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

Soutenance de stage - master 2 LOGOS (U. Paris-Cité)

Luc Pommeret

Laboratoire Interdisciplinaire des Sciences du Numérique
Université Paris-Saclay

1er avril - 1er octobre 2025

Plan de la présentation

Contexte et Problématique

État de l'art : Granularité en RAG

Propositions Atomiques

AtomicEval : Framework d'évaluation

Architecture proposée

Contributions et perspectives

Contexte et Problématique

Environnement de recherche

LISN - Laboratoire Interdisciplinaire des Sciences du Numérique

- Fusion du Limsi et du LRI
- 5 départements de recherche
- Stage dans le département *Sciences et technologies des langues*

Encadrement :

- Sahar Ghannay (Maîtresse de conférences)
- Christophe Servan (Chercheur)
- Sophie Rosset (Directrice de recherche)

Problématique centrale

Question de recherche

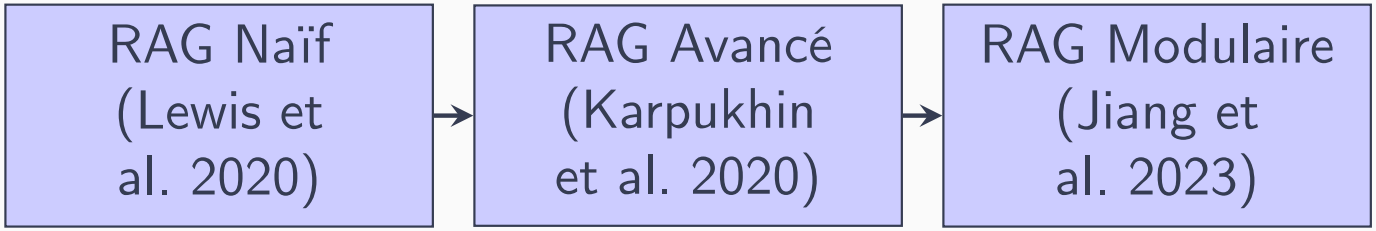
« Jusqu'où peut-on goinfrer le LLM pour la génération de réponses contraintes dans le cadre d'un chatbot ? »

Deux dimensions d'analyse :

1. **Quantitative :** Volume d'informations contextuelles
   - Nombre de documents récupérés
   - Longueur totale des passages (tokens)
   - Taille du contexte global
2. **Qualitative :** Granularité et pertinence
   - Qualité sémantique des passages
   - Qualité du processus de récupération
   - Qualité rédactionnelle des sources

État de l'art : Granularité en RAG

Évolution des paradigmes RAG



Récupération → Génération

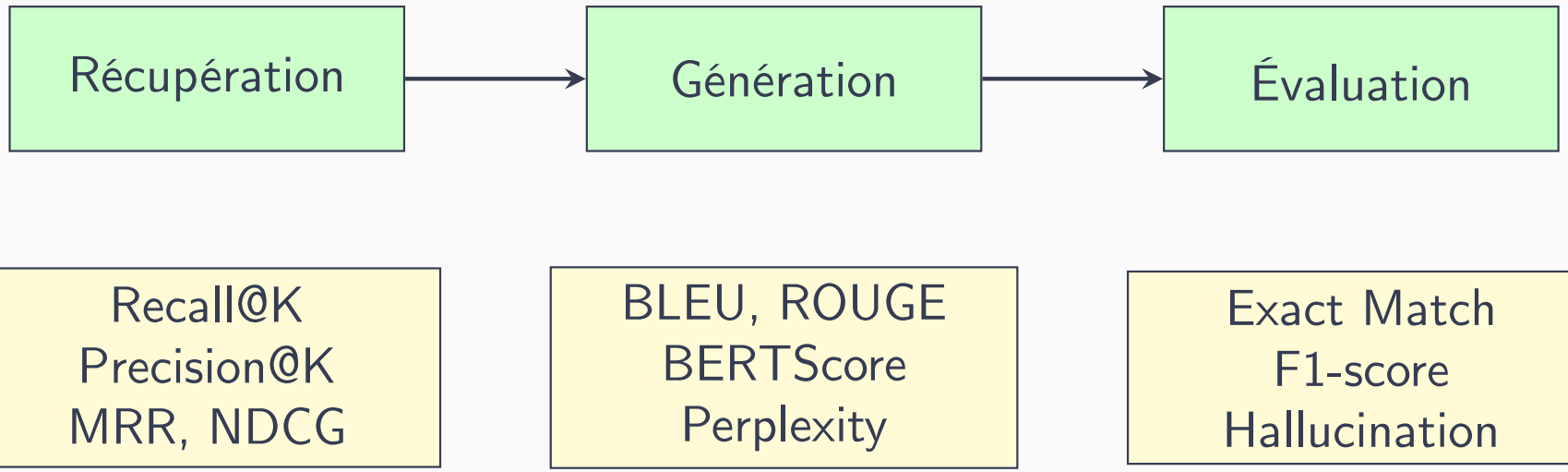
Optimisation pré/post récupération

Adaptation dynamique

Observation clé

Évolution vers une gestion **qualitative** du contexte plutôt que quantitative

Métriques d'évaluation en RAG



Problème

Hétérogénéité des métriques $\Rightarrow$ Comparaisons difficiles

Approches de granularité - Résultats

Approche	Amélioration	Optimum	Efficacité
Dense X (Chen et al.)	+10.1% Recall@20	11.2 mots/prop	
Financial Chunking	+11%	Structure adapt.	-44% chunks
M-RAG (Wang et al.)	+8-12%	4 partitions	
ITER-RETGEN	+8.6% abs	2 itérations	
SELF-ROUTE	95% perf.	Routage adapt.	-65% coût

