مختصر كتابي

The Inference Pipeline



ماذا يحدث بين لحظة الضغط على Send وظهور الإجابة؟

يبدأ فهم أنظمة الذكاء الاصطناعي من خط سير واحد: الخطوات التي تمرّ بها رسالتك منذ إرسال الطلب (Request) حتى وصول الرد (Response). هذه هي الصورة التي ستُركَّب داخلها بقية المفاهيم في السلسلة.

تخيّل أن عميلًا يشتكي من تأخّر طلبه، وأنك كتبت للتطبيق:

```
اكتب رد مختصر، اعتذر فيه عن التأخير، ومتحددش ميعاد تسليم إلا لو عندك ميعاد مؤكد
```

هنا تستخدم Model مدرَّبًا ليُخرج مخرجات (Output) من المدخلات (Input) التي أعطيته إياها، وهذا ما يُسمّى **الاستدلال** (Inference). التدريب (Training) هو وقت التجهيز، والاستدلال هو لحظة التعامل مع الحالة بما تعلّمه الـModel، لكنه لا يعمل كموظف، بل بحسابات على تمثيلات عددية.

أولًا يجهّز التطبيق **الـContext**، وهو كل المعلومات التي أمام الـModel وهو يكتب الرد: رسالة العميل، وتعليمات الأسلوب، وجزء من سجل المحادثة (Chat History)، وحالة الطلب إن جلبها التطبيق من نظام الشركة. وجود المعلومة في الـDatabase الخاصة بك لا يعني أنها وصلت إلى الـModel.

بعد ذلك يقسّم المُجزِّئ (Tokenizer) النص إلى وحدات تُسمّى **الـTokens**، وقد يكون الـToken كلمة أو جزءًا من كلمة أو علامة ترقيم. لكل Token معرّف (ID) يحدّد أي Token نقصد، ثم يتحول إلى **الـEmbedding**، وهو التمثيل العددي الذي يدخل حسابات الـModel.

تمرّ هذه التمثيلات عبر طبقات الـModel وتتغير بحسب الـContext. وفي Models Transformer تربط آلية **الانتباه** (Attention) أجزاء الـContext بعضها ببعض. تُسمّى معالجة المدخلات **Prefill**، وفي نهايتها يصبح الـModel جاهزًا لاختيار أول Token في الرد.

ثم تبدأ مرحلة **Decode**: يحسب الـModel احتمالات الـToken التالي، ويختاره، ثم يبني عليه حساب الـToken الذي بعده، وهكذا حتى يظهر Token النهاية أو يبلغ الرد الحد الأقصى للطول. وتُسمّى هذه الطريقة **التوليد الانحداري الذاتي** (Autoregressive Generation). وإذا فُعِّل Streaming ظهرت الإجابة تدريجيًا.

أما الأوزان (Weights) فهي ما حفظه الـModel في التدريب، وتبقى ثابتة أثناء الاستدلال المعتاد. فإذا طلبت «رد أقل رسمية» فقد أضفت تعليمات إلى الـContext، ولم تُحدِّث الأوزان.

- الاستدلال هو استخدام Model مدرَّب لإنتاج مخرجات من مدخلات، وليس تدريبًا جديدًا مع كل رسالة.
- الـContext هو ما يراه الـModel فقط، فلا تفترض أنه يعرف معلومة لم يرسلها التطبيق ولا توجد أداة (Tool) تجلبها.
- عدد الـTokens لا يساوي بالضرورة عدد الكلمات، والمعرّف (ID) يختلف عن الـEmbedding.
- Prefill يعالج المدخلات، وDecode يولّد المخرجات Token بعد Token، وطول المدخلات وطول المخرجات يؤثران في وقت الانتظار بطريقتين مختلفتين.

- الوقت يتأثر أيضًا بضغط الخدمة والشبكة وأي أدوات يستعين بها النظام.
- الرد المقنع لا يعني معلومة صحيحة، فاحتمال الـToken ليس مقياسًا مباشرًا للصحة.

إذا وعد المساعد عميلك بميعاد من عنده، فابدأ بأسئلة محددة: هل وصلت حالة الطلب إلى الـModel؟ هل التعليمات واضحة؟ هل توجد أداة تجلب الميعاد؟ وكيف نختبر التزامه بالمعلومات المتاحة له؟

الخلاصة: رسالتك تتحول إلى Context، ثم إلى Tokens وEmbeddings، ثم يعالجها الـModel في Prefill ويولّد الرد في Decode خطوة بخطوة بأوزان ثابتة، وجودة النتيجة تعتمد على الـModel وعلى المعلومات التي وصلته وعلى طريقة التحقق من المخرجات.