UNIVERSITÉ PARIS-SACLAY

LISN - Laboratoire Interdisciplinaire des Sciences du Numérique

RAPPORT DE STAGE

Evaluation du contexte dans les systèmes RAG : une approche par propositions atomiques

Étudiant :
Luc Pommeret
Master 2 mathématiques et applications, parcours Logos (U. Paris-Cité, LISN)

Responsable pédagogique :
Michel de Rougemont
Professeur (U. Paris-Cité, IRIF)

Encadrants :
Sahar Ghannay
Maîtresse de conférences (U. Paris-Saclay, LISN)

Christophe Servan
Chercheur (LISN)

Sophie Rosset
Directrice de recherche (LISN)

1er avril - 1er octobre 2025

Remerciements

Je tiens à remercier Sophie, Christophe et Sahar, mes encadrants, qui m'ont suivi et aidé pendant ces six mois.

Merci également au sympathique groupe "Vendredi" (bien que nos réunions étaient les mardi et jeudi) composé de mes encadrants, de Thomas et Thomas, ainsi que des futurs docteurs Aygalic, Iskandar et Pierre, qui m'ont accompagné et aidé.

Merci à Patrick, nos longues discussions sur la logique et la linguistique m'ont beaucoup éclairé.

Merci au CESFO et au groupe du CESFO, dirigé par Manon et sa cloche, avec lequel nous jouions après manger, avec un café (allongé) chez Nede.

Merci à toutes les personnes qui ont facilité mon stage dans l'équipe du LISN, à l'équipe administrative, à l'équipe SAMI (merci Olivier de m'avoir prêté l'engin pour sauver mon disque dur !), tous très réactifs.

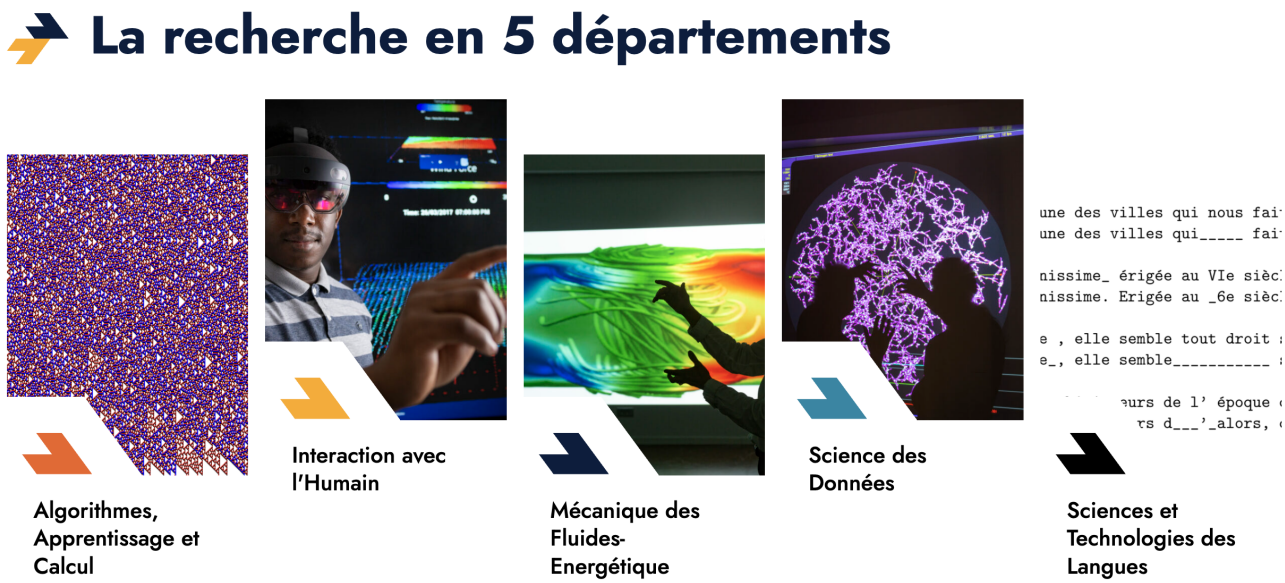
Merci à Michel, qui m'a donné l'opportunité de présenter mon sujet de stage au CRED, et qui suit mon évolution en visio, semaine après semaine.

Table des matières

Introduction générale

Ce rapport présente les travaux réalisés durant un stage de recherche au LISN (Laboratoire Interdisciplinaire des Sciences du Numérique) de l'Université Paris-Saclay, sous la direction de Sahar Ghannay, Christophe Servan et Sophie Rosset.

Environnement



Le laboratoire, issu de la fusion du Limsi et du LRI est situé sur le plateau de Saclay et est divisé en cinq départements :

— Algorithmes, apprentissage et calcul, issu de l'ancien LRI
— Interaction avec l'humain
— Mécanique des fluides, énergétique
— Science des données
— Sciences et technologies des langues, dans lequel j'ai effectué mon stage.

L'environnement de recherche stimulant m'a permis d'avoir de nombreuses discussions avec des personnes d'horizons différents (des informaticiens, des mathématiciens, mais aussi des linguistes, des psychologues et des sociologues).

Problématique

La problématique centrale de ce stage porte sur l'optimisation du contexte dans les systèmes RAG (Retrieval-Augmented Generation). Ces systèmes, qui combinent la récupération d'informations avec la génération de texte par des modèles de langage, soulèvent une question fondamentale : jusqu'où peut-on "goinfrer" un LLM avec des informations contextuelles sans compromettre sa capacité à générer des réponses précises et contraintes ?

Mes travaux se sont articulés autour de deux axes complémentaires :

Premier axe : analyse théorique et empirique des limites du RAG. J'ai mené une revue systématique de l'état de l'art sur les différentes stratégies de granularité en RAG, analysant neuf études récentes pour identifier

les seuils optimaux de performance. Cette analyse révèle que la limite du goinfrage n'est pas quantitative mais qualitative : elle dépend de l'optimisation simultanée de cinq dimensions (granularité propositionnelle, adaptation structurelle, partitionnement intelligent, récupération dynamique, et hybridation contextuelle).

Second axe : développement d'outils d'évaluation des propositions atomiques. Face au manque d'évaluation systématique de l'atomicité dans la littérature TAL actuelle, j'ai conçu AtomicEval, un framework d'évaluation qui mesure l'autonomie des propositions atomiques par cohérence contextuelle.

J'ai de plus entraîné un propositionneur français basé sur FLAN-T5, contribuant à l'extension des capacités de traitement atomique au-delà de l'anglais. oui, j'ai essayé de mieux séparer les deux en faisant deux paragraphes, c'est vrai que c'était un peu trop mélangé...

Ces deux axes convergent vers une théorie unifiée de la granularité en RAG, proposant une architecture à cinq niveaux qui articule les dimensions logique, structurelle, organisationnelle, hiérarchique et temporelle de l'optimisation contextuelle.

Le présent rapport documente cette démarche de recherche depuis les fondements théoriques jusqu'aux contributions expérimentales, en passant par une analyse comparative approfondie des méthodes existantes.

Lors de ce stage j'ai, à la suite d'une phase pratique (mise en place d'un système de chatbot pour les bibliothèques Paris-Saclay, dans la continuité d'un projet qui dure depuis plusieurs années, et qui a utilisé plusieurs approches successives), "zoomé" sur une problématique qui me semblait intéressante : les propositions atomiques.

La deuxième partie était donc plus exploratoire, et m'a permis de développer un cadre théorique, fondé sur la théorie des modèles, qui a débouché sur un système d'évaluation de l'atomicité des propositions (AtomicEval), et l'entraînement d'un modèle frugal permettant de découper un passage textuel en une liste de propositions atomiques, suivant des critères strictes.

Un **chatbot** est un système conversationnel automatisé capable d'interpréter et de répondre à des requêtes en langage naturel. Dans le contexte professionnel, ces systèmes doivent accéder à des bases de connaissances spécialisées pour fournir des réponses précises et contextualisées. Cette nécessité d'accès à l'information externe conduit à l'utilisation du **RAG** (de l'anglais Retrieval Augmented Generation).

Le RAG est une architecture qui combine deux composants essentiels : un système de récupération d'informations (*retrieval*) qui identifie les documents pertinents dans une base de connaissances, et un modèle de génération (*generation*) qui produit une réponse cohérente en s'appuyant sur ces informations récupérées. Cette approche permet aux LLM de dépasser les limitations de leurs connaissances pré-entraînées en intégrant des données actualisées et spécialisées.

La **génération contrainte** est le problème central du RAG : elle désigne la capacité d'un LLM à produire des réponses qui respectent simultanément plusieurs contraintes : fidélité stricte aux sources récupérées (pas d'hallucination), format de réponse imposé par le domaine d'application, respect des limites du domaine d'expertise, et traçabilité des sources utilisées.

Un chatbot juridique illustre parfaitement ces enjeux. Face à la question « Quels sont les délais de prescription en droit du travail ? », le système doit d'abord récupérer les articles pertinents du Code du travail (L. 1471-1, L. 1262-1), puis générer une réponse qui cite précisément ces textes, distingue les différents types de créances (salaires, congés payés, dommages-intérêts), respecte la terminologie juridique exacte, et indique les exceptions applicables. La génération contrainte garantit que la réponse reste fidèle au droit positif sans ajout d'interprétations non-fondées.

La génération contrainte permet au LLM d'extraire des informations pertinentes en regard d'une question spécifique, même lorsque ces informations sont noyées dans un contexte potentiellement non-pertinent ou volumineux.

La question qui guide ce premier chapitre est : « jusqu'où peut-on goinfrer le LLM » ? En d'autres termes, quelle est la limite de la quantité d'informations contextuelles que l'on peut fournir à un LLM sans gêner sa capacité à générer des réponses précises et contraintes ?

Le « jusqu'où » est à prendre en deux sens. D'abord dans le sens de la **quantité** de passages ou de documents que l'on peut donner au LLM, ensuite, dans le sens de la **qualité** de ces passages.

La **quantité** désigne le volume d'informations contextuelles mesurable : nombre de documents récupérés, longueur totale des passages (en mots ou tokens), nombre de chunks traités, et taille du contexte global fourni au LLM. Cette dimension purement volumétrique influence directement les coûts computationnels et la latence du système.

La **qualité** se décompose en trois dimensions distinctes mais interdépendantes :

La qualité sémantique concerne la pertinence informationnelle des passages récupérés par rapport à la requête. Elle englobe la pertinence thématique (adéquation du contenu à la question), la cohérence interne (logique et consistance des informations), la granularité informationnelle (niveau de détail approprié), et la structuration (organisation logique des informations). Cette qualité détermine directement la capacité du LLM à générer des réponses précises et contextualisées.

La qualité du processus de récupération évalue l'efficacité du système de retrieval lui-même, mesurée par des métriques techniques comme la précision (proportion de documents pertinents parmi ceux récupérés), le rappel (proportion de documents pertinents effectivement récupérés), et la vitesse de récupération. Cette dimension reflète les performances algorithmiques du système RAG.

La qualité rédactionnelle concerne les propriétés intrinsèques des documents source indépendamment de leur récupération. Elle inclut la clarté d'expression, l'exhaustivité du traitement des sujets, la variété des sources consultées, la fiabilité des informations, et la mise à jour des contenus. Cette qualité influence la capacité du LLM à produire des réponses bien formulées et complètes.

Cette problématique s'inscrit dans le cadre plus large de la génération augmentée par récupération (RAG), où l'efficacité du système dépend de la qualité du processus de récupération et de la granularité des passages fournis au modèle.

Chapitre 1

Jusqu'où peut-on goinfrer le LLM pour la génération de réponses contraintes dans le cadre d'un chatbot ?

1.1 Introduction

Dans ce chapitre, nous examinerons les conditions de possibilité d'un RAG efficace en analysant les fondements techniques et architecturaux nécessaires. Mais la question de l'efficacité du RAG (section 2) nous amène à la question de l'évaluation, que nous étudierons dans la section 3, en voyant que le paysage de l'évaluation de la génération est loin d'être unifié, et est encore aujourd'hui (juillet 2025) sans consensus. Ces différentes pistes, différentes granularités, nous les explorerons dans la section 4. Ce paysage dispersé appelle un essai de synthèse, que nous proposerons dans la section 5, où nous tenterons de dégager des tendances qualitatives et quantitatives sur les limites du goinfrage des LLM. Nous formulons trois principes directeurs qui semblent se dégager de la littérature.

1.2 Conditions de possibilité d'un RAG efficace

Un système RAG efficace repose sur deux piliers fondamentaux (récupération et génération) qui déterminent sa capacité à fournir des réponses de qualité à l'utilisateur.

