مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٩) العدد (١٢) ٢٠٢٥
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (9) No. (12) 2025

دراسة فعالية منهجيات تقطيع النصوص في أنظمة استرجاع المعلومات الأكاديمية العربية

د. جعفر سلمان*

م.ريم مهنا**

(تاريخ الإيداع ٢٠٢٥/٨/٢١ . قُبِل للنشر في ٢٠٢٥/١٠/١٣)

□ ملخّص □

يواجه استخراج المعلومات الدقيقة من المستندات النصية الطويلة تحديات تتعلق بالسرعة والكفاءة. وتُعد أنظمة الاسترجاع الذكية المعتمدة على تقنيات الذكاء الاصطناعي من الحلول الواعدة، خاصة عند دمجها مع النماذج اللغوية الكبيرة مثلChatGPT ، حيث تزودها بمعلومات منتقاة من مصادر موثوقة وحديثة.

يمثل هذا البحث أول دراسة تحليلية تجريبية تهدف إلى تصميم نظام استرجاع نصي أكاديمي متكامل للغة العربية، يركز على المجال الأكاديمي. يعتمد النظام على تقسيم المستندات إلى مقاطع (Chunks) واختبار أدائه عبر ثلاث منهجيات رئيسية: التقطيع الثابت، التقطيع البنيوي، والتقطيع الدلالي. وقد طُبّق أول منهجين ضمن ستة سيناريوهات مختلفة تراوحت بين ثلاث إعدادات لحجم المقطع (صغير، متوسط، كبير) مع أو بدون تداخل بين المقاطع، في حين شكّل التقطيع الدلالي السيناريو السابع. تم اختبار هذه السيناريوهات على مجموعة من المقالات العلمية المحكمة من مجلة جامعة طرطوس، بهدف تقييم استقلالية المقاطع ونجاح التقطيع، بالإضافة إلى دقة الاسترجاع عند مستويات مختلفة من النتائج المسترجعة.

أظهرت النتائج تفوق التقطيع الدلالي في تحقيق أعلى دقة عند النتيجة الأولى، مع أداء ممتاز عند السماح بإرجاع نتائج متعددة، بينما حققت المنهجيات الأخرى أداءً أقل في النتيجة الأولى لكنه تحسن بوضوح عند النظر في أول ثلاث إلى أربع نتائج.

تسهم هذه النتائج في إبراز أهمية اختيار منهجية التقطيع المناسبة لتحسين أنظمة الاسترجاع النصي، وتشكل خطوة نحو تطوير حلول أكاديمية عربية أكثر دقة وكفاءة.

الكلمات المفتاحية: استرجاع المعلومات، الذكاء الاصطناعي، البحث النصي، تقطيع النص، الأنظمة الذكية، أنظمة التوليد المُعزز بالاسترجاع.

*أستاذ مساعد في قسم تكنولوجيا المعلومات–كلية تكنولوجيا المعلومات والاتصالات–جامعة طرطوس الحكومية–طرطوس–سوريا

**معيدة وماجستير في قسم تكنولوجيا المعلومات–كلية تكنولوجيا المعلومات والاتصالات–جامعة طرطوس الحكومية–طرطوس–سوريا

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (٩) العدد (١٢) ٢٠٢٥
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (9) No. (12) 2025

Study of the Effectiveness of Text Chunking Methodologies in Arabic Academic Information Retrieval Systems

Dr.Jaafar salman*
Eng.Reem Muhana**

(Received 21/8/2025 . Accepted 13/10/2025)

□ ABSTRACT □

Extracting accurate information from long text documents faces challenges related to speed and efficiency. Intelligent retrieval systems based on artificial intelligence techniques are promising solutions, especially when integrated with large language models like ChatGPT, providing them with selected information from reliable and up-to-date sources.

This research represents the first experimental analytical study aimed at designing an integrated academic text retrieval system for the Arabic language, focusing on the academic field. The system relies on dividing documents into chunks and testing its performance across three main methodologies: fixed chunking, structural chunking, and semantic chunking.

The first two methodologies were applied across six different scenarios, which varied between three chunk sizes (small, medium, and large), with and without overlap. Semantic chunking, on the other hand, represented the seventh scenario. These scenarios were tested on a set of peer-reviewed scientific articles from the University of Tartous Journal, with the objective of evaluating chunk independence, segmentation success, and retrieval accuracy at different levels of returned results.

The results showed that semantic chunking achieved the highest accuracy at the top result, with excellent performance when multiple results were allowed. In contrast, the other methodologies produced lower accuracy in the first result but showed clear improvement when the top three to four results were considered.

These findings emphasis the importance of selecting an appropriate chunking methodology to improve text retrieval systems, and represent a step towards developing more accurate and efficient Arabic academic solutions.

Keywords: Information Retrieval, Artificial Intelligence, Text Search, Text Chunking, Intelligent Systems, Retrieval-Augmented Generation (RAG).

*Associate professor, Department of Information Technology, Faculty of Information and Communication Technology, University of Tartous, Syria
**Teaching Assistant and Postgraduate, Department of Information Technology, Faculty of Information and Communication Technology, University of Tartous, Syria.

١-المقدمة

شهد الذكاء الاصطناعي تطوراً مذهلاً مع ظهور النماذج اللغوية الكبيرة التي أثبتت قدرتها على فهم اللغة البشرية وتوليد نصوص طبيعية بدقة ومرونة عالية مثل chatGPT(Generative Pre-trained Transformer). وقد أحدثت هذه النماذج تحولاً جوهرياً في مختلف القطاعات، بما في ذلك التعليم، الرعاية الصحية، الصناعة، والبحث العلمي، حيث يمكنها أداء مهام معقدة مثل الإجابة على الأسئلة، تلخيص الوثائق، وتحليل البيانات الضخمة [١].

رغم هذه الإمكانيات، يواجه استخدام ChatGPT في البحث الأكاديمي تحديات جوهرية، مثل اختلاق المصادر، محدودية التغطية، غياب التقييم النقدي للمصادر، انقطاع المعرفة عند تاريخ معين، والتحيز الناتج عن البيانات التدريبية [٣,٢]. وللتغلب على هذه القيود، تبرز أهمية أنظمة الاسترجاع النصي (Retrievers) التي تدعم النماذج اللغوية الكبيرة من خلال تزويدها بمقاطع دقيقة ومحدثة من مصادر موثوقة، وتشكل الخطوة الأولى في معالجة الاستعلامات ضمن آلاف أو ملايين النصوص.

تعتمد فعالية هذه الأنظمة على منهجيات تقطيع النصوص إلى مقاطع صغيرة (Chunks)، حيث يؤثر حجم المقطع وطريقة التقطيع مباشرة على دقة وسرعة استرجاع المعلومات [٤]. في هذا البحث، تم دراسة ثلاث منهجيات رئيسية للتقطيع: التقطيع الثابت (التقطيع الحرفي)، التقطيع البنيوي (التقطيع التكراري)، والتقطيع الدلالي. وقد طُبّق المنهجان الأولان ضمن ستة سيناريوهات باستخدام ثلاثة أحجام للمقاطع مع أو بدون تداخل، بينما شكّل التقطيع الدلالي السيناريو السابع.

تم اختبار هذه السيناريوهات عملياً على مقالات علمية محكمة من مجلة جامعة طرطوس [٥]، بهدف تقييم استقلالية المقاطع ونجاح التقطيع ودقة استرجاع المعلومات. يمثل هذا البحث أول نظام استرجاع نصي أكاديمي متكامل باللغة العربية، ويعد خطوة مهمة نحو تطوير تقنيات استرجاع نصي دقيقة وفعّالة لدعم البحث العلمي باللغة العربية.

