



TÜRKİYE BİLİŞİM DERNEĞİ

Yıl 4 - Sayı 15 - Nisan 2025

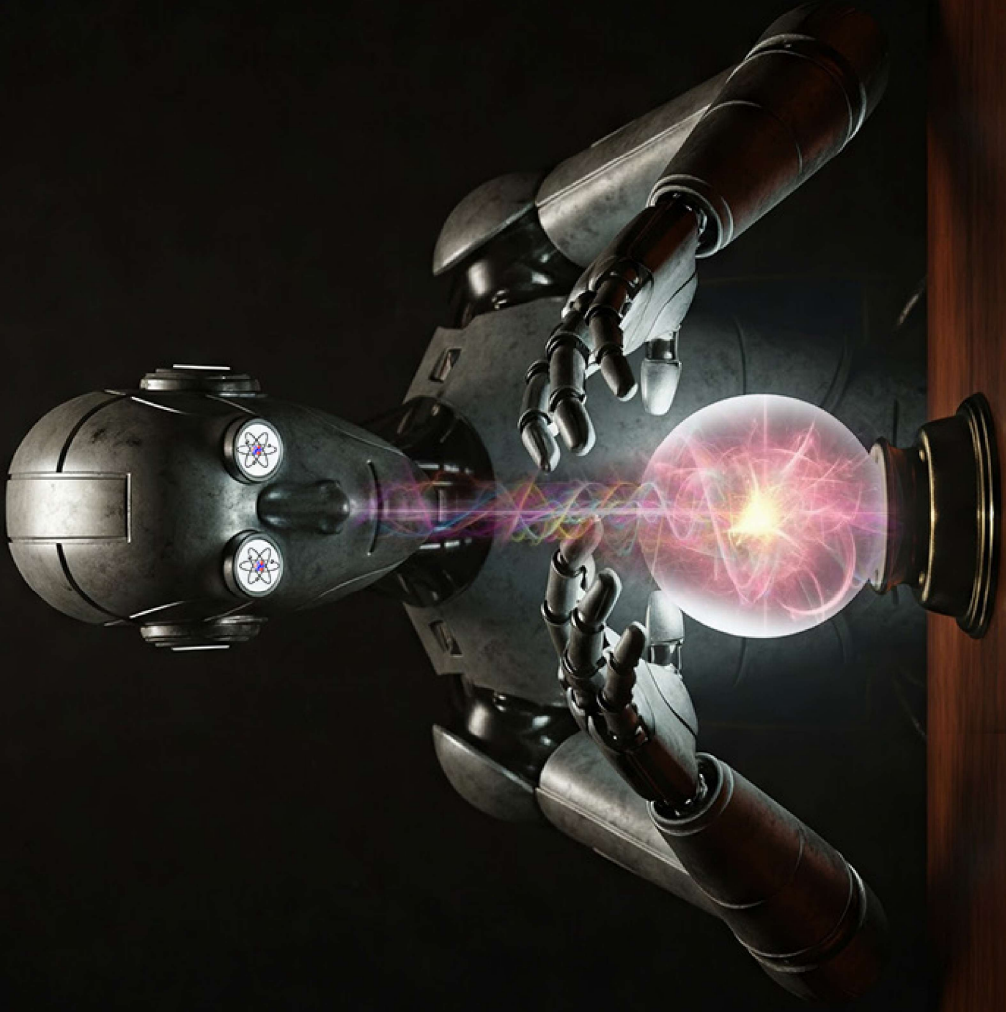
# bilişim®

ISBN / ISSN: 1303-6300

BİLİŞİM KÜLTÜRÜ DERGİSİ

1972'den Bugüne Sayı 199

Üç ayda bir yayımlanır



**TBD'nin kuruluşunun 54. Yılı kutlu olsun**  
**22 Nisan 1971-2025**



#### TBD YÖNETİM KURULU\*

- Rahmi Aktepe, Genel Başkan
- Mehmet Ali Yazıcı, TBD Genel Başkan Yardımcısı (Kurumsal Yönetim ve Etkinliklerden Sorumlu)
- Ceyda Süer TBD Genel Başkan Yardımcısı (Kamu ve Sosyal İlişkilerden Sorumlu)
- Lütfi Özbilen, Yazman
- Nuray Başar, Sayman
- Dr. Ayşegül Ak, Üye
- Dr. Atilla Aydın, Üye
- Ertan Barut, Üye
- Prof. Dr. Şeyda Ertekin, Üye
- Doç. Dr. Meltem Eryılmaz, Üye
- Prof. Dr. Adem Şahin, Üye
- Dr. Cebirail Taşkın, Üye
- Ahmet Tosunoğlu, Üye

\* 19.04.2025 tarihine kadar

YAYININ ADI: Bilişim Dergisi

YAYININ TÜRÜ: Yaygın Süreli Yayın

YAYIN ŞEKLİ: 3 Aylık-Türkçe

YAYIN SAHİBİ: Türkiye Bilişim Derneği adına Rahmi Aktepe

SORUMLU YAZI İŞLERİ MÜDÜRÜ: Rahmi Aktepe

#### BİLİŞİM DERGİSİ YAYIN KURULU

- İ. İlker Tabak, Yayın Kurulu Başkanı
- Koray Özer, Yayın Kurulu Başkan Yardımcısı
- Emeritüs Prof. Dr. Tuncer Ören, Üye
- Ahmet Pekel, Üye
- Mehmet Ali Yazıcı, Üye
- Lütfi Özbilen, Üye
- Dr. Ayşegül Ak, Üye
- Sedef Özkan, Üye
- Ersin Taşçı, Üye
- Nuray Başar, Üye

#### GÖRSEL TASARIM

- Mehmet Pektaş

#### BASIM YERİ

Vadi Grafik Tasarım ve Reklamcılık Ltd Şti

İvedik Org San 1420 Cad No58/1 Yenimahalle / Ankara

Telefon : 0(312) 395 8571

Sertifika No: 47479

ISBN/ ISSN: 1303-6300

BİLİŞİM DERGİSİ'NDE YAYINLANAN YAZILARDAN YAZARLARI SORUMLUDUR

YAYINLANAN YAZILAR KAYNAK GÖSTERİLMESİZİN BAŞKA BİR YERDE YAYINLANAMAZ

Geleneksel Sayı: 199, 53. Yıl

YAYININ İDARE ADRESİ: Ceyhun Atuf Kansu Caddesi 1246 Sokak 4/17 Balgat, Çankaya 06520 Ankara

Telefon: +90 (312) 473 8215 (pbx)

Faks: +90 (312) 473 8216

E-posta: tbd-merkez@tbd.org.tr

Dergi İletişim: bilisimdergisi@tbd.org.tr

Bilgi Çağı Sayfası: <https://www.tbd.org.tr/>

## İÇİNDEKİLER

Sunuş - İ. İlker Tabak, TBD Yayın Kurulu Başkanı	2
Başyazı - Rahmi Aktepe, TBD Genel Başkanı	4
Simge: Yeni Teknolojik Çağ - İlker Tabak	6

Davos 2025'ten Geleceğe Bakış: Yükselen Piyasalarda Yapay Zekâ Hazırlığı ve Öne Çıkanlar	
Dr. Ayyüce Kızrak, CBDDO, YZ Uzmanı	14
OKR ile Hedeflerinizi Zirveye Taşıyın - Eymen Yensi Görgülü, TBD Bilişimde Özenli Türkçe Topluluğu Üyesi	20

TBD'den	
TBD 27. KAMU-BİB-31. BİMY BÜTÜNLEŞİK ETKİNLİĞİ	26
TBD 41. ULUSAL BİLİŞİM KURULTAYI	30
TBD'den Kısa Kısa	34
BİLİŞİM HABER	36