Tendances convergentes

- Améliorations dans la fourchette **8-12%**
- Rendements décroissants au-delà de seuils spécifiques
- Primauté de la **qualité** sur la quantité

Propositions Atomiques

Fondements historiques

Logique classique

- Frege (1879) : Begriffschrift
- Russell : Principia Mathematica
- Wittgenstein : Tractatus Logico-Philosophicus

> *« Le monde est la totalité des faits, non des choses. »*
> — *Wittgenstein, Tractatus 1.1*

TAL moderne (Chen et al. 2024)

> *« 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. »*

Caractéristiques :

- Absence de connecteurs logiques
- Granularité informationnelle la plus petite qui reste autonome

Exemple

« Le chat et le chien sont dans la cuisine. » devient deux propositions atomiques :

- « Le chat est dans la cuisine. »
- « Le chien est dans la cuisine. »

Le problème de l'atomicité réelle

Constat

Aucun travail n'évalue si les propositions sont **réellement** atomiques

Exemple problématique (FACTScore) :

- Proposition : « *He appeared as an actor.* »
- Marquée comme « supported » par FACTScore
- Mais : Qui est « He » ? $\Rightarrow$ **Non-autonome**

Contexte complet :

> « *Joey D. Vieira is an American actor [...] He also appeared as an actor in films such as "The Wild One" and "Viva Las Vegas".* »

Besoin

Framework d'évaluation systématique de l'autonomie

AtomicEval : Framework d'évaluation

AtomicEval

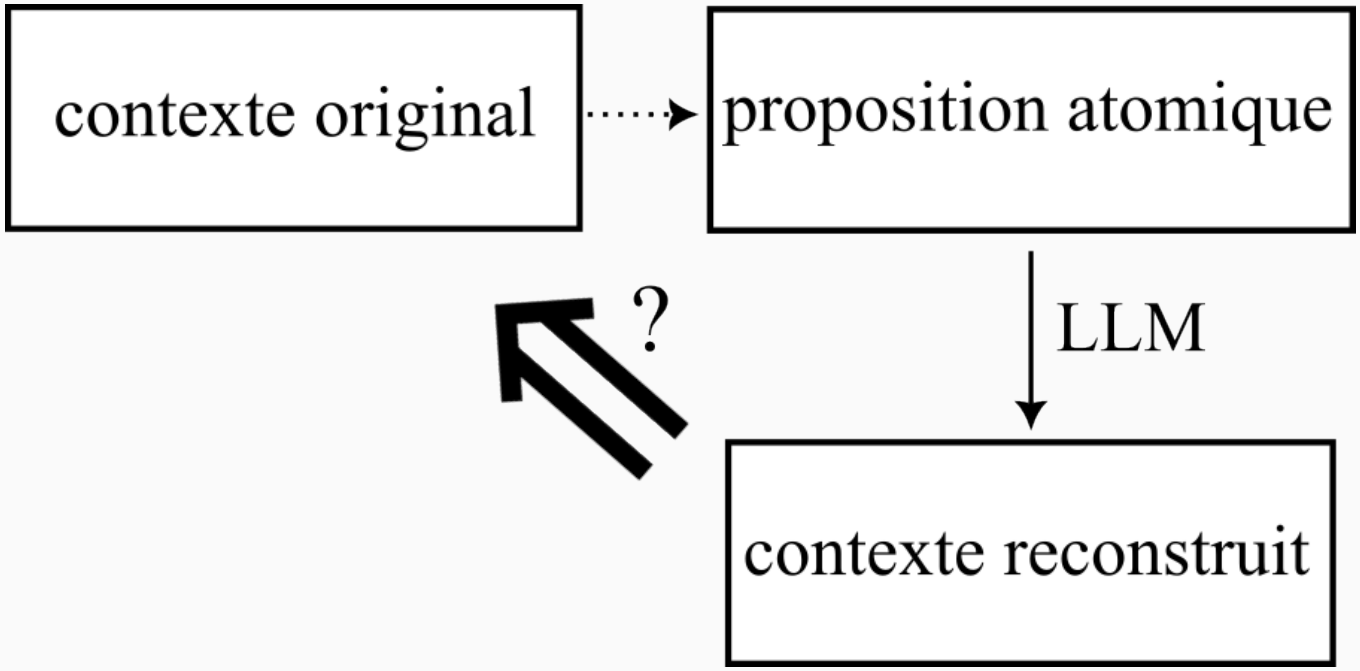


Figure 1: Le *framework* AtomicEval. Le but est de savoir si le contexte reconstruit implique le contexte original.

