Yapay zekâda token nedir?



www.yzokul.com.tr

Yapay zekâda token nedir?



Yapay zekâ dil modellerinin token'larla çalıştığını duymuşsundur. Token nedir? Kısa yanıt: token, bir dil modelinin işlediği küçük bir metin parçasıdır. Modelin metni okurken ya da yazarken ele aldığı temel birimdir. Bu kitap token'ları sıfırdan başlayarak adım adım açıklar; böylece tam olarak ne olduklarını ve neden önemli olduklarını anlayabilirsin.



Bir dil modeli bütün bir cümleyi bir anda okumaz. Bunun yerine metni küçük parçalara böler. Her parçaya token denir. Model sonra bu token'ları tek tek işler. Bu bölme işlemi gereklidir çünkü model sayılarla çalışır ve token'lar sayıya dönüştürülen parçalardır. Yani token'lar, bir modelin metni nasıl ele aldığının ilk adımıdır.



Basit ve kısa sözcüklerde tam bir sözcük tek bir token olabilir. Örneğin İngilizce cat sözcüğü kısa ve yaygındır, bu yüzden genellikle tek bir token'dır. Model cat'i tek bir parça olarak görür. Bu, bir token'ın tam bir sözcük olabileceğini gösterir. Ama bu her zaman böyle değildir; bunu birazdan göreceksin.



Bir token her zaman tam bir sözcük değildir. Uzun ya da az kullanılan sözcükler daha küçük parçalara bölünebilir. Her parça bir token olur. Örneğin unbelievable sözcüğü un, believe ve able olarak bölünebilir. Bu parçaların her biri bir token'dır. Yani bir token yalnızca tam bir sözcük değil, bir sözcük parçası da olabilir.



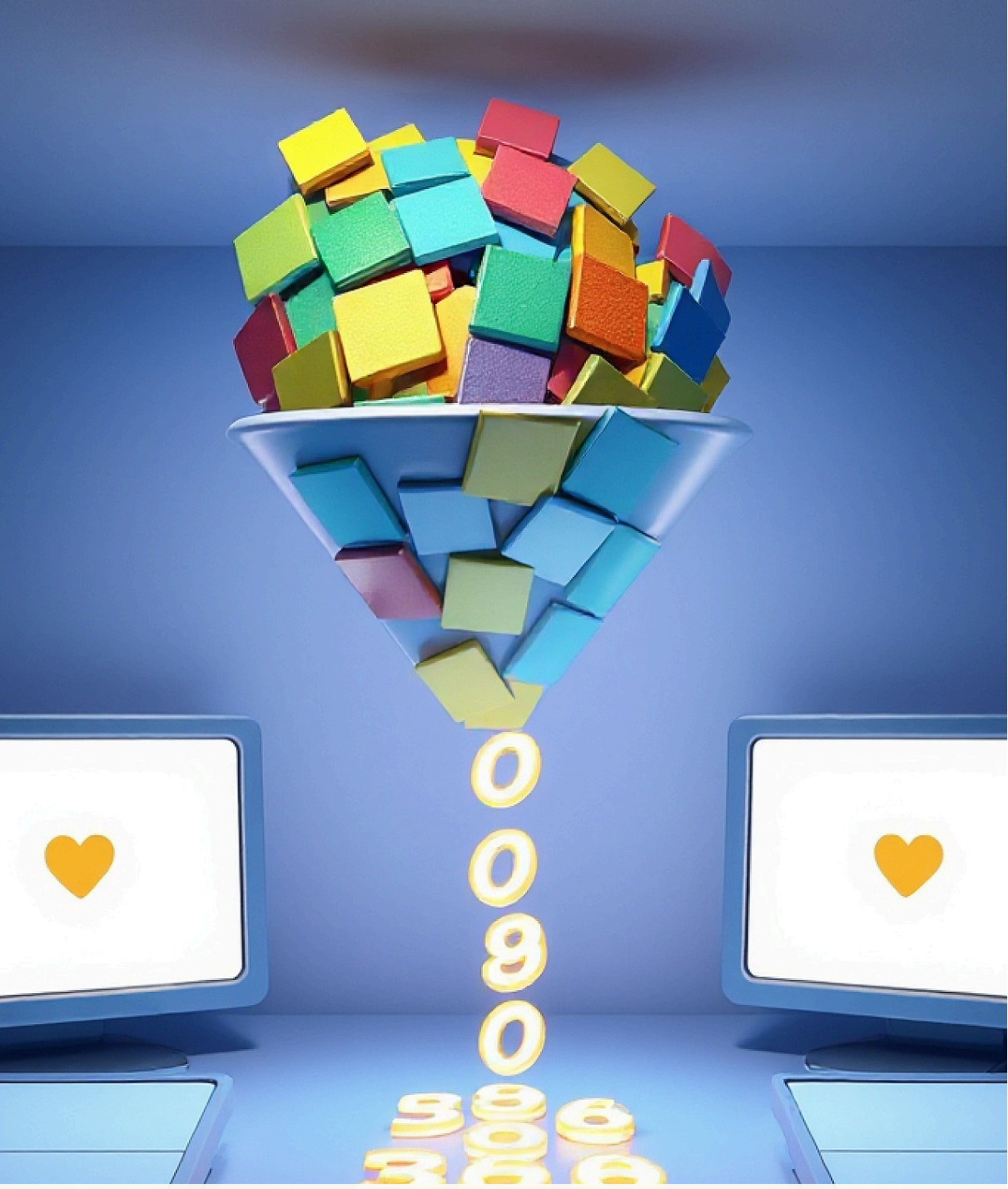
Noktalama işaretleri de token'dır. Bir virgül, bir nokta ya da bir soru işareti, modelin işlediği bir metin parçasıdır. Örneğin Hello, world! cümlesinde virgül ve ünlem işareti ayrı ayrı birer token'dır. Yani token'lar yalnızca harfleri ve sözcükleri değil, noktalama işaretlerini de kapsar. Bunu hatırlamak önemlidir.