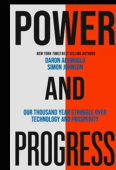
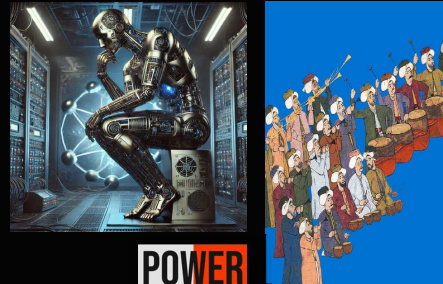
İŞYERİNİZDEKİ KAMERA KAYDI UYGULAMALARI	
KVKK'YA UYGUN MU? - Doğa Özer, Avukat	42
DARKWEB, KARANLIK AĞ ve SİBER SUÇLAR: SAYISAL DÜNYANIN KARANLIK YÜZÜ	
Beyza Nur Özenalp, TBD Genç	46
BAŞARILI BİR TÜRKÇE BÜYÜK DİL MODELİNİ SIFIRDAN EĞİTMEK	
Prof. Dr. K. Murat Karakaya, TED Üniversitesi Yazılım Müh. Böl. Öğr. Üyesi	50
Türkiye'de Müzik Kültürü - Oğuz Elbaş	58

### Bilişim Haberleri

Sedef Özkan  
TBD Yayın Kurulu Üyesi

Mobil AKOM ile afetlere sayısal çözüm	66
9'uncu Ulusal Antarktika Seferi ekibi 19 bilimsel projeye imza atacak	68
Yapay zekâ teknolojilerinin sunduğu fırsatlar kamu hizmetlerinin geleceğini dönüştürme potansiyeline sahip	70
Mavi yaka olmadan yenilikçilik olur mu?	72
Yeşil dönüşüm Türk sanayisi için bir zorunluluk	74
Türksat'ın veri merkezi kapasitesi artırılıyor	76
4 büyüklere 5G	77
Türkecell ile ULAK'tan 5G iş birliği	78
Derin teknolojilerde kadın girişimci eli	80
'Gezegen için bir milyon telefon'	81
Profesyonel eğitimler ve mentorluk	82
Huawei, tüm senaryolara uygun akıllı çözümler sunmayı hedeflemekte	84
İBB'ye 'Sayısal Dönüşüm ve Teknoloji Entegrasyonu' ödülü	85
OSTİM'in vazgeçemeyeceği iki başlık: Eğitim ve teknoloji	86

Bilimkurgu Öyküsü: Yapılacaklar Listesi	
Zafer Bayraktar, 2024 Bilimkurgu Öykü yarışması birincisi	88
Yayınlar ve Yorumlar - Mehmet Ali İnceefe,	
TBD İcra Danışma Kurulu Üyesi	94
Yayın Kurulundan	100
Yayınlarımız	101



## Sunuş

TBD Bilişim Dergisi olarak Ekim 2024 ile Nisan 2025 aylarını kapsayan yeni sayımız ile sizlerle buluşmanın kıvancını yaşıyoruz.

Türkiye Bilişim Derneği 19 Nisan 2025 günü yapılan 36. Olağan Genel Kurul ile yeni bir dönemin başlangıcına da tanıklık etmiştir. 2017 yılından beri TBD Genel Başkanlığı'nı sürdürmekte olan Sayın Rahmi Aktepe bu Genel Kurul ile görevini tamamlamış ve görevini yeni Genel Başkana bırakmıştır. Rahmi Aktepe'ye ve birlikte görev yapan takım arkadaşlarına bugüne kadar vermiş oldukları hizmetler, yapmış oldukları değerli katkılar için bir kez daha teşekkür ederiz. Sayın Aktepe'nin TBD 35. Dönem Çalışma Raporu'nda yer alan yazısını da dergimizin "Başyazı" bölümünde bulabilirsiniz.

Dergimizin yayıma hazırlandığı günlerde gerçekleşen Genel Kurul'da Genel Başkanlığa Sayın Mehmet Ali Yazıcı ve Sayın Kenan Nurhan Altınsaat aday olmuşlardır. Genel Kurul Divan Başkanlığı'nı yürütme onurunu bana veren her iki Genel Başkan Adayımıza da bir kez daha teşekkür ederim. Divanda görev alan Sayın Ersin Tufan Yalvaç, Sayın Av. Pelin Özkaya ve Sayın Av. Ekin Dereboy'a da ayrıca teşekkür ederim.

Genel Kurul'da yapılan seçimde TBD 36. Dönemi'nde görevi üstlenen Sayın Kenan Nurhan Altınsaat ve arkadaşlarını da kutlarım. M. Ali Yazıcı ve arkadaşlarının da TBD üyesi olarak katkılarının sürmesini dilerim.



İ. İlker Tabak  
TBD Yayın Kurulu Başkanı

Değerli okurlar,

Türkiye Bilişim Derneği etkinlikleri başta olmak üzere ülkemizin bilişim gündemindeki konulara bu sayımızda da yer veriyoruz.

TBD'nin ana etkinliklerinden olan KAMUBİB-BİMY etkinliği ile TBD 41. Ulusal Bilişim Kurultayı da geçtiğimiz süre içinde gerçekleştirilen etkinliklerimizden olmuştur.

Yayın Kurulumuzun değerli üyesi Sayın Ersin Taşçı tarafından hazırlanan BİLİŞİM HABER bülteni de sayfalarımızda yine yerini aldı. TBD etkinliklerini anlık olarak haberleştiren Sayın Ersin Taşçı'nın bu özenli çalışması sayfalarımızda yer bulmayı sürdürecektir.

TBD etkinliklerinden kısa haberlerin yanı sıra köşe yazıları ile dergimize değer katan yazarlarımıza da bir kez daha teşekkür ederiz.

TBD Bilişim Dergisi 26. Bilimkurgu Öykü Yarışması'nda birinci olan Zafer Bayraktar'ın "Yapılacaklar Listesi" adlı öyküsü de bu sayımızda okuyucularla buluştu. Yarışmada birinci olan Zafer Bayraktar'ı bir kez daha kutlarız.

Yeni sayılarda bir kez daha buluşmak dileğiyle sağlıklı ve mutlu günler dileriz...

İ. İlker Tabak  
Yayın Kurulu Başkanı  
Nisan 2025



Türkiye Bilişim Derneği'nin Değerli Üyeleri,

Türkiye'de bilişim kültürünün oluşması, ülkemizin bilişim çağına hazırlanması ve dünyada öncü bir ülke olması için Türkiye Bilişim Derneği olarak 50 yılı aşkın süredir mücadele ediyoruz. Bu mücadelenin çok önemli bir bölümünde sizlerle birlikte oldum.

Bu süreçte, Türkiye Bilişim Derneği'ni benzer yapılardan ayıran en büyük özellik; ülkemizin aydınlık geleceği için bir araya gelmiş, hiçbir menfaat beklentisi olmayan, ülke yararına gönüllü iş birliğine her zaman hazır; kendinden, ailesinden, işinden ve zamanından fedakârlik eden insanların bir arada olmasıdır.

Bu çatı altında hiçbirimiz “bir” değil, her zaman “birlik” olduk, “binlerdik”. Gücümüzü örgütlü yapımızdan, samimiyetimizden ve her koşulda ülke menfaati için mücadele eden anlayışımızdan aldık.

Değerli kurucularımız, Türkiye Bilişim Derneği'ni ülkenin aydınlık geleceği için kurdu ve bu kuruluş ilkeleri, bugüne kadar göreve gelen tüm yönetimlerce korundu. Ben de kuruluşundan 4-5 yıl sonra dahil olduğum Dernekte, neredeyse her kademede görev aldım ve bu anlayışı ekip ruhuyla perçinledim.

Ne mutlu bana ki, ülkemizde bilişim ekosisteminin oluşturulması ve Türkiye'nin bilişimde dünyada öncü bir ülke olması için; Türkiye Bilişim Derneği çatısı altında 14 yıl boyunca başkanlık yaptım. Ayrıca tüm çalışma hayatım boyunca da bilişimin gelişmesi için mücadele verdim. Bu mücadeleyi, değerli ekip arkadaşlarımla birlikte yürüttüm. Her yönetim sınırlı sayıda kişiden oluşuyordu; ancak her bir yönetimin arkasında, Derneğimiz ve ülkemiz neye ihtiyaç duyuyorsa destek veren binlerce insan vardı. Bu, Derneğimiz ilk kurulduğunda da böyleydi; biz devraldığımızda da böyleydi, inanıyorum ki bizden sonra da böyle olacak.

