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

De la génération libre à la génération contrôlée : étude de la stabilité des résumés juridiques

Marah Baccari
LS2N, Nantes, France
marah.baccari@etu.univ-nantes.fr

RÉSUMÉ

Les documents juridiques sont souvent longs, complexes et rédigés dans un langage spécialisé. Leur synthèse exige non seulement une réduction du contenu, mais aussi une forte fiabilité factuelle et structurelle. Dans cet article, nous étudions des méthodologies contrôlées pour la synthèse structurée de longues décisions de la Cour suprême des États-Unis. Nous analysons des mécanismes visant à améliorer la robustesse des grands modèles de langage (LLMs). Nous évaluons une synthèse hiérarchique via des prompts, des stratégies de génération à enrichissement contextuel, ainsi qu'une génération agentique sous contraintes structurelles. Nous introduisons notamment une approche de raffinement intra-document et une approche fondée sur Pydantic permettant d'imposer explicitement une structure juridique prédéfinie lors de la génération. Les résultats montrent que les contraintes structurelles et la validation itérative améliorent l'alignement factuel et la conformité formelle.

ABSTRACT

From Free Generation to Controlled Generation : A Study of Stability in Legal Summarization

Legal documents are often long, complex, and written in specialized language. Their summarization requires not only content reduction but also strong factual and structural reliability. In this paper, we investigate controlled methodologies for structured summarization of long U.S. Supreme Court decisions. We study mechanisms that improve the robustness of large language models (LLMs). We evaluate hierarchical prompt-based summarization, retrieval-augmented generation strategies, and agent-based generation under structural constraints. In particular, we introduce an intra-document refinement approach and a Pydantic-based framework that explicitly enforces a predefined legal structure during generation. Experimental results show that structural constraints and iterative validation improve factual alignment, coherence, and format compliance.

MOTS-CLÉS : Résumé automatique, Synthèse juridique, Génération à enrichissement contextuel, Génération contrôlée, IA agentique.

KEYWORDS: Automatic summarization, Legal summarization, Retrieval-augmented generation, Controlled generation, Agent-based AI.

1 Introduction

L'augmentation du nombre de litiges et d'affaires pénales a considérablement accru la charge de travail des tribunaux et des institutions juridiques à travers le monde. Par conséquent, les professionnels du droit doivent analyser de grands volumes de documents juridiques longs et complexes, souvent sous de fortes contraintes de temps. Les décisions de justice peuvent s'étendre sur plusieurs dizaines de

Actes des 19e Rencontres Jeunes Chercheurs en RI (RJCRI) et 28ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), pages 117–133.

pages et contiennent un langage juridique dense, rendant la lecture et la synthèse manuelles longues et exigeantes.

Dans ce contexte, la synthèse automatique de texte a suscité un intérêt croissant comme moyen d'aider les professionnels du droit en automatisant une tâche routinière mais nécessitant un effort important d'analyse et de traitement de l'information. Les premières approches de résumé automatique reposaient principalement sur des méthodes extractives, telles que TextRank et LexRank (Mihalcea & Tarau, 2004; Erkan & Radev, 2004), qui sélectionnent les phrases les plus représentatives d'un document à partir de critères statistiques ou graphes. Bien que robustes, ces méthodes restent limitées car elles ne permettent ni de reformuler l'information ni d'imposer une organisation structurée adaptée aux exigences juridiques.

Les approches abstractives basées sur des modèles neuronaux ont ensuite permis de générer des résumés plus cohérents et synthétiques (Naveed *et al.*, 2024), mais elles demeurent sensibles aux erreurs factuelles et offrent un contrôle limité sur la structure de sortie. Plus récemment, les grands modèles de langage (LLMs) ont démontré des capacités avancées de compréhension et de génération de texte, ouvrant de nouvelles perspectives pour le résumé juridique automatisé.

Toutefois, la synthèse juridique automatisée reste particulièrement difficile en raison de la complexité du raisonnement juridique, de la nécessité de préserver la précision factuelle et du respect de structures rédactionnelles strictes propres aux décisions judiciaires. Les LLMs présentent ainsi plusieurs limites en contexte juridique : ils peuvent générer des *hallucinations structurelles* (Huang *et al.*, 2025), manquer de contrôle formel sur la sortie et produire des résumés dont la fiabilité et la conformité aux normes juridiques sont incertaines (Shukla *et al.*, 2022).

Pour résoudre ces défis, nous proposons de formaliser la tâche comme une **synthèse juridique contrôlée**, où l'objectif est de produire des résumés fidèles au texte source tout en respectant un schéma juridique strict. En introduisant des contraintes structurelles explicites dans le processus de génération, cette formalisation permet de guider le modèle et de renforcer la cohérence argumentative des résumés produits.

Plus précisément, la tâche consiste à générer, à partir d'une *entrée longue* comprenant les différentes sections d'un dossier judiciaire, une *sortie structurée* organisée selon un schéma juridique prédéfini : *faits de l'affaire*, *question juridique*, et *conclusion de la Cour*. Cette structuration s'inspire des analyses rhétoriques des décisions judiciaires, qui identifient généralement des sections majeures telles que *Facts*, *Issues*, *Arguments* et *Decision* (Bonnard *et al.*, 2025). Dans ce travail, nous nous concentrons sur un sous-ensemble de ces composantes afin de produire des synthèses juridiques concises et comparables. L'évaluation se base à la fois sur la précision factuelle et la conformité structurelle, offrant un cadre scientifique pour comparer des méthodes extractives, des approches génératives hiérarchiques, des stratégies de *génération à enrichissement contextuel* (RAG), et des modèles agentiques avec contraintes de schéma.

Cet article vise donc à analyser comment différents mécanismes de contrôle peuvent améliorer la stabilité, la fiabilité et la cohérence des résumés juridiques, en identifiant les forces et les limites de chaque approche dans le contexte de décisions longues de la Cour suprême des États-Unis.

2 État de l'art

La synthèse automatique de documents juridiques a fait l'objet de nombreux travaux, qui peuvent être regroupés selon l'évolution des paradigmes de résumé automatique.

Les premières approches reposent sur la synthèse extractive, consistant à sélectionner les phrases les plus importantes du document source afin de former un résumé. Des méthodes fondées sur des graphes telles que TextRank et LexRank (Mihalcea & Tarau, 2004; Erkan & Radev, 2004) évaluent l'importance des phrases à l'aide de mécanismes de centralité inspirés de PageRank (Wachsmuth *et al.*, 2017), tandis que des modèles plus récents exploitent des représentations contextuelles issues de modèles de langage, comme BERTSum (Shukla *et al.*, 2022). Ces méthodes présentent l'avantage de préserver fidèlement le contenu original, un critère particulièrement important dans le domaine juridique. Toutefois, elles produisent souvent des résumés fragmentés et ne garantissent pas une organisation logique conforme aux structures juridiques attendues.

