Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Clustering de résumés LLM guidés par l'utilisateur : vers une approche constructiviste et réaliste unifiée

Carl Hatoum[1,2] Catherine Combes[1] Virginie Fresse[1] Christophe Gravier[1]
Mathieu Orzalesi[2]
(1) Laboratoire Hubert Curien, UMR CNRS 5516, Saint-Étienne, France
(2) Segula Technologies, 5 Rue Simone Veil, 69200 Vénissieux, France
carl.hatoum@univ-st-etienne.fr, catherine.combes@univ-st-etienne.fr,
virginie.fresse@univ-st-etienne.fr, christophe.gravier@univ-st-etienne.fr,
mathieu.orzalesi@segula.fr

RÉSUMÉ

Nous introduisons un cadre hybride combinant grands modèles de langage et techniques de regroupement pour extraire, résumer, évaluer et structurer automatiquement les connaissances de larges collections textuelles. Après avoir sélectionné, via une métrique d'entropie sémantique, la stratégie de prompt la plus stable, un LLM génère des résumés modulables qui font l'objet d'une évaluation factuelle assurant leur fiabilité. Ces résumés validés sont ensuite vectorisés, projetés en basse dimension et regroupés en thématiques. Optionnellement, un second LLM affine ensuite leurs libellés pour renforcer l'interprétabilité. Expérimentée sur un corpus majeur d'incidents aériens, cette approche augmente la cohérence et la granularité des clusters thématiques par rapport à une analyse directe des textes, ouvrant de nouvelles perspectives pour la recherche d'information et l'exploration de bases documentaires.

ABSTRACT

User-Guided LLM Summary Clustering : Towards a Unified Constructivist and Realistic Approach

We introduce a hybrid framework combining large language models with clustering techniques to automatically extract, summarize, evaluate, and structure knowledge from large text collections. After selecting the most robust prompting strategy via a semantic entropy metric, a LLM produces modular summaries that undergo factual evaluation to ensure their reliability. These validated summaries are then vectorized, dimensionally reduced, and grouped into thematic clusters. Optionally, a secondary LLM refines their labels to enhance interpretability. Evaluated on a major corpus of aviation incident reports, this approach improves the coherence and granularity of thematic clusters compared to direct text analysis, opening new avenues for information retrieval and exploration of document collections.

MOTS-CLÉS : Résumé constructiviste, Génération de texte contrôlée, Entropie sémantique, Regroupement, Modélisation thématique.

KEYWORDS: Constructivist Summarization, Controlled Text Generation, Semantic Entropy, Clustering, Topic Modeling.

ARTICLE : **Accepté à** CORIA.

1 Introduction

Face à l'explosion du volume de données textuelles, l'extraction et la formalisation des connaissances à partir de grandes collections de documents posent des défis à la fois techniques et méthodologiques. Les approches classiques de modélisation thématique (*topic modeling*), telles que LDA (*Latent Dirichlet Allocation*) (Blei *et al.*, 2003), et leurs évolutions vers les *Neural Topic Models*, comme *BERTopic* (Grootendorst, 2022), permettent de représenter les documents par des distributions sur des thématiques ou via des représentations contextuelles issues de modèles pré-entraînés. Toutefois, ces méthodes présentent des limites importantes : elles peinent à modéliser la sémantique fine des thématiques et à s'aligner sur les préférences des utilisateurs finaux.

Des alternatives fondées sur l'utilisation des grands modèles de langage (LLMs) ont récemment émergé (Pham *et al.*, 2024; Kapoor *et al.*, 2024) visant à produire directement des résumés thématiques ou des représentations synthétiques de documents. Toutefois, ces approches n'intègrent pas explicitement les préférences des utilisateurs et offrent un contrôle limité sur le niveau de granularité ou la structuration des contenus.

Dans ce contexte, nous introduisons la notion de **résumé constructiviste**, que nous définissons comme suit :

Ce résumé, pour un document d, est une sortie textuelle générée par un LLM conditionnée sur un prompt P spécifiant :

- une **préférence utilisateur** (ex. extraire une information, une thématique, etc.),
- une **stratégie de génération** optionnelle (ex. few-shot, chaîne de pensée).

Cette définition permet de produire des résumés adaptés à des besoins spécifiques, en précisant non seulement *quoi* résumer, mais aussi *comment* le faire. Elle repose sur une approche **constructiviste**, où l'information est activement construite en fonction des intentions de l'utilisateur.

Dans un second temps, nous structurons ces résumés à l'aide d'une approche que nous qualifions de **réaliste**, c'est-à-dire fondée sur la structuration thématique intrinsèque des résumés obtenus. Il s'agit ici de faire émerger des regroupements pertinents à partir des données elles-mêmes, sans imposer de catégories *a priori*.

Cette dichotomie fait écho à la distinction constructiviste / réaliste décrite par Hennig (2015), qui souligne que la notion de « vérité » en regroupement dépend tantôt des objectifs et préférences de l'utilisateur (**constructivisme**), tantôt de structures intrinsèques aux données (**réalisme**).

