Yapay zekâ sistemi neden bazen yavaş yanıt verir?



www.yzokul.com.tr

Yapay zekâ sistemi neden bazen yavaş yanıt verir?



Bir yapay zekâ sistemine soru sorduğunda yanıt bazen hızlı gelir, bazen çok daha uzun sürer. Kısa yanıt şudur: bekleme süresine tek bir neden değil, birçok şey birlikte karar verir. Modelin boyutu, girdinin uzunluğu, yanıtın uzunluğu, sunucu yükü ve kullanılan araçların hepsi rol oynar. Bu kitap her birini adım adım açıklar.



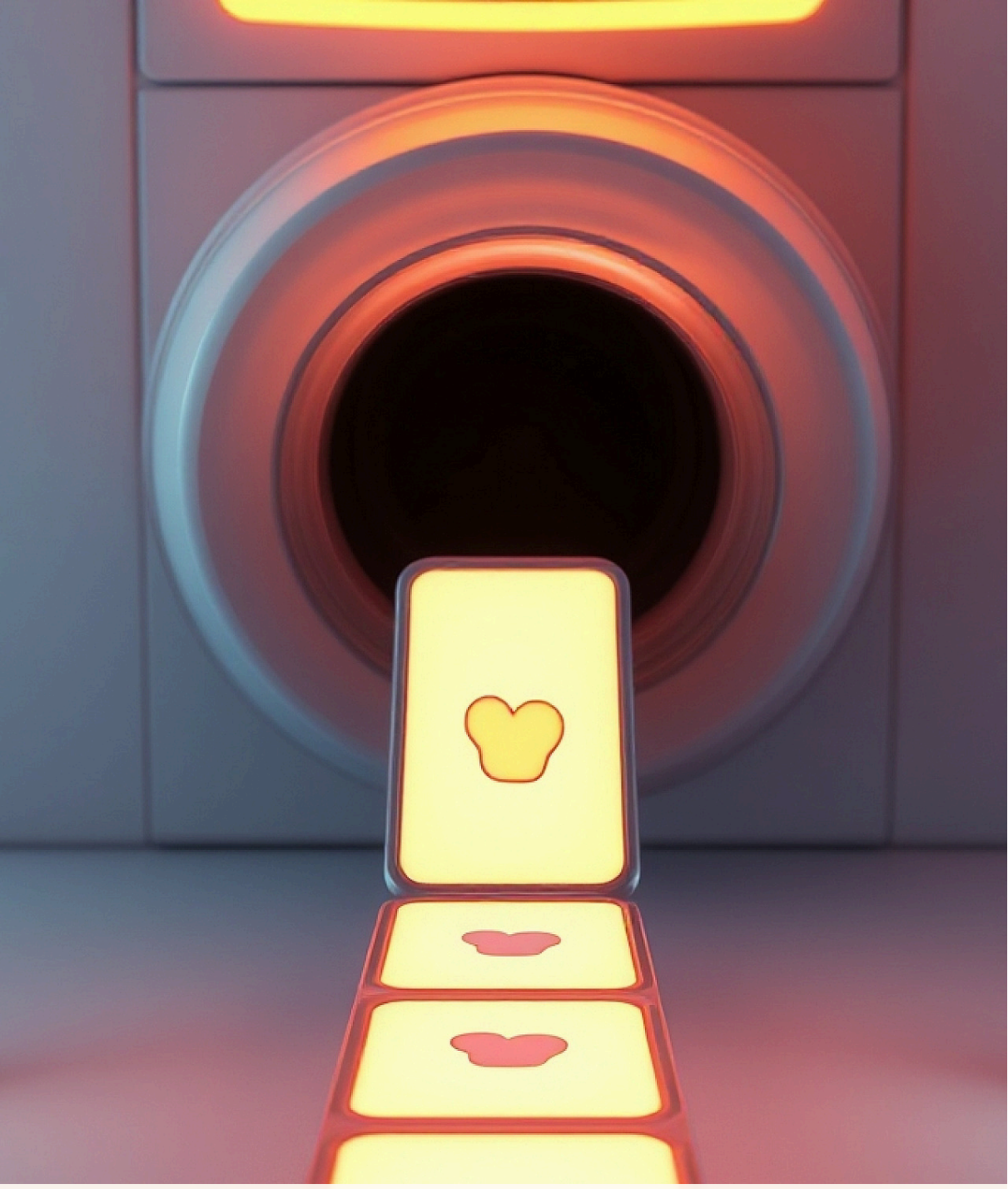
Bir yapay zekâ sistemi üç ana aşamada çalışır. Birincisi, isteğini okur ve işleyebileceği bir biçime dönüştürür. İkincisi, modelini kullanarak bir yanıt hesaplar. Üçüncüsü, yanıtı sana geri gönderir. Bu eylemlerin her biri biraz zaman alır ve toplam bekleme süresi hepsinin toplamıdır.



