Ensemble of In Context Learning
Bağlamda Öğrenme Kararlarının Birleştirilmesi

Ahmed UĞUR
Bilgisayar Mühendisliği A.B.D.
Yıldız Teknik Üniversitesi
İstanbul, Türkiye
ahmed.ugur@std.yildiz.edu.tr

***Özetçe*—Bu çalışma, In Context Learning (CNL) yöntemlerinin farklı veri kümeleri üzerindeki performansını incelemektedir. Ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1 modeli ile gerçekleştirilen deneylerde, ARC ve Hellaswag veri kümeleri için ensemble yöntemlerinin bireysel model performansını bir miktar artırdığı gözlemlenmiştir. GSM8K veri kümesinde ise zaman kısıtlamaları nedeniyle ensemble yöntemi uygulanmamıştır. Çalışmanın sonuçları, CNL yöntemlerinin pratik uygulamalardaki etkinliğini ve ensemble yaklaşımları ile sağlanabilecek iyileştirmeleri ortaya koymaktadır. Gelecekte daha geniş veri kümeleri ve farklı dil modelleri ile yapılacak araştırmalar, bu alanın potansiyelini daha da ileriye taşıyacaktır.**

Anahtar Kelimeler*—*Ensemble, Few-Shot Learning, LLM Eval.

***Abstract*—This study comprehensively examines the performance of In Context Learning (CNL) methods on various datasets. Experiments conducted using the ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1 model revealed that ensemble methods significantly enhanced individual model performance for the ARC and Hellaswag datasets. Due to time constraints, ensemble methods were not applied to the GSM8K dataset. The findings highlight the practical applicability of CNL methods and the potential improvements achievable through ensemble approaches. Future research with larger datasets and diverse language models will further explore and expand the potential in this domain.**

Keywords*—*Ensemble, Few-Shot Learning, LLM Eval.

I. Giriş

Günümüz yapay zeka çalışmaları, farklı problemleri çözmek için yaratıcı ve verimli yaklaşımlar geliştirme çabasındadır. In Context Learning (CNL), öğrenmenin sadece bilgiyi ezberlemekten öte, bu bilginin gerçek dünya durumlarına ve bağlamlarına uygulanmasını gerektiren bir yaklaşımdır. Bu öğrenme türünde, bilgi soyut bir kavram olmaktan çıkarak, somut deneyimler ve pratik uygulamalarla anlam kazanır. CNL, Bir modelin öğrenme sürecini ek eğitim veya ince ayarlama gereksinimi olmaksızın, yalnızca birkaç örnekten öğrenerek gerçekleştirmesine olanak tanır. Bu özellik, özellikle büyük dil modellerinin çeşitli görevlerde kolaylıkla kullanılabilmesini sağlamaktadır.

Çalışmada, CNL yöntemlerinin farklı veri kümelerindeki performansını incelemeyi ve bu yöntemlerin ensemble modeller ile nasıl birleştirilebileceğini araştırmayı amaçlamaktadır. Çalışmada, ARC, Hellaswag ve GSM8K veri kümelerinin Türkçe versiyonları kullanılmıştır. Her bir veri kümesi, farklı problem türlerini temsil etmekte olup projede, 10-shot öğrenme yöntemleriyle değerlendirilmiştir.

Bu çalışmada, cosmos-DPO modeli kullanılmıştır [1]. Bu model, Türkçe dilinde üstün performans sağlayan bir büyük dil modeli olarak, CNL sürecinin etkin bir şekilde uygulanmasını desteklemiştir.

II. Veri Kümesi

Bu projede kullanılan veri kümeleri; ARC, Hellaswag ve GSM8K olmak üzere üç ana başlıkta incelenmiştir. Her veri kümesi, farklı problem türlerini ve zorluk seviyelerini temsil etmektedir. Tüm veri kümelerinin Türkçe versiyonları kullanılmıştır.

