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

CoRAFIG : corpus multi-genre pour la méta-évaluation du résumé automatique en français

Skander Hellal[1] Louis Jourdain[1]
(1) ChapsVision, France
{shellal, ljourdain}@chapsvision.com

RÉSUMÉ

CoRAFIG est le premier corpus français multi-genre de résumés automatiques, annoté sur cinq dimensions qualitatives et une couche d'erreurs explicites. Près de 1 000 résumés issus de 10 systèmes et 100 textes de six genres (presse, scientifique, juridique, financier, oral, littéraire) ont été annotés professionnellement. Contrairement aux corpus de presse existants, CoRAFIG évalue la robustesse des métriques face à la variation générique et sert de référence pour la méta-évaluation en français.

ABSTRACT

CoRAFIG : a multi-genre corpus for automatic summarization evaluation in French

CoRAFIG is the first multi-genre French corpus of automatic summaries, annotated on five qualitative dimensions with an explicit error layer. Nearly 1,000 summaries from 10 systems and 100 texts spanning six genres (news, scientific, legal, financial, oral, literary) are professionally annotated. Unlike existing news-focused corpora, CoRAFIG assesses metric robustness against genre variation and provides a meta-evaluation benchmark for French.

MOTS-CLÉS : résumé automatique, corpus multi-genre, annotation humaine, méta-évaluation.

KEYWORDS: automatic summarization, multi-genre corpus, human annotation, meta-evaluation.

1 Introduction

Les grands modèles de langue ont profondément transformé le résumé automatique : là où les modèles spécialisés produisaient des résumés laborieux sur des textes de presse courts, les LLMs semblent désormais capables de résumer n'importe quel document avec fluidité et cohérence. Mais *semblent*, car les outils pour l'affirmer font défaut, particulièrement en français. Les métriques automatiques d'évaluation du résumé automatique ont été massivement calibrées sur SummEval (Fabbri *et al.*, 2021) : un corpus de presse anglophone, homogène en longueur et en genre. En français, aucun corpus équivalent annoté humainement n'existe à ce jour. Cette lacune empêche de répondre à deux questions pourtant centrales : les corrélations entre métriques automatiques et jugement humain se maintiennent-elles en français ? Résistent-elles à la variation générique ? Ce travail présente CoRAFIG, un corpus de près de 1 000 résumés automatiques en français générés par 10 systèmes d'IA à partir de 100 textes de genres variés. Ces résumés ont été notés de 1 à 5 par des annotateurs professionnels, qui ont de plus relevé les erreurs de langue et de factualité. Cette ressource a pour objectif de déterminer si la tâche de résumé automatique a réellement été saturée en français, et d'aider à concevoir de nouvelles métriques automatiques mieux adaptées.

Actes de l'atelier sur l'évaluation des modèles génératifs (LLM) et challenges (EvalLLM 2026), pages 13–37.

Nos principales contributions sont :

- **CoRAFIG** (Corpus de Résumés Automatiques Francophones Inter-Genres), premier benchmark français multi-genre pour l'évaluation du résumé automatique : ∼1 000 résumés issus de 10 systèmes, 100 textes et 6 genres, annotés professionnellement sur 5 dimensions Likert et une couche d'erreurs explicites, en libre accès ;
- constatation d'une **séparation bimodale LLM / extractif** robuste sur cinq dimensions, confirmant en français la hiérarchie observée en anglais ;
- mise en évidence d'**effets différentiels du genre** (défaillance sur le juridique, ancrage sur la presse) et d'une **corrélation longueur-score faible**, en contradiction avec UniSumEval ;
- établissement d'un **désalignement score / erreur** : les erreurs factuelles dégradent significativement la consistance, démontrant les limites de l'évaluation Likert seule.

2 Travaux connexes

2.1 Une évolution des modes de résumé automatique, une tâche saturée ?

L'arrivée des modèles d'IA génératifs a entraîné une reconfiguration subite de la recherche sur le résumé automatique, au point qu'elle est souvent jugée saturée. Toutefois, les méthodes d'évaluation et les ressources n'ont pas su suivre ces avancées, rendant le développement de ressources pour mesurer ces performances encore plus crucial. En quinze ans, la discipline a muté d'un paradigme extractif fondé sur la sélection statistique de segments sources vers une approche abstractive par l'essor des architectures neuronales encodeurs-décodeurs avant d'aboutir à la synthèse conceptuelle des LLMs, capables de reformulations fluides. L'évaluation holistique des LLMs a assez tôt impliqué leur évaluation sur la tâche de résumé (Guo *et al.*, 2023) mais a été réalisée sur des benchmarks de 2020 et 2021 qu'ils saturent. L'étude de Goyal et al. (Goyal *et al.*, 2022) souligne que les LLMs échappent désormais aux mesures classiques et que les métriques automatiques s'avèrent incapables de juger leur inédite qualité. D'où le constat humoristique de Pu et ses coauteurs : *Summarization is (Almost) Dead* (Pu *et al.*, 2023) : les résumés générés par LLM sont en général préférés aux résumés écrits par des humains notamment en raison d'une plus grande fluidité. Ce décalage rend les jeux de données traditionnels obsolètes et force une remise en question profonde des protocoles de validation scientifique. Ce changement de paradigme a aussi entraîné un décalage grandissant entre l'évaluation rigide et la diversification de la tâche. Si le résumé de textes de presse semble saturé par les LLMs, l'attention se porte sur le résumé de textes longs (Wang *et al.*, 2022), multi-documents, ou guidé par instruction (Liu *et al.*, 2024). Les travaux manquent en langue française ce qui ne permet pas de déterminer si la tâche est saturée en français.

2.2 L'évaluation des résumés : une évolution en décalage

L'évaluation du résumé automatique a longtemps reposé sur des métriques de recouvrement lexical avant l'introduction de métriques sémantiques comme BERTScore (Zhang *et al.*, 2020), capables de saisir la synonymie. Toutefois, la pertinence de ces outils par rapport au jugement humain est restée longtemps débattue. Un tournant majeur, (Fabbri *et al.*, 2021), a permis de stabiliser l'évaluation autour de quatre dimensions qualitatives (Kryściński *et al.*, 2019) sur lesquelles nous reviendrons. SummEval est dès lors devenu le corpus de référence pour la méta-évaluation, servant de base à la

validation de toute nouvelle métrique en anglais, en particulier les nouvelles approches *LLM-as-a-judge* dont la capacité à fournir des scores numériques discrets et stables est contestée, l'obtention de jugements nuancés nécessitant des stratégies coûteuses à l'inférence (test-time scaling) (Gao *et al.*, 2024). De plus, ces juges automatiques restent calibrés sur les dimensions classiques de 2019, désormais saturées par les LLMs, ce qui masque des failles comme les hallucinations ou les omissions.

2.3 Limites des méthodes actuelles de calibration

Dai et al. (Dai *et al.*, 2024) identifient quatre étapes déterminantes pour toute méta-évaluation : choix des données, définition des dimensions de qualité, collecte des jugements humains et comparaison avec les métriques automatiques. Ils soulignent que ces métriques ont été calibrées quasi-exclusivement sur des textes de presse, avec un focus sur la consistance (*faithfulness*), et que les études restent cantonnées à des évaluations intrinsèques, ignorant tout impact extrinsèque. Le standard de référence depuis 2020, SummEval, cristallise ces dérives. Son ancrage exclusif dans CNN/DailyMail expose l'évaluation à un risque de contamination, ses résumés de référence, de simples *highlights* journalistiques, manquent de cohérence logique et intègrent des informations externes au texte source (Bachey *et al.*, 2025). Son protocole d'annotation, fondé sur un lissage supervisé des désaccords, a sacrifié le jugement humain naturel sur l'autel du consensus. Ces biais rendent impérative la création de nouveaux benchmarks, d'autant que cette limitation aux news pénalise particulièrement l'évaluation des LLMs (Zhang *et al.*, 2024). La calibration anglophone rend par ailleurs toute généralisation multilingue hasardeuse. Forde et al. (Forde *et al.*, 2024) montrent que les corrélations s'effondrent pour les langues typologiquement distantes. Le français, mieux doté, échappe partiellement à ce constat, mais Braun et al. (Braun *et al.*, 2021) ont montré, en traduisant SummEval en sept langues dont le français, que si la pertinence et la cohérence survivent pour les langues proches, la fluidité et la consistance factuelles, elles, ne résistent pas. L'absence de ressources conçues nativement pour le français, couplée à l'obsolescence des protocoles hérités de SummEval, constitue un angle mort critique pour la validation des LLMs francophones.

2.4 Un manque de ressources pertinentes dans un domaine en ébullition

L'anglais dispose de corpus spécialisés annotés qui font cruellement défaut au français : résumés de dialogues (Gao & Wan, 2022), documents longs, domaines spécifiques. D'autres langues comblent ce manque (Clark *et al.*, 2023) ; le corpus BASSE (Barnes *et al.*, 2025) offre pour l'espagnol et le basque des jugements humains (Likert, cinq critères) sur cinq LLMs, révélant que les métriques standards calibrées sur l'anglais y échouent et que les LLMs open-source peinent encore. Le français reste, lui, cantonné à sa part dans des corpus multilingues massifs mais non annotés comme MLSum (Scialom *et al.*, 2020). Des travaux récents comme UniSumEval (Lee *et al.*, 2024) commencent à dépasser l'hypothèse d'un genre unique, mais leur approche « IA-assistée » soulève un risque de biais circulaire. Pour garantir la neutralité de l'évaluation face aux LLMs eux-mêmes évalués, seule une annotation humaine strictement indépendante appliquée à des données multi-genres fait office de référence fiable. C'est pour répondre à cette exigence et pallier l'absence de ressources natives françaises qu'a été conçu le corpus CoRAFIG, dont le protocole est détaillé dans la section suivante.

3 Constitution du corpus CoRAFIG

3.1 Vue d'ensemble et motivations

Les corpus d'évaluation existants reposent en général sur un genre unique (la presse) ce qui biaise autant les métriques calibrées sur eux que les conclusions tirées sur la saturation de la tâche. La robustesse d'une métrique ne s'établit qu'en présence de variation générique réelle : longueur, registre, densité informationnelle, structure rhétorique. CoRAFIG cible 100 textes issus de six genres distincts, résumés par 10 modèles et annotés par des professionnels sur cinq dimensions qualitatives avec une couche d'erreurs. Les sous-corpus (Table 1) couvrent la presse (Scialom *et al.*, 2020), les retranscriptions orales (Rennard *et al.*, 2023), le juridique (Aumiller *et al.*, 2022), le financier (Zmandar *et al.*, 2022), le scientifique (HAL.fr) et la littérature francophone (XIXe–XXe siècle), avec des résumés de référence adaptés aux conventions de chaque genre.

TABLE 1 – Composition du corpus CoRAFIG par genre (longueurs en mots).

Genre	Sous-corpus	Prévus	Long. moy.	Médiane	Résumé de référence
Presse	MLSum	50	992	906	Chapeau journalistique
Retranscriptions	FREDSum	10	4 176	4 416	3 résumés humains
Scientifique	HAL.fr	10	2 994	2 970	Abstract auteur
Financier	CoFiF Plus	10	7 751	7 323	Résumé humain (exécutif)
Juridique	EUR-Lex-Sum	10	12 196	10 668	Résumé humain (officiel)
Littéraire	corpus interne	10	—	—	
Total		**100**	**3 564**	**1 329**	

La diversité des longueurs, allant de 992 mots (presse) à 12 196 (juridique), constitue un test de robustesse explicite des modèles au-delà du seul genre journalistique. La surreprésentation de la presse (50/90 textes) est délibérée : elle isole l'effet linguistique de l'effet générique en restant comparable à SummEval. Les résumés de référence juridiques et financiers sont institutionnels (EUR-Lex, CoFiF Plus) et servent uniquement aux métriques automatiques ; l'annotation humaine est absolue, sans référence. La phase 2 d'annotation couvre 90 textes sur les 100 prévus (367 résumés, 1 072 annotations) ; le corpus littéraire n'a pas encore été livré. Les résultats présentés ici portent sur ces 90 textes.

