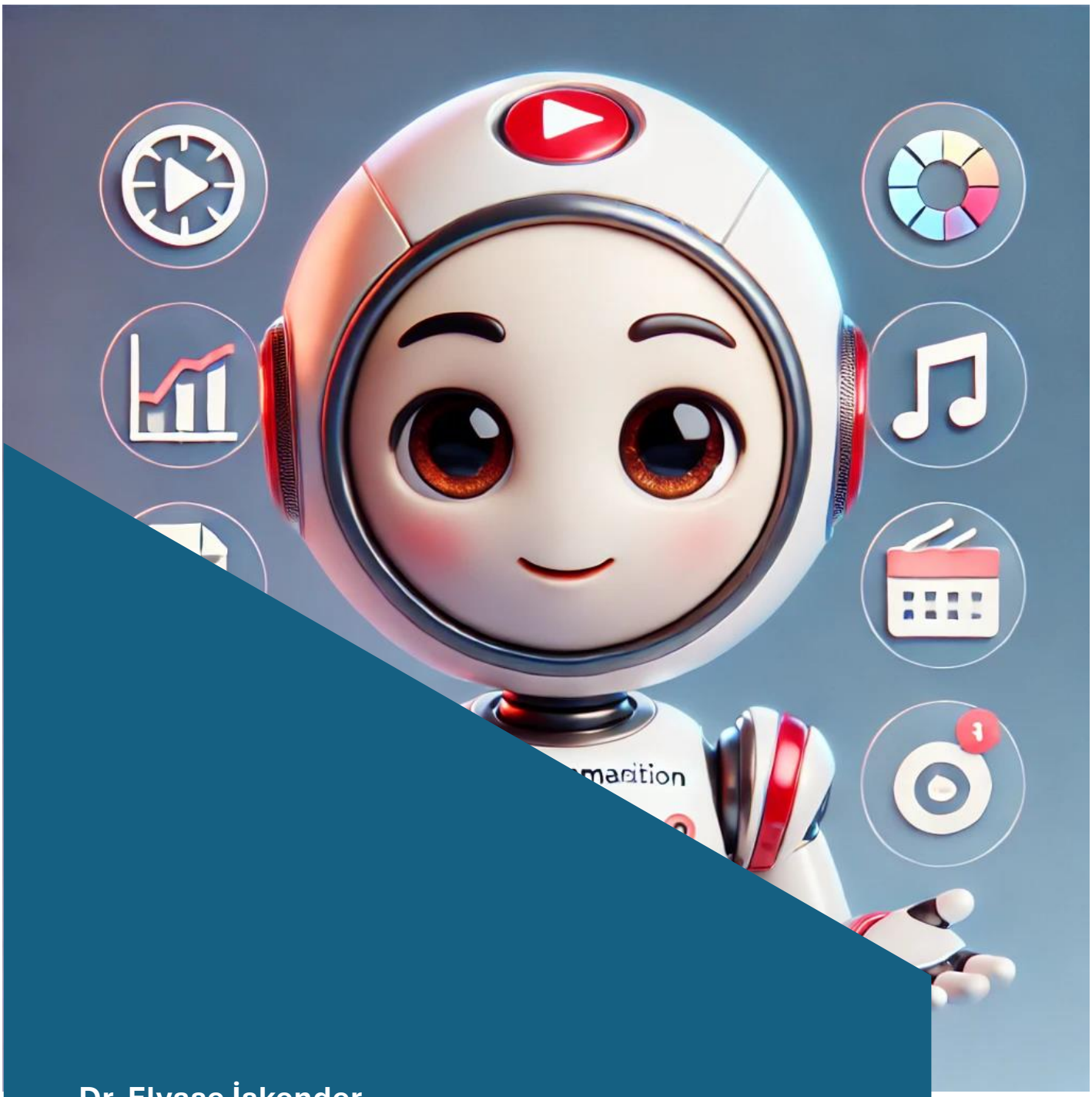
Aimyy Raporu : Yönlendirme / İstem (Prompting)