Logique d'évaluation :

- Si contexte généré $\approx$ contexte réel $\Rightarrow$ **AUTONOME**
- Si contexte généré minimal/inexact $\Rightarrow$ **NON-AUTONOME**

Cadre théorique - Représentations vectorielles

Définition (Représentation vectorielle)

Soit g un LLM et s une suite de symboles en langue naturelle. s et la couche c induisent une représentation vectorielle :

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

Définition (Représentation matricielle)

Lorsqu'on considère toutes les couches à la fois d'un transformeur à n couches, on obtient la matrice :

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

Intuition : Les représentations capturent la sémantique des suites de symboles dans l'espace vectoriel du modèle.

Sondes et définissabilité

Définition (Sonde pour propositions d'ordre 0)

Pour une relation R d'arité k sur des **propositions d'ordre 0** et une représentation $M_g(s)$, une *sonde* est une famille de fonctions :

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

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

Définition (Sonde linéaire)

Une *sonde linéaire* pour une relation R et des constantes $\vec{a}$ est :

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

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

Point clé : Seules les propositions d'ordre 0 (atomiques) sont sondées. Le reste est structurel.

Définissabilité et précision

Définition (Vérité terrain)

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

Définition (Précision d'une sonde linéaire)

La précision d'une sonde pour une relation R d'arité k et une liste de constantes $\vec{a}$ est :

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

où la probabilité est prise sur toutes les phrases s.

Définition (Définissabilité proposition d'ordre 0)

Une proposition $R(\vec{a})$ est *linéairement définissable* dans les représentations de g si et seulement si :

$$\text{acc}(R(\vec{a})) = 1$$

pour tout s.

Modèles induits par les LLMs

Théorème (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 :

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

Conséquence : Les représentations neuronales induisent des structures formelles interprétables.

Satisfaction et conséquence

Définition (Satisfaction d'une proposition d'ordre 0)

Le modèle $\mathfrak{M}_{g,s}$ satisfait une proposition d'ordre 0 $R(\vec{a})$ si et seulement si :

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

On note : $\mathfrak{M}_{g,s} \models R(\vec{a})$

Définition (Conséquence naturelle)

Une formule ϕ est **conséquence naturelle** d'une suite de symboles s par un LLM g, lorsque $\mathfrak{M}_{g,s} \models \phi$. On note :

$$s \models_g \phi$$

Postulats fondamentaux :

1. Si $t \equiv t'$ par g, alors g peut générer une confirmation
2. Hypothèse de représentation platonique (Huh et al. 2024)

Métriques formelles d'évaluation

Autonomie : Pour évaluer l'autonomie des propositions atomiques :

Soit t la proposition testée par le modèle g. Si t a du sens, on aura $t \models \phi$.

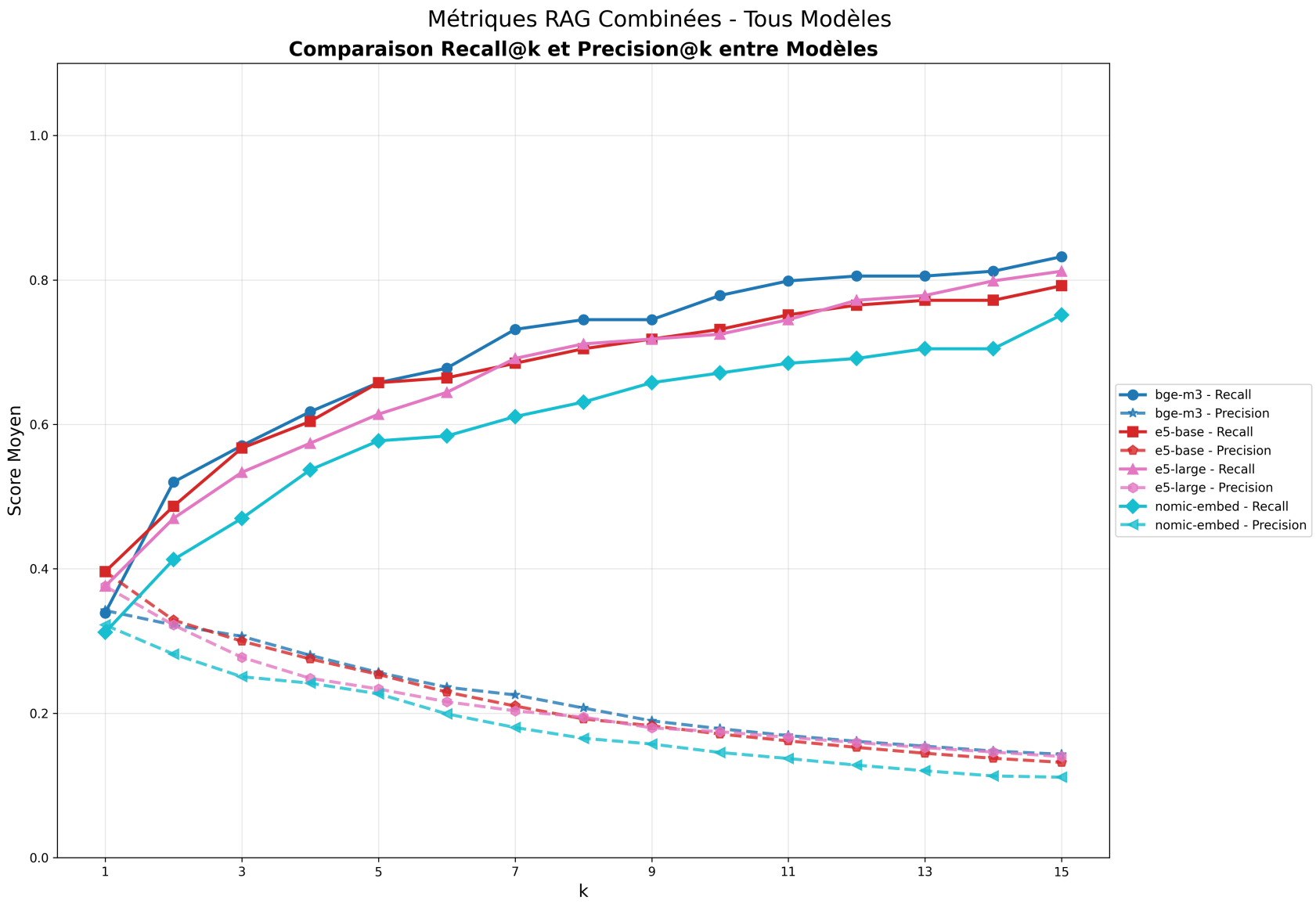
Procédure :

