TR-DocVQA-Synth: Türkçe İş Belgeleri için Büyük Ölçekli Sentetik bir Görsel Soru-Cevap Veri Seti ve Görsel-Dil Modelleriyle Değerlendirme

Anonim Yazar(lar)
Anonim Kurum
Anonim Şehir, Türkiye
Çift-kör değerlendirme için gizlenmiştir

***Abstract*—Türkçe iş belgelerinden otomatik bilgi çıkarma, mevcut çok-dilli sistemler için açık bir boşluk olmayı sürdürmektedir; çünkü bu alanda hâlâ kamuya açık, büyük ölçekli ve değerlendirilebilir bir Türkçe Belge Görsel Soru-Cevap (DocVQA) referansı bulunmamaktadır. Bu çalışmada TR-DocVQA-Synth adıyla yayımladığımız sentetik bir Türkçe DocVQA veri setini ve üç farklı görsel-dil mimarisinin karşılaştırmalı değerlendirmesini sunuyoruz. Veri seti, üç ana iş belgesi ailesinde (e-Fatura/e-Arşiv tarzı faturalar, ticari sözleşmeler ve teklif/sipariş belgeleri) sentetik olarak üretilmiş 15.000 belge görüntüsü ve 235.000 soru-cevap çifti içermektedir. Veri seti tasarımı, gerçek kişisel ve kurumsal verileri ifşa etmeden çeşitli Türkçe belge yerleşimlerini ve alan-düzeyinde akıl yürütmeyi kapsayacak biçimde yapılmıştır. Üç model değerlendirilmiştir: ince ayar uygulanan OCR-bağımsız bir kodlayıcı-çözücü (Donut), LoRA tabanlı bir görsel-dil modeli ince ayarı (PaliGemma-3B) ve İngilizce DocVQA üzerinde önceden eğitilmiş bir model üzerinde sıfır-atış değerlendirme (Pix2Struct). LoRA ile ince ayarlı PaliGemma-3B modeli, 2.000 örneklik test setinde 72,05 % normalleştirilmiş EM ve 87,45 % ANLS sonucu elde ederek diğer iki temel modeli belirgin bir farkla geride bırakmıştır. Buna karşın OCR-bağımsız ince ayarlı Donut ve sıfır-atış Pix2Struct yapılarının Türkçe sayısal alanlar ve para birimi biçimlendirmede zorlandığını gözlemledik. Hem veri seti hem değerlendirme protokolü kamuya açık olarak yayımlanmaktadır.**

***Index Terms*—Belge görsel soru-cevap, Türkçe NLP, sentetik veri seti, PaliGemma, Donut, görsel-dil modelleri, belge anlama**

I. Giriş

Belge Görsel Soru-Cevap (DocVQA), bir belge görüntüsünü ve doğal dilde bir soruyu girdi olarak alıp doğrudan görsel kanıttan cevap üreten çok-modlu bir görevdir [1]. Bu görev, klasik OCR→NLP zincirinin aksine, OCR hatalarının aşağı akışa yayılmasını azaltır ve belge yerleşim bilgisinden faydalanır. Görev tanımı, kaynak verisi büyük ölçüde İngilizce olan çok sayıda benchmark ile olgunlaşmıştır [1]–[4].

Türk iş ekosistemi yüksek hacimli yapılandırılmış belgeler üretmektedir: GİB e-Fatura ve e-Arşiv standartlarına uygun faturalar, ticari sözleşmeler ve teklif/sipariş belgeleri bunların başında gelmektedir. Bu belgeler dilsel ve yapısal olarak ayırt edici özelliklere sahiptir: (i) sözcük sonlarını değiştiren Türkçe eklemeli morfoloji [5], (ii) Türk şirket adlandırma kalıpları ve VKN/TCKN biçimindeki kimlik numaraları, (iii) Türk Lirası tutarları ve gün-ay-yıl formatında tarihler, (iv) e-fatura kalem tablolarının ve sözleşme maddelerinin kendine özgü yerleşimleri.

Buna karşın Türkçeye özgü kamuya açık ve büyük ölçekli bir DocVQA referansı bilgimiz dahilinde bulunmamaktadır. Bu boşluk, çok-dilli modellerin Türk belgeler üzerinde değerlendirilmesini güçleştirmekte ve alana özgü yöntemlerin geliştirilmesini geciktirmektedir.