Afin d'améliorer la cohérence et la fluidité des résumés, les approches abstractives ont introduit des modèles neuronaux capables de générer de nouvelles formulations. Des architectures telles que BART (Lewis *et al.*, 2019), Pegasus (Zhang *et al.*, 2020a) et plus récemment les grands modèles de langage (LLMs) (Naveed *et al.*, 2024) permettent de produire des résumés plus synthétiques et mieux structurés. Cependant, ces modèles sont sujets aux hallucinations et aux incohérences factuelles, ce qui limite leur adoption dans des contextes juridiques sensibles (Ji *et al.*, 2023).

Pour pallier ces limitations, plusieurs travaux ont exploré des approches hybrides combinant extraction et génération, où un modèle extractif sélectionne d'abord les passages pertinents avant leur reformulation par un modèle génératif (Habu *et al.*, 2023). D'autres recherches s'appuient sur l'apprentissage par renforcement afin d'optimiser la qualité des résumés à l'aide de fonctions de récompense spécifiques (Verma *et al.*, 2025). Le *fine-tuning* de modèles pré-entraînés a également montré des améliorations notables, notamment en raison du caractère spécialisé du langage juridique (Mullick *et al.*, 2024).

Un défi majeur demeure toutefois le traitement des documents juridiques longs, souvent incompatibles avec les limites de contexte des modèles neuronaux. Des architectures adaptées aux longs contextes, telles que Longformer-Encoder-Decoder (LED) (Beltagy *et al.*, 2020), ainsi que des stratégies de découpage de documents (*chunking*), ont été proposées pour mieux gérer ces entrées volumineuses (Liu *et al.*, 2023).

Malgré ces avancées, les approches existantes ne garantissent pas explicitement le contrôle structurel des résumés générés. Les modèles extractifs manquent de flexibilité, tandis que les modèles génératifs, bien que plus expressifs, produisent des sorties dont la structure et la fiabilité restent difficiles à maîtriser. Cette limitation motive l'exploration de mécanismes de contrôle explicites visant à améliorer la stabilité et la conformité des résumés juridiques générés.

3 Cadre expérimental et préparation des données

3.1 Cadre méthodologique général

À partir de l'état de l'art, nous adoptons une démarche expérimentale progressive visant à analyser l'impact de différents mécanismes de contrôle sur le résumé de décisions judiciaires longues, comme

illustré par la Figure 1. Nous comparons plusieurs stratégies de complexité croissante : une synthèse extractive, une synthèse hiérarchique basée sur des prompts, des approches de génération à enrichissement contextuel (RAG), un mécanisme de raffinement intra-document, puis une génération agentique intégrant des contraintes structurelles explicites.

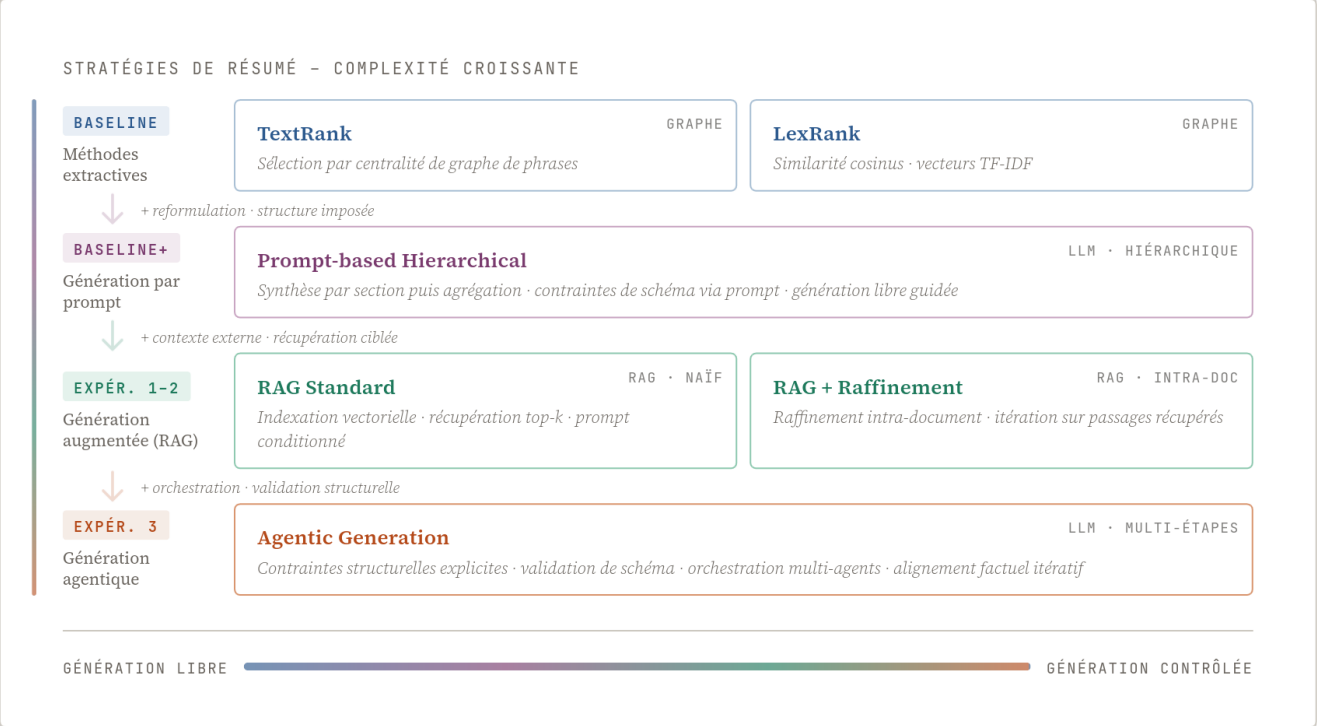


FIGURE 1 – Vue d'ensemble de la démarche expérimentale progressive.

Cette progression permet d'évaluer dans quelle mesure l'ajout de contexte et de validations structurelles successives améliore l'alignement factuel et la cohérence formelle des résumés produits.
Il est important de préciser que les différentes approches évaluées ne correspondent pas toutes à des expériences indépendantes. Certaines méthodes réutilisent les sorties intermédiaires des étapes précédentes dans une logique progressive de raffinement. En particulier, l'approche RAG 2 utilise le résumé généré par la méthode basée sur les prompts comme requête de récupération afin d'affiner le résumé initial à l'aide de segments pertinents issus du document source. De même, l'approche fondée sur PydanticAI reprend le pipeline hiérarchique de résumé par segments avant d'appliquer des contraintes structurelles explicites lors de la génération finale. Les approches extractives (TextRank, LexRank) restent en revanche entièrement indépendantes des pipelines génératifs.

