**METU NLP BENCHMARK**

**XCOPA VERİ KÜMESİNİN KALİTE DEĞERLENDİRMESİ**

*Hazırlayanlar:*

*Abdullah Topraksoy (İstanbul Üniversitesi Dilbilimi Bölümü)*

*Ahmet Kaan Sever (Bilkent Üniversitesi Bilgisayar Mühendisliği Bölümü)*

*Ayşe Aysu Cengiz (ODTÜ Bilgisayar Mühendisliği Bölümü)*

*Münire Büşra Tufan (Hacettepe Üniversitesi Sosyoloji Bölümü)*

*Çağrı Toraman (ODTÜ Bilgisayar Mühendisliği Bölümü)*

*Esra Darıcı (ODTÜ Türk Dili Bölümü)*

Bu doküman, ODTÜ Türkçe LLM Leaderboard için belirlediğimiz mevcut veri kümelerinden biri olan XCOPA’nın kalitesinin etiketleyenler tarafından değerlendirilmesi sırasında yol göstermesi amacıyla hazırlanmıştır.

Değerlendirme için veri kümesinden rastgele seçilen 100 satırla bir excel tablosu oluşturuldu. Değerlendirmenin nasıl yapılacağının detaylı açıklaması “**Kriterler**” bölümünde, veri kümesinin detaylı açıklaması ise “**XCOPA**” bölümünde verilmiştir.

**KRİTERLER**

Tüm değerlendirmeler iki tabanında yapılacak olup kriter sağlanıyorsa 1, sağlanmıyorsa 0 olarak etiketleme bekleniyor.

Kriterin sağlanması için o kriter altındaki tüm alt soruların evet olarak cevaplanması gerekir. En az bir soruya hayır diyorsanız o kriter sağlanmamıştır.

Bu kriterlerin XCOPA için özel açıklamaları “**XCOPA**” başlığı altında yer almaktadır.

1. **Doğruluk:** Sorunun cevabı doğru mu? Doğru olarak etiketlenen cevap ile sorulan sorunun gerçek cevabı aynı mı? Bu kriter bazı sorular için araştırma gerektirebilir. Soru doğrudan yanlış olabilir veya cevabın doğruluğu çeviri sırasında kaybolmuş olabilir.
2. **Dil bilgisi:** Metin Türkçe dil bilgisine uygun mu? Yazım kurallarına ve noktalama işaretlerine uyulmuş mu?

Özetleme metinlerinde bulunan yazım kuralları hatalarına örnekler:

* “Filistin topraklarından fırlatılan bir füzenin İsrail'in güneyine isabet etmesinin ardından İsrail, Cumartesi gecesi Gazze Şeridi'ne 12 defa hava saldırısı düzenledi.”
  + Belirli bir tarih bildirmeyen ay ve gün adları küçük harfle yazılır.
  + **Doğrusu:** …İsrail, cumartesi gecesi…
* “Venezuela Devlet Başkanı Nicolas Maduro, yeni anayasayı yazacak Kurucu Meclis'in üyelerini belirleyen seçimler sonrası zaferini ilan etti.”
  + Kurum, kuruluş, kurul, birleşim, oturum ve iş yeri adlarına gelen ekler kesmeyle ayrılmaz.
  + **Doğrusu:** …Kurucu Meclisin üyelerini…

Detaylı bilgi için ["Yazım Kuralları"](https://drive.google.com/file/d/1NJ3-SzXrQIRiFoLRw8oHc1qrzoeqH8HN/view?usp=drive_link) dokümanına bakabilirsiniz.

1. **Bağdaşıklık ve Tutarlılık (Cohesion and Coherence):** Metni oluşturan sözcük ve cümleler arasındaki dil bilgisi kuralları uyumlu mu? Cümle ve paragrafların anlam bütünlüğü mantık düzenine göre yapılandırılmış mı? Terimler ve dil yapıları çeviri boyunca tutarlı bir şekilde kullanılıyor mu?