Katkılarımız. Bu çalışmanın katkıları üç başlık altında özetlenebilir:

- **TR-DocVQA-Synth:** Üç iş belgesi ailesini kapsayan, 15.000 belge görüntüsü ve 235.000 soru-cevap çiftinden oluşan, kamuya açık sentetik bir Türkçe DocVQA veri seti. Veri seti, alan-düzeyinde sağlama (provenance) kayıtları ve birden çok yerleşim şablonu ile birlikte sunulmaktadır.
- **Karşılaştırmalı temel modeller:** Üç farklı paradigmanın aynı protokol altında değerlendirilmesi: (a) ince ayarlı OCR-bağımsız kodlayıcı-çözücü (Donut-TR), (b) LoRA ile ince ayarlı görsel-dil modeli (PaliGemma-3B), (c) İngilizce DocVQA üzerinde önceden eğitilmiş bir modelin sıfır-atış uygulaması (Pix2Struct).
- **Değerlendirme protokolü ve hata analizi:** Türkçe karakter farkındalıklı normalleştirme ile EM, ANLS ve Token F1 metrikleri; belge türüne, cevap türüne ve hata kategorisine göre ayrıştırılmış sonuçlar.

II. İlgili Çalışmalar

A. Belge Anlama Modelleri

LayoutLM [6] ve ardılı LayoutLMv3 [7], metin gömmelerini 2B konum ve görüntü yaması gömmeleriyle birleştirerek belge anlama görevlerinde temel mimariler haline gelmiştir. Bu modellerin uygulamada OCR çıktısına ihtiyaç duyması, bağımlı OCR motorlarının kalitesini doğrudan modelin tavanına dönüştürmektedir.

Donut [8], OCR-bağımsız uçtan-uca bir kodlayıcı-çözücü mimarisi önererek görüntüden hedef diziye dönüşümü tek bir modelde gerçekleştirmektedir. Pix2Struct [9] ise ekran görüntülerini ayrıştırmaya odaklanan bir ön eğitim hedefi tanımlamış ve DocVQA dahil birçok görevde güçlü sıfır-atış başlangıçları sağlamıştır. Daha yakın zamanda PaliGemma [10], görsel-dil aktarımı için tasarlanmış 3B parametreli bir model olarak çıkmıştır; bu çalışmada PaliGemma-3B'yi LoRA [11] ile uyarlamaktayız.

B. DocVQA Veri Setleri

İngilizce DocVQA [1], ardından gelen InfographicVQA [2], VisualMRC [12] ve DUDE [4] alanı tanımlayan benchmarklar olarak öne çıkmıştır. Çok-dilli sürümler ve Belge Koleksiyonu VQA [3] gibi uzantılar bulunmakla birlikte, Türkçeye özgü kamuya açık bir referans bulunmamaktadır. OCR-IDL [13] gibi büyük ölçekli OCR koleksiyonları da Türkçe içeriği marjinal düzeyde temsil etmektedir.

C. Türkçe NLP Kaynakları

Türkçe NLP, sözdizimsel çözümleme [5] ve dil modelleme [14] alanlarında belirli bir olgunluğa ulaşmıştır. Buna karşın belge anlama ve özel olarak DocVQA için Türkçe kaynakların eksikliği, hem akademik karşılaştırmaları hem de Türk iş otomasyon uygulamalarını sınırlamaktadır.

III. TR-DocVQA-Synth Veri Seti

A. Tasarım İlkeleri

Veri setinin temel tasarım amaçları şunlardır:

- *Gizlilik ve telif uyumluluğu:* Tüm belgeler sentetik olarak üretilmiştir; gerçek kişisel veya kurumsal veriler kullanılmamıştır.
- *Yerleşim çeşitliliği:* Her belge ailesi için birden çok şablon kullanılarak modelin tek bir görsel yapıya aşırı uyum sağlamasının önüne geçilmesi hedeflenmiştir.
- *Alan-düzeyinde sağlama:* Her cevap, kaynak yapılandırılmış kayıttaki bir alana bağlanmıştır; bu bağlantı yeniden üretilebilir değerlendirme ve hata çözümlemesi için kayıt altına alınmıştır.
- *Türkçe biçimsel öğeler:* Türk Lirası işaretleri, gün.ay.yıl tarih biçimi, VKN/TCKN benzeri kimlik şablonları ve şirket adlandırma kalıpları açıkça modellenmiştir.

