مساعد ذكي تعليمي

عمر ماريني , محمد سمير الأطرش , يزن منذر , عمار حسين

جامعة دمشق: كلية الهندسة المعلوماتية

1. مقدمة

تتمثل المسألة في الحاجة لتطوير نظام ذكي للإجابة على أسئلة الطلاب حول المحتوى التعليمي المتاح على المنصة التعليمية. يهدف النظام إلى تقديم إجابات دقيقة ومفيدة بناءً على مصادر تعليمية. لتحقيق هذا الهدف، تم تحليل وتقييم منهجيات مختلفة، بما في ذلك النماذج التقليدية والحديثة.

توصيف المسألة:

الإجابة على الأسئلة (Question Answering) هي إحدى مسائل معالجة اللغات الطبيعية التي تهدف إلى تمكين الأنظمة من فهم سؤال مكتوب بلغة طبيعية (Natural Language) وتقديم إجابة دقيقة وصحيحة بناءً على مصادر معطاة. يمكن لهذه المصادر أن تكون نصوصًا غير مهيكلة (unstructured data) (مثل مقاطع نصية من المقالات أو الكتب) أو قواعد بيانات منظمة (structured data) أو حتى مزيجًا من الاثنين.
تندرج هذه المسألة ضمن sequence-to-sequence tasks، حيث يتمثل الهدف في تحويل تسلسل من الكلمات (السؤال) إلى تسلسل آخر (الإجابة). لتحقيق ذلك، يجب أن تتضمن المنهجية خطوات تحليل وفهم عميق للسؤال، استخراج المعلومات ذات الصلة، ومن ثم صياغة الإجابة بلغة طبيعية مفهومة.

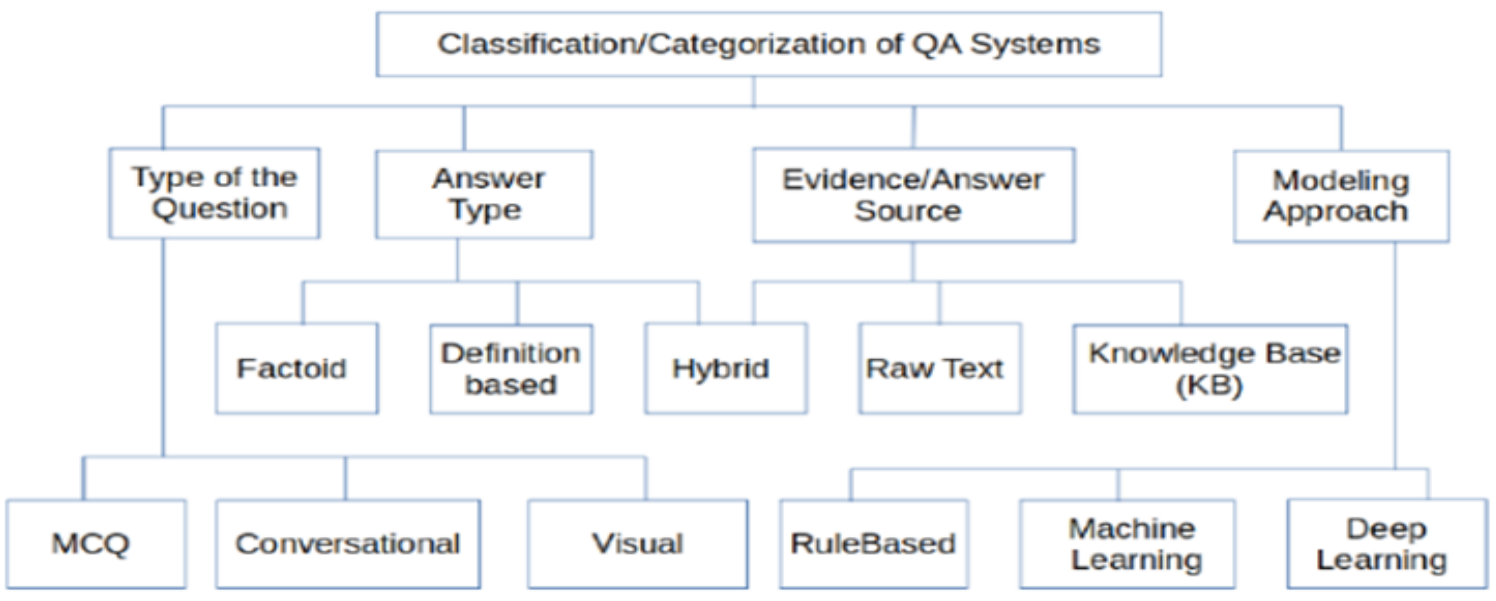
طبيعة الدخل والخرج :[1]