٢-الدراسات المرجعية

يعد التقطيع خطوةً أساسيةً في المعالجة المسبقة في أنظمة التوليد المعزز بالاسترجاع، حيث يؤثر بشكل كبير على فعالية الاسترجاع عبر مجموعات البيانات المتنوعة. قيّمت الدراسة [٦] استراتيجيات التقطيع ذات الحجم الثابت وتأثيرها على أداء الاسترجاع باستخدام مجموعات بيانات قصيرة وطويلة، وأظهرت النتائج أن حجم القطعة يلعب دوراً حاسماً في فعالية الاسترجاع؛ إذ إن القطع الأصغر أكثر ملاءمة للبيانات ذات الإجابات الموجزة والقائمة على الحقائق، بينما تُحسّن القطع الأكبر الاسترجاع في المجموعات التي تتطلب فهماً سياقياً أوسع. **وقد تميّزت هذه الدراسة بتركيزها على الفرق بين البيانات القصيرة والطويلة، غير أنها اقتصرت على الجانب الكمي للحجم دون النظر إلى بنية النص أو معناه أو تطرق إلى منهجيات تقطيع أخرى.**

في المقابل، ناقشت الدراسة [٧] التقطيع البنيوي في مجال التقارير المالية، منتقلةً من تقسيم يعتمد على حدود الفقرات والجمل إلى تقسيم قائم على العناصر البنيوية للتقارير، مما حسن دقة النتائج بشكل ملحوظ. **ورغم أهميتها، إلا أنها ظلت مقتصرة على المجال المالي ولم تُعمم على نصوص أكاديمية أو لغات أخرى ولم تقارن مع منهجيات تقطيع أخرى.**

وفيما يتعلق بالتقطيع الدلالي، فرغم انتشاره المتزايد، لا يزال من غير الواضح مدى تفوقه على منهجيات التقطيع التقليدية، إذ يبقى تحديد الاستراتيجية الأمثل تحدياً بين جودة الاسترجاع والكفاءة الحاسوبية. **فقد أكدت الدراسة [٨] أن اختيار استراتيجية التقطيع يعتمد على طبيعة البيانات وسيناريو الاسترجاع**، مع ضرورة الموازنة بين الأداء والتكلفة الحسابية عند تصميم أنظمة الاسترجاع. وأظهرت النتائج أن فوائده تعتمد بدرجة كبيرة على نوع المهمة، فهو أكثر فاعلية في البيانات المركبة ذات التنوع الموضوعي الكبير (مثل مقالات من مجالات متعددة: علوم، اقتصاد، طب)، بينما يظل التقطيع التقليدي أكثر كفاءة وموثوقية في البيانات غير المركبة (مثل التقارير المالية) نظراً لبساطته وانخفاض عبئه الحاسوبي،**رغم أن هذه الدراسة تناولت التقطيع الدلالي بعمق إلا أنها تحتاج إلى تحقق تجريبي واسع للتقطيع التقليدي بسيناريوهات مختلفة.**

بناءً على ما سبق، يتضح أن الدراسات السابقة قدمت إضاءات مهمة على فعالية كل من التقطيع الثابت والبنيوي والدلالي، لكنها عانت من محدوديات تتعلق إما بحصر التطبيق في مجالات محددة أو بإهمال اللغة العربية والنصوص الأكاديمية المحكمة. ومن هنا تأتي القيمة المضافة لهذا البحث، **حيث يقدم أول دراسة تجريبية متكاملة تقارن بين هذه المنهجيات الثلاث في سياق أكاديمي عربي وفق سبع سيناريوهات مختلفة، بالاعتماد على مقالات علمية محكمة، مما يوفر أدلة كمية واضحة حول دور التقطيع في رفع دقة الاسترجاع في النصوص العربية.**

٣-أهمية البحث وأهدافه

تنطلق أهمية االبحث من حاجة أنظمة الذكاء الاصطناعي الحديثة إلى الاسترجاع (Retrieval) بسبب وجود قيود جوهرية في النماذج اللغوية الكبيرة، وأهمها:

- **سعة ذاكرة محدودة(Context Window) :**أي النماذج لا تستطيع معالجة نصوص طويلة جداً دفعة واحدة، وإذا زاد حجم النص عن حد معين قد تتجاهل أجزاء مهمة أو تقع في أخطاء.
- **تشتت المعلومات غير المهمة:** حيث أن وجود معلومات كثيرة أو غير مرتبطة بالسؤال يؤدي إلى تشتيت النموذج، فيصعب عليه التركيز على الجزء المطلوب من النص، مما يقلل دقة الإجابة.

وقد أظهرت دراسات حديثة أن تشتت أو ضعف جودة المعلومات في نافذة السياق (context window) يؤدي إلى تراجع أداء النموذج بشكل ملحوظ [10,9]، كما أنها قد تُصاب بالهلوسة (hallucinate) فتُنتج عبارات تبدو معقولة لكنها غير صحيحة أو غير منطقية، بالتالي يمكن حل هذه القيود من خلال تزويد النموذج فقط بالمقاطع الضرورية والمرتبطة بالسؤال، لضمان دقة الإجابة وتركيز النموذج على المعلومات المهمة دون تشويش أو إهدار للموارد.

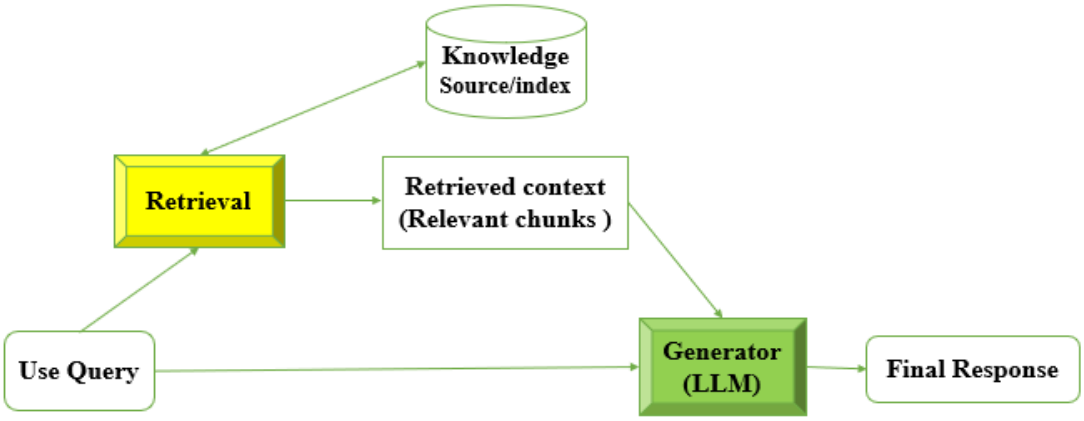
استجابةً لندرة الدراسات التي تهتم بتطوير أنظمة الاسترجاع في المجال الأكاديمي، على الرغم من أهميتها الكبيرة في البحث العلمي، يسعى هذا البحث إلى سد هذه الفجوة من خلال تصميم أول نظام استرجاع نصي متكامل باللغة العربية في المجال الأكاديمي، مما يجعله عملاً فريداً ومتكاملاً.

ويهدف البحث بشكل رئيسي إلى تقييم أثر استراتيجيات تقطيع النص المختلفة على فعالية أنظمة استرجاع المعلومات النصية في المستندات العلمية الأكاديمية، من خلال مقارنة التقطيع الحرفي، التقطيع المتكرر، والتقطيع الدلالي. تم اختبار هذه المنهجيات عبر سبعة سيناريوهات تجريبية شملت ثلاث إعدادات للتقطيع ذو الحجم الثابت والبنيوي (حجم المقطع ٣٠٠ حرف مع تداخل ٠،حجم المقطع ٥٠٠ حرف مع تداخل ٥٠، حجم المقطع ٨٠٠

حرف مع تداخل ١٠٠) والحالة السابعة للتقطيع الدلالي. يضمن تنفيذ هذه السيناريوهات تقديم أدلة عملية على أهمية منهجيات التقطيع في تحسين أداء أنظمة الاسترجاع النصي باللغة العربية، مما يعزز القدرة على الوصول إلى المعلومات الدقيقة والموثوقة في النصوص الأكاديمية.

