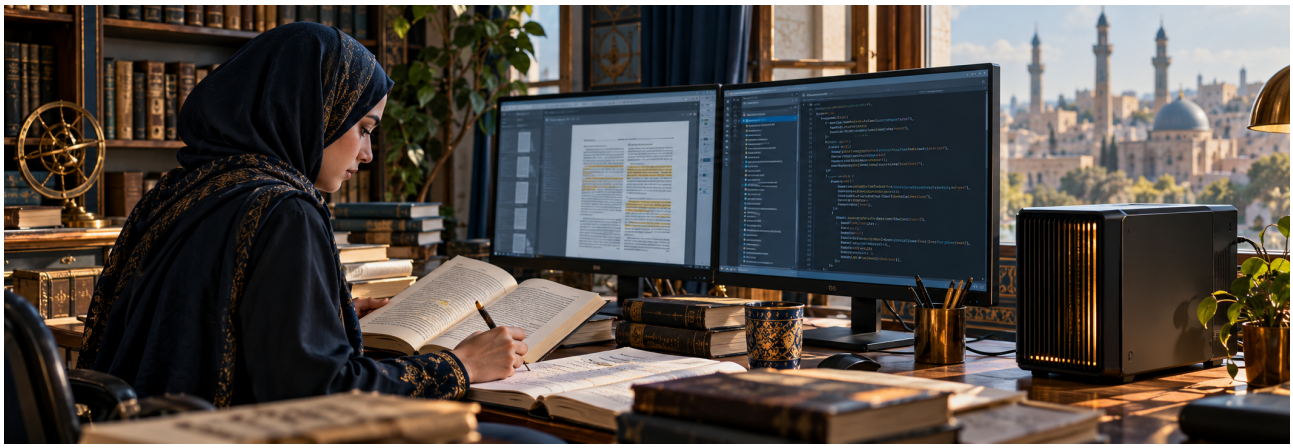
من الاستشهاد النصي إلى (BibTeX): نماذج لغوية تعمل محليًا

إيريث واي. نتزر



الملخص:

يعرض المقال تجربة مكتبات جامعة نورث وسترن في تحويل الاستشهادات النصية إلى بيانات (BibTeX) قابلة للمعالجة آليًا باستخدام نماذج لغوية تعمل محليًا. ويشرح جمع عينات من (Crossref) بصيغ استشهاد متنوعة، واختبار ثمانية نماذج عبر (Ollama)، ومقارنة سلامة المخرجات ومدى تطابق حقولها ووقت توليدها. وحقق النموذج الأكبر أعلى تطابق، بينما قدمت بعض النماذج الأصغر موازنة عملية بين الدقة والسرعة ومتطلبات الحوسبة. ويميز المقال بين صحة البنية النحوية وصحة البيانات، وبين الاختلاف النصي والخطأ الفعلي، ويناقش تحيز العينة وحدود المقارنة والاعتبارات البيئية. ويخلص إلى إمكان استخدام أدوات محلية صغيرة لتسريع إعداد بيانات النشر الجامعي، مع بقاء التحقق البشري ضروريًا وإتاحة الشيفرة والبيانات لدعم إعادة التجربة.

الكلمات المفتاحية: النماذج اللغوية الكبيرة، البيانات الببليوجرافية، تنظيم الاستشهادات، (BibTeX)، النشر الجامعي

خلفية ودافع:

استغرقت مكتبات جامعة نورث وسترن وقتًا وجهدًا كبيرين في تنظيم بيانات الاستشهادات يدويًا. وتبين هذه التجربة أن نماذج اللغة الكبيرة يمكن أن تحول الاستشهادات النصية إلى (BibTeX) بكفاءة، لإنتاج بيانات وصفية قابلة للمعالجة آليًا، وأنها تستطيع العمل محليًا دون نفقات حوسبة سحابية. وعند اقترانها بالمراجعة البشرية، يمكن أن تحسن كفاءة النشر الجامعي مع المحافظة على الجودة.

