Benzer Soruları Çözebilme Yeteneğinin Analizi ve Görselleştirilmesi

1. Muhammed Kayra Bulut
Yıldız Teknik Üniversitesi

***Abstract*—Bu çalışma, ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1 dil modelinin, ytu-ce-cosmos/gsm8k_tr veri kümesindeki matematik problemlerini çözme yeteneğini ve bu problemlerin farklı şekillerde ifade edilmiş "muadil" sürümlerindeki başarımını incelemektedir. Proje kapsamında, modelin başlangıçta doğru ve yanlış çözdüğü 50'şer soru belirlenmiş, bu sorulara anlamsal olarak benzer ancak farklı ifadelerle muadil sorular Google Gemini API kullanılarak üretilmiştir. Üretilen muadil sorular tekrar aynı dil modeline çözdürülmüş ve sonuçlar değerlendirilmiştir. Son olarak, asıl soruların, muadil soruların ve model cevaplarının metin gömme (embedding) vektörleri ytu-ce-cosmos/turkish-e5-large modeli ile hesaplanmış, bu vektörler t-SNE algoritması ile iki boyuta indirgenerek modelin cevap doğruluğuna göre görselleştirilmiştir. Bu analiz, modelin problem çözme tutarlılığını ve farklı ifade biçimlerine karşı hassasiyetini ortaya koymayı amaçlamaktadır.**

***Index Terms*—yapay zeka**

I. Giriş

Büyük dil modellerinin (LLM) yetenekleri gün geçtikçe artmakta ve çeşitli alanlarda kullanılmaktadır. Bu alanlardan biri de matematiksel problem çözmedir. Bu proje, özellikle `ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1` modelinin Türkçe matematik problemlerindeki başarımını derinlemesine incelemeyi hedefler. Modelin sadece asıl soruları çözme başarısı değil, aynı zamanda bu soruların anlamsal olarak benzer fakat farklı şekillerde ifade edilmiş "muadil" sürümlerini çözebilme yeteneği de analiz edilmiştir.

Projenin temel amaçları şunlardır:

- `Turkish-Llama-8b-DPO-v0.1` modelinin `gsm8k_tr` veri kümesindeki başarımını belirlemek.
- Modelin doğru ve yanlış cevapladığı sorulardan yola çıkarak, farklı talimatlar (prompt) aracılığıyla Google Gemini API kullanarak bu sorulara benzer muadil sorular üretmek.
- Üretilen muadil soruları yine `Turkish-Llama-8b-DPO-v0.1` modeline çözdürerek, modelin farklı soru türlerine karşı tutarlılığını ve başarısını ölçmek.
- Asış sorular, muadil sorular ve model cevapları için `ytu-ce-cosmos/turkish-e5-large` modeli ile metin gömme (embedding) vektörleri oluşturmak.
- Bu yüksek boyutlu gömme vektörlerini t-SNE (t-distributed Stochastic Neighbor Embedding) algoritması ile iki boyuta indirgeyerek görselleştirmek.
- Görselleştirmeler aracılığıyla, modelin cevap doğruluğu ile soruların gömme uzayındaki dağılımları arasındaki ilişkiyi incelemek ve farklı soru tiplerinin (asıl, muadil, doğru/yanlış cevaplanmış) nasıl kümelendiğini veya ayrıştığını analiz etmek.

Bu çalışma, modelin sadece ezbere dayalı değil, aynı zamanda anlamsal çıkarım yapabilme ve farklı ifadelere uyum sağlayabilme yeteneğini değerlendirmeyi amaçlamaktadır.

II. Yöntem

Proje, birkaç ana adımdan oluşan bir süreç izlemiştir: asıl soruların seçilmesi ve model tarafından değerlendirilmesi, bu sorulara dayalı muadil soruların üretilmesi, üretilen muadil soruların model tarafından cevaplanması, cevapların doğruluğunun kontrol edilmesi ve son olarak tüm bu verilerin gömme vektörleri aracılığıyla görselleştirilmesi.

A. Veri Kümesi ve İlk Değerlendirme

1) **Veri Kümesi Yükleme:** Çalışmada, Türkçe matematik problemlerini içeren `ytu-ce-cosmos/gsm8k_tr` veri kümesi kullanılmıştır. Bu veri kümesi, Hugging Face Datasets kütüphanesi aracılığıyla yüklenmiş ve rastgelelik için karıştırılmıştır.
2) **Model ile Soru Çözümü:** Karıştırılmış veri kümesinden alınan sorular, `ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1` modeline bir metin üretim (text-generation) görevi olarak sunulmuştur. Modelin görevi, verilen matematik problemini çözmek ve cevabı sayısal bir değer olarak vermektir. Model, Hugging Face Transformers kütüphanesindeki `pipeline` arayüzü ile yüklenmiş ve çıkarım işlemleri gerçekleştirilmiştir.
3) **Cevap Doğruluk Kontrolü:** Modelin ürettiği cevaplar ile veri kümesindeki gerçek cevaplar karşılaştırılmıştır. Bu karşılaştırma için, hem modelin metin çıktısından hem de gerçek cevaptan sayısal değerler ayıklanmıştır. Ayıklama işlemi, metin içindeki sayıları bulmaya yönelik bir fonksiyon (düzenli ifadeler kullanılarak) ile yapılmıştır. İki sayısal değerin eşitliği, cevabın doğruluğunu belirlemiştir.
4) **Soru Seçimi:** 50 doğru ve 50 yanlış cevaplanmış soru, daha sonraki muadil soru üretim aşamasında kullanılmak üzere seçilmiş ve kaydedilmiştir. Bu kayıt, sorunun metnini, asıl cevabını, modelin ürettiği cevabı ve cevabın doğruluk durumunu içermektedir.

B. Muadil Soru Üretimi

İlk aşamada seçilen (50 doğru ve 50 yanlış cevaplanmış) asıl sorular kullanılarak, bu sorulara anlamsal olarak benzer ancak farklı ifadelerle "muadil" sorular üretilmiştir.