3.2 Sélection des textes sources

SummEval recourait à un tirage aléatoire parmi les textes CNN, ce qui présente plusieurs limites : risque de répétition thématique, hypothèse implicite d'uniformité de qualité et d'intérêt, impossible à transposer à des corpus scientifiques ou littéraires. Nous avons opté pour une sélection semi-automatique en deux étapes : (1) annotation des textes par la technique *LLM-as-judge* selon des critères explicites propres à chaque genre de texte (Annexe A), puis agrégation des scores en utilisant l'algorithme Reciprocal Rank Fusion (RRF) (Cormack *et al.*, 2009) ; (2) sélection finale par un auteur du papier via une interface dédiée, par catégorie, par ordre décroissant de score. L'étape algorithmique, dont l'heuristique est critiquable, vise uniquement à filtrer et ordonner les textes, et la décision finale reste humaine. Des critères de taille ont guidé la sélection pour maintenir une homogénéité relative entre genres, ce qui s'est avéré particulièrement difficile pour les textes de lois européens et les articles scientifiques. Pour les textes scientifiques, une présélection a été téléchargée

automatiquement depuis HAL.fr avec des critères de langue et de taille (8 pages PDF maximum). On a privilégié des textes compréhensibles par un non-spécialiste (scores *LLM-as-judge* sur Clarté et Signification, cf. Annexe A, décision finale humaine) portant sur des sujets pertinents (IA, santé, société). Les abstracts et bibliographies ont été retirés manuellement. La diversité des genres a soulevé des questions techniques spécifiques : noms d'orateurs dans les retranscriptions, références et notes de bas de page souvent reprises telles quelles dans les résumés, ce qui a impacté la fluidité.

3.3 Génération des résumés

Les textes ont fait l'objet d'un traitement commun : normalisation, conversion en JSON, extraction des entités nommées. Le choix des modèles posait un dilemme : se limiter aux modèles à l'état de l'art en 2025, LLMs notamment, aurait produit un instantané pertinent mais sans continuité avec les familles évaluées dans SummEval et avec un risque d'homogénéité des scores, alors que reproduire le setup de Fabbri et al. (Fabbri *et al.*, 2021) aurait rendu le corpus inadapté aux enjeux actuels. Nous avons retenu une voie intermédiaire, à savoir un spectre délibérément large couvrant trois générations d'approches, afin que le corpus puisse à la fois documenter l'état de l'art actuel et permettre des comparaisons qualitatives avec les travaux antérieurs. Notre protocole s'inspire du cadre de Fabbri et al. (Fabbri *et al.*, 2021) sans chercher à le répliquer : l'évolution du paysage rend toute comparaison de scores absolus peu pertinente ; l'objectif est de vérifier si les corrélations humain/métrique observées en anglais sont robustes au transfert vers le français et à la variation générique. Les dix modèles retenus couvrent trois familles :

— **4 modèles extractifs :** TextRank (Mihalcea & Tarau, 2004), et trois algorithmes de ChapsVision : Clustering , NER , NER+Clustering . Ces approches sans entraînement sont computationnellement peu coûteuses.
— **2 modèles spécialisés :** BARThez (Kamal Eddine *et al.*, 2021) et mBART (Liu *et al.*, 2020), fine-tunés sur le résumé en français.
— **4 LLMs** sélectionnés pour permettre des comparaisons ciblées : GPT-4o (OpenAI, 2024) (modèle propriétaire, référence) ; Qwen3-8B et Qwen3-32B (Yang *et al.*, 2025) (même famille, tailles différentes, pour mesurer l'effet de taille) ; Mistral-Small (comparé à Qwen3-8B, taille équivalente, pour mesurer l'effet du corpus d'entraînement multilingue de Mistral).

Comparer les performances de modèles sur une tâche suppose de choisir entre optimiser le prompt pour chaque modèle ou conserver le même prompt pour tous (Barnes *et al.*, 2025). L'objectif du travail n'étant pas d'optimiser les scores individuels des LLMs mais d'évaluer leur production dans des conditions comparables, nous avons fixé un prompt unique (cf. Annexe A) ainsi que les hyperparamètres de génération, ne faisant varier que le modèle. L'absence de vérification post-génération est délibérée : mesurer la qualité native des modèles, sans post-édition, est l'objet même de la méta-évaluation. Pour des raisons de limite de contexte, BARThez et mBART, dont l'architecture BART impose une fenêtre d'entrée de 1 024 tokens, n'ont pas pu générer de résumé pour les textes dépassant ce seuil. Ces deux modèles sont donc présents uniquement sur les textes courts, principalement MLSum. Le corpus final compte ainsi environ 900 résumés au lieu de 1 000. Les résumés présentent des taux de compression (longueur résumé / longueur source) très contrastés : LLMs entre 0,14 (Mistral-Small) et 0,25 (Qwen3-32B), extractifs entre 0,20 (TextRank) et 0,36 (NER+Clust).

4 Protocole d'annotation

L'annotation a été confiée à des professionnels de la langue formés à la tâche, rémunérés dans des conditions éthiques — ni *crowd workers* anonymes ni experts du domaine susceptibles d'être biaisés. Chaque résumé a été annoté sur 5 dimensions (échelle de Likert 1–5).

4.1 Dimensions d'évaluation

En s'appuyant sur le cadre défini par Fabbri et al. (Fabbri *et al.*, 2021), qui avaient eux-mêmes adopté les définitions de Kryściński et al. (Kryściński *et al.*, 2019), nous retenons les quatre dimensions fondamentales (cohérence, consistance, fluidité, pertinence) comme base commune.

SummEval amalgame sous le terme *fluency* deux propriétés distinctes : la correction linguistique et la facilité de compréhension. Or un texte peut être grammaticalement irréprochable tout en restant difficile à comprendre — les textes juridiques en offrent l'illustration la plus nette. Nous avons donc scindé cette dimension en *fluidité* (correction grammaticale) et *lisibilité* (accessibilité cognitive) (Gana *et al.*, 2025), choix explicitement discuté avec les annotateurs (Annexe B). Bien que ces deux dimensions soient fortement corrélées en moyenne (r = 0,912), les cas de divergence — résumés linguistiquement corrects mais cognitivement denses — sont précisément ceux que la scission permet de détecter. Les cinq critères retenus sont les suivants :

- **Cohérence (Coherence)** : qualité globale et structurelle du résumé : organisation logique, progression des idées et enchaînement naturel des phrases.
- **Consistance (Consistency)** : alignement factuel entre résumé et document source ; sanctionne hallucinations, inventions, contradictions et déformations d'information.
- **Fluidité (Fluency)** : qualité linguistique des phrases individuelles : grammaire, orthographe, ponctuation ; pénalise artefacts typographiques et mélanges de registres.
- **Lisibilité (Readability)** : facilité de compréhension, indépendamment de la correction grammaticale : phrases claires, vocabulaire accessible, syntaxe non enchevêtrée.
- **Pertinence (Relevance)** : capacité à extraire l'information utile, en évitant détails superflus, redondances et opinions subjectives ajoutées.

4.2 Annotation des erreurs linguistiques et factuelles

Au-delà des cinq dimensions, nous avons demandé aux annotateurs de relever explicitement les erreurs de langue et les erreurs de factualité, en s'inspirant de la méthodologie du benchmark FRANK (Pagnoni *et al.*, 2021) mais avec une typologie simplifiée : erreurs d'entités, erreurs de prédicats, erreurs de circonstances, erreurs de coréférences, autres (Annexe C). L'hypothèse est qu'il est difficile de quantifier la factualité de façon absolue, mais que l'on peut identifier les erreurs ponctuelles et vérifier leur corrélation avec le score de consistance attribué a priori (voir Annexe D pour un exemple concret). L'inconvénient est que cette vérification a forcé les annotateurs à confronter chaque information du résumé au texte source, ce qui a considérablement allongé le temps d'annotation pour les textes longs.

4.3 Organisation de la campagne d'annotation

Les annotateurs ont des compétences approfondies en langue française, linguistique générale et annotation multimédia. Leur tranche d'âge est de 25 à 60 ans. L'équipe est coordonnée par deux doctorants en littérature et linguistique. 4 annotateurs ont été impliqués dans la phase 1, 4 dans la phase 2 (pool d'annotateurs distincts). Chaque résumé a été annoté par 3 annotateurs différents parmi ce pool de 4 (affectation aléatoire, sans ordre fixe). Les annotateurs n'ont pas vu les résumés d'un même texte consécutivement (limitation des biais de comparaison) et n'ont pas eu connaissance du modèle générateur. La campagne a été structurée en deux phases :

— **Phase 1** (calibration) : 15 textes (120 résumés annotés sur les 10 modèles disponibles à ce stade), prise en main du guide, identification et résolution des points de difficulté. Cette phase a permis de préciser le guide d'annotation.
— **Phase 2** (annotation à grande échelle) : 367 résumés sur 90 textes (BARThez et mBART exclus des textes longs), pour 1 072 annotations réalisées. Chaque résumé a été annoté par 3 annotateurs parmi un pool de 4 (couverture : 97,4% des résumés annotés par exactement 3 annotateurs).

Le temps consacré à chaque annotation est enregistré depuis la phase 2 (l'interface de phase 1 ne capturait pas ce champ). Sur la livraison courante, la lecture du texte source prend en moyenne **21,5 min** (médiane 15 min, max 160 min), la lecture du résumé **6,5 min** (médiane 5 min), et l'annotation complète (notation + relevé d'erreurs) **24,9 min** (médiane 18 min, max 122 min). Ces valeurs illustrent la charge cognitive significative liée à la vérification des erreurs factuelles, particulièrement sur les textes longs. Les annotateurs ont généralement trouvé la tâche variée et stimulante. Des difficultés sont apparues pour les textes longs et techniques : « De fil en aiguille, chaque annotateur utilise sa propre méthodologie et s'adapte en fonction du document et de ses connaissances personnelles. »

4.4 Accord inter-annotateurs

L'accord inter-annotateurs est mesuré par l'alpha de Krippendorff ordinal, adapté aux échelles de Likert (Figure 1).

