3. Ders

YZ Çıktılarını Değerlendirme ve Doğrulama Teknikleri

Güvenilirlik, Doğrulama ve İnsan Denetimi

Fettah KIRAN
YZ ve Veri Mühendisliği
fkiran@yildiz.edu.tr

Dersin Kapsamı

- Yapay Zekanın Güçlü ve Zayıf Yönleri
- Yapay Zeka Çıktılarında Yaygın Hatalar
- Hallucination: Yanlış ve Uydurma Bilgi Sorunu
- Yapay Zeka Çıktılarında Mantık ve Tutarlılık Kontrolü
- Kaynak Doğrulama ve Eleştirel Değerlendirme
- İnsan Denetimi ve Bilişsel Becerilerin Korunması

[ETKİLEŞİM] Isınma: Ne Kadar Güveniyorsunuz?

Yapay zekanın size verdiği bir cevabı, başka bir kaynaktan kontrol etmeden doğrudan kullandığınız oldu mu?

- Eğer olduysa, sonucunda bir hata fark ettiniz mi?
- Yanınızdaki arkadaşınızla deneyimlerinizi paylaşın.

Yapay Zekanın Güçlü Yönleri

- **Veri İşleme Hızı:** Devasa miktardaki bilgiyi saniyeler içinde analiz edebilir.
- **Kalıp Tanıma:** Karmaşık veriler arasındaki gizli ilişkileri ve örüntüleri yakalayabilir.
- **7/24 Erişilebilirlik:** Her an kesintisiz destek sunabilir.
- **Ölçeklenebilirlik:** Aynı anda çok sayıda görevi yorulmadan yerine getirebilir.



Yapay Zekanın Zayıf Yönleri

- **Bağlam Eksikliği:** Verilen bilgilerin sosyal, kültürel veya etik derinliğini tam olarak kavrayamaz.
- **Gerçek Zamanlı Bilgi Kısıtı:** Eğitim verisinin kesildiği tarihten sonraki olaylara dair "uydurma" bilgi üretebilir.
- **Duygusal Zeka Yoksunluğu:** İnsani duyguları ve ince detayları (ironi, alt metin vb.) anlamlandırmada zayıftır.



YZ Çıktılarında Hata Türleri: Olgusal Hatalar

- **Tanım:** Tarihler, isimler, yerler veya teknik verilerdeki yanlışlıklardır.
- **Örnek:** Bir elementin erime noktasını yanlış söylemesi veya tarihi bir olayı farklı bir yıla atfetmesi.
- **Neden:** Eğitim verisindeki hatalar veya olasılıksal kelime seçimi.

YZ Çıktılarında Hata Türleri: Mantıksal Hatalar

- **Tanım:** Sebep-sonuç ilişkisinin yanlış kurulması veya hatalı çıkarımlardır.
- **Örnek:** "A maddesi B'den ağırdır, B maddesi C'den ağırdır; o halde C en ağırdır" gibi hatalı akıl yürütmeler.
- **Risk:** Çok düzgün bir dille sunulduğu için bu hataları fark etmek zordur.

YZ Çıktılarında Hata Türleri: Bağlamsal Hatalar

- **Tanım:** Sorulan sorunun kapsamı dışına çıkılması veya bağlamın yanlış anlaşılmasıdır.
- **Örnek:** Kimyasal bir süreci sorarken, cevabın mutfak yemekleri tarifine dönüşmesi.
- **Süreklilik Kaybı:** Uzun diyaloglarda konunun başındaki kısıtlamaların unutulması.