الدخل (Input): السؤال (Question): يتم تقديم السؤال بصيغة نصية مكتوبة بلغة طبيعية. قد يكون السؤال مفتوح النهاية (definition-based) مثل: "ما هو تعريف الذكاء الاصطناعي؟"
أو محددًا (Factoid) مثل: "ما هي عاصمة فرنسا؟". يمكن أن يتضمن السؤال تعبيرات معقدة، أوجه غموض، أو سياقات تحتاج إلى تفسير.
المصدر (Context): مقطع نصي أو مستند يحتوي على الإجابة المطلوبة (مثل فقرة من مقال أو مستند كامل).
قد تكون البيانات هيكلية، مثل قواعد بيانات أو معارف مرتبة (Knowledge Graphs).
في بعض الحالات، لا يتم تقديم مصدر بشكل مباشر، ويُتوقع من النموذج البحث عن الإجابة في مجموعة بيانات كبيرة أو عبر الإنترنت.

الخرج (Output): الإجابة (Answer): تُقدّم الإجابة بصيغة نصية قصيرة ومباشرة، مثل: "عاصمة فرنسا هي باريس."
في حال كان السؤال يتطلب شرحًا، يمكن أن تكون الإجابة أطول ومفصلة. و قد يُتوقع أحيانًا تقديم الإجابة في سياق نصي كامل للحفاظ على الوضوح.
قبل الخوض في مسألة question answering ستواجهنا مجموعة من الاعتبارات و الأسئلة البديهية بما في ذلك تصنيفات أنظمة Question Answering و طبيعة الأسئلة و الخرج المتوقع و قدرة النظام على فهم و تحليل السؤال و تقديم الإجابة المناسبة و معايير التقييم .

تصنيفات ال question answering systems :

ذكرت مجلة جامعة شيديان في دراسة استقصائية عن مسألة ال question answering في الصفحة 3:
" تتنوع أنظمة ال Question Answering لأسباب مثل نوع السؤال و نوع الإجابة المتوقعة و مصدر الإجابة و نهج نمذجة عملية تقديم الإجابة و ما إلى ذلك [1]. و هي مبينة كما في الشكل التالي :



كما ذكرت في الصفحة رقم 8:

بدءًا من نماذج مطابقة الكلمات (Word Matching Models) إلى النماذج الحديثة المعتمدة على المحولات (Transformers) مثل BERT وGPT، يمكن تصنيف النماذج بشكل عام إلى :

1- النماذج المعتمدة على القواعد (Rule-based Models):

في أغلب النهج التي تعتمد على القواعد، يتم تصميم قواعد مختلفة بناءً على نوع السؤال [WH Question Type] .
تلعب مهام معالجة اللغة مثل تصنيف الفئات الدلالية (Semantic Class Tagging) والتعرف على الكيانات (Entity Recognition) دورًا رئيسيًا في مثل هذه الأنظمة. و يمكن أن تكون القاعدة عبارة عن تركيبة منطقية لأي من المهام السابقة. تمنح كل قاعدة نقاطًا لجميع الجمل المدخلة [50].
بعد تطبيق جميع القواعد، يتم إرجاع الجملة التي تحصل على أعلى درجة كناتج للإجابة.
في العصر الحالي، تعتمد النماذج الموجهة للغات ذات الموارد المنخفضة بشكل أساسي على هذه القواعد [96]، بينما تحولت النماذج الأخرى إلى الأنظمة المعتمدة على التعلم الآلي أو التعلم العميق، حيث لوحظت تحسينات كبيرة في الأداء.