٤-طرائق البحث ومواده

نظام التوليد المعزز بالاسترجاع (RAG:Retrieve Augmented Generation) هو نظام متقدم يجمع بين البحث الذكي في مصادر المعرفة (مثل قواعد البيانات أو المستندات) والقدرة على توليد إجابات نصية باستخدام النماذج اللغوية الكبيرة (LLMs:Large Language Model) مثل ChatGPT.يتألف هذا النظام [6] من مرحلتين أساسيتين وفق الشكل (١) كالتالي:

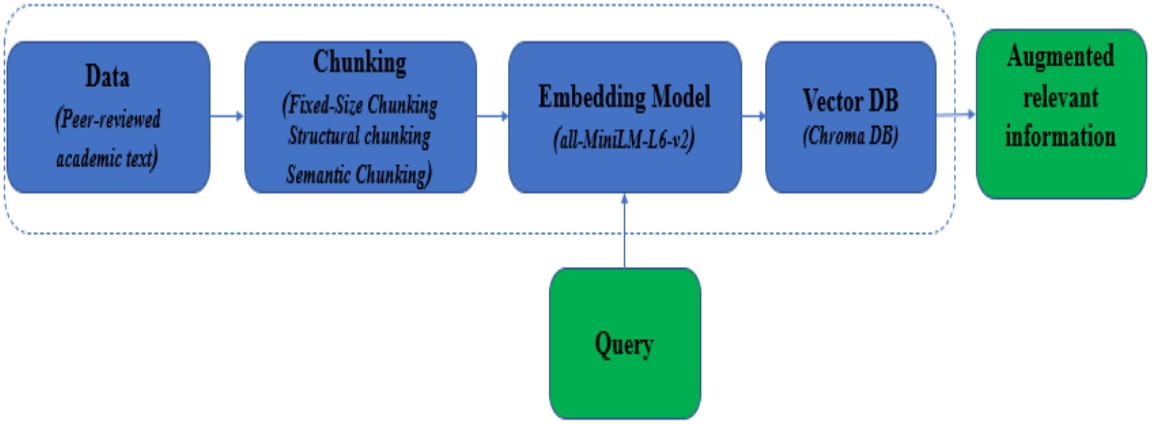


الشكل (١): نظام التوليد المعزز بالاسترجاع

- **الاسترجاع(Retrieval):** عندما يُرسل المستخدم استعلاماً (Use Query)، يستخدم نظام RAG الاستعلام أولاً للبحث ضمن قاعدة معارف مُحدّدة مسبقاً (Knowledge). قد تكون قاعدة المعارف هذه مجموعة من المستندات، أو قاعدة بيانات، أو صفحات ويب، أو مصادر بيانات نصية أخرى ذات صلة بالاستعلامات المتوقعة. الهدف من هذه المرحلة هو العثور على مقتطفات نصية أو مستندات أكثر صلة باستعلام المستخدم (Retrieved context) وهي ما نعني بها (المعزز) أي هي عملية إثراء المُدخلات بالأهم و الأكثر صلة.يُطلق على هذا المُكوّن اسم المُسترجع **(Retriever)**.
- **التوليد(Generation):** تُدمج المعلومات ذات الصلة المُسترجعة في الخطوة الأولى (Relevant chunks) مع استعلام المستخدم الأصلي. يُشكّل هذا النص المُدمَج موجّه معزز يُغذي النموذج اللغوي الكبير (LLM) و الذي نطلق عليه اسم المُولِّد **(Generator)** والذي يُولد بدوره إجابة نهائية دقيقة وملائمة للسؤال المطروح (Final Response).

في هذا البحث، سنركز فقط على مرحلة الاسترجاع، أي على كيفية تقطيع المقالات العربية الأكاديمية واختيار المقاطع الأنسب لاسترجاعها، دون التطرق إلى مرحلة توليد الإجابة نفسها، وذلك بهدف تحسين فعالية ودقة استرجاع المعلومات من المقالات العربية العلمية المحكمة، مما ينعكس مباشرة على جودة الإجابات النهائية في أنظمة **RAG**.

يوضح الشكل (٢) المراحل الأساسية للمُسترجع و التقنيات المستخدمة في هذا البحث، تم تنفيذ النهج المقترح باستخدام المحاكاة الحاسوبية من خلال برنامج Visual Studio Code بإصدار بايثون ٣.٧.



الشكل (2): خطوات البحث

١.**البيانات(Data)**:أول خطوة في تجهيز البيانات هي استخراج النصوص من جميع المصادر بصيغ مختلفة (pdf,word,txt) مع الحفاظ على كامل محتوى المقالات. تم اختيار مجموعة من **المقالات العلمية المحكمة العربية** [5] كنموذج لتطبيق الاسترجاع في المجال الأكاديمي.

٢.**التقطيع(Chunking)**: تمثل المعالجة المسبقة للبيانات بتقطيعها أو تجزئتها، إذ تحتوي معظم نماذج اللغة الكبيرة على نافذة سياقية(context window)، وهي حدّ صارم لكمية النص التي يُمكن أخذها في الاعتبار في أي وقت عند معالجة المدخلات وتوليد المخرجات، يشمل هذا الحدّ الاستعلام الأصلي والمعلومات السياقية التي يتم استرجاعها. بالإضافة إلى ذلك فإن إنشاء تضمين متجه واحد لنص كبير جداً يقلل من التفاصيل الدلالية المحددة داخله. لذلك، فإن التقسيم ليس مجرد ضرورة تقنية بسبب قيود النموذج؛ إنها تقنية أساسية لتحسين جودة ودقة المسترجع إذ أن تقسيم المستندات إلى أجزاء قابلة للفهم، تضمن توافق السياق مع متطلبات LLM، وتزيد بشكل كبير من فرص أن تكون المعلومات المسترجعة هي بالضبط ما تحتاجه للإجابة على استفسار المستخدم بفعالية[7]،و للتقطيع عدة منهجيات منها:

- **منهجية التقطيع الثابت (Fixed Chunking)**: تعد أبسط تقنية لتقطيع النص إذ تقسمه بناءً على الطول مع تجاهل البنية الدلالية،استخدمنا كمثال عليها **التقطيع الحرفي (Character-Based Splitting)** حيث نحدد حجم المقطع (Chunk) ب ١٠٠٠حرف مثلاً ويتم تقسيمه إلى عدة مقاطع بالحجم المذكور.تتميز هذه الطريقة بسهولة التنفيذ وبمقاطع متساوية الحجم إلا أنها تعاني من العمى الدلالي (Semantic Blindness) وهو تقسيم الكلمات أو الجمل إلى نصفين مما يؤدي إلى خلل في المعنى عند حدود الأجزاء بالتالي يصعب على نموذج التضمين استيعاب السياق الكامل للجملة المقطعة[7].
- **منهجية التقطيع البنيوي (Structural chunking)**: تهدف هذه المنهجية إلى الحصول على مقطع يحمل فكرة مكتملة ويمثل وحدة دلالية كاملة وذلك من خلال البحث عن الحدود الطبيعية داخل النص مثل نهاية الفقرات أو الجمل مما يزيد من احتمال أن يكون الجزء المسترجع كافياً للإجابة على السؤال بدقة.استخدمنا كمثال عليها **التقسيم المتكرر (Recursive chunking)** حيث يتم تقطيع النص باستخدام قائمة فواصل مرتبة حسب الأولوية ['\n\n','\n','.',' ',''] ننتقل من فاصل إلى آخر إذا تجاوز المقطع الحجم المحدد مع مراعاة آخر

فاصل موجود حتى لا نخسر المعنى وهو ما نسميه هامش المرونة (wiggle room) والذي يمنع انقسام الجمل أو الكلمات في أماكن غير منطقية[7].