La synergie entre ces deux dimensions — le **constructivisme**, en amont pour la génération de résumés, et le **réalisme**, en aval pour leur organisation thématique — permet d'articuler les niveaux micro (interprétation locale) et macro (structure globale) dans l'analyse de collections textuelles.

La section 2 décrit en détail la méthodologie proposée, la section 3 expose et analyse les résultats obtenus, et la section 4 discute des verrous ainsi que des perspectives de recherche.

2 Méthodologie proposée

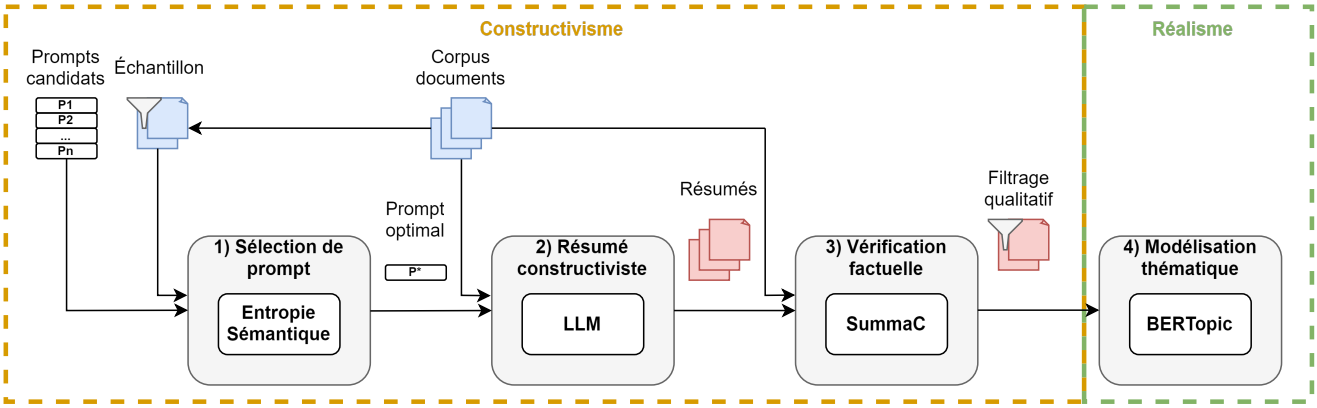


FIGURE 1 – Schéma de la méthodologie proposée

La Figure 1 présente la méthodologie d'extraction et d'organisation de l'information proposée, articulant un volet constructiviste (génération de résumés par LLM) et un volet réaliste (regroupement thématique). Le système sélectionne d'abord un prompt optimal, puis génère des résumés constructivistes, avant de les soumettre à une vérification factuelle. Les résumés validés sont ensuite donnés en entrée pour réaliser une modélisation thématique.

2.1 Étape 1 : Sélection de prompt par entropie sémantique

Un défi central dans l'utilisation des LLMs pour des tâches de génération structurée — comme le résumé orienté par thématique — réside dans leur extrême sensibilité aux instructions reçues. Ce phénomène, connu sous le nom de ***prompt brittleness***, désigne la forte sensibilité des sorties d'un LLM aux variations – même mineures – des prompts fournis.

Cette instabilité complique la conception de prompts fiables, en particulier pour l'extraction d'information à partir de textes complexes. Nous nous intéressons ici à la **variabilité sémantique** des réponses — c'est-à-dire aux différences de contenu informatif, au-delà des variations lexicales ou syntaxiques. L'objectif de cette première étape est d'identifier, parmi un ensemble de prompts candidats, celui induisant la plus faible variabilité sémantique dans les résumés générés. Cette variabilité constitue également un indicateur indirect d'une **incertitude** élevée du LLM quant à la réponse à fournir, révélatrice d'une propension accrue à l'hallucination (Farquhar *et al.*, 2024).

Dans cette optique, nous comparons quatre variantes de prompts notées P_0, P_1, P_2, P_3, correspondant à des paradigmes à complexité progressive d'instruction des LLMs, inspirés des travaux sur le prompting et les chaînes de pensée (Brown *et al.*, 2020; Wei *et al.*, 2022) :

— P_0 **(Zero-shot)** : prompt minimal demandant un résumé fondé sur une préférence utilisateur, sans fournir de contexte ni d'exemples ni d'étapes intermédiaires.
— P_1 **(Few-shot)** : même objectif que P_0, mais enrichi de plusieurs exemples document–résumé pour illustrer l'attendu et guider la génération.
— P_2 **(Chaîne de pensée simple)** : prompt structuré en deux blocs explicites (`<thinking>` suivi de `<answer>`), forçant le modèle à expliciter une réflexion intermédiaire avant de produire le résumé.

— P_3 **(Chaîne de pensée décomposée)** : prompt plus fortement structuré, composé d'étapes atomiques (`<step1>`, `<step2>`, etc.) guidant le modèle dans un raisonnement progressif avant d'aboutir à un résumé final.

Sur un échantillon représentatif du corpus $\mathcal{D}_{\text{sample}} \subset \mathcal{D}$, chaque prompt P_j est évalué en explorant la distribution des sorties du modèle pour chaque document individuellement. Plus précisément, pour chaque document $d \in \mathcal{D}_{\text{sample}}$, on génère un ensemble $S_{d,j} = \{s_{d,j}^{(1)}, \ldots, s_{d,j}^{(n)}\}$ de n résumés indépendants par échantillonnage à partir du LLM, conditionné par P_j. Cet échantillonnage permet de capturer la diversité des interprétations plausibles induites par le prompt sur un même contenu.