1) **Talimat (Prompt) Hazırlığı:** Muadil soru üretimi için 5 farklı talimat şablonu tanımlanmıştır. Bu talimatlar, Google Gemini API'sine (bu çalışmada `gemini-1.5-flash-002` modeli kullanılmıştır) asıl soru ile birlikte verilerek, sorunun farklı açılardan yeniden yazılması istenmiştir. Kullanılan talimatlar:
   - "Aşağıdaki matematik problemini, sayısal cevabı aynı kalacak şekilde farklı kelimeler ve farklı bir senaryo kullanarak yeniden yaz. Bana sadece oluşturduğun yeni soruyu ver. Başına sonuna veya ortasına başka bir şey ekleme."
   - "Bu probleme benzeyen, ancak farklı nesneler veya kişiler içeren ve aynı sayısal sonucu veren bir problem oluştur ve bana yaz."
   - "Verilen problemi daha basit ve anlaşılır bir dille ifade et, ama çözüm adımları ve nihai sayısal cevap kesinlikle değişmesin."
   - "Problemin konusunu (örneğin elmalar yerine kalemler, Ayşe yerine Mehmet gibi) değiştirerek, ana mantığı ve sayısal cevabı koruyarak yeni bir soru yaz."
   - "Bu soruyu, bir ilkokul öğrencisinin anlayabileceği şekilde yeniden formüle et, ancak sorunun asıl sayısal cevabını değiştirmeden yap."
2) **Gemini API ile Üretim:** Her bir asıl soru için, tanımlanan 5 talimatın her biri kullanılarak ve her talimat için 5 farklı deneme yapılarak (toplamda asıl soru başına 25 muadil soru) Gemini API'sine istek gönderilmiştir. API'den dönen yanıtlar, üretilen muadil soru metinleridir.
3) **Veri Kaydı:** Üretilen muadil sorular, asıl sorunun indeksi, kullanılan talimatın indeksi, deneme numarası ve asıl sorunun doğru cevabı ile birlikte ayrı CSV dosyalarına kaydedilmiştir. İşlemin kesintiye uğraması durumunda kalınan yerden devam edebilmek için, daha önce işlenmiş (asıl soru, talimat, deneme) üçlüleri takip edilmiş ve tekrarlı API çağrıları önlenmiştir.

C. Muadil Soruların Cevaplanması

Bir önceki adımda üretilen muadil sorular, yine `ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1` modeline çözdürülmüştür.

1) **Veri Yükleme:** Dosyalar okunmuştur.
2) **Model ile Çözüm:** Her bir muadil soru, `Turkish-Llama-8b-DPO-v0.1` modeline sunulmuş ve modelden cevap üretmesi istenmiştir.
3) **Cevap Doğruluk Kontrolü:** Modelin muadil soruya verdiği cevap ile muadil sorunun (dolayısıyla asıl sorunun) beklenen doğru cevabı karşılaştırılmıştır. Doğruluk kontrolü, *Veri Kümesi ve İlk Değerlendirme* bölümünde anlatılan sayısal değer ayıklama ve karşılaştırma yöntemiyle yapılmıştır.
4) **Sonuçların Kaydı:** Modelin muadil soruya verdiği cevap, ayıklanan sayısal değer ve cevabın doğruluk durumu, ilgili muadil soru bilgileriyle birlikte yeni CSV dosyalarına kaydedilmiştir. Bu aşamada da, işlem kaldığı yerden devam edebilmesi için işlenmiş kayıtlar takip edilmiştir.

D. Cevap Doğruluk Kontrolü Ayrıntıları

Modelin ürettiği cevapların doğruluğunu değerlendirmek için sürekli kullanılan bir yaklaşım izlenmiştir.

1) **Sayısal Değer Ayıklama:** Hem modelin ürettiği metin cevaptan hem de veri kümesindeki cevaptan sayısal değerler ayıklanmıştır. Bu işlem, genellikle metnin sonundaki sayısal ifadeyi (örneğin, "Cevap 42'dir." veya sadece "42") bulmayı amaçlar. Projede, bu amaçla python fonksiyonu yazılıp kullanılmıştır. Bu fonksiyon, metin içindeki çeşitli sayı biçimlerini (tam sayılar, ondalık sayılar) tanıyarak bunları standart bir sayısal türe (float) dönüştürür.
2) **Karşılaştırma:** Ayıklanan iki sayısal değer (modelin cevabı ve gerçek cevap) birbirleriyle karşılaştırılmıştır. Eğer iki değer birbirine eşitse, modelin cevabı "doğru" (True) olarak kabul edilmiş, aksi halde "yanlış" (False) olarak etiketlenmiştir.
3) **Özel Durumlar:** Bazı durumlarda (mesela, cevap "Öğrencilerin yüzde otuzu okulda bulunmaktadır." gibi bir ifade içeriyorsa), yukarıdaki yöntemle sayısal ayıklama imkansızdır. Bu tür durumlar, ilk soru alımı aşamasında gözle kontrol edilip düzeltilmiştir. Ancak muadil soruların cevaplanması aşamasında, modelden sadece sayısal bir çıktı vermesi istendiği için bu soruna rastlanmamıştır. Eğer model veya gerçek cevap metninden sayısal bir değer ayıklanamazsa, bu durum yanlış cevap olarak değerlendirilmiştir.

E. Gömme Vektörlerinin Hesaplanması

Analizin bir sonraki adımı, asıl soruların, üretilen muadil soruların ve bu sorulara model tarafından verilen cevapların metin gömme (embedding) vektörlerini hesaplamaktır.

1) **Gömme Modeli:** Metin gömme vektörlerini oluşturmak için `ytu-ce-cosmos/turkish-e5-large` modeli kullanılmıştır. Bu model, Türkçe metinler için yüksek kaliteli gömme vektörleri üretme yeteneğine sahiptir.
2) **Vektör Hesaplama:**
   - asıl sorular, asıl cevap metinleri ve modelin bu asıl sorulara verdiği cevap metinleri.
   - Gemini tarafından oluşturulan muadil sorular ve modelin bu muadil sorulara verdiği cevap metinleri.

   Yukarıda listelenen tüm metinler için ayrı ayrı gömme vektörleri hesaplanmıştır. Her bir metin, `turkish-e5-large` modeline girdi olarak verilmiş ve modelin çıktısı olan 1024 boyutlu vektör 1024 alınmıştır.
3) **Sonuçların Kaydı:** Hesaplanan gömme vektörleri, ilgili metinler ve doğruluk bilgileriyle birlikte yeni CSV dosyalarına kaydedilmiştir.

F. Boyut İndirgeme ve Görselleştirme

Yüksek boyutlu gömme vektörlerini insan tarafından yorumlanabilir hale getirmek için t-SNE (t-distributed Stochastic Neighbor Embedding) algoritması kullanılarak boyut indirgeme işlemi yapılmıştır.