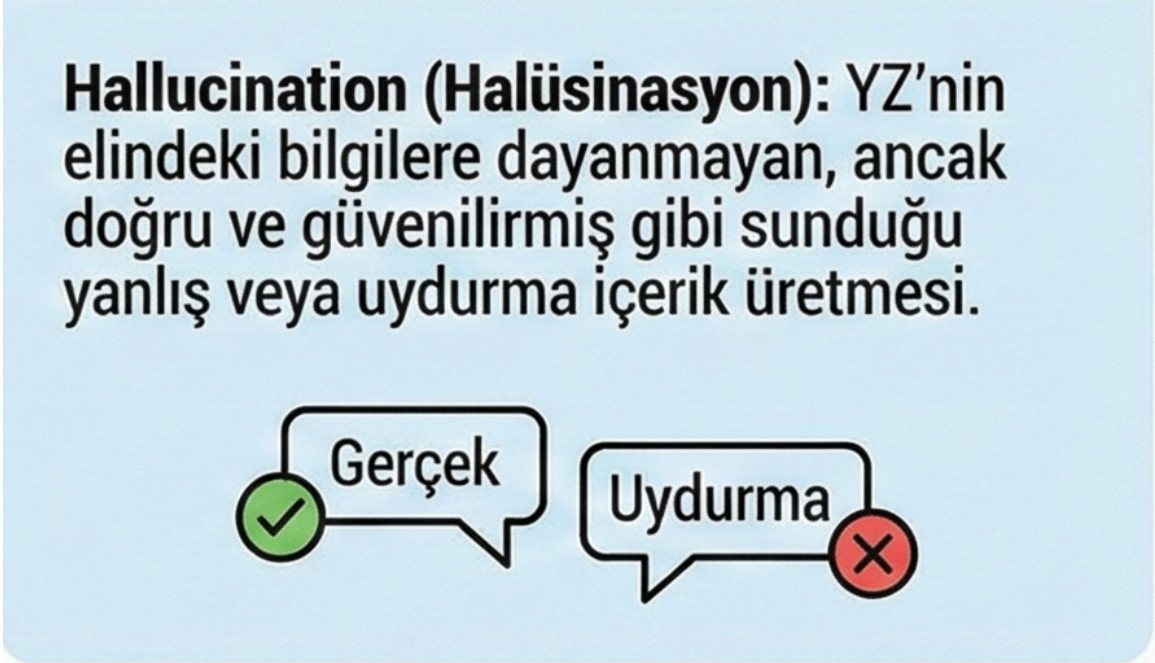
[ETKİLEŞİM] Hata Avcısı

Görev: "Güneş sistemindeki en soğuk gezegen hangisidir?" sorusuna YZ "Mars" cevabını verirse bu ne tür bir hatadır?

1. Olgusal mı?
2. Mantıksal mı?
3. Bağlamsal mı?

Hallucination: Yanlış ve Uydurma Bilgi Sorunu

- **Halüsinasyon Nedir?** YZ'nin elinde veri olmamasına rağmen, son derece emin bir dille, ikna edici ama tamamen uydurma bilgiler üretmesidir.
- **Özellik:** Bilgi uydururken gramer olarak kusursuzdur, bu da kullanıcının güvenini suiistimal eder.



YZ Neden Halüsinasyon Görür?

- **Olasılıksal Tahmin:** YZ bir sonraki kelimeyi tahmin ederken, "doğru" olana değil, "en muhtemel görünene" odaklanır.
- **Eğitim Verisi Boşlukları:** Modelin eğitim setinde bulunmayan konularda boşlukları "yaratıcılık" ile doldurmaya çalışması.
- **Prompt Yönlendirmesi:** Kullanıcının soruyu soruş şeklinin YZ'yi bir cevap uydurmaya zorlaması.

Halüsinasyon Örnekleri: Uydurma Referanslar

- Olmayan makale başlıkları oluşturma.
  - *"Prof. Dr. Ahmet Yılmaz'ın 2022 yılında Nature dergisinde yayımlanan 'Kuantum Biyolojisinde Yeni Yaklaşımlar' başlıklı makalesine göre..." (Gerçekte Ahmet Yılmaz'ın böyle bir makalesi yoktur).*
- Mevcut yazarlara yazmadıkları kitapları atfetme.
  - *Örnek: "Albert Camus'nün Gecenin Sonundaki Işık adlı deneme kitabında belirttiği gibi..." (Camus'nün böyle bir kitabı bulunmamaktadır).*
- Rastgele, geçersiz DOI numaraları veya URL üretme.
  - *Örnek: https://www.who.int (Böyle bir rapor bağlantısı mevcut olmayabilir).*