1.2.1 Document retrieval de qualité

La qualité du retrieval constitue le premier maillon de la chaîne RAG. Elle dépend de plusieurs facteurs techniques :

La représentation vectorielle des documents et des requêtes doit capturer efficacement la sémantique pour permettre une similarité pertinente. L'organisation de la base de connaissances influence directement la rapidité et la précision de la récupération. Le choix de la fonction de distance (cosinus, euclidienne, etc.) impacte la pertinence des documents récupérés.

1.2.2 Architecture des LLM

Depuis l'avènement du mécanisme d'attention en 2018 (Vaswani et al. 2017a), les architectures Transformer ont révolutionné le traitement du langage naturel. Ces modèles présentent des caractéristiques particulièrement adaptées au RAG :

Le mécanisme d'attention permet au modèle de se concentrer sur les parties les plus pertinentes du contexte fourni. Les modèles récents peuvent traiter des contextes de plus en plus longs (jusqu'à plusieurs millions de tokens). L'apprentissage in-context permet d'adapter le comportement en fonction des exemples fournis dans le prompt.

1.2.3 Architecture contextuelle adaptative

L'efficacité d'un système RAG ne dépend pas uniquement de la qualité du retrieval statique, mais également de sa capacité à s'adapter au contexte spécifique de chaque requête. Anantha et Vodianik (2024) identifient une limitation des systèmes RAG traditionnels : l'échec de la recherche sémantique lorsque les requêtes manquent de contexte ou sont implicite.

Leur approche de *Context Tuning* montre qu'un système de récupération contextuelle utilisant des signaux numériques, catégoriels et d'usage habituel peut surmonter ces limitations. Les résultats expérimentaux révèlent des améliorations substantielles : 3.5 fois pour la récupération contextuelle et 1.5 fois pour la récupération d'outils, sans compromettre la capacité de traitement du système (Anantha et Vodianik 2024).

Cela montre que l'optimisation du contexte peut permettre un meilleur usage de l'information disponible sans saturer la capacité de traitement du LLM. Le fine-tuning de la recherche sémantique élimine même le besoin d'augmentation Chain-of-Thought (CoT) tout en maintenant des performances comparables (Anantha et Vodianik 2024).

1.2.4 Paradigmes évolutifs du RAG

La compréhension des limites contextuelles des LLM nécessite une perspective historique sur l'évolution des approches RAG. Gao et al. (2023) explicitent trois paradigmes successifs qui reflètent une sophistication croissante dans la gestion de l'information contextuelle. Cette taxonomie s'appuie sur l'analyse de travaux fondateurs : Lewis et al. (2020) pour le RAG naïf, et Xiong et al. (2021) pour le RAG avancé, et Jiang et al. (2023) pour le RAG modulaire.

Le RAG Naïf est un paradigme séquentiel simple (récupération → génération) caractérisé par une récupération statique unique suivie d'une génération passive. Cette approche présente des limites structurelles : précision de récupération limitée par l'absence de feedback, surcharge informationnelle durant la génération due à l'accumulation aveugle de passages, et incapacité d'adaptation contextuelle (Gao et al. 2023 ; Lewis et al. 2020).

Le RAG Avancé introduit des stratégies d'optimisation pré-récupération (query expansion, reformulation) et post-récupération (compression contextuelle, re-ranking des passages) pour gérer la densité informationnelle. Cette approche améliore l'efficacité par des mécanismes de filtrage et de priorisation statiques, mais maintient une logique séquentielle rigide (Gao et al. 2023 ; Karpukhin et al. 2020).

Le RAG Modulaire est un paradigme adaptatif caractérisé par des capacités de récupération dynamiques qui transcendent les structures fixes, permettant une gestion contextuelle des contraintes informationnelles. Cette approche intègre des mécanismes de feedback, d'adaptation temps-réel, et de routage intelligent des requêtes (Gao et al. 2023 ; Jiang et al. 2023).

Le RAG modulaire représente un saut qualitatif fondamental par rapport aux approches précédentes. Contrairement au RAG naïf qui accumule linéairement les passages récupérés sans discernement, et au RAG avancé qui optimise statiquement les processus selon des règles prédéfinies, le RAG modulaire ajuste dynamiquement ses stratégies selon les caractéristiques contextuelles de chaque requête. Cette adaptabilité permet une gestion intelligente du goinfrage : plutôt que de maximiser la quantité d'information (RAG naïf) ou d'optimiser uniformément la qualité

(RAG avancé), le RAG modulaire calibre précisément le volume et la granularité selon les besoins spécifiques de chaque contexte.

Cette progression paradigmatique du RAG naïf vers le RAG modulaire illustre une évolution qualitative fondamentale dans l'approche du goinfrage des LLM. Chaque paradigme révèle des stratégies différentes pour gérer le défi central de l'optimisation contextuelle : comment fournir suffisamment d'information pour une génération précise sans surcharger les capacités de traitement du modèle.

Le RAG modulaire, en particulier, ouvre des perspectives prometteuses pour une gestion intelligente du goinfrage. Cependant, cette sophistication technologique soulève une question méthodologique fondamentale : comment mesurer objectivement l'efficacité de ces approches évolutives ? Comment quantifier les gains réels en termes de qualité de génération tout en tenant compte des contraintes computationnelles ?

Cette interrogation méthodologique constitue un prérequis essentiel pour évaluer les limites du goinfrage des LLM. Sans instruments de mesure appropriés, il devient impossible de déterminer les seuils optimaux ou de comparer l'efficacité des différentes stratégies de granularité. La section suivante examine donc les métriques et méthodes d'évaluation qui permettent de quantifier rigoureusement la qualité de génération en RAG, établissant les fondements empiriques nécessaires à l'analyse comparative des approches de goinfrage.

1.3 Comment mesurer la qualité de la génération ?

La mesure de la qualité de génération en RAG constitue un défi méthodologique complexe qui nécessite d'évaluer simultanément la pertinence de la récupération et la fidélité de la génération. Cette section examine les métriques utilisées dans l'état de l'art et propose une analyse comparative des performances.

1.3.1 Typologie des métriques d'évaluation

Les études analysées emploient des métriques très hétérogènes qui peuvent être catégorisées selon trois dimensions principales :

Les métriques de récupération évaluent la qualité du processus de récupération avant génération. Chen et al. (2024) utilisent principalement Recall@20 et Recall@100 pour mesurer la capacité à récupérer les documents pertinents. Leurs résultats montrent que la granularité propositionnelle améliore le Recall@20 de +10.1 points pour les récupérateurs non-supervisés et +2.7 points pour les supervisés.

Le Recall@K évalue l'efficacité de la récupération d'informations. Il mesure la proportion de documents pertinents effectivement retrouvés parmi les K premiers résultats renvoyés par le système. Formellement :

$$Recall@K = \frac{\text{Nombre de documents pertinents dans les K premiers résultats}}{\text{Nombre total de documents pertinents}}$$

L'étude de Chen et al. 2024 compare l'efficacité de récupération selon différentes granularités de décomposition textuelle. Leurs expériences portent sur des corpus de questions-réponses factuelles où ils testent l'impact de la granularité propositionnelle (propositions atomiques moyennes de 11.2 mots) face aux approches traditionnelles de chunking par passages (58.5 mots en moyenne).

Les résultats révèlent un gain substantiel : la granularité propositionnelle améliore le Recall@20 de +10.1 points. Cela signifie concrètement que cette approche permet de récupérer 10.1% de documents pertinents supplémentaires dans les 20 premiers résultats par rapport aux approches de granularité classique.

Cette amélioration quantitative révèle plusieurs implications théoriques et pratiques pour le goinfrage des LLM :

La granularité fine élimine les passages non-pertinents qui diluent l'information contextuelle, permettant au LLM de se concentrer sur les éléments réellement utiles pour la génération.

En récupérant des propositions atomiques plus précises, le système fournit un contexte plus riche en informations utiles pour un volume équivalent.

L'amélioration plus marquée pour les récupérateurs non-supervisés (+10.1 vs +2.7 points) suggère que la granularité propositionnelle compense partiellement les limitations des systèmes moins sophistiqués.

Les métriques de génération mesurent la qualité du texte produit par le LLM après intégration du contexte récupéré.

BLEU. (Bilingual Evaluation Understudy) évalue la similitude n-grammes entre le texte généré et les références humaines. Cette métrique privilégie la précision lexicale et la fluidité linguistique. Le score BLEU est calculé selon :

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

où p_n représente la précision des n-grammes, w_n les poids associés, et BP (Brevity Penalty) pénalise les textes trop courts.

ROUGE. (Recall-Oriented Understudy for Gisting Evaluation) mesure le recouvrement lexical entre génération et référence, privilégiant le rappel d'informations importantes. La formule ROUGE-N est :

$$ROUGE - N = \frac{\sum_S \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_S \sum_{gram_n \in S} Count(gram_n)}$$

où $Count_{match}(gram_n)$ est le nombre de n-grammes co-occurents dans la génération et la référence.

Exact Match (EM). mesure la correspondance parfaite entre la réponse générée et la réponse attendue. Cette métrique binaire (0 ou 1) donne la capacité du système à produire des réponses précises sans approximation :

$$EM = \begin{cases} 1 & \text{si réponse générée = réponse attendue} \\ 0 & \text{sinon} \end{cases}$$

F1-score. équilibre la précision et le rappel au niveau des mots, fournissant une mesure qui pénalise les déséquilibres entre ces deux dimensions :

$$F1 = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

où la précision mesure la proportion de mots corrects dans la réponse générée, et le rappel mesure la proportion de mots pertinents effectivement inclus dans la génération.

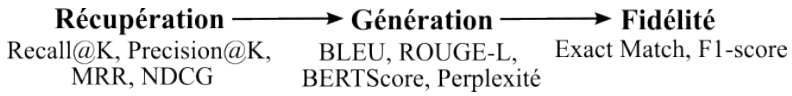


FIGURE 1.1 – Différentes métriques d'évaluations à différents moments

Classification des métriques par étape du pipeline RAG. La mesure de la qualité en RAG nécessite une approche systémique qui distingue les métriques selon leur positionnement dans le pipeline de traitement :

1. **Étape 1 - Récupération** : Recall@K, Precision@K, MRR (Mean Reciprocal Rank), NDCG (Normalized Discounted Cumulative Gain). Ces métriques évaluent la capacité du système à identifier et classer les documents pertinents avant tout traitement par le LLM.
2. **Étape 2 - Génération** : BLEU, ROUGE-L, BERTScore, Perplexity. Ces métriques mesurent la qualité linguistique et sémantique du texte produit par rapport au contexte récupéré.
3. **Étape 3 - Evaluation** : Exact Match, F1-score, métriques d'hallucination, cohérence contextuelle. Ces métriques évaluent la correspondance entre l'information source et la génération finale.

Les comparaisons inter-études sont méthodologiquement compromises en raison de la diversité des métriques employées, mais révèlent néanmoins des tendances convergentes au niveau des gains (relatifs) obtenus par les différentes approches de granularité.

1.3.2 Méthodes d'optimisation de la granularité

Avant d'analyser les performances comparatives, il convient de présenter systématiquement les principales méthodes d'optimisation de la granularité identifiées dans l'état de l'art. Ces approches se distinguent par leurs stratégies de décomposition et leurs contextes d'application.

Méthode 1 (Granularité propositionnelle). : Chen et al. (2024) développent Dense X Retrieval, une approche qui décompose les documents en propositions atomiques via un module "propositioner" spécialisé. Cette méthode cible des unités de 11.2 mots en moyenne, contre 58.5 mots pour les passages traditionnels.

Méthode 2 (Chunking structurel adaptatif). : Jimeno-Yepes et al. (2024a) proposent une approche basée sur l'analyse visuelle des documents via leur modèle Chipper, qui identifie automatiquement les éléments structurels (titres, tableaux, graphiques) pour optimiser le découpage.

Méthode 3 (RAG multi-partitions). : Wang et al. (2024) développent M-RAG, un système qui organise la base de connaissances en partitions multiples optimisées par apprentissage par renforcement avec des agents spécialisés (Agent-S pour la sélection, Agent-R pour le raffinement).

Méthode 4 (Récupération itérative). : Shao et al. (2023) proposent ITER-RETGEN, une approche qui optimise l'interaction récupération-génération par des cycles itératifs contrôlés, traitant l'information de manière holistique.

Chaque méthode présente des résultats spécifiques qui doivent être analysés dans leur contexte expérimental pour une interprétation rigoureuse.

Résultats - Granularité propositionnelle. Chen et al. (2024) obtiennent des gains de Recall@20 de +10.1 points avec des propositions de 11.2 mots en moyenne, évaluées sur Natural Questions. Contexte expérimental : décomposition automatique par module "propositioner" sur corpus de questions-réponses factuelles. Limitation : généralisation incertaine hors contexte QA factuelle.