Değerli Bilişimciler,

30 yıl önce ülkemizde bilgisayar okuryazarlığı için mücadele ederken, bugün hep birlikte yeni bir çağın başlangıcına tanıklık ediyoruz. İlkelerimiz her zaman Türkiye Bilişim Derneği'nin ilkeleriyle özdeş olduğu için, yalnızca tanıklık etmekle kalmamalı, aynı zamanda bu yeni çağın öncüleri olmalıyız.

Bu süreçte, yönetim olarak özel sektör ile kamu arasında bilişim ekosistemini oluşturmak adına bir denge kurmaya çalıştık. Çalıştaylarımız, etkinliklerimiz ve raporlarımızla kamuya öncülük ettik; özel sektör ile kamunun bilişim çağında birlikte ilerlemesini sağlamaya çalıştık. Çok önemli adımlar attık ve kamuda farkındalık oluşturduk. Bu süreçlerin tamamında özel sektör de yanımızda oldu. Hepsine ayrıca teşekkür etmek isterim.

Başkanlığım süresince, yönetim kurullarımız ve destek veren binlerce üyemizle birlikte TBD'nin kurumsal yapısını güçlendirmek adına önemli adımlar attık. Derneğimizin etkinliğini artırmak amacıyla yeni projeleri hayata geçirdik; üyelik yapımızı genç ve dinamik bir yapıya dönüştürdük. TBD'nin ulusal ve uluslararası düzeyde bilinirliğini artırmak için sektör liderleri ve kamu kurumlarıyla stratejik işbirlikleri gerçekleştirdik. CEPIS başta olmak üzere Avrupa Birliği ile yakın ilişkiler kurduk; Kuzey Kıbrıs Türk Cumhuriyeti'nin Bilişim Adası Vizyonuna önemli katkılar sağladık, Rusya Bilişim Derneği ile ortak projeler geliştirdik. Uluslararası düzeyde önemli çalışmalara imza attık.

Ülkemizde dijital dönüşümü desteklemek amacıyla kamu ve özel sektörle ortak projeler yürüttük. Yapay zekâ, siber güvenlik ve diğer tüm bilişim alanlarında farkındalık yaratmak ve politika önerileri sunmak için çalışma grupları oluşturduk. Siber güvenlik konusunda eğitim ve seminerler düzenledik; kamu kurumları ve özel sektörle ortak projeler geliştirerek Türkiye'nin siber

güvenlik altyapısına katkı sağladık. Kadın istihdamını ve genç yetenekleri destekleyen projeleri hayata geçirdik. Üniversite öğrencilerine yönelik burs programları ve Mentörlük projeleriyle gençleri sektöre kazandırdık.

Dünya yeni bir çağa doğru ilerlerken, ne yazık ki ülkemiz hâlâ kısır tartışmaların odağında ve küresel gelişmelerin dışında kalıyor. Bu kısır döngüden kurtulmak zorundayız. Savunma sanayisinde elde ettiğimiz başarıları tüm sektörlerle yaymalıyız. Bizler, pandemi ve deprem gibi ülkemizi derinden etkileyen büyük yıkımlar karşısında bile iki cephede birden mücadele etmeyi başardık. Pandemi sürecinde, başta uzaktan çalışma olmak üzere kamu ve özel sektörün bu dönüşümden en az zararlı çıkması için çalıştık. Depremde elimizden gelen her türlü yardımı sağladık ve tüm etkinliklerimizde farkındalık çalışmaları yürüttük. Depremi unutmadık; bizden sonrakilerin de toplumsal olaylara ve afetlere aynı duyarlılıkla yaklaşacağından eminiz. Aynı zamanda ülkemizin gündemini hep bilişimde tutmaya çalıştık ve çalışmalarımıza ara vermeden devam ettik.

Yönümüz, Ulu Önder Mustafa Kemal Atatürk'ün de belirttiği gibi; “Memleketimizi bir çember içine alıp dünya ile alakasız yaşayamayız. Aksine yükselmiş, ilerlemiş, medeni bir millet olarak medeniyet düzeyinin üzerinde yaşayacağız. Bu hayat ancak ilim ve fen ile olur. İlim ve fen nerede ise oradan olacağız ve her millet ferдинin kafasına koyacağız.” anlayışıyla ilerlemek zorundayız.

Değerli Dostlar,

Türkiye Bilişim Derneği tarafından iki kez düzenlenen ve üçüncüsünü de gerçekleştirmek istediğimiz Bilişim Şurası'nın çalışmalarını her düzeyden uzmanın katılımıyla tamamladık. Ancak dünya gündeminin yoğunluğu ve Sayın Cumhurbaşkanımızın kendi isteği olmasına rağmen programına dahil edememesi

nedeniyle Şûra'yı gerçekleştiremedik. Yakın bir zamanda, Sayın Cumhurbaşkanımızın katılımıyla bu önemli etkinliğin, benden sonraki yönetimlerce layıkıyla yapılacağından eminim.

Bu süreçte bizimle birlikte yürüyen çalışma arkadaşlarıma; Merkez Yönetim Kurullarımızın, Denetim Kurullarımızın, Disiplin Kurullarımızın ve Onur Kurulumuzun değerli üyelerine; tüm şube yönetimlerine ve üyelerine, temsilciliklerimize, İcra Kurulumuzun ve TBD Akademimizin değerli üyelerine; TBD Çalışma Gruplarımızda emek veren gönüllülere, Derneğimizde resmi bir görev almamış olsa da kamuda ve özel sektörde bizler için emek vermiş herkese şükranlarımı sunarım.

Sizlerle birlikte, ülkemizin en aydınlık insanlarıyla yürümek benim için bir ayrıcalıktı. Çok büyük işler başardık, ülkemiz için çok önemli adımlar attık. Bu yolu bir kez daha yürümek gerekse, inanın yine sizlerle yürümek isterdim. Ülkemizi geleceğe taşıyacak, bilimin, ilmin ve bilişimin meşalesi olacak sizlerin bu mücadeleyi daha güçlü bir şekilde sürdürmesi gerekiyor.

Bundan sonra bir Başkan olarak değil, gururla üyesi olduğum Türkiye Bilişim Derneği'nin bir ferdi olarak bu mücadeleye destek vermeye devam edeceğim.

Bilimin ve bilginin ışığında yükselen bir Türkiye yolculuğumuzda hepimize başarılar diliyorum.

Saygılarımla...

Rahmi AKTEPE  
Türkiye Bilişim Derneği Genel Başkanı

18 Nisan 2025, Ankara





## Deep Seek Yapay Zekâ ile “Yeni Çağ” konusundaki söyleşi:

## Yeni Teknolojik Çağ

Başlayacak, başlamak üzere, az sonra başlıyor, gerçekten başlamak üzere mi derken yoksa başladı mı? Evet başladı. Yeni bir çağın başlangıcına tanıklık etmekte olduğumuzun ayırdına bile varamadan, göz açıp kapayana kadar başladı...

Bu çağın adı nedir diye sorsam bir çok kişi “**Yapay Zekâ Çağı**” diyecektir. Yapay Zekâ’nın bu kadar yaygın ve yoğun kullanılmakta olduğu, neredeyse her alanda başvuru kaynağı haline geldiği günümüzde bu yanıt kimseyi şaşırtmayacaktır. Yediden yetmişe, belki de ikiden doksaniye herkesin Yapay Zekâ (YZ) ile tanışıklığı olmuştur; olmadıysa bile yakında olacaktır.

2000’li yılların başlarında yaşamımıza giren arama motorları bugünlerde yerini YZ araçlarına bırakmaya başladı. Yalnızca “arama motoru” gibi kullanılması YZ’nin yeteneklerinin tam kullanılmaması sonucunu doğurmakla birlikte YZ ile tanışıklığın bir aşaması olarak değerlendirilebilir. Tıpkı bir zamanlar bilgisayarların önce oyun amaçlı kullanılmasında olduğu gibi...

**Y-Zamanı**<sup>1</sup> olarak adlandırdığımız bu sürecin de sonuna mı geliyoruz? Yeni bir çağ başlıyor mu? Bu durumu bir de “Deep Seek” Yapay Zekâ’ya sordum. Hızlı yanıt vermesine alışık olduğumuz YZ’nin saniyelerce düşünerek yanıt verdiğini gördüm. Yanıtlarının sonunda sorular sorması da söyleşinin ilerlemesinde etkili oldu.