* Bağdaşıklık (Cohesion): “Bir metni oluşturan çeşitli bölümlerin bir dilsel bütünlük sağlayacak biçimde birbirlerine bağlanması” (Onursal, 2003:8) Bağdaşıklık “bir yazının metin olmasını sağlayan, metin içi ilişkileri kuran, dille ilgili özelliklerin tümünü belirten bir kavram” (Günay, 2007: 71) olarak tanımlanabilir.

Bağdaşıklık (doğru örnek):

"Ben dün okula gittim. Okulda arkadaşlarımla görüştüm. Sonra eve geldim."

Bu örnekte, cümleler arasında "ben" zamiri, "dün" ve "sonra" gibi bağlaçlar ve zaman zarfları kullanarak bir dilbilgisel bağlantı kurulmuştur.

Bağdaşıklık (yanlış örnek):

"Ben dün okula gittim. Sonra kitap okumayı seviyorum. Ardından arkadaşlarımla görüştüm."

Bu örnekte, cümleler arasında dilbilgisel bağlantı kurulmaya çalışılmış, ancak "kitap okumayı seviyorum" cümlesi, metnin geri kalanıyla dilbilgisel olarak bağlantılı değildir. Bu cümle, bir genel ifade olarak dururken, diğer cümleler bir hikaye anlatmaktadır.

* Tutarlılık (Coherence): Tutarlılık, “metindeki mantıksal bağlantı” (Subaşı Uzun, 1995: 110); bir metni oluştururken sözcükler arasındaki anlam ve mantık bağıntısı” (Aksan, 1998: 35) olarak tanımlanabilir.

Tutarlılık (doğru örnek):

"Ben dün okula gittim. Okulda arkadaşlarımla görüştüm. Sonra eve geldim ve yemek yedim."

Bu örnekte, cümleler arasında mantıksal bir bağlantı kurulmuştur. Okula gitmek, arkadaşlarla görüşmek ve eve gelmek, ardından yemek yemek, birbiriyle ilgili ve mantıklı bir sıralamadır.

Tutarlılık (yanlış örnek):

"Ben dün okula gittim. Okulda arkadaşlarımla görüştüm. Sonra aniden uzaya gittim ve bir uzaylıyla tanıştım."

Bu örnekte, cümleler arasında mantıksal bir bağlantı kurulmaya çalışılmış, ancak "aniden uzaya gittim ve bir uzaylıyla tanıştım" cümlesi, metnin geri kalanıyla mantıksal olarak bağlantılı değildir. Okulda arkadaşlarla görüştükten sonra uzaya gitmek ve bir uzaylıyla tanışmak, mantıklı bir sıralama değildir.

1. **Anlaşılabilirlik, Akıcılık ve Anlatım Bozukluğu:** Metin doğal ve akıcı bir şekilde okunabiliyor mu? Ana dili Türkçe olan biri bu cümleleri kurar mıydı? Metinde anlatım bozukluğu var mı?

Anlatım Bozukluğuna özetleme metinlerinden örnekler:

* “75 Nobel ödülü sahibi açık mektup yazdı 2017 yılında 75 Nobel ödülü sahibi İranlı yetkililere hitaben kaleme aldıkları bir açık mektupta Ahmadreza Celali'nin derhal serbest bırakılması çağrısı yaptılar.”
  + Cümlede hem noktalama işaretleri eksikliği hem de özne-yüklem uyumsuzluğundan kaynaklanan anlatım bozukluğu var.
  + **Doğrusu:** 75 Nobel ödülü sahibi açık mektup yazdı: 2017 yılında 75 Nobel ödülü sahibi, İranlı yetkililere hitaben kaleme aldıkları bir açık mektupta Ahmadreza Celali'nin derhal serbest bırakılması çağrısı yaptı.
* “Polisin silahsız siyah bir genci vurması nedeniyle 11 gündür protesto ve şiddet olaylarının yaşandığı ABD'nin St Louis bölgesinde, polisin ikinci siyah erkeği öldürmesiyle gerginliğin daha da artmasından korkuluyor.”
* Cümlede tamlama hatasından kaynaklanan anlatım bozukluğu var.
* **Doğrusu:** Polisin silahsız siyahi bir genci vurması nedeniyle 11 gündür protesto ve şiddet olaylarının yaşandığı ABD'nin St Louis bölgesinde, polisin ikinci kez siyahi bir erkeği öldürmesiyle gerginliğin daha da artmasından korkuluyor.