Résultats - Chunking structurel adaptatif. Jimeno-Yepes et al. (2024a) atteignent +11% d'amélioration avec 44% de réduction du nombre de chunks sur rapports financiers. Contexte expérimental : documents visuellement structurés avec éléments tabulaires. Limitation : performance conditionnée par la structure intrinsèque des documents.

Résultats - RAG multi-partitions. Wang et al. (2024) obtiennent des améliorations variables (synthèse : +11%, traduction : +8%, dialogue : +12%) avec 4 partitions optimales. Contexte expérimental : tâches de génération avec références gold standard. Limitation : évaluation limitée aux scenarios contrôlés.

Résultats - Récupération itérative. Shao et al. (2023) observent un gain maximal de +8.6% absolu avec 2 itérations optimales. Contexte expérimental : modèles GPT-3.5/GPT-4 sur tâches multi-hop. Limitation : seuils possiblement dépendants des capacités contextuelles des modèles évalués.

1.3.3 Analyse critique et limitations

Ces résultats, bien que prometteurs individuellement, révèlent plusieurs problématiques méthodologiques qui limitent leur comparabilité et leur généralisabilité :

Hétérogénéité des tâches d'évaluation. Les études évaluent des tâches différentes (VQA, synthèse, traduction, QA factuelle), rendant les comparaisons directes problématiques. Néanmoins, la convergence remarquable autour d'améliorations de 8-12% suggère l'existence d'un plafond de performance indépendant de la tâche spécifique.

Diversité des datasets. : Les corpus d'évaluation varient considérablement : Natural Questions (Chen et al. 2024), rapports financiers (Jimeno-Yepes et al. 2024a), OK-VQA (Adjali et al. 2024a). Cette diversité limite certes la comparabilité directe, mais renforce paradoxalement la généralité des conclusions observées.

Variabilité des modèles LLM. : Les études emploient des architectures différentes (GPT-3.5, GPT-4, Claude, modèles propriétaires), introduisant des biais de performance et posant des problèmes de reproductibilité, particulièrement avec les modèles propriétaires dont les paramètres ne sont pas contrôlables.

Implications pour la généralisation. : Malgré ces limitations méthodologiques, l'émergence de tendances convergentes (rendements décroissants, seuils optimaux similaires, plafonds de performance) suggère l'existence de lois empiriques robustes qui transcendent les spécificités expérimentales.

1.3.4 Méta-analyse des performances

Malgré ces limitations, une méta-analyse montre des constantes remarquables sur les limites du goinfrage :

Rendements décroissants. Toutes les études observent une saturation des gains au-delà de seuils spécifiques : 2 itérations (Shao et al. 2023), 4 partitions (Wang et al. 2024), 11 mots par proposition (Chen et al. 2024).

Dilemme performance/coût. Li et al. (2024) quantifient ce compromis : SELF-ROUTE maintient 95% des performances des LLM long-contexte avec 65% de réduction des coûts pour Gemini-1.5-Pro (Google).

Convergence qualitative. L'identité de 60% des prédictions RAG/Long-Context (Li et al. 2024) montre que les approches optimisées convergent vers des solutions similaires pour la majorité des cas intéressants.

TABLE 1.1 – Comparaison quantitative des approches RAG selon différentes métriques

Approche	Métrique	Amélioration	Optimum	Efficacité
Dense X (CHEN et al. 2024)	Recall@20 Recall@100	+10.1 pts +2.7 pts	11.2 mots/prop	
Financial Chunking (JIMENO-YEPES et al. 2024a)	Performance Nb. chunks	+11% 62,529	Structure adapt. vs 112,155	-44% chunks
M-RAG (WANG et al. 2024)	Synthèse Traduction Dialogue	+11% +8% +12%	4 partitions	
MiRAG (ADJALI et al. 2024a)	EM (VQA) F1 (VQA)	36.6% 41.2%	2-3 entités	
FLARE (JIANG et al. 2023)	Hallucinations	-23%	Conf. dynamique	
ITER-RETGEN (SHAO et al. 2023)	Performance	+8.6% abs	2 itérations	
SELF-ROUTE (LI et al. 2024)	Coût Gemini Coût GPT-4O Convergence	-65% -39% 60% identique	Routage adapt.	95% perf.
Context Tuning (ANANTHA et VODIANIK 2024)	Récup. context. Récup. outils	+3.5x +1.5x	Fine-tuning	

1.3.5 Tableau récapitulatif des performances

Le tableau 1.1 synthétise les métriques quantitatives clés extraites des neuf études analysées, permettant une comparaison directe des approches malgré l'hétérogénéité des tâches et datasets.

Il y a donc plusieurs tendances quantitatives convergentes :

Les améliorations se situent systématiquement dans la fourchette 8-12%, (limite naturelle aux gains obtenables par optimisation de la granularité seule ?).

Chaque approche identifie des points d'équilibre (11 mots, 4 partitions, 2 itérations) au-delà desquels les rendements décroissent.

Les méthodes les plus performantes optimisent simultanément la qualité (+11%) et l'efficacité (-44% de chunks), confirmant la primauté de la qualité sur la quantité.

1.3.6 Limites de la comparabilité

Malgré l'utilité de ce tableau comparatif, plusieurs limitations méthodologiques doivent être dites :

Normalisation impossible. l'absence de datasets communs empêche une normalisation des performances. Les gains de +10.1 Recall@20 (CHEN et al. 2024) et +11% en synthèse (WANG et al. 2024) ne sont pas comparables.

Contextes d'évaluation variables. Les conditions expérimentales diffèrent (taille des corpus, complexité des requêtes, architectures LLM), introduisant des biais qui limitent la généralisation et la reproductibilité.

Métriques hétérogènes. l'emploi de métriques différentes (Recall, BLEU, EM) montre la diversité des tâches mais complique l'établissement d'un mètre-étalon.

Néanmoins, la convergence observée autour de seuils similaires (8-12% d'amélioration, rendements décroissants après optimisation) montre que ces limitations n'invalident pas les tendances générales identifiées.

1.4 État de l'art : différentes granularités en RAG

Il y a plusieurs travaux récents qui explorent différentes stratégies de granularité des passages récupérés, qui permettent d'explorer qualitativement « jusqu'où peut-on goinfrer le LLM » sans compromettre la qualité de génération.

1.4.1 Fondements théoriques de la granularité

La question de la granularité en RAG s'enracine dans des considérations logiques. Selon la taille et la forme syntaxique des passages, la récupération et la génération subséquente présentent des performances variables. Ce problème s'apparente aux questions philosophiques sur la nature des propositions atomiques, telles qu'explorées par Wittgenstein, Frege et Russell.

Une proposition atomique, dans le contexte logique classique, se définit comme une proposition ne contenant aucun connecteur logique. Dans le cadre du RAG, cette notion se transpose avantageusement vers des unités informationnelles qui se suffisent à elles-mêmes pour véhiculer un sens complet et exploitable par le LLM. « Le monde est composé de faits, pas d'objets » (Wittgenstein 1922).

1.4.2 Granularité propositionnelle : l'approche Dense X Retrieval

Chen et al. (2024) établissent dans leur étude EMNLP 2024 que la granularité propositionnelle constitue le niveau optimal pour la récupération. Leur approche « Dense X Retrieval » montre que les performances sur les propositions dépassent celles sur les phrases, et *a fortiori* sur les passages plus larges.

L'innovation réside dans l'utilisation d'un « propositioner » - un module spécialisé capable de décomposer automatiquement les documents en propositions atomiques. Cette décomposition permet d'atteindre une granularité fine tout en préservant la sémantique nécessaire à une bonne récupération.

La théorie sous-jacente postule une correspondance entre l'atomicité de la question et celle de la réponse : une question atomique requiert une réponse de même nature pour optimiser la précision du système RAG.

1.4.3 Granularité structurelle adaptative

Jimeno-Yepes et al. (2024b) proposent un chunking basé sur les éléments structurels des documents plutôt que sur une granularité arbitraire.

Leur technique repose sur l'utilisation d'un modèle de vision-encodeur-décodeur baptisé Chipper, capable d'analyser visuellement les documents pour identifier automatiquement leurs éléments structurels : titres, paragraphes, tableaux, et autres composants organisationnels. Cette approche multimodale permet de dépasser les limitations des méthodes purement textuelles qui ignorent la mise en forme.

Les résultats empiriques sont frappants : une amélioration de 11% des performances par rapport aux méthodes traditionnelles de chunking à taille fixe, tout en réduisant de moitié le nombre de chunks nécessaires. Cette bonne efficacité illustre le principe que la granularité optimale doit s'adapter à la structure des documents plutôt que d'imposer un découpage arbitraire.

Comme le disent les auteurs : *« Element-based chunking achieves the highest retrieval scores with only half the number of chunks required compared to methods that do not consider the structure of the documents. »* Donc que l'intelligence artificielle doit mimer l'approche humaine de lecture structurée pour optimiser la compréhension.

1.4.4 RAG multi-partitions avec optimisation dynamique

Wang et al. (2024) proposent une architecture révolutionnaire avec M-RAG, un RAG multi-partitions qui redéfinit la granularité au niveau organisationnel des bases de connaissances. Dans cette approche, chaque partition de base de données constitue une unité de base pour l'exécution RAG, permettant une granularité adaptative à différents niveaux d'abstraction.

L'innovation technique centrale réside dans l'utilisation d'un système multi-agents basé sur l'apprentissage par renforcement : L'Agent-S (Selection) optimise la sélection des partitions les plus pertinentes pour une requête donnée tandis que l'Agent-R (Refinement) améliore itérativement la qualité des mémoires récupérées.

Cette approche répond à la question du *goinfrage* en montrant qu'une granularité fine au niveau des partitions, couplée à une optimisation dynamique, permet des améliorations : 11%, 8%, et 12% sur trois tâches distinctes.

L'élégance de cette solution tient à son alignement naturel avec les objectifs de génération de texte. Comme le précisent les auteurs : *« M-RAG addresses all of the three challenges... The training objective of M-RAG is well aligned with that of text generation tasks. »* Cette cohérence architecturale explique en partie l'efficacité observée du système.

1.4.5 Retrieval hiérarchique multi-niveaux

Adjali et al. (2024b) développent MiRAG, une architecture hiérarchique sophistiquée qui implémente le concept de granularité multi-niveaux. Cette approche réconcilie les avantages de différents niveaux de granularité au sein d'un même pipeline.

La méthodologie procède en deux étapes (complémentaires) :

Étape 1 - Retrieval d'entités (granularité grossière) : le système identifie d'abord les entités conceptuelles pertinentes, fournissant un cadrage sémantique global pour la requête. (Exemple : Johannes Brahms).

Étape 2 - Retrieval de passages (granularité fine) : les entités récupérées servent ensuite à l'expansion de requête, enrichissant la recherche de passages spécifiques.

Cette orchestration hiérarchique résout un dilemme fondamental du RAG : comment avoir simultanément la couverture conceptuelle large (entités) et la précision informationnelle (passages). Les résultats confirment que cette granularité multi-niveaux surpasse les approches à niveau unique. Et donc qu'une bonne stratégie de *goinfrage* doit être structurée et progressive.

1.4.6 Approches dynamiques et adaptatives

Au-delà de l'optimisation statique de la granularité, une nouvelle génération d'approches RAG utilisent des stratégies dynamiques pour déterminer quand récupérer l'information contextuelle. Ces approches répondent directement à la question du goinfrage en adaptant la quantité d'information selon les besoins réels du processus de génération.

Récupération active durant la génération

Jiang et al. (2023) proposent FLARE (Forward-Looking Active REtrieval augmented generation), une approche qui permet de décider quand récupérer de l'information pendant le processus de génération. Cette méthode utilise un seuil de confiance basé sur la probabilité des tokens générés pour déterminer le moment optimal de récupération (Jiang et al. 2023).

L'innovation est dans l'utilisation de *forward-looking sentences* : le système anticipe les besoins informationnels futurs en analysant le contexte de génération en cours. Cette approche permet d'éviter la récupération excessive d'information qui pourrait surcharger le LLM, tout en garantissant la disponibilité d'informations pertinentes quand nécessaire (Jiang et al. 2023).

Les résultats montrent que FLARE surpasse les approches de récupération passive, et donc que la limite du goinfrage n'est pas fixe mais dépend dynamiquement du contexte de génération et de la confiance du modèle (Jiang et al. 2023).

Synergie itérative récupération-génération

Shao et al. (2023) développent ITER-RETGEN, une approche qui optimise la l'interaction entre récupération et génération par des itérations contrôlées. Contrairement aux approches qui interrompent fréquemment la génération, ITER-RETGEN traite toutes les connaissances récupérées de manière complète « holistique » (Shao et al. 2023).