تنشر المكتبات دوريتين محكمتين: (The Bulletin of Applied Transgender Studies) (Studies in Russiang Philosophy, Literature, and Religious Thought). وتعمل تحت قيود مالية مباشرة وأخرى تتعلق بتكلفة الفرصة البديلة، لكنها تواجه مشكلات شبيهة بمشكلات الناشرين التجاريين بموارد أقل كثيرًا [1، 2].

ومن هذه المشكلات بيانات المراجع القابلة للمعالجة آليًا، التي تستخدم في عد الاستشهادات. ويساعد التقاطها وعدها واستخدامها بدقة جهات التمويل والجامعات والناشرين في قرارات توزيع الأموال، والمحكمين في التحقق من البحوث، والباحثين في تسريع مراجعة الأدبيات.

مثال من سير العمل:

نتلقى المخطوطات في ملفات (Microsoft Word)، ثم نستخدم (Pandoc) لتحويلها إلى (Markdown)، ومن مصدر واحد ننشئ نسختي (PDF) والويب [3]. لكن كتابة المؤلفين في (Word) تعني البحث يدويًا عن كل مرجع، وإضافته إلى (Zotero)، ثم تصديره إلى (BibTeX) لنسخة الويب [4]. ومع تولي شخص واحد عملية نشر الدوريات واتساعها وتعقدها، أصبح البحث عن طريقة أسرع ضرورة.

توجد مشروعات لتحويل النص العادي إلى (BibTeX) برمجيًا، لكنها قد تقتصر على لغات معينة [5] أو تعتمد على بيانات خارجية [6]. وتستشهد مخطوطاتنا كثيرًا بمصادر بلغات لا تستخدم الحروف اللاتينية، كما في دراسات الفلسفة الروسية، مما يقيد صلاحية عدد من تلك الأدوات [7].

بدأنا بتجربة واجهتي (GPT-3) و(GPT-3.5)، لكن دفع رسوم المعالجة يضعف قابلية توسيعها إلى دوريات كثيرة ومئات المراجع في العدد. ونفضل أدوات شفافة يمكن تكرار نتائجها؛ إذ قد يتغير الوصول إلى النماذج التي تديرها جهة خارجية، فنضطر إلى تبديل النظام. أما الأدوات المفتوحة فتسمح بتغيير الإصدارات حسب الحاجة دون رسوم خدمة مماثلة. ولهذا اتجهنا إلى (Ollama) [8].

القيود ومواطن القلق:

تقتصر بيانات التحليل على أعمال تظهر في واجهة (Crossref)، مما يخلق تحيزًا ضد الأعمال القديمة والكتب الأكاديمية. وتناسب النتائج بدرجة أكبر المجالات التي تركز استشهاداتها على أعمال حديثة، مثل العلوم الفيزيائية. وينبغي أن تتوسع الدراسات التالية إلى مراجع تاريخية وغير رقمية وغير أكاديمية.

وتثير نماذج اللغة مخاوف بيئية [9، 10]. وترى الكاتبة أن استخدام نماذج أصغر مثل المتاحة عبر (Ollama) قد يخفف هذه المخاوف. كذلك قد تولد النماذج بيانات مختلقة. لكن الجمع بين مدققات مثل (bibtexparser) والمراجعة البشرية منح فريقنا ثقة كافية لاستخدام النظام في إنتاج الدوريات [11]. ويظل بناء مسارات قابلة للتوسع والتحقق، تحتاج إلى إشراف بشري أقل، مهمة مستقبلية.

المنهجية:

جُمعت البيانات عبر (Crossref) باستخدام دالة العينات لاختيار (DOI) عشوائي، ثم تفاوض المحتوى لجلب استشهاد نصي بنمط يختار عشوائيًا من سبعة أنماط: (Chicago Author-Date) و(Elsevier-Harvard) و(Ecoscience) و(APA) و(MLA) و(IEEE) و(Council of Science Editors) [12].

وجُمعت لكل عمل أربعة متغيرات: (DOI)؛ والاستشهاد بصيغة (BibTeX)؛ والاستشهاد النصي؛ ونمط الاستشهاد النصي. وأنتج الاختيار العشوائي توزيعًا متقاربًا للأنماط.