Les résumés ainsi générés sont ensuite regroupés selon leur équivalence sémantique via un clustering fondé sur des tests d'implication bidirectionnelle. Chaque groupe sémantiquement cohérent est considéré comme une classe d'interprétation possible du document sous le prompt P_j. L'entropie de Shannon associée à cette partition, notée $H(d; P_j)$, quantifie la dispersion sémantique de ces interprétations. Une faible entropie indique une réponse stable et déterminée du modèle pour le document considéré, tandis qu'une entropie élevée signale une incertitude ou ambiguïté interprétative.

Ce processus est décrit en détail dans l'Algorithme 1, qui formalise l'évaluation de chaque prompt à travers l'entropie moyenne $\bar{H}(P_j)$ sur l'échantillon $\mathcal{D}_{\text{sample}}$, servant de critère pour sélectionner le prompt optimal P^* sur un critère de faible variabilité sémantique des réponses.

Algorithm 1: Sélection de prompt par entropie sémantique

Input: Corpus complet $\mathcal{D} = \{d_1, \ldots, d_N\}$; prompts candidats $\mathcal{P} = \{p_1, \ldots, p_K\}$; taille échantillon M ; nombre de résumés échantillonnés par doc/prompt n ; seuil NLI δ.
Output: Prompt optimal P^*

Étape 1 : Constitution de l'échantillon;
$\mathcal{D}_{\text{sample}} \leftarrow$ échantillon_uniforme$(\mathcal{D}, M)$;

foreach $P_j \in \mathcal{P}$ **do**
  foreach $d \in \mathcal{D}_{\text{sample}}$ **do**
    `// Génération de` n `résumés`
    $S_{d,j} \leftarrow \{s_{d,j}^{(1)}, \ldots, s_{d,j}^{(n)}\}$;
    `// Clustering NLI bidirectionnel`
    Construire graphe NLI sur $S_{d,j}$:
      — pour tout $(s_a, s_b) \in S_{d,j}^2$, calculer NLI$(s_a \to s_b)$ et NLI$(s_b \to s_a)$;
      — lier s_a, s_b si les deux $\geq \delta$;
    Extraire clusters $\mathcal{C}_{d,j} = \{C_{d,j}^{(1)}, \ldots\}$;
    `// Calcul d'entropie`
    $p_\ell \leftarrow |C_{d,j}^{(\ell)}|/n$ pour chaque cluster;
    $H(d; P_j) \leftarrow -\sum_\ell p_\ell \log p_\ell$;
  end
  $\bar{H}(P_j) \leftarrow \frac{1}{M} \sum_{d \in \mathcal{D}_{\text{calib}}} H(d; P_j)$;
end

Étape 3 : Sélection;
$P^* \leftarrow \arg\min_{P_j \in \mathcal{P}} \bar{H}(P_j)$;
return P^*;

2.2 Étape 2 : Résumé constructiviste

Pour chaque document d du corpus, un résumé constructiviste est généré à l'aide du prompt optimal P^* sélectionné à l'étape 1. Cette génération prend en compte une préférence utilisateur et une stratégie de génération optionnelle, explicitées dans le prompt, ainsi que les hyperparamètres du LLM (température, etc.).

2.3 Étape 3 : Vérification factuelle

La génération de résumés par des modèles de langage de grande taille (LLMs) est sujette aux hallucinations, c'est-à-dire l'introduction d'éléments absents ou infondés par rapport au document source. Cette dérive compromet la fidélité factuelle des résumés, un critère particulièrement critique dans les domaines sensibles. Dès lors, évaluer la cohérence entre le résumé généré et le document source devient un enjeu central. Même si l'entropie sémantique associée à un prompt donné est faible – indiquant une stabilité des réponses –, cela n'exclut pas complètement la présence d'hallucinations : il reste indispensable de vérifier la cohérence résumé / document source sur le plan factuel.

Plusieurs approches non supervisées ont été proposées pour estimer automatiquement, sans référence, la fidélité factuelle des résumés. Parmi elles, QuestEval (Scialom *et al.*, 2021) repose sur la génération automatique de questions à partir du résumé et du document, suivie d'une évaluation croisée des réponses. D'autres méthodes s'appuient sur des modèles NLI (Laban *et al.*, 2022; Steen *et al.*, 2023), qui évaluent les relations d'implication entre les segments du résumé et du document source.