1) **t-SNE Uygulaması:** Bir önceki adımda hesaplanan gömme vektörleri (asıl soruların gömmeleri, muadil soruların gömmeleri vb.) t-SNE algoritmasına girdi olarak verilmiştir. t-SNE, bu yüksek boyutlu vektörleri 2 boyutlu bir uzaya haritalamıştır. Rastgelelik içeren t-SNE sürecinin tekrarlanabilirliği için `random_seed` değeri 571 olarak ayarlanmıştır.
2) **Görselleştirme:** Elde edilen 2 boyutlu t-SNE koordinatları, Matplotlib kütüphanesi kullanılarak saçılım grafikleri (scatter plots) şeklinde görselleştirilmiştir. Her bir nokta, bir soruyu (veya cevabı) temsil etmektedir. Noktalar, modelin ilgili soruya verdiği cevabın doğruluğuna göre renklendirilmiştir (doğru cevaplar mavi, yanlış cevaplar kırmızı).
3) **Farklı Veri Kümeleri İçin Grafikler:** Aşağıdaki veri alt kümeleri için ayrı ayrı t-SNE grafikleri oluşturulmuştur.
   - Asıl soruların t-SNE dağılımı.
   - Modelin doğru cevapladığı asıl sorulardan türetilen muadil soruların t-SNE dağılımı.
   - Modelin yanlış cevapladığı asıl sorulardan türetilen muadil soruların t-SNE dağılımı.
   - Tüm soruların (asıl ve muadillerin birleştirilmiş hali) t-SNE dağılımı.
   - Talimat türüne göre gruplandırılmış muadil soruların t-SNE dağılımları.

III. Kullanılan Modeller

Proje boyunca farklı amaçlar için çeşitli önceden eğitilmiş dil modelleri kullanılmıştır. Bu modeller ve kullanım amaçları Tablo I de özetlenmiştir.

Tablo I
Projede Kullanılan Modeller ve Amaçları

Model Adı	Kullanım Amacı
`gemini-1.5-flash-002`	Veri artırımı (Muadil soru üretimi)
`ytu-ce-cosmos/turkish-e5-large`	Gömme hesaplama
`ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1`	Soru cevaplama (Asıl ve Muadil)

`gemini-1.5-flash-002`, Google tarafından geliştirilen ve metin üretimi konusunda yetenekli bir model olup, bu projede asıl sorulara benzer yeni sorular (muadiller) oluşturmak için kullanılmıştır. `ytu-ce-cosmos/turkish-e5-large`, özellikle Türkçe metinler için etkili gömme vektörleri üreten bir modeldir ve t-SNE görselleştirmeleri için temel oluşturmuştur. `ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1` ise, projenin ana odak noktası olan ve hem asıl `gsm8k_tr` sorularını hem de üretilen muadil soruları çözmek için kullanılan dil modelidir.

IV. Sonuçlar ve Tartışma

A. İlk Değerlendirme ve Soru Seçimi Safhası

Projenin ilk aşamasında, `ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1` modeline `gsm8k_tr` veri kümesinden rastgele seçilen sorular çözdürülmüştür. Hedef, modelin 50 soruyu doğru ve 50 soruyu yanlış cevapladığı bir alt küme oluşturmaktır.

- Model 50. yanlışı yaptığında henüz toplamda 74 soru cevaplamıştır. Bu, modelin başarımının matematik için pek de iyi olmadığını olmadığını ve yanlış cevap oranının dikkate değer olduğunu göstermektedir.
- Cevap doğruluk kontrolü sırasında, "Öğrencilerin yüzde otuzu okulda bulunmaktadır." gibi metinsel veya yüzde içeren cevapların otomatik olarak sayısal değere dönüştürülmesinde zorluklar yaşanmıştır. Bu tür durumlar, ilk soru alımı aşamasında gözle kontrol edilerek ve cevaplar standart bir sayısal biçime elle getirilerek çözülmüştür. Bu, otomatik değerlendirme sistemlerinin sınırlarını ve bazı durumlarda insan müdahalesinin gerekliliğini göstermektedir.

B. Muadil Soru Üretim Safhası

Seçilen 100 asıl soru (50 doğru, 50 yanlış cevaplanmış) için `gemini-1.5-flash-002` modeli ve 5 farklı talimat kullanılarak muadil sorular üretilmiştir.

- Kullanılan talimatlar, sorunun senaryosunu, kelimelerini, kullanılan nesne/kişileri değiştirmeyi veya soruyu daha basit bir dille ifade etmeyi amaçlıyordu, ancak temel mantık ve sayısal cevap korunmalıydı.
- Sunumda belirtilen bir zorluk, özellikle "Bu probleme benzeyen, ancak farklı nesneler veya kişiler içeren ve aynı sayısal sonucu veren bir problem oluştur ve bana yaz." talimatı kullanıldığında, Gemini modelinin bazen cevaplarına "Orijinal Problem:" veya "Yeni Problem:" gibi istenmeyen ön ekler eklemesiydi. Bu durum, üretilen soruların yaklaşık %5'inde gözlemlenmiş ve bu ön ekler elle temizlenmiştir. Bu, Büyük DİL Modellerinin talimatları harfiyen takip etme konusunda bazen zorlanabildiğini ve çıktıların bir miktar son işleme gerektirebileceğini göstermektedir.
- Bir diğer gözlem, modelin bazen tekrarlayan sorular üretmesiydi. Bu tür durumlar tespit edildiğinde, sorular elenmiş ve modelden tekrar üretim yapması istenmiştir. Modelin tekrarda ısrarcı olduğu nadir durumlarda ise soru elle değiştirilmiştir. Bu, üretim sürecinde çeşitliliği sağlamanın ve kalite kontrolünün önemini vurgulamaktadır.

Sonuç olarak, her bir asıl soru için 25 muadil soru (5 talimat x 5 deneme) olmak üzere toplamda $50 \times 25 = 1250$ "doğru cevapların muadili" ve $50 \times 25 = 1250$ "yanlış cevapların muadili" soru üretilmiştir.

C. Muadil Soruların Cevaplanma İstatistikleri

Üretilen muadil sorular, `ytu-ce-cosmos/Turkish-Llama-8b-DP` modeline çözdürülmüş ve cevaplanma istatistikleri Tablo II de gösterilmiştir.