- **منهجية التقطيع الدلالي (Semantic Chunking):** تعتمد هذه المنهجية على تقسيم النصوص وفقاً لمحتواها الدلالي، بحيث تُدمج الجمل أو العبارات المتقاربة في المعنى في مقطع واحد، مما يحافظ على الترابط المعنوي ويعزز قدرة أنظمة الاسترجاع على فهم السياق الكامل. على عكس التقطيع الثابت أو البنيوي، لا تستند هذه الطريقة إلى الطول أو الفواصل النصية فقط، بل تستخدم نماذج التضمين الدلالي (Sentence Embeddings) لقياس درجة التشابه بين الجمل المتتالية، ثم تحدد نقاط الفصل ديناميكياً باستخدام عتبة إحصائية قائمة على المعيار الإحصائي (z-score) [11] كما موضح في المعادلة (١):

$$\tau = similarity^{\mu} - 0.5\, similarity^{\sigma} \quad (1)$$

- حيث $similarity^{\mu}$:هو متوسط قيم التشابه بين الجمل المتتالية(Cosine Similarity) .
- $similarity^{\sigma}$هو الانحراف المعياري لقيم التشابه.
- تُعتبر الجملة بداية مقطع جديد إذا كان التشابه أقل من العتبة τ و إلا تُدمج في المقطع الحالي.

يتم حساب التشابه بين الجمل [12] باستخدام تشابه جيب التمام (Cosine Similarity) بين متجهات التضمين، وفق المعادلة (٢):

$$\cos(\theta) = similarity(i, i+1) = \frac{A.B}{||A||\,||B||} = \frac{embeddings[i].embeddings[i+1]}{||embeddings[i]||\,||embeddings[i+1]||} \quad (2)$$

حيث تمثل A متجه التضمين للجملة i، و B متجه التضمين للجملة التالية i+1.

لتوليد هذه التضمينات، استخدمنا نموذج all-MiniLM-L6-v2 وهو نموذج صغير الحجم مبني على معمارية المحولات(Transformer) ومدرب على مهام التشابه الدلالي (Semantic Textual Similarity).يتميز هذا النموذج بتوازن جيد بين الدقة والسرعة، وحجم متجه ٣٨٤ بعداً، مما يجعله مناسباً للتطبيقات التي تتطلب أداءً عالياً مع استهلاك منخفض للموارد[13]. توفر هذه المنهجية مقاطع متسقة المعنى، مما يقلل من فقدان السياق ويزيد من دقة الاسترجاع، لكنها تتطلب موارد حسابية أكبر مقارنة بالتقطيع التقليدي نظراً لاعتمادها على النماذج الدلالية الجاهزة، مما يلغي الحاجة إلى تدريب نموذج من الصفر على بيانات ضخمة ويختصر الوقت مع الحفاظ على جودة التضمينات.

تُحدد التجربة المنهج الأمثل للتقطيع النصي من خلال مقارنة أداء الاسترجاع على عينة من النصوص والاستعلامات، بهدف الوصول إلى توازن بين مقاطع صغيرة كفاية لمعالجة فعّالة، ومقاطع كبيرة كفاية لاحتواء معلومات مترابطة وذات صلة بالاستعلامات المحتملة للمستخدم.كما تجدر الإشارة إلى أن **تحديد درجة التداخل بين المقاطع (Overlap) إلى جانب حجم المقطع (Chunk Size)** يعد عاملاً مهماً للحفاظ على ترابط الأفكار وعدم فقدان المعلومات الحرجة الواقعة عند حدود المقاطع، حيث يتيح التداخل لكل مقطع أن يحتوي على بعض الجمل أو الكلمات من نهاية المقطع السابق،مما يعزز فهم السياق الكامل أثناء البحث.علاوة على ذلك، فإن القيم المثالية للتداخل وحجم المقاطع تعتمد على طبيعة النصوص

المستخدمة، ويجري تحديدها تجريبياً عبر تقييم أثرها على جودة عملية الاسترجاع و في هذه الدراسة تم اختبار عدة قيم مختلفة واختيار الأنسب منها بناءً على نتائج التقييم.

٣.نموذج التضمين (Embedding Model): التضمين الشعاعي هو التمثيل الرقمي للنصوص في فضاء متجهي، بحيث تتحول الجمل أو المقاطع النصية إلى متجهات عددية يمكن مقارنتها من حيث المعنى. يتيح هذا التمثيل لأنظمة الاسترجاع الذكية الانتقال من مطابقة الكلمات المفتاحية إلى فهم العلاقات الدلالية. عند طرح المستخدم سؤالًا يتم تضمينه إلى شعاع باستخدام النموذج (all-MiniLM-L6-v2) المذكور مسبقاً ثم حساب درجة التشابه وفق المقياس المستخدم في المعادلة (٢)، وذلك لقياس مدى تقارب الاستعلام (A) مع تضمينات المقاطع المخزّنة مسبقاً (B) في قاعدة بيانات شعاعية والتي تم تقطيعها باستخدام المنهجيات السابقة، مما يضمن دقة واتساق عملية المقارنة بين النصوص.

٤.قواعد البيانات الشعاعية (Vector DB): إن مقارنة تضمين الاستعلام مع كل تضمين في قاعدة البيانات لاختيار الأعلى تشابهاً يسمى (Brute-force similarity search)، يتميز بالدقة العالية لأننا نقارن مع كل التضمينات لكنه بالمقابل بطيء جداً وغير عملي في التطبيقات الواقعية التي تتطلب استجابة فورية.من ناحية أخرى قواعد البيانات التقليدية مثل (SQL) غير مصممة للتعامل مع فضاء التضمينات.

بناء عليه تم استخدام قواعد بيانات متخصصة في تخزين التضمينات والبحث عنها بكفاءة وسرعة هائلة وتعتمد على:

- **البحث عن الجار الأقرب التقريبي(Approximate Nearest Neighbor - ANN) :**هذه التقنية تُستخدم بدلاً من مقارنة كل زوج من التضمينات، وهو أمر مكلف جداً زمنياً في قواعد البيانات الضخمة، تعتمد على هياكل الفهرسة للعثور بسرعة على المتجهات الأقرب إلى متجه الاستعلام، مع الحفاظ على دقة عالية في النتائج توازي تقريباً البحث الشامل في معظم التطبيقات العملية،مما يجعل البحث عملياً وقابلاً للتوسع حتى مع ملايين أو مليارات المتجهات.
- **هياكل الفهرسة(Indexing Structures) :**تُبنى هياكل فهرسة متخصصة لتنظيم المتجهات داخل قاعدة البيانات،مما يقلل عدد العمليات الحسابية اللازمة عند كل عملية بحث، مما يؤدي إلى استرجاع النتائج بسرعة كبيرة جداً مقارنة بالبحث التقليدي.

إحدى قواعد البيانات الشعاعية الشائعة التي تقدم هذه الخدمات هي Chroma والتي تستخدم الرسم البياني (Graph-based HNSW) كنوع للفهرسة أي أنه يتم تنظيم المتجهات (التضمينات) على شكل طبقات من الرسوم البيانية، بحيث تبدأ عملية البحث من طبقة تحتوي على روابط طويلة المدى بين المتجهات البعيدة، ثم تنتقل تدريجياً إلى طبقات أدنى ذات روابط أقصر وأكثر كثافة بين المتجهات القريبة، مما يسمح بالوصول بسرعة ودقة إلى المتجهات الأكثر تشابهاً مع الاستعلام دون الحاجة لفحص كل المتجهات في القاعدة.

