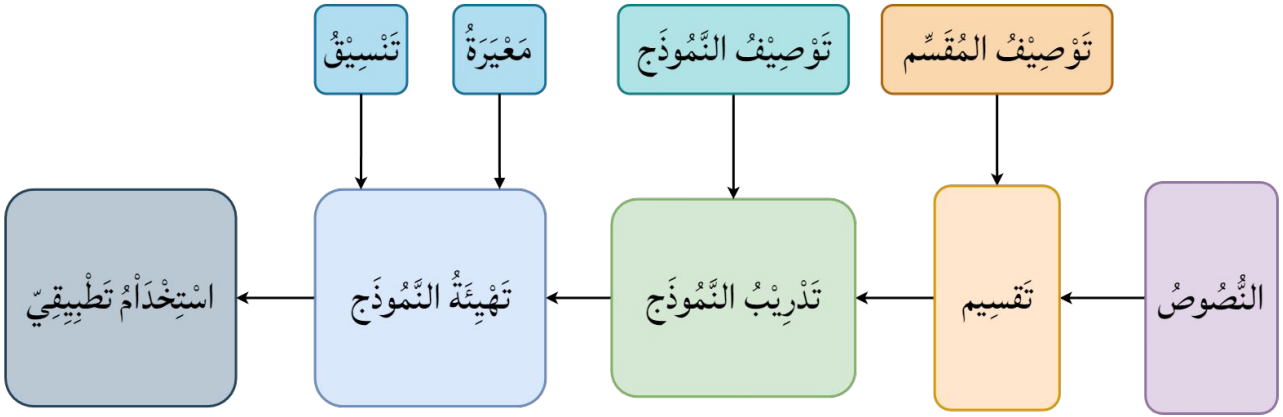
مقدمة لنماذج المحوِّلات

سري السباعي

فهرس العرض

مقدمة عن العرض

المقصد من هذا العرض تصوير أُسُس النماذج اللغوية من أصلها إلى النماذج الحديثة مع ذكرِ أشهر النماذج والمدوَّنات العربية. وتبيين أهم المصطلحات في المجال وترجمتها. وكذا الإحالة على مصادر مفيدة لمن أراد التوسّع فما مجلِسُنا هذا إلا قطرةً من يَمّ وقليلًا من جم.



أهم المصطلحات المُستعملة

Transformers	المحولات	Language model	نموذج لغوي
Decoder	مفكك	FineTune	معيرة (ضبط دقيق)
Encoder	مشفّر	Understand	فهم
Corpus\Dataset	مدونة (المُدخلات)	Generative	توليد
Tokenizer	مُقسِّم (مُجَزِّء)	Token	كُليمة (جزء لغوي)

أصل العرض

المحولات التوليدية العربية بنظرة تقنية رياضية بحثيّة مُختصرة

سري بن تيسير السباعي

مهندس باحث في معمل الروبوتات وإنترنت الأشياء في جامعة الأمير سلطان في السعودية

مقدمة

الحمدُ لله وصلاة وسلاما على رسولِ الله أما بعدُ: فهذه ورقة معتصرة ألخِّص فيها أسسَ نماذجِ المحولاتِ المولِّدةِ فاكَّةِ التشفير[1] مُخصِّصًا منها ما كان متعلقا بالعربية. قسّمتها على ثلاث مقالات كلُّ مقالةٍ فصولٌ وكلُّ فصلٍ أبوابٌ وقبلهنَّ تقدمةٌ في ترجمة المصطلحات المُسْتعملة في الورقة.

أما تفصيل المقالات الثلاث:

فأولهن: في الكلام عن المدوّنات النصية العربية أنواعِها وطرقِ استنصاصها[2] وتجميعها ثم نتطرّق إلى طرق تمثيلها الرقمية للنموذج،

والثانية: نَصِفُ فيها التركيبَ الرياضيَّ وهيكلةَ النموذجِ المُحوّل فاكّ التشفير وأنواع بناءِ مكوّناتِه واختلافَها وألمح إلى مبحث سبب تعلُّم هذه النماذج، مختتمًا بذكرِ أسماء أشهرِ النَّماذج اللغوية العربية.

والثالثة في مَعْيَرة النماذج اللغوية الكبيرة واستعمالها على بيانات مخصوصةٍ والتطبيقات المُختلفة لذلك.

ويتخلل الفصول تنبيهات وفوائد حسب المناسب ان شاء الله.