3.2 Jeu de données

Le résumé de décisions judiciaires longues nécessite un corpus de documents intégraux, structurés et annotés selon des critères juridiques explicites, idéalement accompagnés de résumés rédigés par des experts servant de référence. C'est pourquoi nous utilisons le jeu de données issu des travaux de (Srun *et al.*, 2024; Zakaria, 2025), collecté à partir du site officiel de la Cour suprême des États-Unis (SCOTUS). Il contient 3468 affaires juridiques jugées entre 1791 et 2024, et ces travaux détaillent la construction du corpus ainsi que les critères de sélection des affaires.

Chaque affaire comprend des métadonnées de base telles que l'*année de la décision*, le *numéro de rôle* (docket number) représentant la référence unique de l'affaire, le *nom de l'affaire*, ainsi que des

liens vers les documents originaux. Le contenu principal de chaque affaire est divisé en trois parties.

La première partie contient les opinions rédigées par la Cour. Celles-ci sont disponibles à la fois en format HTML brut et dans une version structurée et analysée. Les données structurées incluent l'*opinion majoritaire*, une opinion *per curiam*, les opinions *concordantes* et *dissidentes*, ainsi qu'un *syllabus* qui fournit un bref aperçu de l'affaire.

La deuxième partie se compose des transcriptions, incluant les *plaidoiries orales* (oral arguments), les *annonces* de l'opinion de la Cour et, dans de rares cas, les transcriptions des *opinions dissidentes*.

La troisième partie contient des résumés de référence (vérité terrain) pour l'évaluation. Ces résumés ont été rédigés par des experts juridiques et reflètent une interprétation faisant autorité de chaque affaire. Chaque résumé de référence est structuré en trois sections : les *faits de l'affaire*, la *question juridique* et la *conclusion finale* de la Cour.

Une propriété importante de ce jeu de données est la longueur des documents juridiques. La plupart des affaires contiennent entre 10 000 et 30 000 tokens, tandis qu'un nombre plus restreint dépasse 40 000 tokens.

3.3 Analyse des données

Avant d'appliquer les modèles de résumé, nous avons mené une analyse exploratoire du jeu de données afin de mieux comprendre sa structure, son contenu et ses éventuels problèmes. Les données originales étaient fournies au format JSON et ont été converties en une représentation tabulaire (CSV) en extrayant les champs pertinents dans des colonnes dédiées. Cette transformation a facilité l'inspection, le nettoyage et l'analyse des champs textuels et des métadonnées.

Une première analyse exploratoire a permis d'examiner la taille et la structure du jeu de données, d'identifier les valeurs manquantes, de détecter les doublons et de repérer d'éventuelles entrées inhabituelles ou incohérentes.

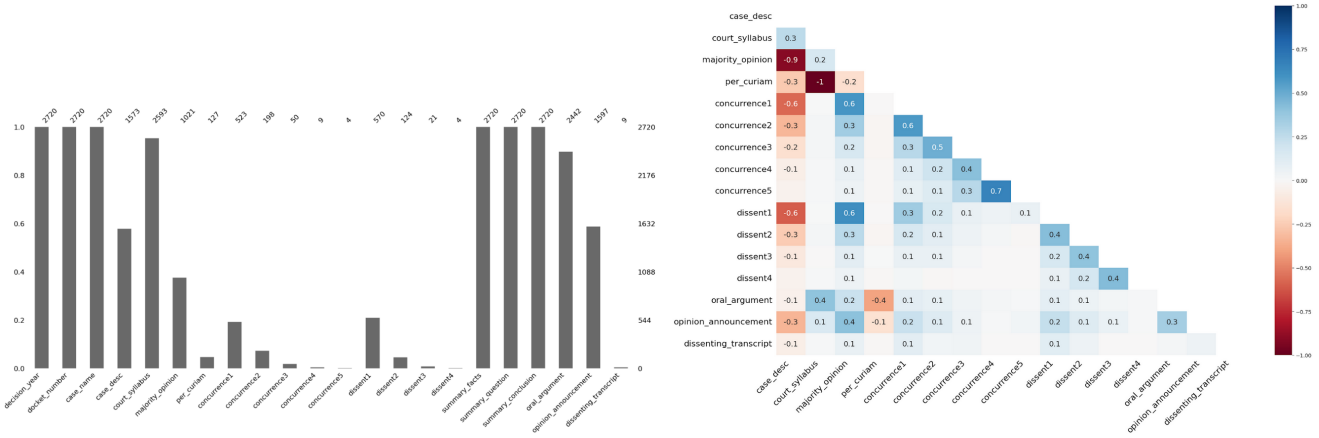
Le prétraitement textuel s'est concentré sur le nettoyage des artefacts introduits lors du web scraping. Cela comprenait la suppression des balises HTML, des retours à la ligne, des tabulations et d'autres annotations bruitées, suivie d'une normalisation des espaces. Ces étapes étaient nécessaires pour garantir la cohérence entre les documents et éviter l'introduction de biais lors du traitement.

L'analyse des doublons a révélé que certaines affaires apparaissaient plusieurs fois sous des titres identiques, certaines étant présentes jusqu'à quatre fois. Après vérification, les entrées redondantes ont été supprimées afin d'éviter une surreprésentation d'une même affaire dans les expériences.

Nous avons également réalisé des analyses exploratoires visuelles afin de mieux caractériser le jeu de données. La Figure 2 présente la distribution des valeurs manquantes par colonne ainsi que la matrice de corrélation entre les différentes sections textuelles. Ces visualisations permettent d'identifier les éventuelles similarités entre les colonnes et de vérifier la complétude des données.
Nous avons par ailleurs identifié un sous-ensemble d'affaires consolidées ou liées, traitées conjointement par la Cour. Bien que ces relations n'aient pas été explicitement modélisées dans nos expériences, elles mettent en évidence la complexité structurelle du jeu de données.

3.4 Construction des entrées et des cibles

Pour la tâche de résumé, nous avons défini explicitement le texte d'entrée et le résumé cible utilisés pour l'entraînement et l'évaluation.



(a) Distribution des valeurs manquantes

(b) Matrice de corrélation entre colonnes

FIGURE 2 – Analyses exploratoires du jeu de données.

Le texte d'entrée est construit en sélectionnant et en concaténant uniquement les champs contenant un raisonnement juridique significatif et un contexte factuel pertinent : *case name* et *description*, *majority opinion*, *concurrences*, *dissents* et *syllabus*. Nous avons exclu les *oral argument transcripts*, jugés plus bruités et moins structurés.

Toutes les sections sélectionnées sont fusionnées en une seule entrée textuelle représentant l'ensemble du dossier écrit de l'affaire. Cette entrée unifiée est ensuite traitée à l'aide de la stratégie de segmentation (chunking) décrite ultérieurement.