Nous reprenons ici la méthode SummaC *zero-shot* proposée dans (Laban *et al.*, 2022), dont la sortie est un score compris entre 0 et 1 (une valeur proche de 1 témoigne d'une forte cohérence factuelle entre le résumé et le document source). La procédure se décompose en cinq grandes étapes :

1. On segmente le document source d en I phrases $\{x_1, \ldots, x_I\}$ et le résumé s en J phrases $\{y_1, \ldots, y_J\}$.
2. Pour chaque paire (x_i, y_j), on calcule le score d'implication unidirectionnel :

$$E_{i,j} \;=\; \mathrm{NLI}(x_i \to y_j)$$

   Ce qui définit la matrice d'implication

$$E \;=\; [E_{i,j}]_{1\le i\le I,\; 1\le j\le J}.$$

3. Pour chaque phrase y_j du résumé, on extrait le score local maximal :

$$m_j \;=\; \max_{1\le i\le I} E_{i,j}$$

4. On agrège ces scores pour obtenir le score global de cohérence :

$$\mathrm{SummaC}(d, s) \;=\; \frac{1}{J}\sum_{j=1}^{J} m_j,$$

2.4 Étape 4 : Regroupement hiérarchique et extraction de thématiques

Les résumés validés sont encodés en vecteurs de plongement puis soumis à une réduction de dimension par UMAP (McInnes *et al.*, 2020), puis à un regroupement hiérarchique via HDBSCAN (McInnes *et al.*, 2017). Pour chaque cluster ainsi formé, nous extrayons les mots-clés les plus représentatifs au moyen d'un TF-IDF par classe dans le cadre de BERTopic (Grootendorst, 2022), obtenant une structure hiérarchique de thématiques.

2.5 Étape 5 (optionnelle) : Raffinement des thématiques par LLM

Afin de renforcer l'interprétabilité et la pertinence de la structure thématique obtenue, nous exploitons un LLM de grande capacité (par exemple GPT-4 (Achiam *et al.*, 2023)) qui reçoit directement la structure thématique en entrée afin de contextualiser les mots-clés et expliciter les concepts sous-jacents, et de proposer des libellés synthétiques et cohérents pour chaque thématique.

3 Résultats Préliminaires

Dans cette étude préliminaire, nous exploitons un jeu de données relatif aux incidents aériens issu de la base de la *National Transportation Safety Board (NTSB)* [1] . Les quatre stratégies de génération de résumés, décrites dans la section 2 et adaptées à notre contexte applicatif, sont présentées en Annexe A.

Pour chacune de ces variantes, nous générons des résumés avec *Mistral 7B v0.3* (Jiang *et al.*, 2023). Les mesures d'entropie sémantique et le filtrage de cohérence résumé / document source sont réalisés avec le modèle *deberta-large-mnli* (He *et al.*, 2021). Nos choix de modèles sont guidés par des critères d'accessibilité, en privilégiant ceux pouvant fonctionner sur du matériel local.

Après génération des résumés et leur filtrage, nous les vectorisons dans un espace de plongement, puis réalisons un regroupement hiérarchique, suivi d'une extraction de thématiques et leur raffinement, comme décrit dans les étapes 3, 4, et 5 de la section 2.

3.1 Sélection de prompt par entropie sémantique

Le Tableau 1 présente l'estimation des scores d'entropie sémantique obtenus pour chacune des quatre stratégies de génération de résumé. Les résultats révèlent une diminution nette de l'entropie entre le *zero-shot* et les autres configurations, la *chaîne de pensée décomposée* atteignant le meilleur score, traduisant une convergence vers une sémantique unique et une plus grande stabilité des résumés.

La mesure d'entropie sémantique confirme deux mécanismes déjà mis en lumière dans la littérature : (i) l'ajout d'exemples explicites (*Few-shot*) aligne le modèle sur une préférence utilisateur et contraint l'espace de réponse (Brown *et al.*, 2020), tandis que (ii) la structuration du raisonnement par des *chaînes de pensées* canalise l'espace de réponse, même sans exemples, en forçant le passage par des étapes intermédiaires cohérentes (Wei *et al.*, 2022).

1. https://www.ntsb.gov/safety/data/Pages/Data_Stats.aspx

Stratégie de génération	Entropie sémantique ↓
Zero-shot	1.277
Few-shot	0.518
Chaîne de pensée simple	0.485
Chaîne de pensée décomposée	**0.067**

TABLE 1 – Estimation des entropies sémantiques obtenues par méthode de résumé (plus bas est meilleur)

Cependant, la réduction de l'entropie sémantique ne garantit pas nécessairement une amélioration de la qualité des résumés. Par exemple, le *Few-shot* peut conduire à une reproduction mécanique des schémas de réponse présentés en exemple. Une stabilité apparente peut aussi résulter d'une simple mémorisation des données d'entraînement. Ce type de comportement souligne l'importance d'enrichir l'évaluation par des mesures de cohérence des résumés.

3.2 Évaluation de la cohérence résumé / document source

Nous expérimentons deux méthodes : QuestEval (Scialom *et al.*, 2021) et SummaC (Laban *et al.*, 2022) . Nous constatons que SummaC est plus robuste et pertinent dans notre contexte, et décidons de le retenir pour la suite. Sur l'ensemble du corpus, nous analysons les scores SummaC obtenus en confrontant les résumés générés à leurs documents sources. La Figure 2 illustre la distribution empirique de ces scores selon les différentes stratégies de génération.