واعلم أن هذه الورقة لم تُبيَّت فيها نية الاستقصاء ولا همة الجمع (فهذا يحتاج كتابا أو أكثر) وإنما غايتنا التنبيه والإلماح لعدد كبيرٍ من المصادر والمجالات والأفكار البحثية وربطها بالعربية نسأل الله العون وحسن الاختيار وصلى الله وسلم على نبيِّنا محمد وعلى آله وصحبه أجمعين.

وما من كاتبٍ إلا سيفنى ... ويُبقي الدهر ما كتبت يداه

فلا تكتب بكفِّك غير شيء ... يسرُّك في القيامة أن تراه[3]

نظرةٌ عامّة عن النماذج اللغوية

طريقة ومنهج لتمثيل اللغة حوسبيًّا إما **لتوليد أو فهم** بعضِ **خصائصها اللغوية** حسْبَ مرتبتها (بدايةً من التحليل النُّطقي وصولًا إلى التحليل المُعجمي الشامل الفهم الدلاليَّ).

فمن الأمثلة على الفهم المَبني على قواعد ثابتة نظام دردشَة "إلِزَا" (الخمسينات) وكذا أنظمة فهم أقسام الكلام العربي وتصريفها

وأما أمثلة التوليد فلَم تظهَر إلا بعد طفرة **التعلّم الإحصائي** بأن يتوقَّع النموذج احتمال الكُلَيمة التالية بناءً على سياق محدد

https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

أمثلة على الطُّرُق القديمة

باستعمال "الأنماط الطبيعية" REGEX

البرمجة المُباشرة

البحث بالكلمات

معجم تركيبي للتوليد والتحليل الصرفي والنحوي

التحليل الصرفي للجذور والجذوع

المُتَّجهات العَدِّيَّة

المُتْجِهات الكلماتية

الشبكات التكراريّة

الشبكات العصبيّة الصِّرفَة

أهمّ المحطات التاريخية باختصار

1997 الشبكات التكرارية
RNN

2013 الكلمجات
word2vec

2017 المحولات

2022 متمم-3

متمم

أنواع المحولات الثلاثة

متفهم (المُشَفِّر)

مُترَجِم (مُشَفِّر فاكّ)

متمم (مفكك)

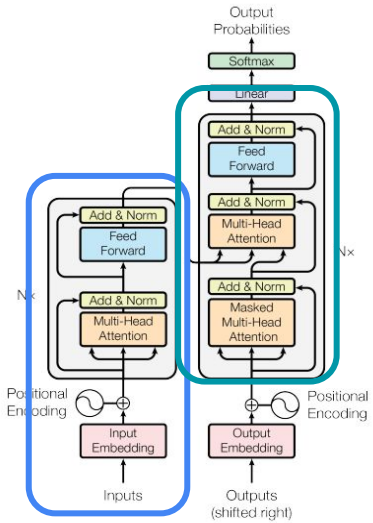
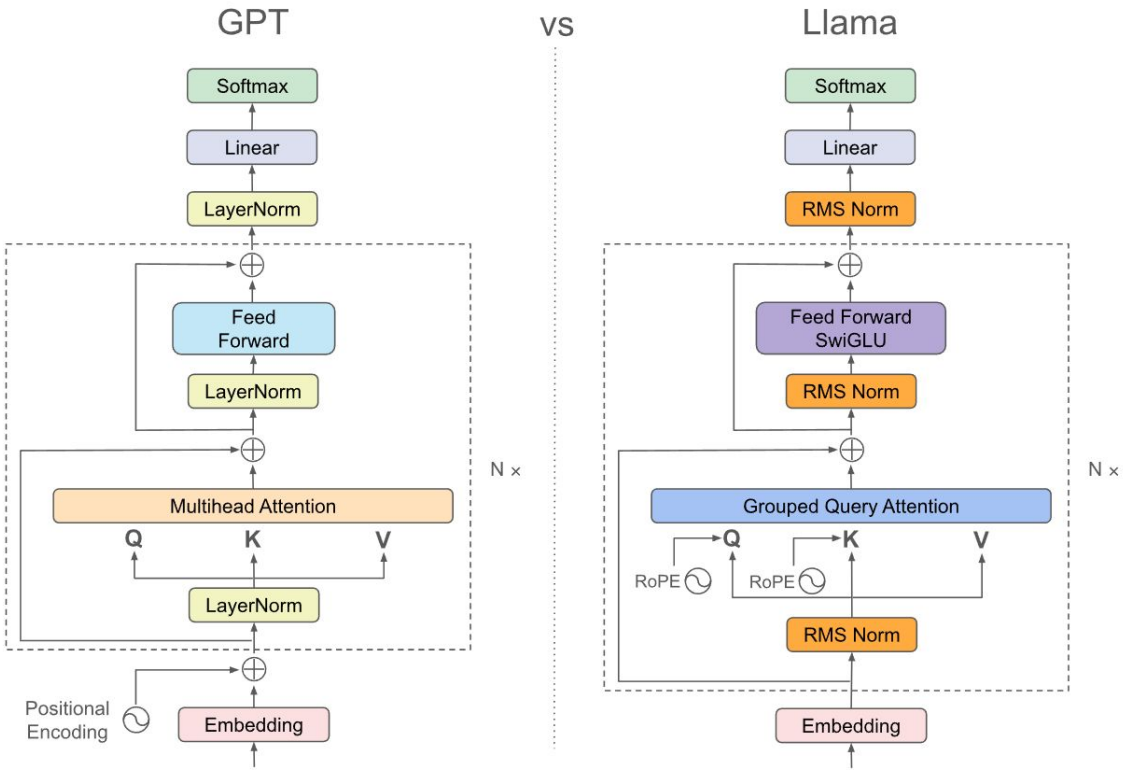


