Proje Adı: Güncel bilgiye uyumlu, akıl yürütebilen, kapsamlı Türkçe dil modeli geliştirme

Özet: Bu projenin amacı, Türkçe dilinde akıl yürütebilen, güncel bilgiyle uyum sağlayabilen ve çok çeşitli görevlerde yüksek performans gösterebilen kapsamlı bir büyük dil modeli (LLM) geliştirmektir. Büyük dil modelleri, metin çevirisi, özetleme, duygu analizi ve soru-cevap sistemleri gibi pek çok doğal dil işleme (NLP) görevlerinde yaygın olarak kullanılır. Bu modeller ayrıca sohbet robotları, dijital asistanlar, eğitim ve programlama desteği, alana özel dokümantasyon ve finansal raporlama gibi birçok alanda etkili şekilde uygulanmaktadır. Bu çok yönlü kullanım, büyük dil modellerini hem akademik hem endüstriyel bağlamda güçlü bir araç haline getirmiştir. Bu proje ile, Türkçe'nin yapısal ve anlamsal özelliklerini derinlemesine öğrenebilen, çok sayıda görevde yüksek performans gösterebilen ve güncel bilgiye dayalı akıl yürütme yeteneğine sahip bir büyük dil modeli geliştirilecektir. Model, yalnızca dilsel yeterliliğe değil; aynı zamanda Türkçe'nin sosyo-kültürel bağlamını dikkate alan bir anlayışla tasarlanacaktır. Bu sayede hem genel amaçlı doğal dil işleme görevlerinde (örneğin özetleme, çeviri, metin üretimi, soru-cevap) hem de alan-özgü uygulamalarda (örneğin sağlık, hukuk, eğitim, kamu yönetimi) kullanılabilecek esnek ve güvenilir bir altyapı oluşturulacaktır. Modelin eğitimi için hem modern hem de tarihi Türkçe metinleri kapsayan, dengeli ve yüksek kaliteli bir veri kümesi oluşturulacak ve model eğitimin ardından kapsamlı bir şekilde değerlendirilecektir. Ayrıca modelin performansı, yalnızca İngilizce tabanlı ya da çok dilli LLM'lerle değil, Türkçe için geliştirilmiş mevcut modellerle de karşılaştırılarak ortaya konacaktır. Bu proje, Türkçe'nin dijital çağda daha güçlü ve bağımsız bir şekilde temsil edilmesini sağlamayı, aynı zamanda yerli yapay zeka ekosistemine katkıda bulunarak ulusal teknoloji kapasitesini artırmayı hedeflemektedir. Geliştirilecek Türkçe büyük dil modeli; kamu hizmetleri, hukuk, sağlık, finans ve eğitim gibi alanlarda doğrudan ticarileşebilir çözümler üretme potansiyeline sahiptir. Aynı zamanda kültürel bağlamı bilen bir model, müşteri destek sistemlerinden akıllı asistanlara kadar, kullanıcı memnuniyetini ve kabul edilebilirliği artırarak doğrudan ekonomik fayda sağlayacaktır. Ayrıca Türkçe büyük dil modeli geliştirilmesi yerelleşmiş yapay zekâ ekosisteminin oluşumuna katkı sunarak Türkiye'nin teknoloji bağımsızlığı ve rekabet gücü açısından stratejik bir avantaj yaratacaktır.

Kazanım ve Sonuçlar: Bu proje sonunda elde edilmesi beklenen en önemli kazanımlardan biri, Türkçe'nin morfolojik yapısına daha uygun ve yüksek verimlilik sağlayan bir tokenizer geliştirilmesidir. Türkçe için özel olarak tasarlanacak bu sözcükleme yönteminin, kelime başına düşen alt birim sayısını azaltarak hem eğitim verimliliğini hem de model performansını artırması beklenmektedir. Tokenizer değerlendirmeleri sonucunda yalnızca intrinsik metriklerde değil, aynı zamanda adlandırılmış varlık tanıma, metin sınıflandırma ve özetleme gibi farklı görevlerde de anlamlı performans artışları sağlanacaktır. Bu sayede, Türkçe dil modelleri için altyapısal bir