B. Belge Aileleri

TR-DocVQA-Synth üç belge ailesini kapsamaktadır:

e-Fatura / e-Arşiv biçimli faturalar. Satıcı/alıcı bilgileri, VKN, fatura numarası, ürün/hizmet kalem tablosu, miktar, birim fiyat, KDV oranı ve ödeme tutarı gibi alanları içeren faturalar.

Ticari sözleşmeler. On farklı sözleşme tipi (hizmet, mal alımı, kira, yazılım lisansı, danışmanlık, gizlilik, tedarik, abonelik, bakım/destek, eğitim) için sözleşme numarası, taraflar, VKN'ler, konu, bedel, başlangıç/bitiş tarihleri, ödeme süresi, cezai şart ve yetkili mahkeme alanlarını içeren belgeler.

Teklif, proforma ve sipariş belgeleri. Fiyat teklifi, proforma fatura, satın alma siparişi ve benzeri belgelerde belge numarası, geçerlilik tarihi, ürün/hizmet tablosu, indirim, KDV, net/genel toplam ve teslimat/ödeme koşulları gibi alanlar.

C. Üretim Süreci

Veri seti iki aşamalı bir sentetik veri üretim hattıyla hazırlanmıştır. **Birinci aşamada**, her belge için yapılandırılmış kaynak kayıt üretilmiştir: şirket adları, sektörler ve metin alanları sentetik olarak; sayısal alanlar (toplamlar, KDV, vade) ise deterministik kod tarafından hesaplanmıştır. **İkinci aşamada**, kaynak kayıtlar HTML şablonlarına dökülerek PDF ve PNG çıktıları üretilmiş; her belge için soru-cevap JSON eki ve manifest girdisi oluşturulmuştur. Soru-cevap çiftleri, OCR veya model tahminine değil, doğrudan yapılandırılmış kaynak kayıttaki alanlara dayanmaktadır; bu yöntem cevapların kesin bilinmesini ve belge içeriğiyle tutarlı kalmasını garanti eder.

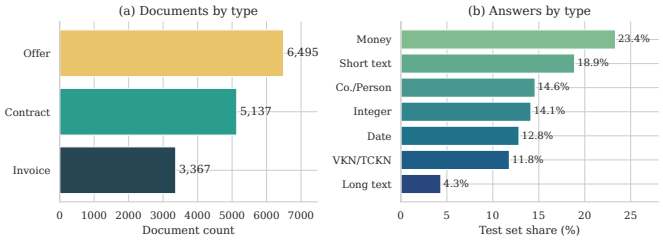

Fig. 1: TR-DocVQA-Synth dağılım istatistikleri. (a) Belge türlerine göre belge sayısı. (b) Test seti üzerinde cevap türlerinin yüzdelik dağılımı (n=2000). *Money* en geniş cevap türünü, *Long text* en daraltıcısını oluşturmaktadır.

TABLE I: Model Konfigürasyonları

Model	Strateji	OCR	Eğit. param.
Donut-TR	İnce ayar	Yok	Tam
PaliGemma-3B	LoRA ince ayar	Yok	90,4 M
Pix2Struct	Sıfır-atış	Yok	0

D. İstatistikler ve Bölümleme

Veri seti standart belge-düzeyinde bölümleme protokolüyle 80 %/10 %/10 % oranında ayrılmıştır: **eğitim** 12.000 belge / 188.026 soru, **geliştirme** 1.500 belge / 23.504 soru, **test** 1.500 belge / 23.470 soru. Aynı belgeye ait soruların farklı bölümlere düşmesi engellenmiştir.

Şekil 1, belge ve cevap türlerinin dağılımını sunmaktadır. Veri setinde teklif belgeleri en sık temsil edilen aileyken (%43,3), faturalar (%22,4) en düşük orana sahiptir; cevap türleri arasında parasal değerler ve kısa metinler baskındır.

IV. Modeller ve Değerlendirme Protokolü

A. Değerlendirilen Modeller

Üç model, aynı protokol altında 2.000 örneklik test alt kümesinde değerlendirilmiştir (Tablo I).