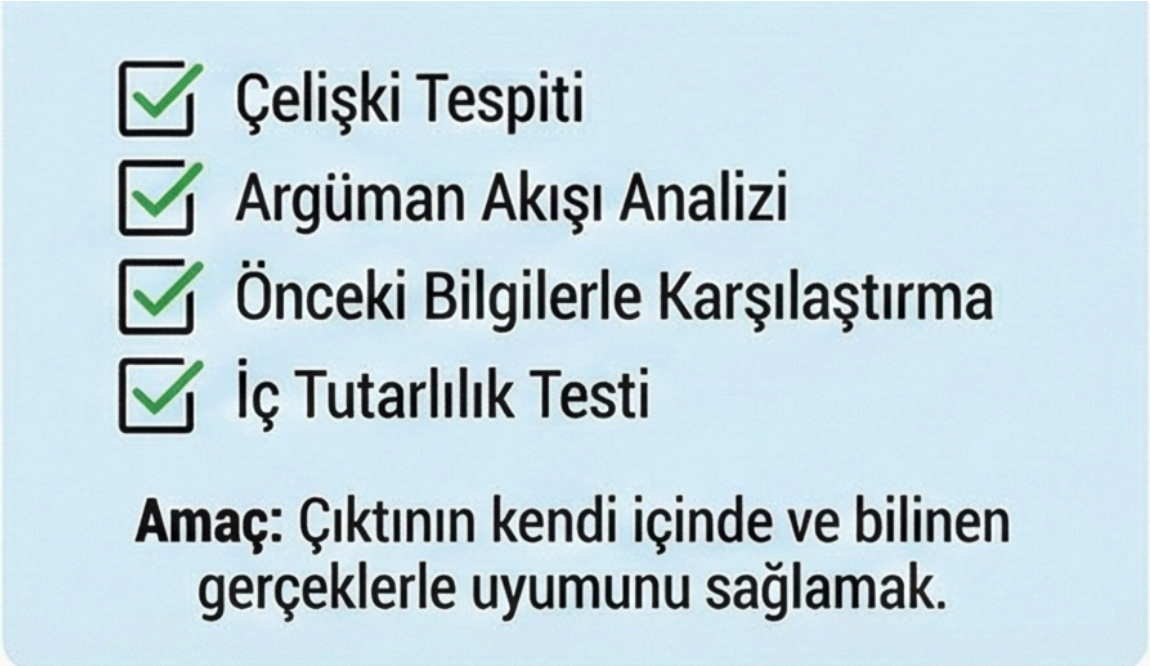
[ETKİLEŞİM] Tartışma: İkna Edici Yalanlar

Soru: YZ'nin çok kibar ve emin bir tonla "Dünya aslında üçgen şeklindedir" dediğini varsayalım. Tonlamanın güvenilirliğimiz üzerindeki etkisi nedir?

Yanlış bilgiye neden daha kolay inanıyoruz?

Mantık ve Tutarlılık Kontrolü: Öz-Denetim

- **Tekrar Sorma:** YZ'ye verdiği cevabı başka bir açıdan tekrar kontrol ettirmek.
- **Tersinden Sorgulama:** "Bu cevabın yanlış olma ihtimali nedir?" veya "Bu sonuca nasıl ulaştın?" diye sormak.



Adım Adım Akıl Yürütme (Chain of Thought)

- Kullanıcının YZ'den cevabı verirken aşamalarını açıklamasını istemesi.
- Her bir aşamanın doğruluğunun ayrı ayrı kontrol edilmesi.
- Hatalı adımın tespit edilerek düzeltilmesi.

Kendi İçinde Tutarlılık (Self-Consistency)

