensions de qualité de la Section 3.2 à l'aide de modèles d'évaluation disponibles publiquement. Ces propriétés appellent naturellement des architectures de modèles différentes ; utiliser un modèle unique risquerait de les confondre, là où des modèles spécialisés permettent de les mesurer indépendamment. Ces modèles fournissent des signaux *proxy* imparfaits : nous nous concentrons délibérément sur la sélection sous des signaux partiels et bruités, ce qui constitue le régime réaliste de déploiement. Ce choix est motivé par les contraintes du régime industriel ciblé : le coût d'inférence par candidat et le besoin d'auditabilité par dimension y justifient des modèles spécialisés légers plutôt qu'un modèle généraliste produisant un score scalaire global. Des modèles spécialisés et légers permettent une traçabilité par dimension, un déploiement sur site, et une mise à jour indépendante de chaque composant.

Concrètement, nous approximons :

— la **Couverture** en utilisant BERTScore (Zhang *et al.*, 2020) entre chaque résumé candidat $s_i \in S_{\delta,\alpha}$ et le document source δ, calculé avec `roberta-large`[2] ;
— l'**Adhérence** en utilisant le *cross*-encodeur `jina-reranker-v3`[3], qui score la pertinence en encodant conjointement (α, δ, s_i) (requête (α, δ) ; candidats $s_i \in S_{\delta,\alpha}$). Un modèle de reranking mesure dans quelle mesure un candidat répond à une requête dans son contexte : cela constitue une approximation naturelle de l'adhérence à l'aspect, qui évalue précisément si le résumé traite bien la dimension thématique demandée au vu du document source ;
— la **Cohérence** en utilisant `deberta-large-mnli`[4], estimant la cohérence factuelle entre chaque résumé candidat $s_i \in S_{\delta,\alpha}$ et le document δ via l'inférence d'implications au niveau des phrases pour pénaliser les contradictions et les affirmations non étayées (Laban *et al.*, 2022).

4.2 Objectif 1 : Complémentarité des dimensions

Nous testons si les dimensions instanciées forment des axes complémentaires plutôt qu'un facteur latent unique. Nous travaillons délibérément avec des dimensions à interprétation sémantique fixe (couverture, adhérence, cohérence), car l'interprétabilité par dimension est une propriété essentielle pour l'audit et la diagnosticabilité en déploiement. La faible corrélation empirique ci-après valide que les dimensions ne se recoupent pas de façon redondante.

Complémentarité par paires. Après centrage de la moyenne dans chaque ensemble $S_{\delta,\alpha}$, les corrélations de Spearman intra-ensemble restent faibles ; la plus grande, Couverture vs. Adhérence ($\rho = 0{,}162$), demeure non significative après correction de Benjamini-Hochberg (BH-FDR). Les résumés se distribuent sur des régions de compromis distinctes dans l'espace à trois dimensions, sans se regrouper sur un axe latent unique.

Ablation des dimensions de qualité. La suppression d'une dimension avant le calcul de la Z-Moyenne réduit l'accord avec le classement complet ($\bar{\tau}$ de Kendall $= 0{,}612$ sans Couverture, $0{,}603$ sans Adhérence, $0{,}609$ sans Cohérence), indiquant une information de classement non redondante.

2. https://huggingface.co/FacebookAI/roberta-large
3. https://huggingface.co/jinaai/jina-reranker-v3
4. https://huggingface.co/microsoft/deberta-large-mnli

4.3 Objectif 2 : Robustesse de l'agrégation

En déploiement industriel, les signaux de qualité proviennent de modèles *proxy* imparfaits susceptibles de produire des scores bruités ou aberrants selon le domaine ou le document. Il est donc essentiel que la règle d'agrégation maintienne un classement stable malgré ces perturbations. Nous comparons les règles de décision Z-Moyenne et Borda sous des métriques stables et sous une corruption contrôlée.