Donut-TR. `naver-clova-ix/donut-base` [8] kontrol noktasından başlatılmış, 50.000 örneklik bir alt kümede 3 dönem boyunca BF16 hassasiyetiyle ince ayar uygulanmıştır. Görüntü çözünürlüğü 1280×960, kod çözücü maks. uzunluğu 256 ve öğrenme oranı 3×10^{-5} olarak belirlenmiştir.

PaliGemma-3B LoRA. `google/paligemma-3b-pt-224` [10] taban modelinin dil çözücüsündeki dikkat projeksiyonlarına LoRA [11] uygulanmıştır (r=32, α=64, dropout 0,05). 50.000 örneklik alt kümede 2 dönem, öğrenme oranı 2×10^{-4} ve BF16 hassasiyetiyle eğitilmiştir. Eğitilebilir parametre sayısı 90,4 M, yaklaşık 3B toplam parametrenin %3'üne karşılık gelmektedir.

Pix2Struct (sıfır-atış). `google/pix2struct-docvqa-base` [9] kontrol noktası, İngilizce DocVQA üzerinde önceden eğitilmiş haliyle TR-DocVQA-Synth test setine herhangi bir adaptasyon yapılmaksızın uygulanmıştır.

TABLE II: TR-DocVQA-Synth test seti (n=2000) üzerinde ana sonuçlar. EM, ANLS ve Token F1 0–1 aralığındadır.

Model	EM	ANLS	Tok F1	Boş %
Donut-TR (ince ayar)	0,0150	0,0474	0,0435	21,65
Pix2Struct (sıfır-atış)	0,2105	0,5538	0,3539	0,00
PaliGemma-3B (LoRA)	**0,7205**	**0,8745**	**0,7294**	0,00

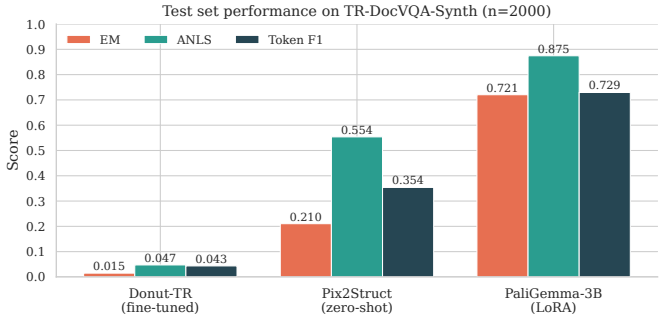

Fig. 2: Test setindeki ana metriklerin model bazında karşılaştırması. PaliGemma-3B LoRA üç metrikte de açık ara önde gelmektedir.

Tüm eğitim ve çıkarım işleri 2×NVIDIA H200 80GB GPU üzerinde TRUBA altyapısında yürütülmüştür.

B. Metrikler

Standart DocVQA metriklerini Türkçe karakter farkındalıklı bir normalleştirme katmanıyla birlikte raporluyoruz:

- **Normalleştirilmiş Tam Eşleşme (EM):** Cevap dizesi büyük/küçük harf, baş/son boşluk ve noktalama temizliğinden sonra eşit ise 1, değilse 0.
- **ANLS:** [1]'da tanımlanan DocVQA standardı; Levenshtein benzerliği 0,5'in altındaki tahminler 0 olarak puanlanır.
- **Token F1:** Normalleştirme sonrası belirteç düzeyinde F1.
- **Boş ve geçersiz tahmin oranları:** Cevap üretmeyen veya bozuk çıktı veren örneklerin oranı.
- **Hata taksonomisi:** Sayısal, biçimsel, halüsinasyon ve kopya hatalarının kategorize edilmiş dağılımı.

Türkçeye özgü karakter dönüşümleri (ör. ï/Ï, ç/Ç) korunmuş; ASCII'ye düşürülmüş versiyonlar yalnızca yardımcı bir referans olarak sunulmuştur.

V. Sonuçlar

A. Genel Performans

Tablo II ve Şekil 2 üç modelin test setindeki performansını özetlemektedir.