٥-المناقشة

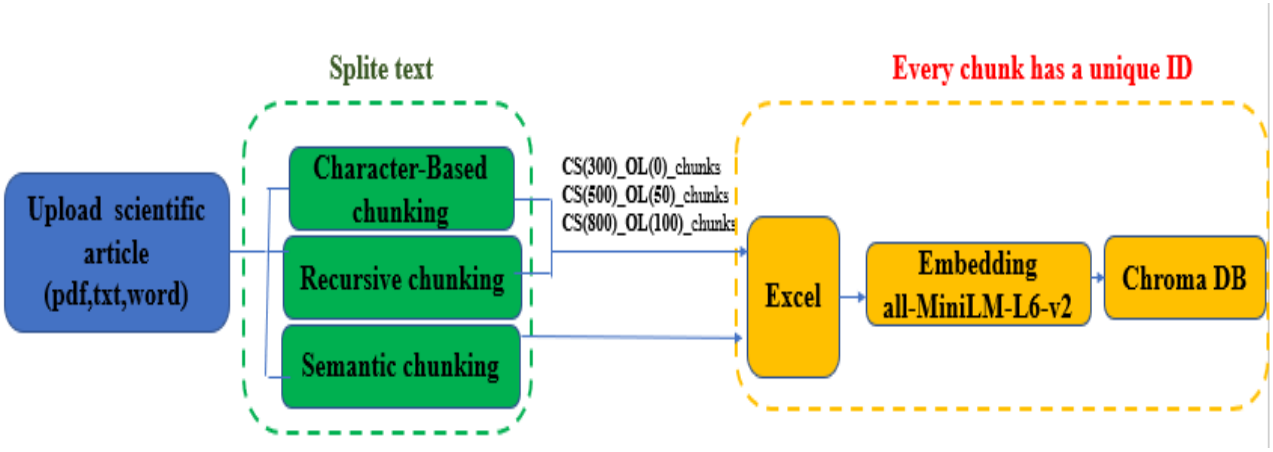
تم تنفيذ الدراسة على مرحلتين أساسيتين:

المرحلة الأولى: تنفيذ نظام الاسترجاع الشكل(3).

المرحلة الثانية: تقييم نظام الاسترجاع الشكل(4).

٥-١ تنفيذ نظام الاسترجاع

تم تنفيذ الدراسة وفق الخطوات التالية والموضحة بالشكل (٣) كالتالي:



الشكل (3): نظام الاسترجاع.

١. **رفع الملفات النصية:** تم إدخال الملفات بصيغ PDF، Word، وTXT، حيث يقوم النظام تلقائياً باستخراج النصوص منها. وقد تم تطبيق هذه الخطوة على مجموعة من المقالات العلمية المحكمة [٥] بوصفها عينة تجريبية، وذلك بهدف تحديد أنسب نوع من التقطيع والمعاملات المرتبطة به بما يتلاءم مع طبيعة هذا النوع من المقالات.

٢. **تقطيع النص وفق نوع التقطيع وحجم المقطع (CS) والتداخل المطلوب (OL):** في هذه الدراسة تم استخدام ثلاث منهجيات للتقطيع (التقطيع الحرفي والتقطيع التكراري و التقطيع الدلالي)، ومن أجل أول منهجيتين تم اعتماد ثلاث حالات تجريبية مختلفة:

- حجم مقطع 300حرف بدون تداخل.
- حجم مقطع 500 حرف بتداخل ٥٠.
- حجم مقطع 800 حرف بتداخل ١٠٠.

وبهذا، بلغ مجموع السيناريوهات المنفّذة **سبعة سيناريوهات** مختلفة.

٣. **تخزين المقاطع** :تم تخزين المقاطع الناتجة في ملف Excel مخصص لكل نوع ومعاملاته، مع توليد معرف فريد لكل مقطع لضمان إمكانية المطابقة لاحقاً.

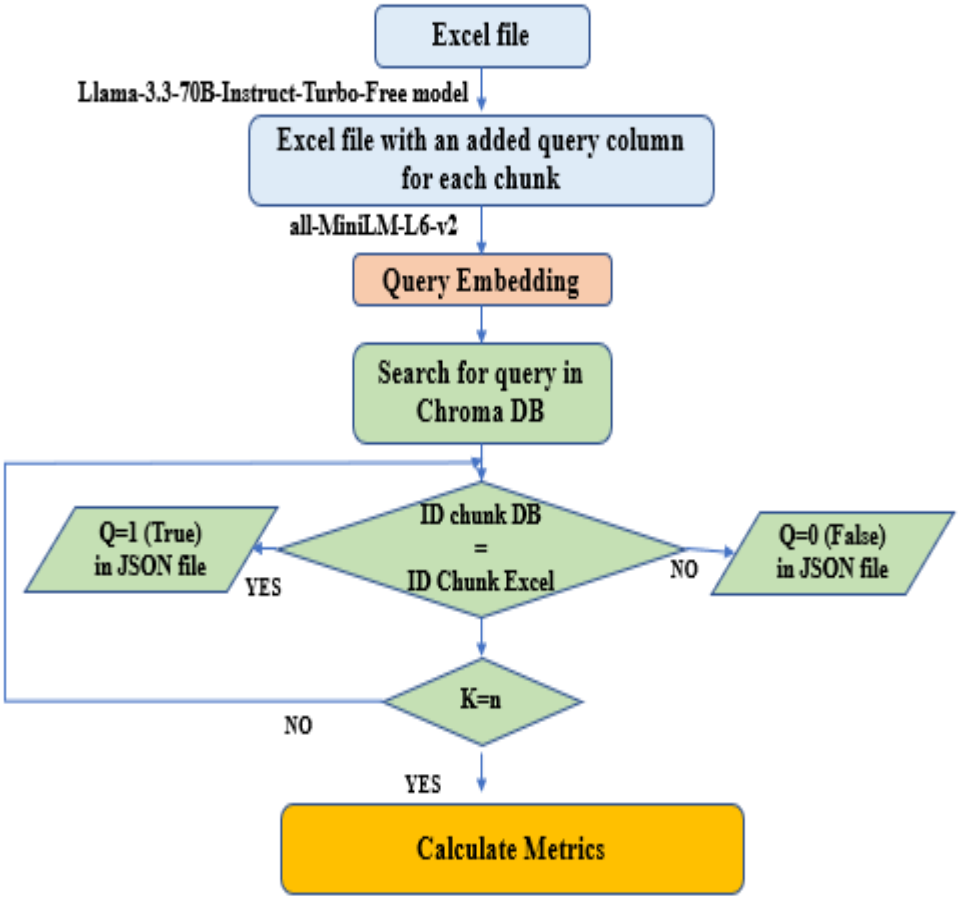
٤. **تضمين المقاطع:** تم تحويل المقاطع إلى تمثيلات عددية (Embeddings) باستخدام نموذج all-MiniLM-L6-v2، وتخزينها في قاعدة بيانات Chroma كمتجهات عددية حيث لكل نوع ومعاملاته قاعدة بيانات خاصة به، وقد رُبط كل تضمين بمعرف المقطع ذاته الموجود في ملف Excel، مما يضمن تكامل البيانات بين المصدرين ويُسهل عمليات البحث والتحليل بدقة وكفاءة عالية.

تضمن هذه العملية المؤتمتة تكامل البيانات بين ملف Excel وقاعدة بيانات Chroma، مع الحفاظ على معرفات موحدة لكل مقطع، مما يسهل عمليات البحث والاسترجاع والتحليل لاحقاً بدقة وكفاءة عالية.

٥-٢ تقييم نظام الاسترجاع

يتم تقييم أداء نظام استرجاع المعلومات النصية المعتمد على ChromaDB والموضح بالشكل (٤) من خلال:

- قياس دقة استرجاع المقاطع النصية (Chunks) بناءً على استعلامات محددة.
- حساب مقاييس الأداء (accuracy@k,F1-score@K,Recall@K,Precision@K) حيث يشير k إلى عدد النتائج التي يتم أخذها بعين الاعتبار عند التقييم؛ فعندما تكون k=1 يتم الاعتماد فقط على أفضل نتيجة، بينما عند k=4 يتم التقييم استناداً إلى أفضل أربع نتائج مسترجعة.