iyileştirme elde edilmiş olacaktır. İkinci olarak, damıtma (knowledge distillation) tabanlı yöntemlerle geliştirilecek Türkçe odaklı modeller, sınırlı GPU kaynaklarıyla dahi yüksek kaliteli sonuçlar üretebilecek kapasiteye ulaşacaktır. Qwen3-4B, DeepSeek-R1-Distill-Llama-8B ve GPT-OSS-20B (MoE) modelleri ile yapılacak üç farklı koşu sonucunda, Türkçe'nin farklı alanlarda (hukuk, fen bilimleri, beşeri bilimler vb.) performansının ölçülmesi sağlanacaktır. Bu modeller, hem TÜBİTAK'ın kapalı benchmark setlerinde hem de Cetvel, OpenLLM Turkish Leaderboard ve TARA gibi açık kaynaklı değerlendirmelerde test edilerek, mevcut ulusal ve uluslararası standartlara göre kapsamlı biçimde raporlanacaktır. Buradan elde edilecek sonuç, Türkçe'nin çok-dilli modeller içinde temsil gücünü artırmak ve bağımsız Türkçe odaklı modellerin potansiyelini ortaya koymak olacaktır. Üçüncü büyük çıktısı, RAG tabanlı mimarinin Türkçe büyük dil modellerine entegrasyonuyla ilgilidir. Bu yöntem, modelin yalnızca parametrelerine gömülü bilgiyle sınırlı kalmayıp, harici bilgi kaynaklarından güncel veriye erişebilmesini sağlayacaktır. Böylece, özellikle hızlı değişen bilgi alanlarında güncellik, doğruluk ve güvenilirlik artacaktır. Bu kapsamda geliştirilecek RAG sistemlerinin Türkçe'nin eklemeli yapısına uyarlanması, hem geri getirme (retrieval) hem de üretim (generation) bileşenlerinde önemli kazanımlar sağlayacaktır. Beklenen sonuç, halüsinasyon oranlarının azalması, doğruluk oranlarının yükselmesi ve gerçek hayatta kullanılabilirliğin artmasıdır. Dördüncü kazanım alanı, tarihî Türkçe metinlerin işlenmesinde kaydedilecek ilerlemelerdir. Tarihî varyantların modern Türkçe dil modellerine entegrasyonu sayesinde, kültürel mirasın dijital ortamda daha etkin biçimde işlenmesi ve analiz edilmesi mümkün hale gelecektir. Bu yaklaşım yalnızca doğal dil işleme alanında yeni bir katkı sunmakla kalmayacak, aynı zamanda tarih, dilbilim ve kültürel çalışmalar açısından da güçlü bir araştırma altyapısı sağlayacaktır. Böylece, modern Türkçe için geliştirilen modellerin tarihî metinlere uygulanabilirliği ciddi biçimde artırılacaktır. Sonuç olarak, bu proje yalnızca Türkçe dil modellerinin teknik kapasitesini artırmayı değil, aynı zamanda ulusal araştırma ekosistemine sürdürülebilir bir katkı sağlamayı hedeflemektedir. Geliştirilecek tokenizer, damıtma tabanlı modeller, RAG entegrasyonu ve tarihî Türkçe uyarlamaları; hem akademik literatüre hem de pratik uygulamalara doğrudan katkı sunacaktır. Böylece, Türkçe için bugüne kadar eksik kalan büyük ölçekli dil modeli altyapısının güçlendirilmesi, ulusal stratejik hedefler açısından da önemli bir adım olacaktır.

Detaylar: Bu proje Prof. Suayb S. Arslan (suayb.arslan@bogazici.edu.tr), Prof. Taha Koç Yiğit ve Prof. Betül Özateş ortaklığında yürütülmektedir.

Destek: TÜBİTAK BİLGEM

Destek Miktarı: Bogazici Universitesi (>6M TL), Toplam (>10M TL)

Destek Süresi: 2026-2027

ENGLISH

Project Title: Development of a Comprehensive Turkish Language Model Capable of Reasoning and Adapting to Up-to-Date Knowledge

Abstract:
The aim of this project is to develop a comprehensive large language model (LLM in Turkish that is capable of reasoning, adapting to up-to-date knowledge, and achieving high performance across a wide range of tasks. Large language models are widely used in numerous natural language processing (NLP) tasks such as machine translation, summarization, sentiment analysis, and question answering. They are also effectively applied in many domains, including chatbots, digital assistants, educational and programming support, domain-specific documentation, and financial reporting. This versatility has established large language models as powerful tools in both academic and industrial contexts.