Tablo II
MUADIL SORULARIN CEVAPLANMASI SAFHASI İSTATISTIKLERI

Soruların Türü	Toplam Soru Sayısı	Yanlış Cevaplananların Sayısı	Doğru Cevaplananların Sayısı
Yanlış Cevaplanan Soru Muadilleri	1250	1018 (%81,44)	232 (%18,56)
Doğru Cevaplanan Soru Muadilleri	1250	675 (%54)	575 (%46)

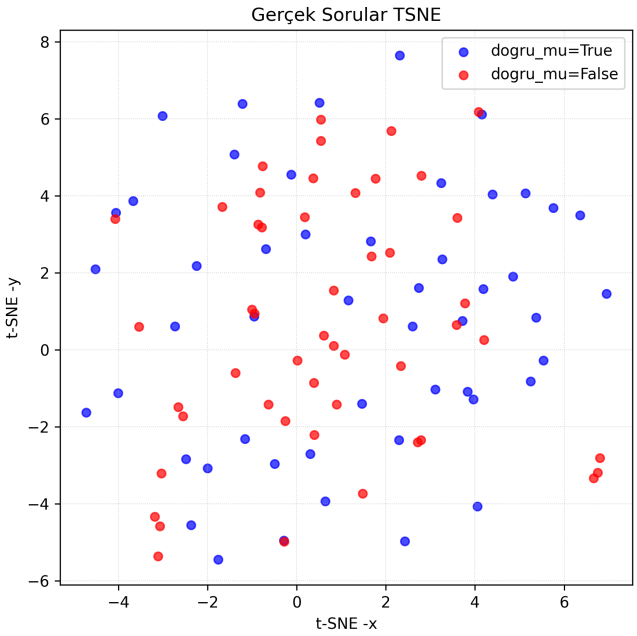
Yorumlar:

- **Yanlış Cevaplanan Soru Muadilleri:** Asılında modelin yanlış cevapladığı soruların muadillerinde, modelin yine büyük bir çoğunlukla (%81,44) yanlış cevap verdiği görülmektedir. Sadece %18,56'sını doğru cevaplayabilmiştir. Bu, modelin bir soruyu anlamakta veya çözmekte zorlandığı temel bir nokta varsa, sorunun farklı şekillerde ifade edilmesinin bu zorluğu her zaman aşamadığını göstermektedir. Temel kavramsal eksiklik veya yanlış akıl yürütme, ifade değişikliğine rağmen devam edebilir.
- **Doğru Cevaplanan Soru Muadilleri:** Asılında modelin doğru cevapladığı soruların muadillerinde ise başarı oranı %46'ya düşmüştür. Yani, model asıl soruyu doğru çözebilmiş olmasına rağmen, bu sorunun farklı bir ifadesini çözmekte zorlanmış ve %54 oranında yanlış cevap vermiştir. Bu durum, modelin çözümünün bir dereceye kadar asıl sorunun özel ifadesine ve yapısına bağlı olduğunu, anlamsal eşdeğerliği tam olarak koruyamadığını ve ifade değişikliğinin model için yeni bir zorluk unsuru getirdiğini düşündürmektedir. Aslında modelin "anlayışı" yüzeyseldir ve farklı ifade biçimleri bu yüzeyselliği ortaya çıkarmıştır.

Bu sonuçlar, LLM'lerin problem çözme yeteneklerinin sadece doğru cevabı bulmaktan öte, sorunun farklı ifade biçimlerine karşı ne kadar dirençli ve tutarlı olduğuyla da ölçülmesi gerektiğini göstermektedir.

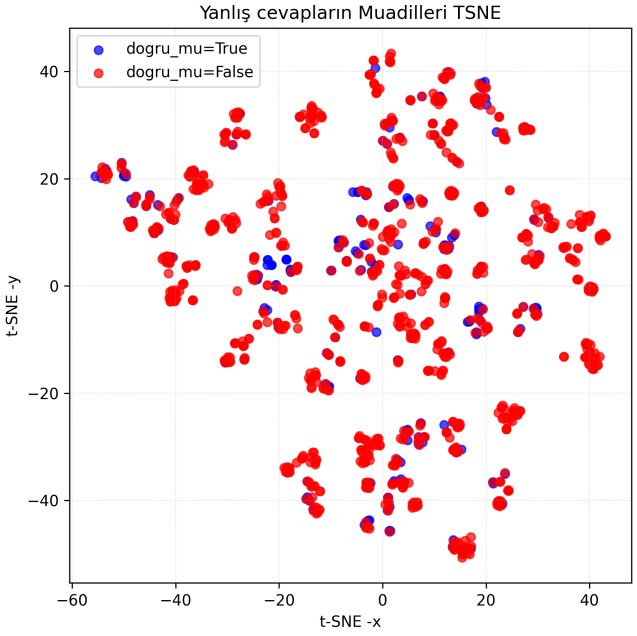
D. t-SNE Görselleştirmeleri ve Yorumları

Hesaplanan gömme vektörleri t-SNE ile 2 boyuta indirgenerek çeşitli soru grupları için görselleştirmeler yapılmıştır. Bu görsellerde mavi noktalar modelin doğru cevapladığı, kırmızı noktalar ise yanlış cevapladığı soruları/cevapları temsil etmektedir.



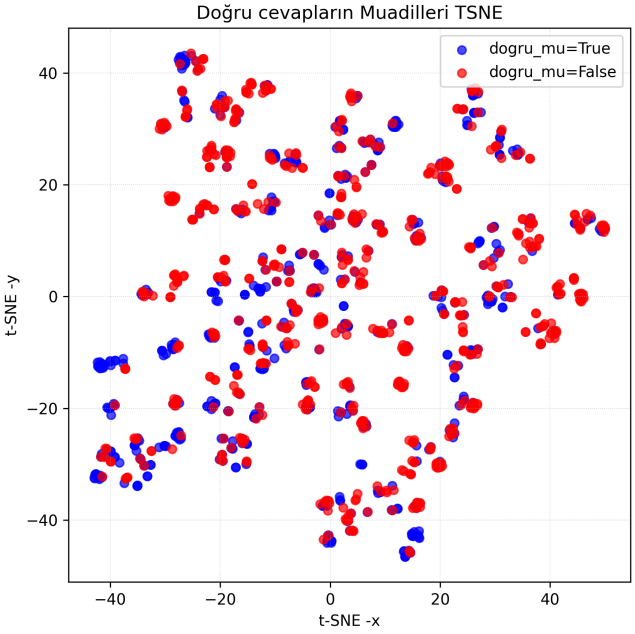
Şekil 1. Gerçek Sorular t-SNE Dağılımı

Şekil 1 (Gerçek Sorular): Asıl 100 sorunun (50 doğru, 50 yanlış cevaplanmış) t-SNE dağılımı incelendiğinde, doğru (mavi) ve yanlış (kırmızı) cevaplanan sorular arasında belirgin bir kümelenme veya ayrışma gözlenmemektedir. Noktalar büyük ölçüde iç içe geçmiş durumdadır. Bu, sadece soru metninin gömme vektörüne bakarak modelin o soruyu doğru çözüp çözemeyeceğini tahmin etmenin zor olduğunu gösterir. Sorunun zorluğu veya modelin başarısızlığı, gömme uzayında basitçe ayrıştırılabilir özelliklere sahip değildir.