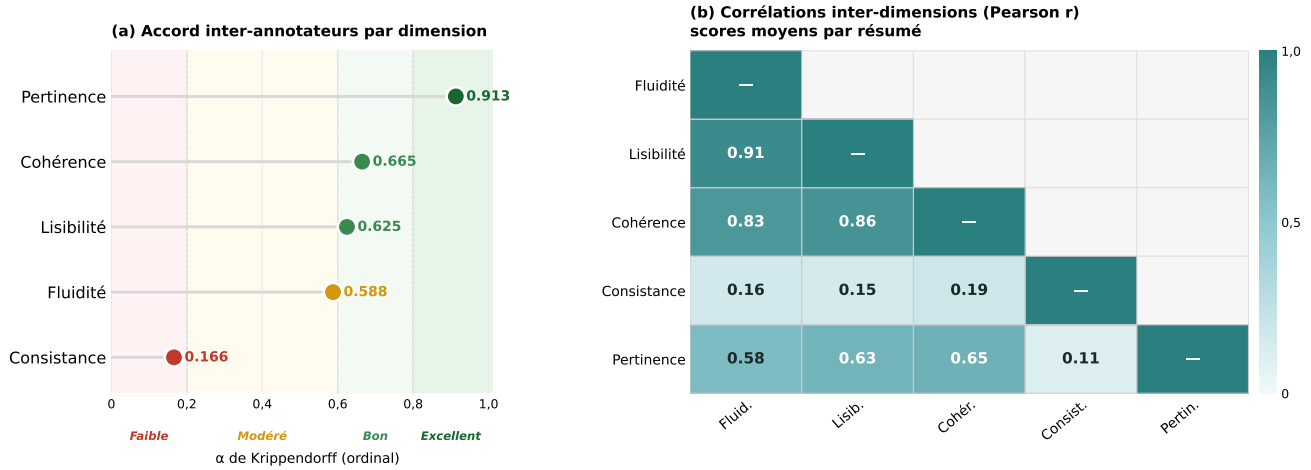


FIGURE 1 – Accord inter-annotateurs (α de Krippendorff) et corrélations inter-dimensions (Pearson).

Les résultats révèlent une forte hétérogénéité : la pertinence, dimension la plus sévèrement notée, est aussi la mieux accordée (α = 0,913), ce qui valide son usage comme critère robuste. À l'inverse, la consistance présente un accord faible (α = 0,166) malgré des scores élevés. Ce paradoxe s'explique par deux effets conjugués. D'une part, la distribution des scores est extrêmement concentrée en haut de l'échelle : la médiane est de 5,00, 60,3 % des jugements valent exactement 5 et 90,4 % sont ≥ 4 ; les désaccords inter-annotateurs portent donc presque exclusivement sur la frontière $\{4, 5\}$, là où la granularité ordinale est minimale. D'autre part, la détection d'une erreur factuelle exige une confrontation active avec le texte source, tâche cognitivement coûteuse : deux annotateurs peuvent manquer des erreurs différentes, ou identifier les mêmes tout en leur attribuant des poids différents sur l'échelle. Fabbri et al. (Fabbri *et al.*, 2021) observent eux-mêmes que la consistance est la dimension la moins reproductible du cadre SummEval. Ce faible accord plaide précisément pour le relevé systématique des erreurs factuelles en complément du score Likert (cf. section 4.2 et Annexe F). Le score final est la moyenne des 3 annotations ; pour la consistance (faible α, distribution en plateau), cet agrégateur reste valide car les désaccords portent quasi-exclusivement sur la frontière $\{4, 5\}$. La stabilité intra-phase 2 (écarts < 0,13 point entre lots) confirme l'efficacité de la calibration. Les corrélations inter-dimensions (Pearson) montrent que fluidité et lisibilité sont fortement liées (r = 0,912), tandis que la consistance reste faiblement corrélée à toutes les autres (r < 0,22), confirmant qu'elle capture une propriété distincte.

5 Résultats

Les résultats portent sur la livraison actuelle de la phase 2 (90 textes, 367 résumés, 1 072 annotations). Pour la méta-évaluation, l'unité pertinente est le nombre de résumés annotés : ces 367 résumés issus de 10 modèles, chacun évalué 3 fois, sont comparables aux 1 600 de SummEval, avec une diversité générique que ce dernier ne possède pas. Les annotations sont disponibles sur Hugging Face : CoRAFIG-phase1 et CoRAFIG-phase2. Le dashboard d'exploration est disponible sur GitHub.

5.1 Classement des modèles selon les annotateurs

La Figure 2 présente le classement des 10 modèles sur chacune des cinq dimensions et en score global (dot plots ordonnés par score) ; les valeurs complètes sont en Annexe E. LLMs et extractifs forment deux familles statistiquement séparées sur cinq dimensions (fluidité : p = $1{,}5\times10^{-38}$, r = 0,909 ; cohérence : p = $8{,}8\times10^{-44}$, r = 0,975 ; pertinence : p = $1{,}1\times10^{-47}$, r = 0,998 ; test de Mann-Whitney). Exception : la consistance, où les extractifs obtiennent des scores comparables aux LLMs (p = 0,90, n.s.), résultat mécaniquement attendu, les extractifs reproduisant les phrases source sans reformulation. On retrouve ainsi en français la hiérarchie observée sur SummEval et les travaux récents sur LLMs. L'écart est maximal sur la pertinence et strictement bimodal : tous les LLMs dépassent 4,25, tous les extractifs restent sous 1,35, là où se mesure la distance entre synthèse et sélection. Au sein des LLMs, Qwen3-32B (4,48) domine sur les dimensions linguistiques, GPT-4o (4,51) sur la pertinence et la consistance. Des différences significatives sont observées entre Mistral-Small et Qwen3-8B sur la lisibilité (p = 0,034) et la pertinence (p = 0,045), mais non sur les trois autres dimensions. La hiérarchie est cohérente entre les annotateurs : le classement relatif des modèles est identique pour les 4 annotateurs de la phase 2.

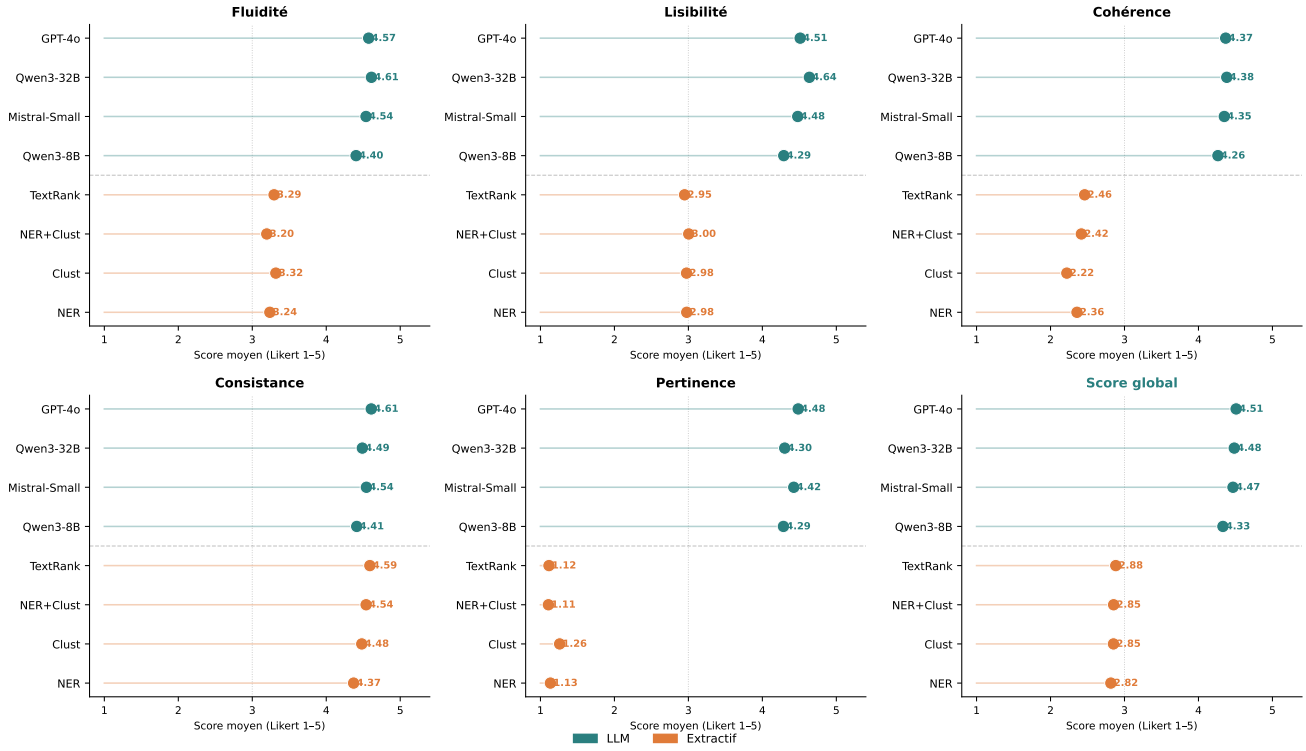


FIGURE 2 – Classement des 10 modèles par dimension et score global (phase 2, ordre fixe). BARThez et mBART, limités aux textes courts, sont inclus sur leur sous-ensemble de documents.

5.2 Corrélation avec les métriques automatiques

Sept métriques sont calculées sur les 366 résumés annotés : ROUGE-1, BERTScore-F1, BLEU, METEOR, chrF, ainsi que deux indicateurs structurels issus du framework SummEval : *coverage* (part de n-grammes du résumé présents dans le source) et *compression* (rapport longueur résumé/source). La Figure 3 présente les corrélations de Spearman ρ entre ces métriques et les cinq dimensions humaines (les corrélations de Pearson sont concordantes sauf pour *compression*, cas discuté ci-après).

Le désalignement est quasi total : sur 35 paires dimension/métrique, seules cinq corrélations atteignent $p < 0{,}05$. BERTScore-F1 est la seule métrique à déceler un signal cohérent : cohérence ($\rho = 0{,}11^{*}$) et consistance ($\rho = 0{,}12^{*}$), deux dimensions liées à la fidélité sémantique au source. *Coverage* corrèle modestement avec la pertinence ($\rho = 0{,}15^{**}$) : les résumés couvrant davantage le contenu source sont jugés plus pertinents. ROUGE-1, BLEU, METEOR et chrF n'atteignent aucun seuil de significativité sur aucune dimension. La *compression* constitue un cas limite : Pearson $r = +0{,}16^{**}$ mais Spearman $\rho = -0{,}11^{*}$ sur la pertinence, signe opposé révélant un effet de levier dû aux résumés de Mistral-Small qui excèdent la longueur du source sur les textes courts de MLSum. Ces résultats confirment que les métriques de recouvrement lexical et sémantique calibrées sur l'anglais échouent à capturer le jugement humain en français multi-genre. Les résumés de référence (Table 1) constituent la vérité terrain pour les métriques à recouvrement ; la moyenne des 3 jugements humains par résumé tient lieu de gold standard pour la méta-évaluation, conformément au protocole de Fabbri et al. (Fabbri *et al.*, 2021). Une extension aux métriques sans référence (QuestEval, SummaQA) via le framework AllSummedUp (Herserant & Guigue, 2025) est prévue.

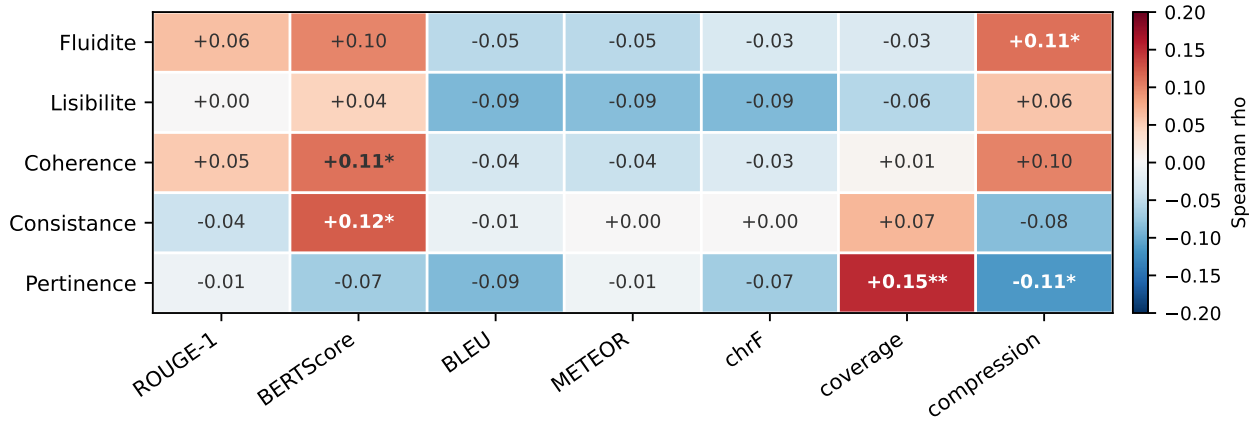


FIGURE 3 – Corrélations de Spearman entre métriques automatiques et jugements humains (CoRAFIG, $n = 366$ résumés). $^{*}p < 0{,}05$, $^{**}p < 0{,}01$.