Detaylı bilgi için ["Anlatım Bozuklukları"](https://drive.google.com/file/d/1NHQyZXApRtmwbtvALU4ARt8rvO13amE2/view?usp=drive_link) dokümanına bakabilirsiniz.

1. **Terim Kullanımı:** Teknik ve özel terimler doğru ve yerinde kullanılıyor mu?

Örnekler:

* MMLU-pro-TR veri kümesinde Jovian Planets ibaresi Jüpiter Gezegenleri olarak çevrilmiş fakat Jüpiter Gezegenleri diye bir gezegen sınıfı yok. Jovian Planets Dev gezegenler demek.
* Prime facie duty ibaresi birincil görev diye çevrilmiş fakat felsefede prime facie ilk bakışta ödevler olarak çevrilir.

1. **Türkiye’deki Ortak Kültürel Sağduyu Bilgisine Uygunluk:** Aşağıdaki kriterlerden her ikisi de (**a** ve **b**) doğru ise 1, değilse 0 veriniz.
   1. Bu kriter doğrudan, salt, değişmez, ülkenin konumu, kuralları, yasaları, siyasi işleyiş biçimi ile ilgili verilen bilgileri kapsamaktadır. Belirtilen veri Türkiye’nin sosyal, ekonomik, kültürel, coğrafi kuralları ile ilgili mi?

Örnekler:

* “İngiliz hukukunda geleneğin hukuka dahil edilmesinin önkoşulu nedir? ….” başka bir ülkenin kurallar bütünü ile ilgilidir.
* "4 feet uzunluğunda bir tel parçası kullandı. Tel, 6 inç uzunluğunda parçalara kesildi…Yerel bir benzin istasyonu galon başına 3 dolar fiyatla benzin satıyor" örneklerinde feet, inç, galon metrikleri ülkemizde kullanılmamaktadır.  
  1. Aşağıda belirtilen kriterlerle ilgili yargılar içeriyorsa aşağıdaki kurallara göre yanıtlayınız. Bu kriterler belirlenirken Nguyen vd. (2023), Anacleto vd. (2006), Acharya vd. (2020) kaynaklarından yararlanılmıştır. Bunlar elbette sınıflara, statülere, inançlara, eğitim düzeyine, toplumsal cinsiyete, ırk ve etnik kökenlere göre değişiklik göstermektedir. Bu yüzden amaç kesin bir bilimsel anket sunmak değil, makul varsayımlara ulaşmak olacaktır.
* Yemek ve Yiyecekler