PaliGemma-3B LoRA modeli üç metrikte de açık ara önde yer almakta ve Pix2Struct sıfır-atış performansının yaklaşık 3,4 katı EM, 1,6 katı ANLS skoruna ulaşmaktadır. Donut-TR yapılandırması ise hem EM hem ANLS açısından kullanılabilir bir seviyenin altında kalmış;

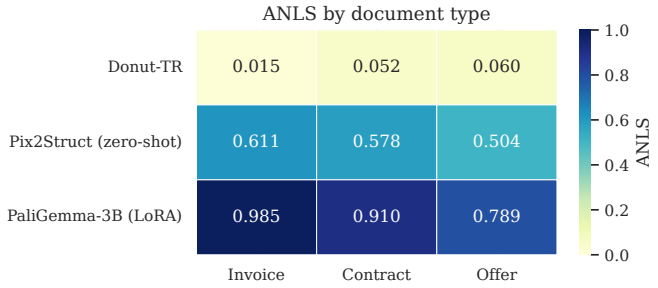

Fig. 3: Belge türlerine göre ANLS. PaliGemma-3B LoRA faturalarda 0,98 ANLS ile neredeyse doyma noktasına ulaşırken, teklif belgelerinde 0,79 düzeyine inmektedir.

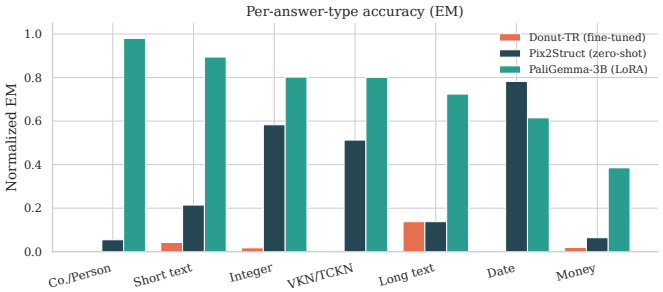

Fig. 4: Cevap türlerine göre normalleştirilmiş EM. PaliGemma-3B LoRA özel adlarda 0,98 EM ile doyma noktasındadır; parasal değerler 0,39 ile en zorlu kategoridir.

B. Belge Türüne Göre Ayrıştırma

Şekil 3, ANLS metriğinin belge türü kırılımını sunmaktadır. PaliGemma-3B LoRA modeli faturalar üzerinde **0,9847 ANLS** (test n=449) ile neredeyse mükemmel sonuca ulaşmakta, sözleşmelerde **0,9097** (n=685) ve teklif belgelerinde **0,7895** (n=866) skoruyla devam etmektedir. Performans farkı, teklif belgelerinin daha çeşitli kalem tabloları ve daha fazla parasal alan içermesinden kaynaklanmaktadır.

C. Cevap Türüne Göre Ayrıştırma

Şekil 4 cevap türüne göre normalleştirilmiş EM performansını göstermektedir.

PaliGemma-3B LoRA için en yüksek performans *şirket/kişi adı* (**0,9795 EM**, n=292) ve *kısa metin* (0,8942, n=378) kategorilerindedir. Sayısal kategoriler içinde *tamsayı* (0,8021) ve *VKN/TCKN benzeri kimlik* (0,8008) yüksek performans gösterirken *tarih* (0,6148) ve özellikle *para* (**0,3854**) belirgin biçimde geride kalmaktadır. Pix2Struct sıfır-atış modeli ise *tarih* kategorisinde sürpriz bir biçimde 0,7821 EM'e ulaşmakta; bu durum İngilizce DocVQA ön eğitiminde tarih biçimlerinin nispeten dilden bağımsız öğrenilmiş olabileceğine işaret etmektedir.

D. Hata Taksonomisi

Şekil 5 her model için hata kategorilerinin dağılımını sunmaktadır.

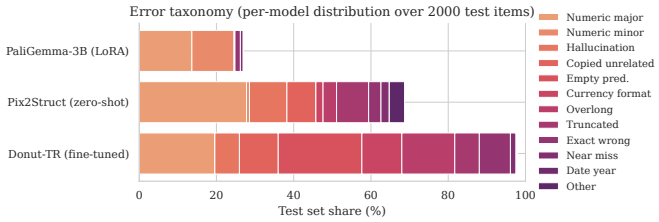

Fig. 5: Hata kategorilerinin dağılımı (test setinde 2000 örnek üzerinde her hatanın yüzdesi). Donut-TR'de baskın hata tipi boş tahmin ve büyük sayısal hatalardır; Pix2Struct'ta halüsinasyon ve kopya metin baskındır; PaliGemma-3B LoRA'da temel sınırlama parasal alanlarda sayısal hatalardır.