Figure 1: The Transformer - model architecture.

https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

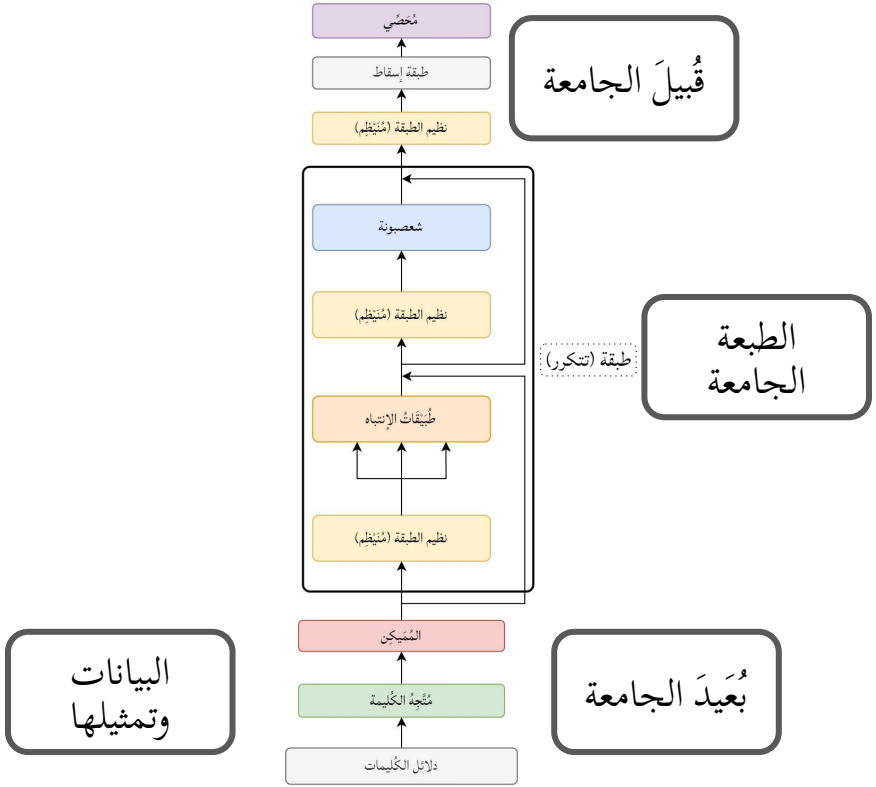
أشهر المفكّكات



- Jais
- AYA
- ALLAM
- QWEN
- ArabianGPT
- AraGPT

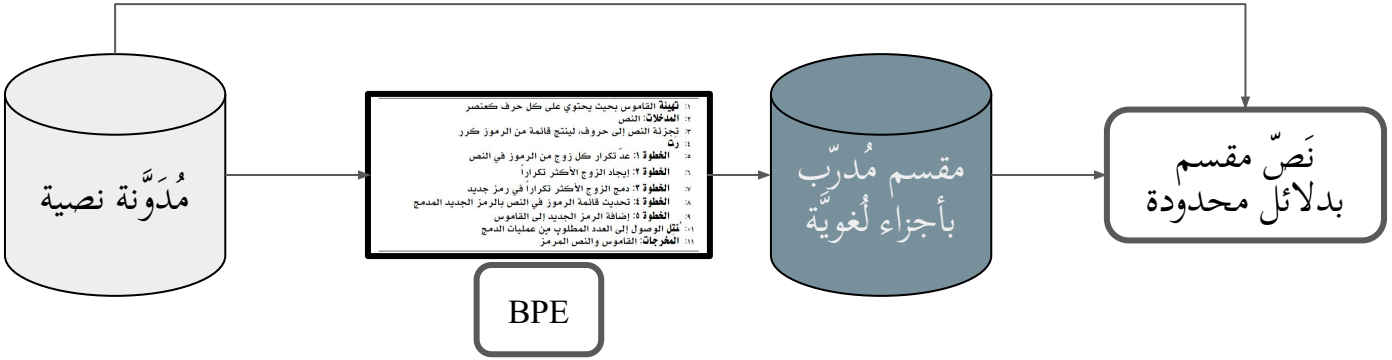
https://docs.nvidia.com/deeplearning/transformer-engine/user-guide/examples/te_llama/tutorial_accelerate_hf_llama_with_te.html

نموذج المحوّل الفاكّ



البيانات وتمثيلها

أشهر تمثيل للنصوص المكتوبة يكون باستعمال خوارزمية **التزويج البايتِيُّ** وعليها بُنيَت الباقيات وغالب النماذج العربية استعملتها



مدونة CaltureX	400 جيجا
مدونة ArabicText2022	200 جيجا
مدونة 101 مليار	-

قُبَيل الجامعة

$$E = \text{Embedding}(X) \quad E \in \mathbb{R}^{T \times d_{model}}$$