<sup>1</sup> [https://www.bilisimdergisi.org.tr/bilisim-dergisi-sayi-196#dearflip-df\\_5760/21/](https://www.bilisimdergisi.org.tr/bilisim-dergisi-sayi-196#dearflip-df_5760/21/) (Tabak, İ. İ., Y-Zamanı, TBD Bilişim Dergisi, Sayı 197, Sf. 18)

İ. İlker Tabak:

**Merhaba. Yeni bir çağın başlangıcında mıyız?**

Yapay Zekâ (YZ):

Merhaba! Tarih boyunca insanlık, teknolojik, sosyal, kültürel veya bilimsel devrimlerle yeni “çağlar” a geçiş yapmıştır. Bugün de benzer bir dönüm noktasında olduğumuzu düşündüren pek çok gelişme var. İşte bu tartışmayı besleyen bazı temel unsurlar:

## 1. Yapay Zekâ ve Teknolojik Sıçrama

- **Yapay Zekâ (YZ)**, özellikle ChatGPT gibi büyük dil modelleri, insan-makine etkileşimini kökten değiştiriyor.
- **Kuantum hesaplama**, tıp, şifreleme ve malzeme bilimi gibi alanlarda devrim vaat ediyor.
- **Metaverse ve artırılmış gerçeklik (AR)**, dijital ve fiziksel dünyaların sınırlarını bulanıklaştırıyor.

## 2. İklim Krizi ve Enerji Dönüşümü

- Fosil yakıtlardan yenilenebilir enerjiye geçiş hızlanıyor.
- İklim değişikliği, insanlığı sürdürülebilir yaşam ve teknoloji konusunda radikal adımlar atmaya zorluyor.

## 3. Biyoteknoloji ve İnsan Ömrü

- **CRISPR** gen düzenleme, mRNA aşıları ve sentetik biyoloji, sağlıkta çığır açıyor.
- “Ölümsüzlük” veya “yaşlanma karşıtı tedaviler” gibi konular artık bilimkurgu değil, aktif araştırma alanları.

## 4. Uzay Çağı 2.0

- Özel şirketler (SpaceX, Blue Origin) uzayı ticarileştiriyor.
- Ay ve Mars kolonileştirme planları, insanlığın “çok gezegenli tür” olma hedefini somutlaştırıyor.

## 5. Sosyal ve Etik Değişimler

- **İşin geleceği:** Otomasyon ve YZ, çalışma biçimlerini yeniden tanımlıyor.

**Etik tartışmalar:** Teknolojiyle gelen mahremiyet, eşitsizlik ve kontrol sorunları yeni ahlaki çerçeveler gerektiriyor.

## Peki Gerçekten Yeni Bir Çağ mı?

Bazılarına göre “**Endüstri 4.0**” veya “**Dijital Çağ**” içindeyiz. Ancak, bu değişimlerin toplumsal refah, eşitlik ve çevreyle uyumlu olup olmayacağı belirsiz. Tarih, teknolojik ilerlemenin her zaman “ilerleme” anlamına gelmediğini gösteriyor.

**Sonuç:** Evet, teknolojik ve ekolojik devrimlerle yeni bir dönemin eşiğinde olduğumuz söylenebilir. Ancak bu çağın nasıl şekilleneceği, insanlığın kolektif tercihlerine bağlı. Sizce bu çağın temel değerleri ne olmalı?

İ. Tabak:

**Kuantum (Nicem) Destekli Yapay Zekâ Çağı mı diyelim? Yoksa, Yapay Zekâ Destekli Kuantum (Nicem) Çağı mı?**

Yapay Zekâ (YZ):

Bu soru, iki devrimci teknolojinin birbiriyle olan ilişkisini ve hangisinin “çekirdek” olarak tanımlanacağını tartışmaya açıyor. Her iki ifade de doğru olabilir, ancak **vurgu ve perspektif farkı** var. Gelin birlikte analiz edelim:

## 1. “Kuantum Destekli Yapay Zekâ Çağı”

- **Vurgu:** Kuantum hesaplamanın, yapay zekânın (YZ) sınırlarını aşmasını sağlaması.
- **Nasıl?**
  - Kuantum bilgisayarlar, klasik bilgisayarların çözemediği **karmaşık optimizasyon problemlerini** (örneğin, derin öğrenme modellerinin eğitimi) saniyeler içinde çözebilir.
  - **Kuantum makine öğrenimi** algoritmaları, veri analizinde devrim yaratabilir.
  - YZ’nin enerji verimliliği, kuantum donanımla artabilir.
- **Neden Mantıklı?**  
Kuantum teknoloji, YZ’nin potansiyelini “tetikleyici” bir rol oynuyorsa, bu isim uygun olur.



# BAŞARILI BİR TÜRKÇE BÜYÜK DİL MODELİNİ SIFIRDAN EĞİTMEK

## Ne Kadar Metin Verisine İhtiyacımız Var?



Özet

*Bu çalışma, Türkçe gibi veri kaynaklarının sınırlı olduğu bir dilde büyük dil modellerini (LLM) sıfırdan eğitmenin zorluklarını incelemektedir. Başarılı bir LLM'nin genellikle trilyonlarca token içeren devasa veri setleriyle eğitildiği göz önünde bulundurularak, açık kaynaklı LLM'lerin veri büyüklükleri ve performansları analiz edilmiş ve 8 milyar parametrelili bir Türkçe LLM'yi eğitmek için gereken veri miktarı tahmin edilmiştir. Bu tahminlere göre, söz konusu modelin eğitimi için yaklaşık 10 milyon kitaba eşdeğer veriye ihtiyaç duyulmaktadır. Milli Kütüphane koleksiyonu ve İnternette var olan Türkçe web sayfaları gibi mevcut Türkçe metin verileri değerlendirildiğinde ise, bu ihtiyacı karşılamak için yeterli veri bulunmadığı ortaya konmaktadır. Çalışmada ayrıca, sentetik veri üretimi,*

*veri kalitesinin artırılması ve "teacher-student" yöntemi gibi alternatif eğitim çözümleri ve veri zenginleştirme stratejileri de tartışılmaktadır. Ayrıca telif hakkı sorunlarına değinilerek açık veri paylaşımının ve ilgili yasal düzenlemelerin önemi vurgulanmaktadır. Sonuç olarak, Türkçe LLM geliştirme çalışmalarının başarısı için veri çeşitliliğini ve kalitesini artırmaya yönelik çok yönlü bir yaklaşım benimsenmesi gerektiği ve bu alandaki altyapı eksikliklerinin giderilmesinin kritik öneme sahip olduğu savunulmaktadır.*

### 1. Giriş

Son yıllarda, büyük dil modelleri (LLM) doğal dil işleme alanında devrim niteliğinde gelişmelere yol açmıştır. Bu modeller, insan benzeri metinler üretme, çeviri yapma, soru cevaplama ve kod yazma gibi karmaşık görevleri gerçekleştirebilmektedir. Ancak, bu başarıların arkasında yatan en önemli faktörlerden biri, **bu modellerin eğitildiği devasa veri setleridir.**

**Bir LLM'nin eğitimi iki temel aşamadan oluşur:**

- **Ön eğitim (Pre-training):** Bu aşama, bir çocuğun dili öğrenme sürecine benzer. Model, milyarlarca kelime ve cümle içeren devasa bir metin veri setiyle "beslenir". Bu veri seti, internette toplanan makaleler, kitaplar, kodlar ve diğer metin türlerini içerebilir. Bu aşamada model, dilin temel yapısını, kelimelerin ve cümlelerin nasıl bir araya geldiğini, dil bilgisi kurallarını ve genel dünya bilgisini öğrenir. Henüz belirli bir göreve odaklanmaz, genel bir dil anlayışı geliştirir.