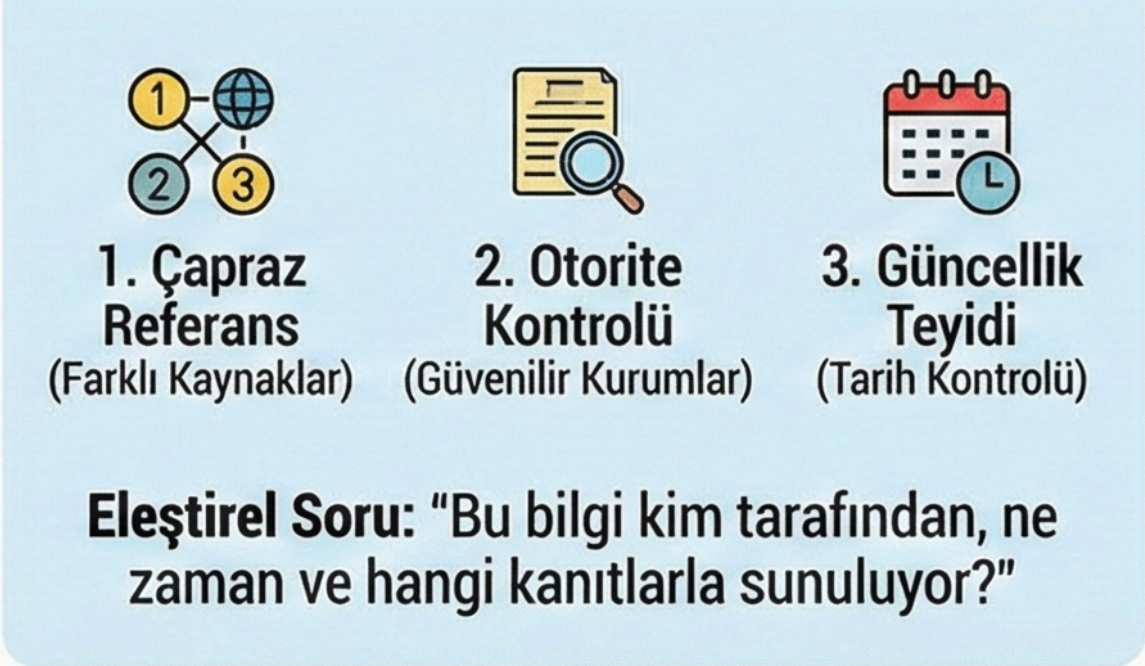
- YZ'nin aynı soruya farklı zamanlarda veya farklı pencerelerde verdiği cevapların karşılaştırılması.
- Çelişkili ifadelerin ayıklanması.

[ETKİLEŞİM] Uygulama: Mantık Zinciri

Görev: Karmaşık bir matematik problemini veya kimyasal bir denklemi YZ'ye çözdürün. Ardından "Çözümünü adım adım açıkla ve her adımın doğruluğunu kanıtla" diyerek süreci denetleyin.

Bazı Kaynak Doğrulama Yöntemleri

- **Birincil Kaynaklara Dönüş:** YZ'nin sunduğu bilgiyi orijinal el kitapları, makaleler veya ders kitaplarından teyit etmek.
- **Akademik Veritabanları:** Google Scholar, Scopus veya kütüphane kataloglarında ilgili atıfları aratmak.
- **DOI Kontrolü:** Verilen dijital nesne tanımlayıcılarının (DOI) geçerliliğini test etmek.



Güvenilir İnternet Kaynakları ile Çapraz Kontrol

- Bilginin sadece YZ'den değil, güvenilir kurumların (NASA, TÜBİTAK, WHO vb.) resmi sitelerinden kontrol edilmesi.
- Wiki sayfaları yerine hakemli yayınlara odaklanmak.

YZ Çıktılarının Eleştirel Değerlendirilmesi

- **Şüpheci Yaklaşım:** Her çıktıyı "doğru" değil, bir "iddia" olarak kabul etmek.
- **Dil ve Üslup Analizi:** Gereksiz süslü veya kaçamak cevapların arkasındaki bilgi eksikliğini fark etmek.
- **Önyargı (Bias) Tespiti:** Çıktının belirli bir görüşü kayırıp kayırmadığını incelemek.

[ETKİLEŞİM] Kaynak Dedektifi

Görev: YZ'den "İklim değişikliği ile ilgili 3 önemli makale önerisi" isteyin. Verilen yazarları ve başlıkları internette aratın. Kaçı gerçekten mevcut?