$$PE = \text{PositionalEncoding}(T, d_{model})$$

$$Z_0 = E + PE \quad Z_0 \in \mathbb{R}^{T \times d_{model}}$$

الطبقة الجامعة

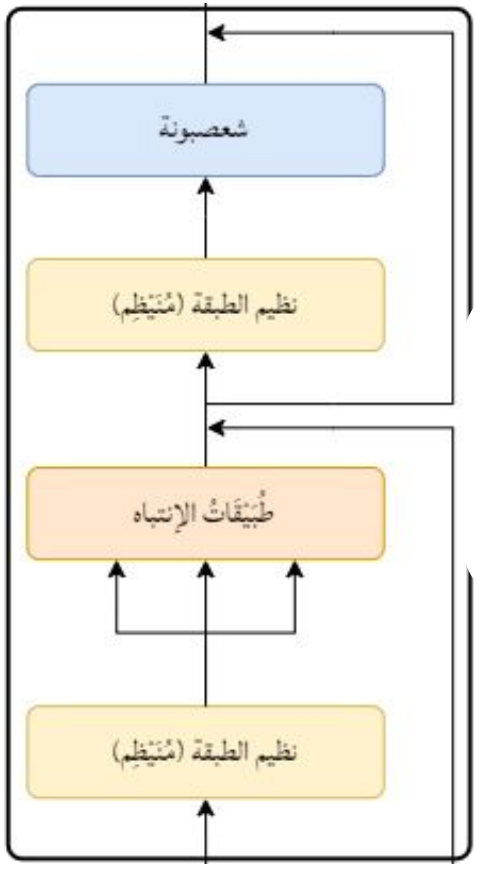
$$F_1 = \hat{Z}_1 W_1 \quad F_1 \in \mathbb{R}^{T \times d_{ff}}$$

$$F_2 = f(F_1)$$

$$Z_1 = Z_0 + \text{MultiHead}(Q, K, V)$$

$$\hat{Z}_1 = \text{LayerNorm}(Z_1) \quad \hat{Z}_1 \in \mathbb{R}^{T \times d_{model}}$$

$$\hat{Z}_0 = \text{LayerNorm}(Z_0) \quad \hat{Z}_0 \in \mathbb{R}^{T \times d_{model}}$$



$$Q = \hat{Z}_0 W_Q \quad K = \hat{Z}_0 W_K \quad V = \hat{Z}_0 W_V$$

$$Q, K, V \in \mathbb{R}^{T \times d_k}$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right) V \quad M \in \mathbb{R}^{T \times T}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W_O$$