A. ARC-TR

Bu veri kümesi, bilimsel sorgulama ve mantık yürütme becerilerini ölçmeye yönelik hazırlanmıştır. İçerdiği sorular, genellikle bilimsel bilgiler gerektirir ve çoğu zaman çıkarımsal düşünmeyi zorunlu kılar. Bu veri kümesi, CNL yöntemlerinin mantık tabanlı sorulara nasıl yanıt verdiğini test etmek için idealdir. Veri kümesinin ARC-Challenge ve ARC-Easy olmak üzere toplamda 7,787 adet örnek içeren iki alt seti mevcuttur. Çalışmada ARC-Easy bölümünün test için kullanılan Türkçe dilindeki çeviri veri seti kullanılmıştır [2]. Tablo I'de ARC-TR veri kümesinden örnekler yer almaktadır.

B. Hellaswag-TR

Bu veri kümesi, dil modelinin bağlamsal anlam çıkarmadaki başarısını ölçmek için geliştirilmiştir. Veri kümesi, bir bağlam cümlesini tamamlamak için en uygun seçeneği seçmeyi gerektirir. Bu özellik, dil modelinin bağlamsal bütünlüğü koruma ve mantıksal devamlılık sağlama yeteneklerini değerlendirmek için kullanılmıştır. Veri kümesinde 59,944 adet örnek bulunmaktadır.

C. GSM8K-TR

Özellikle matematiksel problem çözme ve mantıksal akıl yürütme becerilerini test etmek amacıyla tasarlanmış 8,790 soru ve cevap çiftinden oluşmaktadır. Bu veri kümesi, bir dil modelinin karmaşık matematiksel soruları anlama ve çözme kapasitesini değerlendirmek için önemli bir araçtır [3]. Cevap olarak her bir sorunun açıklamalı bir şekilde tüm işlem adımları detaylıca ifade edildikten sonra nihai cevap verilmiştir.

Bu veri kümeleri, projede 10-shot öğrenme yöntemleriyle değerlendirilmiş ve her bir veri kümesindeki başarım, CNL yöntemlerinin farklı problem türlerindeki etkinliğini ölçmek

TABLO I. ARC-TR VERİ KÜMESİ

question	choices	answerKey
Periyodik tabloda hangi grup en az reaktif olan elementleri içerir?	{"label": ["A", "B", "C", "D"], "text": ["Grup 1 (1A)", "Grup 3 (3B)", "Grup 16 (6A)", "Grup 18 (8A)"]}	D
Hangi ifade bilimsel bir keşfin olumlu etkisini tanımlar?	{"label": ["A", "B", "C", "D"], "text": ["Bu bazı insanları üzer.", "Yararlı olması uzun zaman alır.", "İşlerin nasıl yürüdüğünü açıklamaya yardımcı olur.", "İşin daha zor olmasına neden olur."]}	C
Aşağıdakilerden hangisi televizyon yapımında kullanılan doğal kaynakları en iyi şekilde korur?	{"label": ["A", "B", "C", "D"], "text": ["bozuk televizyonu onarmak", "indirimde olan bir televizyon satın almak", "eski televizyonları çöpe atmak", "okula yeni bir televizyon bağışlamak"]}	A

TABLO II. HELLASWAG-TR VERİ KÜMESİ

ctx	endings	label
Bir grup adam karanlık bir kulübün içindedir. bir adam	["masa tenisi oyunuyla meşgul.", "ekranın dışındaki bir şeye dart atmaya başlar.", "bir direğe bağlı ipin ucunda takla atar.", "bir kız tarafından öpülür ve baş aşağı yere düşer."]	1
Bir adam karda duruyor. O	["büyük bir kürek tutuyor.", "ağaçlarla dolu bir alandadır ve başka bir adam tarafından tutulmaktadır.", "bir kepçe alır ve kürekle dışarı atar.", "bir makineye bağlı bir ipe tutunuyor."]	0
Siyah şapkalı bir adam sokakta duruyor. Bir kişi	["buzdolabının üstünde oturup bira içiyor.", "oyun koyma oyunu oynuyor.", "bir sopayla pinataya vurur.", "sokakta dans etmeye başlar."]	3

için kullanılmıştır. Veri kümelerine aşağıdaki ön işlemler uygulanmıştır.

ARC-TR: Soruya ilişkin şıklar soru ile birleştirilmiş ve "question" sütunu oluşturulmuştur. Cevaba ilişkin şıkkı ifade eden "answerKey" sütunu korunarak, cevabın içeriği "answer" olarak yeni bir sütuna eklenmiştir.