Cette approche révèle un principe important pour le goinfrage : le traitement « holistique » de l'information récupérée est plus efficace que les interruptions fréquentes de récupération. Les résultats montrent des gains de performance allant jusqu'à 8.6% absolus par rapport aux méthodes état de l'art, avec des surcoûts computationnels réduits (Shao et al. 2023).

L'analyse empirique montre que deux itérations fournissent les gains de performance optimaux, suggérant une limite naturelle à l'accumulation d'information : au-delà de ce seuil, les gains marginaux diminuent et peuvent même devenir contre-productifs (Shao et al. 2023).

1.4.7 Optimisation comparative : RAG vs. Long-Context

Une question cruciale pour déterminer les limites du goinfrage concerne le choix entre l'augmentation de la récupération RAG et l'utilisation de LLM à contexte long. Li et al. (2024) fournissent une analyse comparative.

Analyse coût-performance

L'étude montre que les LLM à contexte long surpassent toujours les approches RAG lorsque les ressources computationnelles sont suffisantes. Cependant, cette supériorité s'accompagne de coûts computationnels plus élevés (Li et al. 2024)...

Les résultats quantitatifs montrent que les prédictions RAG et Long-Context sont identiques pour plus de 60% des requêtes, et donc qu'il existe des limites contextuelles où la récupération sélective équivaut au traitement de contexte complet (Li et al. 2024).

Approche hybride SELF-ROUTE

Li et al. (2024) proposent SELF-ROUTE, une méthode qui route dynamiquement les requêtes vers l'approche optimale (RAG ou Long-Context) selon leurs caractéristiques. Cette approche hybride atteint des performances comparables aux LLM à contexte long tout en réduisant les coûts de 65% pour Gemini-1.5-Pro et 39% pour GPT-4o (Li et al. 2024).

Cette innovation répond à la question du goinfrage en montrant que la limite optimale n'est pas universelle mais dépend des caractéristiques de chaque requête (le monde réel ne répond pas à une seule méthode). L'approche SELF-ROUTE : plutôt que de déterminer une limite globale, le système adapte dynamiquement sa stratégie selon le contexte (Li et al. 2024).

1.5 Synthèse et perspectives

1.5.1 Convergence des approches : vers une granularité intelligente

L'analyse comparative des travaux récents révèle une convergence vers un principe fondamental : la limite du goinfrage des LLM n'est pas *quantitative* mais *qualitative*. Cela remet en question l'approche naïve qui consisterait à simplement augmenter le volume d'informations contextuelles !

Les axes d'innovation identifiés dans l'état de l'art convergent vers :

La granularité propositionnelle atomique : Chen et al. (2024) démontrent que les propositions (11.2 mots en moyenne) surpassent les passages traditionnels (58.5 mots) avec des gains de +10.1 Recall@20. La granularité structurelle adaptative : Jimeno-Yepes et al. (2024a) changent l'approche classique avec un chunking basé sur les éléments visuels, atteignant 11% d'amélioration avec moitié moins de chunks. La granularité organisationnelle dynamique : Wang et al. (2024) optimisent l'organisation multi-partitions avec des améliorations de 11%, 8% et 12% sur trois tâches distinctes. La granularité hiérarchique progressive : Adjali et al. (2024a) combinent différents

niveaux dans un pipeline unifié pour la VQA. La granularité adaptative temporelle : Jiang et al. (2023) et Shao et al. (2023) introduisent des approches dynamiques qui ajustent la récupération selon les besoins en temps réel.

1.5.2 Implications théoriques et pratiques

Ces résultats imposent une reformulation de la question de recherche initiale. Plutôt que de chercher à déterminer « jusqu'où peut-on goinfrer le LLM », il convient de s'interroger sur « comment peut-on goinfrer intelligemment (en optimisant l'espace) le LLM ».

L'optimisation de la génération contrainte est un problème d'ingénierie où la pertinence et la structuration priment sur la quantité. Cette perspective, validée par l'ensemble des études analysées, ouvre plusieurs pistes de recherche prometteuses :

Les travaux de Jiang et al. (2023) et Anantha et Vodianik (2024) démontrent la faisabilité de systèmes ajustant dynamiquement la granularité selon la complexité de la requête et la confiance du modèle. L'approche de Jimeno-Yepes et al. (2024a) avec leur modèle Chipper illustre l'intégration de signaux visuels, structurels et sémantiques pour optimiser le chunking. Les méthodes multi-agents de Wang et al. (2024) utilisent l'apprentissage par renforcement pour découvrir les granularités optimales spécifiques à chaque domaine. L'approche SELF-ROUTE de Li et al. (2024) montre comment optimiser dynamiquement le choix entre récupération sélective et contexte long selon les caractéristiques de la requête.

1.5.3 Vers une théorie de la granularité en RAG ?

L'émergence de ces différentes approches révèle la nécessité pressante d'une théorie unifiée de la granularité en RAG. Cette théorie devrait articuler les dimensions fondamentales identifiées dans cette revue, maintenant enrichies par les contributions des neuf études analysées.

Fondements d'une théorie de la granularité

Une théorie cohérente de la granularité pourrait s'articuler autour de trois principes unificateurs :

1. Principe d'adaptabilité contextuelle : la granularité optimale doit s'adapter dynamiquement aux propriétés du corpus (structure documentaire), aux caractéristiques de la requête (complexité sémantique), et aux contraintes computationnelles (latence, précision).

2. Principe de monisme sémantique : chaque unité de granularité doit constituer une entité informationnelle autonome et cohérente, capable de répondre à une classe de requêtes sans nécessiter de contexte externe.

3. Principe d'efficacité computationnelle : l'architecture de granularité doit optimiser la précision de récupération, la vitesse de traitement, et l'utilisation des ressources mémorielles.

Architecture théorique proposée

Contribution personnelle : Nous proposons une implémentation de cette théorie unifiée selon l'architecture à cinq niveaux suivante, constituant notre contribution principale à l'état de l'art :

Niveau 1 - Granularité *logique* : Décomposition en propositions atomiques selon les principes de Chen et al. (2024), garantissant l'indivisibilité sémantique des unités d'information ($\approx$ 11 mots optimaux).

Niveau 2 - Granularité *structurelle* : Reconnaissance et préservation des éléments organisationnels du document, avec les innovations de Jimeno-Yepes et al. (2024a) avec leur modèle Chipper pour l'analyse visuelle.

Niveau 3 - Granularité *organisationnelle* : Partitionnement intelligent selon les critères thématiques et contextuels, inspiré de l'approche multi-partitions de Wang et al. (2024) avec optimisation par apprentissage par renforcement.

Niveau 4 - Granularité *hiérarchique* : Orchestration multi-niveaux avec Adjali et al. (2024a), permettant la récupération progressive du général (entités) au spécifique (propositions).

Niveau 5 - Granularité *temporelle* : Adaptation dynamique de la récupération selon les approches de Jiang et al. (2023) et Shao et al. (2023), optimisant le timing et la quantité d'information selon le contexte.

Mécanismes d'orchestration

Validation expérimentale

Les benchmarks multi-domaines nécessitent une évaluation sur des corpus diversifiés (juridique, scientifique, technique) pour tester l'adaptabilité contextuelle. Les métriques holistiques requièrent le développement de métriques intégrant précision, efficacité computationnelle et cohérence sémantique. Les analyses comparatives nécessitent une confrontation systématique avec les approches mono-niveau existantes pour quantifier les gains de performance.

1.5.4 Réponse synthétique à la question de recherche

L'analyse des neuf études permet de formuler une réponse à jour à la question « jusqu'où peut-on goinfrer le LLM pour la génération de réponses contraintes ? » :

La limite du goinfrage n'est pas une frontière quantitative fixe, mais un optimum qualitatif adaptatif qui dépend de cinq faits interdépendants :

1. les propositions atomiques ($\approx$ 11 mots) constituent l'unité optimale (Chen et al. 2024)
2. l'adaptation à la hiérarchie visuelle améliore l'efficacité de 11% (Jimeno-Yepes et al. 2024a)
3. le partitionnement (4 partitions optimales) optimise la sélection (Wang et al. 2024)
4. la récupération active surpasse les approches statiques (Jiang et al. 2023)
5. l'approche SELF-ROUTE montre l'optimalité contextuelle (Li et al. 2024)

Corollaire pratique : on peut « goinfrer » le LLM aussi loin que nécessaire, à condition d'optimiser simultanément la qualité (granularité propositionnelle), la structure (adaptation visuelle), l'organisation (partitionnement intelligent), la temporalité (récupération dynamique) et la stratégie (hybridation contextuelle).

On transforme donc la question de recherche initiale : plutôt que de chercher une limite absolue, l'enjeu devient l'orchestration intelligente de multiples dimensions qualitatives de la granularité pour maximiser la pertinence informationnelle tout en respectant les contraintes computationnelles du système.

1.6 Le RAG pour la recherche d'information pour la Bibliothèque Paris-Saclay

La Dibiso est ... (à compléter, (présenter le stage), le chatbot, etc.)

1.6.1 Architecture expérimentale proposée

Pipeline de granularité adaptive

Nous proposons d'implémenter un pipeline à 5 niveaux inspiré des meilleures pratiques identifiées :

1. **Pré-traitement intelligent** : extraction et normalisation multimodale (texte, tableaux, images) avec conservation de la structure documentaire.

2. **Granularité propositionnelle** : décomposition en propositions atomiques via un module *propositioner* entraîné sur le corpus Dibiso, ciblant 8-12 mots par proposition selon les résultats de Chen et al. (2024).
3. **Indexation hiérarchique** : double indexation entités/propositions permettant la récupération multi-niveaux de Adjali et al. (2024a).
4. **Sélection adaptative** : module de routage SELF-ROUTE adapté déterminant dynamiquement la stratégie optimale (RAG vs contexte long) selon Li et al. (2024).
5. **Génération contrainte** : intégration FLARE pour récupération active durant la génération selon Jiang et al. (2023).

Protocole d'évaluation

Le dataset d'évaluation Dibiso-QA est composé de 149 questions portant sur des questions courantes des utilisateurs de la bibliothèque ("comment faire un quitus ?", "comment imprimer ?", etc.).

Protocole d'annotation : j'ai annoté moi-même les questions qui préexistaient au stage.

Métriques d'évaluation :

RECUPERATION :

— Recall@5, Recall@10 : proportion de documents pertinents récupérés

— précision moyenne@5, 10, 15

— NDCG@10 : qualité du classement (pondérée par la pertinence)

GENERATION :

1.6.2 Plan d'implémentation technique détaillé

Architecture logicielle

Stack technologique :

— **Backend** : Python 3.11+

— **Embedding et récupération** : différents modèles que l'on a sur Ollama (comme Mge, e5, etc.)

— **LLM** : Ollama, avec différents modèles open-weight (mais pas toujours open-source).