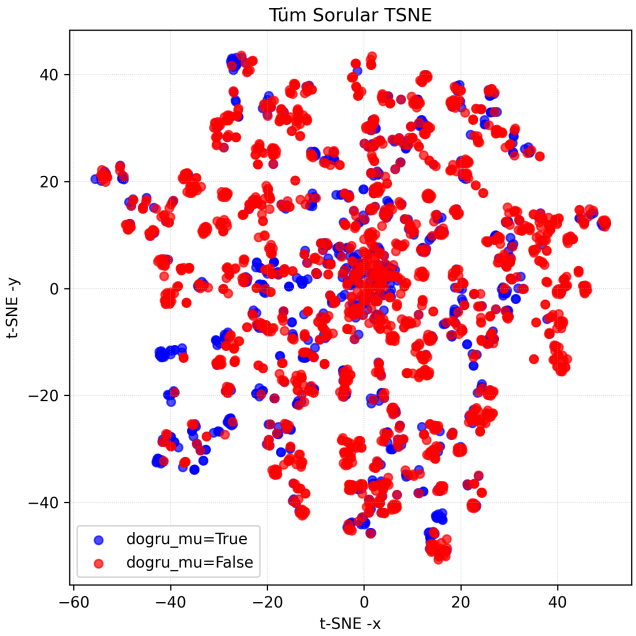
Şekil 2. Yanlış Cevapların Muadilleri t-SNE Dağılımı

Şekil 2 (Yanlış Cevapların Muadilleri): Asılında yanlış cevaplanan soruların muadillerinin t-SNE dağılımında, kırmızı noktaların (yanlış cevaplar) yoğunluğu dikkat çekicidir. Mavi noktalar (doğru cevaplar) daha dağınık ve azınlıktadır. Yine de, doğru ve yanlış cevaplar arasında net bir ayrım veya kümelenme gözlenmemektedir. Bu, muadil soruların da gömme özelliklerinin, modelin başarısını doğrudan yansıtmadığını gösterir.



Şekil 3. Doğru Cevapların Muadilleri t-SNE Dağılımı

Şekil 3 (Doğru Cevapların Muadilleri): Asılında doğru cevaplanan soruların muadillerinin t-SNE dağılımında, kırmızı (yanlış) ve mavi (doğru) noktaların sayısı birbirine daha yakındır (Tablo II'de %54 yanlış, %46 doğru). Bu grafikte de belirgin bir ayrışma veya kümelenme görülmemektedir. Bu, modelin asıl soruyu doğru çözmüş olmasının, muadilinin gömme uzayında "kolay" bir bölgeye düşeceği anlamına gelmediğini gösterir.



Şekil 4. Tüm Sorular (Asıl ve Muadiller) t-SNE Dağılımı

Şekil 4 (Tüm Sorular): Tüm sorular (asıl ve tüm muadiller) birleştirildiğinde elde edilen t-SNE dağılımı oldukça yoğundur. Kırmızı noktaların genel bir hakimiyeti vardır, bu da modelin muadil sorularda genel olarak zorlandığını gösterir. Bu birleşik grafikte de doğru ve yanlış cevaplar arasında net bir ayrım çizgisi çekmek mümkün değildir. Bu, kullanılan gömme modelinin (`turkish-e5-large`) yakaladığı anlamsal özelliklerin, `Turkish-Llama-8b-DPO-v0.1` modelinin problem çözme mekanizmasındaki başarı/başarısızlık nedenlerini tam olarak ayırt edemediğini veya bu nedenlerin gömme uzayında basitçe ayrışmadığını gösterir.

E. Talimat Bazlı Sonuçlar ve Görselleştirmeler

Muadil soruların cevaplanma başarısı, kullanılan üretim talimatına göre de incelenmiştir.

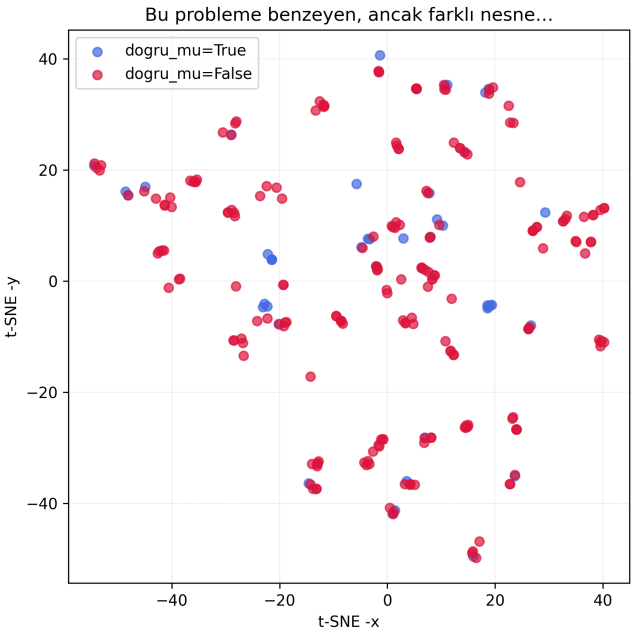
1) Yanlış Cevaplanan Asıl Soruların Muadilleri (Talimat Bazlı): Tablo III, asılında yanlış cevaplanan sorular için üretilen muadillerin, hangi talimatla üretildiklerine bağlı olarak `Turkish-Llama-8b-DPO-v0.1` tarafından ne kadarının doğru veya yanlış cevaplandığını göstermektedir. Her talimat için 250 muadil soru üretilmiştir (50 asıl yanlış soru x 5 deneme).

Tablo III
TALIMATLARA GÖRE OLUŞTURULAN YANLIŞ CEVAP MUADIL SORULARININ CEVAPLANMASI

Talimat	Yanlış Sayısı	Doğru Sayısı
Aşağıdaki matematik problemini, sayısal cevabı aynı kalacak şekilde farklı kelimeler ve farklı bir senaryo kullanarak yeniden yaz. Bana sadece oluşturduğun yeni soruyu ver. Başına sonuna veya ortasına başka bir şey ekleme.	203	47
Bu probleme benzeyen, ancak farklı nesneler veya kişiler içeren ve aynı sayısal sonucu veren bir problem oluştur ve bana yaz.	206	44
Verilen problemi daha basit ve anlaşılır bir dille ifade et, ama çözüm adımları ve nihai sayısal cevap kesinlikle değişmesin.	205	45
Problemin konusunu (örneğin elmalar yerine kalemler, Ayşe yerine Mehmet gibi) değiştirerek, ana mantığı ve sayısal cevabı koruyarak yeni bir soru yaz.	196	54
Bu soruyu, bir ilkokul öğrencisinin anlayabileceği şekilde yeniden formüle et, ancak sorunun asıl sayısal cevabını değiştirmeden yap.	208	42

