# الجمهورية العربية السورية


**الجامعة السورية الخاصة**

**كلية الهندسة المعلوماتية**

**قسم الذكاء الصنعي وعلوم البيانات**


**عنوان المشروع :**

**التقييم الآلي لمقالات اللغة الإنجليزية**


**أعداد الطالب : ياسر الدغيم**

**اشراف الدكتورة : ماجدة البكور**

# المحتويات


الملخص : ..... ٣
الفصل الأول : ..... ٤
مقدمة : ..... ٥
1.1 دوافع العمل : ..... ٥
1.2 فكرة المشروع : ..... ٥
1.3 اهداف المشروع : ..... ٥
الفصل الثاني : ..... ٧
2 . الدراسة المرجعية : ..... ٨
الفصل الثالث : ..... ١٠
الدراسة التحليلية ..... ١١
3.1 المتطلبات الوظيفية : ..... ١١
3.2 المتطلبات غير الوظيفية : ..... ١١
الفصل الرابع : ..... ١٢
التجارب والاختبار ..... ١٣
4.1 نموذج تقييم المقال : ..... ١٣
الفصل الخامس : ..... ١٥
بنية النظام : ..... ١٥
5.1 الموارد المستخدمة: ..... ١٥
5.2 بيانات التدريب: ..... ١٦
5.3 معايير الأداء ..... ١٦
5.4 مرحلة معالجة النص الطبيعية : ..... ١٦
5.5يتألف النظام من: ..... ١٧
5.6 النتائج والمقارنة: ..... ١٧
الفصل السادس : ..... ١٩
6.1الخاتمة ..... ١٩
6.2الآفاق المستقبلية ..... ٢٠
6.3 المراجع ..... ٢٠

## الملخص :

يهدف هذا المشروع إلى بناء نظام آلي لتقييم مقالات اللغة الإنجليزية (Automated Essay Scoring) بالاعتماد على تقنيتين مختلفتين في معالجة اللغة الطبيعية. التقنية الأولى تعتمد على نموذج تعلم عميق تقليدي وهو LSTM باستخدام تمثيل الكلمات عبر Embedding Layer أو Word2Vec، بهدف التنبؤ بالدرجات الرقمية للمقالات. أما التقنية الثانية فتعتمد على نماذج اللغة الكبيرة (LLMs) مدعومة بأسلوب هندسة الموجهات (Prompt Engineering) لتوجيه النموذج نحو مهمة التقييم وفق معايير محددة. تم تصميم النظام بحيث ينتج تقييمًا متعدد الأبعاد يشمل عدة معايير خاصة بجودة المقال، ثم تتم مقارنة مخرجات النموذج مع الدرجات الحقيقية المعطاة من قبل مقيمين بشريين. ولقياس جودة التقييم ومدى توافقه مع الحكم البشري، تم استخدام مقياس كابا الموزونة تربيعيًا (Quadratic Weighted Kappa – QWK) باعتباره المقياس الأساسي في أنظمة تقييم المقالات.أظهرت النتائج أن نموذج LSTM قادر على تعلم أنماط لغوية أساسية في النصوص، بينما ساهم استخدام نماذج اللغة الكبيرة وهندسة الموجهات في تعزيز الفهم الدلالي للمقالات. ويبين المشروع أن الدمج بين الأساليب التقليدية والنماذج الحديثة يشكل إطارًا واعدًا لتطوير أنظمة تقييم آلي أكثر دقة وقربًا من التقييم البشري.

# الفصل الأول :

## مقدمة :

### 1.1 دوافع العمل :

تنبع دوافع هذا المشروع من الحاجة المتزايدة إلى تطوير أنظمة ذكية قادرة على تقييم مقالات اللغة الإنجليزية بشكل آلي وموضوعي، خاصة في البيئات التعليمية التي تتطلب تصحيح عدد كبير من المقالات خلال وقت قصير. يعتمد التقييم التقليدي على الجهد البشري، مما يجعله عرضة للتفاوت في الأحكام بين المصححين، إضافة إلى كونه عملية مكلفة زمنيًا وتتطلب موارد بشرية كبيرة.كما أن التطور السريع في مجال معالجة اللغة الطبيعية، وظهور نماذج اللغة الكبيرة (LLMs)، فتح آفاقًا جديدة لفهم النصوص وتحليلها على مستوى دلالي متقدم، وهو ما يشكل فرصة حقيقية لتحسين دقة أنظمة التقييم الآلي. في المقابل، ما تزال النماذج العميقة التقليدية مثل LSTM تمتلك أهمية كبيرة في التعلم من البيانات المصنفة والتنبؤ بالقيم العددية، خاصة عند توفر بيانات تقييم بشرية