5.3 Influence du genre et de la longueur sur la qualité perçue

Les scores par genre révèlent des tendances cohérentes avec nos hypothèses : Les textes juridiques présentent la pertinence la plus faible (2,66) et la consistance la plus élevée (4,64), les résumés restent factuellement fidèles mais peinent à capturer l'essentiel de textes longs et denses. Les textes de presse obtiennent les meilleures fluidité et lisibilité (4,07 / 3,90), reflet de la familiarité des modèles avec ce genre très représenté à l'entraînement. Sur la dimension longueur, les corrélations de Pearson entre la longueur du texte source (en mots) et les scores par dimension restent faibles sur toutes les dimensions (Fluidité r = $-0{,}10$; Lisibilité r = $-0{,}10$; Cohérence r = $-0{,}05$; Pertinence r = +0,11 ; Consistance r = +0,16 ; $|r| < 0{,}16$). Ce résultat confirme que les scores humains ne dépendent pas significativement de la longueur des textes sources, contrairement au constat fait dans UniSumEval (Lee *et al.*, 2024).

5.4 Analyse des erreurs

Sur les 1 072 annotations, 1 835 labels d'erreur ont été relevés (424 annotations avec au moins une erreur, soit 39,6% ; voir Annexe F). Les **erreurs linguistiques dominent** (97,2 %) et opposent nettement les deux familles : les modèles extractifs cumulent entre 8 et 13 erreurs de langue par résumé (NER+Clust : 12,53 ; Clust : 13,06), conséquence de la juxtaposition de phrases hors contexte, tandis que les LLMs en produisent très peu (GPT-4o : 0,30 ; Qwen3-32B : 0,70). Les **erreurs factuelles** (entités, prédicats, coréférences, circonstances) présentent le profil inverse : quasi-absentes des modèles purement extractifs (NER+Clust : 2 labels typés sur l'ensemble du corpus), elles croissent avec la taille décroissante des LLMs (GPT-4o : 0,09/résumé ; Qwen3-32B : 0,19 ; Mistral-Small : 0,26 ; Qwen3-8B : 0,32) et représentent surtout des hallucinations de dates, de noms ou de chiffres. **BARThez atteint le taux le plus élevé** (0,35/résumé, 16 % des résumés affectés), dépassant tous les LLMs malgré son fine-tuning spécialisé sur corpus français ; ce paradoxe illustre la propension des modèles seq2seq supervisés à interpoler les données d'entraînement plutôt qu'à en vérifier la cohérence factuelle. mBART présente à l'inverse un taux très faible (0,04/résumé), proche de GPT-4o. La présence d'au moins une erreur factuelle est associée à une consistance moyenne de 3,38, contre 4,50 pour les résumés sans erreur ($\Delta = 1{,}12$, Mann-Whitney $p = 9{,}6 \times 10^{-12}$). Cet écart confirme

que le score de consistance sous-détecte les erreurs ponctuelles lorsqu'elles sont peu nombreuses, ce qui justifie la couche d'annotation séparée (cf. section 4.2 et Annexe F).

5.5 Comportement et subjectivité des annotateurs

Les quatre annotateurs impliqués dans la phase 2 (livraison courante) présentent des profils de sévérité distincts, malgré un niveau d'expertise homogène. L'équipe de phase 1 comportait un effectif différent. L'annotateur A5 est globalement le plus indulgent (global 3,84 vs 3,35–3,62 pour les autres), mais l'examen par dimension révèle que ce biais est très inégalement réparti : il est maximal sur la consistance et la cohérence, et quasi nul sur la pertinence. Ce biais additif ne remet pas en cause le classement relatif des modèles, qui est cohérent chez les 4 annotateurs.

TABLE 2 – Scores moyens par annotateur et par dimension (phase 2).

Annotateur	Fluidité	Lisibilité	Cohérence	Consistance	Pertinence	**Global**
A4	3,787	3,708	3,174	3,719	2,337	**3,345**
A5	4,259	4,034	3,891	4,782	2,214	**3,836**
A6	3,854	3,732	3,496	4,439	2,557	**3,616**
A7	3,913	3,672	3,284	4,600	2,234	**3,541**

6 Discussion et perspectives

Les résultats de la livraison actuelle permettent de consolider statistiquement les tendances observées dans la version préliminaire. Sur le plan conceptuel, le protocole suppose implicitement que le résumé est une tâche homogène, alors que le domaine évolue vers des résumés guidés et orientés ; la diversité générique compense partiellement cette limite, sans l'effacer. Les textes juridiques et financiers révèlent deux stratégies empiriques : résumé *synthétique* (reformulation) et résumé *énumératif* (compression séquentielle sans réorganisation). Le second obtient une bonne consistance mais une cohérence et une pertinence faibles (EurLex : 4,64 vs 2,66), ce qui soulève la question du destinataire et plaide pour des protocoles à prompt guidé que ce corpus, à prompt unique, n'est pas conçu pour tester. Malgré ce morcellement, qualité de langue et fidélité au source restent essentiels quelle que soit la tâche. Nos résultats suggèrent qu'un modèle 8B bien calibré peut approcher un modèle propriétaire sur les dimensions linguistiques, l'écart intra-LLM restant limité sur la pertinence et la cohérence. Les 1 072 annotations fournissent des préférences pairées exploitables pour un fine-tuning tâche-spécifique par DPO (Rafailov *et al.*, 2023) sur le résumé en français, domaine où quelques centaines à quelques milliers de paires suffisent (Zhou *et al.*, 2023). Une limite à noter : les modèles n'ont pas été appliqués à l'ensemble des documents (couverture de 40,8% des paires document/système prévues), les comparaisons inter-modèles portent donc sur des sous-ensembles de documents partiellement distincts. L'extension à un spectre plus large de LLMs (tailles et architectures variées) et une comparaison directe des corrélations humain/métrique avec SummEval en anglais constituent les priorités de la prochaine phase. Enfin, CoRAFIG est une invitation à développer des métriques multilingues frugales vérifiables à la fois sur SummEval et sur des ressources natives, afin de ne plus calibrer l'évaluation sur le seul étalon CNN/DailyMail.

Références

AUMILLER D., CHOUHAN A. & GERTZ M. (2022). EUR-Lex-Sum : A Multi- and Cross-lingual Dataset for Long-form Summarization in the Legal Domain. DOI : 10.48550/arXiv.2210.13448.

BACHEY F., RODRIGUES C. & BOSSARD A. (2025). Étude critique du corpus cnn/dailymail pour le résumé automatique. In *Actes de CORIA-TALN-RJCRI-RECITAL 2025*, p. 348–359 : Association pour le Traitement Automatique des Langues.

BARNES J., PEREZ N., BONET-JOVER A. & ALTUNA B. (2025). Summarization Metrics for Spanish and Basque : Do Automatic Scores and LLM-Judges Correlate with Humans ?

BRAUN S., VASILYEV O., ISKENDER N. & BOHANNON J. (2021). Does Summary Evaluation Survive Translation to Other Languages ? DOI : 10.48550/arXiv.2109.08129.

CLARK E., RIJHWANI S., GEHRMANN S., MAYNEZ J., AHARONI R., NIKOLAEV V., SELLAM T., SIDDHANT A., DAS D. & PARIKH A. (2023). SEAHORSE : A Multilingual, Multifaceted Dataset for Summarization Evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 9397–9413 : Association for Computational Linguistics. DOI : 10.18653/v1/2023.emnlp-main.584.

CORMACK G. V., CLARKE C. L. A. & BUETTCHER S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 758–759 : ACM. DOI : 10.1145/1571941.1572114.

DAI X., KARIMI S. & FANG B. (2024). A Critical Look at Meta-evaluating Summarisation Evaluation Metrics. In *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 14795–14808 : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-emnlp.869.

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

FORDE J. Z., ZHANG R., SUTAWIKA L., AJI A. F., CAHYAWIJAYA S., WINATA G. I., WU M., EICKHOFF C., BIDERMAN S. & PAVLICK E. (2024). Re-Evaluating Evaluation for Multilingual Summarization. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 19476–19493 : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.1085.

GANA B., ALLENDE-CID H., RÜPING S., BECERRA-ROZAS M. & ZAMORA J. (2025). A systematic review of long document summarization methods : Evaluation metrics and approaches. *Neurocomputing*, **655**, 131287. DOI : 10.1016/j.neucom.2025.131287.

GAO M. & WAN X. (2022). DialSummEval : Revisiting Summarization Evaluation for Dialogues. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 5693–5709 : Association for Computational Linguistics. DOI : 10.18653/v1/2022.naacl-main.418.

GAO Y., XU G., WANG Z. & COHAN A. (2024). Bayesian Calibration of Win Rate Estimation with LLM Evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 4757–4769. DOI : 10.18653/v1/2024.emnlp-main.273.

GOYAL T., LI J. J. & DURRETT G. (2022). News Summarization and Evaluation in the Era of GPT-3.

GUO Z., JIN R., LIU C., HUANG Y., SHI D., YU L., LIU Y., LI J., XIONG B. & XIONG D. (2023). Evaluating Large Language Models : A Comprehensive Survey.

HERSERANT T. & GUIGUE V. (2025). Allsummedup : un framework open-source pour comparer les métriques d'évaluation de résumé. In F. BECHET, A.-G. CHIFU, K. PINEL-SAUVAGNAT, B. FAVRE, E. MAES & D. NURBAKOVA, Éds., *Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM)*, p. 11–21 : ATALA \textbackslash\textbackslash& ARIA.

KAMAL EDDINE M., TIXIER A. & VAZIRGIANNIS M. (2021). BARThez : A Skilled Pretrained French Sequence-to-Sequence Model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 9369–9390 : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.740.

KRYŚCIŃSKI W., MCCANN B., XIONG C. & SOCHER R. (2019). Evaluating the Factual Consistency of Abstractive Text Summarization. DOI : 10.48550/arXiv.1910.12840.

LEE Y., YUN T., CAI J., SU H. & SONG H. (2024). UniSumEval : Towards Unified, Fine-grained, Multi-dimensional Summarization Evaluation for LLMs. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 3941–3960 : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-emnlp.227.

LIU Y., FABBRI A., CHEN J., ZHAO Y., HAN S., JOTY S., LIU P., RADEV D., WU C.-S. & COHAN A. (2024). Benchmarking Generation and Evaluation Capabilities of Large Language Models for Instruction Controllable Summarization. In *Findings of the Association for Computational Linguistics : NAACL 2024*, p. 4481–4501 : Association for Computational Linguistics. DOI : 10.18653/v1/2024.findings-naacl.280.