La sortie cible correspond à un résumé de référence obtenu par la combinaison des trois champs suivants : *faits de l'affaire*, *question juridique*, *conclusion finale* de la Cour. La colonne contenant le résumé de référence sert de sortie attendue pour les modèles de résumé.

3.5 Stratégie d'échantillonnage

En raison de la grande taille et de la forte variabilité des documents juridiques du jeu de données, exécuter les expériences sur l'ensemble du corpus s'est avéré coûteux en ressources computationnelles. Afin de garantir une évaluation équilibrée et représentative tout en maintenant des expériences réalisables, nous avons adopté une stratégie d'échantillonnage stratifié fondée sur deux critères complémentaires : la couverture temporelle et la longueur des documents.

Le critère temporel divise les affaires en trois périodes : avant 2003, de 2003 à 2010, et après 2010. Le second critère regroupe les documents selon leur longueur en trois catégories : courts (moins de 20 000 tokens), moyens (entre 20 000 et 40 000 tokens) et longs (plus de 40 000 tokens).

En combinant ces deux axes, chaque affaire juridique est assignée à une strate unique, garantissant une diversité à la fois en termes de périodes historiques et de tailles de documents. À partir de ce jeu de données stratifié, plusieurs sous-ensembles ont été créés afin de mener des expériences à différentes échelles. En particulier, un sous-ensemble de 100 affaires a été utilisé pour l'évaluation principale. Les distributions détaillées des strates et la composition des sous-ensembles sont présentées en Annexe A.

3.6 Stratégie de segmentation

Les documents juridiques du jeu de données sont souvent trop longs pour être traités directement par la plupart des modèles de résumé, rendant la segmentation nécessaire afin de respecter les contraintes d'entrée des modèles. Les expériences initiales ont montré qu'une segmentation naïve de taille fixe introduit souvent des coupures abruptes dans le texte, tandis qu'une segmentation purement sémantique génère un nombre excessif de segments, augmentant le coût computationnel et pouvant parfois réduire la cohérence globale.

Pour pallier ces problèmes, nous adoptons une stratégie de segmentation adaptative avec chevauchement de phrases. Cette approche ajuste dynamiquement la taille des segments en fonction de la longueur du document tout en préservant la continuité contextuelle grâce à un chevauchement contrôlé. Les documents courts (<300 tokens) ne sont pas segmentés. Pour les documents plus longs, la taille des segments est ajustée en fonction de la longueur totale : 1200 tokens avec un chevauchement d'une phrase par défaut, 1500 tokens (chevauchement de 2 phrases) pour les documents entre 10 000 et 30 000 tokens, 2200 tokens (2 phrases) entre 30 000 et 60 000 tokens, et 3000 tokens avec un chevauchement de 3 phrases au-delà de 60 000 tokens. Le chevauchement inter-segments vise à préserver la continuité contextuelle. Cette stratégie permet d'équilibrer préservation du contexte et efficacité computationnelle et s'est révélée robuste sur des documents de longueurs variées. Les détails de l'implémentation sont fournis en Annexe C.

4 Configuration expérimentale et méthodes

4.1 Sélection des modèles

Le choix du modèle de résumé a été guidé par plusieurs critères pratiques et scientifiques : efficacité computationnelle avec des ressources GPU limitées, bonnes performances sur les tâches de résumé en anglais, longueur de contexte suffisante pour traiter de longs documents juridiques, et utilisation de modèles open-source matures afin d'assurer la reproductibilité.

Sur la base de ces critères, plusieurs familles de LLMs open-source ont été considérées, notamment Mistral (Jiang *et al.*, 2023), LLaMA (Touvron *et al.*, 2023) et Qwen (Yang *et al.*, 2025), avec une préférence pour des modèles de taille réduite adaptés à des environnements aux ressources limitées. Des travaux antérieurs (Aly *et al.*, 2025) montrent que Mistral-7B-Instruct atteint des scores ROUGE (Lin, 2004) comparables à des modèles beaucoup plus volumineux tels que LLaMA-2-70B-Chat, tout en offrant des performances BERTScore (Zhang *et al.*, 2020b) plus équilibrées, ce qui indique une meilleure couverture de l'information.

À partir de ces résultats et de nos expériences préliminaires, Mistral-7B-Instruct v0.3 a été sélectionné comme modèle de référence pour toutes les expériences. Des détails supplémentaires sur les variantes de modèles testées et les comparaisons préliminaires sont fournis en Annexe B.

4.2 Métriques d'évaluation

Nous utilisons des métriques de recouvrement lexical et de similarité sémantique en rapportant les scores F1.

ROUGE et BLEU (Lin, 2004; Papineni *et al.*, 2002) sont utilisés pour mesurer la similarité au niveau superficiel entre les résumés générés et les résumés de référence. Nous rapportons ROUGE-1, qui capture le recouvrement des unigrammes et reflète la couverture basique du contenu, ainsi que ROUGE-L, qui mesure la plus longue sous-séquence commune et fournit une indication de similarité

structurelle. BLEU est inclus comme métrique complémentaire, se concentrant sur la précision des n-grammes et offrant une évaluation plus stricte de la correspondance lexicale, bien qu'elle soit moins tolérante au paraphrasage.

Pour aller au-delà de la simple correspondance de mots, nous utilisons BERTScore (Zhang *et al.*, 2020b), qui calcule la similarité sur la base des embeddings contextuels et est plus robuste au paraphrasage et aux variations lexicales. Cela est particulièrement pertinent pour les textes juridiques, où les mêmes concepts peuvent être exprimés avec des formulations différentes. Cependant, BERTScore est moins intuitif à interpréter que ROUGE ou BLEU, car il reflète la similarité au niveau des embeddings plutôt qu'une correspondance lexicale ou factuelle explicite.

4.3 Modèles de référence (Baseline)

Afin d'établir des points de comparaison, nous considérons deux types de baselines :

Approches extractives : TextRank et LexRank sélectionnent des phrases via des graphes de similarité et un calcul de centralité de type PageRank. TextRank repose sur le recouvrement lexical, tandis que LexRank utilise la similarité cosinus TF–IDF. Ces méthodes évitent les hallucinations mais ne permettent pas d'imposer une structure juridique explicite.

Baseline générative par prompting : Nous adoptons une pipeline hiérarchique en deux étapes adaptée aux documents longs : résumé par segments, puis fusion finale. Plusieurs stratégies (zero-shot, one-shot, few-shot, chain-of-thought) ont été évaluées, le chain-of-thought produisant les résultats les plus cohérents. Toutefois, cette approche ne garantit pas strictement le respect d'un schéma juridique prédéfini.

Les détails d'implémentation sont fournis en Annexe D.

