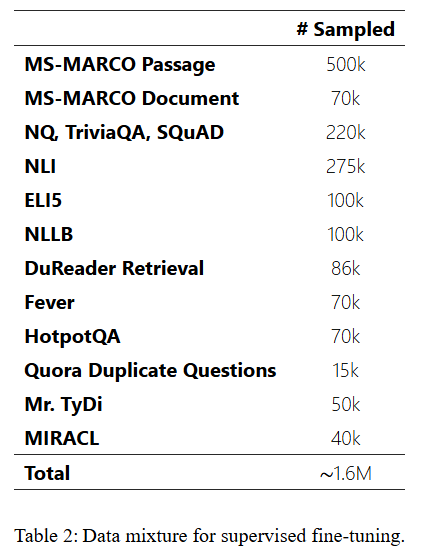
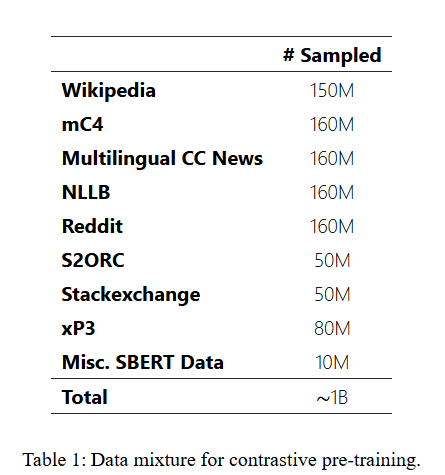
Çok Dilli E5 (Multilingual E5):  
Kısacası, E5 bir metin anlamlandırma aracıdır ve özellikle arama (retrieval) ve anlamsal benzerlik görevlerinde oldukça yüksek performans gösteren popüler bir açık kaynak modeldir. Esas olarak, bir metni (cümle, paragraf veya belge) **sayısal bir vektöre** (gömülme) dönüştüren bir modeldir.

**Multilingual E5 Text Embeddings: A Technical Report**

Bu çalışma, bilgi getirme ve anlamsal arama için yüksek performanslı ve modern modelleri **çok dilli bağlamlara** taşıyarak dil engellerini aşmayı hedeflemektedir.

**Training Methodology**  
Modelin **büyük ve geniş** bir bilgi tabanı oluşturulur. Bu aşamada **etiketsiz (zayıf denetimli)**, yani sadece metin çiftlerinden oluşan devasa veri kullanılır (**1 milyar çift**). Amaç, metinlerin genel anlam ilişkilerini, yani temel yeteneği kazanmasını sağlamaktır. **Sonra İnce İşçilik Yapılır (İnce Ayar):** Modelin gerçek dünya görevlerinde (**bilgi getirme, sorgu cevaplama**) daha hassas ve doğru sonuçlar vermesi için **etiketli (denetimli)**, daha küçük ama **yüksek kaliteli** veriler (**1.6 milyon çift**) kullanılır. Bu, modelin becerilerini görev odaklı olarak keskinleştirmektir.