Within the scope of this project, a large language model will be developed that can deeply learn the structural and semantic characteristics of Turkish, demonstrate strong performance across multiple tasks, and perform reasoning based on up-to-date information. The model will be designed not only with linguistic proficiency in mind, but also with an understanding that takes into account the socio-cultural context of Turkish. In this way, a flexible and reliable infrastructure will be established that can be used both for general-purpose NLP tasks (e.g., summarization, translation, text generation, question answering) and for domain-specific applications (e.g., healthcare, law, education, public administration).

For model training, a balanced and high-quality dataset covering both modern and historical Turkish texts will be constructed, and the model will be comprehensively evaluated after training. Furthermore, the model's performance will be compared not only with English-based or multilingual LLMs, but also with existing models specifically developed for Turkish. This project aims to enable Turkish to be represented more strongly and independently in the digital age, while also contributing to the domestic artificial intelligence ecosystem and increasing national technological capacity. The Turkish large language model to be developed has the potential to produce directly commercializable solutions in areas such as public services, law, healthcare, finance, and education. At the same time, a model that is aware of cultural context will provide direct economic benefits by increasing user satisfaction and acceptability, from customer support systems to intelligent assistants. Moreover, the development of a Turkish large language model will contribute to the formation of a localized AI ecosystem, creating a strategic advantage for Türkiye in terms of technological independence and competitiveness.

Expected Outcomes and Results:
One of the most significant expected outcomes of this project is the development of a tokenizer

that is better suited to the morphological structure of Turkish and offers higher efficiency. This tokenization method, to be specifically designed for Turkish, is expected to reduce the number of subword units per word, thereby improving both training efficiency and model performance. As a result of tokenizer evaluations, meaningful performance improvements are anticipated not only in intrinsic metrics, but also across downstream tasks such as named entity recognition, text classification, and summarization. In this way, an infrastructural improvement for Turkish language models will be achieved.

Second, Turkish-focused models developed using knowledge distillation–based methods will reach the capacity to produce high-quality results even with limited GPU resources. Through three different training runs conducted with the Qwen3-4B, DeepSeek-R1-Distill-Llama-8B, and GPT-OSS-20B (MoE) models, the performance of Turkish across different domains (e.g., law, natural sciences, humanities) will be measured. These models will be evaluated both on TÜBİTAK's closed benchmark sets and on open-source evaluation platforms such as Cetvel, the OpenLLM Turkish Leaderboard, and TARA, and the results will be comprehensively reported according to current national and international standards. The outcome of this evaluation will be to enhance the representational strength of Turkish within multilingual models and to demonstrate the potential of independent, Turkish-focused models.

The third major output concerns the integration of retrieval-augmented generation (RAG)–based architectures into Turkish large language models. This approach will allow the model not to be limited solely to the knowledge embedded in its parameters, but also to access up-to-date information from external knowledge sources. As a result, improvements in timeliness, accuracy, and reliability will be achieved, particularly in rapidly evolving knowledge domains. Adapting the RAG systems developed within this scope to the agglutinative structure of Turkish is expected to yield significant gains in both the retrieval and generation components. The anticipated outcomes include reduced hallucination rates, increased accuracy, and improved real-world usability.

The fourth area of contribution relates to progress in processing historical Turkish texts. By integrating historical variants into modern Turkish language models, it will become possible to more effectively process and analyze cultural heritage in digital environments. This approach will not only provide a novel contribution to the field of natural language processing, but will also establish a strong research infrastructure for history, linguistics, and cultural studies. Consequently, the applicability of models developed for modern Turkish to historical texts will be substantially enhanced.

In conclusion, this project aims not only to increase the technical capacity of Turkish language models, but also to provide a sustainable contribution to the national research ecosystem. The tokenizer to be developed, the distillation-based models, RAG integration, and adaptations for

historical Turkish will contribute directly to both the academic literature and practical applications. In this way, strengthening the large-scale language model infrastructure that has thus far been lacking for Turkish will represent an important step toward achieving national strategic objectives.

Details:
This project is conducted jointly by Prof. Suayb S. Arslan (suayb.arslan@bogazici.edu.tr), Prof. Taha Koç Yiğit, and Prof. Betül Özateş.

Funding Agency: TÜBİTAK BİLGEM

Funding Amount: Boğaziçi University (> 6 million TRY), Total (> 10 million TRY)

Project Duration: 2026–2027