LIU Y., GU J., GOYAL N., LI X., EDUNOV S., GHAZVININEJAD M., LEWIS M. & ZETTLEMOYER L. (2020). Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, **8**, 726–742. DOI : 10.1162/tacl_a_00343.

MIHALCEA R. & TARAU P. (2004). TextRank : Bringing Order into Text. In D. LIN & D. WU, Éds., *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, p. 404–411 : Association for Computational Linguistics.

OPENAI (2024). GPT-4o System Card. DOI : 10.48550/arXiv.2410.21276.

PAGNONI A., BALACHANDRAN V. & TSVETKOV Y. (2021). Understanding Factuality in Abstractive Summarization with FRANK : A Benchmark for Factuality Metrics. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 4812–4829 : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.383.

PU X., GAO M. & WAN X. (2023). Summarization is (Almost) Dead.

RAFAILOV R., SHARMA A., MITCHELL E., MANNING C. D., ERMON S. & FINN C. (2023). Direct Preference Optimization : Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36.

RENNARD V., SHANG G., GRARI D., HUNTER J. & VAZIRGIANNIS M. (2023). FREDSum : A Dialogue Summarization Corpus for French Political Debates. In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 4241–4253 : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.280.

SCIALOM T., DRAY P.-A., LAMPRIER S., PIWOWARSKI B. & STAIANO J. (2020). MLSUM : The Multilingual Summarization Corpus. In *Proceedings of the 2020 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), p. 8051–8067 : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.647.

WANG A., PANG R. Y., CHEN A., PHANG J. & BOWMAN S. R. (2022). SQuALITY : Building a Long-Document Summarization Dataset the Hard Way. DOI : 10.48550/arXiv.2205.11465.

YANG A., LI A., YANG B., ZHANG B., HUI B., ZHENG B., YU B., GAO C., HUANG C., LV C., ZHENG C., LIU D., ZHOU F., HUANG F., HU F., GE H., WEI H., LIN H., TANG J., YANG J., TU J., ZHANG J., YANG J., YANG J., ZHOU J., ZHOU J., LIN J., DANG K., BAO K., YANG K., YU L., DENG L., LI M., XUE M., LI M., ZHANG P., WANG P., ZHU Q., MEN R., GAO R., LIU S., LUO S., LI T., TANG T., YIN W., REN X., WANG X., ZHANG X., REN X., FAN Y., SU Y., ZHANG Y., ZHANG Y., WAN Y., LIU Y., WANG Z., CUI Z., ZHANG Z., ZHOU Z. & QIU Z. (2025). Qwen3 Technical Report. DOI : 10.48550/arXiv.2505.09388.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). BERTScore : Evaluating Text Generation with BERT. DOI : 10.48550/arXiv.1904.09675.

ZHANG T., LADHAK F., DURMUS E., LIANG P., MCKEOWN K. & HASHIMOTO T. B. (2024). Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, **12**, 39–57. DOI : 10.1162/tacl_a_00632.

ZHOU C., LIU P., XU P., IYER S., SUN J., MAO Y., MA X., EFRAT A., YU P., YU L., ZHANG S., GHOSH G., LEWIS M., ZETTLEMOYER L. & LEVY O. (2023). LIMA : Less Is More for Alignment. *Advances in Neural Information Processing Systems*, **36**.

ZMANDAR N., DAUDERT T., AHMADI S., EL-HAJ M. & RAYSON P. (2022). CoFiF Plus : A French Financial Narrative Summarisation Corpus. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1622–1639 : European Language Resources Association.

Annexe A — Prompts utilisés

A.1 — Prompt de génération des résumés

commun à tous les LLMs, zéro-shot

Prompt de génération

```
Tu es un assistant qui génères des résumés des textes qui lui sont soumis.
Fais un résumé sur la totalité du texte en faisant une paraphrase du texte.
Ne reprends pas les phrases du texte directement.
Essaie de conserver autant que possible les événements, les dates
et valeurs numériques, ainsi que les noms de lieux, d'organisations et de personnes.
N'ajoute pas dans le résumé des faits qui n'existent pas dans le texte.
Ce résumé doit être en français. Le résumé ne doit pas être en anglais.
Résume le texte sans phrases d'introduction ni phrases supplémentaires
comme « Voici un résumé du texte » ou toute autre dérivée.
```

Contraintes clés : *paraphrase impose la reformulation (évite la copie) · entités/dates/valeurs réduit les hallucinations factuelles · no hallucination instruction explicite · verrou FR assure la langue de sortie · format clean supprime les méta-phrases d'introduction*

A.2 — Prompt *LLM-as-judge*

sélection des 100 textes sources par genre

Les instructions sont en anglais : les LLMs sont mieux alignés pour l'évaluation structurée (*role prompting*) dans cette langue ; la génération cible reste en français (A.1). Le template est paramétré par genre via quatre variables :

Template *LLM-as-judge*

```
You are a {{ role }}.
Your goal is to {{ goal }}.
Analyze the text provided.
CRITICAL INSTRUCTIONS:
- {{ calibration }}
- A score of 8+ must be EXCEPTIONAL.
Dimensions to Score:
{% for dim in dimensions %}
{{ loop.index }}. {{ dim.name }} (1-10): {{ dim.definition }}
{% endfor %}
Output strictly valid JSON with fields:
topics (list), summary (str), <dim>_score (int 1-10) for each dimension.
TEXT TO ANALYZE: {{ text }}
```

Variable	Presse (MLSum)	Financier (CoFiF+)	Scientifique (HAL)	Juridique (EurLex)
Rôle	Rédacteur en chef d'un corpus de haute qualité	Analyste financier senior en banque d'investissement	Reviewer pour revue académique à fort impact	Juriste senior dans un cabinet européen
Objectif	Filtrer les textes sans intérêt ; identifier les textes remarquables	Identifier les rapports à forte valeur stratégique ; écarter le *boilerplate*	Évaluer qualité, originalité et portée du texte	Identifier règlements et décisions à fort impact
Calibration	Sois sévère. Le texte moyen obtient un 5. Un score $\geq$8 doit être exceptionnel. Un 10 est réservé aux textes historiquement significatifs.	La plupart des rapports annuels sont du remplissage générique (score 3–4). Un score $\geq$8 exige une clarté stratégique exceptionnelle.	*(ancrage sémantique des extrêmes, sans directive de sévérité)*	*(ancrage sémantique des extrêmes, sans directive de sévérité)*
Dimensions	Intérêt · Humour · Utilité · Complexité · Intensité émotionnelle	Vision stratégique · Clarté financière · Densité informationnelle · Force du sentiment	Originalité · Clarté/Structure · Signification · Pertinence · Évaluation globale	Impact juridique · Clarté des obligations · Complexité administrative · Pertinence citoyenne

FREDSum (retranscriptions) et littéraire : sélection manuelle directe, sans notation algorithmique préalable. Les scores sont agrégés par Reciprocal Rank Fusion pour ordonner les candidats.

Annexe B — Extraits du journal des annotateurs (phase 2, 30/03/2026)

Extraits du rapport remis par l'équipe d'annotation à l'issue des trois premiers lots de la phase 2, illustrant les points de friction rencontrés et les décisions prises en cours de campagne.

Charge cognitive — textes longs

« Un des principaux freins à l'avancée rencontrés est le temps pris pour la vérification des détails dans le résumé. Vérifier l'exactitude des résumés implique un retour au texte source pour confirmer la véracité de certains éléments très précis (dates, noms, titres, fonctions…). Ce mouvement de comparaison ralentit la production, surtout dans les textes longs ou comportant beaucoup d'éléments de ce genre. Par exemple, les textes de lois et autres directives européennes, composés de plusieurs dizaines d'alinéas faisant eux-mêmes référence à d'autres textes de lois, sont très laborieux à lire et à annoter. »
— Annotateur 3

Artefacts des modèles extractifs

« Le LLM semble également rencontrer des difficultés lorsque le texte source est une retranscription d'un échange oral, puisqu'il laisse apparaître les noms des orateurs. On retrouve aussi des difficultés si le texte source contient des références ou notes de bas de page, celles-ci sont souvent inclues dans le résumé, impactant drastiquement la fluidité. »
— Annotateur 2

« La mise en page d'un résumé peut parfois être chaotique : espaces, retours à la ligne, mots isolés, chiffres isolés. »
— Annotateur 2

Paradoxe consistance / pertinence

« Dans la mesure où un résumé est un copié-collé du texte source, est-il entendable d'annoter 5 étoiles en factualité, malgré la note catastrophique en pertinence ? Bien qu'il ne soit pas pertinent, un résumé copié-collé est factuellement fidèle au texte source. En ce sens, il serait possible de retrouver des évaluations qualitatives comprenant des critères tous très mauvais, sauf la factualité. »
— Annotateur 1

Résumés d'énumération — textes juridiques

« Dans le cas de longs textes comportant beaucoup d'énumérations comme les textes de loi, il existe deux types de résumés : ceux qui reprennent les grandes idées du texte, en en donnant seulement un aperçu, et les résumés qui raccourcissent la plupart des points soulevés mais ne les réorganisent pas. Comment traiter un résumé d'énumération ? Un résumé sous forme d'énumération est-il considéré comme étant de plus grande valeur qu'une synthèse d'idée ? »
— Annotateur 3

Limites de la typologie d'erreurs

« La catégorie « Autre erreur » a notamment une fois été utilisée pour signaler l'emploi erroné du mot « amour » à la place de l'idée d'altruisme et de bonté. Il ne s'agit pas à proprement parler d'un problème de « Circonstance » car l'idée est présente, mais il s'agit plutôt d'un problème de synonyme et de compatibilité des termes. »
— Annotateur 3

Conclusion du rapport

« Nous avons essayé d'avancer un peu plus rapidement par rapport à la première phase d'annotation, tout en gardant présent à l'esprit la cohérence avec les consignes du guide. Finalement, malgré l'existence, encore, de quelques questions, nous continuons à annoter d'une manière pragmatique et fidèle au guide les résumés. »
— Équipe d'annotation

Annexe C — Guide d'annotation

I) Cadre général de la mission

1. Présentation du projet

Dans le cadre du projet DGA RAPID RAFFAL, qui s'intéresse à l'évaluation de la tâche de résumé automatique en français, nous souhaitons constituer un corpus de référence de résumés annotés permettant de vérifier la corrélation de différentes métriques automatiques avec les scores humains, et de valider ainsi la pertinence des métriques développées par les chercheurs sur des résumés en langue française. Le corpus doit être suffisamment vaste pour calculer des corrélations (Spearman, Pearson) fiables, et présenter une qualité variée (de médiocre à excellent) afin de vérifier la capacité des métriques à juger la qualité des résumés sans biais distributionnel.

2. Intérêt de la mission d'annotation

La quantité d'informations à assimiler pour effectuer une tâche ou prendre une décision ne cesse d'augmenter. Avec l'amélioration des systèmes d'intelligence artificielle générative, l'usage de résumés de documents générés automatiquement devient de plus en plus fréquent dans le monde du travail. Il convient de s'assurer que ces résumés sont de bonne qualité et correspondent aux attentes des utilisateurs finaux. Les chercheurs en TAL mettent en place des métriques qui servent à mesurer la qualité d'un résumé selon plusieurs critères (cohérence, factualité, lisibilité), mais il est nécessaire de pouvoir calibrer ces métriques et vérifier qu'elles retranscrivent bien les attentes humaines. Pour cela, il est nécessaire de disposer de corpus présentant différents résumés d'un même texte, de qualité différente, et notés sur plusieurs critères. Si de tels corpus existent en anglais, ils font défaut en français — d'où le présent travail.