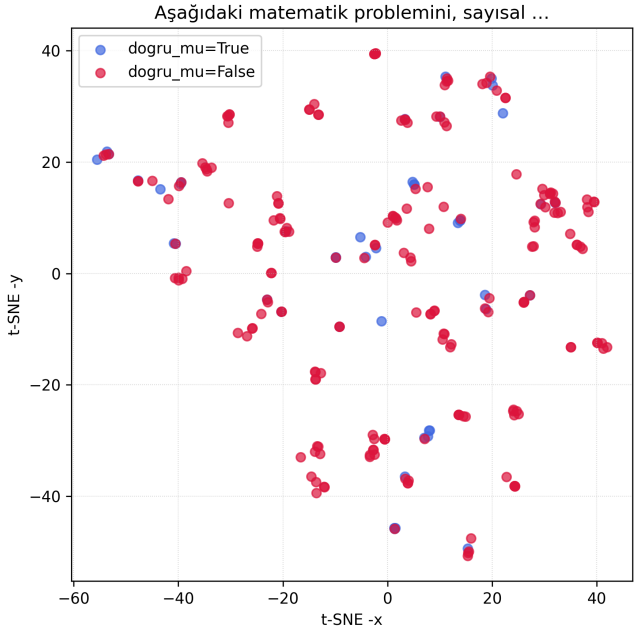
Yorumlar (Tablo III):

- Genel olarak, tüm talimat türlerinde modelin başarı oranı düşüktür (doğru cevap sayısı 42 ile 54 arasında değişmektedir), bu da asıldaki zorluğun muadillere de yansıdığını teyit eder.
- ”Problemin konusunu ... değiştirerek ... yeni bir soru yaz.” talimatıyla üretilen muadillerde görece en yüksek doğru cevap sayısına (54) ulaşılmıştır. Bu, belki de sadece konunun değişmesinin, temel mantık aynı kaldığı için modelin bazen çözüme ulaşmasını bir nebze kolaylaştırdığını düşündürebilir.
- En düşük başarı (”Bu soruyu, bir ilkokul öğrencisinin anlayabileceği şekilde ... formüle et...”) talimatında görülmüştür. Soruyu basitleştirme çabası, model için yeni belirsizlikler veya yanlış yorumlamalara yol açmış olabilir ya da basitleştirilmiş ifade, modelin alıştığı problem yapılarından uzaklaşmış olabilir.

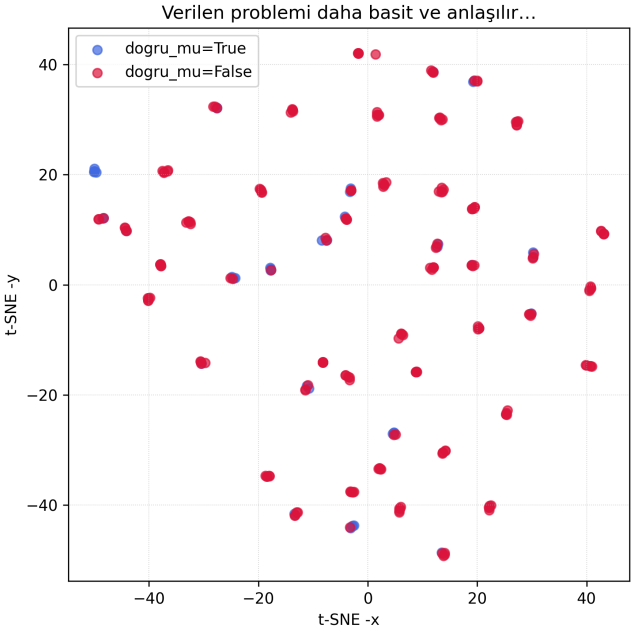


Şekil 6. Yanlış Muadiller - Talimat 2 t-SNE

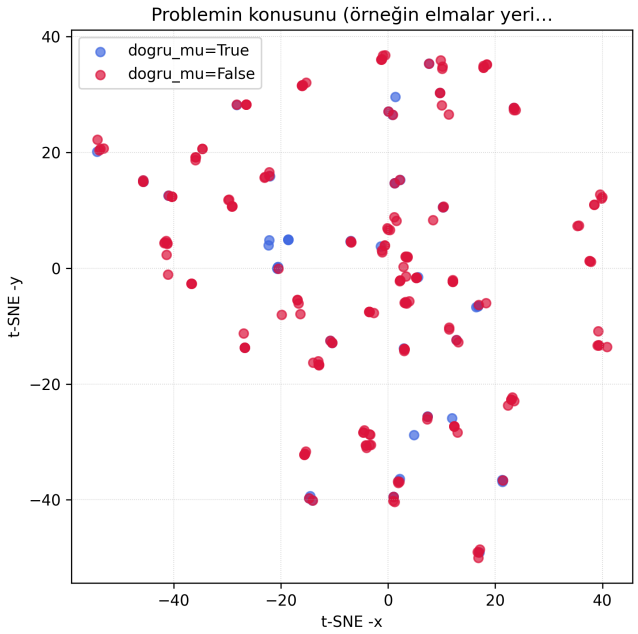
Bu talimatlara karşılık gelen t-SNE grafikleri (Şekil 5 - 9) incelendiğinde, hiçbir talimat türünün doğru ve yanlış cevapları belirgin şekilde ayıran bir kümelenme oluşturmadığı görülmektedir. Kırmızı noktalar her zaman baskındır.



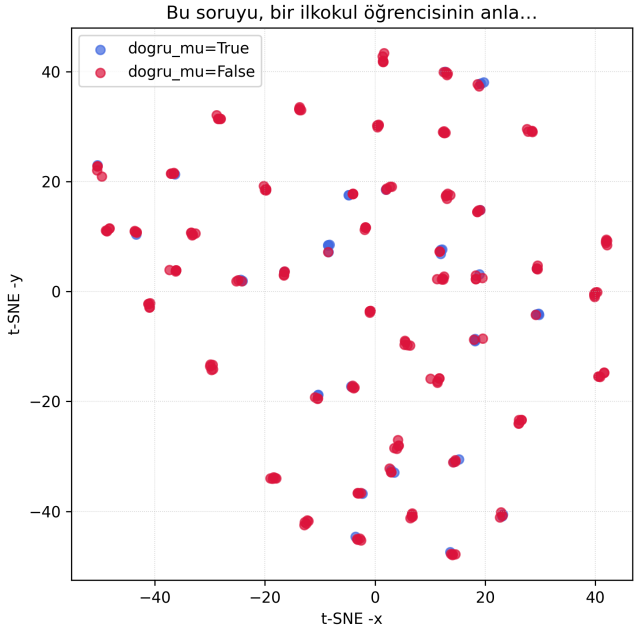
Şekil 5. Yanlış Muadiller – Talimat 1 t-SNE



Şekil 7. Yanlış Muadiller - Talimat 3 t-SNE



Şekil 8. Yanlış Muadiller - Talimat 4 t-SNE



Şekil 9. Yanlış Muadiller - Talimat 5 t-SNE

2) Doğru Cevaplanan Asıl Soruların Muadilleri (Talimat Bazlı): Tablo IV, asılda doğru cevaplanan sorular için üretilen muadillerin, hangi talimatla üretildiklerine bağlı olarak model tarafından ne kadarının doğru veya yanlış cevaplandığını göstermektedir. Her talimat için 250 muadil soru üretilmiştir.

