مختصر كتابي

Tokens and Tokenization



لماذا يحسب الـModel النص بوحدات تختلف عن الكلمات؟

قد ترسل رسالتين بالمعنى نفسه، إحداهما بالعربية والأخرى بالإنجليزية، فتجد أن عدد الـ Tokens مختلف. وقد ترسل النص نفسه إلى Model آخر فيتغير العدد مرة أخرى. السبب أن الـ **Token** ليس وحدة لقياس المعنى، وليس مرادفًا للكلمة، بل هو الوحدة التي يتعامل بها الـModel مع النص.

قبل أن تدخل جملتك إلى الـModel، يحوّلها **المُجزِّئ** (Tokenizer) إلى سلسلة من الـ Tokens، ولكل Token رقم تعريف (ID). قد يكون الـ Token كلمة كاملة، أو جزءًا من كلمة، أو علامة ترقيم، وقد تكون المسافة جزءًا من Token مع النص الذي يليها. لذلك لا يصح أن تعدّ كلمات الجملة وتستنتج أن عدد الـ Tokens مثله.

لماذا لا نعطي لكل كلمة كاملة رقمًا مستقلًا؟ لأن ذلك يتطلب **مفردات** (Vocabulary) ضخمة جدًا، فهناك لغات مختلفة وتصريفات وأسماء وأخطاء إملائية وكود. وإن عاملنا كل حرف وحده صارت السلسلة أطول بكثير. لذلك تستخدم مُجزِّئات كثيرة **الأجزاء الفرعية** (Subwords)، كمن يملك مكعبات كبيرة للأشكال المتكررة وأخرى أصغر لتركيب أشكال جديدة. مثال توضيحي: كلمة playing قد تُمثَّل بقطعتين هما play وingg، وهذا ليس تقسيمًا ثابتًا في كل المُجزِّئات.

تُبنى المفردات والقواعد مسبقًا، غالبًا من بيانات التدريب، ومن الطرق الشائعة طريقة **BPE** (Byte Pair Encoding). تبدأ بوحدات صغيرة وتدمج الأزواج المتجاورة المتكررة كثيرًا حتى يبلغ حجم المفردات المطلوب. والمُجزِّئ لا يفهم المعنى ثم يقسّم بحسبه، بل يطبّق قواعد محددة، أما معالجة المعنى في الـContext فتحدث بعد ذلك داخل الـModel. ولهذا قد يظهر الـ Token ID نفسه في Contexts اثنين بمعنيين مختلفين.

أما اختلاف العدد بين العربية والإنجليزية فمصدره تغطية المفردات لكل لغة وطبيعة كتابة الكلمات، ولا توجد نسبة ثابتة بين اللغتين. وفي الكود قد ينقسم اسم Variable تراه كلمة واحدة إلى عدة Tokens، وقد تغيّر مسافة أو حالة الأحرف (Capitalization) التقسيم.

وللأمر أثر عملي في ثلاثة مواضع:

- **نافذة الـ**(Context Window) **Context**: تُقاس بالـ Tokens لا بالصفحات ولا بالكلمات. إذا كانت الميزانية عشرة آلاف Token وتركت ألفين للمخرجات، بقي ثمانية آلاف للمدخلات، وتشمل التعليمات وسجل المحادثة ومحتوى الملفات ونتائج الأدوات.
- **التكلفة**: يُحسب الاستهلاك غالبًا بحسب Input Tokens وOutput Tokens، وقد يختلف سعر كل منهما، فاحسب الطلب الكامل لا رسالة المستخدم فقط.
- **السرعة**: Models اثنان بالعدد نفسه من Tokens في الثانية قد لا ينتجان الكمية نفسها من الكلام، فانظر أيضًا إلى الوقت حتى أول جزء من الرد وحتى اكتمال إجابة مناسبة.
- لمعرفة العدد الحقيقي استخدم مُجزِّئ الـModel نفسه أو أداة عدّ الـ Tokens لدى الخدمة، وراجع حقل Usage في رد الـ API.

- لا يصلح أي مُجزِّئ لأي Model، فكل ID مرتبط بطريقة تدريب الـModel، والرقم نفسه في مفردات أخرى قد يشير إلى شيء مختلف تمامًا.

الخلاصة: الـ Token هو الوحدة التي يعالجها الـModel لا الكلمة التي تقرؤها، ولذلك لا يكفي عدد الكلمات لتقدير التكلفة أو حدود الـContext.