Model, yapay zekâ sisteminin yanıt üretmek için kullandığı geniş öğrenilmiş örüntüler kümesidir. Daha büyük bir modelin işlemesi gereken daha fazla parçası vardır, bu yüzden ürettiği her kelime için genellikle daha fazla hesaplama gerekir. Daha fazla hesaplama daha uzun bekleme demektir. Örneğin, çok büyük bir model tek bir cümle üretmek için birkaç saniye alabilirken küçük bir model neredeyse anında yanıt verebilir.



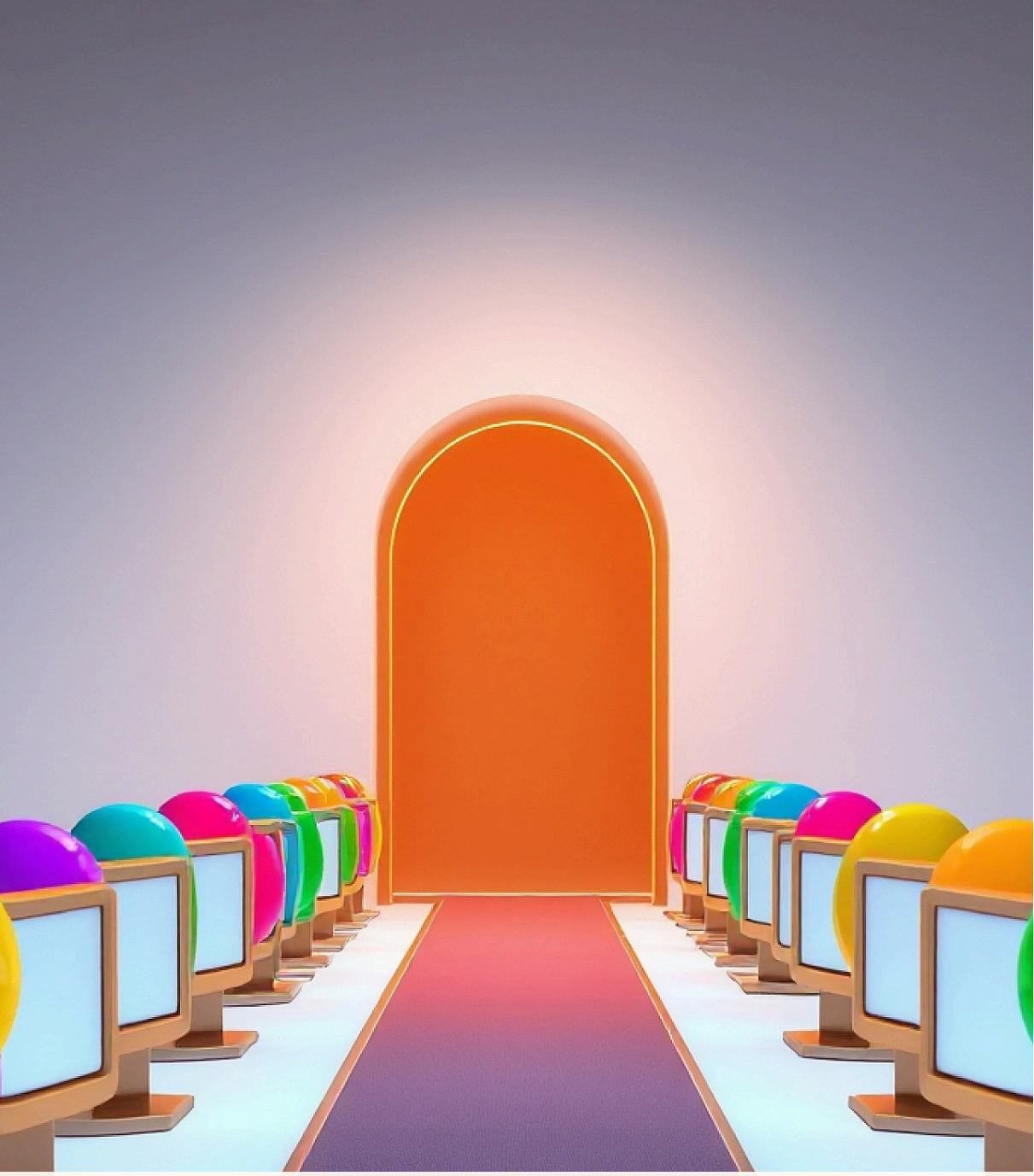
Girdi, sisteme verdiğin her şeydir: sorun, yapıştırdığın metin ve eklediğin dosya. Uzun bir girdi, yanıt başlamadan önce parça parça okunmalı ve işlenmelidir. Daha fazla girdi daha fazla iş demektir, bu yüzden yanıtın ilk kelimesi daha geç görünür. Örneğin, on sayfalık bir belgeyi yapıştırmak, tek satırlık bir soru sormaktan daha fazla işlem gerektirir.



Yanıt bir anda oluşturulmaz. Sistem onu kelime kelime ya da token token oluşturur; token küçük bir metin parçasıdır. Uzun bir yanıt kısa olandan daha fazla adım gerektirir, bu yüzden daha yavaş görünür. Örneğin, iki cümlelik bir yanıt bir anda görünebilirken uzun bir açıklama yavaş yavaş gelir.