— **Base de données** : PostgreSQL avec extension pgvector pour les embeddings

— **Orchestration** : Docker pour la solidité

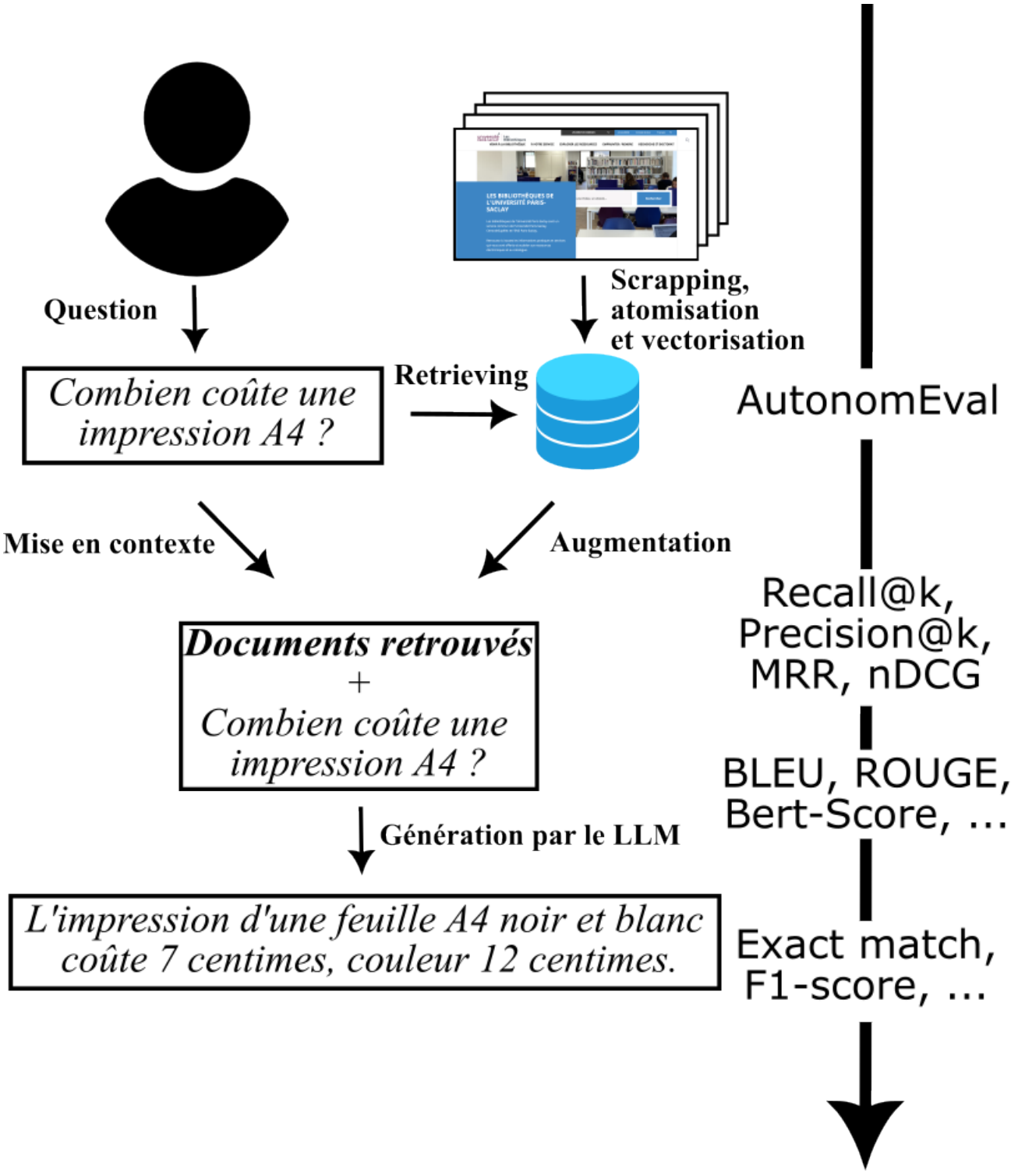


FIGURE 1.2 – L'architecture général du système de RAG proposé pour la Dibiso

Chapitre 2

Mais qu'est-ce qu'une proposition atomique après tout ?

AtomicEval : évaluation de l'autonomie des propositions atomiques

2.1 Introduction

Les propositions atomiques occupent une place centrale dans la logique moderne depuis les travaux fondateurs de Gottlob Frege à la fin du XIXe siècle. Ce concept, né de la nécessité de formaliser le raisonnement mathématique de manière rigoureuse, a traversé plus d'un siècle d'évolutions théoriques pour trouver aujourd'hui une nouvelle actualité dans le domaine du traitement automatique du langage naturel (TAL).

L'expression $H(s) \wedge M(s)$ peut être découpée en $H(s)$ et $M(s)$. On a donc deux unités sémantiques. Elles sont bien autonomes, car on peut les interpréter dans un modèle, et bien sémantiques, car susceptibles d'être satisfaites ou non par ce modèle.

Cependant, malgré l'utilisation généralisée des propositions atomiques dans la recherche NLP récente, aucun travail n'évalue si ces propositions sont réellement atomiques. Cet article propose AtomicEval, un framework d'évaluation systématique qui comble cette lacune en fournissant une méthode pour évaluer l'autonomie des propositions atomiques.

2.2 Fondements historiques et théoriques

2.2.1 Aux origines : Gottlob Frege (1848-1925)

Le problème de Frege était de formaliser le raisonnement mathématique de manière rigoureuse, dans le contexte de la révolution de la logique moderne à la fin du XIXe siècle. Une proposition atomique constitue pour Frege une expression logique élémentaire qui exprime un fait simple sur le monde, ne peut être décomposée en parties plus simples, et possède une valeur de vérité.

Comme l'écrit Frege dans sa *Begriffschrift* (1879) : « [...] les expressions logiques contiennent des signes primitifs et définis [...] les signes primitifs ne peuvent être décomposés davantage par l'analyse. »

« Socrate est mortel » constitue une proposition atomique, contrairement à « Socrate est mortel et sage » qui en combine deux.

2.2.2 L'atomisme logique de Russell et Wittgenstein

Bertrand Russell (1872-1970) développe une théorie complète où la réalité se compose de faits atomiques. Chaque fait atomique correspond à une proposition atomique, et les faits complexes résultent de combinaisons logiques de faits atomiques. Selon Russell dans les *Principia Mathematica* (vol. 2, 1912) : « Étant donnés toutes les propositions atomiques vraies, avec le fait qu'elles sont toutes, toute autre proposition vraie peut théoriquement être déduite par des méthodes logiques. »

Ludwig Wittgenstein précise dans le *Tractus Logico-Philosophicus* (1921) la nature des propositions atomiques. Elles se caractérisent par leur simplicité (absence de connecteurs), leur atomicité sémantique (expression d'un fait indivisible), leur autonomie (sens complet) et leur correspondance directe au monde :

> « 1.1 Le monde est la totalité des faits, non des choses. »
> « 2 Ce qui est le cas, le fait, est l'existence de faits atomiques. »
> « 2.061 Les faits atomiques sont indépendants les uns des autres. »
> « 4.21 La proposition la plus simple, la proposition élémentaire, affirme l'existence d'un fait atomique. »

2.2.3 Pérennité du concept en logique moderne

Le concept de proposition atomique perdure dans tous les développements de la logique moderne. En logique propositionnelle, les variables propositionnelles servent d'atomes pour construire des formules complexes par des connecteurs. La théorie des modèles utilise les formules atomiques comme briques élémentaires pour l'interprétation dans des structures formelles de manière récursive.

Les applications modernes confirment cette importance : les bases de données utilisent des requêtes atomiques, l'intelligence artificielle manipule des faits de base, la programmation logique s'appuie sur des clauses de Horn, et les systèmes experts traitent des assertions élémentaires.

2.3 Appropriation en TAL moderne

2.3.1 Redéfinition

Chen et al. 2024 propose une redéfinition des propositions atomiques adaptée aux besoins du TAL moderne :

> « Propositions are defined as atomic expressions within text, where each encapsulates a distinct factoid and is presented in a concise, self-contained natural language format. »

L'accent se déplace vers les « factoids » informationnels, adaptant le concept aux besoins du TAL moderne. Les nouvelles caractéristiques privilégient l'auto-suffisance textuelle, la granularité informationnelle optimale et la possibilité d'extraction automatique.

« Apple a été fondée en 1976 » et « Paris est la capitale de la France » constituent des propositions atomiques selon cette définition. En revanche, « Bien qu'il fasse beau, Marie préfère rester à la maison car elle est fatiguée » ne l'est pas.

L'utilisation du terme « factoid » est en soi problématique. En effet, ce terme est créé et défini par Malcolm Mailer, un journaliste, comme étant des "facts which have no existence before appearing in a magazine or newspaper". Ce n'est pas ce que nous voulons...

Nous préférons le terme "état de fait" ou "état de choses" qui correspond mieux au balisage qu'en a fait Wittgenstein. La proposition atomique est le corrélat langagier de l'état de fait.

2.3.2 Graphes de connaissance

Diagrammes de Peirce, extraction sémantique, etc.

2.3.3 Applications

Ces propositions atomiques alimentent désormais l'extraction d'information, le résumé automatique, les systèmes de question-réponse et la construction de graphes de connaissances. Citons :

— **FACTScore** Min et al. 2023 : Décompose les textes en faits atomiques pour évaluer la précision factuelle
— **Inférence Atomique** Stacey et al. 2024 : Utilise des faits générés pour le raisonnement NLI
— **Recherche Dense** Chen et al. 2024 : Emploie des propositions atomiques comme unités de recherche
— **Evaluation de la qualité d'un résumé** Herserant et Guigue 2025 : Utilise les propositions atomiques pour évaluer les résumés

2.3.4 Différentes définitions et leurs implications

La littérature présente diverses approches de l'atomicité, avec ou sans exigence d'autonomie (Chen et al. 2024 vs. Min et al. 2023). Cette diversité révèle un manque de consensus théorique sur ce que constitue réellement une proposition atomique dans le contexte du TAL.

Voici quelques distinctions à faire :

1. Clause d'insécabilité : la proposition ne peut pas être découpée en autres propositions sans perdre le sens.
2. Clause d'autonomie : la proposition doit avoir une interprétation univoque.

Comme nous l'avons vu, la clause d'autonomie n'est pas toujours nécéssaire dans la littérature.

2.4 Le problème de l'atomicité réelle

Malgré l'utilisation généralisée des propositions atomiques dans la recherche NLP récente, **aucun travail n'évalue si ces propositions sont réellement atomiques**. Cette lacune est problématique car elle remet en question l'efficacité des méthodes actuelles d'évaluation factuelle.

2.4.1 Exemple problématique

Considérons cet exemple tiré de FACTScore :

« He appeared as an actor. »

Cette proposition est marquée comme « supported » par FACTScore mais n'est clairement pas autonome : **qui est « He » ?** Le contexte complet révèle : « Joey D. Vieira is an American actor [...] He also appeared as an actor in films such as "The Wild One" and "Viva Las Vegas". »

Problème central : Les méthodes actuelles supposent l'atomicité sans l'évaluer systématiquement.

2.5 AtomicEval : Un framework d'évaluation de l'autonomie

2.5.1 Principe et méthodologie

AtomicEval est un framework d'évaluation qui comble cette lacune en fournissant une méthode systématique pour évaluer l'autonomie des propositions atomiques. Le principe clé est qu'une proposition atomique doit être

autonome, c'est-à-dire compréhensible sans contexte supplémentaire.

La méthodologie se déroule en deux étapes :

1. **Développement du contexte** : Un LLM développe le contexte nécessaire à la compréhension de la proposition
2. **Évaluation de l'autonomie** : Comparaison avec le contexte original pour déterminer l'auto-suffisance

2.5.2 Logique d'évaluation

La logique d'évaluation repose sur la comparaison entre contexte généré et contexte réel :
— Si le contexte généré correspond étroitement au vrai contexte → NON-AUTONOME
— Si le contexte généré était minimal ou inexact → AUTONOME

Exemples :
— « Il a pris la décision hier. » (non-autonome)
— « La Tour Eiffel mesure 330 mètres. » (autonome)

2.5.3 Résultats empiriques

L'évaluation d'AtomicEval sur FACTScore donne un **score d'autonomie de 52,14%**. Cette mesure indique que :

— **52,14%** des propositions atomiques de FACTScore sont autonomes
— **47,86%** nécessitent un contexte supplémentaire pour être comprises

Ces résultats remettent en question l'efficacité des méthodes actuelles d'évaluation factuelle et soulignent la nécessité d'une évaluation systématique de l'atomicité.

2.6 Applications en RAG

2.6.1 Implémentation pour la recherche d'information

Les bonnes performances obtenues par Chen et al. 2024 sur Wikipédia dans le cadre d'un RAG utilisant un *propositioner* encouragent son utilisation pour des projets de recherche d'information. Les résultats établissent que les propositions de 11,2 mots en moyenne surpassent les passages de 58,5 mots avec un gain de Recall@20 de +10,1.

2.6.2 Méthode de promptage

Le prompt structure explicitement les critères de Chen et al. 2024 pour guider le modèle de langage :

```
Tu es un expert en segmentation de texte pour des systèmes de recherche avancés. Transforme
la phrase suivante en propositions atomiques selon les critères suivants :
CRITÈRES DES PROPOSITIONS :
1. ATOMIQUES : Chaque proposition contient exactement UN fait distinct
2. AUTO-CONTENUES : Chaque proposition est compréhensible sans contexte externe
3. MINIMALES : Non décomposables davantage ( 8-15 mots idéalement)
4. PRÉCISES : Préservent l'information exacte de la phrase originale
```

2.6.3 Exemple concret

Texte d'entrée (extrait documentation Paris-Saclay) :

> « Focus vous permet d'accéder à l'ensemble des livres, ebooks, revues scientifiques, articles, bases de données et documents numériques disponibles au sein des bibliothèques et centres de documentation de l'Université Paris-Saclay. »

Résultat : 60 propositions générées automatiquement à partir de 3352 caractères, avec une longueur moyenne de 10,8 mots par proposition.

2.7 Justification théorique : Retrouver la sémantique formelle avec les LLMs

2.7.1 Cadre théorique général

Il nous semble que le probing permet de faire rentrer à nouveau la théorie des modèles en linguistique après le bouleversement causé par les LLMs. L'objet premier est la suite de symboles en langue naturelle s, à partir de laquelle tout est construit.