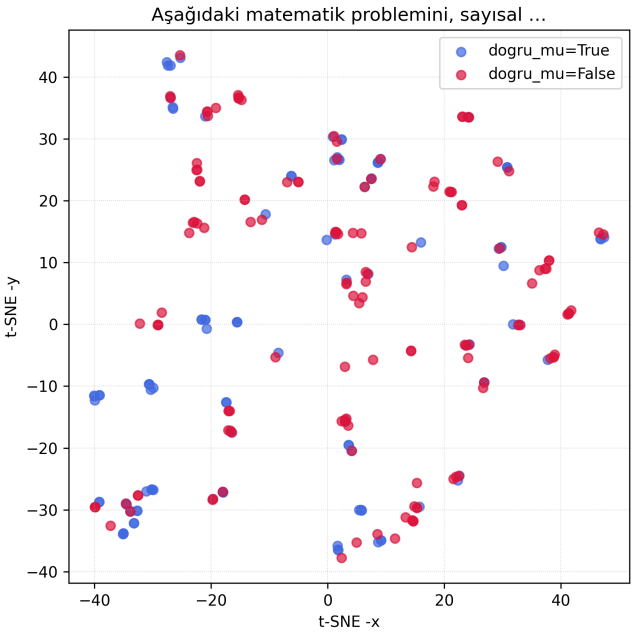
Tablo IV
TALIMATLARA GÖRE OLUŞTURULAN DOĞRU CEVAP MUADIL SORULARININ CEVAPLANMASI

Talimat	Yanlış Sayısı	Doğru Sayısı
Aşağıdaki matematik problemini, sayısal cevabı aynı kalacak şekilde farklı kelimeler ve farklı bir senaryo kullanarak yeniden yaz. Bana sadece oluşturduğun yeni soruyu ver. Başına sonuna veya ortasına başka bir şey ekleme.	137	113
Bu probleme benzeyen, ancak farklı nesneler veya kişiler içeren ve aynı sayısal sonucu veren bir problem oluştur ve bana yaz.	128	122
Verilen problemi daha basit ve anlaşılır bir dille ifade et, ama çözüm adımları ve nihai sayısal cevap kesinlikle değişmesin.	113	137
Problemin konusunu (örneğin elmalar yerine kalemler, Ayşe yerine Mehmet gibi) değiştirerek, ana mantığı ve sayısal cevabı koruyarak yeni bir soru yaz.	137	113
Bu soruyu, bir ilkokul öğrencisinin anlayabileceği şekilde yeniden formüle et, ancak sorunun asıl sayısal cevabını değiştirmeden yap.	160	90

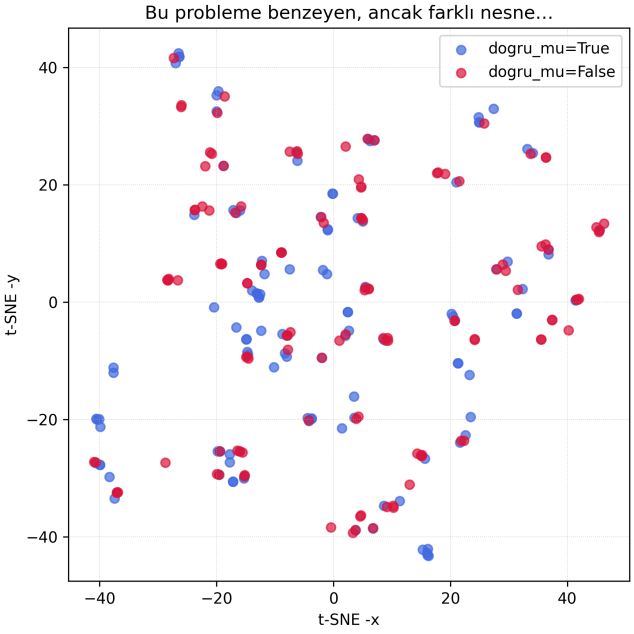
Yorumlar (Tablo IV):

- Asılı doğru çözülmüş soruların muadillerinde, ”Verilen problemi daha basit ve anlaşılır bir dille ifade et...” talimatıyla üretilen sorular en yüksek doğru cevap oranına (137 doğru) sahiptir. Bu, sorunun basitleştirilmesinin, modelin doğru çözüme ulaşmasını kolaylaştırabildiğini göstermektedir.
- İlginç bir şekilde, yine ”Bu soruyu, bir ilkokul öğrencisinin anlayabileceği şekilde ... formüle et...” talimatı en düşük başarıyı (90 doğru) vermiştir. Bu, ”basit ve anlaşılır dil” ile ”ilkokul seviyesine indirgeme” arasında modelin farklı tepkiler verdiğini, ikincisinin beklenenin aksine kafa karıştırıcı olduğunu göstermektedir. Muhtemelen ”ilkokul seviyesi” ifadesi, modelin aşırı basitleştirmesine veya problemdeki önemli nüansları kaybetmesine neden olmuştur.
- Diğer talimatlar (senaryo değiştirme, nesne/kişi değiştirme, konu değiştirme) birbirine yakın sonuçlar vermiştir ve başarı oranları %50 civarındadır. Bu, ifade biçimindeki değişikliklerin model için hala önemli bir zorluk teşkil ettiğini göstermektedir.