- **İnce ayar (Fine-tuning):** Ön eğitimden sonra model, belirli bir görevde uzmanlaşması için "ince ayar" aşamasına geçer. Bu aşama, bir öğrencinin üniversitede belirli bir alanda uzmanlaşmasına benzer. Model, yapacağı göreve özel, daha küçük ve "etiketlenmiş" veri setleriyle eğitilir. Örneğin, bir çeviri modeli eğitmek için, kaynak dil ve hedef dildeki cümle çiftlerinden oluşan bir veri seti kullanılır. Bu veri seti "etiketlenmiş" olarak adlandırılır çünkü her veri örneği (kaynak dildeki cümle) karşılık gelen doğru çıktı (hedef dildeki cümle) ile eşleştirilmiştir. İnce ayar aşaması, modelin genel dil bilgisini belirli bir göreve uyarlamasını ve o görevde yüksek performans göstermesini sağlar.

**Ön eğitim aşaması, LLM'lerin başarısı için kritik öneme sahiptir.** Model ne kadar çok ve kaliteli veriyile eğitilirse, dil anlayışı ve üretme yeteneği o kadar gelişmiş olur. Ancak, özellikle Türkçe gibi kaynakların kısıtlı olduğu dillerde, bu büyüklükte veri setlerine ulaşmak oldukça zordur. Bu durum, Türkçe doğal dil

işleme teknolojilerinin gelişimini ve küresel rekabetteki yerini doğrudan etkilemektedir.

Bu makalenin amacı, 8 milyar parametrelili bir Türkçe LLM'yi sıfırdan eğitmek için (Ön eğitim (Pre-training)) gereken metin miktarını kitap sayısı üzerinden tahmin etmektir. Bu amaçla, öncelikle açık kaynaklı LLM'lerin kullandığı veri setlerinin büyüklükleri ve performansları incelenecektir. Daha sonra, bu verilerden yola çıkarak Türkçe için gereken veri miktarı hesaplanacaktır. Son olarak, Türkiye'deki mevcut Türkçe metin kaynakları değerlendirilerek, bu ihtiyacı karşılanıp karşılanamayacağı tartışılacaktır. Ayrıca, Türkiye'de büyük dil modeli geliştirilmesinde karşılaşılan veri ve altyapı eksikliklerine değinilip bu eksikliklerin üstesinden gelinmesine yönelik bazı öneriler sunulacaktır.

### 2. Açık Kaynak Dil Modellerinin Eğitiminde Kullanılan Veri Setlerinin İncelenmesi

Doğal dil işleme alanında büyük dil modelleri (LLM), trilyonlarca token içeren devasa veri setleriyle eğitilerek üstün başarılar elde etmektedir. Bu veri setlerinin büyüklüğü ve çeşitliliği, modelin dil anlayışı ve metin üretim yeteneklerini doğrudan etkileyen kritik faktörlerdir. Farklı kaynaklardan gelen metinlerin bir araya getirilmesi, modellerin çok dilli destek sunabilmesi ve anlam çeşitliliği taşıyan metinler oluşturulabilmesi açısından büyük önem taşır. Ancak, bu tür veri setleri her zaman eşit kalitede olmayabilir ve dilsel önyargılar, tekrar eden içerikler veya düşük kaliteli metinler model performansını olumsuz etkileyebilir.

Bu bölümde, Meta, Alibaba, Google gibi büyük teknoloji firmalarının açık kaynaklı modelleri olan **Llama**, **Qwen**, ve **Gemma** gibi örneklerin veri seti boyutları, içerikleri ve performansları incelenerek, Türkçe dil modeli geliştirmek için hedeflenmesi gereken veri miktarı hakkında fikir sahibi olunacaktır.

### Llama Modelleri

Meta firması tarafından geliştirilen **Llama 3.1 modeli**, yaklaşık **15 trilyon token** ile eğitilmiştir [1]. Bu model çok dilli olmasına rağmen, eğitim verilerinin büyük bir kısmını İngilizce metinler oluşturmaktadır. Bu durum, Türkçe gibi kaynakların daha az yer aldığı diller için daha fazla veriye ihtiyaç duyulduğunu göstermektedir.

Bir sonraki sürüm olan **Llama 3.2 modeli**, 9 trilyon token ile eğitilmiştir. Bu modelde dikkat çeken bir özellik, daha küçük modeller (1 milyar ve 3 milyar parametrelili) için daha büyük modellerden (8 milyar ve 70 milyar parametrelili Llama 3.1) kullanılarak elde edilmiştir. Bu yöntem, daha küçük modellerin daha büyük modellerin bilgi birikiminden yararlanmasını sağlayarak eğitim sürecinin verimliliğini artırmaktadır.

### Qwen Modelleri

Alibaba firması tarafından geliştirilen **Qwen model ailesi**, 0,5 milyar ile 72 milyar parametre arasında değişen boyutlara sahip modellerden oluşmaktadır [2]. İlk sürüm Qwen modeli, **18 trilyon token** ile eğitilmiştir ve çok dilli bir yapıya sahiptir. Ancak bu veri setinde de İngilizce metinlerin baskın olduğu görülmektedir. Daha sonraki sürüm olan **Qwen 2.5**, yine 18 trilyon token kullanılarak eğitilmiştir ve modelin dil çeşitliliği artırılmıştır.

### Google Gemma Modelleri

Google firması tarafından geliştirilen **Gemma modeli**, başlangıçta 2 milyar ve 7 milyar parametrelili iki farklı boyutta sunulmuştur ve **6 trilyon token** ile eğitilmiştir. Daha sonraki sürüm olan **Gemma 2 modeli** ise 2 milyar, 9 milyar ve 27 milyar parametrelili varyantlar içermektedir. Bu sürümde kullanılan token miktarları sırasıyla 2 trilyon, 8 trilyon ve 13 trilyon olarak belirlenmiştir.

### Veri Setlerinin Performansa Etkisi

Bu modellerin eğitildiği devasa veri setleri, modelin dil anlayışını ve karmaşık görevlerdeki başarı oranını önemli ölçüde artırmıştır. Ancak bu veri setlerinin oluşturulması sırasında dikkat edilmesi gereken bazı noktalar bulunmaktadır:

- **Dil Çeşitliliği:** Modeller genellikle İngilizce ağırlıklı

veri setleriyle eğitildiğinden, Türkçe gibi dillerin temsiliyeti sınırlı kalabilmektedir.

- **Veri Kalitesi:** Tekrarlayan içeriklerin, düşük dilbilgisi kurallarına sahip metinlerin ve yanlış bilgiler içeren veri setlerinin modelin öğrenme sürecini olumsuz etkileyebileceği göz önünde bulundurulmalıdır.

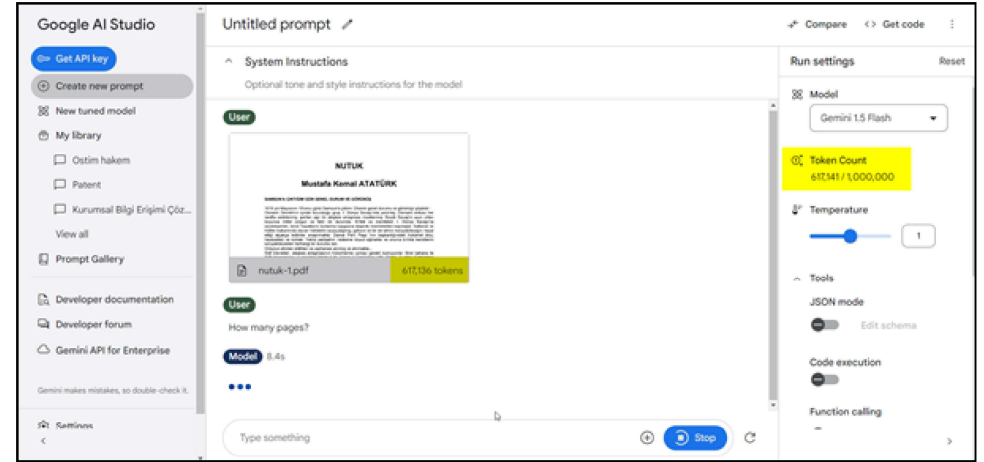
Bu bölümde ele alınan modellerin veri seti büyüklükleri ve eğitim stratejileri, Türkçe dil modeli geliştirme süreçlerinde hedeflenmesi gereken veri miktarı ve stratejileri hakkında değerli ipuçları sunmaktadır.