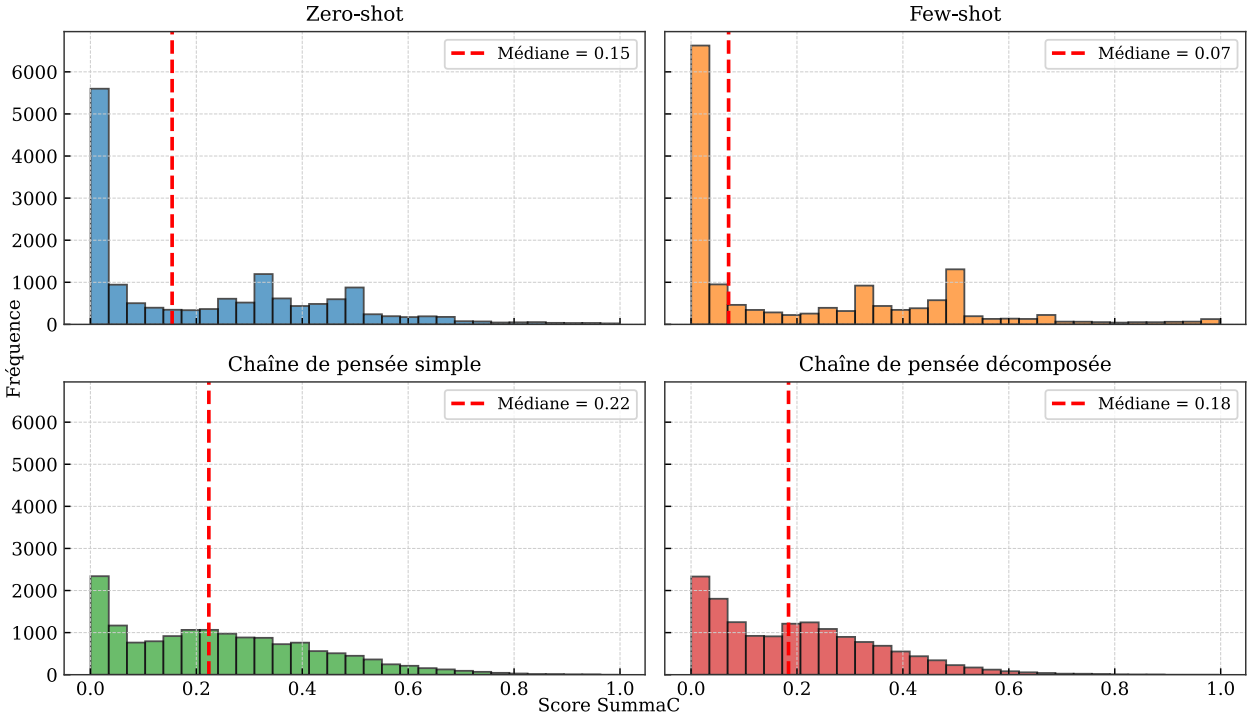


FIGURE 2 – Distribution empirique des scores SummaC par méthode

— ***Few-shot*** : médiane = 0.07, forte concentration autour de zéro, faible fidélité.
— ***Zero-shot*** : médiane = 0.15, amélioration légère, scores toujours bas.
— ***Chaîne de pensée décomposée*** : médiane = 0.18, amélioration modérée.
— ***Chaîne de pensée simple*** : médiane = 0.22, distribution plus étalée, offrant les meilleurs scores.

Pour illustrer ce que reflète un score SummaC, nous réalisons en analyse qualitative des résumés, et identifions quatre tendances majeures, chacune donnée en exemple dans le Tableau 2 :

Document source	**Résumé**	**Score**
« [..] The reason for the loss of engine power on both engines could not be determined. »	« The document suggests that the incident may have been related to a loss of engine power on both engines, which was preceded by the reduction of engine power under rich mixture conditions on two previous occasions. »	0.000
« [..] It is likely that the fuel system was not properly flushed after the restoration process to remove any possible foreign object/liquid contamination. »	« The root cause of the aviation incident was a perforated carburetor float due to corrosion, likely caused by improper flushing of the carburetor after restoration, leading to engine failure during takeoff. »	0.001
« [..] he lost control of the helicopter. [. . .] the wind was from 360° at 19 knots, gusting to 26 knots. »	« The root cause of the aviation incident was the pilot's loss of control during downwind flight in high wind conditions (19 knots, gusting to 26 knots) while conducting rice field pollination operations. »	0.154
« [..] The pilot reported that the accident was due to him not removing the bolt in the yoke. »	« The root cause of the aviation incident was the pilot's failure to remove a bolt he had placed in the yoke for a gust lock during takeoff. »	0.991

TABLE 2 – Exemples d'extraits de résumés et du document source, ainsi que le score SummaC associé

— **Hallucination factuelle** (score très faible) : divergence majeure, apparition d'un fait inexistant.
— **Spéculation / hypothèses** (score faible) : hypothèses non vérifiées, baisse du score malgré pertinence exploratoire.
— **Inférence plausible** (score intermédiaire) : reformulation cohérente sans mention explicite.
— **Extraction de l'information** (score élevé) : reformulation fidèle des faits explicites.