لذلك، جاءت فكرة ه بناء نظام قابل للتطوير ليقترب من أسلوب التصحيح البشري من الرغبة في استكشاف إمكانية تقييم مقالات اللغة الإنجليزية باستخدام نماذج تعلم عميق تقليدية.ودراسة أثر نماذج اللغة الكبيرة وهندسة الموجهات في دعم عملية التقييم. بناء نظام قابل للتطوير يمكن تحسينه مستقبلًا ليقترب أكثر من أسلوب التصحيح البشري.

### 1.2 فكرة المشروع :

تقوم فكرة هذا المشروع على تطوير نظام آلي لتقييم مقالات اللغة الإنجليزية اعتمادًا على تقنيات معالجة اللغة الطبيعية والتعلم العميق، بهدف محاكاة عملية التقييم البشري وتقليل الاعتماد على التصحيح اليدوي. يعتمد النظام المقترح على دمج منهجين مختلفين في تقييم النصوص، هما النماذج العميقة التقليدية ونماذج اللغة الكبيرة، للاستفادة من نقاط القوة في كل منهما.في الجزء الأول من المشروع، يتم استخدام نموذج LSTM لمعالجة النصوص المقالية وتحويلها إلى تمثيل عددي قادر على التقاط الأنماط اللغوية والسياقية داخل المقال. يعمل هذا النموذج على التنبؤ بالدرجات الرقمية للمقالات بشكل مباشر، مع دعم الإخراج متعدد الأبعاد لتقييم أكثر من معيار في آن واحد.أما الجزء الثاني من المشروع، فيعتمد على نماذج اللغة الكبيرة (LLMs)، حيث يتم توظيف أسلوب هندسة الموجهات (Prompt Engineering) لتوجيه النموذج نحو مهمة تقييم المقالات وفق معايير محددة. تتيح هذه الطريقة الاستفادة من القدرة العالية لنماذج اللغة الكبيرة على فهم السياق والمعنى العميق للنصوص الإنجليزية.يتم في النهاية مقارنة مخرجات النظام مع الدرجات الحقيقية المعطاة من قبل المقيمين البشر باستخدام مقياس كابا الموزونة تربيعيًا (QWK)، بهدف قياس مدى توافق التقييم الآلي مع الحكم البشري. وتسعى فكرة المشروع إلى تقديم إطار عملي يجمع بين الأساليب التقليدية والحديثة في تقييم المقالات، مع إمكانية تطويره وتحسينه مستقبلًا.

### 1.3 اهداف المشروع :

يهدف هذا المشروع إلى بناء نظام آلي لتقييم مقالات اللغة الإنجليزية اعتمادًا على تقنيات معالجة اللغة الطبيعية والتعلم العميق، مع السعي إلى تقليل الفجوة بين التقييم الآلي والتقييم البشري. ويمكن تلخيص أهداف المشروع فيما يلي:

١. تصميم وتنفيذ نظام لتقييم مقالات اللغة الإنجليزية بشكل آلي، قادر على التعامل مع النصوص المقالية ذات الأطوال المختلفة.

٢. استخدام نموذج LSTM لتعلّم الأنماط اللغوية والسياقية في المقالات الإنجليزية والتنبؤ بالدرجات الرقمية للمقالات.

٣. توظيف نماذج اللغة الكبيرة (LLMs) في عملية التقييم للاستفادة من قدرتها العالية على فهم السياق والمعنى الدلالي للنصوص.

٤. تطبيق تقنيات هندسة الموجهات (Prompt Engineering) لتوجيه نماذج اللغة الكبيرة نحو مهمة التقييم وفق معايير محددة وواضحة.

٥. دعم الإخراج متعدد الأبعاد لتقييم عدة معايير خاصة بجودة المقال في آن واحد، بدلاً من الاكتفاء بدرجة واحدة عامة.