### 3. 8 Milyar Parametrelili Türkçe Dil Modeli için Veri Setinin Tahminlenmesi:

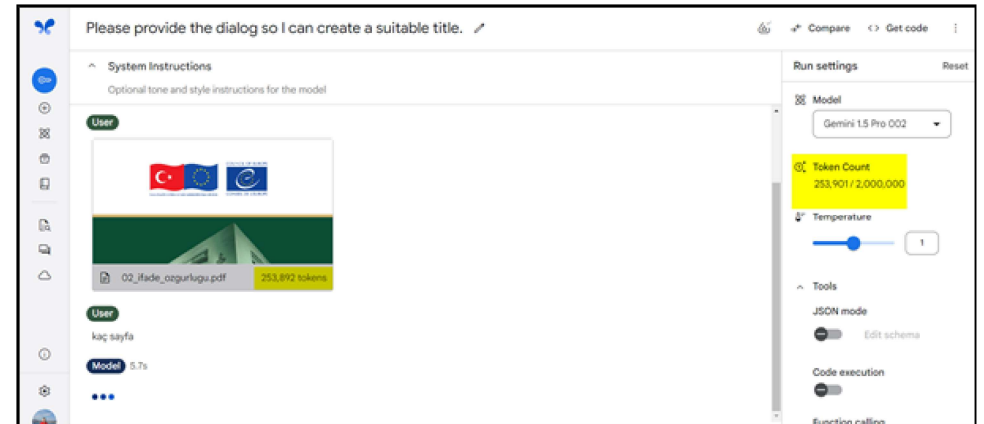
Bir önceki bölümde açık kaynaklı büyük dil modellerinin kullandığı veri setlerinin büyüklükleri incelendi. Bu incelemelerden yola çıkarak 8 milyar parametrelili bir Türkçe LLM'yi eğitmek için **6 ila 9 trilyon token** arasında veriye ihtiyaç duyulacağı tahmin edilmektedir. Sadece Türkçe için eğitilen bir model olacağı düşünülürse, bu rakam **4 ila 6 trilyon token** aralığına da düşürülebilir. Bu bölümde, 6 trilyon tokenlık bir veri setinin kaç kitaba karşılık geldiği hesaplanarak, bir Türkçe LLM'nin eğitimi için ne kadar büyük bir veri setine ihtiyaç duyulduğu somutlaştırılmaya çalışılacaktır.

Bu tahmini somutlaştırmak adına, token sayısının kitap karşılığı hesaplanmıştır. Öncelikle bir kitabın ortalama kaç token içerdiği gözlemlenmiştir. Bu hesaplama için üç farklı örnek kitap kullanılmıştır. Bu kitaplar ve içerdikleri yaklaşık token sayıları şunlardır: 430 sayfalık *Nutuk* (600.000 token) [4], 258 sayfalık *Anayasa Mahkemesi* ifade Özgürlüğü (250.000 token) [5] ve 403 sayfalık *Türk Aile Ansiklopedisi* (210.000 token) [6]. Bu örneklerdeki sayfa sayısı ve token sayısı arasındaki ilişki incelendiğinde, sayfa başına ortalama 1000 token olduğu varsayılmıştır. Dolayısıyla, 600 sayfalık bir kitabın yaklaşık 600.000 token içerdiği öngörülmüştür. Bu doğrultuda:

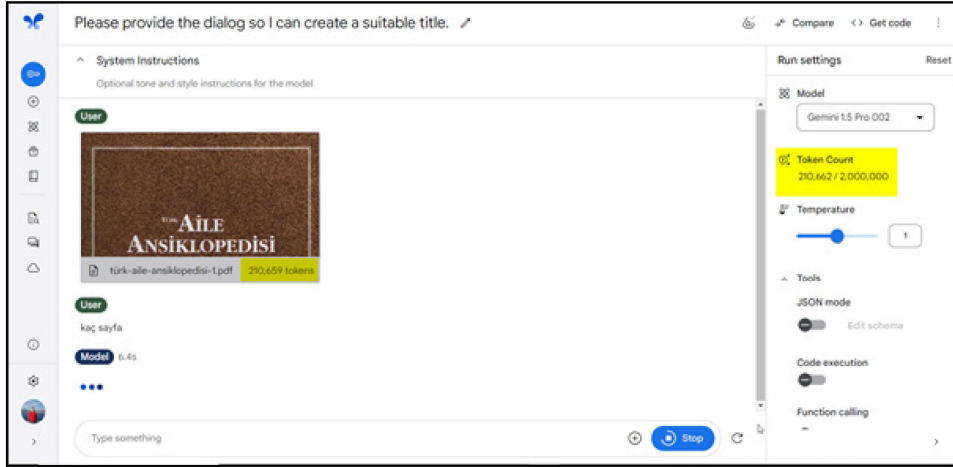
- **1.000 kitap:** Yaklaşık **600 milyon token**,
- **1 milyon kitap:** Yaklaşık **600 milyar token** anlamına gelmektedir.



Bu varsayım ışığında, **6 trilyon token** büyüklüğünde bir veri setinin yaklaşık **10 milyon kitaba eşdeğer** olduğu hesaplanmıştır. Diğer bir deyişle, 8 Milyar Parametrelili Türkçe Dil Modelinin sıfırdan eğitimi için yaklaşık **10 milyon kitaba eşdeğer bir metne ihtiyaç vardır**.



Sonuç olarak, 8 milyar parametrelili bir Türkçe LLM'yi eğitmek için oldukça büyük bir veri setine ihtiyaç duyulmaktadır. Bu veri setinin büyüklüğü, Türkçe dil kaynaklarının sınırlı olduğunu göz önünde bulundurduğumuzda, önemli bir zorluk teşkil etmektedir.

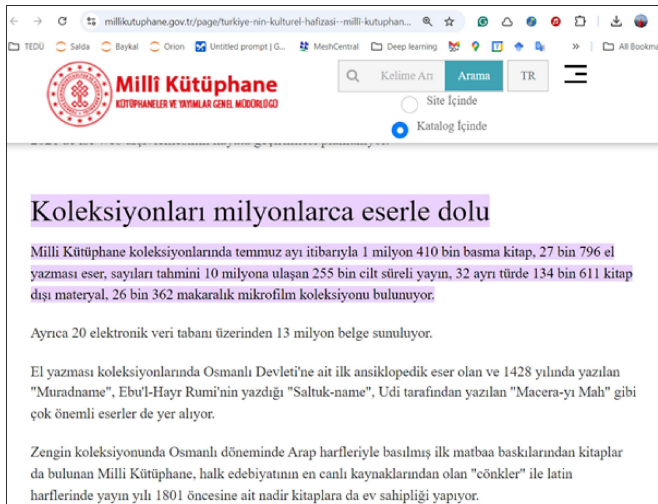


#### 4. Türkçe Yazılmış ve Üretilmiş Metin Miktarının Tahminlenmesi

8 milyar parametrelili bir Türkçe LLM'nin başarılı bir şekilde eğitilmesi için yaklaşık **10 milyon kitaba eşdeğer** metin verisine ihtiyaç duyulmaktadır. Ancak Türkçe metin kaynaklarının bu ihtiyacı karşılayıp karşılayamayacağı sorusu, dil modeli geliştirme sürecinde en önemli zorluklardan birini oluşturmaktadır. Bu bölümde, Türkiye'deki mevcut metin kaynaklarının kapasitesi değerlendirilmiş ve bu ihtiyacı karşılamak üzere önerilerde bulunulmuştur.

##### Basılı Kitap Kaynakları

Türkiye'deki basılı kitap kaynakları, Milli Kütüphane'nin verilerine göre yaklaşık **1.4 milyon kitap** ile sınırlıdır [7]. Bu miktar, hedeflenen 10 milyon kitaba kıyasla oldukça düşük kalmakta ve Türkçe dil modeli geliştirme sürecinde önemli bir veri açığı oluşturmaktadır. Üstelik bu kitapların tamamının dijitalleştirilmemesi olması ve büyük bir kısmının telif hakkı koruması altında bulunması, bu kaynakların doğrudan kullanımını daha da zorlaştırmaktadır.



