Türkçe Ağırlıklı RAG Sistemleri için Embedding Model Karşılaştırma ve Benchmark Raporu

1. Giriş

Bu rapor, Türkçe ağırlıklı Retrieval-Augmented Generation (RAG) sistemlerinde all-MiniLM-L6-v2'nin üzerine çıkabilecek yedi embedding modelini karşılaştırmakta, benchmark verilerini özetlemekte ve pratik bir pipeline önerisi sunmaktadır. Model seçiminde Türkçe uyum, doğruluk, performans, verimlilik ve reranking stratejileri gibi kriterler göz önünde bulundurulmuştur.

2. Seçim Kriterleri

1. 1. Türkçe ve çok-dilli performans
2. 2. Semantic similarity / retrieval doğruluğu
3. 3. Verim / gecikme / model boyutu
4. 4. Türkçeye özel adaptasyon
5. 5. Pratik entegrasyon ve genişletilebilirlik

3. Modellerin Detaylı İncelemesi

multilingual-e5-large-instruct

XLM-R tabanlı çok-dilli instruct-tuned embedding modeli. Düşük kaynaklı dillerde (örneğin Türkçe) güçlü performans sergiler. MMTEB benchmark'ında özellikle düşük kaynaklı senaryolarda öne çıkar. Yüksek boyutlu (1024) embedding, üst seviye doğruluk sağlarken latency ve kaynak maliyeti artar.

Kaynaklar:

- - HuggingFace model card: intfloat/multilingual-e5-large-instruct
- - MMTEB (Massive Multilingual Text Embedding Benchmark) makalesi ve sunumları (ICLR 2025)
- - Çok-dilli embedding arama üzerine blog yazıları

multilingual-e5-base

Multilingual E5'in base varyantı; large'a göre daha hafif ve üretim ortamları için dengeli doğruluk/hız kombinasyonu sunar. Türkçe dahil 100+ dili destekleyen çok-dilli ön-eğitimli yapısı sayesinde karışık dil içeren sorgularda güvenilir retrieval sağlar.

Kaynaklar:

- - HuggingFace model card: intfloat/multilingual-e5-base
- - MMTEB benchmark bağlamı ve çok-dilli performans analizleri

all-mpnet-base-v2

Sentence-Transformers ekosisteminde genel amaçlı embedding kalitesi yüksek olan model. all-MiniLM-L6-v2'ye kıyasla daha iyi semantic similarity sonuçları verir; reranking ile birlikte güçlü performans sağlar. STS-B benzeri görevlerde genelde ~87-88 skorlar göstererek MiniLM serisine göre iyileşme sunar.

Kaynaklar:

- - HuggingFace model card: sentence-transformers/all-mpnet-base-v2
- - Milvus karşılaştırma referansı: all-mpnet-base-v2 ve MiniLM farkları
- - Akademik incelemeler ve benchmark dokümanları (MTEB/MMTEB)

paraphrase-multilingual-mpnet-base-v2

Çok-dilli paraphrase/similarity için optimize edilmiş model. Farklı ifade edilmiş benzer anlamlı cümleleri eşlemede avantaj sağlar. Multilingual olması sebebiyle Türkçe için zero-shot iyi performans verir.

Kaynaklar:

- - Sentence Transformers dokümantasyonu
- - HuggingFace model card: sentence-transformers/paraphrase-multilingual-mpnet-base-v2

emrecan/bert-base-turkish-cased-mean-nli-stsb-tr

Türkçe'ye özel, NLI ve STS-B benzeri veriyle fine-tune edilmiş SBERT türevi. Türkçe cümle çiftlerinde semantic similarity ve retrieval'ta daha tutarlı embedding'ler üretir. Monolingual olduğu için İngilizce içeriklerde çeviri gerekebilir.

Kaynaklar:

- - HuggingFace model card: emrecan/bert-base-turkish-cased-mean-nli-stsb-tr

atasoglu/turkish-e5-large-m2v

Türkçeye özel distill edilmiş Model2Vec varyantı; static embedding kullanımı sayesinde çok hızlıdır ve gerçek zamanlı RAG pipeline'larında latency'yi düşük tutar. E5 tabanlı altyapıdan türetilmiştir.

Kaynaklar:

- - HuggingFace model card: atasoglu/turkish-e5-large-m2v

instructor-xl

Instruction-finetuned embedding modeli. Görev talimatı (instruction) vererek alan ve niyet bazlı özelleştirilmiş embedding'ler üretir. Reranking ve intent çözümlemesi için güçlü bir tamamlayıcıdır.

Kaynaklar:

- - HuggingFace model card: hkunlp/instructor-xl
- - Instructor embedding GitHub repository

4. Benchmark Verileri ve Analizi

Aşağıda yer alan benchmark bilgileri, MMTEB (Massive Multilingual Text Embedding Benchmark) ve ilgili kaynaklardan derlenmiş model bazlı performans özetlerini içerir. Bazı Türkçe'ye özel modellerin (örneğin emrecan ve atasoglu modelleri) kamuya açık geniş karşılaştırmalı puanları sınırlı olabilir; bu durumda tasarım ve fine-tune amaçlı çıkarımlardan bahsedilmiştir.