$$W_O \in \mathbb{R}^{(h \times d_k) \times d_{model}}$$

بُعَيدة الجامعة

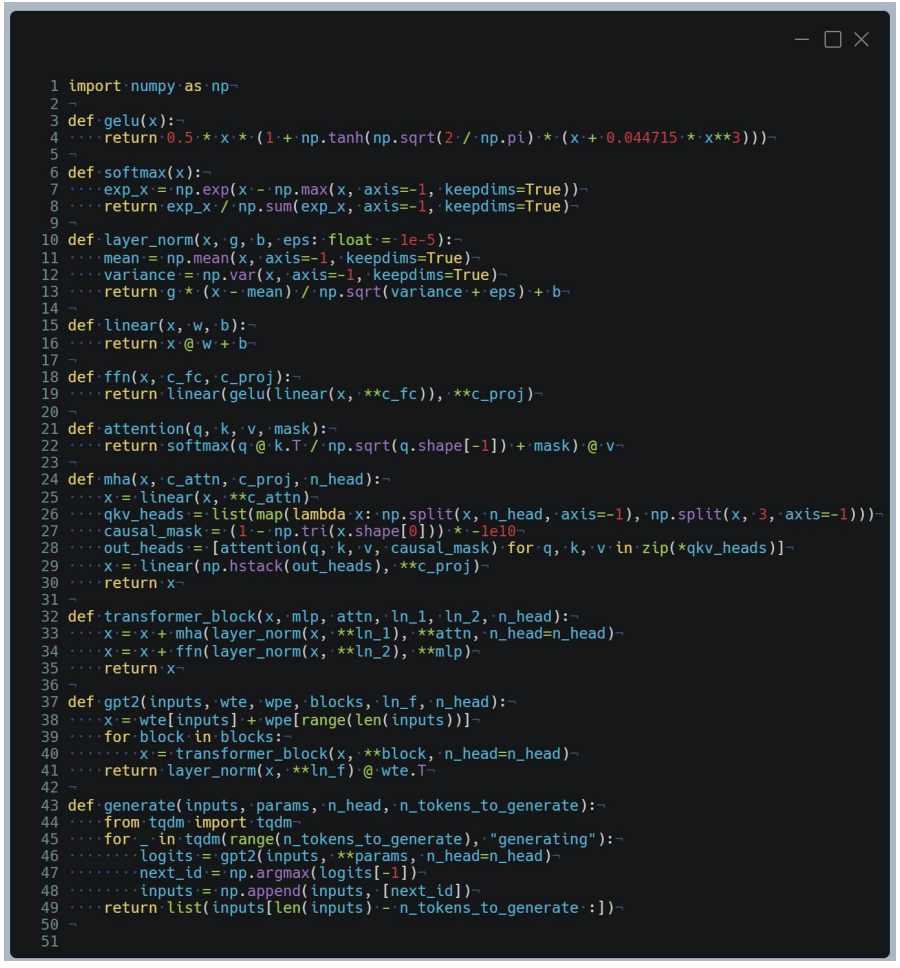
$$\mathcal{L}(\hat{y}, y) = \text{CrossEntropy}(\hat{y}, y)$$