Transformers

Un modèle transformer est caractérisé par plusieurs paramètres clés :
— d - dimension des vecteurs d'embedding
— n - nombre de tokens dans le vocabulaire
— N - longueur maximale de séquence que le modèle peut traiter
Les paramètres additionnels qui définissent l'architecture incluent :
— H - nombre de têtes d'attention
— D - dimension de la couche cachée du réseau feed-forward
— L - nombre de couches transformer
— p - précision numérique des poids
Un Transformer prend N tokens consécutifs en entrée et produit une distribution de probabilité sur le vocabulaire pour le $(N+1)$-ème token. L'architecture consiste en deux composants principaux : le mécanisme d'auto-attention et le réseau de neurones feed-forward.

Pour simplifier, considérons d'abord une seule tête d'attention ($H = 1$). Le mécanisme implique trois matrices apprises :
— $Q \in \mathbb{R}^{d \times d}$ - matrice de transformation de requête
— $K \in \mathbb{R}^{d \times d}$ - matrice de transformation de clé
— $V \in \mathbb{R}^{d \times d}$ - matrice de transformation de valeur
Considérons une séquence de N tokens : $t_1, ..., t_N$ où $x_i \in \mathbb{R}^d$ est l'embedding du token t_i. Le calcul procède comme suit :

1. D'abord, l'encodage positionnel est ajouté à chaque embedding : $x'_i = x_i + \text{pos}(i)$ où $\text{pos}(i)$ est l'encodage positionnel pour la position i
2. Pour chaque position i, calculer les vecteurs de requête, clé et valeur :

$$q_i = (Q \cdot x'_i)^T, \quad k_j = K \cdot x'_j, \quad v_j = V \cdot x'_j$$

3. Calculer les poids d'attention en utilisant le produit scalaire mis à l'échelle :

$$(r_{i,1}, ..., r_{i,N}) = \text{Softmax}\left(\frac{q_i \cdot k_1}{\sqrt{d}}, ..., \frac{q_i \cdot k_N}{\sqrt{d}}\right)$$

4. Calculer la somme pondérée des valeurs :

$$y_i = \sum_{j=1}^{N} r_{i,j} \cdot v_j$$

5. Appliquer la connexion résiduelle et la normalisation de couche :

$$y'_i = \text{LayerNorm}(x'_i + y_i)$$

Ce processus transforme la séquence d'embeddings $x_1, ..., x_N$ en nouveaux vecteurs $y'_1, ..., y'_N$ de même dimension d.

Après le mécanisme d'auto-attention, chaque position passe par un réseau de neurones feed-forward :

1. Appliquer un perceptron à deux couches avec activation ReLU :

$$z'_i = W_2 \cdot \text{ReLU}(W_1 \cdot y'_i)$$

où $W_1 \in \mathbb{R}^{D \times d}$ et $W_2 \in \mathbb{R}^{d \times D}$

2. Appliquer la connexion résiduelle et la normalisation de couche :

$$z_i = \text{LayerNorm}(y'_i + z'_i)$$

La fonction $\text{ReLU}(x) = \max(0, x)$ est appliquée élément par élément.

La transformation $x_i \to z_i$ constitue une couche. En pratique, les transformers utilisent plusieurs têtes d'attention et empilent plusieurs couches :

— Pour $H > 1$ têtes, l'embedding est partitionné en H vecteurs de dimension d/H, avec chaque tête ayant son propre ensemble de matrices $Q, K, V \in \mathbb{R}^{d/H \times d/H}$

— Les sorties de toutes les têtes sont concaténées et transformées linéairement

— Le processus entier est répété L fois avec des paramètres différents pour chaque couche

Après traitement par toutes les couches, la sortie finale est transformée en distribution de probabilité sur le vocabulaire :

$$D_{\text{output}}(x_1, ..., x_N) = \text{Softmax}(W_0 \cdot z_N)$$

où $W_0 \in \mathbb{R}^{n \times d}$ est une matrice de paramètres apprise et z_N est l'embedding du dernier token après toutes les transformations.

Le modèle transformer complet est ainsi paramétré par les matrices $Q, K, V \in \mathbb{R}^{d \times d}$ (pour chaque tête et couche), $W_0 \in \mathbb{R}^{n \times d}$, $W_1 \in \mathbb{R}^{D \times d}$, $W_2 \in \mathbb{R}^{d \times D}$ (pour chaque couche), et les paramètres des opérations LayerNorm.

Une propriété clé des modèles transformer est leur capacité à générer des séquences en prédisant itérativement le token suivant. Étant donnée une séquence $x_1, \ldots, x_i$ avec $i \leq N$, le modèle sélectionne x_{i+1} selon la distribution conditionnelle $p(x_{i+1} \mid x_1, \ldots, x_i)$.

Le processus de génération autorégressif fonctionne comme suit :

— En commençant par une séquence vide, nous échantillonnons x_1 avec probabilité $p(x_1)$

— Nous échantillonnons ensuite x_2 avec probabilité $p(x_2 \mid x_1)$

— Ceci continue jusqu'à ce que nous ayons généré la longueur de séquence désirée

Ce processus définit implicitement une distribution de probabilité jointe sur les séquences :

$$p(x_1, \ldots, x_{N+1}) = p(x_1) \cdot p(x_2 \mid x_1) \cdot \ldots \cdot p(x_{N+1} \mid x_1, \ldots, x_N)$$

Une observation importante est que ceci représente une compression hautement efficace de la distribution jointe complète. Le support de la distribution complète est de taille n^{N+1}, ce qui nécessiterait un espace exponentiel pour être représenté explicitement. Cependant, le transformer paramétrise cette distribution en utilisant seulement :

$$O(L \cdot (d^2 + n \cdot d) + N \cdot d)$$

paramètres, où :

— L est le nombre de couches
— d est la dimension d'embedding
— n est la taille du vocabulaire
— N est la longueur maximale de séquence

Cette compression efficace permet aux transformers de modéliser des distributions de séquences complexes avec un nombre gérable de paramètres, ce qui est crucial pour leur application pratique en modélisation de langage et tâches connexes Manning et al. 2014 ; Hewitt et Manning 2019 ; Vaswani et al. 2017b.

Définition 1 (Représentation vectorielle)**.** *Soit g un LLM et s une suite de symboles en langue naturelle. s et la couche c induisent une représentation vectorielle :*

$$M_{g,c}(s) = g.encode(s) \in \mathbb{R}^d$$

Définition 2 (Représentation matricielle)**.** *Lorsqu'on considère toutes les couches à la fois d'un transformeur à n couches, on obtient la matrice notée ainsi :*

$$M_g(s) = concat(M_{g,1}, M_{g,2}, \ldots, M_{g,n})$$

2.7.2 Définissabilité d'une propriété dans un LLM

Définition 3 (Sonde)**.** *Pour une relation R d'arité k et une représentation $M_g(s)$, une* sonde *est une famille de fonctions :*

$$probe_{R(\vec{a})} : \mathbb{R}^d \to \{0, 1\}$$

où $\vec{a} = (a_1, \ldots, a_k)$ sont les arguments de la relation.

Définition 4 (Sonde linéaire)**.** *Une* sonde linéaire *pour une relation R et des constantes $\vec{a}$ est une fonction :*

$$probe_{R(\vec{a})} = \mathbf{1}[\langle w, x \rangle + b > 0]$$

où $w \in \mathbb{R}^d$, $b \in \mathbb{R}$, et $\mathbf{1}[\cdot]$ est la fonction indicatrice.

Définition 5 (Vérité terrain)**.** *Pour une relation $R(\vec{a})$ et une phrase s, on note $R_s^{réel}(\vec{a}) \in \{0, 1\}$ la valeur de vérité de la relation R appliquée aux arguments $\vec{a}$ dans la situation décrite par s.*

Définition 6 (Précision d'une sonde linéaire)**.** *La précision d'une sonde pour une relation R d'arité k est :*

$$acc(R) = \mathbb{P}_{s,\vec{a}}[probe_{R(\vec{a})}(M_g(s)) = R_s^{réel}(\vec{a})]$$

où la probabilité est prise sur toutes les phrases s et tous les k-uplets valides $\vec{a}$.

Définition 7 (Définissabilité linéaire d'une relation)**.** *Une relation R est* linéairement définissable *dans les représentations de g si et seulement si :*

$$acc(R) = 1$$

pour tout s et tout $\vec{a}$ valide.

Autrement dit :

Lemme 1 (Séparabilité linéaire)**.** *Si $acc(R) = 1$, alors R est linéairement séparable dans l'espace $\mathbb{R}^d$ (par un hyperplan).*

Démonstration. Supposons que $\text{acc}(R) = 1$. Par définition, cela signifie que pour tout s et tout $\vec{a}$ valide, on a :

$$\text{probe}_{R(\vec{a})}(M_g(s)) = R_s^{\text{réel}}(\vec{a})$$

Considérons les deux ensembles suivants dans $\mathbb{R}^d$:

$$S_1 = \{M_g(s) : R_s^{\text{réel}}(\vec{a}) = 1\} \quad \text{(points positifs)} \tag{2.1}$$

$$S_0 = \{M_g(s) : R_s^{\text{réel}}(\vec{a}) = 0\} \quad \text{(points négatifs)} \tag{2.2}$$

Puisque la sonde linéaire $\text{probe}_{R(\vec{a})} = \mathbf{1}[\langle w, x\rangle + b > 0]$ atteint une précision parfaite, nous avons nécessairement :
— Pour tout $x \in S_1$: $\langle w, x\rangle + b > 0$
— Pour tout $x \in S_0$: $\langle w, x\rangle + b \leq 0$
L'hyperplan $H = \{x \in \mathbb{R}^d : \langle w, x\rangle + b = 0\}$ sépare donc parfaitement S_1 et S_0.
Par définition de la séparabilité linéaire, R est linéairement séparable dans $\mathbb{R}^d$. □

Exemple 1 (ChessGPT)**.** *Pour un LLM entraîné à jouer aux échecs sur des PGN (ChessGPT), on considère sur un PGN s la relation $At_s(c, p)$ avec c la case (64 états) et p la pièce (13 états).*
On peut entraîner une sonde linéaire par case et par pièce sur tous les PGN s :

$$probe_{At_s} : \mathbb{R}^d \to \{0, 1\}$$

telle que $probe_{At_s(c,p)} = 1 \iff At_s^{réel}(c, p)$
pour toute case c, pièce p et PGN s.

2.7.3 Modèles induits

La définissabilité linéaire permet de considérer les matrices de représentations comme des *modèles* au sens de la théorie des modèles, c'est-à-dire des structures satisfaisant des formules construites à partir des relations, fonctions et constantes définissables.

Théorème 1 (La matrice de représentation comme modèle)**.** *Soit $M_g(s)$ une matrice de représentation de s par le modèle g, $\mathcal{R} = (R_1, R_2, \ldots, R_n)$ les n relations définissables dans g, $\mathcal{F}$ et $\mathcal{C}$ l'ensemble des fonctions et constantes considérées. Le domaine E est défini comme l'ensemble des constantes.*
La $\mathcal{L}$-structure suivante est un modèle :

$$\mathfrak{M}_{g,s} = (E, \mathcal{R}, \mathcal{F}, \mathcal{C})$$

Démonstration. Pour montrer que $\mathfrak{M}_{g,s} = (E, \mathcal{R}, \mathcal{F}, \mathcal{C})$ est un modèle au sens de la théorie des modèles, nous devons vérifier que cette structure satisfait les axiomes d'une $\mathcal{L}$-structure valide.

1. **Domaine non-vide :** Le domaine E est défini comme l'ensemble des constantes extraites de la suite de symboles s. Puisque s est une suite non-vide contenant au moins des entités référencées (par hypothèse de travail sur des phrases significatives), nous avons $E \neq \emptyset$.
2. **Interprétation des relations :** Chaque relation $R_i \in \mathcal{R}$ est définissable dans g avec $\text{acc}(R_i) = 1$. Pour chaque R_i d'arité k_i et tout k_i-uplet $\vec{a} \in E^{k_i}$, la valeur de vérité de $R_i(\vec{a})$ est déterminée par :

$$R_i^{\mathfrak{M}_{g,s}}(\vec{a}) = \text{probe}_{R_i(\vec{a})}(M_g(s)) \in \{0, 1\}$$

   Cette interprétation est bien définie car la sonde linéaire produit toujours une valeur (binaire).
3. **Interprétation des fonctions :** Pour chaque fonction $f \in \mathcal{F}$ d'arité n, l'interprétation $f^{\mathfrak{M}_{g,s}} : E^n \to E$ est définie via les transformations apprises dans les représentations M_g. La définissabilité de f dans g garantit que pour tout $\vec{a} \in E^n$, il existe un unique $b \in E$ tel que $f^{\mathfrak{M}_{g,s}}(\vec{a}) = b$.
4. **Interprétation des constantes :** Chaque constante $c \in \mathcal{C}$ est interprétée comme un élément unique $c^{\mathfrak{M}_{g,s}} \in E$. Cette interprétation est directement extraite de la représentation $M_g(s)$ où chaque entité nommée correspond à une constante identifiable.
5. **Clôture du domaine :** Le domaine E est clos sous les opérations de $\mathcal{F}$ par construction : pour toute fonction $f \in \mathcal{F}$ et tout tuple approprié dans E, le résultat appartient à E.