Donut-TR hatalarının dağılımına bakıldığında, kayıpların %21,65'i boş tahmin, %19,50'si büyük sayısal hata ve %13,70'i aşırı uzun çıktıdır. Bu örüntü, kod çözücünün uzun sayısal dizileri tutarlı şekilde üretmekte zorlandığını ve bazı durumlarda eğitim sırasında görülmemiş cevap uzunluklarına aşırı uyum sağladığını düşündürmektedir.

Pix2Struct (sıfır-atış) hata profilinde en belirgin kategori %27,80 büyük sayısal hata olmakla birlikte, %9,70 halüsinasyon ve %7,50 ilgisiz metin kopyalama gözlenmektedir. %17,21'lik boş olmayan ama geçersiz biçimde tahmin oranı, Türkçe karakter ayarlarındaki uyumsuzlukla ilişkilidir.

PaliGemma-3B LoRA ise toplam hatasının çok büyük kısmını yalnızca iki kategoride yoğunlaştırmaktadır: %13,65 büyük sayısal hata ve %10,90 küçük sayısal hata. Halüsinasyon ve boş tahmin neredeyse hiç görülmemektedir (sırasıyla %0,15 ve %0,00). Bu örüntü, modelin metin tabanlı alanlarda neredeyse doyma noktasına ulaşırken, parasal değerlerin tam haneli ezberlenmesinin hâlâ açık bir sınır olduğunu göstermektedir.

VI. Tartışma

A. Lessons Learned

Ön eğitim seçimi belirleyicidir. PaliGemma-3B ile Donut arasındaki büyük performans farkı, görsel-dil modelinin geniş ölçekli çok-dilli ön eğitiminin LoRA ile etkin biçimde aktarılabildiğini göstermektedir. OCR-bağımsız ince ayar her zaman daha basit bir hat sağlasa da, küçük ile orta ölçekli alana özgü veri setlerinde rekabet edebilmesi için çok daha geniş ön eğitim veya çok daha büyük ince ayar bütçesi gerekmektedir.

Parasal alanlar baskın sınırlama olmayı sürdürmektedir. En güçlü modelimizde bile parasal değerler %38,54 EM ile diğer alanların gerisinde kalmaktadır. Bu durum hem on binler basamağında konuşlanmış sayıların yapısal karmaşıklığı, hem de Türkçeye özgü TL biçimlendirme (`TL 1.091.897,32` gibi) nedeniyledir. Karakter düzeyinde özel ödüllendirme veya sayı-farkındalıklı dekoder modifikasyonları gelecek çalışma için açık bir yön sunmaktadır.

Sıfır-atış kalitesi belirli kategorilerde sürpriz olabilir. Pix2Struct İngilizce DocVQA üzerinde ön eğitilmiş olmasına rağmen Türkçe tarih kategorisinde 0,7821 EM'e ulaşmaktadır; bu, yapılı biçimsel öğelerin sıfır-atış aktarımının düşünüldüğünden daha kullanışlı olabileceğini düşündürür.

B. Sınırlılıklar

Bu çalışmanın açık sınırlılıkları vardır. *Birincisi*, veri seti sentetiktir ve gerçek dünyada karşılaşılan tarama gürültüsü, eğiklik, leke, el yazısı ek bilgi, mühür ve folyo gibi etkenleri içermemektedir. *İkincisi*, soru çeşitliliği sınırlıdır; sorular ağırlıklı olarak alan çıkarma niteliğindedir ve tablolar üzerinde çok-adımlı veya karşılaştırmalı muhakeme gerektiren soruların payı düşüktür. *Üçüncüsü*, OCR+LayoutLMv3 ve büyük ölçekli ticari VLM'lerle (ör. GPT-4V) sistematik karşılaştırmalar bu çalışmanın kapsamı dışında bırakılmıştır.