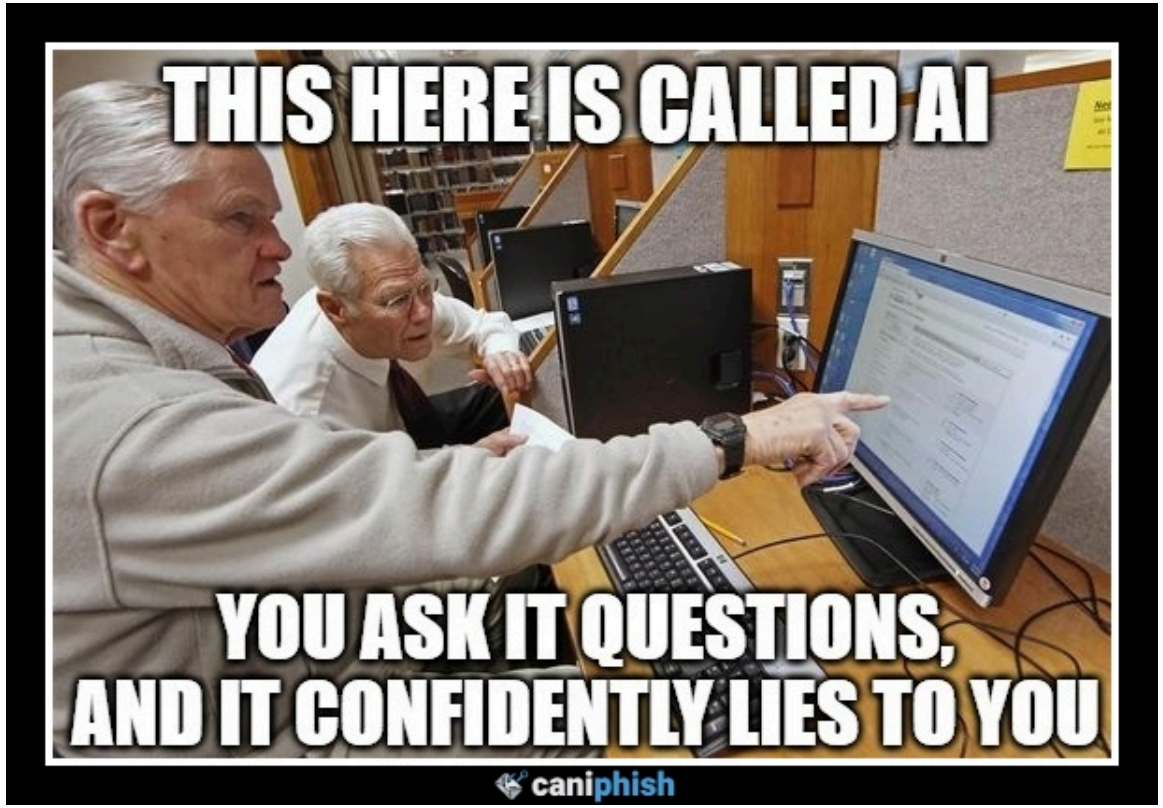
İnsan Denetiminin Önemi (Human-in-the-Loop)

- **Nihai Karar Verici:** YZ sadece bir öneri sunar; kararı veren ve sorumluluğu üstlenen her zaman insandır.
- **Bağlamsal Onay:** Verinin sahadaki gerçekliğe (deney sonuçları, mühendislik pratikleri) uygunluğunu denetlemek.



İnsan Sezgisi, Deneyimi ve Muhakemesi

- **Sezgi:** Yılların tecrübesiyle "bir şeylerin yanlış gittiğini" hissetme becerisi.
- **Deneyim:** Teorik verinin pratik uygulamada neden çalışmayacağını öngörme yetisi.
- **Muhakeme:** Veriler arasındaki ince farkları insani değerlerle tartabilme.