Dr. Elyase İskender

04.05.2025

Büyük Dil Modelleri (LLM'ler), Prompt Mühendisliği ve İlgili Kavramlar

1. LLM'lerin Temelleri ve Mimarileri

- **LLM'ler Ne İşe Yarar?** LLM'ler, büyük veri kümeleri üzerinde eğitilmiş ve metin tamamlama, çeviri, özetleme ve soru yanıtlama gibi çeşitli doğal dil işleme görevlerini gerçekleştirebilen güçlü araçlardır. "LLM'ler genellikle Transfer Öğrenmeye dayalı modeller kullanır. Bir görevden elde edilen bilgi, başka bir göreve aktarılabilir." (Subramanian Venkataraman)
- **Tokenizasyon:** Metnin LLM'ler tarafından işlenmesi için ilk adım tokenizasyondur. Bu, metni kelimeler, alt kelimeler veya karakterler gibi daha küçük birimlere (tokenlara) ayırma işlemidir. Tokenizasyon, modelin bir dizideki bir sonraki tokenı tahmin etmesini sağlar. Farklı modeller (BERT, GPT-2, T5) farklı tokenizasyon stratejileri kullanabilir. "İngilizce'de tokenlar, tek bir karakterden (t) bir kelimeye (great) kadar değişebilir." (James Phoenix ve Michael Taylor).
- **Alt Kelime Tokenizasyonu:** Bu teknik, kelime tabanlı ve karakter tabanlı tokenizasyonun avantajlarını birleştirir. Daha sık görülen kelimeler benzersiz kimliklerle eşlenirken, daha az sıklıkta görülen kelimeler, anlamlarını koruyan alt kelimelere ayrılır. "Alt kelime tokenizasyonu tipik olarak algoritmalara, istatistiksel kurallara ve önemli bir sezgisel yaklaşıma dayanır: Nadir veya seyrek görünen kelimeleri alt kelimelere ayırın ve sık görülen kelimeleri bölmeyin." (Oswald Campesato)
- **Gömme (Embeddings):** Tokenlar, sayısal vektörlere (gömme) dönüştürülür. Bu vektörler, kelimeler arasındaki anlamsal ilişkileri yakalar. "Embeddings contains encoded values for each chunks created" (Subramanian Venkataraman).
- **Dikkat Mekanizmaları:** Transformer mimarisinin temel bir bileşenidir ve modelin bir dizideki farklı tokenların önemini tartmasını sağlar. Bu, modelin kelimelerin anlamını farklı bağlamlarda (örneğin, "banka" kelimesi) anlamasına yardımcı olur. "Transformatör mimarisindeki dikkat mekanizması, bir önceki bölümdeki üç cümledeki 'banka' kelimesi için farklı bir bağlam vektörü (yani, kayan nokta sayılarından oluşan tek boyutlu bir vektör) üretir." (Oswald Campesato).
- **Çoklu Kafa Dikkat (Multi-Head Attention):** Tek bir dikkat mekanizması yerine birden fazla dikkat matrisi hesaplamayı içerir, bu da modelin farklı anlamsal temsilleri yakalamasını sağlar. "Multi-head attention involves calcu-lating multiple attention matrices instead of a single attention matrix." (Oswald Campesato).
- **Konumsal Kodlamalar (Positional Encodings):** Modelin bir dizideki tokenların konumunu anlamasına yardımcı olmak için gömme vektörlerine eklenir. Sinüs ve kosinüs fonksiyonları kullanılarak hesaplanırlar. "Konumsal kodlamalar,

trigonometrik sinüs() ve kosinüs() fonksiyonları aracılığıyla hesaplanır, ardından her kelime gömmeye eklenir." (Oswald Campesato).

- **BERT ve Varyantları:** BERT (Bidirectional Encoder Representations from Transformers), iki yönlü bir modeldir ve tokenları cümle bağlamında anlamak için tasarlanmıştır. Belirli bir tokenı ve iki yönlü bağlamını kullanarak maskelenmiş kelime modelleme (MLM) ve bir sonraki cümle tahmini (NSP) gibi görevlerde eğitilir. "BERT is trained to understand words within the context of sentences." (Oswald Campesato). DistilBERT gibi varyantlar da mevcuttur.

2. Prompt Mühendisliği: LLM'lerle Etkileşim Sanatı

- **Prompt Mühendisliği Nedir?** Prompt mühendisliği, LLM'lerden istenen çıktıyı elde etmek için input (prompt) tasarlama ve iyileştirme sürecidir. "Prompt engineering is the art and science of designing effective prompts to get the desired output from large language models (LLMs)." (Subramanian Venkataraman).
- **Etkili Promptlar Yazmak:** Etkili promptlar, net talimatlar, bağlam ve istenen ton veya stil gibi çeşitli öğeler içerir. "To write effective prompts, it is important to include all the elements of a prompt." (Subramanian Venkataraman).
- **Prompt Türleri:Zero-shot Prompting:** Modele herhangi bir örnek verilmeden direkt soru sormak.
- **Few-shot Prompting:** Modeli belirli bir görevde yönlendirmek için birkaç örnek sağlamak. "Few-Shot Prompting is a technique to guide the model in generating responses tailored to your specific needs." (Subramanian Venkataraman).
- **Chain of Thought Prompting:** Zero-shot prompting'in bir uzantısıdır, burada modelin mantıksal bir düşünce zincirini takip etmesi için interconnected promptlar verilir.
- **Tree of Thoughts Prompting:** Self-consistency prompt'a benzer, ancak önceki adımlardan elde edilen cevaplar bağlam olarak kullanılır. "The Tree of thoughts prompting is similar to self-consistency prompt where the answer received from one question is passed as a context to the next question." (Subramanian Venkataraman).
- **Role Prompting:** Modelden belirli bir role bürünmesini istemek.
- **Prompt Parametreleri:** Prompt tasarımında modeli yapılandırmak için kullanılan parametreler şunlardır:
- **Context:** Modele daha iyi anlaması için ek bilgi veya bağlam sağlamak.
- **Max_tokens:** Cevap için maksimum kelime sayısını ayarlamak. "Setting a maximum word count for the response." (Subramanian Venkataraman).