$\mathfrak{M}_{g,s}$ satisfait donc les conditions pour être une $\mathcal{L}$-structure valide, où $\mathcal{L} = (\mathcal{R}, \mathcal{F}, \mathcal{C})$ est le langage du premier ordre considéré. □

2.7.4 Vérité, interprétation

Définition 8 (Satisfaction d'une proposition atomique)**.** *Le modèle $\mathfrak{M}_{g,s}$ satisfait une proposition atomique $R(\vec{a})$ (avec R une relation et $\vec{a}$ une liste de constantes) si et seulement si :*

$$probe_{R(\vec{a})}(M_g(s)) = 1 \text{ et } R \text{ est bien définie dans } g$$

On notera ce fait $\mathfrak{M}_{g,s} \models R(\vec{a})$.

Théorème 2. *Si $\mathfrak{M}_{g,s} \models R(\vec{a})$ alors $\mathfrak{M}_{g,s} \not\models \neg R(\vec{a})$*

Démonstration. Supposons que $\mathfrak{M}_{g,s} \models R(\vec{a})$. Par la définition 8, $\text{probe}_{R(\vec{a})}(M_g(s)) = 1$. Donc $\langle w, x \rangle + b > 0$ par la définition 4. Comme R est bien définie par 8, $\text{acc}(R) = 1$, c'est-à-dire que $\mathbb{P}_{s,\vec{a}}[\text{probe}_{R(\vec{a})}(M_g(s)) = R_s^{\text{réel}}(\vec{a})] = 1$, donc $\text{probe}_{R(\vec{a})}(M_g(s)) = R_s^{\text{réel}}(\vec{a})$. Par conséquent, $\text{probe}_{\neg R(\vec{a})}(M_g(s)) = 0$, d'où $\mathfrak{M}_{g,s} \not\models \neg R(\vec{a})$. □

2.7.5 Conséquence naturelle

Définition 9 (Conséquence naturelle)**.** *On dira qu'une formule ϕ est **conséquence naturelle** d'une suite de symboles s par un LLM g, lorsque $\mathfrak{M}_{g,s} \models \phi$. On notera cela :*

$$s \models_g \phi$$

voire

$$s \models \phi$$

lorsqu'il n'y a pas d'ambiguïté sur le modèle.

Définition 10 (Conséquence artificielle)**.** *On dira qu'une suite de symboles s est **conséquence artificielle** d'une formule ϕ par un LLM g, lorsque $\mathfrak{M}_{g,s} \models \phi$. On notera cela :*

$$\phi \models_g s$$

voire

$$\phi \models s$$

lorsqu'il n'y a pas d'ambiguïté sur le modèle.

2.7.6 Proposition atomique en langue naturelle

Définition 11 (Proposition atomique)**.** *Une formule est dite atomique si elle ne contient pas de connecteurs logiques.*

Définition 12 (Proposition atomique en langue naturelle)**.** *Une suite de symboles s sera dite **atomique** par g si ϕ est atomique, et $\phi \models_g s$*

Exemple 2. *Soit g un LLM fixé. Soient a = « La bibliothèque conserve des œuvres d'art » et b = « La bibliothèque conserve des livres anciens » deux suites de symboles.*

On note par $C(x, y)$ la relation « x conserve y », b la bibliothèque, o les œuvres d'art et l les livres anciens.

On aura $\mathfrak{M}_a \models C(b, o)$ et $\mathfrak{M}_b \models C(b, l)$. Donc $a \models C(b, o)$ et $b \models C(b, l)$

La suite c = « La bibliothèque conserve des œuvres d'art et des livres anciens » donne : $c \models C(b, o) \wedge C(b, l)$

Définition 13 (Équivalence naturelle)**.** *On dira que deux suites de symboles t et t' sont **naturellement équivalentes** par g s'il existe une formule ϕ telle que $\phi \models_g t$ et $\phi \models_g t'$.*

2.7.7 Postulats fondamentaux

Deux postulats centraux justifient l'approche AtomicEval :

Postulat 1. *Si $t \equiv t'$ par g, g pourra générer une suite de symboles le confirmant.*

Postulat 2 (Hypothèse de représentation platonique)**.** *(Huh et al. 2024) Les LLMs développent des représentations universelles de concepts.*

Avec les postulats 1 et 2, nous pouvons justifier la manière d'évaluer l'autonomie des propositions atomiques.

2.7.8 Métriques

Autonomie Pour évaluer la reformulation des propositions complexes en propositions atomiques, nous pouvons donner la proposition atomique à un LLM, lui demander le contexte, puis, lui donner le véritable contexte, et demander si les contextes correspondent. Si ce n'est pas le cas, il faut recommencer, sinon la clause d'autonomie est bien respectée.

Soit t la proposition testée par le modèle g. Si t a du sens, on aura $t \models \phi$. Pour vérifier son autonomie, on demande à un LLM d'expliciter t (par le postulat 1 cela revient à développer ϕ). Cela donne t' une autre suite qui est l'interprétation de t par g.

Ensuite, on donne le contexte complet c à partir duquel la proposition atomique a été générée (si ce contexte a du sens, on aura $c \models \psi$).

Deux cas :

— Si $\phi \equiv \psi$, par le postulat 1, le LLM sera capable de le dire, et donc on pourra conclure que t satisfait la clause d'autonomie par g.
— Si $\phi \not\equiv \psi$, par le postulat 1, le LLM sera capable de le dire, et donc on pourra conclure que t ne satisfait pas la clause d'autonomie par g.

La métrique d'autonomie prend donc un contexte et une proposition candidate, et renvoie 1 si elle est autonome, et 0 sinon.

Atomicité Pour évaluer l'atomicité, la tâche est plus complexe. Il faudrait développer des méthodes pour vérifier qu'une proposition ne peut être décomposée davantage en éléments plus simples tout en préservant son sens.

2.8 SEval-Ex : extension vers l'évaluation des résumés

Inspiration directe.

L'approche par propositions atomiques trouve une application directe dans l'évaluation de la qualité des résumés de texte. Herserant et Guigue 2025 présente SEval-Ex, un framework qui décompose l'évaluation des résumés en phrases atomiques, permettant à la fois une haute performance et une explicabilité des décisions d'évaluation.

2.8.1 Principe méthodologique

SEval-Ex emploie un pipeline en deux étapes complémentaire à AtomicEval :

1. **Extraction des phrases atomiques** : Un LLM décompose le texte source et le résumé en phrases atomiques autonomes
2. **Mise en correspondance** : Alignement et classification de ces phrases en Vrais Positifs (TP), Faux Positifs (FP) et Faux Négatifs (FN)

Cette approche contraste avec les méthodes existantes qui ne fournissent que des scores globaux, en générant un parcours détaillé des décisions grâce à un alignement au niveau des phrases.

2.8.2 Performances empiriques

Efficacité de SummEval :
— Corrélation de 0.580 avec les jugements humains sur la cohérence
— Surpasse GPT-4 (0.521) tout en maintenant l'interprétabilité
— Robustesse contre différents types d'hallucinations (entités, événements, détails)

2.8.3 Convergence méthodologique

SEval-Ex confirme la pertinence de l'approche atomique dans l'évaluation de textes générés. La décomposition en unités sémantiques élémentaires permet à la fois une précision fine dans l'évaluation factuelle, une explicabilité complète des décisions d'évaluation et une généralisation à différents domaines et types de textes

Cette convergence entre AtomicEval et SEval-Ex illustre la puissance unificatrice du concept de proposition atomique dans l'évaluation de systèmes de TAL.

2.9 Validation expérimentale des postulats

2.9.1 Protocole expérimental

Concernant le premier postulat, nous pouvons mettre en place plusieurs protocoles expérimentaux pour vérifier que la réponse du LLM est bien en accord avec le probing.

AtomicEval s'avère particulièrement utile pour évaluer l'autonomie d'une phrase ou d'un paragraphe. Pour effectuer des tests, l'interface en ligne de commande permet une utilisation directe :

```
python autonomeval.py --model llama3.2:1b \
  --proposition "Il a pris la décision hier." \
  --context "Jean, le directeur de TechCorp, a pris la décision
  d'embaucher 50 employés hier."
```

2.9.2 Application à l'évaluation de résumés

L'idée centrale pour l'évaluation de résumés consiste à faire développer le résumé par le LLM, puis à comparer le contexte original avec le résumé étendu. Si le résumé a été bien compris, il ne devrait pas y avoir de différence significative entre les deux contextes.

Pour utiliser AtomicEval dans ce cadre, on peut placer le résumé dans le paramètre `proposition` et le texte original dans le paramètre `context`. Cependant, quelques ajustements de prompt seraient nécessaires pour optimiser le fonctionnement spécifiquement pour l'évaluation de résumés plutôt que pour des propositions atomiques isolées.

2.9.3 Approche synthétique vs analytique

L'originalité d'AtomicEval réside dans son approche inverse du problème : au lieu de demander de découper en parties atomiques puis de vérifier une à une les propositions (comme dans Herserant et Guigue 2025), nous passons par une approche holistique en étendant au contraire la proposition ou le résumé.

En quelque sorte, là où on pourrait être analytique (découpage en propositions atomiques), la méthode AtomicEval est synthétique (construction d'un contexte plus large).

2.10 Contribution expérimentale : AtomicEval et le propositionneur français

Cette section présente les résultats expérimentaux obtenus durant le stage, comprenant l'implémentation d'AtomicEval sur FACTScore et le développement d'un propositionneur français basé sur FLAN-T5.

2.10.1 Framework AtomicEval

Méthodologie

AtomicEval fonctionne selon un processus en deux étapes : (1) **Développement du contexte** : Étant donnée une proposition p, un LLM génère le contexte minimal c_{gen} nécessaire à sa compréhension. (2) **Évaluation de l'autonomie** : Comparaison entre c_{gen} et le contexte original c_{orig}.

La logique d'évaluation est la suivante : si $c_{gen} \approx c_{orig}$ alors la proposition est AUTONOME, si c_{gen} ne correspond pas alors la proposition est NON-AUTONOME.

Cadre théorique

Nous ancrons cette approche dans la théorie de la représentation. Étant donnés un LLM g et un texte s, les représentations $M_g(s) \in \mathbb{R}^d$ induisent des modèles formels $\mathfrak{M}_{g,s}$. Une relation R est linéairement définissable lorsque la sonde (un classificateur linéaire) atteint une précision parfaite :

$$\text{probe}_{R(\hat{a})} = \mathbf{1}[\langle w, x \rangle + b > 0]$$

Où $\hat{a}$ est un tuple, w les poids du classificateur linéaire, x les entrées, et b le biais.

Cette nouvelle approche s'appuie sur la théorie des modèles Lascar 2009, qui fournit le formalisme et la définition précise de ce qui est « sémantique ». L'autonomie des propositions devient évaluable par la cohérence : $s \models_g \phi$ si $\mathfrak{M}_{g,s} \models \phi$. Deux propositions sont équivalentes lorsqu'elles satisfont la même formule atomique.

2.10.2 Propositionneur français

Nous avons affiné FLAN-T5-large sur 42 000 exemples français suivant les critères de Chen et al. : atomiques (un fait), auto-contenus et minimaux (8-15 mots)Chen et al. 2024. Le modèle a été entraîné sur deux GPU A6000 pendant 3 époques. Il [1] génère des propositions d'en moyenne 10,8 mots avec une gestion syntaxique française appropriée.

Dans l'exemple suivant, « Le chat et le chien sont dans la cuisine » devient « Le chat est dans la cuisine. Le chien est dans la cuisine. » [2] où chaque entité est incluse dans une phrase séparée.

2.10.3 Résultats expérimentaux

Nous avons appliqué AtomicEval sur FACTScore ainsi qu'avec le propositionneur français sur une tâche de résumé.

AtomicEval sur FACTScore

L'évaluation révèle des problèmes d'autonomie significatifs : 52,14% autonomes, 47,86% non-autonomes. Les cas non-autonomes incluent des pronoms non résolus (« Il a obtenu son diplôme en 1995 »), des références temporelles (« La décision a été prise hier ») et des déictiques spatiaux (« La réunion a eu lieu là-bas »).