Ainsi, un score SummaC plus élevé indique un meilleur ancrage factuel du résumé dans le document source, sans toutefois distinguer la nature exacte des écarts ni leur pertinence potentielle dans un contexte constructiviste. Nous postulons que les scores bas obtenus sont liés à la nature constructiviste des résumés – notamment du fait du conditionnement du résumé par les préférences utilisateur – et qu'ils doivent être interprétés avec précaution et complétés par d'autres critères pour une évaluation plus complète.

3.3 Comparaison des regroupements

L'évaluation des modèles thématiques s'appuie essentiellement sur des métriques automatiques de cohérence. Cependant, plusieurs travaux (Hoyle *et al.*, 2021; Doogan & Buntine, 2021) soulignent leurs limites, notamment un décalage fréquent avec l'appréciation des utilisateurs finaux.

Pour mieux cerner l'impact de la génération de résumés constructivistes, nous analysons un tableau croisé comparant deux regroupements : celui issu des documents bruts et celui des résumés, focalisés ici sur les causes et produits selon la méthode *chaîne de pensée décomposée*. Cette comparaison

quantifie la correspondance entre les clusters, c'est-à-dire dans quelle mesure les regroupements des documents bruts et des résumés se recoupent ou divergent.

TABLE 3 – Clusters de documents bruts présentant un fort accord avec ceux des résumés

Clusters documents bruts	Pourcentage d'accord (%)	Clusters résumés
0	100.00 %	0
2	95.16 %	2
4	86.84 %	5
8	90.32 %	11
12	96.21 %	14
19	92.31 %	15

Le Tableau 3 montre, pour chaque cluster issu des documents bruts, le pourcentage d'accord avec le clustering des résumés. Six clusters (0, 2, 4, 8, 12 et 19) affichent une correspondance supérieure à 85 %, témoignant d'une forte stabilité thématique : les documents bruts traitent majoritairement de la préférence utilisateur (la cause de l'incident). Le cluster 0 atteint même 100 % d'accord, ce qui suggère que certains documents étaient déjà alignés *a priori* sur cette préférence avant résumé.

Le Tableau 4 détaille la réaffectation des clusters des documents bruts vers ceux des résumés. Les taux d'accord plus faibles (15,79 %–56,34 %) révèlent une redistribution marquée dans les résumés constructivistes. Les clusters 3, 9 et 17 présentent les taux d'accord les plus bas (15,79 %–36,84 %), indiquant que leur sémantique a été recatégorisée selon des dimensions thématiques plus fines dans les résumés.

Ces observations mettent en évidence deux régimes complémentaires :

— **Clusters stables** (taux d'accord élevés) : la cause d'incident demeure sémantiquement dominante dans le document brut.

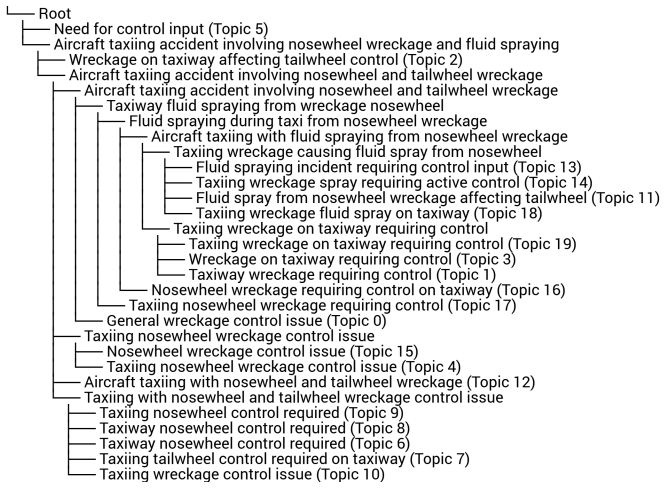
— **Clusters redistribués** (taux d'accord faibles) : le résumé obtenu est réaffecté selon des dimensions thématiques plus fines.

TABLE 4 – Clusters de documents bruts présentant une répartition différente dans le clustering des résumés

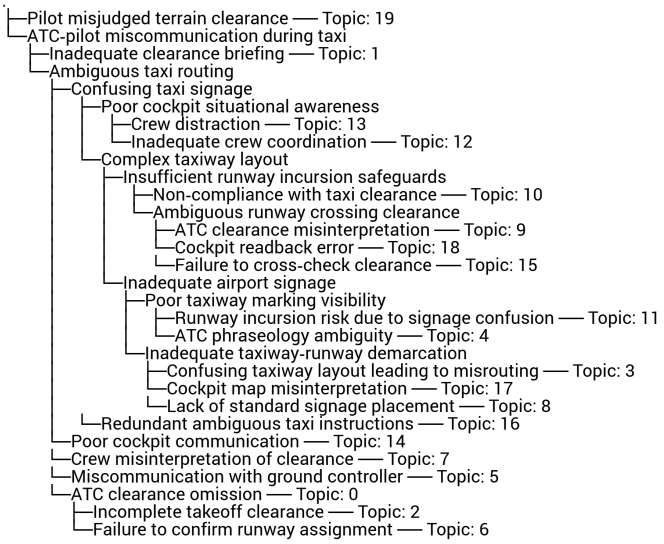
Clusters documents bruts	Pourcentage d'accord (%)	Clusters résumés
5	42.86 %	3
6	41.76 %	
1	56.34 %	4
7	31.69 %	
3	36.84 %	10
9	28.95 %	
17	15.79 %	

Pour vérifier cette hypothèse, nous complétons l'étude par une analyse qualitative des thématiques obtenues, visant à identifier les mots-clés expliquant les redistributions et à vérifier si les nouveaux clusters reflètent réellement la préférence utilisateur conditionnant les résumés.