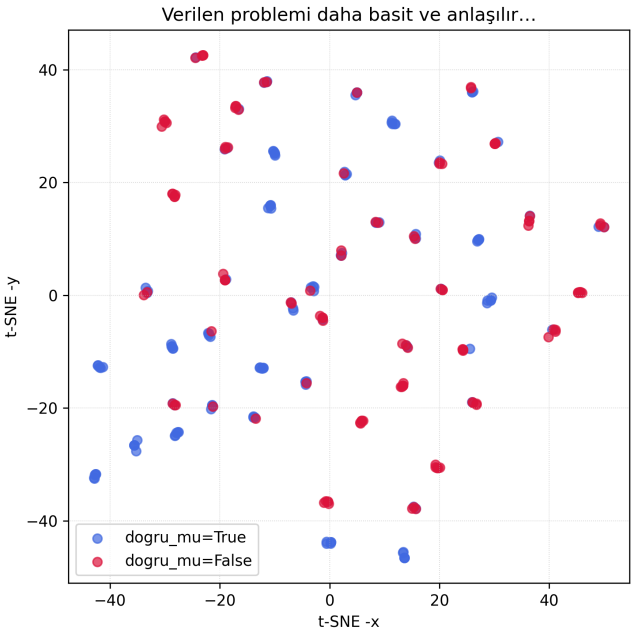
Bu talimatlara karşılık gelen t-SNE grafikleri (Şekil 10 - 14) incelendiğinde, yine belirgin bir ayrışma olmamakla birlikte, ”Verilen problemi daha basit...” talimatına ait grafikte (Şekil 12) mavi noktaların (doğru cevaplar) diğerlerine kıyasla biraz daha belirgin olduğu söylenebilir, ancak bu fark çok çarpıcı değildir.



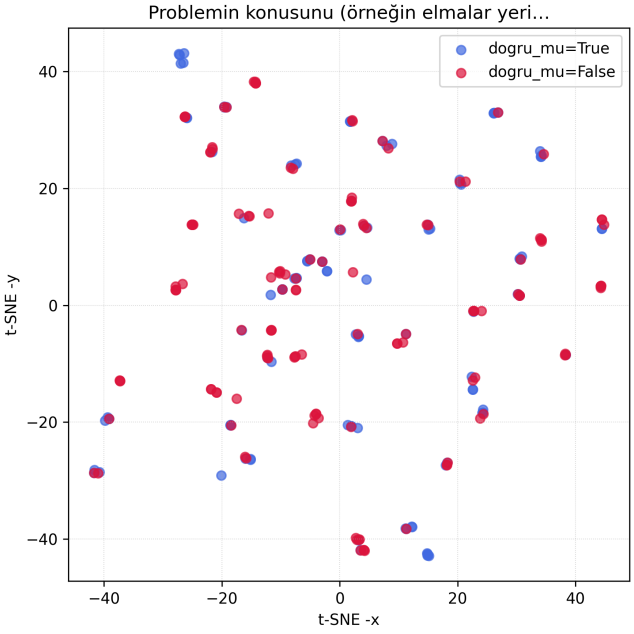
Şekil 10. Doğru Muadiller - Talimat 1 t-SNE



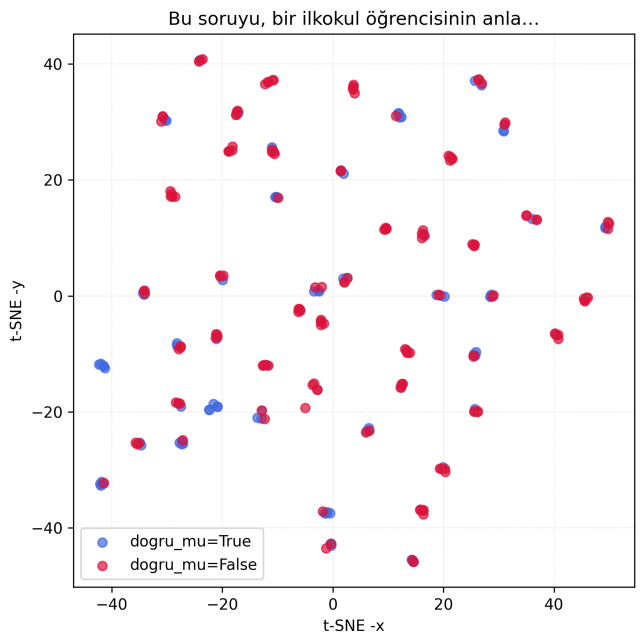
Şekil 11. Doğru Muadiller - Talimat 2 t-SNE



Şekil 12. Doğru Muadiller - Talimat 3 t-SNE



Şekil 13. Doğru Muadiller - Talimat 4 t-SNE



Şekil 14. Doğru Muadiller - Talimat 5 t-SNE

V. Genel Değerlendirme ve Gelecek Çalışmalar

Bu proje, `ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1` modelinin matematiksel problem çözme yeteneğini, özellikle soruların farklı ifade biçimlerine karşı tutarlılığını incelemiştir. Elde edilen sonuçlar, modelin bir soruyu asıl formunda doğru çözebilmesinin, o sorunun anlamsal olarak eşdeğer fakat farklı ifade edilmiş bir sürümünü de doğru çözeceği anlamına gelmediğini göstermiştir. Asılında doğru çözülen soruların muadillerinde bile başarı oranı %46'ya düşmüştür. Asılda yanlış çözülen soruların muadillerinde ise modelin başarısızlığı büyük ölçüde devam etmiştir.

t-SNE görselleştirmeleri, soru metinlerinin gömme vektörleri ile modelin cevap doğruluğu arasında basit ve doğrudan bir ilişki olmadığını ortaya koymuştur. Doğru ve yanlış cevaplanan sorular, gömme uzayında belirgin şekilde ayrışmamıştır. Bu durum, modelin problem çözme mekanizmasının karmaşıklığını ve sadece metin benzerliğinin ötesinde faktörlerin etkili olduğunu düşündürmektedir.

Gelecek Çalışmalar İçin Öneriler:

- **Farklı Gömme Modelleri:** Daha gelişmiş veya farklı mimarilere sahip gömme modelleri kullanarak t-SNE analizleri tekrarlanabilir. Belki farklı bir gömme modeli, doğru/yanlış cevaplar arasında daha iyi bir ayrım sağlayabilir.
- **Modelin İç Temsilleri:** Sadece soru metninin değil, modelin problem çözme sırasında ürettiği iç temsillerin analizi, başarısızlık nedenleri hakkında daha fazla bilgi verecektir.
- **Hata Analizi:** Modelin yanlış cevapladığı muadil sorular detaylı incelenerek, ne tür hatalar yaptığı kategorize edilebilir.
- **Daha Fazla Talimat Çeşidi:** Muadil soru üretimi için daha çeşitli ve özel talimatlar denenebilir.
- **Farklı LLM'ler ile Karşılaştırma:** Aynı analiz, farklı LLM'ler üzerinde tekrarlanarak modellerin ifade değişikliklerine karşı ne kadar dirençli oldukları karşılaştırılabilir.
- **İnce Ayar Yöntemleri:** Modelin, soruların farklı ifade biçimlerine karşı daha dirençli hale gelmesi için özel veri artırma ve ince ayar yöntemleri geliştirilebilir.