4.4 Résumé via RAG

Nous explorons la génération à enrichissement contextuel (RAG) afin d'enrichir le contexte et réduire les hallucinations. L'utilisation du RAG repose sur l'hypothèse que des segments juridiquement similaires peuvent fournir un contexte complémentaire utile pour améliorer la précision terminologique et réduire certaines ambiguïtés présentes dans les résumés générés.

Deux variantes sont évaluées. RAG 1 récupère des segments sémantiquement similaires provenant d'autres affaires juridiques pour enrichir la génération. RAG 2 utilise le résumé initial (généré par prompt) comme requête et récupère des segments issus du même cas afin d'affiner et corriger la sortie.

La différence principale réside dans la portée de la récupération : inter-cas pour RAG 1, intra-cas pour RAG 2. Empiriquement, RAG 2 offre une meilleure précision juridique et limite l'introduction d'informations non pertinentes. Les détails techniques sont fournis en Annexe E.
Dans les deux approches RAG, la récupération vectorielle repose sur une recherche des $k = 5$ segments les plus similaires à l'aide de FAISS et des embeddings bge-m3. Cette valeur a été retenue empiriquement afin d'équilibrer richesse contextuelle et limitation du bruit introduit par des segments moins pertinents.

4.5 Génération structurée sous contraintes

Contrairement à RAG 1, l'approche RAG 2 repose sur une hypothèse de raffinement intra-document : un résumé initial peut servir de représentation condensée de l'affaire afin de guider la récupération des segments les plus pertinents du document source. Cette stratégie vise à de garantir le respect strict

de la structure juridique (*Faits*, *Question*, *Conclusion*), nous adoptons une approche de génération structurée fondée sur des contraintes de schéma. Cette stratégie répond à une limitation observée dans les approches génératives, notamment RAG : malgré des instructions explicites dans les prompts, les résumés produits ne respectent pas toujours la structure attendue.

Contrairement au prompting standard, cette méthode impose un schéma de sortie formel définissant explicitement les champs attendus. La réponse générée est automatiquement validée par rapport à ce schéma, et la génération est relancée en cas de non-conformité. Les contraintes portent ainsi sur la structure des sorties plutôt que sur le vocabulaire généré. Dans nos expériences, ce mécanisme est implémenté à l'aide de *PydanticAI*. Les détails d'intégration sont fournis en Annexe F.

5 Résultats et Discussion

Toutes les méthodes ont été évaluées sur un sous-ensemble de 100 affaires juridiques, de manière stratifiée en combinant la couverture temporelle et la longueur des documents (voir 3.5 et Annexe A). La taille limitée de ce sous-ensemble a été choisie afin de garantir des expérimentations diversifiées tout en respectant les contraintes computationnelles, chaque affaire nécessitant un temps d'exécution conséquent pour les modèles génératifs et les environnements comme Colab imposant un plafond de durée pour éviter la déconnexion automatique. Ces méthodes ont été évaluées à l'aide des métriques automatiques décrites à la Section 4.2, qui capturent des aspects complémentaires de la qualité des résumés, incluant le recouvrement lexical (BLEU, ROUGE-1, ROUGE-L) et la similarité sémantique (BERTScore). L'évaluation compare les approches extractives, génératives, à enrichissement contextuel et agentiques en termes de recouvrement lexical et de similarité sémantique. La Table 1 rapporte les scores F1 moyens pour chaque méthode.

Method	BLEU	R-1	R-L	BERTScore
TextRank	0.0505	0.3809	0.1779	0.8192
LexRank	0.0558	0.3725	0.1839	0.8276
Prompt-based	0.0422	0.3454	0.2013	0.8523
RAG 1	0.0418	0.2869	0.1353	0.8175
RAG 2	0.0391	0.3383	0.2004	0.8524
PydanticAI	**0.0958**	**0.4375**	**0.2340**	**0.8581**

TABLE 1 – Comparaison des méthodes d'après les métriques BLEU, ROUGE-1, ROUGE-L et BERTScore (scores F1 moyens calculés sur 100 affaires juridiques)

Les résultats mettent en évidence une relation claire entre le niveau de contrôle imposé au processus de génération et la stabilité des résumés produits. Les méthodes extractives, représentées par TextRank et LexRank, obtiennent des scores relativement élevés en ROUGE-1 et en BERTScore, indiquant un alignement lexical et sémantique raisonnable avec les résumés de référence. Cette performance s'explique par leur capacité à sélectionner directement des segments du texte source, garantissant ainsi une forte fidélité textuelle. Toutefois, leur nature purement extractive limite leur capacité de synthèse : elles ne permettent pas de contrôler l'organisation du contenu et échouent systématiquement à produire la structure juridique essentielle de type Faits–Question–Conclusion.

Les approches génératives basées uniquement sur des prompts introduisent un premier niveau de structuration grâce aux instructions textuelles fournies au modèle. Elles produisent des résumés globalement plus cohérents et présentent une amélioration de la similarité sémantique, comme en témoignent les scores BERTScore plus élevés. Néanmoins, les scores BLEU et ROUGE-1 restent modérés en raison du paraphrasage et des reformulations générées par le modèle. L'analyse qualitative révèle également des divergences factuelles et une variabilité structurelle importantes, illustrant les

limites d'une génération non contrainte dans un domaine exigeant une forte rigueur formelle.

Parmi les méthodes de génération à enrichissement contextuel, RAG 2 surpasse RAG 1, confirmant que l'affinement d'un résumé initial à l'aide du contexte source récupéré constitue une stratégie plus efficace que la génération directe à partir de la fusion des résumés issus des différents chunks avec du contexte externe. La première approche RAG souffre de la longueur excessive des entrées et de l'introduction de documents externes pouvant perturber la cohérence interne du raisonnement généré. La seconde approche améliore la cohérence globale et l'alignement sémantique en utilisant un résumé intermédiaire comme requête de récupération. Toutefois, malgré ces améliorations, les méthodes RAG présentent encore des limites en termes d'alignement lexical et de stabilité structurelle, leurs performances restant inférieures aux baselines extractives sur certaines métriques de recouvrement.

L'approche agentique basée sur PydanticAI obtient les meilleures performances globales sur l'ensemble des métriques évaluées. En appliquant un schéma de sortie strict et en validant automatiquement les réponses générées, le système impose des contraintes explicites au modèle, permettant de réduire les hallucinations structurelles et d'assurer une meilleure cohérence des résumés. Le mécanisme de validation et de régénération garantit que la sortie finale respecte la structure attendue tout en maintenant une forte similarité sémantique avec les références humaines. Cette amélioration constante à travers toutes les métriques souligne l'importance de mécanismes de contrôle explicites lors de l'application des LLMs à des tâches spécialisées telles que la synthèse juridique

Les résultats indiquent que la qualité des résumés juridiques dépend moins des capacités génératives seules que du degré de contrainte structurelle appliqué. Les méthodes extractives assurent la fidélité lexicale sans abstraction, le prompting apporte de la flexibilité mais une instabilité structurelle, et les approches RAG renforcent l'ancrage informationnel sans résoudre entièrement l'organisation du contenu. L'approche agentique combine génération et contrôle, montrant que l'imposition explicite de contraintes structurelles est clé pour des systèmes de résumé juridique fiables.