1. Demander à un LLM d'expliciter $t \rightarrow$ obtenir t' (interprétation de t par g)
2. Donner le contexte complet $c \rightarrow$ si c a du sens, alors $c \models \psi$
3. Évaluer l'équivalence :
   - Si $\phi \equiv \psi \rightarrow t$ satisfait la clause d'autonomie
   - Si $\phi \not\equiv \psi \rightarrow t$ ne satisfait pas la clause d'autonomie

Atomicité : Vérification qu'une proposition ne peut être décomposée davantage sans perdre son sens.

Résultats expérimentaux

Choix du modèle d'embedding : bge-m3



Résultats expérimentaux

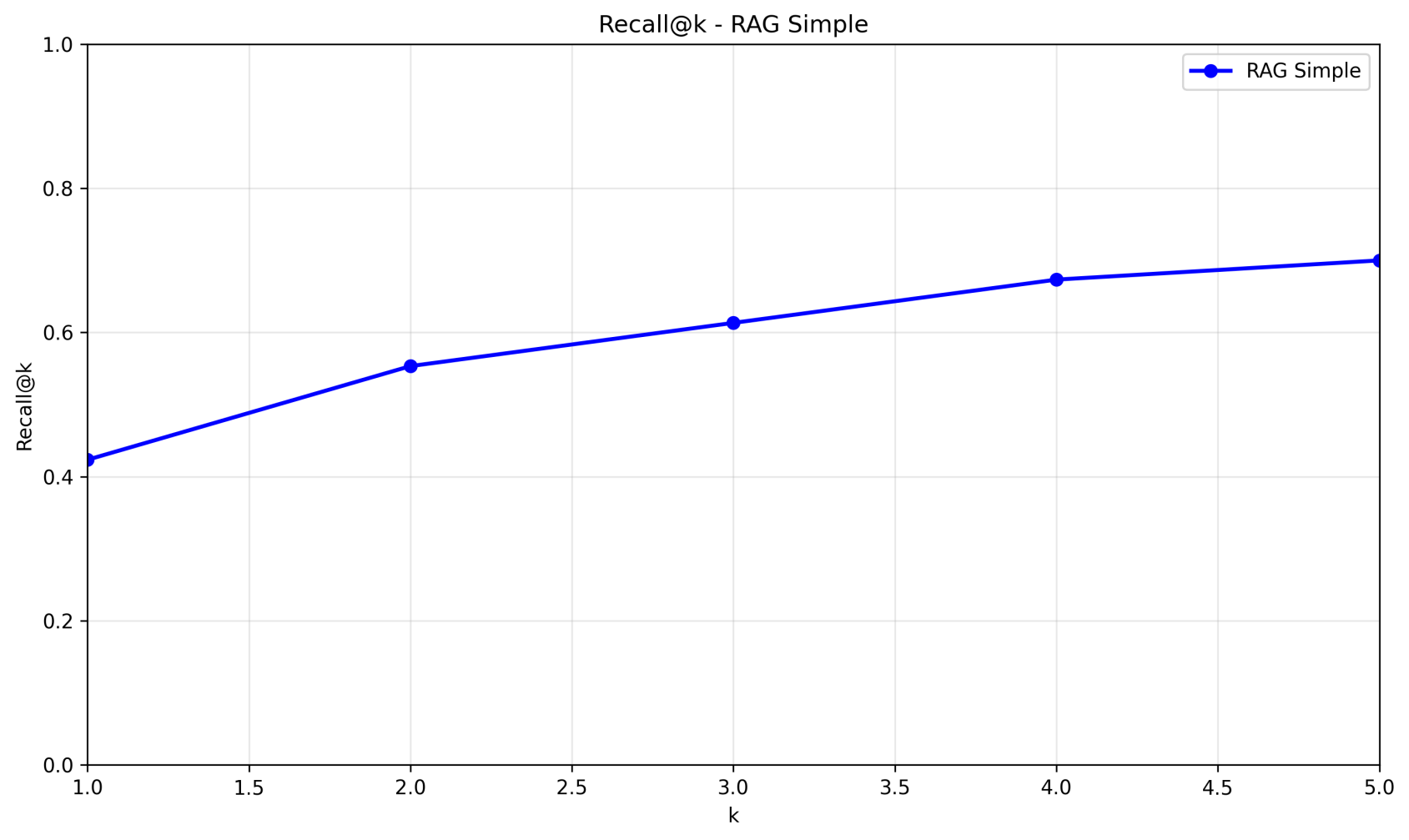


Figure 3: Capacité de *retrieval* d'un RAG simple. Ici le système `rag-dibiso` sur 149 questions.

Résultats expérimentaux

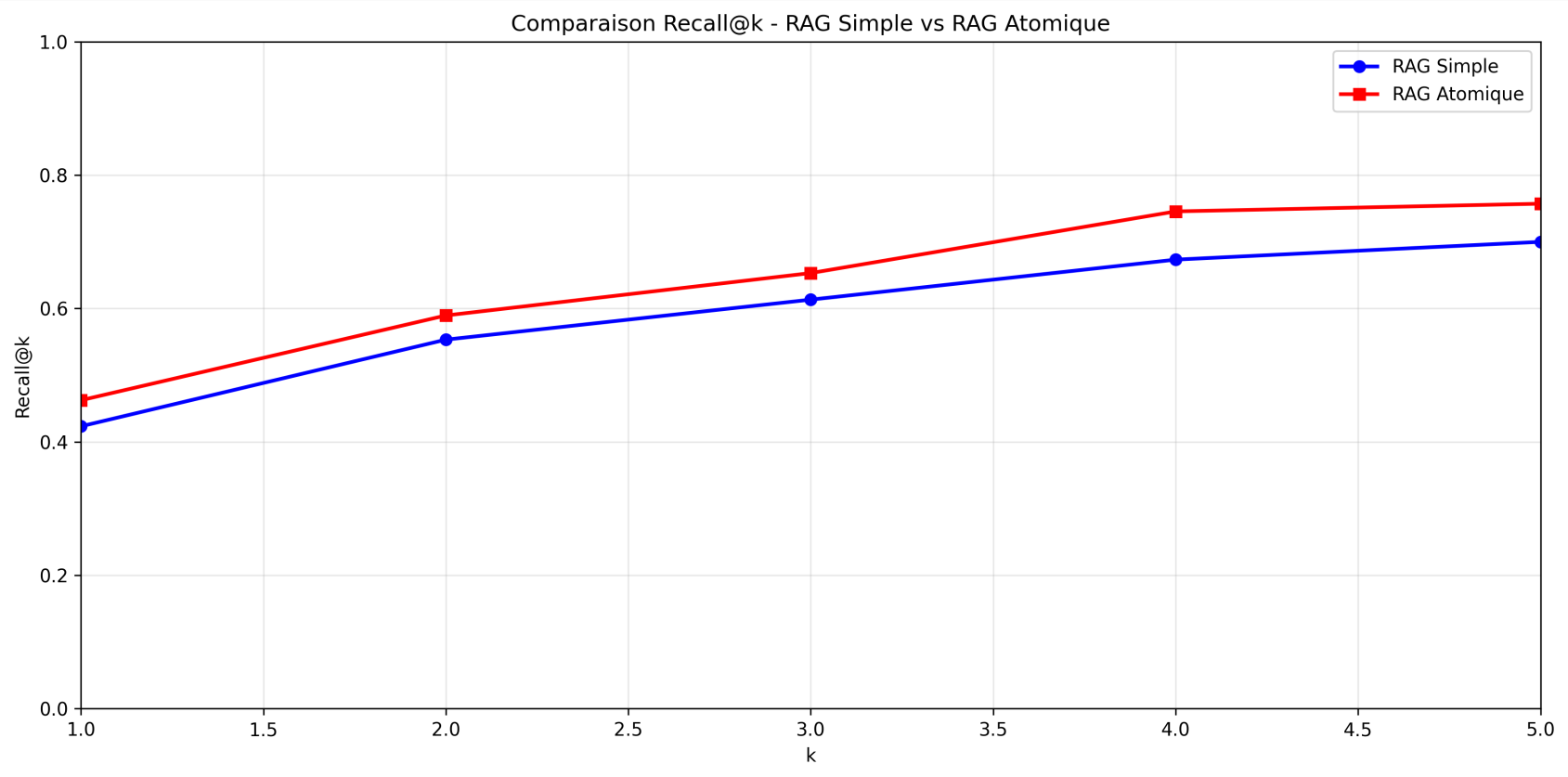


Figure 4: Capacité de *retrieval* d'un RAG nourri de propositions atomiques. Ici le système `rag-dibiso` sur 149 questions.