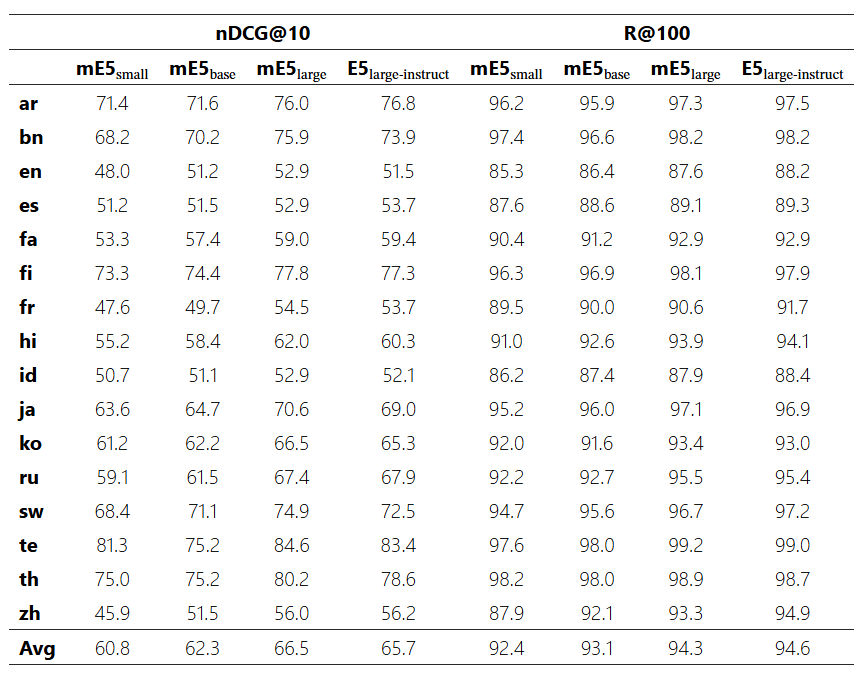
Deneysel Sonuçlar:  
**Çok Dilli E5** modelleri, hem **İngilizce'de** mevcut SOTA (State-of-the-Art / Alanının En İyisi) modelleri geçerek hem de **çok dilli bilgi getirme** ve **diller arası benzerlik** görevlerinde ezici bir üstünlük sağlayarak, metin gömülme alanında lider konuma yükselmiştir.

### **1. nDCG@10 (Normalleştirilmiş İndirgenmiş Kümülatif Kazanç)**

* **Anlamı:** Bilgi getirme (arama) sistemlerinde sonuçların **alaka düzeyini ve sıralamasını** ölçen çok önemli bir metriktir. Yüksek skor, en alakalı sonuçların listenin en başında olduğu anlamına gelir.

### **2. R@100 (Recall @ 100 / Geri Çağırma @ 100)**

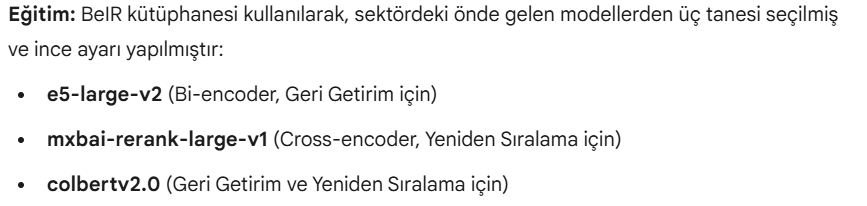
* **Anlamı:** Sorgu için alakalı olan tüm belgelerin **ilk 100 sonuç arasında ne kadarının bulunduğunu** gösterir. Yüksek skor, modelin alakalı sonuçları kaçırmadığını gösterir.



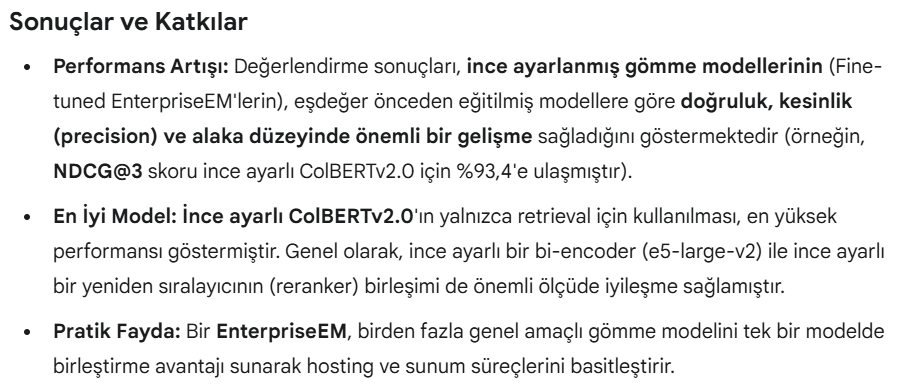
**2)- Fine-tuned Embeddings for Enterprise Semantic Search(Kurumsal Anlamsal Arama İçin İnce Ayarlı Gömülmeler)**  
  
-Genel amaçlı AI modelleri, şirketlerin **özel ve tescilli verileriyle** uyumlu değildir.  
-Önceden eğitilmiş gömülme modellerini **kurumsal görevlere göre ince ayar yapma** metodolojisi.  
  