2- النماذج المعتمدة على التعلم الآلي (Machine Learning-based Models):

يتم تقديم نتائج تحليل السؤال إلى نماذج تعلم الآلة مثل Support Vector Machine (SVM) وDecision Tree (DT) و Naive Bayes (NB) لغرض التصنيف [48, 21, 60].
يمكن أن تكون مشكلات التنبؤ الأخرى باستخدام التعلم الآلي مثل التنبؤ بما إذا كان السؤال الذي يُطرح في مجتمع معين [73] سيحصل على إجابة أم لا، وذلك باستخدام مجموعات ميزات محددة مسبقًا [72, 69].

3- النماذج المعتمدة على التعلم العميق (Deep Learning-based Models):

- مع توفر قوة الحوسبة بشكل أكبر وتقديم نماذج الشبكات العصبية المتكررة (RNN) في مجال معالجة النصوص، تحوّل التقدم البحثي في الإجابة على الأسئلة من النماذج القائمة على التعلم الآلي إلى النماذج القائمة على التعلم العميق [11, 98, 87, 64].
- بالنسبة لمهام الإجابة على الأسئلة، يُعتبر تحسين النماذج المدربة مسبقًا على المحولات (Pretrained Transformer Models) مثل BERT من Google [23] أو GPT من OpenAI [10] هو الأفضل حاليًا (State-of-the-Art)."

تمهيد :
من الواضح أن الأسئلة في أغلب الأحيان لن تكون فقط factoid [1] , لذلك تكمن قدرة النموذج ضمن ما يسمى MC (machine comprehension) (يُستخدم مصطلح "فهم الآلة" (Machine Comprehension) لوصف قدرة الأنظمة على فهم النصوص والإجابة عن الأسئلة بناءً على ذلك)[2] في إيجاد الجواب بعد إعطائه السؤال و النص الخام (context) الذي قد يحوي غالباً جزء (passage) ذات صلة بالإجابة ,حيث يتم التوجه إلى تطوير نماذج للإجابة عن الأسئلة تعتمد على تعزيز المعرفة والتحقق من الإجابات لتحسين أداء الفهم [3]. سنستهل بحثنا هذا في ذكر أبرز تلك النماذج كما سنتطرق لتجريب و اختبار نموذج LLAMA3.1 [7] بهدف استخدامه لتحسين تقديم الإجابات عن الأسئلة التعليمية باستخدام المحتوى التعليمي.

2. منهجيات مرجعية

2.1 Attention-based BiLSTM

الوصف: يستخدم النموذج مزيجًا من CNN وRNN مع آلية انتباه لتحسين التمثيل السياقي للنصوص، مما يعزز مطابقة الأسئلة مع الإجابات.[4]
المزايا: تحسين التمثيلات السياقية باستخدام مزيج من CNN وRNN و آلية انتباه فعّالة لتحديد الأجزاء المهمة من النص و أداء متميز على مجموعتي InsuranceQA و TREC-QA.
القيود: تعقيد حسابي مرتفع و محدودية الأداء مع النصوص الطويلة أو الأسئلة المعقدة و حساسية لجودة البيانات المدخلة.
مجموعة البيانات: InsuranceQA، TREC-QA النتائج: حقق النموذج أداءً جيدًا على TREC-QA و InsuranceQA .

2.2 Retrieval-Augmented Generation (RAG)

- الوصف: يعتمد على الجمع بين استرجاع المعلومات باستخدام TF-IDF أو Dense Retrieval، وتوليد الإجابات باستخدام نماذج توليد النصوص (مثل BART)[5].
- مزاياه: أداء متميز في الأسئلة المفتوحة و القدرة على استرجاع المعلومات من مصادر خارجية.
- القيود: معقد في الإعداد و يتطلب بيانات كبيرة ومصادر معرفة.

مجموعة البيانات: Natural Questions (NQ)، TriviaQA، WebQuestions (WQ)، CuratedTREC

النتائج: Exact Match: سجلت النتائج تحسنًا في الإجابات الصحيحة مقارنة بالنماذج السابقة.

2.3 نموذج Transformer-based (BERT)

- الوصف: يستخدم لتحليل النصوص وتقديم إجابات قائمة على الفهم العميق للسياق باستخدام آلية Attention[6].
- مزاياه: دقة عالية في فهم النصوص القصيرة و قابل للتخصيص للمهام التعليمية.
- القيود: محدود في النصوص الطويلة أو الأسئلة متعددة السياقات.