Résultats expérimentaux

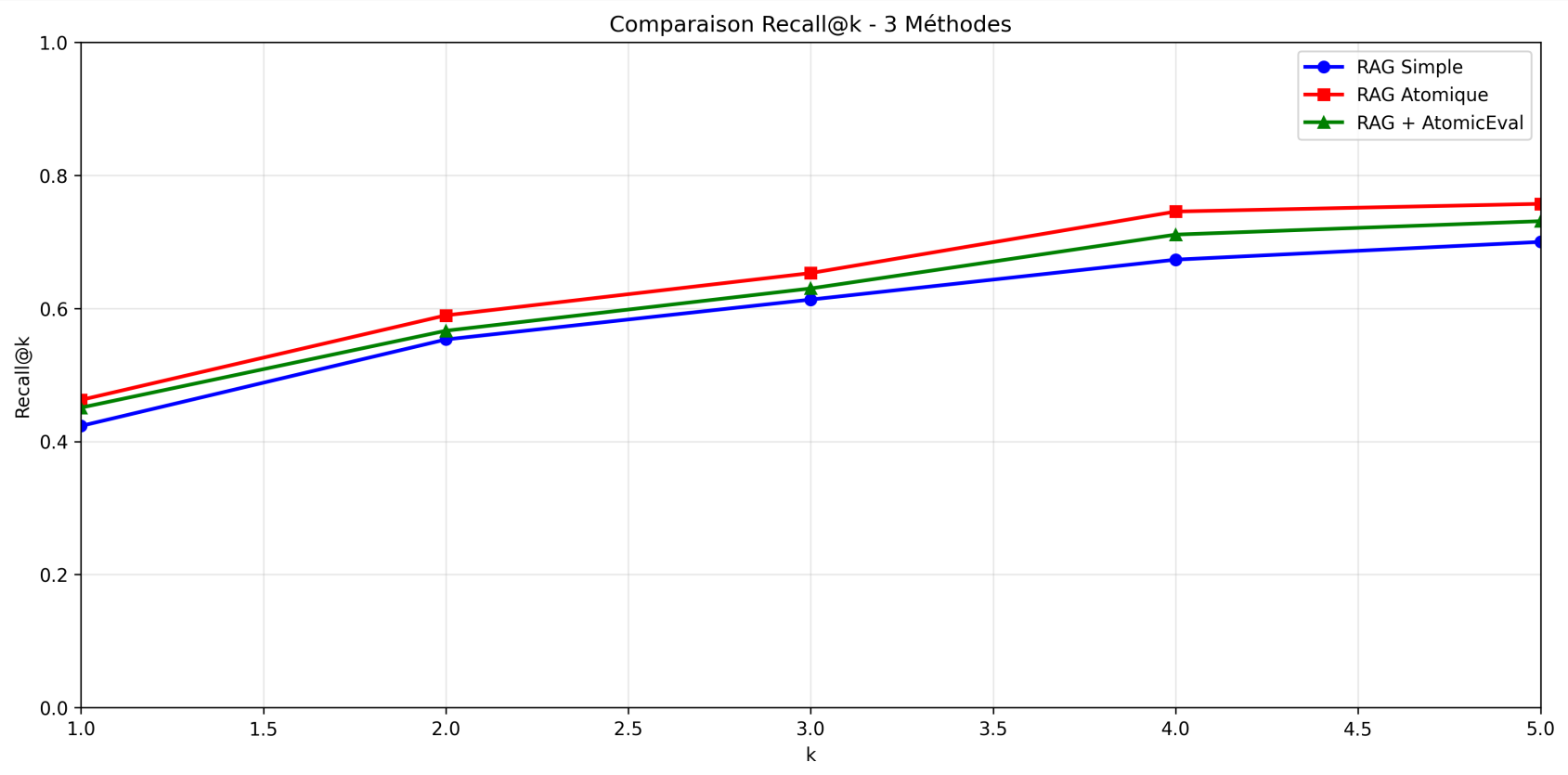


Figure 5: Capacité de *retrieval* d'un RAG nourri de propositions atomiques et filtré par AtomicEval. Ici le système `rag-dibiso` sur 149 questions.