Accord entre agrégateurs. Sur tous les ensembles, les classements Z-Moyenne et Borda s'accordent fortement : τ de Kendall $= 0{,}863 \pm 0{,}035$ et ρ de Spearman $= 0{,}969 \pm 0{,}015$ (tous deux avec $p < 10^{-58}$).

Robustesse à la corruption des dimensions de qualité. Nous corrompons une fraction aléatoire $f \in [0, 0{,}60]$ des candidats par ensemble en ajoutant aléatoirement une perturbation de $+5\sigma_d$ à une ou deux dimensions (où σ_d désigne l'écart-type intra-ensemble de la dimension d), et comparons les classements perturbés aux classements de référence via le τ de Kendall (Figure 2). La Z-Moyenne descend à $\tau = 0{,}538$ pour $f = 0{,}30$ et $0{,}503$ pour $f = 0{,}60$, tandis que Borda atteint $\tau = 0{,}630$ pour $f = 0{,}30$ et $0{,}503$ pour $f = 0{,}60$. Globalement, les deux agrégateurs se dégradent régulièrement à mesure que la contamination augmente, Borda montrant une meilleure robustesse pour des taux de corruption intermédiaires ($f = 0{,}30$) ; les deux agrégateurs convergent à $f = 0{,}60$ (Figure 2).

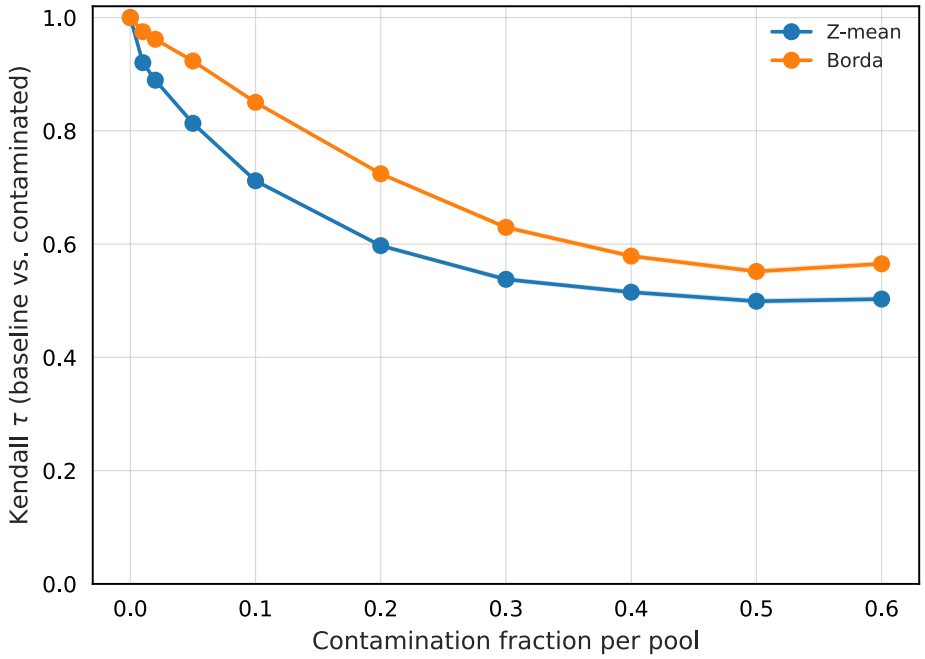

FIGURE 2 – Robustesse sous corruption. τ de Kendall entre classements original et perturbé en fonction de la fraction de corruption f.

4.4 Objectif 3 : Comportement sur les résumés de référence vs. générés

Nous analysons comment l'espace des dimensions de qualité et les règles d'agrégation distinguent les résumés de référence (écrits par des humains) des candidats générés par LLM, et comment les conditionnements de *prompting* et d'échantillonnage affectent le comportement de classement. Les résumés de référence servent ici de points de repère empiriques pour la comparaison relative, et non comme une définition absolue de la qualité des résumés.