6 Limitations

Malgré des résultats encourageants, plusieurs limites doivent être soulignées. Premièrement, bien que les mécanismes de contrôle structurel améliorent la conformité formelle des résumés, ils ne garantissent pas une compréhension juridique profonde. Un modèle peut respecter un schéma prédéfini tout en produisant des simplifications excessives ou des interprétations incomplètes du raisonnement judiciaire. Deuxièmement, les approches fondées sur le RAG et sur l'IA agentique restent dépendantes de la qualité des représentations intermédiaires et du découpage en segments. Le *chunking*, bien qu'utile pour gérer la longueur du contexte, peut fragmenter l'information et affecter la cohérence globale du résumé. Troisièmement, l'évaluation a été conduite sur 100 affaires seulement, ce qui limite la portée des conclusions et ne couvre pas l'ensemble de la diversité des décisions de la Cour suprême. Enfin, les métriques automatiques présentent des limites connues : BLEU et ROUGE mesurent principalement le recouvrement lexical, tandis que BERTScore évalue la similarité sémantique sans capturer explicitement la qualité du raisonnement juridique ni la stabilité structurelle du résumé.

7 Conclusion et Future Directions

Dans cet article, nous avons comparé plusieurs approches de synthèse de longues décisions juridiques de la Cour suprême des États-Unis, incluant des méthodes extractives, des approches génératives basées sur des prompts, des architectures RAG et une génération structurée fondée sur des agents. Les résultats montrent que l'augmentation des capacités génératives seules ne garantit ni la stabilité ni la

conformité structurelle des résumés. Les méthodes extractives assurent un bon recouvrement lexical, tandis que les approches génératives améliorent parfois la similarité sémantique, mais restent sensibles aux incohérences et aux erreurs factuelles. L'intégration d'un schéma de sortie strict via PydanticAI conduit à des gains systématiques sur toutes les métriques et renforce la stabilité structurelle. Cela suggère que le contrôle explicite des sorties est plus déterminant que la seule sophistication du modèle pour traiter des documents juridiques longs et complexes.

Plusieurs perspectives se dégagent de ce travail, le développement de métriques adaptées à la synthèse structurée, l'évaluation sur des ensembles de données plus larges et l'intégration systématique de la génération structurée dans différentes pipelines pourraient améliorer la fiabilité et l'interprétabilité des systèmes de résumé juridique automatisé. Par ailleurs, des protocoles d'évaluation fondés sur des approches de type *LLM-as-a-Judge* pourraient être explorés afin d'évaluer plus explicitement des dimensions difficilement capturées par les métriques automatiques classiques, notamment le respect de la structure juridique, la cohérence argumentative et la qualité rédactionnelle des résumés générés.

Références

ALY W. M., SOLIMAN T. H. A. & ABDELAZIZ A. M. (2025). An Evaluation of Large Language Models on Text Summarization Tasks Using Prompt Engineering Techniques. arXiv :2507.05123 [cs], DOI : 10.48550/arXiv.2507.05123.

BELTAGY I., PETERS M. E. & COHAN A. (2020). Longformer : The Long-Document Transformer. arXiv :2004.05150 [cs], DOI : 10.48550/arXiv.2004.05150.

BONNARD W., LAVISSIÈRE M. C., BELFATHI A., HERNANDEZ N., JACQUIN C. & MONCEAUX-CACHARD L. (2025). "Steps" towards a corpus of SCOTUS opinions annotated using a Swalesian approach. *Ibérica*, (50), 45–80. DOI : 10.17398/2340-2784.50.45.

ERKAN G. & RADEV D. R. (2004). LexRank : Graph-based Lexical Centrality as Salience in Text Summarization. *Journal of Artificial Intelligence Research*, **22**, 457–479. arXiv :1109.2128 [cs], DOI : 10.1613/jair.1523.

HABU R., RATNAPARKHI R., ASKHEDKAR A. & KULKARNI S. (2023). A Hybrid Extractive-Abstractive Framework with Pre & Post-Processing Techniques To Enhance Text Summarization. In *2023 13th International Conference on Advanced Computer Information Technologies (ACIT)*, p. 529–533. ISSN : 2770-5226, DOI : 10.1109/ACIT58437.2023.10275584.

HUANG L., YU W., MA W., ZHONG W., FENG Z., WANG H., CHEN Q., PENG W., FENG X., QIN B. & LIU T. (2025). A Survey on Hallucination in Large Language Models : Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.*, **43**(2), 42 :1–42 :55. DOI : 10.1145/3703155.

JI Z., LEE N., FRIESKE R., YU T., SU D., XU Y., ISHII E., BANG Y., CHEN D., DAI W., CHAN H. S., MADOTTO A. & FUNG P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, **55**(12), 1–38. arXiv :2202.03629 [cs], DOI : 10.1145/3571730.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., 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. arXiv :2310.06825 [cs], DOI : 10.48550/arXiv.2310.06825.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2019). BART : Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv :1910.13461 [cs], DOI : 10.48550/arXiv.1910.13461.

LIN C.-Y. (2004). ROUGE : A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, p. 74–81, Barcelona, Spain : Association for Computational Linguistics.

LIU N. F., LIN K., HEWITT J., PARANJAPE A., BEVILACQUA M., PETRONI F. & LIANG P. (2023). Lost in the Middle : How Language Models Use Long Contexts. arXiv :2307.03172 [cs], DOI : 10.48550/arXiv.2307.03172.

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, Barcelona, Spain : Association for Computational Linguistics.

MULLICK A., BOSE S., SAHA R., BHOWMICK A. K., VEMPATY A., GOYAL P., GANGULY N., DEY P. & KOKKU R. (2024). Leveraging the Power of LLMs : A Fine-Tuning Approach for High-Quality Aspect-Based Summarization.

NAVEED H., KHAN A. U., QIU S., SAQIB M., ANWAR S., USMAN M., AKHTAR N., BARNES N. & MIAN A. (2024). A Comprehensive Overview of Large Language Models. arXiv :2307.06435 [cs], DOI : 10.48550/arXiv.2307.06435.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a Method for Automatic Evaluation of Machine Translation. In P. ISABELLE, E. CHARNIAK & D. LIN, Éds., *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

SHUKLA A., BHATTACHARYA P., PODDAR S., MUKHERJEE R., GHOSH K., GOYAL P. & GHOSH S. (2022). Legal Case Document Summarization : Extractive and Abstractive Methods and their Evaluation. In Y. HE, H. JI, S. LI, Y. LIU & C.-H. CHANG, Éds., *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 1048–1064, Online only : Association for Computational Linguistics. DOI : 10.18653/v1/2022.aacl-main.77.