الشكل (4): تقييم نظام الاسترجاع.

وقد تم تنفيذ عملية التقييم وفق الخطوات التالية:

١.توليد الاستعلامات :

- لكل مقطع نصي في السيناريوهات السبعة، يتم إنشاء استعلام قصير ودقيق بشكل آلي باستخدام نموذج **Llama-3.3-70B-Instruct-Turbo-Free**، يتيح هذا النموذج إنشاء استعلامات تعكس محتوى المقاطع النصية بدقة نظراً لقدراته المتقدمة في الفهم والتوليد متعدد اللغات.
- كل استعلام مرتبط مباشرة بمعرف المقطع الذي تم توليده منه، مما يسمح بتحديد الإجابة الصحيحة أثناء عملية التقييم.

٢.تضمين الاستعلامات: يستخدم نموذج all-MiniLM-L6-v2 لتمثيل الاستعلامات بشكل رقمي، بحيث يمكن مقارنة التشابه بينها وبين المقاطع المخزنة المضمنة في ChromaDB.

٣. البحث والتحقق من النتائج في ChromaDB:

- يتم البحث عن المقاطع الأكثر تشابهاً مع الاستعلام في قاعدة البيانات.

- إذا تم استرجاع مقطع بنفس المعرف ضمن عدد النتائج المطلوبة (n)، تعتبر الإجابة صحيحة (Q=1)، وإلا فهي خاطئة(Q=0).
- تحفظ جميع النتائج في ملفات JSONلكل سيناريو.

٤.حساب مقاييس الأداء (Accuracy@K و F1-score@K و Recall@K و Precision@K)

بناءً على النتائج المسترجعة المخزنة في ملفات JSON لكل سيناريو.

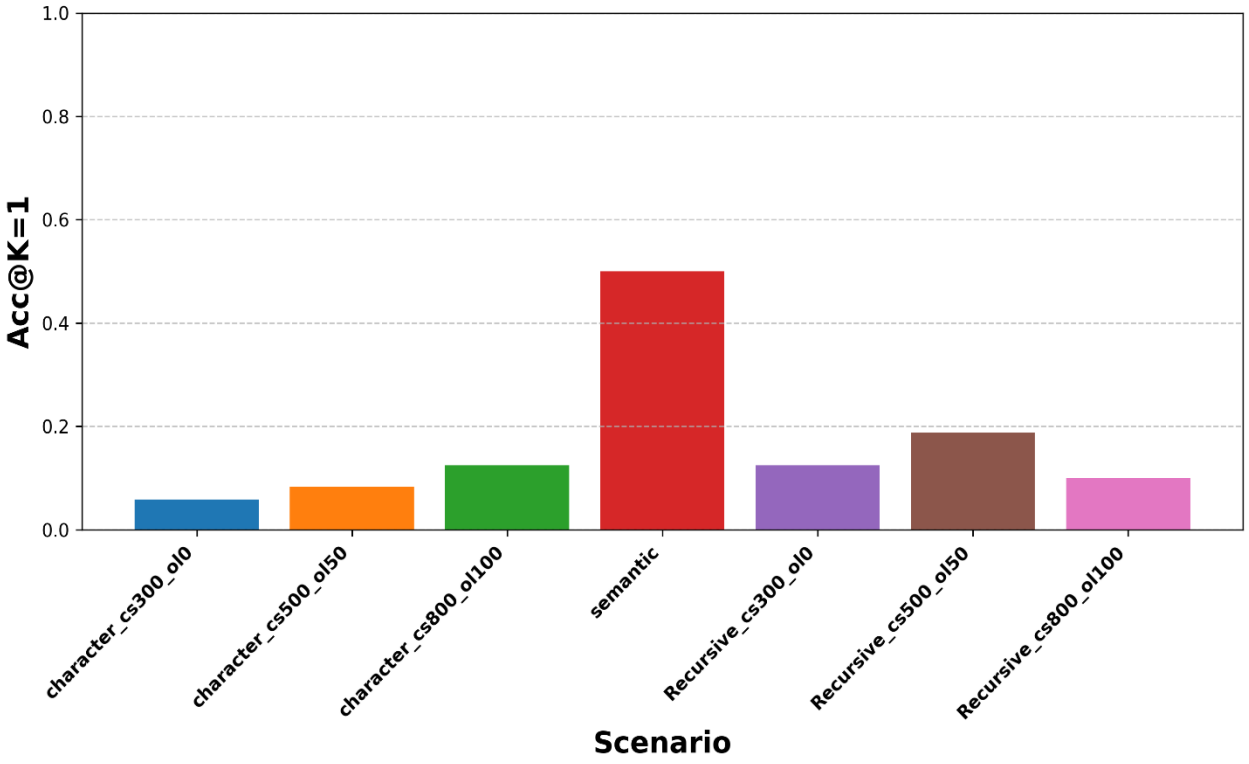
يسهم هذا التقييم في التأكد من كفاءة النظام في الاسترجاع الذكي، ويُظهر مدى فعالية الاعتماد على التضمين الشعاعي والفهرسة المتقدمة في تسريع عملية الوصول إلى المعلومات الأكثر صلة من مجموعات نصية ضخمة.

٥-3 النتائج

يوضح الشكل (٥) المقارنة بين منهجيات التقطيع الثلاث باستخدام معاملات مختلفة وفقاً لمقياس الدقة عند السماح للنظام بإرجاع نتيجة واحدة فقط (k=1). ويُعرَّف مقياس الدقة (Accuracy) بأنه النسبة بين عدد الحالات التي تم تصنيفها بشكل صحيح، أي مجموع الاستجابات الصحيحة الإيجابية (True Positives)والاستجابات الصحيحة السلبية (True Negatives)، إلى العدد الكلي للحالات التي تم اختبارها. ويمكن التعبير عنه رياضياً بالمعادلة التالية[14]:

$$\boldsymbol{Accuracy = \frac{TP + TN}{TP + TN + FP + FN}} \qquad (3)$$

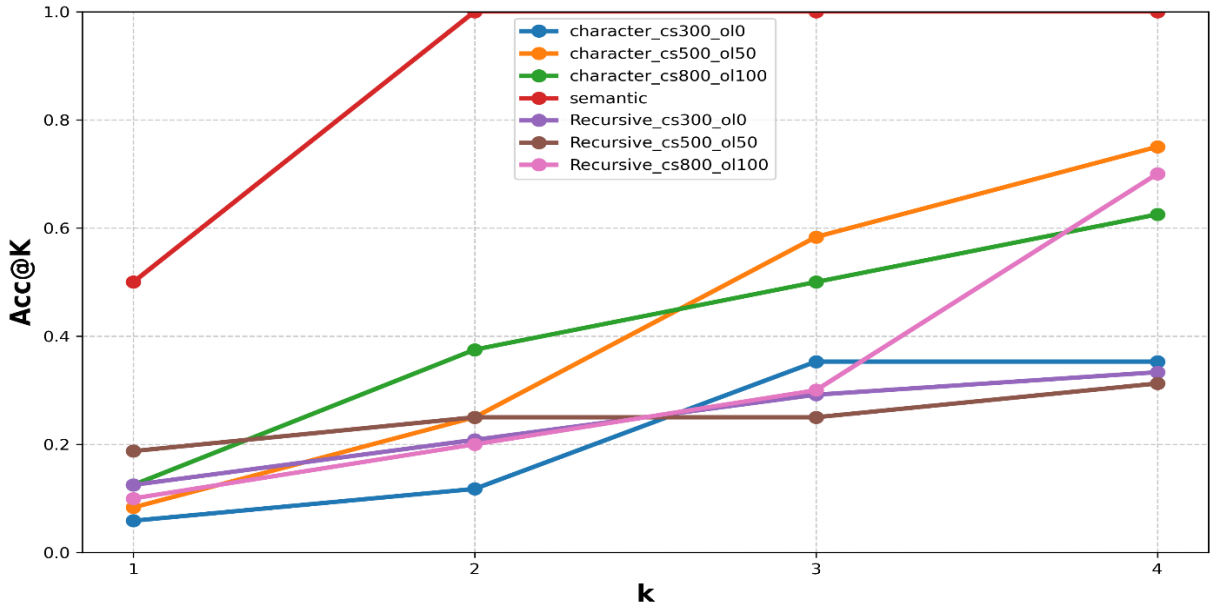
حيث: TP(True Positives)عدد الحالات الصحيحة المسترجعة،TN(True Negatives) عدد الحالات الصحيحة غير المسترجعة،FP(False Positives) عدد الحالات الخاطئة المسترجعة، (False FN Negatives) عدد الحالات الخاطئة غير المسترجعة.