مجموعة البيانات: GLUE، SQuAD v1.1، SQuAD v2.0، MultiNLI

النتائج: حقق BERT نتائج رائدة، حيث سجل:

- GLUE score: 80.5%
- SQuAD v1.1 F1 score: 93.2%
- SQuAD v2.0 F1 score: 83.1%
- MultiNLI accuracy: 86.7%

خلاصة : النماذج القائمة على المحولات (Transformer-based models)، بفضل بنيتها المبتكرة ومرونتها، تعتبر أقوى وأكثر ملاءمة لمهام الإجابة على الأسئلة مقارنة بالنماذج التقليدية. بالتالي إن LLAMA3.1 كونه (Transformer-based architecture) يمثل اختيارًا مثاليًا للإجابة على الأسئلة بفضل الأداء القوي، فهم السياق العميق، والمرونة العالية .

---

4 مجموعة البيانات

تم جمع البيانات المستخدمة في هذا البحث بعناية لضمان شموليتها وجودتها. تمثلت البيانات في مجموعة من الأسئلة والأجوبة (Q&A) الناتجة عن تحليل وتحويل مجموعة من الكورسات التعليمية. العملية تضمنت المراحل التالية:
تحويل الكورسات إلى صوت: تم أخذ مجموعة من الكورسات التعليمية وتحويل محتواها إلى ملفات صوتية.
تحويل الصوت إلى نص: باستخدام تقنية (Deepgram)، تم تحويل الصوت إلى نصوص دقيقة وموثوقة.
إنشاء أسئلة وأجوبة: تم تحليل النصوص المجمعة لاستخراج أسئلة وأجوبة (Q&A) باستخدام (Claude AI)

تحسين البيانات المجمعة:
لضمان تنظيم البيانات بشكل أفضل وتعزيز جودة النماذج المُعتمدة على هذه البيانات، تم إضافة حقل جديد لكل زوج من الأسئلة والأجوبة (Q&A)، وهو حقل يعبر عن الموضوع (Topic) هذا الحقل يساعد في تصنيف الأسئلة والأجوبة ضمن مواضيع محددة، مما يعزز من دقة النماذج القائمة على هذه البيانات ويساهم في تحسين قدرتها على الفهم والتصنيف.

المجالات المشمولة:
تشمل الكورسات مواضيع متعددة مثل مهارات الدراسة و الأدب و الصحة العامة والاعمال و تغيرات المناخ.
تنوع البيانات: تم تصميم الأسئلة لتغطي مستويات مختلفة، بدءًا من الأساسيات وحتى المفاهيم المتقدمة.

4.1 Exploratory data analysis:

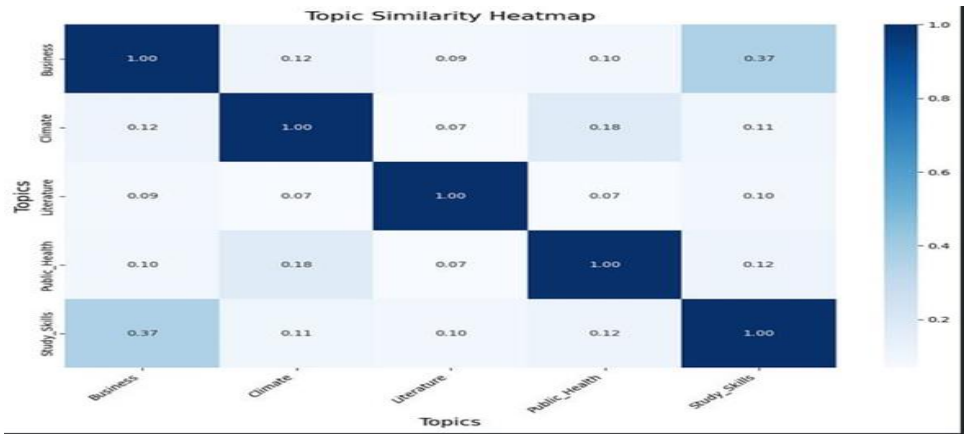
طبقنا بعض العمليات التحليلية على مجموعة البيانات قبل البدء في تنظيفها فكانت النتائج كالتالي:

4.1.1 Word Cloud

المخطط (1)

يظهر المخطط الكلمات الاكثر ورودا و هي الاكثر تأثيرا في البيانات

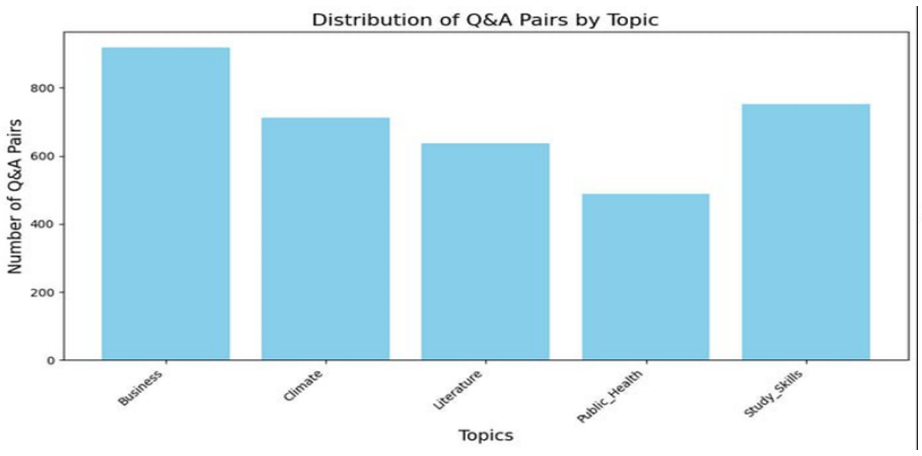
4.1.2 Heatmap



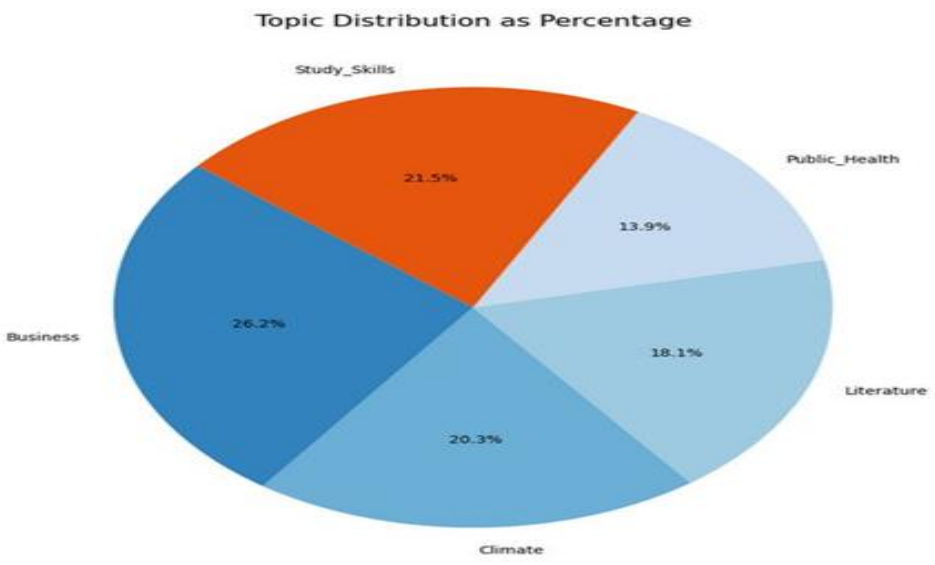
المخطط (2)

نلاحظ من المخطط وجود علاقة ضعيفة بين ال Topics في البيانات

Pie Chart & Histogram 4.1.3



المخطط (4)



المخطط (3)

يمثل المخططين عدد عينات البيانات لكل (Topic) نلاحظ وجود تفاوت في عدد العينات

4.2 تقسيم البيانات

تم تقسيم البيانات بعناية إلى ثلاث مجموعات رئيسية لضمان توزيع متوازن بين الموضوعات (Topics):

مجموعة التدريب 74% :(Training Set) من البيانات، لتدريب النموذج. مجموعة التحقق 13% :(Validation Set) من البيانات، لضبط أداء النموذج.