Détection des résumés de référence. Nous évaluons dans quelle mesure les dimensions de qualité individuelles et leurs agrégations séparent les résumés de référence des candidats générés en utilisant l'AUC-ROC. Parmi les dimensions individuelles, l'Adhérence fournit la discrimination la plus forte (AUC $= 0{,}704$), suivie par la Cohérence ($0{,}654$) et la Couverture ($0{,}597$). L'agrégation améliore substantiellement la séparabilité ; Borda ($0{,}762$) et Z-Moyenne ($0{,}758$) atteignent des niveaux comparables. Les ablations de dimensions soulignent le rôle dominant de l'Adhérence : sa suppression entraîne la plus grande baisse pour la Z-Moyenne ($0{,}698$) et Borda ($0{,}705$), tandis que la suppression de la Couverture ($0{,}743$ Z-Moyenne ; $0{,}750$ Borda) ou de la Cohérence ($0{,}721$ Z-Moyenne ; $0{,}714$ Borda) produit des dégradations plus faibles mais cohérentes, indiquant des contributions complémentaires entre les dimensions.

Proximité par rapport aux résumés de référence. Au-delà de la séparabilité binaire, la distance au résumé de référence fournit une vue quantitative de l'alignement des résumés sélectionnés dans l'espace de qualité induit. Les règles basées sur l'agrégation sélectionnent des candidats substantiellement plus proches du résumé de référence que le hasard, avec des distances euclidiennes médianes dans $\mathbb{R}^3$ de $1{,}87$ pour le gagnant Borda et de $2{,}18$ pour le gagnant Z-Moyenne, comparées à $2{,}44$ sous sélection aléatoire uniforme.

Statistiques de distribution des rangs et entropie sémantique. Pour quantifier la stabilité du classement selon les conditionnements, nous analysons l'entropie sémantique sur les déciles de rang. Au-delà de la confirmation attendue que des températures élevées augmentent la dispersion, l'intérêt pratique est double : (1) l'entropie sémantique constitue un outil de diagnostic pour choisir les paramètres de génération en déploiement, permettant à un praticien d'identifier le régime de conditionnement produisant les classements les plus stables ; (2) les résumés de référence atteignent systématiquement une entropie très faible, ce qui valide que la métrique capture un signal de qualité réel et non du bruit. Le Tableau 1 présente les valeurs résultantes. Parmi les candidats générés, un *prompting* à plus haute spécificité et des températures modérées produisent une entropie plus faible et des distributions de rangs plus concentrées, principalement sur ACLSum.

4.5 Objectif 4 : Alignement avec les préférences d'évaluateurs humains

Nous testons si les ordres induits par les dimensions de qualité s'alignent avec les préférences d'évaluateurs humains, traitant leurs jugements comme un signal comparatif indépendant.

Protocole. Pour chaque document, nous construisons des mini-ensembles de 6 résumés (cinq candidats du modèle, un résumé de référence), produisant 147 ensembles sur LexAbSumm et ACLSum.

TABLE 1 – Entropie sémantique (H_{sem}) et masse de probabilité dans les deux déciles supérieurs ($p_9 + p_{10}$) par conditionnement. **Gras** = 1er rang, souligné = 2e rang, *italique* = 3e rang (par colonne, par groupe).