٦. تقييم أداء النظام باستخدام مقياس QWK لقياس مدى توافق مخرجات النموذج مع التقييمات البشرية، باعتباره مقياسًا مناسبًا لطبيعة التقييم الترتيبية.

٧. تحليل نتائج النموذج ومناقشة التحديات المرتبطة بتقييم المقالات آليًا، وبيان نقاط القوة والقصور في النهج المستخدم.

٨ تقديم إطار قابل للتطوير يمكن تحسينه مستقبلًا عبر تعديل بنية النماذج أو تحسين الموجهات أو توسيع مجموعة البيانات.

# الفصل الثاني :

## 2 . الدراسة المرجعية :

هدف الدراسة [1] تحسين دقّة أنظمة تقييم المقالات الآلية (AES) المعتمدة على النماذج اللغوية الكبيرة (LLMs) من خلال دمج الميزات اللغوية التقليدية (مثل عدد الكلمات الفريدة، طول الجمل، التعقيد اللغوي، عدد الأسماء، الكلمات الطويلة... إلخ) داخل الـ prompt المستخدم في النماذج . تم تقيم النظام على مجموعتي بيانات

ASAP Dataset: تحتوي على ~١٣ ألف مقال من طلاب الصفين ٧–١٠.

ELLIPSE Dataset: تحتوي على ~٦,٥ ألف مقال لطلاب ESL (الإنجليزية كلغة ثانية).

دمج الخصائص اللغوية أدى إلى تحسن ملحوظ في دقّة التقييم مقارنةً باستخدام LLM وحده.نموذج Mistral-7B مع جميع الخصائص اللغوية اقترب أداؤه من أداء GPT-4.نموذج BERT (المُدرّب إشرافياً) ما زال الأفضل أداءً، لكنه يتطلب موارد أكبر. تبين انإدراج الخصائص اللغوية داخل الـ prompt يُحسّن قدرة LLM على محاكاة تقييم الإنسان للمقالات.رغم التحسن، لا تزال هناك تحديات في تعميم النظام على أنواع مختلفة من المقالات.الدمج بين الطرق الإشرافية وLLMs يعدّ اتجاهاً واعداً لتطوير أنظمة تقييم أكثر دقة وشفافية. تتناولت الدراسة [2] مراجعة منهجية للأدبيات العلمية حول أنظمة تقييم المقالات الآلية (AES) خلال الفترة من ٢٠١٨ إلى ٢٠٢٣، بهدف تحديد مدى كفاءتها في البيئات التعليمية الواقعية. الهدف منها تحليل أنظمة تقييم المقالات الإنجليزية آلياً، واكتشاف نقاط القوة والضعف والقيود والاتجاهات البحثية الحديثة. حددت النماذج والتقنيات الأكثر استخداماً في البحوث الحديثة (٢٠١٨–٢٠٢٣) وهي تشمل أنظمة AES الحديثة على التعلم الآلي (ML) والتعلم العميق (DL) ومعالجة اللغات الطبيعية (NLP). ايضا النماذج العميقة مثل BERT وGPT حسّنت الدقة لكنها ما تزال تواجه صعوبات في التفسير ونقص البيانات. لكن منذ ظهور ChatGPT عام ٢٠٢٣، بدأ استخدام النماذج اللغوية الكبيرة في تقييم المقالات بشكل متزايد. بالنسبة لقاعدة البيانات كانت : ASAP (Kaggle) هي الأكثر استخداماً .يهدف البحث [3] الى تطوير نظام آلي لتقييم المقالات الإنجليزية بالاعتماد على تقنيات التعلم الآلي ( Machine Learning) ومعالجة اللغة الطبيعية (NLP). يهدف النظام إلى مساعدة المعلمين على تقليل الجهد والوقت اللازمين لتصحيح المقالات، وتقديم تقييمات دقيقة وموضوعية لأداء الطلاب في الكتابة.تم بناء النظام المقترح باستخدام خوارزميات تعلم الآلة مثل Random Forest (RF) وGradient Boosting Decision Tree (GBDT) وXGBoost، وتمت مقارنة أدائها من خلال معامل الارتباط لبيرسون (PCC) ومؤشر كابا الموزون (QWKV). أظهرت النتائج أن خوارزمية Random Forest تفوقت على النماذج الأخرى من حيث دقة التنبؤ بدرجات المقالات.كما تم استخدام نموذج BERT لتمثيل النصوص بطريقة تمكن من فهم المعنى والسياق بشكل أفضل، مما يساعد في تقدير مدى ترابط المقال مع الموضوع المطروح.تم اختبار النظام على ثلاث مجموعات بيانات: مجموعة ASAP (من مسابقة Kaggle) و بيانات طلاب الصف الرابع و بيانات موقع Critique.com . وأظهرت النتائج أن نموذج RF حقق أعلى معامل ارتباط (٠,٨٢٣) مقارنة بالنماذج الأخرى مثل BiLSTM وRNN، مما يدل على كفاءته العالية في تقييم المقالات الإنجليزية آلياً. يسعى البحث [4] إلى تحسين دقة نظم التقويم الآلي للمقالات (AES) عبر دمج طريقتين: تمثيلات Word2Vec مضبوطة المعاملات ( Hyper-parameter tuned) تُغذى إلى شبكة LSTM. نموذج لغوي كبير (LLM) من نوع Meta-Llama-3.2-1B-Instruct مُعدّل باستخدام مجموعة بيانات ASAP. النهج الأول: Word2Vec + LSTM تم استخدام خوارزمية Word2Vec بتمثيل متجهات من ٣٠٠ بُعد. تم ضبط معاملات مثل window size = 100، min word count = 4، epochs = 10.استخدمت شبكة LSTM بثلاث طبقات خفية، تفعيل ReLU، ودالة تحسين Adam.النهج الثاني: Fine-tuned Llama . النموذج الأساسي هو Meta-Llama-3.2-1B-Instruct.تم تدريبه على بيانات ASAP بعد تنظيف النصوص وتوحيدها. تم ضبط معاملات