3. Comprendre les attentes des utilisateurs

La lecture de documents peut être chronophage et il est souvent nécessaire de prendre connaissance de l'essentiel en un temps très limité. D'où l'emploi fréquent de résumés, notamment dans les administrations, en préambule de publications. L'utilisateur attend en général d'un résumé :

- Qu'il soit plus court que le document original (taille fixe ou proportionnelle).
- Qu'il soit bien formulé, dans une langue claire et intelligible, sans reproduire la technicité éventuelle du texte de départ.
- Qu'il contienne les informations essentielles du texte de façon non redondante.
- Qu'il couvre l'ensemble du texte : début et fin sont souvent plus retranscrits, mais des sections entières ne doivent pas être laissées de côté.

À ces attentes génériques peuvent s'ajouter des attentes spécifiques liées au genre du texte ou à l'usage pratique qui sera fait du résumé.

II) Portée de la tâche

La plupart des études portant sur l'annotation de résumés portent essentiellement sur l'anglais et sur des textes de presse. Les résumés présentés comme « référence » n'étant parfois qu'un simple abstract de l'article, cette catégorie de résumé n'est pas représentative des usages modernes du résumé automatique. Un des objectifs de notre travail est de pallier ce manque de ressources pour des ressources plus diversifiées. Toutefois, pour que nos résultats soient comparables à ceux d'études précédentes, il est nécessaire de conserver un nombre significatif d'articles de presse dans le corpus final.

Le corpus se composera de 100 textes chacun résumés automatiquement par 10 modèles d'IA différents, pour un total d'environ 1 000 résumés à annoter, issus de six genres :

- 50 actualités portant sur des sujets divers ;
- 20 rapports financiers ;
- 20 textes juridiques ;
- 20 articles scientifiques provenant de la plateforme HAL ;
- 20 retranscriptions de dialogues ou de discours ;

— 20 extraits de livres libres de droit (XIX^{e}–XX^{e} siècle, roman, essai, théâtre).

Chaque texte fera entre 2 et 10 pages. Les annotateurs prendront connaissance du texte original, puis évalueront plusieurs résumés du même texte (fournis dans un ordre aléatoire, sans ordre fixe). Les annotateurs n'ont pas vu les résumés d'un même texte consécutivement (limitation des biais de comparaison) et n'ont pas eu connaissance du modèle générateur. Les résumés ne doivent pas être comparés les uns aux autres mais évalués individuellement.

III) Critères d'annotation retenus

On demandera aux annotateurs d'évaluer la qualité des résumés en réalisant trois tâches complémentaires :

* Étude des défauts de langue

On leur demandera d'abord de repérer dans le texte les erreurs de langue jugées gênantes pour un locuteur natif (erreur d'accord ou de conjugaison, orthographe sur un mot courant, syntaxe d'un segment incompréhensible).

* Étude des erreurs factuelles

On demande aux annotateurs de repérer les erreurs factuelles introduites dans les résumés et de les classer selon la typologie suivante :

— **Erreur d'entité** : un nom propre, un nombre, une date ou une organisation est incorrect ou inventé. *Ex. : l'article dit « Emmanuel Macron », le résumé écrit « François Hollande ».*
— **Erreur de prédicat** : l'action ou l'événement décrit est faux. *Ex. : l'article dit « l'équipe a gagné », le résumé dit « l'équipe a perdu ».*
— **Erreur de circonstance** : les informations de contexte (lieu, date, moment, condition) sont fausses ou inventées. *Ex. : l'article dit « la réunion a eu lieu à Paris le lundi », le résumé dit « la réunion a eu lieu à Lyon le mardi ».*
— **Erreur de coréférence** : un pronom ne renvoie pas au bon référent. *Ex. : « Marie a félicité Paul. Elle était heureuse. » → le résumé comprend « Paul était heureux ».*
— **Autres erreurs** : tout problème factuel ne rentrant pas dans les catégories précédentes (mélange d'événements, contradiction globale, incohérence logique).

Chaque erreur doit être identifiée et étiquetée individuellement. Il est tout à fait possible (et même souhaité !) que les résumés ne contiennent aucune erreur factuelle.

* Notation sur 5 critères (échelle de Likert 1–5)

Quatre des critères correspondent à ceux largement utilisés pour la méta-évaluation depuis Fabbri et al. (2020) et le corpus SummEval. Une distinction est introduite pour préciser le sens à donner au terme « fluidité ».

a) Cohérence (*Coherence*)

La cohérence désigne la qualité globale du résumé, en ce que les phrases forment un tout cohérent : ce n'est pas seulement un assemblage de faits sans lien.

— Que le résumé soit bien structuré et bien organisé, avec une progression logique d'idées.
— Que les phrases s'enchaînent de façon logique, avec des connecteurs textuels ou au moins des transitions implicites qui guident le lecteur.
— Que les erreurs grammaticales ou factuelles n'affectent pas la cohérence (on sanctionne uniquement l'enchaînement logique).
— Qu'un résumé qui apparaît comme une simple liste d'événements sans continuité soit pénalisé.

L'idée est de mesurer si le résumé « se tient » comme un texte, s'il raconte quelque chose de fluide et non s'il est juste un collage d'informations.

Remarques : ne pas sanctionner les erreurs factuelles ou grammaticales ici ; si une phrase est isolée (un seul énoncé), l'absence de connecteur n'est pas pénalisée — l'essentiel est le flux global ; si tout le résumé est une

TABLE 3 – Barème de cohérence.

Note	Description	Ce qu'on attend / à sanctionner
1	Résumé très désordonné : « liste à puces » sans lien	Pas de structure discernable, les phrases ou termes semblent placés au hasard
2	Liste d'événements incohérente	Il y a une « liste », mais les éléments ne se lient pas (*ex. « Clavier. Regarder TV. »*)
3	Liste d'événements mieux formulée	Les items sont bien formulés, mais toujours sans transition ou ordre clair
4	Cohérence implicite	Les phrases s'enchaînent de façon logique (temps, cause, progression), sans connecteurs explicites
5	Cohérence explicite avec connecteurs	Utilisation claire de marqueurs textuels (*d'abord, ensuite, donc, car, cependant...*) pour lier les idées

suite d'éléments sans lien, la note doit être basse (1 ou 2) ; un résumé qui débute par un paragraphe puis liste quelques événements sans lien perd en cohérence (3 ou 4 selon le contexte).

b) Consistance / Factualité (*Consistency*)

La consistance (souvent appelée factualité) désigne l'alignement factuel entre le résumé et son document source.

- Un résumé factuellement consistant ne contient que des énoncés qui sont étayés, impliqués ou *entailed* par le texte source.
- Tout ajout inventé (hallucination), contradiction ou déformation d'un fait doit être considéré comme une erreur de consistance.
- Les expressions temporelles relatives (*aujourd'hui, hier, cette année*) doivent être jugées dans le contexte de l'article.
- La consistance est différente de la cohérence : ici on évalue si le contenu est véridique par rapport au texte, non pas s'il « se tient » comme un discours fluide.

TABLE 4 – Barème de consistance.

Note	Description	Exemple typique
1	Le résumé ne reprend **aucune idée correcte** du texte source	Résumé totalement inventé
2	Le résumé contient **beaucoup d'informations incorrectes** ou inventées	Plusieurs phrases décrivent des faits absents ou faux
3	Le résumé contient **plusieurs erreurs factuelles** (mais reprend aussi des infos correctes)	2–3 faits majeurs mal restitués
4	Le résumé contient **une seule erreur factuelle**	Tout est exact sauf une date ou un nom
5	Le résumé est **entièrement fidèle** : toutes les informations proviennent du texte, aucune hallucination ni contradiction	Alignement parfait avec la source

À sanctionner : les ajouts inventés (*hallucinations*) ; les contradictions avec le texte source ; les erreurs de chiffres, noms, dates, lieux, actions. *À ne pas sanctionner :* les reformulations stylistiques qui ne changent pas le sens ; l'omission d'informations (c'est la couverture, pas la consistance) ; les expressions temporelles génériques cohérentes avec le contexte.

c) Pertinence (*Relevance*)

La pertinence désigne la capacité du résumé à sélectionner les informations importantes du texte source et à éviter les informations superflues.

— Le résumé doit inclure les faits ou idées essentiels que le lecteur doit connaître pour comprendre le texte source, sans diluer par des détails triviaux.
— Il doit éviter les redondances, les digressions inutiles, les métaréférences générées par le modèle (ex. *« voici le résumé »*) ou les opinions personnelles non présentes dans le texte.
— Si une phrase du résumé n'apporte pas de valeur informative ou dérive du sujet principal, elle doit être pénalisée.

TABLE 5 – Barème de pertinence.

Note	Description	Ce qu'on attend ou sanctionne
1	Le résumé est essentiellement hors sujet ou reprend le texte intégral	Il ne sélectionne aucune information essentielle, devient un « copié » ou ajoute beaucoup de contenu inutile
2	Le résumé contient beaucoup d'informations non pertinentes	Plusieurs passages sont hors sujet ou ajoutent des détails triviaux
3	Le résumé mélange informations utiles et non utiles	Une part notable des phrases est pertinente, mais beaucoup d'autres sont redondantes ou secondaires
4	Le résumé contient principalement des informations utiles	Très peu de passages non pertinents ou redondants, et l'essentiel du résumé est centré sur les faits clés
5	Tout est pertinent	Toutes les phrases du résumé contribuent aux points essentiels du texte source, aucun contenu accessoire

Remarques : ne pas sanctionner une omission de détail (c'est la couverture, pas la pertinence) ; sanctionner les répétitions inutiles ; sanctionner les phrases injonctives ou métaréférentielles inventées par le modèle (*« Voici un résumé : . . . »*) ; sanctionner les opinions subjectives ajoutées absentes du texte original. En cas de doute : *« Cette phrase contribue-t-elle à la compréhension du sujet principal ou non ? »*

d) Fluidité (*Fluency*)

La fluidité renvoie à la qualité linguistique des phrases prises individuellement.

— Les phrases d'un résumé fluide doivent respecter les règles de grammaire, d'orthographe et de ponctuation du français.
— Le texte ne doit pas contenir d'artefacts typographiques (problèmes de majuscules, ponctuation absente, caractères inappropriés).
— Le mélange de registres, dialectes ou langues doit être sanctionné s'il est inattendu et rend la lecture inconfortable.
— On n'évalue pas le style ou la lisibilité globale (simplicité), mais seulement si le texte est correctement formé en français.

Cette dimension correspond à la question « Est-ce du bon français ? »

TABLE 6 – Barème de fluidité.

Note	Description	Exemple typique
1	Beaucoup d'erreurs majeures de grammaire/orthographe, ou texte pas en français / mélange chaotique de langues	*« Marie aller marché demain. Il bon. »*
2	Plusieurs erreurs graves, ou mélange de registres/dialectes rendant la lecture pénible	*« Elle vont au cinéma et il être contents »*
3	Quelques erreurs mineures (accords, accents, fautes d'orthographe) ou mélange de variétés linguistiques un peu artificiel	*« Les enfants jouaient dans le parc et il étaient content »*
4	Très peu d'erreurs, qui ne gênent pas la compréhension	*« Les enfants jouaient dans le parc et ils étaient content »*
5	Aucune erreur grammaticale ou orthographique ; registre et variété linguistique cohérents	*« Les enfants jouaient dans le parc et ils étaient contents »*

e) Lisibilité (*Readability*)

La lisibilité désigne la facilité avec laquelle un lecteur peut lire et comprendre un résumé sans effort excessif.