3.4 Comparaison qualitative des structures thématiques obtenues



(a) Structure thématique des documents bruts.



(b) Structure thématique des résumés orientés causes.

FIGURE 3 – Extraits des structures hiérarchiques obtenues après raffinement des thématiques : (a) document brut et (b) résumés orientés causes par chaîne de pensée décomposée.

Les deux arbres mettent en lumière deux types de regroupements distincts : l'arbre 3a structure les événements autour d'une description narrative des événements et phénomènes, tandis que l'arbre 3b segmente la collection selon les différentes catégories d'incidents (humains, environnementaux, techniques, etc.). Ainsi, l'orientation de chaque arbre se distingue clairement : narratif pour le premier, causale pour le second.

Cette comparaison révèle que, dans le regroupement issu des documents bruts, les thématiques les plus dominantes des documents orientent leur regroupement, et que d'autres peuvent être « noyées ». À l'inverse, l'approche de résumés constructivistes offre une granularité optimale pour isoler et extraire précisément les thématiques en fonction des préférences des utilisateurs.

4 Conclusion et perspectives

Cette étude préliminaire présente une méthode hybride alliant l'adaptabilité des LLM à la robustesse du regroupement hiérarchique pour extraire et organiser efficacement les connaissances de vastes collections, comme les rapports d'incidents en aviation. Les résumés constructivistes fournissent une interprétation contextualisée selon les besoins des utilisateurs, tandis que le regroupement et l'extraction de thématiques structurent les données de façon cohérente et hiérarchique. Ensemble, ils enrichissent les thématiques et clarifient les représentations, bien au-delà d'une simple analyse des documents bruts.

Cependant, plusieurs défis subsistent quant à l'utilisation de la méthodologie proposée, ouvrant la voie à de nouvelles perspectives de recherche :

Évaluation des résumés constructivistes Le contrôle de la factualité demeure un enjeu central, notamment la gestion des **hallucinations**, qui peut compromettre la confiance dans les contenus générés. Le recours à des approches de raisonnement avec LLM, telles que les chaînes de pensée (Wei *et al.*, 2022), constitue une piste prometteuse pour améliorer la transparence et la fiabilité des inférences réalisées sur le document. Nous avons évalué la cohérence résumé / document source, mais ses limites soulignent la nécessité d'un **cadre multicritère** plus large pour évaluer la qualité des résumés générés, articulé autour de trois axes :

- **Ancrage au document source** : via des métriques de résumés agnostiques.
- **Fidélité à la préférence utilisateur** : mesure de l'adéquation aux éléments demandés.
- **Plausibilité des inférences** : évaluation des raisonnements, hypothèses, et spéculations sur les faits du document source.

Ce cadre évaluatif vise à distinguer les interprétations plausibles, acceptables dans une logique constructiviste, des hallucinations, et à équilibrer **ancrage documentaire** et **souplesse interprétative**.

Évaluation de la structure hiérarchique thématique La structuration thématique requiert le développement de métriques adaptées, capables de mesurer la pertinence des relations de subsomption, de la granularité des regroupements, ainsi que de quantifier la cohérence et la diversité des nœuds à la fois sur le plan hiérarchique et sémantique. Il est également envisageable d'évaluer cette structure thématique en mobilisant un LLM comme évaluateur (*LLM-as-a-Judge*) (Zheng *et al.*, 2023), pour analyser la cohérence globale et la qualité des liens hiérarchiques obtenus.

Validation extrinsèque et usages applicatifs Une perspective importante consiste à valider le cadre proposé dans des contextes applicatifs concrets, tels que la recherche d'information, l'aide à la décision, ou les systèmes de recommandation. La structuration obtenue pourrait y jouer un rôle en facilitant la navigation dans des corpus complexes.

Les travaux futurs envisagés s'orienteraient vers une formalisation plus rigoureuse des résumés constructivistes, ainsi que le développement de protocoles d'évaluation capables de rendre compte des spécificités propres à ce type de résumés. L'articulation entre préférences utilisateur, inférences et ancrage documentaire constituerait un axe de recherche prometteur.

Références

ACHIAM J., ADLER S., AGARWAL S., AHMAD L., AKKAYA I., ALEMAN F. L., ALMEIDA D., ALTENSCHMIDT J., ALTMAN S., ANADKAT S. *et al.* (2023). Gpt-4 technical report. *arXiv preprint arXiv :2303.08774*. DOI : 10.48550/arXiv.2303.08774.

BLEI D. M., NG A. Y. & JORDAN M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**(null), 993–1022.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901. DOI : 10.48550/arXiv.2005.14165.

DOOGAN C. & BUNTINE W. (2021). Topic model or topic twaddle ? re-evaluating semantic interpretability measures. 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. 3824–3848, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.300.