التدريب (learning rate = 3e-5، epochs = 3، batch size = 2...).ة تفوق النموذج المقترح على نماذج متقدمة مثل: ChatGPT-4

الهدف الرئيسي للورقة [5] معالجة التحدي المتمثل في تطوير أنظمة تقييم تلقائية للكتابات (AES) قادرة على التعامل مع مواضيع ومقاييس تقييم متنوعة، من خلال الجمع بين كفاءة النماذج اللغوية الكبيرة (LLMs) في الاستدلال وقوة الأساليب القائمة على السمات اللغوية في الأداء.المشكلة هي طرق التقييم التلقائي التقليدية تنقسم إلى: الطرق القائمة على السمات الخاضعة للإشراف: تحقق أداءً عالياً ولكنها تتطلب تدريباً مكلفاً من حيث الموارد. الطرق القائمة على النماذج اللغوية الكبيرة: فعالة حسابياً أثناء الاستدلال ولكن أداؤها أقل.تم اقتراح طريقة هجينة تقوم على دمج السمات اللغوية في نمط التوجيه (Prompt) المقدم للنماذج اللغوية الكبيرة الطريقة الاولى بناء النمط الموجه: استخدام هيكل محدد يشمل: دور النموذج (باحث تعليمي)، موضوع الكتابة، مهمة التحليل، النص الطلابي، معلومات إضافية (السمات اللغوية). الطريقة الثانية السمات اللغوية: تم اختيار ١٠ سمات لغوية أظهرت دراسات سابقة ارتباطاً قوياً (≥٠,٦) بدرجة الكتابة، مثل: عدد الكلمات الفريدة، عدد الكلمات، عدد الجمل، عدد الأفعال الأساسية (lemmas)، عدد الأسماء، إلخ. الطريقة الثالثة النماذج المستخدمة: تمت التجربة على نموذج مفتوح المصدر (Mistral 7B) ونموذج مغلق المصدر (GPT-4). الطريقة الرابعة تم اعتماد مجموعة ASAP (شاملة لأنواع كتابة متنوعة) للتقييم الداخلي، ومجموعة ELLIPSE (مكتوبة من قبل متعلمي اللغة الإنجليزية) لاختبار القدرة على التعميم على بيانات خارج التوزيع. حقق النموذج الخاضع للإشراف (BERT) أعلى أداء، كما هو متوقع. أداء GPT-4 و Mistral كان متفاوتاً depending على نوع الكتابة ومقياس التقييم.