الشكل (5): مقارنة بين نوعي التقطيع بالمعاملات المختلفة من أجل نتيجة واحدة

- في منهجية التقطيع Character-Based chunkingنلاحظ أن دقة الاسترجاع منخفضة جداً حتى مع تغيير طول المقطع أو التداخل، ويعزى السبب إلى أن هذا النوع من التقطيع لا يلتقط البعد الدلالي بشكل كافٍ، بل يركز فقط على وحدات سطحية (حروف) دون مراعاة المعنى أو السياق، مما يضعف من جودة المطابقة مع الاستعلامات.
- تعطي منهجية التقطيع Recursive chunking أداءً أولياً أعلى مقارنة بالنوع الأول،ويعزى السبب أن التقطيع التكراري يحاول الحفاظ على معنى أكثر من التقطيع الحرفي، لكنه ما زال محدوداً في استرجاع النصوص الدقيقة،كما نفسر أن المعاملات الأفضل هي حجم المقطع (٥٠٠) مع تداخل (٥٠) بالتالي:
  - ✓ يحتوي المقطع (٥٠٠) على معلومات أكثر تركيزاً وذات صلة مباشرة بالسؤال، مما يجعلها أقرب لمحتوى الاستعلام عند البحث وأكثر تشابهاً مقارنة مع مقطع (٨٠٠) الذي قد يحتوي على تفاصيل غير مرتبطة مباشرة بالاستعلام، مما يخفف من وضوح السياق ويقلل من احتمالية تطابقه الدقيق مع الاستعلام.
  - ✓ التداخل (٥٠ كلمة في ٥٠٠) يضمن عدم فقدان المعلومات عند حدود المقاطع، لكن التداخل الأكبر (١٠٠ كلمة في ٨٠٠) قد يؤدي إلى تكرار زائد في المعلومات بين المقاطع، ما يقلل من تميز كل مقطع عن الآخر ويؤدي إلى نتائج أقل دقة لأنها تدمج الكثير من المعلومات غير الضرورية، مما يقلل من جودة المطابقة مع الاستعلامات ويؤثر سلباً على أداء النظام.
- حققت منهجية التقطيع Semantic chunking أعلى أداء بين جميع السيناريوهات حيث يعطي دقة أفضل بمرتين أو أكثر مقارنة بأفضل سيناريو تقليدي. وذلك لأنه يعتمد على المعنى وليس مجرد عدد الحروف أو الطول، وبالتالي يحقق تمثيل أفضل للنصوص في قاعدة البيانات ويزيد فرص استرجاع المقطع الصحيح عند الاستعلام.

يوضح الشكل (6) المقارنة بين منهجيات التقطيع بالمعاملات المختلفة من أجل قيم مختلفة ل k.



الشكل (6): مقارنة بين نوعي التقطيع بالمعاملات المختلفة حتى ٤ نتائج (accuracy@k)

• **تبدأ منهجية التقطيع الحرفي (Character-Based chunking)** بأداء ضعيف جداً عند k=1 لكن يتحسن تدريجياً مع زيادة عدد النتائج المسترجعة k، كما يُلاحظ تحسن الأداء عند زيادة حجم المقطع والتداخل. ومع ذلك، يبرز من البيانات أن السيناريو الثاني(character_cs500_ol50) حقق أداء أفضل من(character_cs800_ol100)، مما يشير إلى أن تحسين التقطيع الحرفي لا يعتمد فقط على زيادة الحجم أو التداخل، بل على تحقيق توازن مناسب بينهما لضمان أفضل أداء.

• **إن أداء منهجية التقطيع التكراري أقل من التقطيع الحرفي و حقق السيناريو** Recursive_cs800_ol100 أفضل النتائج بين السيناريوهات الثلاثة لهذه المنهجية.

• **منهجية التقطيع الدلالي(Semantic Chunking) :** تفوقت هذه المنهجية بشكل واضح، حيث حققت استرجاعاً مثالي بدءاً من (k=2)، مما يدل على فعالية النهج الدلالي في الوصول إلى النتائج الصحيحة بشكل أكبر مقارنة بالطرق التقليدية التي احتاجت إلى k=4 حتى تصل إلى أقصى أداء ممكن.

بالتالي نلاحظ أن زيادة عدد النتائج المسترجعة (k) تحسن الأداء بشكل عام، إذ يزيد احتمال وجود المقطع الصحيح ضمن النتائج المسترجعة. ومع ذلك، يظهر أن تأثير زيادة k يكون أقوى في الطرق التقليدية (الحرفي والتكراري)، بينما يصل النهج الدلالي إلى الأداء الأمثل بسرعة ويصبح أقل تأثراً بزيادة k.

إن مقياس الدقة (accuracy) يعطي فكرة عامة عن أداء النظام،لذلك نناقش أيضاً المقياس (F1-Score) لتأكيد النتائج والذي نعبر عنه بالعلاقة التالية[14]:

$$\boldsymbol{F1 = \frac{2\ X\ Precision\ X\ Recall}{Precision + Recall}} \qquad (4)$$

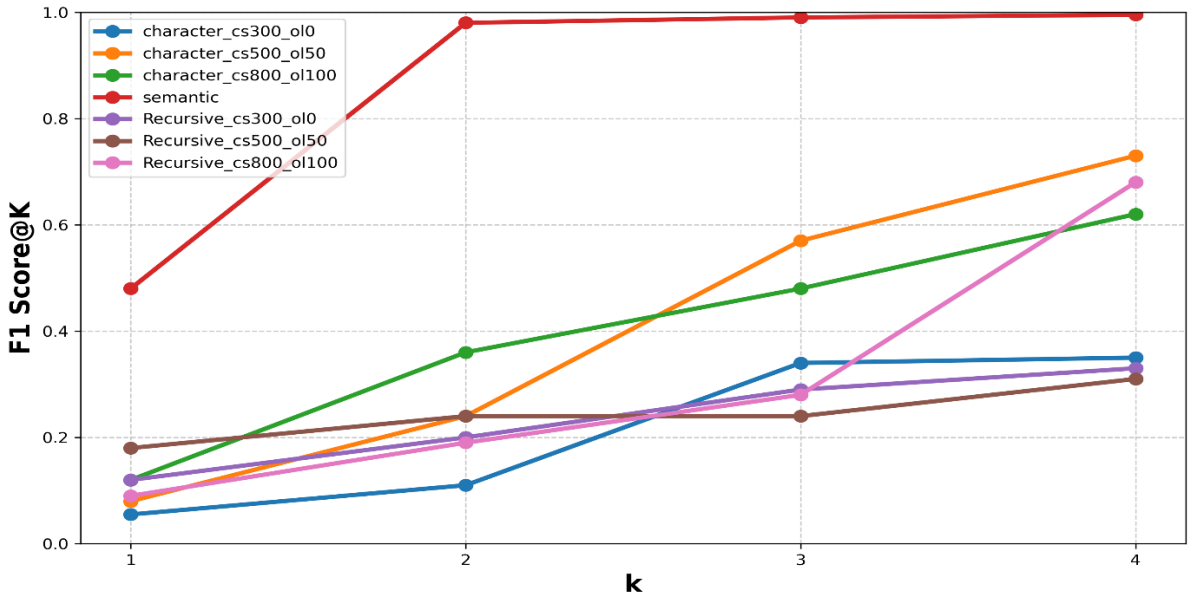
• **Recall@k:** يشير إلى نسبة المقاطع الصحيحة التي تم استرجاعها من بين جميع المقاطع الصحيحة الموجودة. كلما كانت هذه القيمة أكبر، زادت احتمالية استرجاع جميع المقاطع المهمة، وهو أمر حاسم لأن فقدان أي مقطع قد يؤدي إلى إجابة غير مكتملة أو خاطئة. ويُحسب Recall بالعلاقة التالية [14]:

$$\boldsymbol{Recall = \frac{TP}{TP + FN}} \qquad (5)$$

• **Precision@k**: : يشير إلى نسبة المقاطع الصحيحة بين جميع المقاطع المسترجعة. كلما كانت هذه القيمة أكبر، دل ذلك على أن أغلب النتائج التي تظهر للمستخدم في البداية دقيقة وذات صلة بالاستعلام. ويُحسب Precision بالعلاقة التالية (٦) [14].