Évaluation de résumés

Nous utilisons la méthode de Herserant et Guigue 2025 pour évaluer notre modèle affiné, le propositionneur français. Les résultats présentés dans le tableau 2.1 lors du remplacement de la phase de segmentation en propositions atomiques (effectuée par Qwen 2.5 dans l'article original) par le propositionneur français, montrent que le propositionneur n'est pas aussi performant que Qwen.

Cependant, le propositionneur est 75,1% plus rapide, ce qui en fait une option intéressante à considérer. Ce résultat est particulièrement intéressant car le propositionneur français n'a pas été entraîné pour effectuer l'évaluation de résumés, mais pour créer des propositions atomiques. Ce résultat montre une différence de performance modeste, mais le gain en vitesse est considérable.

1. Le modèle est disponible à : https ://huggingface.co/Zual/propositioneur
2. « Le chien et le chat sont dans la cuisine » devient « Le chien est dans la cuisine » et « Le chat est dans la cuisine ».

TABLE 2.1 – Comparaison des métriques d'évaluation

Métrique	Original	Propositionneur	Différence
Précision	0,8378	0,7502	-0,0875
Rappel	0,8420	0,7731	-0,0689
F1	0,8307	0,7388	-0,0919

2.10.4 Conclusion et travaux futurs

Nous pensons qu'AtomicEval fait le pont entre la logique formelle et le TAL pratique. Le framework établit que l'autonomie des propositions est mesurable par la cohérence contextuelle plutôt que par l'analyse syntaxique. Cela relie les faits atomiques wittgensteiniens (WITTGENSTEIN 1972) à l'apprentissage de représentations moderne : les propositions correspondent aux formules atomiques lorsqu'elles induisent des interprétations de modèles cohérentes.

Notre hypothèse est que si les propositions $t \equiv t'$ sous le modèle g, alors g peut générer des preuves confirmatives. Cela permet une vérification systématique de l'autonomie par introspection du LLM plutôt que par annotation externe.

AtomicEval améliore les systèmes d'évaluation factuelle en filtrant les propositions non-autonomes qui confondent les algorithmes de vérification. La recherche dense (CHEN et al. 2024) bénéficie de l'élimination des références dépendantes du contexte qui dégradent les calculs de similarité. Le propositionneur français permet des applications de propositions atomiques françaises précédemment limitées à l'anglais.

Les systèmes d'évaluation de résumés comme SEval-Ex (HERSERANT et GUIGUE 2025) gagnent en fiabilité grâce aux décompositions vérifiées pour l'autonomie. Les systèmes RAG atteignent de meilleures performances avec des unités de récupération proprement atomiques qui maintiennent la cohérence sémantique (CHEN et al. 2024).

Les premières expériences montrent que le propositionneur français traite avec succès les structures syntaxiques complexes tout en préservant le contenu sémantique. Les propositions générées maintiennent la plage requise de 8-15 mots et démontrent une décomposition factuelle appropriée à travers divers types de textes.

De plus, l'atomicité appropriée réduit les exigences de stockage tout en améliorant la granularité sémantique. Tandis que les études (CHEN et al. 2024) ont montré une amélioration dans les expériences RAG qui montrent +10,1% (CHEN et al. 2024) d'amélioration du Rappel@20 par rapport à la récupération basée sur les passages lors de l'utilisation de propositions vérifiées pour l'autonomie, nous évaluons notre méthode sur la tâche de résumé.

Chapitre 3

Conclusion et perspectives

3.1 Synthese des contributions

Ce travail de recherche a aborde deux problematiques fondamentales dans l'optimisation des systemes RAG : les limites du greffage contextuel des LLM et l'evaluation de l'atomicite reelle des propositions.

3.1.1 Contribution theorique principale

Notre analyse comparative de neuf etudes recentes revele que la limite du greffage des LLM n'est pas quantitative mais qualitative. Nous proposons une architecture unifiee a cinq niveaux (granularite logique, structurelle, organisationnelle, hierarchique et temporelle) qui optimise simultanement la pertinence informationnelle et les contraintes computationnelles.

3.1.2 Contribution methodologique

AtomicEval constitue le premier framework d'evaluation systematique de l'autonomie des propositions atomiques. Les resultats sur FACTScore (52,14% de propositions reellement autonomes) remettent en question l'efficacite des methodes actuelles et etablissent de nouveaux standards d'evaluation.

3.1.3 Contribution technique

Le developpement d'un propositionneur francais base sur FLAN-T5 etend les capacites de traitement atomique au-dela de l'anglais, avec des performances comparables (75,1% plus rapide) aux solutions existantes.

3.2 Perspectives de recherche

3.2.1 Extensions theoriques

L'architecture a cinq niveaux proposee necessite une validation experimentale sur des corpus diversifies. Le developpement de metriques holistiques integrant precision, efficacite computationnelle et coherence semantique constitue une priorite.

3.2.2 Applications pratiques

L'implementation complete du pipeline propose pour le systeme RAG de la Bibliotheque Paris-Saclay offrirait un banc d'essai reel pour valider les approches developpees. L'extension d'AtomicEval a d'autres langues et domaines permettrait une generalisation des resultats.

3.2.3 Recherches futures

Trois axes de recherche emergent : l'integration des signaux multimodaux dans l'evaluation de l'atomicite, le developpement d'architectures RAG auto-adaptatives utilisant les principes identifies, et l'extension de la theorie des modeles aux representations neuronales pour une formalisation complete du lien entre logique et TAL.

Bibliographie

Adjali, Omar et al. (2024a). « Multi-Level Information Retrieval Augmented Generation for Knowledge-based Visual Question Answering ». In : *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*. Sous la dir. d'Yaser Al-Onaizan, Mohit Bansal et Yun-Nung Chen. Association for Computational Linguistics, p. 16499-16513. doi : 10.18653/V1/2024.EMNLP-MAIN.922. url : https://doi.org/10.18653/v1/2024.emnlp-main.922.

— (2024b). « Multi-Level Information Retrieval Augmented Generation for Knowledge-based Visual Question Answering ». In : *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*. Sous la dir. d'Yaser Al-Onaizan, Mohit Bansal et Yun-Nung Chen. Association for Computational Linguistics, p. 16499-16513. doi : 10.18653/V1/2024.EMNLP-MAIN.922. url : https://doi.org/10.18653/v1/2024.emnlp-main.922.

Anantha, Raviteja et Danil Vodianik (mars 2024). « Context Tuning for Retrieval Augmented Generation ». In : *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*. Sous la dir. de Raúl Vázquez et al. St Julians, Malta : Association for Computational Linguistics, p. 15-22. url : https://aclanthology.org/2024.uncertainlp-1.2/.

Chen, Tong et al. (nov. 2024). « Dense X Retrieval : What Retrieval Granularity Should We Use ? » In : *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Sous la dir. d'Yaser Al-Onaizan, Mohit Bansal et Yun-Nung Chen. Miami, Florida, USA : Association for Computational Linguistics, p. 15159-15177. doi : 10.18653/v1/2024.emnlp-main.845. url : https://aclanthology.org/2024.emnlp-main.845/.

Gao, Yunfan et al. (2023). « Retrieval-Augmented Generation for Large Language Models : A Survey ». In : *CoRR* abs/2312.10997. doi : 10.48550/ARXIV.2312.10997. arXiv : 2312.10997. url : https://doi.org/10.48550/arXiv.2312.10997.

Herserant, Tanguy et Vincent Guigue (2025). « SEval-Ex : A Statement-Level Framework for Explainable Summarization Evaluation ». In : *CoRR* abs/2505.02235. doi : 10.48550/ARXIV.2505.02235. arXiv : 2505.02235. url : https://doi.org/10.48550/arXiv.2505.02235.

Hewitt, John et Christopher D. Manning (juin 2019). « A Structural Probe for Finding Syntax in Word Representations ». In : *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota : Association for Computational Linguistics, p. 4129-4138. doi : 10.18653/v1/N19-1419. url : https://aclanthology.org/N19-1419.

Huh, Minyoung et al. (2024). « Position : The Platonic Representation Hypothesis ». In : *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. url : https://openreview.net/forum?id=BH8TYy0r6u.

Jiang, Zhengbao et al. (2023). « Active Retrieval Augmented Generation ». In : *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*. Sous la dir. d'Houda Bouamor, Juan Pino et Kalika Bali. Association for Computational Linguistics, p. 7969-7992. doi : 10.18653/V1/2023.EMNLP-MAIN.495. url : https://doi.org/10.18653/v1/2023.emnlp-main.495.

Jimeno-Yepes, Antonio et al. (2024a). « Financial Report Chunking for Effective Retrieval Augmented Generation ». In : *CoRR* abs/2402.05131. DOI : 10.48550/ARXIV.2402.05131. arXiv : 2402.05131. URL : https://doi.org/10.48550/arXiv.2402.05131.

— (2024b). « Financial Report Chunking for Effective Retrieval Augmented Generation ». In : *CoRR* abs/2402.05131. DOI : 10.48550/ARXIV.2402.05131. arXiv : 2402.05131. URL : https://doi.org/10.48550/arXiv.2402.05131.

Karpukhin, Vladimir et al. (2020). « Dense Passage Retrieval for Open-Domain Question Answering ». In : *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Sous la dir. de Bonnie Webber et al. Association for Computational Linguistics, p. 6769-6781. DOI : 10.18653/V1/2020.EMNLP-MAIN.550. URL : https://doi.org/10.18653/v1/2020.emnlp-main.550.

Lascar, Daniel (2009). *La théorie des modèles en peu de maux*. Nouvelle bibliothèque mathématique. Paris : Cassini, p. 352. ISBN : 978-2-84225-137-6.

Lewis, Patrick et al. (2020). « Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks ». In : *Advances in Neural Information Processing Systems 33 : Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Sous la dir. d'Hugo Larochelle et al. URL : https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Li, Zhuowan et al. (2024). « Retrieval Augmented Generation or Long-Context LLMs ? A Comprehensive Study and Hybrid Approach ». In : *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing : EMNLP 2024 - Industry Track, Miami, Florida, USA, November 12-16, 2024*. Sous la dir. de Franck Dernoncourt, Daniel Preotiuc-Pietro et Anastasia Shimorina. Association for Computational Linguistics, p. 881-893. DOI : 10.18653/V1/2024.EMNLP-INDUSTRY.66. URL : https://doi.org/10.18653/v1/2024.emnlp-industry.66.

Manning, Christopher D. et al. (2014). « The Stanford CoreNLP Natural Language Processing Toolkit ». In : *Association for Computational Linguistics (ACL) System Demonstrations*, p. 55-60. URL : http://www.aclweb.org/anthology/P/P14/P14-5010.

Min, Sewon et al. (2023). « FActScore : Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation ». In : *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, p. 12076-12100.

Shao, Zhihong et al. (2023). « Enhancing Retrieval-Augmented Large Language Models with Iterative Retrieval-Generation Synergy ». In : *Findings of the Association for Computational Linguistics : EMNLP 2023, Singapore, December 6-10, 2023*. Sous la dir. d'Houda Bouamor, Juan Pino et Kalika Bali. Association for Computational Linguistics, p. 9248-9274. DOI : 10.18653/V1/2023.FINDINGS-EMNLP.620. URL : https://doi.org/10.18653/v1/2023.findings-emnlp.620.

Stacey, Joe et al. (2024). « Atomic Inference for NLI with Generated Facts as Atoms ». In : *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 10188-10204.

Vaswani, Ashish et al. (2017a). « Attention is All you Need ». In : *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Sous la dir. d'Isabelle Guyon et al., p. 5998-6008. URL : https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

— (2017b). « Attention is All you Need ». In : *Advances in Neural Information Processing Systems 30 : Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, p. 5998-6008. URL : https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, Zheng et al. (2024). « M-RAG : Reinforcing Large Language Model Performance through Retrieval-Augmented Generation with Multiple Partitions ». In : *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*. Sous la dir.

de Lun-Wei Ku, Andre Martins et Vivek Srikumar. Association for Computational Linguistics, p. 1966-1978. doi : 10.18653/V1/2024.ACL-LONG.108. url : https://doi.org/10.18653/v1/2024.acl-long.108.

Wittgenstein, Ludwig (1922). *Tractatus Logico-Philosophicus*. London : Kegan Paul, Trench, Trubner & Co.

— (1972). *Tractatus logico-philosophicus*. Trad. de l'allemand par Gilles-Gaston Granger. Tel 311. Introduction de Bertrand Russell. Préambule et notes de Gilles-Gaston Granger. Paris : Gallimard.

Xiong, Lee et al. (2021). « Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval ». In : *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. url : https://openreview.net/forum?id=zeFrfgyZln.