| رقم الدراسة | التقنيات المستخدمة | قاعدة البيانات | النتائج |
|---|---|---|---|
| [1] | LLMs , prompt ,BERT<br>Mistral-7B | ASAP<br>ELLIPSE | دمج الخصائص اللغوية أدى إلى تحسن ملحوظ في دقّة التقييم مقارنةً وحده.LLMباستخدام |
| [2] | NLP , DL , ML<br>BERTالنماذج العميقة مثل<br>GPTو | ASAP | ChatGPTمنذ ظهور عام ٢٠٢٣، بدأ استخدام النماذج اللغوية الكبيرة في تقييم المقالات بشكل متزايد |
| [3] | NLP , ML | بيانات طلاب الصف الرابع<br>بيانات موقع Critique.com<br>ASAP | حقق أعلى RFنموذج معامل ارتباط (٠,٨٢٣) مقارنة بالنماذج الأخرى مثل RNN وBiLSTM |
| [4] | LLM , LSTM | ASAP | LSTM + Word2Vec التقليدي<br>0.445<br>LSTM + Word2Vec مضبوطة<br>0.596<br>Fine-tuned Llam<br>0.72 |

| | | | |
|---|---|---|---|
| [5] | LLM , PROMPT | ASAP<br>ELLIPSE | BERT<br>0.545 |

# الفصل الثالث :

## الدراسة التحليلية

### 3.1 المتطلبات الوظيفية :

تم تحديد المتطلبات الوظيفية للنظام المقترح على النحو الآتي:

- إدخال المقالات النصية المكتوبة باللغة الإنجليزية إلى النظام بصيغة نصية.
- معالجة المقالات مسبقًا (تنظيف، تقسيم، وتحويل إلى تمثيل عددي).
- تقييم المقالات باستخدام نموذج LSTM لإنتاج درجات رقمية متوقعة.
- تقييم المقالات باستخدام نموذج لغة كبير (LLM) موجَّه عبر Prompt Engineering.
- دعم التقييم متعدد الأبعاد لإخراج أكثر من درجة لكل مقال.
- حفظ نتائج التقييم لكل مقال لإجراء التحليل والمقارنة.
- مقارنة مخرجات النظام مع الدرجات الحقيقية باستخدام مقياس QWK.
- عرض نتائج التقييم بشكل منظم وقابل للتحليل.

### 3.2 المتطلبات غير الوظيفية :

- الدقة: تحقيق أعلى توافق ممكن مع التقييم البشري.
- المرونة: إمكانية تعديل النموذج أو الموجهات بسهولة.
- قابلية التوسع: دعم تقييم عدد كبير من المقالات.
- قابلية الصيانة: سهولة فهم الكود وتطويره مستقبلًا.
- الكفاءة الزمنية: تنفيذ التقييم خلال وقت معقول.
- الموثوقية: ثبات النتائج عند تكرار التقييم.

# الفصل الرابع :

## التجارب والاختبار

### 4.1 نموذج تقييم المقال :

### النموذج الأول :

تقييم باستخدام نماذج اللغة الكبيرة (LLM)

للحصول على تنبؤات دقيقة لدرجات المقالات، استخدمنا نهج Few-shot prompting مع نماذج اللغة المتقدمة. نظرًا لأن النماذج اللغوية لا تستطيع معالجة النصوص بشكل مباشر، قمنا بتصميم نظام تقييم متعدد الأبعاد يعتمد على:

- مخطط التقييم: ٤ أبعاد أساسية (الأفكار، التنظيم، الأسلوب، الاتفاقيات)
- نظام العلامات: مقياس من ٠-٣ لكل بُعد
- أمثلة توضيحية: ٣ أمثلة تمهيدية لكل مستوى أداء

تم استخدام نموذج Nemotron-3-nano-30b-a3b عبر OpenRouter API، مع تخصيص prompt يحتوي على:

- تعليمات واضحة للمقيم
- أمثلة توضيحية لكل مستوى تقييمي
- تنسيق إخراج محدد بدقة

**نتائج التقييم:**

البعد QWK الدقة الملاحظات

الأفكار ٠,٠٠٠ - تحتاج تحسين

التنظيم ٠,٠٠٠ - تحتاج تحسين

الأسلوب ٠,٠٠٠ - تحتاج تحسين

الاتفاقيات ٠,٠٠٠ - تحتاج تحسين

ملاحظة: النتائج الأولية تشير إلى حاجة لتعديل نظام التقييم وزيادة الأمثلة التوضيحية.