$$Z_2 = F_2 W_2 \quad Z_2 \in \mathbb{R}^{T \times d_{model}}$$

$$Z_3 = Z_1 + Z_2 \quad Z_3 \in \mathbb{R}^{T \times d_{model}}$$

$$Y = \text{Softmax}(Z_3 W_Y) \quad Y \in \mathbb{R}^{T \times d_{vocab}}$$

مُبَرمجة

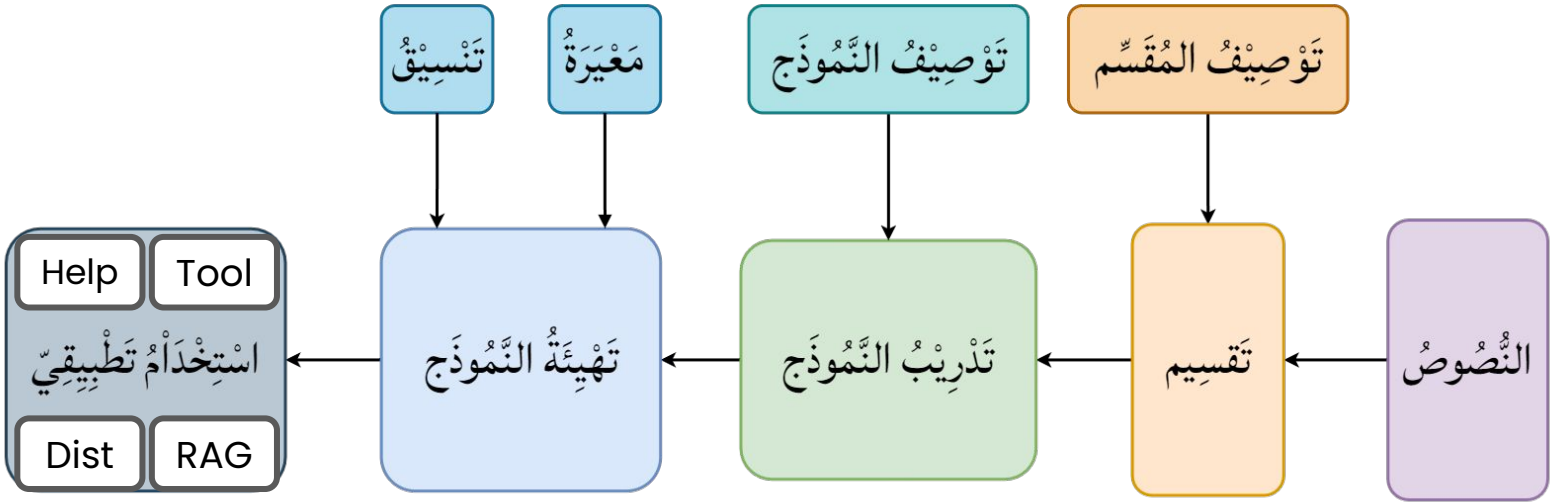


```python
import numpy as np

def gelu(x):
    return 0.5 * x * (1 + np.tanh(np.sqrt(2 / np.pi) * (x + 0.044715 * x**3)))

def softmax(x):
    exp_x = np.exp(x - np.max(x, axis=-1, keepdims=True))
    return exp_x / np.sum(exp_x, axis=-1, keepdims=True)

def layer_norm(x, g, b, eps: float = 1e-5):
    mean = np.mean(x, axis=-1, keepdims=True)
    variance = np.var(x, axis=-1, keepdims=True)
    return g * (x - mean) / np.sqrt(variance + eps) + b

def linear(x, w, b):
    return x @ w + b

def ffn(x, c_fc, c_proj):
    return linear(gelu(linear(x, **c_fc)), **c_proj)

def attention(q, k, v, mask):
    return softmax(q @ k.T / np.sqrt(q.shape[-1]) + mask) @ v

def mha(x, c_attn, c_proj, n_head):
    x = linear(x, **c_attn)
    qkv_heads = list(map(lambda x: np.split(x, n_head, axis=-1), np.split(x, 3, axis=-1)))
    causal_mask = (1 - np.tri(x.shape[0])) * -1e10
    out_heads = [attention(q, k, v, causal_mask) for q, k, v in zip(*qkv_heads)]
    x = linear(np.hstack(out_heads), **c_proj)
    return x

def transformer_block(x, mlp, attn, ln_1, ln_2, n_head):
    x = x + mha(layer_norm(x, **ln_1), **attn, n_head=n_head)
    x = x + ffn(layer_norm(x, **ln_2), **mlp)
    return x

def gpt2(inputs, wte, wpe, blocks, ln_f, n_head):
    x = wte[inputs] + wpe[range(len(inputs))]
    for block in blocks:
        x = transformer_block(x, **block, n_head=n_head)
    return layer_norm(x, **ln_f) @ wte.T

def generate(inputs, params, n_head, n_tokens_to_generate):
    from tqdm import tqdm
    for _ in tqdm(range(n_tokens_to_generate), "generating"):
        logits = gpt2(inputs, **params, n_head=n_head)
        next_id = np.argmax(logits[-1])
        inputs = np.append(inputs, [next_id])
    return list(inputs[len(inputs) - n_tokens_to_generate :])