Résultats expérimentaux

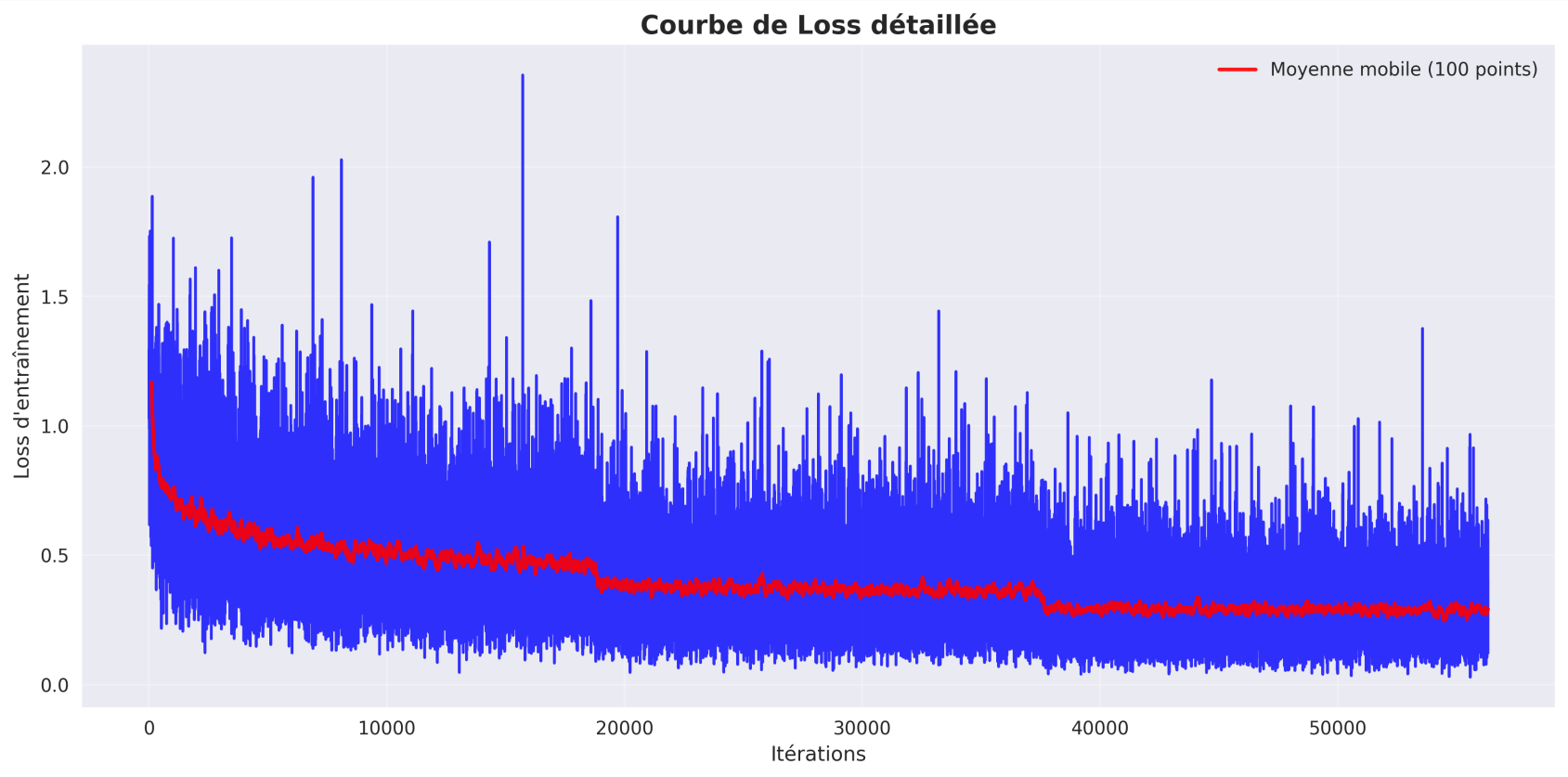


Figure 6: Courbe de *loss* du `propositioneur`. On remarque qu'elle est globalement décroissante, et que chaque époque permet de passer un palier.

Résultats expérimentaux en général

AtomicEval sur FACTScore (Min et al. 2023)

- Score d'autonomie : **52,14%**
- 47,86% des propositions nécessitent un contexte supplémentaire

Evaluation de résumé (Herserant et Guigue 2025) :
Propositionneur (finetuning FLAN-T5) :

- 42 000 exemples d'entraînement
- Propositions moyennes : 10,8 mots
- 75,1% plus rapide que les solutions existantes

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
Temps par exemple	5'38"55	1'24"30	-0,7510