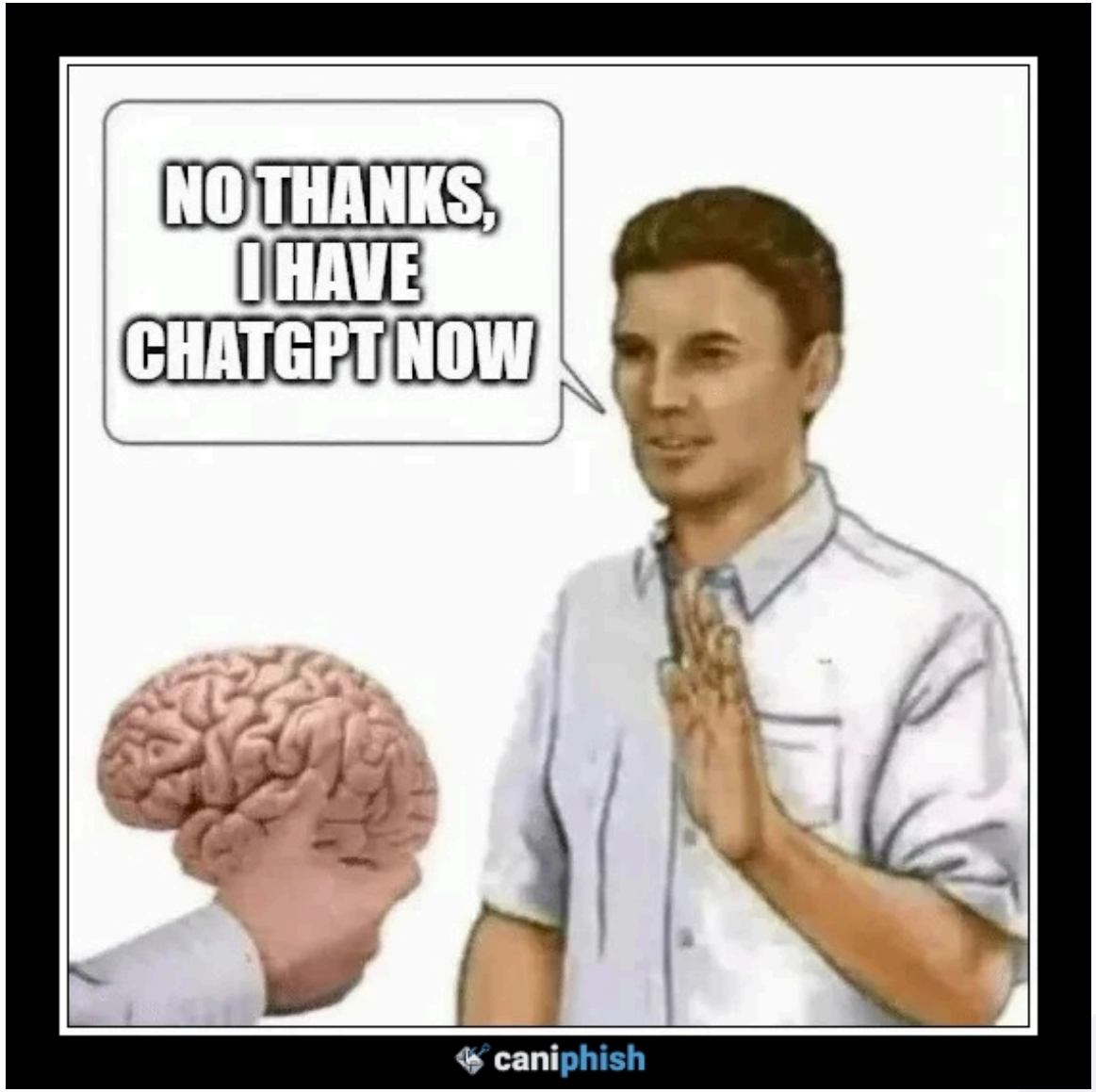


[ETKİLEŞİM] Vaka Analizi: Mühendislik Kararı

Görev: Bir YZ sistemi, bir köprü tasarımı için standartların dışında ama çok ekonomik bir malzeme önerdi. Tüm simülasyonlar "başarılı" diyor. Uzman olarak bu kararı onaylar mısınız yoksa kendi deneyiminize mi güvenirsiniz?

Yapay Zekaya Aşırı Bağımlılık Riski

- **Bilişsel Tembellik:** Sorgulama yeteneğinin körelmesi.
- **Yetenek Kaybı:** Temel analizleri ve hesaplamaları YZ olmadan yapamaz hale gelmek.
- **Hata Riskinde Artış:** Yanlış çıktıların denetlenmeden onaylanması sonucu oluşabilecek facialar.



Bilişsel Becerilerin Korunması Stratejileri

- **Zihinsel Egzersiz:** Önce problemi kendiniz çözmeye çalışın, sonra YZ desteği alın.
- **Sürekli Öğrenme:** Temel mühendislik ve bilimsel ilkeleri güncel tutmak.
- **Sorgulayıcı Kültür:** "Neden?" sorusunu sormaktan asla vazgeçmemek.