$$\boldsymbol{Precision = \frac{TP}{TP + FP}} \qquad (6)$$

يوضح الشكل (7) المقارنة بين نوعي التقطيع بالمعاملات المختلفة من أجل قيم مختلفة ل k وذلك بالنسبة للمقياس F1-Score@k. تعكس قيم F1-Score نفس النتائج التي لوحظت في مقياس الدقة، حيث يتفوق التقطيع الدلالي بوضوح على الطرق التقليدية. يعود سبب هذا التفوق إلى أن التقطيع الدلالي يحافظ على استقلالية المقاطع وتماسكها الدلالي، أي أن كل مقطع يمثل وحدة مفهومية مكتملة وقريبة من معنى الاستعلام. كلما كانت المقاطع مستقلة ومتسقة دلالياً، كلما زادت فرص استرجاع المقطع الصحيح بدقة، مما يعكس نجاح نظام الاسترجاع في الوصول إلى المعلومات الأكثر صلة بالاستعلام.



الشكل (7): مقارنة بين نوعي التقطيع بالمعاملات المختلفة حتى ٤ نتائج (F1-Score@k)

٦-الاستنتاجات والتوصيات

- تم تطوير نظام استرجاع متكامل للمقالات العلمية العربية بالاعتماد على دراسة تحليلية وتجريبية شاملة لفعالية منهجيات التقطيع ضمن **سبع سيناريوهات مختلفة**.
- أظهرت النتائج تفوّق التقطيع الدلالي بوضوح على المناهج التقليدية، **إذ حقّق دقة ٥٠٪ عند k=1 مقابل أقل من ٢٠٪ للمناهج الأخرى**، وواصل الارتفاع مع قيم k الأكبر حتى بلغ مستويات شبه مثالية، **كما سجّل قيم F1 تراوحت بين ٠.٤٨ و٠.٩٥، وهي الأعلى مقارنةً بباقي المناهج التي لم تتجاوز ٠.٧٥ في أفضل الحالات.** مما يجعله الخيار الأكثر كفاءة في استرجاع المعلومات النصية العربية بعاً لبنية النصوص العلمية.
- أثبتت التجارب أن الاعتماد على التقطيع التقليدي فقط يؤدي إلى نتائج ضعيفة نسبياً، **ورغم أن زيادة حجم المقطع (من ٣٠٠ إلى ٨٠٠) والتداخل (من ٠ إلى ١٠٠) حسّنت الأداء جزئياً(من ٠.٠٦ إلى ٠.١٢ عند k=1)، إلا أنها لم تصل إلى المستوى الذي وفره التقطيع الدلالي.**
- عند السماح باسترجاع عدة نتائج، تحسن أداء الطرق التقليدية تدريجياً ليصل **إلى مستوى جيد عند k=4تجاوز ٠.٦ في بعض الحالات، بينما أظهر التقطيع الدلالي أداءً ممتازاً بالفعل عندk=2 (٠.٩٥)، مما يعكس قدرته على تحديد المقطع الصحيح بدقة أعلى وبكفاءة أكبر.**
- خلافاً للدراسات السابقة التي ركّزت إما على التقطيع الثابت فقط [٦] أو على التقطيع البنيوي في مجالات ضيقة كالمالية [٧] أو على تحليلات نظرية للتقطيع الدلالي دون تحقق تجريبي موسّع [٨]، فإن هذا البحث قدّم **أول دراسة تحليلية تجريبية متكاملة** تقارن بين التقطيع الحرفي، والتكراري، والدلالي في **النصوص الأكاديمية العربية**، بالاعتماد على مقالات علمية محكّمة وضمن سيناريوهات متعددة. هذا يعزّز موثوقية النتائج ويوضح بشكل فريد أن التقطيع الدلالي هو الخيار الأكثر كفاءة لاسترجاع المعلومات من النصوص العربية.

التوصيات

- دمج مرحلة التوليد (Generator) لبناء نظام RAG قادر على إنتاج إجابات دقيقة وموثوقة.
- تطوير واجهة تفاعلية تسهّل على الباحثين والطلاب إدخال المقالات وطرح الأسئلة دون الحاجة للمعرفة التقنية بالتقطيع.

المراجع

[1] Raza, M., Jahangir, Z., Riaz, M. B., Saeed, M. J., & Sattar, M. A. (2025). *Industrial applications of large language models.* Scientific Reports, 15(1), 13755.

[٢] Ambrosio, L., Schol, J., La Pietra, V. A., Russo, F., Vadalà, G., & Sakai, D. (2023). *Threats and opportunities of using ChatGPT in scientific writing-The risk of getting spineless*. JOR spine, 7(1), e1296. https://doi.org/10.1002/jsp2.1296

[٣] Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). *A SWOT analysis of ChatGPT: Implications for educational practice and research. Innovations in education and teaching international*, 61(3), 460-474.

[٤] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). *Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:*2312.10997, 2(1).

[5] Tartous University Journal - Series of Engineering Science, http://tartous-univ.edu.sy/mjlh-trtws/ar/get-contents

[6] Bhat, S. R., Rudat, M., Spiekermann, J., & Flores-Herr, N. (2025*). Rethinking Chunk Size For Long-Document Retrieval: A Multi-Dataset Analysis.* arXiv preprint arXiv:2505.21700.

[7] Yepes, A. J., You, Y., Milczek, J., Laverde, S., & Li, R. (2024). *Financial report chunking for effective retrieval augmented generation.* arXiv preprint arXiv:2402.05131.

[8] Qu, R., Tu, R., & Bao, F. (2024). *Is semantic chunking worth the computational cost*?. arXiv preprint arXiv:2410.13070.

[9] Jiang, M., Huang, T., Guo, B., Lu, Y., & Zhang, F. (2024). *Enhancing robustness in large language models: Prompting for mitigating the impact of irrelevant information*. arXiv preprint arXiv:2408.10615.

[10] Lyu, X., Grafberger, S., Biegel, S., Wei, S., Cao, M., Schelter, S., & Zhang, C. (2023*). Improving retrieval-augmented large language models via data importance learning*. arXiv preprint arXiv:2307.03027.

[11] Radim ˇRehurek, R. (2011). *Scalability of semantic analysis in natural language processing (Doctoral dissertation,* Masaryk University).

[12] Steck, H., Ekanadham, C., & Kallus, N. (2024, May*). Is cosine-similarity of embeddings really about similarity?*. In Companion Proceedings of the ACM Web Conference 2024 (pp. 887-890).

[13] Reimers, N., & Gurevych, I. (2019*). Sentence-bert: Sentence embeddings using siamese bert-network*s. arXiv preprint arXiv:1908.10084.

[14] Manning, C. D. (2008). *Introduction to information retrieval*. Syngress Publishing.