```

معيرة النماذج

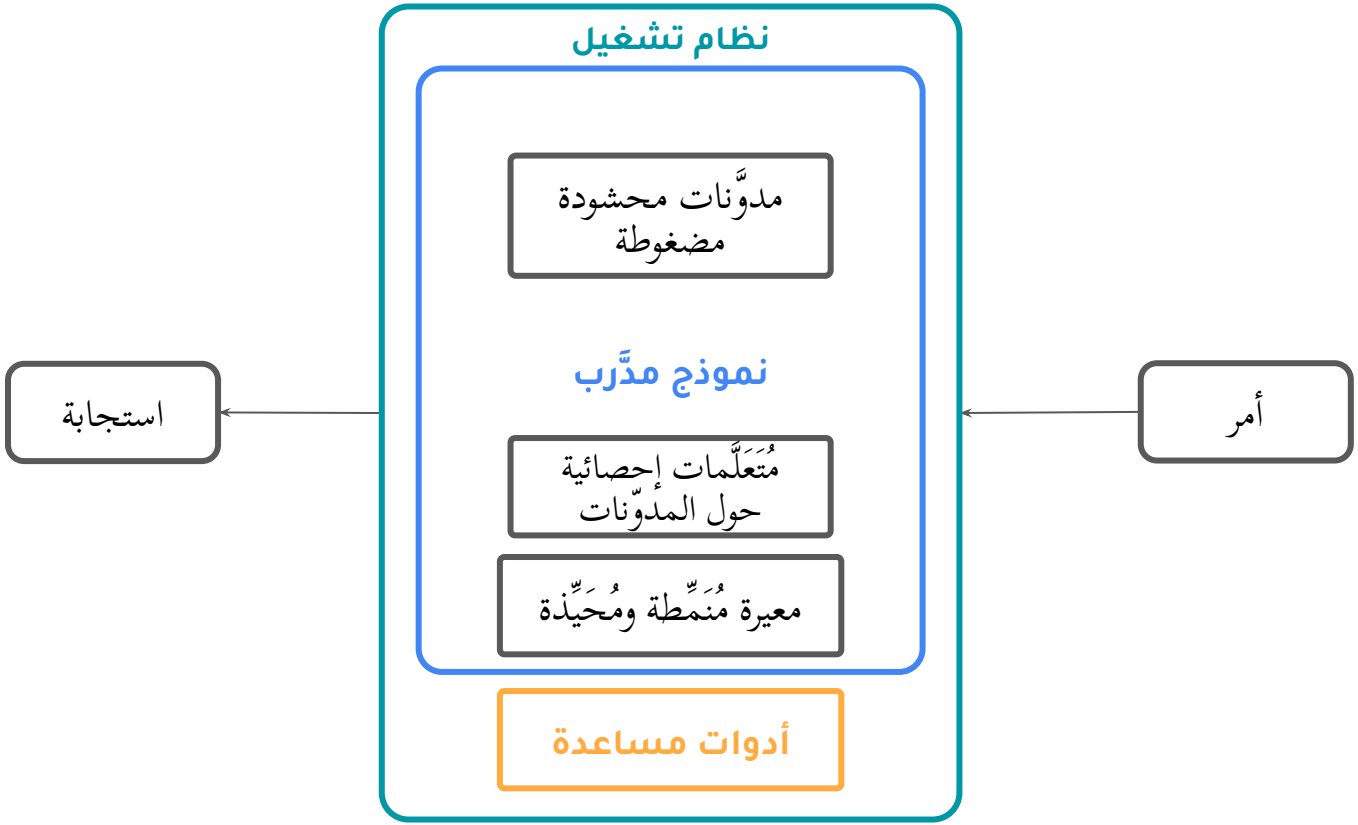
هي عملية تدريب عاديّة ولكنّ المُدخلات (المدوّنة) تكون بصيغة مُهَيكلة ذات بُنية محددة ليتعَلَّمها النموذج بإحاطة المُدخلات بأجزاء لغوية مخصصة يتبع النموذج أنماطها بعد التعلم

عَوْدًا على بدء



رَأَيْتُ النَّاسَ غَالِبَهُمْ رُسُومِيْ... ولَيْسَ بِقَارِئٍ يَفْرِيْ السُّطُورَا