FARQUHAR S., KOSSEN J., KUHN L. & GAL Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, **630**(8017), 625–630. DOI : 10.1038/s41586-024-07421-0.

GROOTENDORST M. (2022). Bertopic : Neural topic modeling with a class-based tf-idf procedure. DOI : 10.48550/arXiv.2203.05794.

HE P., LIU X., GAO J. & CHEN W. (2021). Deberta : Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

HENNIG C. (2015). What are the true clusters ? *Pattern Recognition Letters*, **64**, 53–62. Philosophical Aspects of Pattern Recognition, DOI : https ://doi.org/10.1016/j.patrec.2015.04.009.

HOYLE A., GOEL P., PESKOV D., HIAN-CHEONG A., BOYD-GRABER J. & RESNIK P. (2021). Is automated topic model evaluation broken ? the incoherence of coherence. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA : Curran Associates Inc.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., DE LAS CASAS D., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L., LAVAUD L. R., LACHAUX M.-A., STOCK P., SCAO T. L., LAVRIL T., WANG T., LACROIX T. & SAYED W. E. (2023). Mistral 7b. DOI : 10.48550/arXiv.2310.06825.

KAPOOR S., GIL A., BHADURI S., MITTAL A. & MULKAR R. (2024). Qualitative insights tool (qualit) : Llm enhanced topic modeling. DOI : 10.48550/arXiv.2409.15626.

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. DOI : 10.1162/tacl_a_00453.

MCINNES L., HEALY J. & ASTELS S. (2017). hdbscan : Hierarchical density based clustering. *The Journal of Open Source Software*, **2**. DOI : 10.21105/joss.00205.

MCINNES L., HEALY J. & MELVILLE J. (2020). Umap : Uniform manifold approximation and projection for dimension reduction. DOI : 10.21105/joss.00861.

PHAM C. M., HOYLE A., SUN S., RESNIK P. & IYYER M. (2024). TopicGPT : A prompt-based topic modeling framework. In K. DUH, H. GOMEZ & S. BETHARD, Éds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 2956–2984, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.naacl-long.164.

SCIALOM T., DRAY P.-A., LAMPRIER S., PIWOWARSKI B., STAIANO J., WANG A. & GALLINARI P. (2021). QuestEval : Summarization asks for fact-based evaluation. In M.-F. MOENS, X. HUANG, L. SPECIA & S. W.-T. YIH, Éds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 6594–6604, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.529.

STEEN J., OPITZ J., FRANK A. & MARKERT K. (2023). With a little push, NLI models can robustly and efficiently predict faithfulness. In A. ROGERS, J. BOYD-GRABER & N. OKAZAKI, Éds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 914–924, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-short.79.

WEI J., WANG X., SCHUURMANS D., BOSMA M., XIA F., CHI E., LE Q. V., ZHOU D. *et al.* (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, **35**, 24824–24837. DOI : 10.48550/arXiv.2201.11903.

ZHENG L., CHIANG W.-L., SHENG Y., ZHUANG S., WU Z., ZHUANG Y., LIN Z., LI Z., LI D., XING E. P., ZHANG H., GONZALEZ J. E. & STOICA I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA : Curran Associates Inc. DOI : 10.48550/arXiv.2306.05685.

A Prompts de génération de résumés

A.1 Zero-shot

```
Analyze the provided document to identify and summarize
the root cause(s) or contributing factors of an aviation
incident. Return the result as a single paragraph directly
stating the cause(s).

Document: Document
```

A.2 Few-shot

```
Analyze the provided document to identify and summarize
the root cause(s) or contributing factors of an aviation
incident. Return the result as a single paragraph directly
stating the cause(s).

Example:      Document: "According to the certified flight
instructor (CFI), they had departed Tamiami-Executive
Airport (KTMB) at 0800 en route to the training area. ...
When the throttle in the cockpit was moved from idle to
full power, the carburetor throttle arm could be held in
idle with little force with one finger."   Output: "A loss
of engine power due to inadequate maintenance inspection
```

resulting in a worn throttle housing going undetected and failing."

Document: Document

A.3 Chaîne de pensée simple

Analyze the provided document to identify and summarize the root cause(s) or contributing factors of an aviation incident. Return the result using the following exact format:

1. Thinking process enclosed within <thinking> and </thinking> tags. For example: <thinking>Let's think step by step... </thinking>

2. Final answer enclosed within <answer> and </answer> tags.

Do not generate multiple reasoning or answer sections. Only one of each.

Document: Document

A.4 Chaîne de pensée décomposée

Analyze the provided document to identify and summarize the root cause(s) or contributing factors of an aviation incident. Return the result using the following exact format:

<step1> Extract Key Entities and Events: Identify the main actors, objects, and significant events from the narrative. </step1>

<step2> Sequence and Link Events: Arrange the events in order and determine the cause-and-effect relationships between them. </step2>

<thinking>Let's think step by step by breaking down the narrative into its core components and connecting the events logically to trace the causal chain.</thinking>

<answer>Summarize the Causal Chain: Provide a concise summary that explains how each event led to the next, forming a clear chain of reasoning that identifies the root cause(s) or contributing factors.</answer>

Document: Document