SRUN N., HERNANDEZ N. & JACQUIN C. (2024). *Automated Generation of Argumentative Roles from Legal Opinions Using Pertinent Information Recognition*. Rapport interne, LS2N.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F., RODRIGUEZ A., JOULIN A., GRAVE E. & LAMPLE G. (2023). LLaMA : Open and Efficient Foundation Language Models. arXiv :2302.13971 [cs], DOI : 10.48550/arXiv.2302.13971.

VERMA J. K., VERMA R., VERMA B., ARORA D. & SINGH P. (2025). Context-Aware Legal Summarization using Reinforcement Learning. In *2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)*, p. 1010–1016. DOI : 10.1109/CICTN64563.2025.10932588.

WACHSMUTH H., STEIN B. & AJJOUR Y. (2017). "PageRank" for argument relevance. In M. LAPATA, P. BLUNSOM & A. KOLLER, Éds., *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, p. 1117–1127, Valencia, Spain.

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. arXiv :2505.09388 [cs], DOI : 10.48550/arXiv.2505.09388.

ZAKARIA B. (2025). *Generating rhetorically coherent summaries of long legal documents*. Thèse de doctorat, Faculty of Sciences Semlalia Marrakech.

ZHANG J., ZHAO Y., SALEH M. & LIU P. J. (2020a). Pegasus : pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20 : JMLR.org.

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

A Détails sur l'échantillonnage du jeu de données

En combinant la couverture temporelle et la longueur des documents, nous avons défini un ensemble de strates, chacune correspondant à une combinaison spécifique de période temporelle et de longueur de document. Chaque affaire a été assignée à une unique strate, enregistrée dans une colonne dédiée. Le Tableau 2 présente la distribution des affaires selon les différentes strates. Il montre que la majorité des décisions appartiennent aux catégories de longueur courte et moyenne, tandis que les documents très longs restent relativement rares.

Strate	**Nombre**
before_2003_short	786
before_2003_medium	693
after_2010_medium	340
after_2010_short	252
2003_2010_short	189
2003_2010_medium	172
before_2003_long	94
after_2010_long	53
2003_2010_long	14

TABLE 2 – Distribution des affaires selon les strates.

B Expériences supplémentaires de sélection du modèle

Conformément aux critères de sélection décrits dans l'article principal, plusieurs versions du modèle Mistral-7B ont été évaluées lors d'expérimentations préliminaires. Celles-ci incluaient le modèle de base Mistral-7B ainsi que plusieurs variantes adaptées aux instructions : Mistral-7B-Instruct v0.1, v0.2 et v0.3.

Au cours de ces évaluations préliminaires, les versions adaptées aux instructions ont systématiquement surpassé le modèle de base en termes de cohérence et de pertinence des résumés. En particulier, Mistral-7B-Instruct v0.2 et v0.3 ont produit des résumés de meilleure qualité, la version v0.3 fournissant les résultats les plus stables et les plus cohérents. Ces observations ont motivé la sélection de Mistral-7B-Instruct v0.3 comme modèle final utilisé dans l'évaluation expérimentale.

C Détails de l'implémentation du découpage adaptatif

La stratégie de découpage adaptatif fonctionne au niveau des phrases en utilisant SpaCy pour la segmentation. Les phrases sont ajoutées séquentiellement à un segment (*chunk*) jusqu'à ce qu'un seuil de taille prédéfini soit atteint. Afin de prendre en compte la variabilité de la longueur des documents, la taille des segments est ajustée dynamiquement en fonction du nombre total de mots de chaque document, garantissant la compatibilité avec la fenêtre de contexte effective du modèle de résumé.
Plus précisément, pour les documents contenant entre 10 000 et 30 000 mots, la taille des segments est limitée à 1 500 mots, avec un chevauchement de deux phrases entre segments consécutifs. Pour les documents compris entre 30 000 et 60 000 mots, la taille des segments est portée à 2 200 mots, tout en conservant le même chevauchement de phrases. Pour les documents très longs dépassant 60 000 mots, la taille des segments est augmentée à 3 000 mots, sans accroître le chevauchement, afin de limiter le coût computationnel.
Lorsqu'un segment atteint sa taille maximale, il est enregistré et le segment suivant est initialisé en conservant les phrases de chevauchement situées à la fin du segment précédent. Ce mécanisme permet de préserver le contexte local et de réduire la perte d'information aux frontières des segments. Le processus se poursuit jusqu'au traitement de l'ensemble des phrases, le contenu restant formant le dernier segment.
Dans l'ensemble, cette conception adaptative permet au processus de découpage de rester flexible pour des documents juridiques courts, moyens et longs, tout en équilibrant cohérence contextuelle, robustesse et efficacité computationnelle.

D Détails du résumé basé sur des prompts

D.1 Conception des prompts

Au niveau des segments (*chunks*), le prompt demande au modèle d'agir comme un assistant de synthèse juridique et de produire un résumé concis et factuel d'un segment unique. Les instructions insistent sur la précision, l'identification des parties impliquées, les faits essentiels et l'issue de l'affaire. Une attention particulière est portée à la préservation exacte de la terminologie juridique, des noms d'affaires et des formulations formelles, tout en décourageant la paraphrase ou l'introduction de nouveau vocabulaire.
Lors de l'étape finale de synthèse, un second prompt est appliqué à l'ensemble des résumés intermédiaires issus des segments. Ce prompt impose une structure fixe composée de trois sections : *Facts of the Case*, *Question* et *Conclusion*, et rappelle l'exigence de fidélité au langage juridique original.
Le prompt final utilisé pour la génération des résumés globaux dans la baseline basée sur les prompts est présenté ci-dessous :

```
You are a legal summarization expert.

Format your response EXACTLY as:

1. Facts of the Case: [Summary of what happened]
2. Question: [The legal question being decided]
3. Conclusion: [The court's decision]

IMPORTANT:
- Be accurate and specific
- Include case name, parties, and key facts
```

```
- Clearly state the legal question
- Provide the final ruling/decision
- PRESERVE the exact terminology, case names, and legal terms used in
- Do NOT introduce new vocabulary or paraphrase legal terms
```

D.2 Stratégies de prompting

Plusieurs stratégies de prompting ont été évaluées, notamment le zero-shot, le one-shot, le few-shot ainsi que le chain-of-thought. Le prompting de type chain-of-thought, sans exemples explicites, a donné les meilleurs résultats, car il encourage un raisonnement plus structuré tout en évitant un surapprentissage lié à des démonstrations spécifiques.