Hellaswag-TR: Bu veri kümesi soru-cevap içermemektedir. Ancak, çalışmada ortak isimlendirme standartı oluşturmak üzere "ctx" sütunu, "question" olarak belirlenmiştir. Cümle tamamlama için kullanılan "endings" sütun içeriğindeki seçenekler ARC-TR veri kümesinde yapıldığı gibi soru ile birleştirilmiştir. Cevabın indisini belirten "label" sütunu doğru şıkkın harfini belirtecek şekilde düzenlenmiş ve "answerKey" sütununda saklanmıştır. Bağlamdaki cümlenin devamını tamamlayan doğru cümle ise "answer" sütununda saklanmıştır.

TABLO III. GSM8K-TR VERİ KÜMESİ

question	answer
Çiftçi Brown'ın çiftliğinde hepsi tavuk veya inek olmak üzere 20 hayvan var. Toplamda 70 bacağı var. Hayvanların kaçı tavuktur?	Tavukların sayısı C olsun. 20-C'lik inekler var. İneklerin 4*(20-C) bacakları vardır. İneklerin 2C bacakları vardır. Toplam bacak sayısı 2C+4(20-C)=70'dir. 2C+80-4C=70 2C=10 C=«5=5»5 #### 5
Indras'ın adında 6 harf var. Kız kardeşinin adında Indras'ın adındaki harflerin yarısından 4 harf daha fazla var. İndras ve kız kardeşinin adlarında kaç harf var?	ben= «6=6»6 Kardeş= 6/2 + 4= «6/2+4=7»7 6 + 7= «6+7=13»13 harf Indras ve kız kardeşinin adlarında 13 harf var. #### 13
Shiela'nın 15 sayfalık bir araştırma makalesi sunması gerekiyor. Makalenin 1/3'ünü yazmayı çoktan bitirdi. Yazacak kaç sayfası kaldı?	Shiela, 15 sayfa x 1/3= «15*1/3=5»5 sayfa olan makalenin üçte birini zaten yazdı. Yani yine de 15 sayfa - 5 sayfa= «15-5=10»10 sayfa yazması gerekiyor. #### 10

GSM8K-TR: Matematiksel problemin çözümüne yönelik işlem adımlarını detaylıca ifade ettikten sonra nihai cevabı içeren "answer" sütunu, "####" belirtecinden önceki işlem adımları kullanılmamış olup sadece nihai cevap saklanmıştır.

Bağlamı oluşturmak üzere her bir veri kümesinde 10 örnek, cevapları tahmin etmek üzere ise seçilen 10 örneği içermeyen 400'er adet örnek kullanılacak şekilde ayrılmıştır. Veri ayırma işlemi için soru uzunlukları 250 karakteri aşmayacak şekilde seçim yapılmıştır. Kullanılacak büyük dil modeli için 256 token uzunluğu belirlendiği için uzun sorular sürece dahil edilmemiştir.

III. DENEYSEL ANALİZ

Çalışmada, ytu-ce-cosmos Turkish-Llama-8b-DPO-v0.1 modeli kullanılarak ön işlemlerin uygulandığı veri setleri üzerinde 10-shot sonuçları elde edilmiştir. Aynı 10 örneğe ilişkin soru-cevap çiftini içeren bağlamla birlikte, tüm sorular için cevap tahminleri elde edilmiştir. Her bilinmeyen yeni soru için bağlamı oluşturan 10 adet soru-cevap çifti modele girdi olarak verilerek yeni bir soru sorulmuştur. ARC ve Hellaswag veri setleri için bilinmeyen soru ve şıklar verildikten sonra "*Sadece yukarıdaki Şıklar arasından doğru şık harfini belirt.*", GSM8K veri seti içinse "*Lütfen açıklama yapmayın, cevabınız sadece sayı içersin.*" şeklinde özel bir istemde bulunulmuştur.