##### Web Tabanlı Türkçe Metin Kaynakları

Web tabanlı metin kaynakları, büyük dil modelleri için önemli bir veri kaynağıdır. Düzenli olarak internetteki sayfaları tarayarak büyük veri kümeleri oluşturan **Common Crawl** projesi bu alandaki en kapsamlı veri kaynaklarından biridir [8]. Common Crawl verilerine göre, web sayfalarının dil dağılımında Türkçe sayfalar yaklaşık **%1.2** oranında yer almaktadır. 2024 yılı Eylül ayı tarama verilerine göre:

- **Toplam indirilen web sayfası:** 2.8 milyar,
  - **Toplam veri boyutu:** 410 TB,
- Bu verilere dayanarak, Türkçe

içeriklerin boyutu yaklaşık **5.41 TB** olarak hesaplanabilir.

Literatürde **1 TB HTML metninin yaklaşık 250 milyar token içerdiği** varsayılmaktadır. Bu doğrultuda:

- **5.41 TB'lık Türkçe metin verisi:** Yaklaşık **1.35 trilyon token** içerir ve bu da **2.25 milyon kitaba eşdeğer** metin anlamına gelmektedir.

##### Türkçe Metin Kaynaklarının Toplam Değerlendirilmesi

Milli Kütüphane'deki basılı kitaplar ve Common Crawl verilerinden gelen metinler birlikte değerlendirildiğinde, toplamda yaklaşık **3.7 milyon kitaba eşdeğer** metin elde edilmektedir. Bu sayı, hedeflenen 10 milyon kitabın yalnızca **%37'sini** oluşturmaktadır. Dolayısıyla, **6.3 milyon kitaba eşdeğer** ek metin kaynağına daha ihtiyaç duyulmaktadır.

##### Web Verisinin Kalitesi ve Sorunlar

Web tabanlı Türkçe metinlerin miktarı belirli bir düzeyde ihtiyacı karşılayabilse de bu verilerin kalitesi, dil modeli eğitiminde önemli bir engel oluşturmaktadır. Web verilerinin düşük kalitesi aşağıdaki sorunları beraberinde getirebilir:

- **Tekrarlayan İçerikler:** Haber siteleri, forumlar ve bloglar aynı bilgileri farklı biçimlerde sunarak veri setinde tekrarlara neden olabilir.
- **Dilbilgisi ve Yazım Hataları:** İnternet metinlerinin önemli bir kısmı dilbilgisi kurallarına uygun olmayan

##### < Back to Blog

September 24, 2024

## September 2024 Crawl Archive Now Available

The crawl archive for September 2024 is now available. The data was crawled between September 7th and September 21st 2024, and contains **2.8 billion web pages (or 410 TiB of uncompressed content)**.



Thom Vaughan

Thom is Principal Technologist at the Common Crawl Foundation.

veya yazım hataları içeren metinlerden oluşur.

- **Yanlış Bilgi ve Önyargılar:** Web kaynaklı metinler yanlış bilgi veya önyargılı içerik barındırabilir, bu da modelin olumsuz eğilimler geliştirmesine yol açabilir.

##### Veri Temizleme ve Kalite Artırma Yöntemleri

Türkçe LLM geliştirme sürecinde, kaliteli bir veri seti oluşturmak için dikkatli veri temizleme ve filtreleme yöntemleri uygulanmalıdır:

- **Tekrarlayan İçeriklerin Elenmesi:** Veri setlerinde aynı bilgilerin tekrar edilmesini önlemek için benzersiz metinlerin seçilmesi gerekmektedir.
- **Dilbilgisi ve Yazım Hatalarının Düzeltilmesi:** Otomatik dil işleme teknikleri kullanarak metinlerin dilbilgisi hataları giderilmeli ve metin kalitesi artırılmalıdır.
- **Yanlış Bilgilerin Ayıklanması:** Güvenilir kaynaklardan gelen metinlerin seçilmesi, yanlış bilgi içeren veri kümelerinin elenmesi için önemlidir.

Bunlara ek olarak, **Türkiye'de büyük dil modeli geliştirmek için gerekli altyapı eksikliği** de önemli bir engel teşkil etmektedir. Büyük dil modellerini eğitmek için güçlü donanım ve yazılım altyapısı gerekmektedir. Bu altyapının eksikliği, Türkiye'deki araştırmacıların ve geliştiricilerin çalışmalarını sınırlandırmaktadır.



## 5. Değerlendirmeler

Bu çalışmada, 8 milyar parametrelili bir Türkçe büyük dil modeli (LLM) geliştirmek için gerekli veri miktarı ve Türkiye’de mevcut veri kaynaklarının bu ihtiyacı karşılama potansiyeli incelenmiştir. Elde edilen bulgular, Türkçe metin kaynaklarının sınırlı olduğunu ve dil modeli eğitimi sürecinde çeşitli zorluklarla karşılaşıldığını göstermektedir. Bu bağlamda veri eksikliğinin üstesinden gelmek ve kaliteli bir Türkçe dil modeli geliştirmek için önerilen çözüm stratejileri ele alınmıştır.

### Veri Eksikliği

Yukarıdaki tartışmalar ışığında, başarılı bir Türkçe LLM eğitimi için en az 6 trilyon token’ın gerekliliği ifade edilebilir. Bu miktar, örnek Türkçe kitaplar üzerinden yapılan analizler ve sayfa başına ortalama 1000 token varsayımına dayanarak, yaklaşık 10 milyon kitaba eşdeğer bir veri büyüklüğüne denk gelmektedir. Mevcut veri kaynakları incelendiğinde ise bu hedefe ulaşmanın ne kadar zorlu olduğu ortaya çıkmaktadır. Milli Kütüphane ve basılı kaynaklar yaklaşık 1.4 milyon kitap içerirken, Common Crawl verileri yaklaşık 2.25 milyon kitaba eşdeğer metin sunmaktadır. Bu iki kaynaktan elde edilebilecek toplam veri, yaklaşık 3.7 milyon kitaba denk gelmektedir. Teorik olarak hem Milli Kütüphane’deki tüm kitaplara hem de web’deki tüm Türkçe web sitelerine ulaşılabilirse dahi, elde edilebilecek veri miktarı bu seviyededir. Ancak, telif hakkı kısıtlamaları, teknik zorluklar, web verilerinin kalitesizliği ve tekrar içerme olasılığı gibi çeşitli nedenlerden dolayı bu teorik limitin pratikte erişilebilir olmadığı düşünülmektedir. Sonuç olarak, hedeflenen 10 milyon kitabın yalnızca %37’sine denk gelen 3.7 milyon kitaplık veri, büyük bir veri açığına işaret etmekte ve Türkçe LLM eğitimi için yaklaşık 6.3 milyon kitaba daha ihtiyaç duyulduğunu göstermektedir.

### Alternatif Çözüm Önerileri

Veri eksikliğinin üstesinden gelmek için aşağıdaki alternatif yöntemler önerilmektedir:

#### 1. Sentetik Veri Üretimi

Derin öğrenme teknikleri kullanılarak yapay metin üretimi yapılabilir. Bu yöntem, gerçek dünyadaki verilerin yapısını taklit eden veri setleri oluşturarak veri eksikliğini gidermeyi amaçlar. Sentetik veri ile dil bilgisini içerecek örnekler üretilip eğitimde kullanılabilir. Ancak bu yöntemde dikkat edilmesi gereken en önemli husus, sentetik verinin dilin doğallığını tam anlamıyla yansıtmaması ve gerçek

dünyaya dair bilgileri içermemesidir. Örneğin, İstanbul’un ne zaman fethedildiği gibi tarihsel bir bilgi veya belirli bir kültürel pratiğe dair detaylar sentetik veri üretimi ile elde edilemez. Sentetik veri, dilin yapısını ve istatistiksel özelliklerini öğrenebilir, ancak tarih, sosyal, kültürel, siyasi vb. alanlarda anlamlı ve doğru bilgi üretemez. Bu tür bilgilerin sentetik olarak üretilmesi girişimleri ya başarısızlıkla sonuçlanır ya da hatalı, kendini tekrarlayan veya mevcut verilerdeki önyargıları yansıtan taraflı bilgilere yol açabilir. Bu durum, modelin gerçekçi olmayan ve olgusal hatalar içeren metinler üretmesine neden olabilir. Dolayısıyla, sentetik veri üretimi, veri eksikliğine bir çözüm olarak görülse de gerçek dünya bilgisine dayalı içerik üretme konusunda sınırlı bir çözümdür.