[ETKİLEŞİM] Öz-Değerlendirme

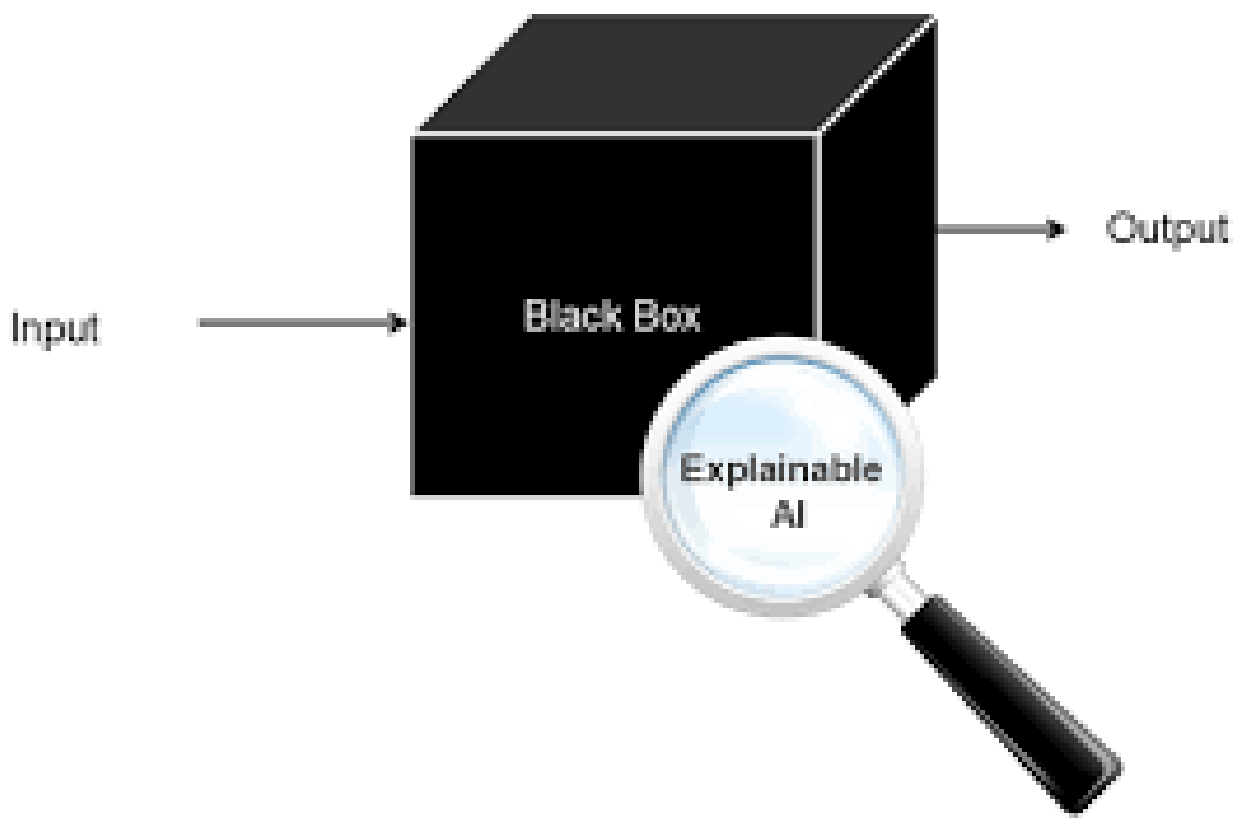
> Yapay zeka hayatınızda yokken kullandığınız hangi zihinsel yeteneğinizin (Örn: Ezber, hızlı hesaplama, planlama) zayıfladığını düşünüyorsunuz? Bunu korumak için ne yapabilirsiniz?

Eleştirel Düşünme Becerileri

- Bilginin doğruluğunu, geçerliliğini ve güvenilirliğini tarafsız bir şekilde analiz etme süreci.
- YZ çağında eleştirel düşünme, bilgiyi "bulmaktan" ziyade "doğrulamak" üzerine yoğunlaşır.

YZ ve "Kara Kutu" (Black Box) Problemi

- YZ'nin bir cevaba nasıl ulaştığını tam olarak açıklayamaması.
- Şeffaf olmayan bir kararın denetlenmesinin zorluğu.
- Mühendisin "açıklanabilir" sonuçlara olan ihtiyacı.



Güçlü Bir Doğrulama Protokolü Oluşturma

1. Çıktıyı oku ve ana iddiaları belirle.
2. Mantık hatalarını tara.
3. Teknik verileri el kitaplarından kontrol et.
4. Kaynakların gerçekliğini sorgula.
5. Kendi tecrübenle kıyasla.

[ETKİLEŞİM] Grup Çalışması: Doğrulama Rehberi

Görev: Laboratuvar raporu yazarken YZ kullanan bir öğrenci için 5 maddelik bir "Doğrulama Listesi" (Checklist) hazırlayın.