- **Temperature:** Çıktının yaratıcılığını veya rastgeleliğini ayarlamak. "Adjusting the creativity or randomness of the response." (Subramanian Venkataraman).
- **Model:** Prompta göre belirli bir AI modeli seçmek.
- **Constraints:** Modelin yanıtını belirli sınırlar içinde tutmak.
- **Prompt Geliştirme Süreci:** Problem çözme, değişkeni izole etme, basitleştirme, çözme, kontrol etme ve özel durumları dikkate alma gibi adımları içerir. "To solve a generic first-degree equation, follow these steps: 1. Identify the Equation... 2. Isolate the Variable..." (Valentina Alto).
- **Modelin Evcilleştirilmesi:** İstenen çıktıyı elde etmek için promptun tamamlanmasının anatomisini anlamak ve modelin davranışını yönlendirmek önemlidir.

3. İleri Prompt Mühendisliği Teknikleri ve LLM Uygulamaları

- **Tool Kullanımı:** LLM'ler, belirli görevleri yerine getirmek için harici araçları (örneğin, hesap makinesi) kullanacak şekilde eğitilebilir. "{tool: calculator, expression: 10 * 4 * 2} gibi bir dize, ifadenin sonucunu hesaplamak için bir matematiksel yorumlayıcının çağrılmasını tetikler." (prompting.pdf).
- **Retrieval Augmented Generation (RAG):** Harici bilgi kaynaklarından (örneğin, belgeler, veritabanları) relevant bilgileri alarak LLM'lerin yeteneklerini geliştiren bir tekniktir. "Retrieval-augmented generation (RAG) involves providing the language model with additional information or context from external sources." (Subramanian Venkataraman). Bu, modelin belirli bir bilgi alanında daha doğru ve bağlama duyarlı cevaplar üretmesini sağlar.
- **Vektör Veritabanları:** RAG'de önemli bir rol oynar. Yapılandırılmamış verileri (metin, sayı, resim vb.) yüksek boyutlu uzayda sayısal vektörler olarak saklamak için kullanılır. "It's highly used for storing vector embeddings, a numerical representation of data in high dimensional space." (Subramanian Venkataraman). Kosinüs benzerliği gibi metrikler, vektörler arasındaki benzerliği ölçmek ve sorgularla eşleşen relevant belgeleri almak için kullanılır.
- **Chunking:** Metin verilerini yönetilebilir parçalara bölme tekniğidir. Kaydırma penceresi (sliding window) chunking, metni belirli sayıda karaktere dayalı olarak örtüşen parçalara ayırır. "Sliding window chunking is a technique used for dividing text data into overlapping chunks, or windows, based on a specified number of characters." (James Phoenix ve Michael Taylor).
- **Fine-tuning:** Belirli bir görev veya etki alanı için mevcut bir LLM'i daha fazla eğitmeyi içerir. Bu, modelin endüstriye özgü bir taksonomi gibi belirli bir anlayışa sahip olmasını sağlamak için kullanılabilir. "If this is the case, we might want to

leverage LLMs that have been trained and fine-tuned for this purpose, like Microsoft's BioGPT." (Valentina Alto).

- **Ön Ek Ayarlaması (Prefix Fine-tuning):** Her katmanın başına birkaç vektör ekleyerek modelin ince ayarını yapmanın bir yoludur.
- **Özel Modeller:** Belirli ihtiyaçlar için optimize edilmiş modellere (örneğin, Microsoft'un BioGPT'si gibi endüstriye özgü taksonomiler için eğitilmiş modeller) erişim ihtiyacı olabilir.
- **Lokal Çalıştırma:** İnternet bağlantısı olmayan senaryolarda (örneğin, açık denizdeki bir santral), modellerin lokal olarak çalıştırılması gerekebilir. Bu, modelin tescilli olup olmaması veya gerekli işlem gücünün (trilyonlarca parametreye sahip bir modeli barındırabilecek bir süper bilgisayar gibi) mevcut olup olmamasına bağlıdır.
- **Kullanım Alanları:** LLM'ler ve prompt mühendisliği, veri analizi, karar verme, yazılım geliştirme, soru yanıtlama, özetleme, çeviri, metin sınıflandırması ve duygu analizi gibi çeşitli alanlarda uygulanmaktadır. (Ajantha Devi Vairamani & Anand Nayyar, Subramanian Venkataraman).
- **Etik Hususlar ve Güvenlik:** LLM'lerin kullanımıyla ilgili etik hususlar ve veri güvenliği önemli konulardır. "Ethical considerations" ve "Data security" (Ajantha Devi Vairamani & Anand Nayyar). Modellerin zararlı içerik üretmesini önlemek için guardrail'ler (koruyucular) kullanılabilir. "the refusal breaker pattern to test these guardrails out." (Shivendra Srivastava ve Naresh Vurukonda). Siber güvenlik bağlamında, modellerin kötü amaçlı yazılımlar hakkında bilgi vermemesi gibi konular önemlidir.