مجموعة الاختبار 13% :(Test Set) من البيانات، لتقييم النموذج.

منهجية التقسيم تقسيم طبقي (Stratified Splitting): لضمان توازن الموضوعات في كل مجموعة.

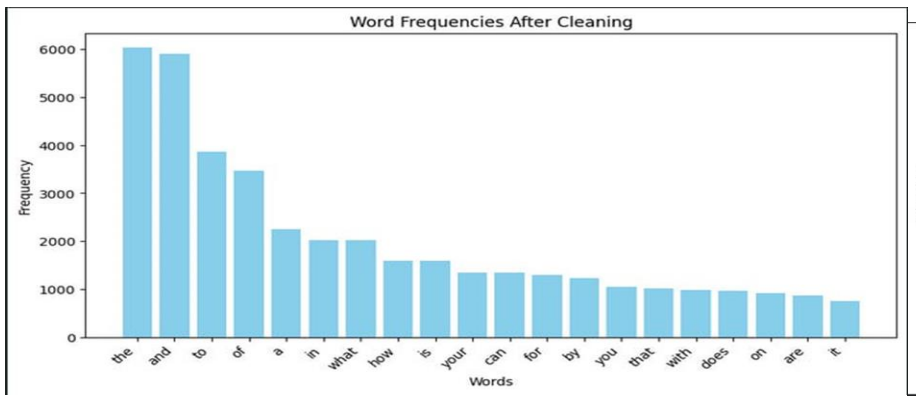
حيث تم تقسيم بيانات كل موضوع بشكل منفصل بنسبة متساوية. بعد ذلك، تم دمج البيانات الموزعة لتكوين المجموعات الثلاث.

أهمية التقسيم: يضمن هذا التقسيم تدريب النموذج على بيانات متنوعة من جميع الموضوعات واختباره على تمثيل شامل لها، مما يعزز دقة النموذج وأداءه.

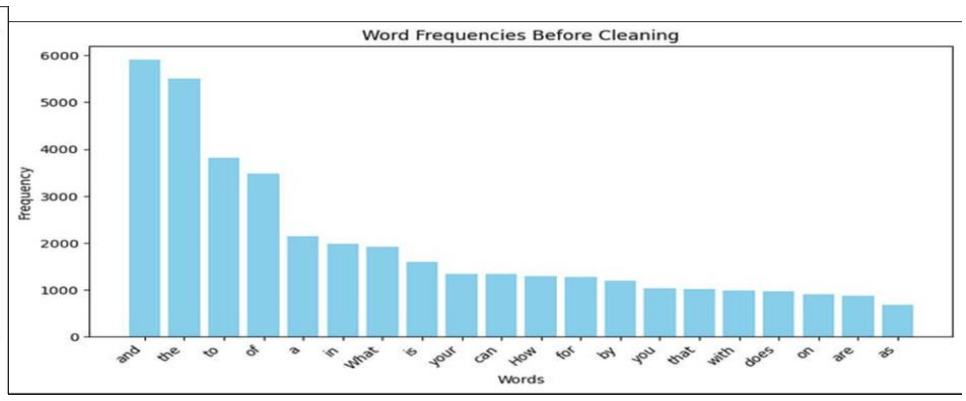
4.3 معالجة البيانات

اتبعنا العمليات التالية عند القيام بمعالجة وتنظيف البيانات:

- تحويل النص إلى أحرف صغيرة: نقوم بتحويل جميع الأحرف الكبيرة إلى أحرف صغيرة لضمان التجانس في تحليل البيانات.
- إزالة الرموز غير المرغوب فيها: نزيل جميع الرموز غير الحروف والأرقام والمسافات وعلامات الاستفهام والتعجب والنقاط والفواصل. يساهم هذا في إزالة أي ضوضاء قد تؤثر على دقة التحليل.
- التخلص من الأحرف المكررة أكثر من مرة واستبدالها بمحرف واحد.
- إزالة المسافات الزائدة: نقوم بتقسيم النص إلى كلمات منفصلة ثم نجمعها مجددًا مع مسافة واحدة بين كل كلمتين، وذلك للتخلص من أي مسافات زائدة أو غير ضرورية.