Cosmo DPO modelini değerlendirmek üzere LM Evaluation Harness [4] kullanılarak üç veri kümesinde de 10 shot sonuçlara yönelik performans ölçümü yapılmıştır. Bu ölçümler için aracın kendi üzerinde tanımlı arc_easy, hellaswag ve gsm8k görevlere ait yapılandırma dosyalarında ön işlemlerin yapıldığı yerel veri setinin kullanımına ilişkin ayarlar düzeltilmiştir. Ancak ölçümlere ait çıktılar incelendiğinde, değerlendirmenin Hugging Face üzerindeki orijinal ve İngilizce metinlerden oluşan veri setleri üzerinden yapıldığı görülmüştür. Bu ölçümlere ait sonuçlar Tablo IV'te yer almaktadır.

Bunun üzerine, performans değerlendirmesi için çalışmada değişikliğe gidilmiştir. Modelin ürettiği tahmini cevap üzerinden sadece tahmini şık alınarak gerçek şık ile karşılaştırılmıştır. Bunun nedeni, bilinmeyen yeni soruları modele tahmin

TABLO IV. ORİJİNAL VERİ KÜMELERİNE AİT DEĞERLENDİRME SONUÇLARI

Tasks	Version	Filter	nshot	Metric	Value	Stderr
arc_easy	1	none	10	acc	0.8657	0.0070
		none	10	acc_norm	0.8678	0.0069
hellaswag	1	none	10	acc	0.6038	0.0049
		none	10	acc_norm	0.8107	0.0039
gsm8k	3	flexible-extract	10	exact_match	0.7854	0.0113
		strict-match	10	exact_match	0.7839	0.0113

ettirebilmek üzere ilave olarak kullanılan özel isteme rağmen model, tahminlerinde sadece doğru şık harfini belirtmeyerek ilave bazı açıklamalarda da bulunmasıdır. Bu durumun başarı ölçümünde sorun oluşturmaması için üretilen cevapta yalnızca tahmini şıkkı içerecek şekilde ön işlem uygulanmıştır. Bunun sonucunda ARC-TR, Hellaswag-TR ve GSM8K-TR veri kümelerinin tümü için oluşan tahminler Tablo V'te gösterilmiştir.

TABLO V. MODEL TAHMİNLERİ

dataset	question	true_ans	pred_ans
ARC-TR	Periyodik tabloda hangi grup en az reaktif olan elementleri içerir? Şıklar: A. Grup 1 (1A) B. Grup 3 (3B) C. Grup 16 (6A) D. Grup 18 (8A)	D	D
	Hangi ifade bilimsel bir keşfin olumlu etkisini tanımlar? Şıklar: A. Bu bazı insanları üzer. B. Yararlı olması uzun zaman alır. C. İşlerin nasıl yürüdüğünü açıklamaya yardımcı olur. D. İşin daha zor olmasına neden olur.	C	B
Hellaswag-TR	Kırmızı gömlekli bir adam ellerini çırpıyor. Bir pistte koşuyor ve kuma atlıyor. insanlar Şıklar: A. Stadyumda durup onu izliyorlar. B. bekleme odasında oturup izliyorlar. C. daha sonra kumu tırmıklayın. D. onun koşmasını izliyoruz.	C	D
	İnsanlar spor salonunda bisiklete biniyorlar. iki kadın Şıklar: A. Bisikletler ahşap namlulu bir sahada sıralanmıştır. B. tulumlarla spor salonundayız. C. arka tekerlekte vagon. D. kondisyon bisikletleri üzerinde çalışıyorlar.	D	D
GSM8K-TR	Çiftçi Brown'un çiftliğinde ya tavuk ya da inek olan toplam 20 hayvanı var.	5	15
	Bir sepetin içinde 1'i bozuk, %20'si olgunlaşmamış, 2'si ekşi ve geri kalanı iyi durumda olan 25 portakal bulunmaktadır. Kaç portakal iyidir?	17	17

Performans ölçümünün tahmini şık ile gerçek şık ile karşılaştırılarak manuel olarak yapıldığındaki sonuçlar Tablo VI'da gösterilmiştir.

