Ekonomiye eLaLeM ne der: Küçük Ölçekli Alan Uyarlamasında Akademik Görevler için QLoRA ve PEFT ile LLaMA-2 Dil Modelinin Ekonomi Metinlerinde İnce Ayarlaması

Bu çalışmada, birçok farklı dil görevine uygulanabilirliği ile bilinen açık kaynaklı LLaMA-2-7B (Katmanlı Dil Modeli Mimarisi) [1] isimli büyük dil modeli, QLoRA (Kuantize Düşük Rütbe Uyarlaması) [2] ile PEFT (Parametre Verimli İnce Ayar) yöntemleri kullanılarak ekonomi alanında küçük ölçekli alan uyarlamasına yönelik ince ayar yapılmıştır. Bu anlamda LLaMA-2-Ekon isminde ince ayarlanmış model üç farklı görev tipi için İngilizce dilindeki ekonomi makaleleriyle eğitilmiştir: (i) bir özete dayalı başlık oluşturma, (ii) özet sınıflandırma ve (iii) ekonomi alanında soru-cevap (S&C). LLaMA-2-Ekon dil modelimiz daha fazla ekonomi alanında araştırma yapanların yararlanabilmesi için Streamlit ve Langchain kütüphaneleri kullanılarak bir internet uygulaması haline getirilmiştir.

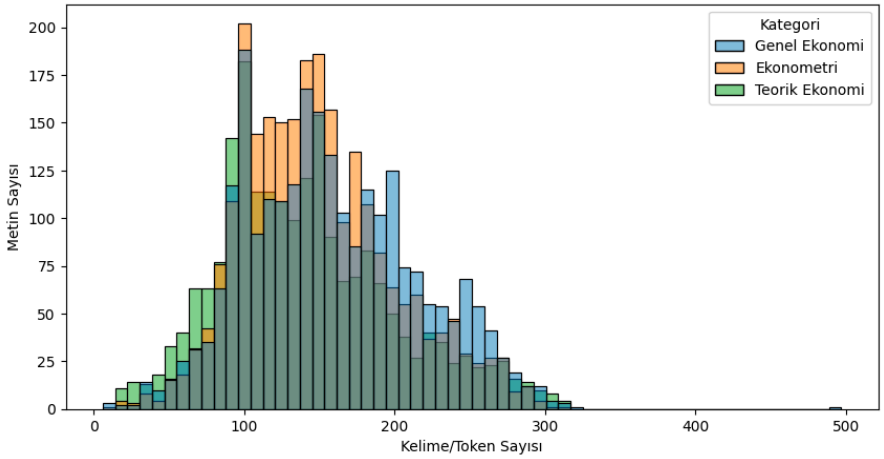
RNN [3] ve LSTM [4] mimarileri gibi sinir ağlarının ortaya çıkışını takip eden dönüştürücü mimarileri [5], GPT-4, Bard ve Gemini gibi kapalı modeller ile LLaMA-2, Bloom ve Mistral gibi açık kaynaklı büyük dil modellerinin gelişimine yol açmıştır. Ayrıca, QLoRA ve PEFT gibi yeni yöntemler daha az bellek gerektiren ve daha küçük donanımlarda daha hızlı eğitim süreleri sağlayıp parametre sayısını önemli ölçüde azaltarak modellerin ölçeklenebilmesi ve görev bazlı ince ayarlanmasına imkân sağlamıştır. Bu şekilde ince ayarlanmış modeller çeşitli görevlere uyarlanmıştır [6]–[9]. Ancak bilgimiz dahilinde ekonomi alanında çeşitli akademik görevlere yönelik ince ayarlanmış açık kaynaklı bir model mevcut değildir.

Çalışma kapsamında toplamda arXiv veri tabanından 6719 örnek eğitim seti ve 350 örnek test seti için toplanmıştır. Ekonometri, genel ekonomi ve teorik ekonomi kategorilerinde ekonomi makalelerinin başlık, özet ve diğer meta verileri çekilmiş olup Şekil 1'de verinin kategorilere göre dağılımı paylaşılmıştır. LLaMA-2-Ekon modelimiz ekonomi makale özetleri ve bu özetlere dayalı başka bir büyük dil modeliyle hazırlanmış sentetik soru-cevap diyalogları üzerinde [10] Gözetimli İnce Ayarlama (Supervised Fine-Tuning) yöntemi ile etiketlerine göre Python'da A100 GPU kullanılarak ince ayarlanmıştır.

Başlık oluşturma görevi için, LLaMA-2-Ekon modelimizi, temel (baseline) ince ayarsız LLaMA-2 modeli ve tek atış öğrenme (one shot) metodunun kullanan GPT-Neo ve OPT gibi diğer benzer kod-çözücü (decoder-only) modellerle ROUGE ve BLEU değerlendirme ölçütlerinde karşılaştırdık. Tablo 1'deki sonuçlar ince ayarlanmış LLaMA-2-Ekon modelinin daha başarılı olduğunu göstermektedir. Özet sınıflandırma görevi için ise LLaMA-2-Ekon farklı makine ve derin öğrenme algoritmalarıyla karşılaştırılmıştır. Tablo 2'de modelimizin Lojistik Regresyon, K En Yakın Komşu Sınıflandırma, Rastgele Orman Sınıflandırıcı, XGB Sınıflandırıcı, Karar Ağacı Sınıflandırıcı ve SVC (Destek Vektör Sınıflandırıcıları) dahil farklı makine öğrenimi modellerine ve LSTM ile RNN gibi derin öğrenme algoritmalarına kıyasla daha iyi performans gösterdiği görülmektedir.