#### 2. Damıtma (Distillation) Eğitim Yöntemleri

Daha küçük bir modelin (“student”) sınırlı veri ile eğitilmesinin ardından daha büyük bir modelin (“teacher”) rehberliğinde bu küçük modelin performansı artırılabilir. Bu yöntem damıtma (distillation) denir. Bu yöntem özellikle veri miktarının kısıtlı olduğu durumlarda etkili bir öğrenme stratejisi sunmaktadır. Türkçe LLM’ler için bu yöntemin uygulanabilirliği, uygun bir “teacher” modelin bulunup bulunmadığına bağlıdır. Çok dilli bir BDM (LLM) öğretmen (teacher) olarak kullanılabilir, ancak Türkçe dil becerisi değerlendirilmelidir. Öğrenci (Student) model, çeşitli ve temizlenmiş Türkçe verilerle eğitilirken, bilgi aktarımı bilgi damıtması (knowledge distillation) gibi yöntemlerle gerçekleştirilebilir. Başarısı, Öğrenci (Student) modelin çeşitli NLP görevlerindeki performans artışı ile ölçülür. Bu yöntemin potansiyeli olsa da etkinliği dikkatlice değerlendirilmelidir.

#### 3. Gelişmiş Embedding Modelleri

Kelimeleri ve cümleleri sayısal vektörlere dönüştüren “embedding” teknikleri, anlam benzerliklerini öğrenmek için önemlidir. Türkçe dil modeli eğitimi sırasında çok dilli “embedding” yöntemlerinden faydalanılarak diğer dillerdeki verilere erişim sağlanabilir. Bu, veri çeşitliliğini artırarak modelin daha zengin dilsel yapılar öğrenmesine olanak tanır.

#### 4. Çoklu Dil Modellerinin Kullanımı

Büyük miktarda İngilizce verilerle eğitilmiş çok dilli modellerden Türkçe dil bilgisini aktarmak, Türkçe dil modelinin eğitimi için faydalı olabilir. Bu yöntem, veri eksikliğini kısmen gidermeyi ve modelin performansını artırmayı amaçlamaktadır.

## 5. Veri Temizliği ve Kalite Artırımı

Web tabanlı verilerin düşük kalitesi nedeniyle, metinlerin dikkatlice filtrelenmesi ve temizlenmesi gerekmektedir. Tekrarlayan içerikler, dilbilgisi hataları ve yanlış bilgiler içeren metinler ayıklanmalı ve kaliteli veri kaynakları önceliklendirilmelidir.

### Telif Hakları Sorunu ve Açık Veri Paylaşımı

Veri toplama ve kullanımı sürecinde karşılaşılan önemli bir engel de telif haklarıdır. Basılan birçok kitabın telif hakkı koruması altında olması, bu kaynakların dijital ortamda kullanılmasını zorlaştırmaktadır. Telif hakkı sahiplerinden izin almak uzun ve maliyetli bir süreçtir. Açık veri paylaşımının teşvik edilmesi elbette önemli bir adım olmakla birlikte, telif hakkı sorununu aşmak için daha somut ve uygulanabilir çözümler gerekmektedir. Bu noktada şu öneriler değerlendirilebilir:

● **Yasal Düzenlemeler:** Araştırma projeleri için telif hakkı koruması altındaki eserlere erişimi kolaylaştıran ve açık veri paylaşımını teşvik eden yasal düzenlemeler yapılmalıdır.

● **Açık Kaynak Projelerinin Desteklenmesi:** Daha fazla dil verisinin paylaşılmasını sağlayacak açık kaynak projelerine maddi ve manevi destek verilmelidir.

● **Telif Hakkı Olmayan Metinlerin Derlenmesi:** Telif hakkı olmayan Türkçe metinlerin toplanması, sınıflandırılması ve araştırmacıların kullanımına sunulması için bir platform oluşturulabilir. Bu platform, kamu malı eserleri, telif süresi dolmuş kitapları, açık lisanslı metinleri ve Creative Commons gibi lisanslarla paylaşılan içerikleri barındırabilir.

● **Telif Hakkı Sahipleriyle İş Birliği:** Telif hakkı sahipleriyle, eserlerinin araştırma amaçlı kullanımına izin verecek anlaşmalar yapılması için girişimlerde bulunulabilir. Bu, belirli bir ücret karşılığında veya belirli kullanım koşullarıyla eserlere erişim sağlanmasını içerebilir.

Bu şekilde çok yönlü bir yaklaşım benimsenerek, telif hakkı engelleri minimize edilebilir ve Türkçe dil modeli geliştirme çalışmaları için gereken veriye daha kolay erişim sağlanabilir.

### Türkçe NLP Alanındaki İlerleme İçin Öneriler

Bu araştırma sonucunda ortaya çıkan bulgular, Türkiye’nin doğal dil işleme (NLP) alanında küresel rekabet gücünü artırmak için stratejik adımlar atılması gerektiğini göstermektedir:

## ● Milli Kütüphane ve Akademik Kurumlar İş Birliği:

Veri dijitalleştirme çalışmalarının hızlandırılması ve Türkçe metin kaynaklarının araştırmacılara sunulması.

● **Endüstri-Akademi Ortaklıkları:** Büyük dil modeli projelerine destek vermek için özel sektör ve akademik kuruluşlar arasında iş birliğinin teşvik edilmesi.

● **Eğitim ve Farkındalık Çalışmaları:** Veri bilimi ve NLP alanında yetişmiş insan kaynağının artırılması için üniversiteler ve teknoloji şirketlerinin desteklediği programların oluşturulması.

## 6. Sonuç

Bu çalışma, Sayısal ortamdaki Türkçe kaynakların henüz yeterli düzeyde olmaması nedeniyle bir dille büyük bir dil modeli (BDM - LLM) eğitmenin önemli veri zorluklarını ortaya koymaktadır. 8 milyar parametrelili bir Türkçe BDM için yaklaşık 10 milyon kitaba eşdeğer veriye ihtiyaç duyulmasına rağmen, mevcut kaynaklardan yalnızca 3.7 milyon kitaplık veri elde edilebilmektedir. Bu büyük veri açığı, telif hakkı sorunları, web verilerinin kalitesi ve tekrar içerme gibi etkenlerle daha da derinleşmektedir. Sentetik veri üretimi, “teacher-student” eğitimi, gelişmiş gömme modelleri ve çok dilli modellerden bilgi aktarımı gibi alternatif çözümler, veri kısıtlamalarını hafifletme potansiyeline sahip olsa da her birinin kendine özgü sınırlamaları bulunmaktadır. Sonuç olarak, başarılı bir Türkçe BDM geliştirmek için, veri zenginleştirme stratejilerinin yanı sıra, açık veri paylaşımını teşvik eden yasal düzenlemeler, veri dijitalleştirme çalışmaları, endüstri-akademi ortaklıkları ve doğal dil işleme (NLP) alanında uzman yetiştirmeye yönelik programlar gibi çok yönlü bir yaklaşım gerekmektedir.

[1] Meta Llama resmi web sayfası, <https://www.llama.com/>

[2] Qwen Github hesabı, <https://github.com/QwenLM>

[3] Google Gemma resmi web sayfası, <https://ai.google.dev/gemma>

[4] Nutuk, PDF kopyası, <https://ekitap.ktb.gov.tr/TR-273376/nutuk.html>

[5] Anayasa Mahkemesi ifade Özgürlüğü, [https://anayasa.gov.tr/media/3545/02\\_ifade\\_ozgurlugu.pdf](https://anayasa.gov.tr/media/3545/02_ifade_ozgurlugu.pdf)

[6] Türk Aile Ansiklopedisi, <https://ekutuphane.aile.gov.tr/media/x2bffa1/t%C3%BCrk-aile-ansiklopedisi-1.pdf>

[7] Milli Kütüphane resmi web sayfası, <https://www.millikutuphane.gov.tr/>

[8] Common Crawl, <https://commoncrawl.org/>