المخطط (6)



المخطط (5)

نلاحظ من الشكلين السابقين ان هناك اختلاف في بعض نسب ظهور الكلمات بعد اجراء عمليات التنظيف على البيانات

5 التجريب و النتائج :

تم استخدام نموذج LLAMA3.1 المدرب مسبقًا (Fine-Tuned) لتوليد إجابات على أسئلة تعليمية ضمن نطاق منصة تعليمية. تم مقارنة أداء النموذج مع نموذج مرجعي (Baseline Model) وهو SentenceTransformer ('all-MiniLM-L6-v2') [8] الذي يعتمد على تضمين النصوص وحساب التشابه بينها باستخدام قياسات مثل cosine similarity.

منهجية التقييم:

- النموذج المرجعي: يتم توليد الإجابات بواسطة النموذج المرجعي بناءً على مطابقة تضمينات الأسئلة مع تضمينات الأسئلة السابقة في مجموعة البيانات.

النموذج LLAMA3.1: تم تحسين النموذج باستخدام بيانات التدريب المخصصة للنظام التعليمي (Fine-Tuning)، حيث يتعامل النموذج مع النصوص ويولد إجابات مباشرة بناءً على السياق المقدم.

معايير التقييم المستخدمة:

- ROUGE: لتقييم مدى التشابه بين النصوص المتوقعة والنصوص المولدة.
- BLEU: لتقييم جودة الترجمة أو الإجابات بناءً على مقارنة تسلسلات الكلمات.
- BERTScore: لقياس الجودة الدلالية بين النصوص المولدة والنصوص المرجعية.

	Model_Name	Model_Epochs	Testing on	Preprocessing_Methods	Rouge1	Rouge2	RougeL	Bleu	Bert_Precision	Bert_Recall	Bert_f1
0	SentenceTransformer 'all-MiniLM-L6-v2'	default	Test Data	clean_text_function	0.348100	0.141500	0.271200	0.100100	0.894200	0.893200	0.893600
1	Fine-tuned Llama3.1-8B Model	60 Steps (0.2 epoch)	Test Data	clean_text_function	0.433200	0.157800	0.325100	0.074900	0.907500	0.903500	0.905400
2	Fine-tuned Llama3.1-8B Model	1 Full Epoch	Test Data	clean_text_function	0.458200	0.192300	0.357800	0.104500	0.895100	0.905300	0.900000
3	Fine-tuned Llama3.1-8B Model	1 Full Epoch	Train Data	clean_text_function	0.481800	0.207500	0.372800	0.111500	0.898400	0.911700	0.904900
4	Fine-tuned Llama3.1-8B Model	2 Full Epochs	Test Data	clean_text_function	0.461800	0.182900	0.355800	0.095400	0.890400	0.907600	0.898800
5	Fine-tuned Llama3.1-8B Model	2 Full Epochs	Train Data	clean_text_function	0.505900	0.266500	0.417600	0.160700	0.898100	0.916600	0.907200

النموذج LLAMA3.1 (Fine-Tuned):

- أظهر أداءً جيدًا من حيث BERTScore، مما يشير إلى جودة دلالية عالية للإجابات المولدة مقارنة بالنصوص المرجعية.
- قيم ROUGE-1 و ROUGE-L مقبولة، مما يعكس محدودية التغطية النصية للإجابات المولدة.
- BLEU Score كان منخفضًا نسبيًا، مما يدل على وجود اختلاف في تسلسل الكلمات بين الإجابات المرجعية والإجابات المولدة.

SentenceTransformer ('all-MiniLM-L6-v2'):

- أداء أقل ROUGE وBLEU، مما يشير إلى قدرة اقل على إنتاج نصوص مطابقة هيكليًا للنصوص المرجعية.
- يعاني من ضعف في الدلالات العميقة مقارنة بـ LLAMA3.1، حيث كان أداء BERTScore أقل.

مقارنة النماذج:

النموذج المرجعي كان مقبولا في النصوص التي تطابق النصوص المرجعية هيكليًا (Exact Matching).
LLAMA3.1 يتفوق دلاليًا، لكنه يحتاج إلى تحسين التغطية النصية للأجوبة.