— Un résumé lisible se caractérise par des phrases claires, pas trop longues, bien structurées, avec un vocabulaire accessible au lecteur cible.
— Même si toutes les phrases sont grammaticalement correctes (bonne fluidité), un texte peut être peu lisible s'il est surchargé, trop technique ou alourdi par des constructions complexes.
— La lisibilité ne sanctionne pas la correction linguistique, mais le confort de lecture et la clarté du style.

Cette dimension correspond à la question « Est-ce que ce texte se lit facilement et se comprend sans effort ? »

TABLE 7 – Barème de lisibilité.

Note	Description	Exemple typique
1	Texte extrêmement difficile à lire : phrases très longues, syntaxe enchevêtrée, vocabulaire opaque ou jargon technique	*« La systématisation heuristique multifactorielle des méthodologies paradigmatiques…»*
2	Texte difficile à lire : nombreuses phrases longues, vocabulaire compliqué ou tournures confuses	*« L'organisation procédurale fut subséquemment engagée dans une dynamique de contextualisation…»*
3	Lisibilité moyenne : phrases grammaticalement correctes mais trop denses ou style maladroit, nécessite un effort soutenu	*« Le rapport présente plusieurs résultats d'importance, dont certains pourraient être perçus comme redondants…»*
4	Texte globalement facile à lire, malgré quelques formulations lourdes ou une phrase trop longue	*« Le rapport présente les résultats essentiels et souligne les aspects majeurs, même si certains détails alourdissent légèrement la lecture. »*
5	Texte très clair et agréable à lire : phrases courtes à moyennes, style fluide, vocabulaire simple et approprié	*« Le rapport présente clairement les résultats principaux et explique les points importants de façon concise. »*

À pénaliser : phrases trop longues ou à la syntaxe compliquée ; utilisation excessive de mots rares, jargon ou termes techniques non expliqués ; redondances qui alourdissent inutilement la lecture ; formulations maladroites ou obscures qui obligent à relire. *À ne pas pénaliser :* la simple omission d'informations (c'est la couverture, pas la lisibilité) ; un style neutre ou peu élégant (tant qu'il reste clair) ; de légères répétitions si elles n'entravent pas la compréhension ; de petites erreurs d'orthographe ou de grammaire (cela relève de la fluidité, pas de la lisibilité).

Annexe D — Étude de cas : un même texte, trois modes de résumé

≥4,5 3,5–4,49 <3,5 Err [F] erreur factuelle Err [L] erreur linguistique

Document source | MLSum · portrait journalistique · janv. 2012 `doc: c6534788d6320e5ffdcb8cad79f2aef7`

[Légende photo] Baltasar Garzon, à la sortie du Tribunal suprême à Madrid, mardi 31 janvier. (Reuters/Susana Vera)

En Espagne, il existe un terme pour désigner, avec une pointe de mépris, les coups d'éclat médiatico-judiciaires : les « garzonadas », du nom du célèbre magistrat Baltasar Garzon. Tour à tour surnommé « Bulldozer », « Juge étoile », ou alors « Super Garzon », le juge fascine autant qu'il agace. Suspendu de ses fonctions depuis mai 2010, il est aujourd'hui, à 56 ans, sous le coup de trois procédures judiciaires pour prévarication (abus de pouvoir) et corruption, et risque jusqu'à vingt ans d'interdiction d'exercer. Accusé d'écoutes illégales dans un procès de corruption impliquant la droite espagnole, il est aussi soupçonné avoir favorisé un banquier qui l'avait généreusement rémunéré pour des conférences données à New York. Mais surtout, on lui reproche d'avoir instruit un procès contre les crimes du franquisme, amnistiés en 1977. Sur ce volet, le magistrat a été entendu par le Tribunal suprême, **mardi 31 janvier**. Cette conjonction d'instructions pourrait faire croire à un acharnement judiciaire, mais elle s'explique aussi par la personnalité flamboyante de ce juge.

Une allure sobre, une avide ambition. Avec ses discrètes lunettes, ses cheveux cendrés plaqués en arrière et son embonpoint, l'homme — qui vit sous forte escorte policière — a pourtant tout l'air d'un homme de droit austère. Celui qui a conduit à l'arrestation de l'ancien dictateur chilien Augusto Pinochet en 1998 — le fait d'armes qui lui a valu une notoriété mondiale — se livre peu et n'accorde que très rarement des interviews. Tout juste a-t-il écrit, à l'âge de 50 ans, un récit introspectif sur sa carrière sous forme de lettre à ses enfants (*Un monde sans peur*, Calmann-Lévy, 2006). Entré en 1988 à l'Audience nationale, la plus haute juridiction espagnole, ce fils de pompiste andalou, qui a poursuivi des études de droit grâce à l'obtention d'une bourse, a rapidement gravi les échelons pour prendre la tête de la chambre d'instruction n° 5. Il n'hésite pas à s'emparer des dossiers les plus sulfureux : du terrorisme basque ou islamiste au trafic de drogue, en passant par des enquêtes de corruption au sein de diverses formations de droite et de gauche.

Militant d'une justice universelle. L'arrestation d'Augusto Pinochet en 1998 a prouvé au monde entier la détermination de cet homme à l'allure discrète et à la voix frêle. Au Chili, en Argentine, il s'adjoint la présence de victimes espagnoles des juntes pour enquêter sur les crimes des dictatures des années 1970–1980. Il vante alors les mérites de l'ingérence au nom de la compétence universelle et contribue à doter l'Espagne en 2005 de la législation la plus avancée en la matière. En l'absence même de victimes espagnoles, l'Espagne ouvre par la suite des enquêtes sur des crimes de « génocide » au Guatemala, puis au Tibet ou encore sur des cas de torture à Guantanamo sous l'administration Bush. En 2009, les compétences universelles de l'Espagne sont toutefois réduites à la suite de démêlés diplomatiques avec la Chine et Israël. Plusieurs fois pressenti pour prendre la tête de la Cour pénale internationale (il y est aujourd'hui conseiller), Baltasar Garzon reste très populaire en Amérique latine, où il a contribué à faire juger les années noires de dictature. En Argentine, le magistrat a reçu des milliers de signatures de soutien ; lundi 30 janvier, un rassemblement en sa faveur s'est tenu à Buenos Aires.

[Légende photo] Lundi 30 janvier, les Mères de la place de Mai ont affiché leur soutien à Buenos Aires au juge espagnol Baltasar Garzon. (AFP/Sergio Goya)

L'impossible instruction des crimes du franquisme. Au sein même de son pays, briser le tabou de la dictature est bien plus compliqué. Répondant aux critiques qui lui reprochaient de ne s'intéresser qu'aux crimes commis outre-Atlantique, Baltasar Garzon décide en 2006 de rompre le « pacte de l'oubli » conclu en 1977 (une amnistie générale avait alors été décrétée en échange de l'acceptation par les caciques du régime de Franco de la transition démocratique). Saisi par des familles de victimes, Garzon requalifie les disparitions de républicains pendant la guerre civile (1936–1939) et sous les premières années de la dictature en crimes contre l'humanité, donc imprescriptibles, et ouvre l'enquête. Il exige l'ouverture de fosses communes et l'identification des corps d'environ 114 000 disparus. Il profite alors d'une volonté politique de briser l'omerta, s'appuyant sur la loi « sur la mémoire historique » votée en 2006 par le gouvernement de José Luis Rodriguez Zapatero. Mais la justice ne suit pas forcément la volonté politique : le parquet a dénoncé une « inquisition » incompatible avec un État de droit. En novembre 2009, le juge se trouve dessaisi du dossier, confié à soixante-deux tribunaux provinciaux, lesquels ensevelissent tous l'affaire. Début 2010, il est accusé d'avoir abusé de ses fonctions et d'avoir ignoré « de façon inexcusable les lois d'amnistie ». Deux associations d'extrême droite portent plainte ; le juge Luciano Varela mène l'enquête. Ironie de l'histoire, l'Argentine, qui a bénéficié dans les années 1990 des compétences universelles espagnoles, a ouvert une enquête fin décembre sur les crimes commis pendant la guerre civile espagnole.

Résumé 1 — Qwen3-8B | LLM 8B · 160 mots `0bea75e8dc8a355f46edbc8e266bc155`

	Fluidité	Lisibilité	Cohérence	Consistance	Pertinence
Score	**5,00**	**4,67**	4,33	4,33	4,33

Baltasar Garzon, un juge espagnol connu pour ses actions médiatiques et judiciaires, est souvent surnommé « garzonadas » en Espagne. Suspendu depuis mai 2010, il fait face à trois procédures pour prévarication et corruption, risquant jusqu'à vingt ans d'interdiction d'exercer. Accusé d'écoutes illégales dans un procès de corruption impliquant la droite espagnole et de favoriser un banquier, il est surtout critiqué pour avoir instruit un procès contre les crimes du franquisme, amnistiés en 1977. Il a été entendu par le Tribunal suprême le 31 janvier 2012 [F]. Son parcours inclut l'arrestation d'Augusto Pinochet en 1998, et il a contribué à la création d'une législation avancée sur la justice universelle en Espagne. En 2006, il a rompu le pacte de l'oubli en ouvrant une enquête sur les crimes du franquisme, malgré des critiques et une démission du dossier en 2009. L'Argentine a récemment ouvert une enquête sur les crimes de la guerre civile espagnole.

> ***Analyse :*** *La source mentionne « mardi 31 janvier » sans année. Le modèle hallucine « 2012 » — erreur factuelle annotée. Par ailleurs, « garzonadas » désigne les* coups d'éclat *du juge, non le juge lui-même : glissement sémantique subtil, non flagué par les annotateurs.*

Résumé 2 — GPT-4o | meilleur LLM · 236 mots `ad55d1409cdc39850985386112f7813a`

	Fluidité	Lisibilité	Cohérence	Consistance	Pertinence
Score	4,33	4,33	4,00	4,33	4,33

Baltasar Garzon, magistrat espagnol renommé, est au cœur de plusieurs affaires judiciaires qui pourraient lui interdire d'exercer pendant vingt ans. À 56 ans, il est accusé de prévarication et de corruption, notamment pour des écoutes illégales dans un procès impliquant la droite espagnole et pour avoir favorisé un banquier en échange de rémunérations pour des conférences à New York. Il est également critiqué pour avoir instruit un procès contre les crimes du franquisme, amnistiés en 1977. Garzon, connu pour son allure austère, a acquis une notoriété mondiale en 1998 avec l'arrestation de l'ancien dictateur chilien Augusto Pinochet. Militant pour une justice universelle, il a mené des enquêtes sur les crimes des dictatures en Amérique latine et contribué à la législation espagnole sur la compétence universelle. En 2009, ces compétences ont été réduites après des tensions diplomatiques. Garzon reste populaire en Amérique latine. En Espagne, il a tenté de briser le « pacte de l'oubli » concernant les crimes du franquisme, mais son enquête a été entravée par la justice. En 2010, il est accusé d'abus de pouvoir et dessaisi du dossier.