### النموذج الثاني:

تقييم باستخدام التعلم العميق (LSTM + Word2Vec)

لتحقيق دقة أعلى في التقييم، قمنا ببناء نموذج تعلم عميق متعدد المهام. نظرًا لأن خوارزميات التعلم الآلي لا يمكنها معالجة النصوص بشكل مباشر، قمنا بتحويل المقالات إلى تمثيلات عددية من خلال:

١. استخراج السمات النصية:
   - تمثيل الكلمات باستخدام Word2Vec
   - معالجة النصوص: تحويل إلى أحرف صغيرة، إزالة stop words، tokenization
   - بناء vocabulary
٢. بناء النموذج:
   - طبقة Embedding باستخدام أوزان Word2Vec المُدرَّبة
   - طبقة LSTM (١٢٨ وحدة) لالتقاط السياق النصي
   - ٤ مخرجات مستقلة (Dense layers) لكل بُعد تقييمي
   - النموذج يحتوي على ٣,٣٦٠,٨٦٤ بارامتر (١٢,٨٢ MB)
٣. تدريب النموذج:
   - تقسيم البيانات: ٨٠% تدريب، ٢٠% اختبار
   - ١٠ epochs مع batch size = 32
   - استخدام MSE كدالة خسارة

· محسن Adam

نتائج الأداء:

```
Ideas: 0.4742
Organization: 0.3074
Style: 0.2895
Conventions: 0.2073
 Average QWK: 0.319582445224596
```

# الفصل الخامس :

بنية النظام :

5.1 الموارد المستخدمة:

الموارد العتادية:

- الحواسيب المستخدمة: حواسيب محمولة ومجهزة بنظام تشغيل Google Colab بيئة سحابية
- الذاكرة: تم استخدام الذاكرة السحابية المقدمة من Google Colab (~12.7 GB RAM)
- المعالج: وحدة معالجة مركزية سحابية (CPU)

· وحدة معالجة الرسوميات: تم استخدام وحدة معالجة رسوميات Tesla T4 مجانًا عبر Colab

الموارد البرمجية:

· اللغة الرئيسية: Python

· بيئة التطوير: Google Colab Notebook

### 5.2 بيانات التدريب:

تم استخدام مجموعة بيانات ASAP (Automated Student Assessment Prize) من منصة Kaggle، والتي تم إطلاقها عام ٢٠١٢ لتطوير أنظمة التقييم الآلي للمقالات. تحتوي المجموعة على مقالات طلابية مع تقييمات بشرية متعددة.

### 5.3 معايير الأداء

تم تقييم أداء النظام المُقترح باستخدام ثلاث مقاييس إحصائية رئيسية، وهي:

١. معامل كابا الرباعي المرجح (Quadratic Weighted Kappa – QWK)

يُعتبر QWK المقياس الرئيسي المستخدم في تقييم أنظمة التصحيح الآلي للمقالات، وهو قياس لمدى الاتفاق بين درجات النظام الآلي ودرجات المقيمين البشريين، مع مراعاة الفروق بين الدرجات بترجيح تربيعي.

يتم حساب QWK على النحو التالي:

· يتم إنشاء مصفوفة الارتباك O بأبعاد N \times N، حيث N هو عدد الدرجات الممكنة (في حالتنا من ٠ إلى ٣ لكل معيار: الأفكار، التنظيم، الأسلوب، الاتفاقيات).

· تمثل كل خلية O_{i,j} عدد المقالات التي حصلت على الدرجة i من المقيم البشري والدرجة j من النظام.

· تُحسب مصفوفة التوقع E بافتراض عدم وجود ارتباط بين التقييمين.

· أخيراً، يتم حساب QWK كالتالي:

تتراوح قيمة QWK بين -١ (اتفاق عكسي) و +١ (اتفاق تام).

· تشير القيمة ٠ إلى أن أداء النظام لا يختلف عن التوقع العشوائي.

· تشير القيمة ١ إلى تطابق تام مع التقييم البشري.

في هذا المشروع، تم حساب QWK لكل معيار على حدة (الأفكار، التنظيم، الأسلوب، الاتفاقيات) وكذلك كمتوسط عام للأداء.

٢. جذر متوسط الخطأ التربيعي (Root Mean Square Error – RMSE)

يُستخدم RMSE لقياس مقدار الخطأ في التنبؤ بالقيم العددية (الدرجات)، وهو يعطي وزنًا أكبر للأخطاء الكبيرة.