D.3 Modèle et ajustement des paramètres

Les paramètres de génération ont été ajustés de manière systématique, notamment la taille du faisceau (*beam size*, nombre de séquences candidates générées en parallèle), la température (degré d'aléa dans la sélection des tokens), la longueur maximale en tokens (limite supérieure du texte généré) ainsi que la pénalité de répétition (mécanisme décourageant la répétition de tokens ou d'expressions).
La configuration finale a été retenue sur la base d'un compromis entre les métriques d'évaluation automatique, l'analyse qualitative de la préservation de la terminologie juridique, la cohérence des résumés et l'efficacité computationnelle.

D.4 Expériences sur le format d'entrée

Les prompts ont été testés à la fois sur des textes bruts et sur des documents extraits de fichiers PDF afin d'évaluer l'impact des différences de formatage sur le comportement du modèle. Aucune différence significative de performance n'a été observée après normalisation du texte.

D.5 Pipeline final

Le pipeline final intègre le découpage des documents en segments, la génération de résumés au niveau de chaque segment, ainsi que la production d'un résumé final structuré.

E Détails d'implémentation du RAG

E.1 RAG – Première approche (RAG 1)

La première approche RAG suit un pipeline classique basé sur l'enrichissement contextuel :

- Les documents juridiques sont segmentés en *chunks* à l'aide de la stratégie de découpage adaptatif décrite précédemment.
- Chaque segment est converti en une représentation vectorielle (*embedding*).
- Les représentations vectorielles sont indexées dans une base de vecteurs (*vector store*).
- Des segments pertinents sont récupérés à partir d'autres documents juridiques sur la base de la similarité sémantique.
- Le contexte récupéré est combiné avec le document cible afin de générer un résumé structuré.

Plusieurs modèles d'embeddings ont été testés, notamment les embeddings Mistral, Nomic et NV-Embed. En raison de contraintes de licence et de problèmes techniques persistants, ces modèles n'ont pas pu être utilisés de manière fiable. Le modèle d'embedding bge-m3 a finalement été retenu en raison de ses bonnes performances sur le benchmark MTEB et de son intégration stable.
Plusieurs bases de vecteurs ont également été évaluées, notamment FAISS, Qdrant et ChromaDB. FAISS a été sélectionné pour sa simplicité et son efficacité.

Des expérimentations initiales ont été menées sur des configurations réduites, puis étendues à des configurations plus larges.

E.2 Adaptation des prompts pour RAG

Les prompts utilisés dans le RAG suivent le même format structuré que celui employé pour la synthèse basée sur les prompts (Faits, Question, Conclusion). Des contraintes supplémentaires garantissent que le contenu récupéré est uniquement utilisé pour clarifier ou compléter des informations manquantes et ne doit en aucun cas contredire le document principal.

Le prompt utilisé dans RAG 1 impose explicitement que le contexte récupéré soit uniquement utilisé pour clarifier ou compléter des informations présentes dans le document principal :

```
IMPORTANT:
- Use the retrieved context only to CLARIFY or COMPLETE missing infor
- Never invent or contradict the main document
- Preserve exact legal terminology whenever possible
```

E.3 RAG – Deuxième approche (RAG 2)

La deuxième approche de RAG s'appuie sur la technique basée sur les prompts et suit un pipeline de type retrieval adapté au raffinement des résumés :

- Les documents juridiques originaux sont segmentés en chunks selon la stratégie de découpage adaptatif décrite précédemment.
- Chaque segment est représenté par un vecteur d'embedding.
- Tous les embeddings sont indexés dans une base vectorielle.
- Le résumé généré par la méthode basée sur les prompts est utilisé comme requête de recherche.
- Les segments les plus pertinents sont récupérés à partir des documents originaux de l'affaire.
- Le contexte récupéré est combiné avec le résumé initial pour produire un résumé structuré raffiné.

Plutôt que de combiner les résumés des segments, cette approche utilise la sortie du résumé basé sur les prompts (la baseline) pour récupérer les segments les plus pertinents à partir des documents originaux de l'affaire afin d'affiner le résumé. Cette méthode repose sur l'hypothèse que le résumé initial capture déjà la structure globale de l'affaire, tandis que l'enrichissement contextuel contribue à améliorer la précision, l'exhaustivité et la terminologie juridique.
Le prompt retenu pour RAG 2 demande explicitement au modèle de corriger et régénérer le résumé initial uniquement à partir du contexte récupéré :

```
RULES:
- Use the RAG document as the single source of truth
- Fix errors in the initial summary
- Add missing facts, parties, legal question, and conclusion
- Remove unsupported information
- Do NOT invent information
```

F IA agentique avec PydanticAI : Détails d'implémentation

F.1 Pipeline basé sur des agents

Le pipeline basé sur des agents suit une structure similaire aux méthodes précédentes :

— Segmentation des documents juridiques selon la stratégie de découpage adaptatif
— Résumé des segments avec un prompt optimisé
— Combinaison des résumés de chunks en une entrée unifiée
— Génération agentique basée sur un schéma de sortie prédéfini

F.2 Définition du schéma de sortie

Le schéma définit explicitement trois champs obligatoires : faits, question et conclusion. Ce schéma est fourni à l'agent comme contrainte lors de la génération.
Le prompt utilisé avec PydanticAI impose explicitement une structure juridique fixe :

```
1. Facts of the Case: [Summary of what happened]
2. Question: [The legal question being decided]
3. Conclusion: [The court's decision]

IMPORTANT:
- Preserve exact legal terminology
- Do NOT introduce new vocabulary
- Clearly state the legal question
- Provide the final ruling/decision
```

F.3 Mécanisme de validation et de régénération

PydanticAI valide chaque sortie générée par rapport au schéma. Si la validation échoue en raison de champs manquants, d'un format incorrect ou d'incohérences structurelles, l'agent demande automatiquement au modèle de régénérer la sortie. Ce processus est répété jusqu'à ce qu'une réponse valide soit produite. Le nombre de régénérations n'est pas fixé a priori et dépend dynamiquement de la conformité de la sortie générée au schéma Pydantic défini. En pratique, la régénération n'est déclenchée qu'en cas d'échec de validation structurelle (champ manquant, format incorrect ou structure invalide), et le processus s'arrête automatiquement dès qu'une sortie valide est produite.

F.4 Application à travers les pipelines

Le cadre PydanticAI a été testé de manière exploratoire sur les pipelines de résumé basés sur les prompts et sur RAG, bien que ces expériences ne constituent pas le cœur méthodologique de ce projet. Bien qu'il améliore systématiquement la conformité structurelle, son impact est particulièrement notable dans les contextes RAG, où la génération libre conduit souvent à des écarts de formatage.