> ***Analyse :*** *Résumé le plus complet : les trois fils narratifs (procédures, carrière internationale, affaire du franquisme) sont couverts avec fidélité factuelle. Aucune erreur annotée. Scores homogènes (4,00–4,33) — la densité narrative compense légèrement en fluidité.*

Résumé 3 — Clust | extractif · 185 mots `6b345bd0e068e30124b44db1146f4a6b`

	Fluidité	Lisibilité	Cohérence	Consistance	Pertinence
Score	3,00	3,00	2,33	4,33	1,00

Baltasar Garzon, à la sortie du Tribunal suprême à Madrid, mardi 31 janvier. Il n'hésite pas à s'emparer des dossiers les plus sulfureux : du terrorisme basque ou islamiste au trafic de drogue, en passant par des enquêtes de corruption au sein de diverses formations de droite et de gauche. Militant d'une justice universelle. Au Chili, en Argentine, il s'adjoint la présence de victimes espagnoles des juntes pour enquêter sur les crimes des dictatures des années 1970–1980. [...] Plusieurs fois pressenti pour prendre la tête de la Cour pénale internationale (il y est aujourd'hui conseiller), Baltasar Garzon reste très populaire en Amérique latine. [...] Début 2010, il est accusé d'avoir abusé de ses fonctions et d'avoir ignoré « de façon inexcusable les lois d'amnistie ».

> ***Analyse :*** *Deux artefacts de mise en page extraits comme phrases ordinaires :* légende photo *et* intertitre de section. *Ces éléments — absents de la prose source — brisent la cohérence (2,33) et annulent la pertinence (1,00). Cas typique de l'extraction aveugle à la segmentation non textuelle. La consistance reste haute (4,33) : l'extractif copie le texte source.*

Comparaison inter-modèles

Modèle	Fluidité	Lisibilité	Cohérence	Consistance	Pertinence	Erreurs
Qwen3-8B	**5,00**	**4,67**	4,33	4,33	4,33	fact.
GPT-4o	4,33	4,33	4,00	4,33	4,33	—
Clust	3,00	3,00	2,33	4,33	1,00	ling.

La consistance est la seule dimension robuste à travers les trois types (4,33) : l'extractif copie le texte source, préservant la factualité malgré ses artefacts de structure.

Annexe E — Résultats complets par modèle (scores humains)

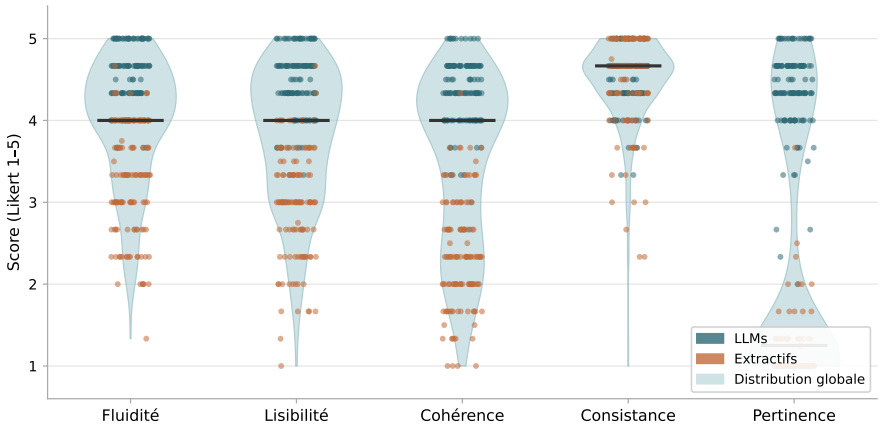
≥4,5 3,5–4,49 <3,5 (colonne *moy* uniquement)

Phase 1 — 15 documents, 120 résumés (calibration, tous modèles)

		Fluidité			Lisibilité			Cohérence			Consistance			Pertinence			Global		
Modèle	*N*	moy	méd	σ	moy	méd	σ	moy	méd	σ	moy	méd	σ	moy	méd	σ	moy	méd	σ
TextRank	12	3,31	3,33	0,63	2,67	2,67	0,47	2,61	2,67	0,57	4,00	4,00	0,38	1,17	1,00	0,25	2,75	2,77	0,31
Clust	12	3,14	3,00	0,67	2,94	3,00	0,56	2,39	2,33	0,43	3,83	4,00	0,48	1,08	1,00	0,14	2,68	2,63	0,37
NER	14	3,31	3,33	0,77	3,07	3,00	0,84	2,81	2,83	0,66	4,19	4,17	0,52	1,21	1,00	0,43	2,92	2,83	0,47
NER+Clust	19	3,18	3,33	0,70	2,93	2,67	0,70	2,53	2,67	0,70	3,93	4,00	0,58	1,14	1,00	0,39	2,74	2,80	0,42
BARThez	4	3,92	4,00	0,36	3,17	2,83	0,69	3,25	3,17	0,49	3,92	3,83	0,60	1,58	1,33	0,68	3,17	3,23	0,36
mBART	4	4,33	4,33	0,53	4,33	4,33	0,47	3,50	3,17	0,96	4,00	3,83	0,41	1,33	1,00	0,58	3,50	3,33	0,55
GPT-4o	12	**4,83**	5,00	0,32	**4,67**	5,00	0,45	**4,36**	4,33	0,25	**4,58**	4,67	0,41	**4,42**	4,67	0,56	**4,57**	4,67	0,30
Qwen3-8B	14	**4,62**	4,67	0,37	4,33	4,33	0,44	4,29	4,33	0,28	4,43	4,67	0,53	4,21	4,33	0,59	4,38	4,40	0,35
Qwen3-32B	14	4,38	4,33	0,35	4,14	4,17	0,39	4,14	4,00	0,43	4,33	4,33	0,52	3,98	4,17	0,61	4,20	4,33	0,33
Mistral-S.	15	**4,71**	5,00	0,40	4,38	4,33	0,38	4,22	4,33	0,23	4,38	4,33	0,54	4,16	4,33	0,38	4,37	4,40	0,21
Ensemble	120	3,93	4,00	0,89	3,64	4,00	0,93	3,40	3,33	0,96	4,19	4,33	0,56	2,56	1,83	1,57	3,54	3,47	0,86

Phase 2 — 90 documents, 367 résumés annotés dont 267 pour les 8 modèles présentés ci-dessous (BARThez et mBART limités aux textes courts)

		Fluidité			Lisibilité			Cohérence			Consistance			Pertinence			Global		
Modèle	*N*	moy	méd	σ	moy	méd	σ	moy	méd	σ	moy	méd	σ	moy	méd	σ	moy	méd	σ
TextRank	26	3,29	3,33	0,79	2,95	3,00	0,75	2,46	2,33	0,74	**4,59**	4,67	0,34	1,12	1,00	0,24	2,88	3,00	1,29
Clust	34	3,32	3,33	0,62	2,98	3,00	0,56	2,22	2,00	0,64	4,48	4,67	0,57	1,26	1,00	0,56	2,85	3,00	1,23
NER	34	3,24	3,33	0,54	2,98	3,00	0,50	2,36	2,33	0,57	4,37	4,67	0,73	1,13	1,00	0,24	2,82	3,00	1,19
NER+Clust	34	3,20	3,33	0,71	3,00	3,00	0,77	2,42	2,33	0,67	**4,54**	4,67	0,50	1,11	1,00	0,24	2,85	3,00	1,27
BARThez	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
mBART	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
GPT-4o	43	**4,57**	4,67	0,38	**4,51**	4,67	0,34	4,37	4,50	0,48	**4,61**	4,67	0,33	**4,48**	4,67	0,49	**4,51**	4,67	0,42
Qwen3-8B	38	4,40	4,33	0,40	4,29	4,33	0,43	4,26	4,33	0,42	4,41	4,33	0,36	4,29	4,33	0,42	4,33	4,33	0,41
Qwen3-32B	27	**4,61**	4,67	0,34	**4,64**	4,67	0,34	**4,38**	4,33	0,48	4,49	4,50	0,35	4,30	4,33	0,65	**4,48**	4,50	0,47
Mistral-S.	31	**4,54**	4,67	0,39	4,48	4,67	0,43	4,35	4,33	0,38	**4,54**	4,67	0,36	4,42	4,50	0,49	4,47	4,50	0,42
Ensemble	267	3,92	4,00	0,83	3,75	4,00	0,92	3,39	3,67	1,14	**4,50**	4,67	0,47	2,84	3,33	1,67	3,68	4,00	1,21



Distribution des scores par dimension (phase 2) — trait = médiane, points = scores moyens.

Annexe F — Erreurs factuelles annotées

81 labels factuels bruts (dont 51 typés) sur 1 072 annotations (267 résumés, phase 2) — *Cst.* : score de consistance moyen attribué *avant* le relevé d'erreurs. Les erreurs génératives sont des hallucinations sémantiques ; les erreurs extractives, des artefacts structurels.

≥4,5 3,5–4,49 <3,5 (colonne *Cst.*)

Modèles génératifs (20 entrées, 6 modèles)		
Modèle	**Cst.**	**Extrait annoté**
GPT-4o	4,00	*«En 1990»* (date)
GPT-4o	**4,67**	*«500 danseurs/j, 70 % d'étrangers»* (valeur)
Qwen3-32B	4,00	*«mère de Jean-Yves»* (relation)
Qwen3-32B	4,00	*«et le frère du narrateur»* (relation)
Mistral-S.	3,33	*«CA Automobile à 37 172 M…»* (valeur)
Mistral-S.	4,33	*«amour»* (lexique)
Mistral-S.	4,33	*«vice-président» / «vice-présidente»* (genre)
Mistral-S.	4,33	*«le 9 mars 2011»* (date)
Qwen3-8B	3,67	*«T. Judt, désormais quadriplégique»* (état)
Qwen3-8B	3,67	*«une dizaine d'années» / «dizaine»* (date)
Qwen3-8B	4,00	*«Le candidat» / «Il»* (coréf.)
Qwen3-8B	4,33	*«31 janvier 2012» / «2012»* (date)
Qwen3-8B	**4,67**	*«d'une de s résultats opérationnels»* (génération)
Modèles seq2seq (BARThez, mBART)		
BARThez	1,00	*«jouer les meilleures équipes du monde»* (prédicat)
BARThez	1,33	*«du vent, de la houle, de l'eau»* (circonstance)
BARThez	3,00	*«"L'appel à la grève"» / «manifestation»* (confusion)
BARThez	3,33	*«Debout Debout»* (prédicat)
BARThez	4,33	*«Fukushima est une ville de la pref. de Fukushima»* (tautologie)
mBART	4,33	*«Elles» / «-ves»* (coréf.)

Modèles extractifs (14 entrées, 4 modèles)		
Modèle	**Cst.**	**Extrait annoté**
NER+Clust	**4,67**	*«1 114» / «1 116»* (n° page)
NER+Clust	**5,00**	*«Le ratio combiné tous les exercices»* (fragment)
NER+Clust	**5,00**	*«Page 2»* (mise en page)
NER+Clust	**5,00**	*«ce couple de viticulteurs» / «saisonniers»* (coréf.)
NER+Clust	**5,00**	*«nous n'arrivons» / «sommes contraints»* (1re pers.)
NER+Clust	**5,00**	*«qu'une personne»* (fragment)
TextRank	**4,67**	*«euh»* (marqueur oral)
TextRank	**5,00**	*«un salarié sur deux, un salarié sur deux»* (répétition)
TextRank	**5,00**	*«à l'État, l'État»* (répétition)
TextRank	**5,00**	*«revenu des Français, revenu du travail»* (répétition)
Clust	3,00	*«l'euro.»* (fragment)
Clust	**4,67**	*«94 %»* (valeur isolée)
NER	2,33	*«l'» / «people»* (tokenisation)
NER	3,67	*«LA BANDE-ANNONCE Film documentaire…»* (méta-texte)