Uzun bir Türkçe sözcük düşün: muzikolojik. Bu sözcük yaygın değildir, bu yüzden birkaç parçaya bölünebilir. Model onu mu, ziko ve lojik olarak bölebilir. Her parça bir token'dır. Özgün sözcüğün daha küçük parçalara ayrıldığını ve her parçanın model tarafından ayrı ayrı işlendiğini görebilirsin.



Önceki örnekteki parçalara bak: mu, ziko, lojik. Bu parçaların hiçbiri tek başına tam bir sözcük değildir. Onlar yalnızca parçacıklardır. Bu, bir token'ın boyutu ne olursa olsun basitçe bir metin parçası olduğunu gösterir. Anlamlı bir sözcük olması gerekmez. Model her parçayı bir token olarak ele alır.



Model doğrudan harflerle çalışmaz. Sayılarla çalışır. Bu yüzden her token bir sayıya dönüştürülür. Bu sayıya token ID denir. Model bu sayıları işler. Örneğin cat token'ı 1234 sayısına dönüşebilir. Metin, modelin hesaplama yapabileceği bir sayı listesine dönüşür.



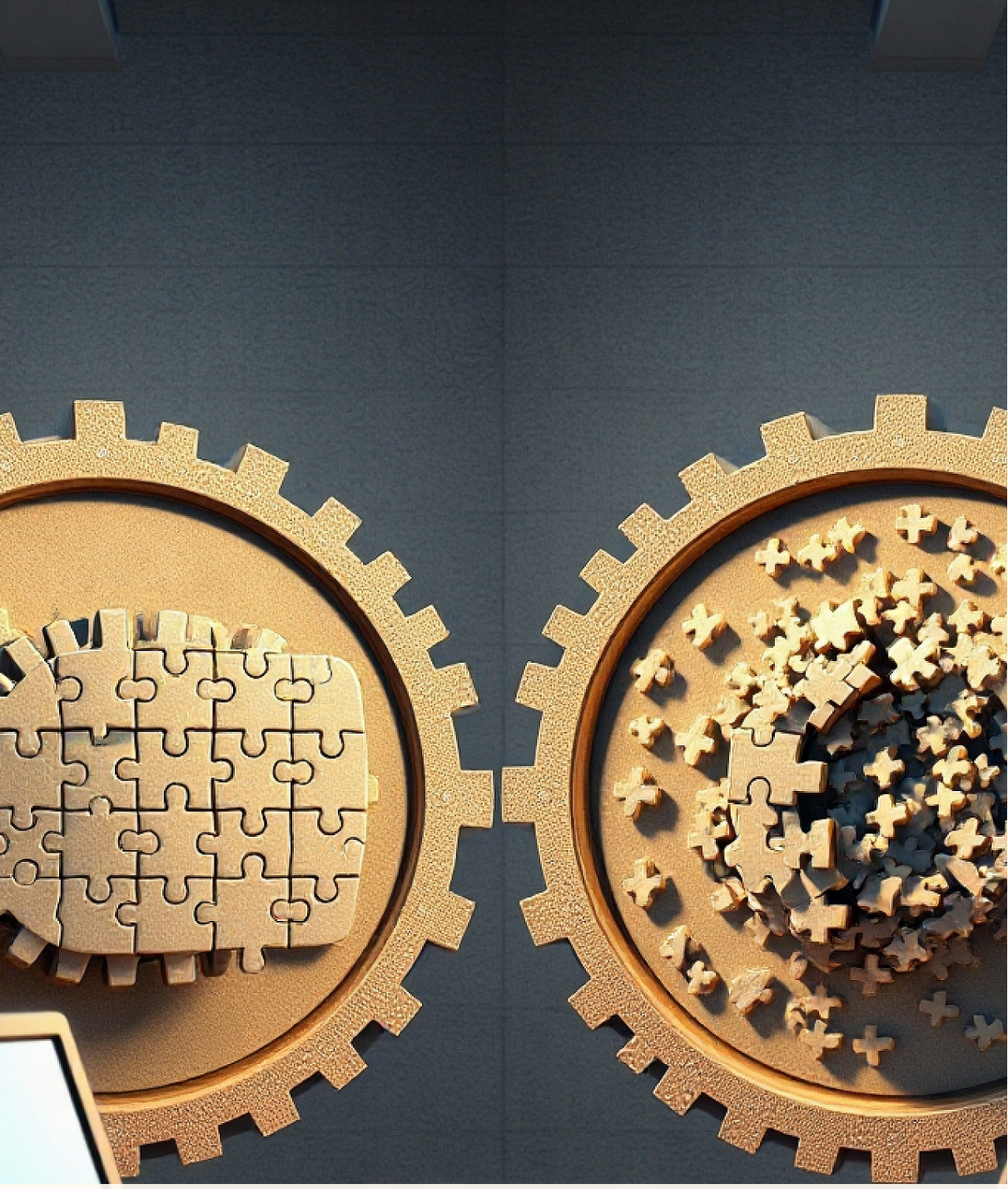
Neden sayılar? Çünkü model matematik yapar. Bir sonraki token'ı tahmin etmek için matematiksel işlemler kullanır. Harfler ve sözcükler matematikte doğrudan kullanılamaz. Bu yüzden her token bir sayı alır ve metin bir sayı dizisine dönüşür. Model metni bu şekilde anlar ve üretir. Sayılar modelin dilidir.