التحليل:

اختُبرت النماذج الثمانية بأدوات (Ollama) على عنقود (Quest) في الجامعة، داخل حاوية (Singularity) [13، 14]. واستغرق الاختبار أربع عشرة ساعة على وحدتي رسوميات (NVIDIA A100)، وعقدة ذات ثماني نوى وذاكرة 128 غيغابايت . وكُتبت الشيفرة بايثون في بيئة (Anaconda) لتيسير إعادة تشغيل البيئة البرمجية [12].

استُخدم الاستشهاد النصي من (Crossref) مدخلًا، وبياناته مرجعًا للمقارنة، ووُجه إلى كل نموذج الطلب نفسه: «أنت محلل استشهادات محترف. حوّل الاستشهاد النصي المعطى إلى مدخل (BibTeX) منظم، مع الحقول المناسبة، مثل المؤلف والعنوان والدورية والمجلد والصفحات والسنة. أخرج مدخل (BibTeX) وحده دون شروح أو مقدمات أو نص إضافي».

وشملت النماذج (codegemma:2b) و(codegemma:7b) و(llama2:7b) و(llama3.3:70b) و(llama3:8b) و(mistral:7b) و(starcoder2:3b) و(tinyllama)، لتمثيل أحجام وطرائق تدريب مختلفة.

وحُفظت متغيرات النموذج، والاستشهاد النصي، ومدة التوليد، و(BibTeX) الفعلي من (Crossref)، وعدد الحقول المقارنة، وعدد الحقول المتطابقة، ونسبة التطابق التي تساوي عدد الحقول المتطابقة مقسومًا على إجمالي الحقول. ونتجت ثمانية ملفات (CSV) يضم كل منها نحو ثلاثة آلاف صف، يمثل كل صف (DOI) واحدًا، لتحليل الكفاءة والفاعلية.

فاعلية النماذج:

حقق (llama3.3:70b)، وهو الأكبر بين النماذج المختبَرة، أفضل أداء. ولم ينتج (starcoder2:3b) أي مدخل (BibTeX) صحيح البنية، بينما أنتجت بقية النماذج مداخل قابلة للتحليل لكل الاستشهادات.

لكن سلامة (BibTeX) هنا تعني إمكانية تحليله دون خطأ نحوي، ولا تعني صحة البيانات التي يتضمنها. وقد أنتج (llama3.3:70b) أكثر المداخل مطابقة للمرجع، لكن الاختلاف عن تمثيل (Crossref) لا يدل دائمًا على خطأ.

اقترب (Mistral) و(Codegemma) من أدائه مع عدد معلمات أقل وكلفة حوسبية أدنى. وقد يكون التنازل عن 0.2% من الدقة الإجمالية مقابل سرعة تقارب خمسة أمثالها مناسبًا لهذه الحالة.

كانت حقول المجلد والسنة والدورية الأكثر دقة، والمؤلف والناشر والجامعة الأقل. وتسمح الحقول الأخيرة بحرية أكبر في الصياغة، مما قد يجعل القيم الصحيحة مختلفة نصيًا. فـ(National Academy of Sciences, The) و(The National Academy of Sciences) يدلان على الجهة نفسها، لكن المقارنة النصية قد تعاملهما كاختلاف. ويجب تطوير مدقق للتكافؤ بين أسماء الأشخاص والناشرين والجامعات.

لذلك ينبغي فهم النتائج باعتبارها تطابقًا مع (Crossref)، لا حكمًا بأن كل حقل مختلف خطأ. وتفيد المقارنة في اختيار النموذج الملائم أكثر مما تفيد في الاستنتاج أن نصف الحقول غير صحيح.

إعادة التجربة على أجهزة شخصية:

استُخدم الحاسوب الفائق لتيسير المقارنة، لكن النماذج الأصغر من عشرة مليارات معلمة يمكن تشغيلها على أجهزة شخصية حديثة. وجُرّبت على جهاز ببطاقة (NVIDIA 3080 Ti) ومعالج (AMD Ryzen) ذي ست نوى وذاكرة 32 غيغابايت، وعلى (Apple M3 Max) بذاكرة 36 غيغابايت.