Sunucu, yapay zekâ modelini çalıştıran ve aynı anda birçok kullanıcıya hizmet veren güçlü bir bilgisayardır. Aynı anda birçok kişi için yemek pişiren ortak bir mutfak düşün. Mutfak yoğunken her siparişin hazırlanması daha uzun sürer. Aynı şekilde, birçok kullanıcı aynı sunucuya istek gönderdiğinde her isteğin tamamlanması daha uzun sürebilir.



Sunucu dolduğunda yeni isteklerin hepsi aynı anda işlenemez. Bir kuyruğa girer ve sıralarını bekler. Kuyrukta geçirilen zaman, herhangi bir hesaplama başlamadan önce harcadığın bekleme zamanıdır. Örneğin, yoğun bir saatte isteğin, sunucu üzerinde çalışmaya başlamadan önce birkaç saniye sırada bekleyebilir.



Bazı yapay zekâ sistemleri sorunu yanıtlamaya yardımcı olmak için araçlar kullanabilir. Belge arama, kod çalıştırma veya başka bir hizmeti çağırma ayrı eylemlerdir. Her ek eylem toplama kendi süresini ekler. Örneğin, sistem önce bir bilgi tabanında arama yapıp sonra bir hesaplama çalıştırmak zorundaysa, yalnızca birini değil her iki adımı da beklersin.



Belge arama, doğru malzemeyi bulmak için birçok dosyaya bakmak demektir. Sistem, en iyi eşleşmeyi kullanabilmeden önce isteğini her dosyanın içeriğiyle karşılaştırmalıdır. Yanıt ancak doğru malzeme bulunduktan sonra başlayabilir. Örneğin, büyük bir rapor koleksiyonunda arama yapmak, tek bir kısa notta arama yapmaktan daha uzun sürer.



Kod çalıştırma, sistemin bir sonuç elde etmek için gerçekten bir program yürütmesi demektir. Kod adım adım çalışmalıdır ve yanıt, sonucun geri gelmesini bekler. Yürütme zaman alır, özellikle karmaşık hesaplamalarda. Örneğin, basit bir toplama neredeyse anında dönerken büyük bir sayı tablosunu işlemek birkaç saniye sürebilir.



İki görev düşün: uzun bir dosyayı incelemek ve kısa bir soruyu yanıtlamak. Uzun dosya okunmalı, parçalara ayrılmalı ve tek tek işlenmelidir. Kısa soru yalnızca hızlı bir geçiş gerektirir. Bu yüzden uzun dosya çok daha fazla işlem gerektirir ve daha uzun sürer. Bu fark normaldir ve beklenir; bir şeyin bozuk olduğunun işareti değildir.



Telefonun veya bilgisayarın da iş yapar. İsteğini göndermeli, yanıtı almalı ve ekranda göstermelidir. Yavaş bir cihaz toplam bekleme süresine kendi gecikmesini ekler. Örneğin, eski bir telefon uzun bir yanıtı yeni bir telefondan daha yavaş görüntüleyebilir, sunucu aynı hızda yanıt verse bile.



Hızlı bir internet bağlantısı yardımcı olur ama hızlı bir yanıtı garanti etmez. Sunucudaki ve cihazındaki hesaplama da önemlidir. Kusursuz bir bağlantıyla bile büyük bir model, uzun bir girdi veya yoğun bir sunucu seni bekletebilir. Örneğin, hızlı bir bağlantıda kısa bir soru, sunucu aşırı yüklüyse yine de yavaş olabilir.



Yaygın bir hata, her yavaş yanıt için interneti suçlamaktır. Ağ, yolun yalnızca bir parçasıdır. Bağlantını suçlamadan önce model boyutunu, girdinin uzunluğunu, beklenen yanıtın uzunluğunu ve sunucu yükünü kontrol et. Örneğin, uzun bir dosya çok hızlı bir bağlantıda bile yavaş olabilir.



Bir başka yaygın hata, yavaşlığın tek bir nedeni olduğunu varsaymaktır. Gerçekte birkaç neden aynı anda birlikte etkili olabilir. Büyük bir modelle yoğun bir sunucuda uzun bir girdi ve bir araç araması, tüm bu nedenlerin birleşimiyle yavaş olabilir. Örneğin, iki kullanıcı girdileri ve sunucu yükleri farklı olduğu için farklı hızlar görebilir.



Özetle, model boyutu, girdi uzunluğu, yanıt uzunluğu, sunucu yükü ve araç kullanımı hızı etkiler. Yavaşlık her zaman internet değildir, çünkü cihazındaki ve sunucudaki hesaplama da sayılır. Bir yanıt yavaş olduğunda tüm bu etkenleri birlikte düşün. Bunları anlamak, gerçekçi beklentiler belirlemene ve hız önemli olduğunda daha kısa girdiler seçmene yardımcı olur.