Sözcük sayısı ile token sayısı aynı değildir. Bir sözcük birkaç token olabilir. Örneğin unbelievable sözcüğü üç token olabilir. Yani beş sözcüklü bir cümlenin beşten fazla token'ı olabilir. Sözcükleri saymak sana token sayısını vermez. Token'ları ayrıca saymalısın.



Kullanılan dil, metnin token'lara nasıl bölündüğünü etkiler. Bazı dillerde uzun sözcükler ya da karmaşık sözcük biçimleri vardır. Örneğin Türkçe sözcükler çok uzun olabilir ve birçok token'a bölünebilir. İngilizce sözcükler genellikle daha kısadır ve daha az token'a bölünebilir. Yani aynı fikir farklı dillerde farklı token sayıları verebilir.



Tokenization yöntemi de sonucu etkiler. Tokenization, metni token'lara bölme işlemidir. Farklı yöntemler metni farklı şekillerde böler. Bir yöntem bir sözcüğü iki token'a bölebilirken başka bir yöntem onu tek token olarak tutabilir. Yani token sayısı kullanılan yönteme bağlıdır.



Yaygın bir hata, bir token'ın her zaman tam bir sözcük olduğunu düşünmektir. Ama gördüğün gibi bir token bir sözcük, bir sözcük parçası ya da bir noktalama işareti olabilir. Örneğin unbelievable sözcüğü üç token olabilir ve bir virgül tek bir token'dır. Yani token eşittir sözcük diye varsayma. Token'lar metin parçalarıdır.



Bir başka yaygın hata, sözcükleri sayıp bunun token sayısı olduğunu varsaymaktır. Sözcükler ve token'lar farklı şekillerde sayılır. On sözcüklü bir cümlenin on beş token'ı olabilir. Yani token sayısına ihtiyacın varsa sözcükleri değil token'ları saymalısın. Bu, dil modellerini kullandığında önemlidir çünkü modellerin token sınırları vardır.



İşte kısa bir metin: I love AI. Onu sözcüklere göre bölersek üç token elde ederiz: I, love, AI. Ama karakterlere göre bölersek daha çok token elde ederiz: I, boşluk, l, o, v, e, boşluk, A, I. Metin aynıdır ama parçalar farklıdır. Bu, yöntemin token'ları değiştirdiğini gösterir.



Özetle, token bir dil modelinin işlediği metin parçasıdır. Tam bir sözcük, bir sözcüğün parçası ya da bir noktalama işareti olabilir. Her token bir sayıya dönüştürülür. Token'lar sözcüklerle aynı değildir ve token sayısı dile ve tokenization yöntemine bağlıdır. Bu temel noktaları hatırla.