Yemek Saatleri

* + - Sabah: 07:00-09:00; Öğlen 12:00-13:30; Akşam 17:00-21:00: bu yargılar görüldüğünde (1)
    - Bu saatler dışında saatlere rastlandığında (0)
* Öğünlerde Yenilen Yiyecekler
  + Sabah: Yumurta, sosis, sucuk, salam, peynir, zeytin, ekmek, tost, reçel; Öğlen ve Akşam: et, tavuk, balık yemekleri, sulu yemekler görüldüğünde (1)
  + Sabah: Bacon, kruvasan, mantar, fasulye, marmelat; Öğlen ve Akşam club sandviç, fish and chips, tofu vs. görüldüğünde (0)
* Özel Gün Yemekleri
* Bayramlarda, kutlamalarda, düğünlerde yenilen yemeklere dikkat edilebilir. Örneğin yeni yıl kutlamalarında hindi yenilmesi bize daha sonradan gelen geleneklerden bir tanesidir.
* Yemekler
  + Domuz eti gibi ürünler görüldüğünde (0)
* İçecekler
  + Türk kahvesi, rakı, bira, şarap, şalgam, ayran (1)
  + Uzo, noel içecekleri vs. (0)
* Kıyafetler
  + İskoçların giydiği Kilt gibi geleneksel kıyafetler (0)
  + Başörtüsü, şalvar, kaftan, takke, bindallı, yemeni, mayo, etek, elbise, pantolon (1)
* Ritüeller
  + Cenaze, doğum günü, yılbaşı, doğum, sünnet, evlilik, bayramlar gibi ritüellere dikkat edilmeli. Bizim için düğünlerin kalabalık olması, bayramlarda el öpme, diş hediği, kırkı çıkartmak vs. gibi alışkanlıklar ortak kültürel sağduyu bilgisine uygundur (1)
  + Bu gibi ritüellerde şampanya patlatmak, düğünlerde resepsiyon vermek, bayramlarda kiliseye gitmek vs. bizim ortak kültürel sağduyu bilgimize uygun değildir (0)
  + Bu kriteri etiketlerken lütfen bahsedilen ritüellerdeki davranışları dikkatlice düşünerek etiketleyiniz.
* Gelenekler
  + El öpme, harçlık alma vs. (1)
  + Harakiri, ötenazi, diş doldurma, bebek atma vs. (0)
* Davranışlar
  + Sokağa tükürme, trafik kuralları, topluluk içinde gaz çıkartma vs. (0)
  + Büyüklerle saygılı konuşma, bacak bacak üstüne atmama, bahşiş bırakma vs. (1)
* Sporlar
  + Amerikan Futbolu vs. (0)
  + Futbol, Yüzme, Güreş vs. (1)

1. **Yorum:** Değerlendirme sütunları arasında yukarıdaki altı kritere ek olarak bir de “Yorum” sütunu yer almaktadır. Bu sütunun her satırını doldurmanız beklenmemektedir. Neden bir kritere 0 verdiğiniz ya da soruda kriterlerin kapsamadığı bir problemin varlığı gibi durumları not almak ve yorumlarınız varsa iletmek adına bu sütunu kullanabilirsiniz.

**XCOPA**

**Açıklama:** XCOPA, sağduyuya dayalı akıl yürütme üzerine bir veri kümesidir. Bu veri kümesinde bir cümle verilir ve sorunun etiketine göre devamına en uygun gelecek olan cümleyi verilen iki şık arasından seçilmesi beklenir. İki çeşit etiket var:

* Effect (Etki): Bu etikete sahip sorularda verilen cümlenin sonucunda oluşabilecek durumu veren şık doğru cevap olarak kabul ediliyor. Cevapların doğruluğunu ölçerken şu şekilde cümleleri birleştirebilirsiniz: (verilen cümle) Bu yüzden (doğru şık)
* Cause (Sebep): Bu etikete sahip sorularda iseverilen cümlede yazan durumun oluşma sebebini veren şık doğru cevap olarak kabul ediliyor. Cevapların doğruluğunu ölçerken şu şekilde cümleleri birleştirebilirsiniz: (verilen cümle) Çünkü (doğru şık)

**Veri kümesi sütunları:**

1. **idx:** Soruya özel numara
2. **premise:** Soruda yaşanan olayı belirten cümle.
3. **question type:** Sorunun tipini belirten etiket. “effect” ya da “cause” olabiliyor.
4. **choice1/choice2:** Cevap şıkları
5. **label:** 0 ise cevap “choice1”, 1 ise cevap “choice2” anlamına geliyor.
6. **changed:** Sorunun daha önce değiştirilip değiştirilmediğini belirtiyor.