Architecture proposée

Architecture

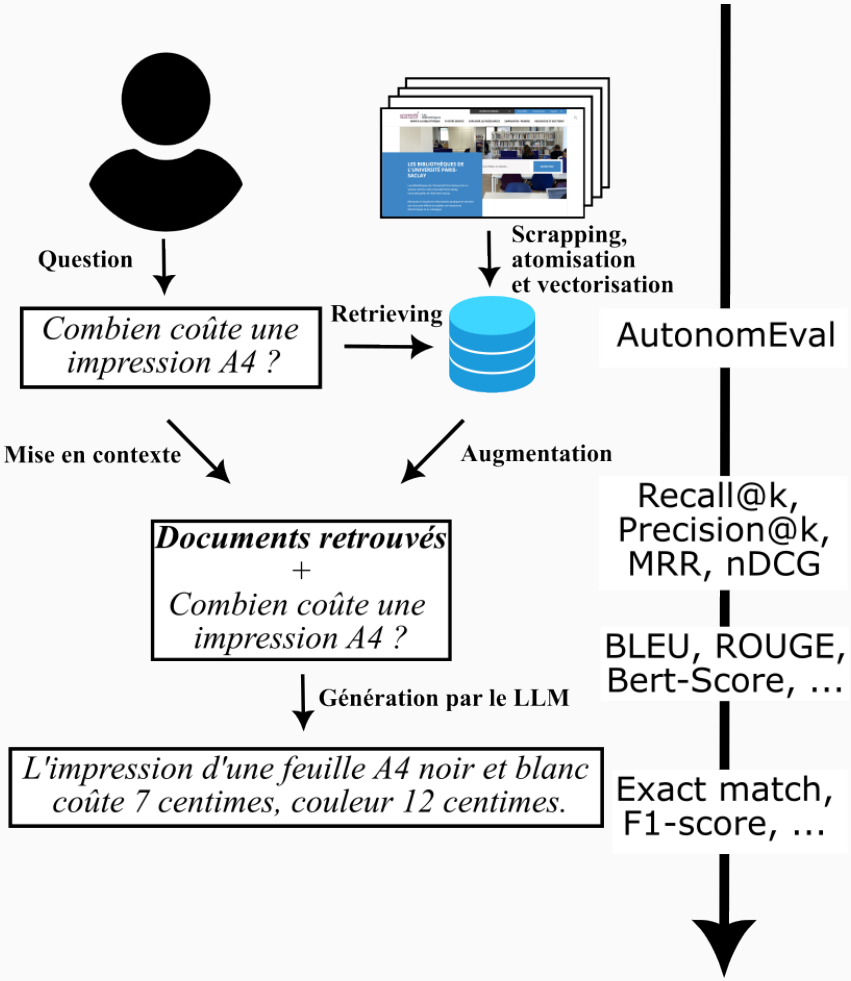


Figure 7: Architecture du système RAG pour la Bibliothèque Paris-Saclay.

Contributions et perspectives

Contributions principales

1. **Contributions théoriques**
   - Cadre théorique pour tenter de relier logique et TAL moderne
   - Analyse des différentes granularités

2. **Contribution méthodologique**
   - AtomicEval : framework d'évaluation systématique de l'autonomie

3. **Contributions techniques**
   - Propositionneur français basé sur FLAN-T5
   - Système de RAG pour la bibliothèque Paris-Saclay

Réponse à la question de recherche

« Jusqu'où peut-on goinfrer le LLM ? »

La limite n'est pas quantitative mais qualitative !

On peut « goinfrer » le LLM aussi loin que nécessaire, à condition d'optimiser simultanément :

1. La **qualité** (filtrage AtomicEval)
2. La **structure** (propositions atomiques)
3. L'**organisation** (index optimisé)

Corollaire
Transformation de la question : *comment goinfrer intelligemment* plutôt que *jusqu'où goinfrer*

Perspectives de recherche

Extensions théoriques :

- Validation expérimentale sur corpus diversifiés
- Approfondir l'étude de la sémantique des LLMs

Applications pratiques :

- Implémentation complète pour la Bibliothèque Paris-Saclay
- Extension d'AtomicEval à d'autres langues

Recherches futures :

- Intégration de signaux multimodaux dans l'évaluation
- Extension de la théorie des modèles aux représentations neuronales, *via probing*

La question qui m'anime depuis un peu plus d'un an : « **Peut-on formaliser la sémantique des LLMs ?** » Et si oui, quelle est la classe des modèles apprenables par les LLMs ?

Références

 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.

 Huh, Minyoung et al. (2024). “Position : The Platonic Representation Hypothesis”. In : *Proceedings of the 41st International Conference on Machine Learning*. Sous la dir. de Ruslan Salakhutdinov et al. T. 235. Proceedings of Machine Learning Research. PMLR, p. 20617-20642. url : https://proceedings.mlr.press/v235/huh24a.html.

Bibliographie ii

 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.

Atomic Propositions : Applications

Atomic Propositions : Core Concept

Definition (Chen et al. 2024)

Atomic propositions are self-contained, factual statements that cannot be decomposed further without losing their semantic meaning.

Key Properties :

- **Autonomy :** Each proposition is interpretable in isolation
- **Atomicity :** Cannot be broken down into smaller meaningful units
- **Factuality :** Contains a single, verifiable fact

Example Transformation

Original : "The cat and the dog are in the kitchen."
Atomic :

- "The cat is in the kitchen."
- "The dog is in the kitchen."

Applications of Atomic Propositions

1. Fact-Checking
(FACTScore)
Fine-grained verification of generated content

→ 52.14% autonomy
Identifies context dependencies

2. Summary Evaluation
(Herserant et al.)
Precise assessment of summary quality

→ 75.1% faster
Maintained accuracy

3. Dense Retrieval
(Chen et al. 2024)
Enhanced RAG with atomic granularity

→ +10.1% Recall@20
Optimal : 11.2 words/prop

Merci pour votre attention !

Questions ?