Son olarak, modelin araştırmacılar tarafından kullanılabilmesi için Streamlit kütüphanesi ile bir internet uygulaması hazırlanmıştır. Bu kapsamda, RAG (geri alma artırımlı üretim) ve FAISS (Facebook AI Benzerlik Araması) yöntemleri kullanılarak yukarıdaki görevlere ek olarak kullanıcıların uygulamada kendi belgelerini yüklemelerine ve ekonomi alanında makalelerle etkileşime geçmesi sağlanacaktır.

Şekil 1. Eğitim verilerindeki kelime/token sayısı ve kategorik dağılımı



Tablo 1. LLaMA-2-Econ modelinin başlık oluşturma görevinde diğer büyük dil modelleriyle karşılaştırmalı performans ölçümleri

Model	**BLEU**	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
LLaMA-2 (temel)	0.09	0.31	0.14	0.28
LLaMA-2-Ekon	**0.12**	**0.40**	**0.19**	**0.37**
GPT Neo (tek atış öğrenme)	0.03	0.19	0.05	0.17
OPT (tek atış öğrenme)	0.06	0.25	0.10	0.22

Tablo 2. LLaMA-2-Econ modelinin özet sınıflandırma görevinde makine ve derin öğrenme algoritmalarıyla karşılaştırmalı performans ölçümleri

Model	**Doğruluk**	**Kesinlik**	**Duyarlılık**	**F1**
Karar Ağacı Sınıflandırıcı	0.7714	0.7191	0.7714	0.7139
K En Yakın Komşu Sınıflandırıcı	0.7857	0.7879	0.7857	0.7881
LLaMA-2-Ekon	**0.8800**	**0.8822**	**0.8800**	**0.8801**
Lojistik Regresyon	0.8543	0.8553	0.8543	0.8547
Rastgele Orman Sınıflandırıcı	0.8543	0.8577	0.8543	0.8545
SVC	0.8571	0.8593	0.8571	0.8578
XGB Sınıflandırıcı	0.8271	0.8277	0.8286	0.8276
RNN	0.8057	0.8098	0.8029	0.8077
LSTM	0.8143	0.8166	0.8143	0.8152

Referanslar: [1] H. Touvron et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models." arXiv, Jul. 19, 2023. Accessed: Jan. 14, 2024. [Online]. Available: http://arxiv.org/abs/2307.09288. [2] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs." arXiv, May 23, 2023 Accessed: Jan. 14, 2024. [Online]. Available: http://arxiv.org/abs/2305.14314 [3] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," nature, vol. 323, no. 6088, pp. 533–536, 1986, Accessed: Jan. 14, 2024. [Online]. Available: https://www.nature.com/articles/323533a0 [4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997, Accessed: Jan. 14, 2024. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6795963/ [5] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017, Accessed: Jan. 14, 2024. [Online]. Available: https://proceedings.neurips.cc/paper/7181-attention-is-all [6] D. Gavrilov, P. Kalaidin, and V. Malykh, "Self-attentive Model for Headline Generation," in Advances in Information Retrieval, vol. 11438, L. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff, and D. Hiemstra, Eds., in Lecture Notes in Computer Science, vol. 11438. , Cham: Springer International Publishing, 2019, pp. 87–93. doi: 10.1007/978-3-030-15719-7_11. [7] L. Loukas, I. Stogiannidis, P. Malakasiotis, and S. Vassos, "Breaking the Bank with ChatGPT: Few-Shot Text Classification for Finance." arXiv, Aug. 28, 2023. Accessed: Jan. 14, 2024. [Online]. Available: http://arxiv.org/abs/2308.14634 [8] T. T. Nguyen, C. Wilson, and J. Dalins, "Fine-Tuning Llama 2 Large Language Models for Detecting Online Sexual Predatory Chats and Abusive Texts." arXiv, Aug. 28, 2023. Accessed: Jan. 14, 2024. [Online]. Available: http://arxiv.org/abs/2308.14683 [9] C. Wu, W. Lin, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "PMC-LLaMA: Towards Building Open-source Language Models for Medicine." arXiv, Aug. 25, 2023. Accessed: Jan. 14, 2024. [Online]. Available: http://arxiv.org/abs/2304.14454 [10] A. Chernyavskiy, M. Bregeda, and M. Nikiforova, "PaperPersiChat: Scientific Paper Discussion Chatbot using Transformers and Discourse Flow Management," in Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, 2023, pp. 584–587. Accessed: Jan. 14, 2024. [Online]. Available: https://aclanthology.org/2023.sigdial-1.54/