**Değerlendirme sütunları:**

1. **Doğruluk:** Aşağıdaki iki soruya da cevabınız “evet” ise 1 verin, aksi takdirde 0 verin.
   1. Verilen etikete göre (cause/effect) cevap olarak gösterilen şık uyumlu mu? Yani aşağıdaki şekilde cümleleri birleştirdiğinizde mantıklı bir anlam akışı oluşuyor mu?
      * Effect: (premise) Bu yüzden (cevap)
      * Cause: (premise) Çünkü (doğru cevap)
   2. İki şık arasından doğru bir cevap seçilebiliyor mu?
2. **Dil bilgisi:** Aşağıdaki iki sorunun cevabı da “hayır” ise kutuya 1 yazın, aksi takdirde 0 yazın.
   1. Yazım hatası var mı?
   2. Noktalama hatası var mı?
3. **Bağdaşıklık ve Tutarlılık:** Aşağıdaki üç sorunun cevabı da “evet” ise kutuya 1 yazın, aksi takdirde 0 yazın.
   1. Metni oluşturan sözcük ve cümleler arasındaki dil bilgisi kuralları uyumlu mu?
   2. Cümle ve paragrafların anlam bütünlüğü mantık düzenine göre yapılandırılmış mı?
   3. Terimler ve dil yapıları çeviri boyunca tutarlı bir şekilde kullanılıyor mu?
4. **Anlaşılabilirlik, Akıcılık ve Anlatım Bozukluğu:** Aşağıdaki dört sorunun cevabı da “evet” ise kutuya 1 yazın, aksi takdirde 0 yazın.
   1. Metin doğal ve akıcı bir şekilde okunabiliyor mu?
   2. Ana dili Türkçe olan biri bu cümleleri kurar mıydı?
   3. Metin anlatım bozukluğu içermeyen cümlelerden mi oluşuyor?
   4. “Premise”, “choice1” ve “choice2” sütunlarındaki cümlelerde anlatılan olayların ne olduğu anlaşılıyor mu?
5. **Terim kullanımı:** Paragraf ve soruda teknik ve özel terimlerin çevirisi doğru yapılmış mı? Bu soruya cevabınız evet ise 1, hayır ise 0 verin. Örnek için bu dokümanın “Kriterler” bölümüne bakabilirsiniz.
6. **Türkiye’deki Ortak Kültürel Sağduyu Bilgisine Uygunluk:** Belirtilen veri, Türkiye’nin sosyal, ekonomik, kültürel, coğrafi kuralları ile ilgili mi? Kriterler kısmındaki koşulları içeriyorsa belirlenen kurallar çerçevesinde mi içeriyor? Cevabınız her ikisine de evet ise 1, hayır ise 0 verin.
7. **Yorum**

**REFERANSLAR**

Onursal, İrem (2003). Türkçe Metinlerde Bağdaşıklık ve Tutarlılık. Günümüz Dilbilim Çalışmaları. Multilingual Yayınları.

Günay, Doğan (2007). Metin Bilgisi. Multilingual Yayınları.

Acharya, A., Talamadupula, K., & Finlayson, M. A. (2020). Towards an atlas of cultural commonsense for machine reasoning. *arXiv preprint arXiv:2009.05664*.

Anacleto, J., Liebennan, H., Tsutsumi, M., Neris, Y, Can'alho, A. et. al., 2006, in IFIP International Federation for Information Processing, Volume 217, Artificial Intelligence in Theory and Practice, ed. M. Bramer, (Boston: Springer), pp. 1-10.

Nguyen, T. P., Razniewski, S., Varde, A., & Weikum, G. (2023, April). Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023* (pp. 1907-1917).

["Yazım Kuralları"](https://drive.google.com/file/d/1NJ3-SzXrQIRiFoLRw8oHc1qrzoeqH8HN/view?usp=drive_link) dokümanı.

["Anlatım Bozuklukları"](https://drive.google.com/file/d/1NHQyZXApRtmwbtvALU4ARt8rvO13amE2/view?usp=drive_link) dokümanı.