يعرض الأصل أمر تثبيت (Ollama) في أنظمة (UNIX) باستخدام (cURL)، ثم تنزيل نموذج (Mistral)، والتعامل مع خادم محلي يستمع إلى المنفذ 11434. ويمكن إرسال المدخلات إلى الدالة للحصول على استجابة (BibTeX)، وتوجد الشيفرة الكاملة في مستودع الكاتبة.

خاتمة وإتاحة البيانات:

يمكن للناشرين الجامعيين استخدام نماذج صغيرة تعمل محليًا لإنشاء (BibTeX) سليم البنية وأتمتة خطوة من إعداد بيانات الاستشهادات. وبحسب التجربة وقت نشرها، توفر نماذج سبعة مليارات معلمة، خصوصًا (Mistral)، أداءً مقبولًا على حواسيب محمولة حديثة بتكلفة منخفضة، مع بقاء التحقق البشري ضروريًا.

تسعى الكاتبة إلى الالتزام بمبادئ (FAIR) لإتاحة البيانات وإعادة استخدامها؛ وتتوافر بيانات التحليل والشيفرة في (GitHub).

جدول مدة التوليد:

النموذج	وسيط الزمن بالثانية	الانحراف المعياري بالثانية
(codegemma:2b)	1.18	3.24
(codegemma:7b)	1.08	0.29
(llama2:7b)	0.96	0.27
(llama3.3:70b)	5.45	1.90
(llama3:8b)	1.00	0.31
(mistral)	1.00	0.27
(starcoder2:3b)	0.00	0.00
(tinyllama)	0.57	0.64

القيمة الصفرية لـ(starcoder2) لا تدل على سرعة إنتاج ناجح؛ فقد فشل في توليد مداخل صالحة في الاختبار.

الأوامر والدالة الواردة في الأصل:

```
curl -fsSL https://ollama.com/install.sh | sh
ollama pull mistral
```

```
def generate_text_with_ollama(model_name, prompt):
    url = 'http://localhost:11434/api/generate'
    payload = {
        "model": model_name,
        "prompt": prompt,
        "temperature": 0,
        "stop": ["\n\n"]
    }
    headers = {"Content-Type": "application/json"}
    response = requests.post(url, headers=headers,
                             data=json.dumps(payload), stream=True)
    generated_text = ''
```

```
    for line in response.iter_lines():
        if line:
            data = json.loads(line)
            if data.get('done', False):
                break
            else:
                generated_text += data.get('response', '')
    return generated_text.strip()
```

المراجع:

1. ACRL (2024). The State of U.S. Academic Libraries: Findings from the ACRL 2023 Annual Survey.
2. RELX (2023). Market Segments.
3. Pandoc، 2024 مارس 8 تاريخ الاطلاع.
4. Zotero: Your Personal Research Assistant، 2024 ديسمبر 31 تاريخ الاطلاع.
5. Makino Takaki. Generate BiBTeX Entry from Plain Text. 2024 مارس 9 النص، تاريخ الاطلاع.
6. Text2bib، 2024 مارس 9 تاريخ الاطلاع.
7. Williams, R. (2024). Sergeii Bulgakov, Socialism, and the Church. Northwestern University Studies in Russian Philosophy, Literature, and Religious Thought.
8. Ollama (2025). Ollama/Ollama.
9. Ding, Y., and Shi, T. (2024). Sustainable LLM Serving: Environmental Implications, Challenges, and Opportunities. IGSC, 37–38. DOI:10.1109/IGSC64514.2024.00016.
10. Chien, A. A., et al. (2023). Reducing the Carbon Impact of Generative AI Inference (Today and in 2035). Proceedings of the 2nd Workshop on Sustainable Computer Systems, 1–7. DOI: 10.1145/3604930.3605705.
11. About the Journal – Bulletin of Applied Transgender Studies، 2024 مارس 8 تاريخ الاطلاع.
12. Netzer, A. (2025). Biblatex-Transformer: الشيفرة والبيانات.
13. Northwestern University. Quest High-Performance Computing Cluster، 2025 يناير 1 تاريخ الاطلاع.
14. Kurtzer, G. M., Sochat, V., and Bauer, M. W. (2017). Singularity: Scientific Containers for Mobility of Compute. PLOS ONE, 12(5), e0177459. DOI:10.1371/journal.pone.0177459.