الخلاصة:

LLAMA3.1 (Fine-Tuned):
نموذج قوي في الفهم الدلالي والإجابات السياقية، مما يجعله مثاليًا للاستخدام في التطبيقات التعليمية التي تتطلب إجابات دقيقة ومدعومة بالسياق. ومع ذلك، فإن الأداء في معايير مثل ROUGE وBLEU يظهر حاجة لتحسين التغطية النصية للإجابات.
SentenceTransformer ('all-MiniLM-L6-v2'):
خيار مرجعي مناسب لتقديم إجابات نصية مشابهة للنصوص المرجعية، لكنه يفتقر إلى العمق الدلالي الذي يميز LLAMA3.1.

[1] Question Answering Survey: Directions, Challenges, Datasets, Evaluation Matrices

[2] University of York / York, UK Question Answering Systems Approaches and Challenges. Reem Alqifari King Saud University /Riyadh, Saudi Arabia
[3] Question answering model based on machine reading comprehension with knowledge enhancement and answer verification. Ziming Yang, Yuxia Sun, Qingxuan Kuang

[4] Improved Representation Learning for Question Answer Matching. Ming Tan, Cicero dos Santos, Bing Xiang & Bowen

Zhou. IBM Watson Core Technologies ,Yorktown Heights, NY, USA.

[5] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela.

[6] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova.

[7] Jiayu Li1,3,*, Xuan Zhu2, Fang Liu2, Yanjun Qi2,3, (2024), AIDE: Task-Specific Fine Tuning with Attribute Guided MultI-Hop Data Expansion 1Syracuse University, 2AWS Bedrock Science

[8] Wenhui Wang Hangbo Bao Shaohan Huang Li Dong Furu Wei∗, (2020), MINILMv2: Multi-Head Self-Attention Relation for Compressing Pretrained Transformers Microsoft Research Distillation

[11] Chen, D., Fisch, A., Weston, J., and Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

[21] Cuteri, B., Reale, K., and Ricca, F. (2019). A logic-based question answering system for cultural heritage. In Calimeri, F., Leone, N., and Manna, M., editors, Logics in Artificial Intelligence, pages 526–541, Cham. Springer International Publishing.

[48] Joachims, T. (2002). Learning to classify text using support vector machines, volume 668. Springer Science & Business Media.

[50] Kahaduwa, H., Pathirana, D., Arachchi, P. L., Dias, V., Ranathunga, S., and Kohomban, U. (2017). Question answering system for the travel domain. In 2017 Moratuwa Engineering Research Conference (MERCon), pages 449–454.

[60] Lahbari, I., Alaoui, S. O. E., and Zidani, K. A. (2018). Toward a new arabic question answering system. Int. Arab J. Inf. Technol., 15:610–619.

[64] Lei, J., Yu, L., Bansal, M., and Berg, T. (2018). TVQA: Localized, com- positional video question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

[69] Liu, Q., Agichtein, E., Dror, G., Gabrilovich, E., Maarek, Y., Pelleg, D., and Szpektor, I. (2011). Predicting web searcher satisfaction with existing community-based answers. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pages 415–424.

[72] Molino, P., Aiello, L. M., and Lops, P. (2016). Social question answering: Textual, user, and network features for best answer prediction. ACM Transactions on Information Systems (TOIS), 35(1):1–40.

[73] Morris, M. R., Teevan, J., and Panovich, K. (2010). What do people ask their social networks, and why? a survey study of status message q&a behavior. In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 1739–1748.

[87] Reddy, S., Chen, D., and Manning, C. D. (2019). Coqa: A conversational question answering challenge. Transactions of the Association for Computational Linguistics, 7:249–266.

[96] Terdalkar, H. and Bhattacharya, A. (2019). Framework for question- answering in Sanskrit through automated construction of knowledge graphs. In Proceedings of the 6th International Sanskrit Computational Linguistics Symposium, pages 97–116, IIT Kharagpur, India. Association for Computational Linguistics

[98] Vakulenko, S., Fernandez Garcia, J. D., Polleres, A., de Rijke, M., and Cochez, M. (2019). Message passing for complex question answering over knowledge graphs. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM ’19, page 1431–1440, New York, NY, USA. Association for Computing Machinery.