Model	**Benchmark/Kaynak**	**Öne Çıkan Performans Notları**	**Türkçe Desteği**
multilingual-e5-large-instruct	MMTEB 2025 (ICLR 2025 sunumu & makale)	Düşük kaynaklı dillerde (Türkçe gibi) küçük XLM-R-tabanlı varyantlar bile daha büyük modellere karşı üstünlük gösteriyor; çok-dilli retrieval'te yüksek doğruluk.	Çok iyi
multilingual-e5-base	MMTEB / MTEB leaderboard	Dengeli doğruluk/hız; karışık dillerle güvenilir retrieval, production için tercih edilebilir.	İyi
all-mpnet-base-v2	MTEB/MMTEB ve STS-B (semantic	STS-B'de ~87-88 (MiniLM ~84-85)	Ortalama (İngilizce ağırlıklı ancak zero-

	similarity) karşılaştırmaları	skorları; genel semantic retrieval'da MiniLM'den anlamlı üstünlük.	shot iyi)
paraphrase-multilingual-mpnet-base-v2	Sentence Transformers dokümantasyonu / MTEB	Paraphrase eşleştirmede güçlü; farklı ifade edilmiş aynı anlamı yakalama üstünlüğü.	İyi
emrecan/bert-base-turkish-cased-mean-nli-stsb-tr	Türkçe özel fine-tune; açık leaderboard verisi sınırlı	Türkçe cümle çiftlerinde semantic similarlity için tasarlanmış; domain'a göre fine-tune ile yüksek doğruluk beklenir.	En iyi (Türkçe özel)
atasoglu/turkish-e5-large-m2v	HuggingFace modeli açıklaması (public benchmark sınırlı)	Distill edilmiş static embedding sayesinde düşük latency; pratik senaryolarda hızlı candidate retrieval için ideal.	Türkçeye özel
instructor-xl	Instructor embedding çalışmalar / topluluk kaynakları	Niyet bazlı embedding ile intent ve reranking'te fayda; task-aware adaptasyonla retrieval doğruluğu artar.	Orta (prompt ile adaptasyon)

5. Karşılaştırmalı Özet Tablo

Model	Türkçe Desteği	Doğruluk	Hız/Verim	Öne Çıkan	Embedding Boyutu

multilingual-e5-large-instruct	Çok iyi	Çok yüksek	Yavaş / Ağır	Instruct + çok-dilli	1024
multilingual-e5-base	İyi	Yüksek (denge)	Orta	Dengeli prod	1024
all-mpnet-base-v2	Ortalama	Çok yüksek	Orta	En iyi genel kalite	768
paraphrase-multilingual-mpnet-base-v2	İyi	Yüksek (paraphrase)	Orta	İfade çeşitliliği	768
emrecan/bert-base-turkish-cased-mean-nli-stsb-tr	En iyi (Türkçe özel)	Yüksek (Türkçe)	Orta	Türkçeye özel SBERT	768
atasoglu/turkish-e5-large-m2v	Türkçeye özel	İyi-yüksek	Çok hızlı	Distill edilmiş static embed	—
instructor-xl	Orta	Yüksek (niyet bazlı)	Değişken	Task-aware embedding	—

6. Önerilen Pipeline ve Uygulama Notları

6. Hızlı retrieval için atasoglu/turkish-e5-large-m2v gibi düşük latency'li model ile candidate'lar getir.
7. Reranking için all-mpnet-base-v2 veya niyet bazlı instructor-xl ile en iyi chunk'ı seç.
8. Çok-dilli içeriklerde multilingual-e5-large-instruct kullan.
9. Türkçeye özel kalibrasyon gerekiyorsa emrecan modelini kendi veri setinle fine-tune et.
10. Top-N sonucu cross-encoder ile yeniden değerlendirerek doğruluğu artır.

7. Sonuç

Bu yedi model, all-MiniLM-L6-v2'nin ötesinde farklı ihtiyaçlara göre avantajlar sunar. En yüksek doğruluk için multilingual-e5-large-instruct ve reranking kombinasyonları, düşük latency için turkish-e5-large-m2v, Hibrit pipeline hem doğruluk hem performans açısından dengeli çözüm olacaktır.

8. Kaynakça

- MMTEB: Massive Multilingual Text Embedding Benchmark (ICLR 2025)

- HuggingFace model card: intfloat/multilingual-e5-large-instruct

- HuggingFace model card: intfloat/multilingual-e5-base

- HuggingFace model card: sentence-transformers/all-mpnet-base-v2

- Milvus comparison article: all-mpnet-base-v2 vs MiniLM

- HuggingFace model card: sentence-transformers/paraphrase-multilingual-mpnet-base-v2

- HuggingFace model card: emrecan/bert-base-turkish-cased-mean-nli-stsb-tr

- HuggingFace model card: atasoglu/turkish-e5-large-m2v

- HuggingFace model card: hkunlp/instructor-xl

- Instructor embedding GitHub repository