Bu makale kurumsal bilgi erişiminde kullanılan **RAG (Geri Getirim Artırılmış Üretim)** ve **ince ayarlı Büyük Dil Modelleri (LLM'ler)** yaklaşımlarını karşılaştırmaktadır. RAG, faktüel doğruluk sağlasa da, önceden eğitilmiş modellere dayandığı için kurumsal verilere tam olarak uyum sağlayamama sorunu yaşamaktadır.

**BioBERT** ve **LEGAL-BERT** gibi alana özgü gömme modelleri başarılı olmasına rağmen, bu modellerin oluşturulması için **çok büyük miktarda (milyarlarca kelime)** alana özgü veri gerekmektedir, bu da çoğu işletme için zorluk yaratmaktadır.

Bir kurumsal metin kümesiyle bir gömme modelini ince ayar yaparak bir **EnterpriseEM (Kurumsal Gömme Modeli)** oluşturmak. Bu özel modelin RAG ardışık düzeninde kullanılması, kurumsal bağlamlarda anlamsal aramanın doğruluğunu artırmanın anahtarı olarak sunulmaktadır.  
  
Sadece metinsel veriler üzerinde çalışılmış.

Kullanılan kütüphaneler  
  


Sonuç olarak, bu makale, önceden eğitilmiş gömme (embedding) modellerinin ince ayarını yaparak kurumsal ortamlardaki bilgi erişim çözümlerinin performansını artırmayı amaçlayan bir metodoloji önermiştir. Tescilli yapılandırılmamış verilerin verimli bir şekilde yönetilmesi zorluğuna değinerek, yaklaşımımız arama sonuçlarının kesinliğini ve alaka düzeyini iyileştirmek için umut verici olanaklar sunmaktadır. Özellikle, açıklama eklenmiş özel varlıklara dayalı **sentetik soru üretimi** tekniklerimizle birleştirildiğinde, metodolojimiz, gömmeleri kurumsal ortamlardaki yaygın bilgi erişim görevlerine daha iyi uyacak şekilde adapte etme potansiyelini göstermekte, böylece daha verimli bilgi keşfi ve karar alma süreçlerini kolaylaştırmaktadır.



**3)- EASE: Entity-Aware Contrastive Learning of Sentence Embedding** (Varlık Bilinçli Cümle Gömmesi için Kontrastif Öğrenme)

Makale, hem tek dilli (monolingual) hem de çok dilli (multilingual) ortamlarda cümle gömmelerinin kalitesini artırmak için **EASE** (**E**ntity-**A**ware **S**entence **E**mbedding - Varlık Bilinçli Cümle Gömmesi) adında yeni bir kontrastif öğrenme yöntemi sunmaktadır.

#### **Temel Amaç ve Motivasyon**

Cümlelerin içeriğindeki varlık (entity) köprü metni (hyperlink) açıklamalarının (özellikle Wikipedia'dan alınan) metin semantiği için güçlü bir sinyal olduğu fikrine dayanarak, bu denetimi cümle gömme modellerine entegre etmeyi amaçlamaktadır.

#### **Başlıca Sonuçlar**

* **Tek Dilli Performans:** EASE, Semantik Metinsel Benzerlik (STS) ve Kısa Metin Kümeleme (STC) görevlerinde, SimCSE gibi güçlü kendinden gözetimli yöntemler de dahil olmak üzere mevcut en iyi yöntemlerden daha iyi veya onlarla rekabet edebilir performans göstermiştir.
* **Çok Dilli ve Düşük Kaynaklı Diller:** EASE, çok dilli ayarlarda ve özellikle düşük kaynaklı dillerde etkili olduğunu kanıtlamıştır. Varlıklar arası köprüler (interlanguage links) ve Wikipedia varlık denetimi kullanılarak, çapraz dil hizalamasını (cross-lingual alignment) kolaylaştırır ve sıfır-atış (zero-shot) transfer yeteneğini artırır.
* **Ablasyon Çalışmaları:** Hem varlık tabanlı kaybın hem de kendinden gözetimli kaybın performansta önemli bir rol oynadığını ve her ikisinin de modelin başarısı için gerekli olduğunu doğrulamıştır.