Bağlam için rastgele belirlenen ve 10-shot sonuçlarını elde etmek üzere kullanılan soru-cevap çiftleri üzerinde aşağıdaki yöntemler kullanılarak her bir veri kümesinde ensemble boyutu 5 olarak belirlenerek her bir soru için 5 farklı tahmini cevap

TABLO VI. 10-SHOT SONUÇLARI

Dataset	Accuracy	F1 Score
ARC-TR	0.3475	0.3475
Hellaswag-TR	0.2950	0.2950
GSM8K-TR	0.1125	0.1125

TABLO VII. TEKİL VE ENSEMBLE BAŞARILARI

Veri Kümesi	Method	Tekil Başarı	Ensemble Başarısı
ARC-TR	10'lu seçim	0.3225	0.435
		0.3625	
		0.41	
		0.2925	
		0.4375	
	5'li seçim	0.3975	**0.525**
		0.4725	
		0.47	
		0.46	
		0.4075	
	Yer değiştirme	0.4325	0.4975
		0.43	
		0.4225	
		0.3775	
		0.435	

oluşturulmuştur.

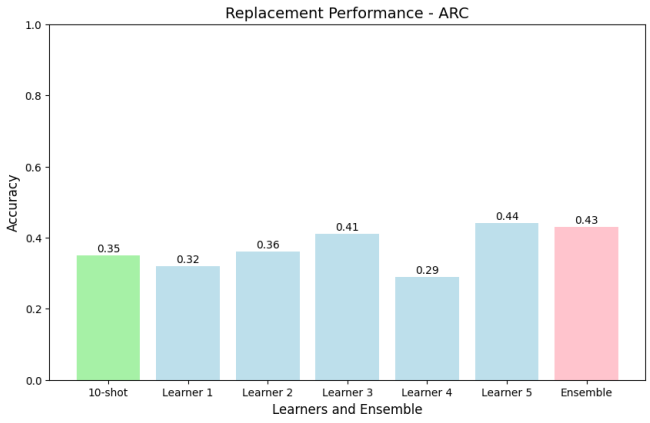
- 10'lu seçim: 10-shot sonuçlarını elde etmek üzere rastgele seçilen 10 adet bağlam örneği üzerinden rastgele ve yerine koyarak seçim yapılmıştır. Bu yöntem ile belirlenen 10 soru-cevap çifti arasında aynı örneğin birden fazla kez yer alma durumu mevcuttur. Her defasında aynı 400 soruya, belirlenen farklı 10 adet soru-cevap çifti içeren örnekler bağlam olarak verilerek model tarafından soruların cevapları tahmin edilmiştir.
- 5'li seçim: Sabit olarak tanımlanan 10 soru-cevap çifti içinden, tekrar etmeyecek şekilde 5 adet örnek bağlam olarak belirlenmiştir. Belirlenen farklı 5 adet soru-cevap çifti ile her defasında aynı 400 soruya model tarafından cevap verilmiştir.
- Yer değiştirme: Sabit olarak tanımlanan 10 soru-cevap çifti üzerinden yalnızca örneklerin sıraları karıştırılarak sonuca etkisi araştırılmıştır. kendi arasında içinden, tekrar etmeyecek şekilde 5 adet örnek bağlam olarak belirlenmiştir. Sıraları değişen farklı 10 adet soru-cevap çifti ile her defasında aynı 400 soruya model tarafından cevap verilmiştir.

Üç farklı yöntemle farklı bağlam örnekleriyle oluşturulan tahminler demokrasi usulüne göre birleştirilmiştir. ARC veri kümesinin 5 farklı tekil model ile ensemble başarısına ait metrikler Tablo VII'de Hellaswag veri kümesinin tekil modelleri ile ensemble başarısına ait sonuçlar Tablo VIII'de gösterilmiştir.

TABLO VIII. TEKİL VE ENSEMBLE BAŞARILARI

Veri Kümesi	Method	Tekil Başarı	Ensemble Başarısı
Hellaswag-TR	10'lu seçim	0.265	**0.275**
		0.27	
		0.215	
		0.2425	
		0.2525	
	5'li seçim	**0.2525**	0.23
		0.23	
		0.25	
		0.24	
		0.23	
	Yer değiştirme	**0.2625**	0.245
		0.205	
		0.255	
		0.2375	
		0.245	

Tablodaki veriler aşağıda grafik halinde gösterilerek veri setleri üzerinde 3 farklı yöntemin uygulanmasıyla elde edilen tekil ve ensemble başarı sonuçları gösterilmiştir (2).