كلما اقتربت قيمة RMSE من الصفر، كان أداء النموذج أفضل.

٣. متوسط الخطأ المطلق (Mean Absolute Error – MAE)

يقيس MAE متوسط حجم الأخطاء في التنبؤ، دون اعتبار للاتجاه (موجب أو سالب).

يُعد MAE مقياسًا سهل التفسير لأنه يعطي الخطأ بنفس مقياس الدرجات الأصلية (من ٠ إلى ٣).

في هذا المشروع:

في الجزء الأول (نموذج LLM باستخدام OpenRouter)، تم استخدام QWK لمقارنة درجات النظام مع متوسط درجات المقيمين البشريين (rater1_domain1 و rater2_domain1).

في الجزء الثاني (نموذج LSTM + Word2Vec)، تم حساب QWK و RMSE و MAE لكل من المعايير الأربعة بشكل منفصل، وكذلك كمتوسط عام.

### 5.4 مرحلة معالجة النص الطبيعية :

قبل البدء في بناء النماذج التنبؤية، تم تنفيذ سلسلة من خطوات المعالجة الأولية للنصوص لاستخراج السمات اللغوية ذات الأهمية في تقييم المقالات.

خطوات المعالجة الأساسية:

١. تنظيف البيانات الأولية:

· تم التحقق من وجود قيم فارغة (NULL) في نص المقالات وأعمدة التقييم.

· تمت إزالة المقالات ذات البيانات الناقصة لضمان جودة مجموعة التدريب.

· تمت معالجة الرموز الخاصة والاختصارات (مثل @CAPS1، @NUM1) الموجودة في نصوص ASAP.

٢. التحليل اللغوي الأساسي:

· تجزئة النصوص (Tokenization): تقسيم كل مقال إلى كلمات فردية باستخدام nltk.word_tokenize.

· إزالة كلمات التوقف (Stop Words Removal): حذف الكلمات الشائعة غير الدلالية باستخدام قائمة stop words من مكتبة nltk.
· التصغير (Lowercasing): تحويل جميع النصوص إلى أحرف صغيرة لتوحيد المعالجة.

### 5.5 يتألف النظام من:

نموذجين متكاملين لتقييم المقالات تلقائيًا باستخدام تقنيات الذكاء الاصطناعي والتعلم الآلي.

#### النموذج الأول:

نظام التقييم باستخدام نماذج اللغة الكبيرة (LLM-Based)
· استخدام نموذج Nemotron-3 عبر واجهة OpenRouter API
· تطبيق تقنية Few-Shot Learning لتوجيه التقييم
· تقييم المقالات على ٤ أبعاد أساسية:
  · Ideas (جودة الأفكار والمحتوى)
  · Organization (التنظيم والترتيب المنطقي)
  · Style (الأسلوب اللغوي والتنوع)
  · Conventions (الدقة النحوية والإملائية)
· تقييم المقالات بناءً على معايير تصنيف من ٠ إلى ٣ لكل بُعد
· استخراج التبريرات التفصيلية لكل درجة ممنوحة

#### النموذج الثاني:

نظام التقييم باستخدام الشبكات العصبية (LSTM-Based)
· استخدام نموذج LSTM مع طبقات تضمين مسبقة (Word2Vec)
· بنية متعددة المهام (Multi-task) لتقييم الأبعاد الأربعة بشكل متزامن
· معالجة مسبقة للنصوص تشمل:
  · تحويل النص إلى أحرف صغيرة
  · إزالة علامات الترقيم والأرقام
  · تجريد الكلمات من التوقف (Stop Words Removal)
  · تقسيم النصوص إلى كلمات (Tokenization)
· تمثيل الكلمات باستخدام Word2Vec
· تسلسل النصوص إلى أطوال ثابتة باستخدام Padding

#### مصادر البيانات:

· استخدام ASAP Dataset، خاصة Set 7
مقالات بمتوسط طول متفاوت
· درجات تقييم من ٠-٣ لكل بُعد من أربعة قيمين (rater1_trait1 إلى rater1_trait4)