Yanlış Bilgiyle Mücadele: Toplumsal Sorumluluk

- YZ tarafından üretilen yanlış bilgilerin yayılmasını engellemek.
- Bilgi kirliliğine karşı mesleki etik çerçevesinde hareket etmek.
- Doğru bilginin savunucusu olmak.

YZ Çıktılarında Taraflılık (Bias)

- Modellerin eğitildiği veri setlerindeki ön yargıların çıktıya yansıması.
- Teknik konularda bile bazı yöntemlerin veya ülkelerin daha çok öne çıkarılması.
- Objektifliğin korunması için farklı bakış açılarının aranması.

[ETKİLEŞİM] Tartışma: Önyargı Tespiti

> **Soru:** Bir YZ sistemine "Başarılı bir mühendis profili çiz" dediğinizde verdiği cevap (cinsiyet, etnik köken, yaş) sizce tarafsız mıdır? Bu, mesleki kararları nasıl etkiler?

Teknik Verilerin Doğrulanması (Mühendislik Örneği)

- **Faz Diyagramları:** YZ'nin çizdiği veya tarif ettiği diyagramların standartlarla karşılaştırılması.
- **Reaksiyon Verimleri:** Teorik verimin YZ tarafından aşırı iyimser hesaplanma riski.
- **Güvenlik Veri Formları (MSDS):** Kimyasal güvenlik bilgilerinin asla sadece YZ'ye bırakılmaması gerekliliği.

Doğrulama İçin Yardımcı Teknolojiler

- İntihal tespit yazılımları.
- Faktör kontrolü (Fact-checking) yapan eklentiler.
- Özel bilimsel hesaplama araçları (WolframAlpha vb.) ile çapraz sorgulama.

[ETKİLEŞİM] Uygulama: İki Model Arasındaki Fark

> **Görev:** Aynı teknik soruyu iki farklı YZ modeline (Örn: ChatGPT ve Claude) sorun. Cevaplar arasındaki farkları ve hangisinin daha "tutarlı" olduğunu analiz edin.

YZ'nin "Eminlik" Aldatmacası

- YZ, cevabı bilmiyorsa bile "bilmiyorum" demek yerine uydurma eğilimindedir.
- Üslup profesyonelliği, doğruluğun garantisi değildir.

İnsan ve Makine: Symbiotic (Ortak Yaşam) İlişki

- Makine hız kazandırır, insan yön verir.
- Makine veri sunar, insan anlam katar.
- Makine hesap yapar, insan etik sınırı çizer.

[ETKİLEŞİM] Sokratik Sorgulama

> "Eğer yapay zeka bir gün yanılmaz hale gelirse, insana hala ihtiyaç duyulur mu?" - Bu soru üzerine eleştirel bir tartışma başlatın.

3. Haftanın Özeti: Temel İlkeler

- YZ'nin her çıktısı bir "taslak" niteliğindedir.
- Halüsinasyon, sistemin doğal bir parçasıdır ve her an karşınıza çıkabilir.
- Doğrulama yapılmadan kullanılan her bilgi, mesleki bir risktir.

[ETKİLEŞİM] Haftalık Öz-Değerlendirme

> Bugünkü dersten sonra, YZ çıktılarına olan güven seviyenizde (1-10 arası) bir değişim oldu mu? Neden?

Gelecek Haftaya Bakış

- 4. Hafta: Akademik Dürüstlük ve YZ Kullanımı.
- Ödevlerde ve raporlarda dürüstlük sınırları nelerdir?
- YZ kullanımı nasıl beyan edilir?

Haftalık Görev

- Bu hafta kullandığınız bir YZ aracının verdiği bir cevabı eleştirel bir gözle okuyun. Varsa en az bir hatayı veya eksikliği raporlayın.

Kapanış Notu

"Sorgulanmayan bilgi, sahibini yanlış yola sürükleyen bir pusula gibidir. Yapay zeka çağında en büyük pusulanız eleştirel aklınızdır."

Teşekkürler & Sorular

YZM4015 - 3. Hafta Sonu

> **Analoji:** Yapay zekadan gelen bilgileri değerlendirmeyi, **kaynağı belirsiz bir haberi** kontrol etmeye benzetebiliriz; haber ne kadar çarpıcı ve ikna edici olursa olsun, güvenilirliğini doğrulamadan başkalarına yaymak veya bu habere göre önemli bir karar almak büyük riskler taşır. Eleştirel düşünme, bu süreçteki **teyit mekanizmanızdır**.