Corpus	Méthode	Conditionnement	H_{sem}	$p_9 + p_{10}$
LexAbSumm	Z-Moy.	HS–HT	2,28	0,16
		HS–TM	2,30	*0,22*
		FS–HT	2,30	0,18
		FS–TM	*2,29*	0,25
		Référence	**1,91**	**0,49**
	Borda	HS–HT	2,29	0,15
		HS–TM	2,30	*0,22*
		FS–HT	2,29	0,18
		FS–TM	2,29	0,25
		Référence	**1,91**	**0,53**
ACLSum	Z-Moy.	HS–HT	2,29	*0,23*
		HS–TM	*2,22*	0,34
		FS–HT	2,21	0,11
		FS–TM	2,28	0,14
		Référence	**1,80**	**0,58**
	Borda	HS–HT	2,29	*0,22*
		HS–TM	2,21	0,34
		FS–HT	2,21	0,11
		FS–TM	2,29	0,14
		Référence	**1,80**	**0,57**

L'agrégation (après normalisation ; Sections 3.3 et 3.4) ordonne les cinq candidats du système. Les six résumés sont mélangés et classés par des évaluateurs humains pour produire l'ordre de référence.

Alignement NDCG et accord par paires. Nous rapportons le Gain Cumulatif Normalisé et Actualisé (NDCG) et l'accord par paires (Tableau 2). Le NDCG est retenu car il pondère davantage les erreurs en tête de classement : dans un contexte de sélection du meilleur résumé, les erreurs sur les premiers rangs ont un coût pratique plus élevé que les erreurs en bas de liste, ce que ne capturent pas les corrélations de rang classiques. L'accord par paires reflète la cohérence globale de l'ordonnancement. Les deux règles d'agrégation atteignent un NDCG similaire ($0{,}742$ vs $0{,}739$), indiquant un alignement comparable avec les préférences humaines sur les résumés les plus appréciés. La différence d'accord par paires ($0{,}544$ vs $0{,}537$) est faible et doit être interprétée avec précaution en l'absence de test de significativité.

TABLE 2 – Alignement entre les classements induits par les dimensions de qualité et les préférences humaines sur 147 mini-ensembles. Plus élevé est meilleur.

Méthode	NDCG (moy. $\pm$ éc.-t.)	Par paires (moy. $\pm$ éc.-t.)
Borda	$0{,}742 \pm 0{,}149$	$0{,}544 \pm 0{,}212$
Z-Moyenne	$0{,}739 \pm 0{,}150$	$0{,}537 \pm 0{,}213$

5 Conclusion et limites

Nous avons formalisé la sélection de résumés par aspect comme une décision multi-objectif sans résumé de référence, définie par trois dimensions de qualité complémentaires (couverture, adhérence, cohérence). Nos expériences montrent que : (1) les trois dimensions capturent des axes indépendants et non redondants ; (2) l'agrégation de Borda est plus robuste que la Z-Moyenne face à des signaux *proxy* bruités pour des taux de corruption intermédiaires ; (3) les candidats sélectionnés par agrégation sont substantiellement plus proches des résumés de référence que la sélection aléatoire, Borda présentant la distance médiane la plus faible ; (4) l'entropie sémantique constitue un diagnostic opérationnel pour choisir les paramètres de génération en déploiement ; (5) les classements obtenus s'alignent modérément avec les préférences d'évaluateurs humains, validant la pertinence pratique de la méthode. Ce cadre est conçu pour des contextes industriels où les données sont confidentielles, les références indisponibles, et le recours à des modèles externes contraint.

Limites. La méthode attribue un poids égal à toutes les dimensions ; l'adaptation des poids selon les préférences des utilisateurs ou les contraintes de la tâche est une direction naturelle. L'analyse s'appuie sur deux corpus anglophones ; une évaluation sur des corpus multilingues et des domaines plus variés renforcerait la généralisation.

Considérations éthiques. Ce travail propose un cadre d'évaluation automatique de résumés sans supervision humaine directe. Les modèles *proxy* utilisés pour mesurer les dimensions de qualité peuvent présenter des biais liés à leur corpus d'entraînement, conduisant à favoriser systématiquement certains styles ou registres de rédaction. Dans un déploiement industriel, des décisions automatisées fondées sur ces scores pourraient amplifier ces biais à grande échelle. Nous recommandons de traiter les classements obtenus comme des outils d'aide à la décision plutôt que comme des verdicts définitifs, et d'intégrer des audits humains périodiques, notamment lors de changements de domaine ou de modèle générateur.