#### خوارزميات وتقنيات مستخدمة:

· LangChain لبناء أنظمة الـ Prompt
· OpenAI API (عبر OpenRouter) للتواصل مع النماذج الكبيرة
· TensorFlow/Keras لبناء الشبكات العصبية
· Gensim لتدريب Word2Vec
· Scikit-learn لقياس الأداء (QWK - Quadratic Weighted Kappa)
· Pandas/Numpy لمعالجة البيانات

### 5.6 النتائج والمقارنة:

من أجل تقييم أداء نظام تصنيف المقالات الآلي، تم تطبيق نهجين مختلفين ومقارنة أدائهما مع أنظمة تقييم معيارية. اعتمد النظام الأول على نماذج اللغة الكبيرة (LLM) والثاني على الشبكات العصبية (LSTM مع Word2Vec).

النهج الأول:

نظام التقييم باستخدام LLM

· استخدام نموذج Nemotron-3-nano-30b عبر OpenRouter API
· تطبيق تقنية Few-Shot Prompting مع ٣ أمثلة تدريبية
· تقييم المقالات على ٤ أبعاد (٠-٣ لكل بُعد)
· نتائج QWK الأولية: ٠,٠٠٠ لجميع الأبعاد
التحديات الملاحظة:
١. مشاكل في تكامل LangChain مع الإصدارات الحديثة
٢. صعوبة في استخراج النتائج بشكل منظم من استجابات النموذج
٣. حاجة لتحسين نظام الـ Prompt لضمان استجابة متسقة
٤. تكاليف API وتحديدات معدل الاستخدام
النهج الثاني:
نظام التقييم باستخدام LSTM + Word2Vec
· بناء نموذج LSTM
· تدريب Word2Vec على بيانات المقالات الخاصة بـ Set 7
· بنية متعددة المهام لتقييم الأبعاد الأربعة
· ١٠ عصور تدريب مع batch size = 32
· نتائج QWK: 0.000 لجميع الأبعاد

مقاييس الأداء أثناء التدريب:
· التدريب (Training Loss): ~2. - 2.2
· التحقق (Validation Loss): ~2.1 - 2.2
· الأفكار (Ideas Loss): 0.78 - 0.84
· التنظيم (Organization Loss): 0.50 - 0.63
· الأسلوب (Style Loss): 0.34 - 0.40
· فقدان الاتفاقيات (Conventions Loss): 0.50 - 0.53

# الفصل السادس :

6.1 الخاتمة

يقدم هذا المشروع إطارًا عمليًا لتقييم مقالات اللغة الإنجليزية باستخدام LSTM و LLM مع هندسة الموجهات.أظهرت النتائج أن الجمع بين التعلم العميق التقليدي وفهم السياق اللغوي عبر نماذج اللغة الكبيرة يعزز دقة التقييم ويقربه من التقييم البشري.
كما أظهر المشروع جدوى استخدام الإخراج متعدد الأبعاد لتقييم عدة معايير في آن واحد، وبيّن أهمية مقياس QWK كمعيار موثوق لقياس توافق النموذج مع المقيمين البشر.

## 6.2 الآفاق المستقبلية

يمكن تطوير هذا المشروع مستقبلًا من خلال عدة اتجاهات، منها:

١. زيادة حجم البيانات لتقوية قدرة LSTM على التعلم من الأنماط اللغوية المختلفة.
٢. تحسين هندسة الموجهات لنماذج اللغة الكبيرة لتحقيق فهم أعمق ودقة أعلى.
٣. استخدام نماذج متقدمة مثل BiLSTM أو Transformer لتقليل الأخطاء في التنبؤ بدرجات المقالات.
٤. توسيع نطاق التقييم ليشمل مقالات أطول ومعقدة أو مواضيع مختلفة لتعزيز قابلية النظام للتعميم.
٥. دمج التقييم التفسيري (Rationale) لتقديم تفسير نصي لكل درجة، مما يدعم عملية التعليم والتغذية الراجعة للطلاب.

## 6.3 المراجع


**[1] Improve LLM-based Automatic Essay Scoring with Linguistic Features**
**[2] Are Automated Essay Scoring Systems Competent in Real-Life Education Scenarios image**
**[3] Machine Learning-Based Automatic English Essay Scoring System**
**[4] Enhancing automated essay scoring by leveraging LSTM networks with hyper-parameter tuned word embeddings and fine-tuned LLMs )**
**[5] Improve LLM-based Automatic Essay Scoring with Linguistic Features**