4. Veri Analizi ve Finansal Uygulamalar

- **Veri Analizi:** LLM'ler, veri analizi süreçlerinde kullanılabilir. Örnekler arasında, satın alma değerine göre en iyi 10 müşterinin belirlenmesi, aylık satış trendlerinin analizi ve satışların farklı ülkelere göre dağılımının incelenmesi yer alır. "Here are the top 10 customers based on their total purchase value for the last year:" (Dr. Harald Gunia et al.).
- **Finansal Risk Değerlendirmesi:** LLM destekli araçlar, finansal riskleri değerlendirmek için kullanılabilir. Bu, yatırımların incelenmesini, iş ortaklarının kredi geçmişlerinin derinlemesine incelenmesini ve risk faktörlerinin neden-sonuç analizini içerebilir. "MoneyWatch's advanced Assessment capability, he meticulously evaluates the financial risks associated with a myriad of business operations..." (Dr. Harald Gunia et al.).
- **Finansal Metrikler:** Finansal sağlığı değerlendirmek için borç/öz sermaye oranı, borç oranı ve operasyonel marj gibi solvabilite ve karlılık metrikleri kullanılır.

- **Pazar Analizi:** LLM'ler, belirli sektörlerdeki (örneğin, lityum rezervleri) pazar analizi için kullanılabilir. Dünya çapındaki rezervler ve önemli oyuncular hakkında bilgi çıkarılabilir. "In 2022, reserves of lithium in Chile amounted to an estimated 9.3 million metric tons, making it the largest worldwide." (Dr. Harald Gunia et al.).

5. Teknik Uygulama Detayları

- **Çevresel Kurulum:** Python tabanlı projeler için sanal ortamların (venv) oluşturulması ve gerekli paketlerin (requirements.txt) yüklenmesi standart bir uygulamadır.
- **Vektör Depolama (Vector Stores):** Vektör depolama için ChromaDB gibi açık kaynaklı araçlar kullanılabilir. Metin verileri parçalara ayrılır, gömme modelleri kullanılarak vektörlere dönüştürülür ve vektör veritabanına yüklenir. Sorgular, vektör benzerliği kullanılarak eşleşen belgeleri almak için vektör veritabanına gönderilir.
- **SQL Ajanları:** LLM'ler, veritabanlarıyla etkileşim kurmak için SQL ajanları olarak kullanılabilir. Bu ajanlar, kullanıcı sorularına dayanarak SQL sorguları oluşturabilir, çalıştırabilir ve sonuçları yorumlayabilir. CRUD (create, read, update, delete) işlemleri dahil olabilir. "You are an agent designed to inte Given an input question, create a syntactically c run, then look at the results of the query and re" (James Phoenix ve Michael Taylor).
- **Guidance:** Promptları veya programları iç içe yerleştirmeye olanak tanıyan bir çerçevedir.
- **Logit Bias:** Tokenlarla ilişkili olasılıkları etkilemeye olanak tanır. Belirli tokenların üretilme olasılığını azaltmak için kullanılabilir.
- **Kelime Ağırlıklandırma:** Promptlardaki kelimelerin önemini etkilemek için parantezler veya (anahtar kelime: faktör) formatı kullanılabilir. (pirate: 1.5) gibi, modelin belirli tokenlara %50 daha fazla dikkat etmesini sağlar.
- **Özel Tokenlar:** BERT gibi modellerde [CLS], [SEP], [PAD] ve [UNK] gibi özel tokenlar kullanılır.

Bu brifing belgesi, sağlanan kaynaklardaki ana temaları ve önemli bilgileri özetlemektedir. LLM'lerin temel çalışma prensiplerinden, prompt mühendisliği tekniklerine, ileri uygulamalardan veri analizi ve finansal kullanımlara kadar geniş bir yelpazeyi kapsamaktadır. Belgeler ayrıca, LLM'lerle çalışırken dikkate alınması gereken etik ve güvenlik konularına da değinmektedir.