نظام تشغيل جديد



خاتمة وتوصيات

عرضنا (بل إن شئت قُل ضغطنا وحشرنا) في هذه الورقة عُلالَةً مُختصرة وإشارات مُقتصرة حول نماذج المحولات التوليدية العربية وما يتعلق بها من مُدوَّنات وخرزمات. فقد شرحنا مركبات النموذج ومن ثَمَّ أشهر النماذج العربية ثم ختمت بتطبيقات النماذج اللغوية. وأختم بأمر يجول بخاطري حول علاقة العربية بهذا المجال من مُنطلقين: الأول حول ضرورة الإهتمام بها ابتداءًا قبل التسارع إلى إدخالها في النماذج اللغوية أو غيرها من التطبيقات الحاسوبية فإنه لا يُتصور أن تكتب الأبحاث وتُنشر المقالات بلغة غيرها والمُتكَلَّمُ عنه هي. والثاني حول المجالات المُهتَمِّ بها وأولويَّتُها في خدمة العربية، فإن الناظر يُدرك بلا شك أن غالب العلوم الآن منشورة بالكلزية ولا بُدَّ أن تُعرَّب وتُعدَّل المناهج كذلك فأظن أنه ينبغي الإهتمام بدءًا بالترجمة وتكثير بحوثها وتطويرها حتى نقدر على تقديم هذه العلوم بأصولها وأُسُسُها بلسان عربي ثم ننطلق بعد ذلك لضغط هذه المعارف وصهرها في النماذج اللغوية. وأخيرا أختم بنصيحة استللتها من كلام تشومسكي لمَّا سُئل عن هذه النماذج فقال "أنها لم تُقَدِّم شيئا في المجال العلمي وإنما هي أدوات مفيدة هندسيا وتطبيقيًّا" [74] أقدمِّها للباحث لألا يترك ما بين يديه من بحوث نظرية ومسائلَ علمية ويَهرَع لهذه المُندَلقات يَكتُب حولها بل فليَلزَم كراسته وقلمَه وبَحثه فإنما هي زوبعةٌ في فنجان وجعجعة بلا إطحان إذ:

(من راقب الناس لم يظفر بحاجته... وفاز بالطيبات الفاتك اللهج)[28]

و (يا باريَ القوس بريا ليس يُحسنُه ... لا تظلم القوس أعط القوس باريها)[29]

للتواصل:

/https://www.linkedin.com/in/serry-sibaee

https://github.com/serrysibaee/

serrytowork@gmail.com

المراجع

1. ورقة المحولات التوليدية العربية بنظرة تقنية رياضية بحثيّة مُختصرة، سري بن تيسير السباعي 2024.
https://drive.google.com/file/d/1rP8jpJOVM82OISqIT0Kq0PUmsoab_bEU/view?usp=sharing
2. كتاب "كيف يعمل متمم" (مقدمة ممتازة)
https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/
3. ورقة المحولات الأصلية Attention is all you need
https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
4. دراسة مسحية عن النماذج العربية
https://arxiv.org/abs/2410.20238
5. برمجة "متمم" بطريقة مختصرة
https://github.com/jaymody/picoGPT/blob/a750c145ba4d09d5764806a6c78c71ffaff88e64/gpt2_pico.py#L3-L58
6. موقع Hugginface لتجربة النماذج والمدوّنات
https://huggingface.co/
7. مختصر تعريفي ببعض النماذج العربية المشهورة
https://docs.google.com/document/d/1pNt4siq6cbTeeT7CEBXJER8yxBQQPxd5-rUbDLVGhHQ/edit?usp=sharing
8. معجم سدايا للذكاء الإصطناعي
https://sdaia.gov.sa/ar/MediaCenter/KnowledgeCenter/ResearchLibrary/SDAIAPublications15.pdf

المراجع

1. مقالات سري السباعي على كراميل (فيها مقدمة عن المجال وكذا مقالتان أخريان)
https://caramellaapp.com/serrysibaee
2. كتاب الدكتور نزار حبش "مقدمة إلى معالجة اللغة العربية"
3. كتاب جرفاسكي https://web.stanford.edu/~jurafsky/slp3/
4. مقرر ستانفورد CS124 وهناك مرئيات له https://web.stanford.edu/class/cs124/
5. معاجم الذكاء الصناعي المذكورة: معجم طعيمة
https://dlarabic.com/%d9%83%d8%aa%d8%a7%d8%a8-%d9%85%d8%b9%d8%ac%d9%85-%d9%85%d8%b5%d8%b7%d9%84%d8%ad%d8%a7%d8%aa-%d8%a7%d9%84%d8%aa%d8%b9%d9%84%d9%85-%d8%a7%d9%84%d8%a2%d9%84%d9%8a-%d9%88%d8%a7%d9%84%d8%aa%d8%b9%d9%84/
6. محاضرات كرباثي https://www.youtube.com/playlist?list=PLAqhIrjkxbuWI23v9cThsA9GvCAUhRvKZ