Bu sınırlılıklara rağmen, sentetik veri setinin kontrollü doğası — yani her cevabın yapılandırılmış bir alana bağlanabilir olması — alan-düzeyinde değerlendirme ve hata izlemesi için gerçek belgelerle eşit ölçüde kullanılabilir bir referans sağlamaktadır. Gelecek çalışmamızda gerçek anonimleştirilmiş bir test alt kümesi ile harici geçerlik analizini yayımlamayı planlamaktayız.

VII. Sonuç

Bu çalışmada Türkçe iş belgeleri için ilk büyük ölçekli kamuya açık DocVQA referansı olan TR-DocVQA-Synth'i ve üç model paradigmasının — ince ayarlı OCR-bağımsız Donut, LoRA ile ince ayarlı PaliGemma-3B ve sıfır-atış Pix2Struct — karşılaştırmalı değerlendirmesini sunduk. Sentetik veri tasarımı, gerçek kişisel veya kurumsal verileri ifşa etmeden geniş bir yerleşim ve alan çeşitliliği sağlamaktadır. PaliGemma-3B LoRA modelinin 72,05 % EM ve 87,45 % ANLS skoruyla diğer temel modelleri belirgin biçimde geride bıraktığını gösterdik. Parasal alanlardaki tam değerli doğruluk, en güçlü modelimizde bile açık bir gelişme alanı olmayı sürdürmektedir.

Hem veri setinin hem değerlendirme protokolünün kamuya açık yayımlanması, Türkçe belge anlama araştırmalarının ölçeklenmesini ve Türk iş otomasyon uygulamaları için bilgi çıkarma sistemlerinin değerlendirmesini desteklemeyi amaçlamaktadır.

Teşekkür

TÜBİTAK ULAKBİM TRUBA Yüksek Başarımlı Hesaplama altyapısının kullanımı için teşekkür ederiz.

References

[1] M. Mathew, D. Karatzas, and C. V. Jawahar, "DocVQA: A dataset for VQA on document images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 2200–2209.

[2] M. Mathew, V. Bagal, R. Tito, D. Karatzas, E. Valveny, and C. V. Jawahar, "InfographicVQA," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022, pp. 1697–1706.

[3] R. Tito, M. Mathew, C. V. Jawahar, E. Valveny, and D. Karatzas, "Document collection visual question answering," in *International Conference on Document Analysis and Recognition (ICDAR)*, 2021, pp. 778–792.

[4] J. Van Landeghem, R. Tito, Ł. Borchmann, M. Pietruszka, P. Joziak, R. Powalski, D. Jurkiewicz, M. Coustaty, B. Anckaert, E. Valveny, M. Blaschko, S. Moens, and T. Stanisławek, "Document understanding dataset and evaluation (DUDE)," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 19 528–19 540.

[5] U. Sulubacak, M. Gokirmak, F. Tyers, C. Coltekin, J. Nivre, and G. Eryiğit, "Universal dependencies for Turkish," in *Proceedings of COLING 2016*, 2016, pp. 3444–3454.
[6] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of text and layout for document image understanding," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2020, pp. 1192–1200.
[7] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "LayoutLMv3: Pre-training for document AI with unified text and image masking," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 4083–4091.
[8] G. Kim, T. Hong, M. Yim, J. Nam, J. Park, J. Yim, W. Hwang, S. Yun, D. Han, and S. Park, "OCR-free document understanding transformer," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 498–517.
[9] K. Lee, M. Joshi, I. R. Turc, H. Hu, F. Liu, J. M. Eisenschlos, U. Khandelwal, P. Shaw, M.-W. Chang, and K. Toutanova, "Pix2Struct: Screenshot parsing as pretraining for visual language understanding," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.
[10] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello *et al.*, "PaliGemma: A versatile 3B VLM for transfer," *arXiv preprint arXiv:2407.07726*, 2024.
[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.
[12] R. Tanaka, K. Nishida, and S. Yoshida, "VisualMRC: Machine reading comprehension on document images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, 2021, pp. 13 878–13 888.
[13] A. F. Biten, R. Tito, L. Gomez, E. Valveny, and D. Karatzas, "OCR-IDL: OCR annotations for industry document library dataset," in *European Conference on Computer Vision Workshops (ECCVW)*, 2022.
[14] S. Schweter, "BERTurk – BERT models for Turkish," *Zenodo*, 2020.