Références

BRANDT F., CONITZER V., ENDRISS U., LANG J. & PROCACCIA A. D., Éds. (2016). *Handbook of Computational Social Choice*. Cambridge : Cambridge University Press. procaccia.info.

CHAWLA K., ZHU C., CAI P., CHO S., NOVOTNEY S., SINGH A., LEWIS J., SAFEWRIGHT K., SAMUEL A., BABINSKY E., ZHANG S. & SAHU S. (2026). Lessons from the field : An adaptable lifecycle approach to applied dialogue summarization. arXiv : 2601.08682.

DEUTSCH D., DROR R. & ROTH D. (2022). Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of NAACL-HLT 2022*, p. 6037–6052 : Association for Computational Linguistics. aclanthology.org/2022.naacl-main.450/.

FABBRI A., KRYŚCIŃSKI W., MCCANN B., XIONG C., SOCHER R. & RADEV D. R. (2021). SummEval : Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, **9**, 391–409. aclanthology.org/2021.tacl-1.27/, DOI : 10.1162/tacl_a_00373.

HUANG Y., LI Y., WANG Y., ZHANG Y., LIU Z. & ZHOU W. (2024). LexAbSumm : A benchmark for aspect-based lexical abstract summarization. In *Proceedings of EMNLP 2024*. arXiv : 2405.02234.

JAIN S., KESHAVA V., SATHYENDRA S. M., FERNANDES P., LIU P., NEUBIG G. & ZHOU C. (2023). Multi-dimensional evaluation of text summarization with in-context learning. In *Findings of the Association for Computational Linguistics : ACL 2023*, p. 8487–8495. aclanthology.org/2023.findings-acl.537/, DOI : 10.18653/v1/2023.findings-acl.537.

KOTO F., BALDWIN T. & LAU J. H. (2022). FFCI : A framework for interpretable automatic evaluation of summarization. *Journal of Artificial Intelligence Research*, **73**, 1553–1607. DOI : 10.1613/jair.1.13167.

LABAN P., SCHNABEL T., BENNETT P. N. & HEARST M. A. (2022). SummaC : Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, **10**, 163–177. aclanthology.org/2022.tacl-1.10/, DOI : 10.1162/tacl_a_00453.

OPENAI (2023). GPT-4 technical report. arXiv : 2303.08774.

SHANNON C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**(3), 379–423. DOI : 10.1002/j.1538-7305.1948.tb01338.x.

TAKESHITA S., GREEN T., REINIG I., ECKERT K. & PONZETTO S. (2024). ACLSum : A new dataset for aspect-based summarization of scientific publications. In *Proceedings of NAACL 2024 (Long Papers)*, p. 6660–6675. aclanthology.org/2024.naacl-long.371/, DOI : 10.18653/v1/2024.naacl-long.371.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F., RODRIGUEZ A., JOULIN A., GRAVE E. & LAMPLE G. (2023). LLaMA : Open and efficient foundation language models. arXiv : 2302.13971.

WANG Y., ZHANG H. *et al.* (2024). JudgeLM : Fine-grained preference modeling and evaluation with large language models. arXiv : 2310.17631.

WANG Y., ZHANG Y., HE C., XU Y. & TANG B. (2023). Automatic evaluation of medical multi-document summarization : Limitations of overlap and embedding metrics. In *Proceedings of EMNLP 2023* : Association for Computational Linguistics. aclanthology.org/2023.emnlp-main.412/.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). BERTScore : Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*. iclr.cc.

ZHONG M., LIU Y., YIN D., MAO Y., JIAO Y., LIU P., ZHU C., JI H. & HAN J. (2022). Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 2023–2038. aclanthology.org/2022.emnlp-main.131/, DOI : 10.18653/v1/2022.emnlp-main.131.