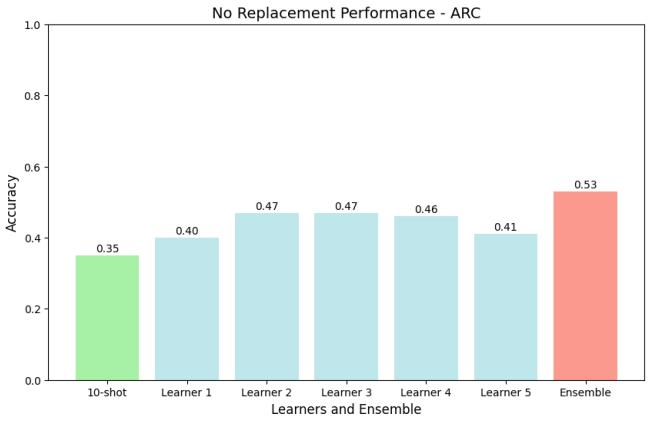


Şekil 1. ARC veri kümesinde 10'lu seçim sonuçları

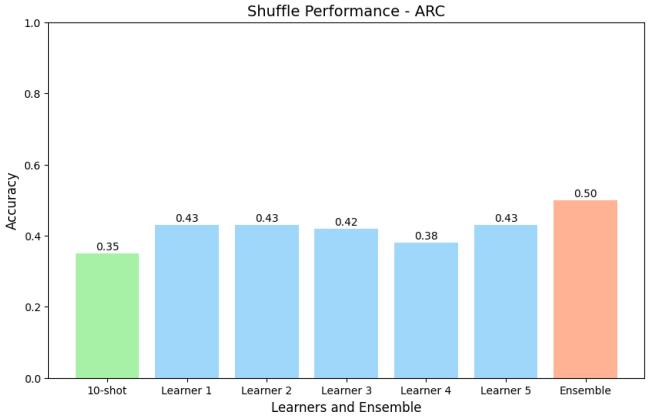
IV. SONUÇ

Bu çalışmada, bağlamda öğrenme yöntemi ytu-ce-cosmos/Turkish-Llama-8b-DPO-v0.1 modeli kullanılarak ARC, Hellaswag ve GSM8K veri kümeleri üzerinde performans ölçümleri yapılmıştır. Üç veri kümesi üzerinde 10-shot sonuçlar elde ediltikten sonra ARC ve Hellaswag veri kümeleri için ensemble yöntemleri de uygulanmış ve bireysel model performansını bir miktar arttırmıştır.

Farklı veri kümelerindeki sonuçlar, yöntemlerin çeşitliliğinin ve ensemble yaklaşımlarının güçlü yanlarını ortaya koymuştur. ARC veri kümesinde mantıksal sorulara verilen cevapların doğruluğu artırılmış, Hellaswag veri kümesinde bağlamsal bütünlük korunmuştur. GSM8K veri kümesinde ise bireysel model matematiksel problem çözme becerilerini test etmek için kullanılmıştır.



Şekil 2. ARC veri kümesinde 5'li seçim sonuçları



Şekil 3. ARC veri kümesinde yer değiştirme seçim sonuçları

KAYNAKLAR

[1] Kesgin, H., Yuce, M., Dogan, E., Uzun, M., Uz, A., İnce, E., Erdem, Y., Shbib, O., Zeer, A. & Amasyali, M. Optimizing Large Language Models for Turkish: New Methodologies in Corpus Selection and Training. *2024 Innovations In Intelligent Systems And Applications Conference (ASYU)*. pp. 1-6 (2024)

[2] Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C. & Tafjord, O. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *ArXiv:1803.05457v1*. (2018)

[3] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C. & Schulman, J. Training Verifiers to Solve Math Word Problems. *ArXiv Preprint ArXiv:2110.14168*. (2021)

[4] Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K. & Zou, A. A framework for few-shot language model evaluation. (Zenodo,2024,7), https://zenodo.org/records/12608602