İnsan verisiyle karşılaştırmalı bir çalışma: Büyük dil modelleri dili ne kadar yüzeysel işliyor? A comparative study with human data: Do LLMs have superficial language processing?

Onur Keleş
Department of Linguistics
Boğaziçi University
Istanbul, Turkey

Nazik Dinçtopal Deniz
Department of Foreign Language Education
Boğaziçi University
Istanbul, Turkey

***Özetçe* —Bu çalışmada, psikodilbilimdeki yüzeysel dil işleme hipotezini bir doğal dil üretimi (NLG) görevinde kod-çözücü büyük dil modelleriyle (LLM) test edilip insan performansıyla karşılaştırılması amaçlanmıştır. Çalışmada LLM'lerin insanlar gibi cümleleri ne kadar sezgi ile ve ne kadar sözdizimsel kurallara göre işledikleri incelenmiştir. Ferreira vd.'nin İngilizce konuşanlarla gerçekleştirdiği Yeterince-İyi (Good-Enough) dil işleme deneyi, bu çalışmada birebir olarak OpenAI''nın GPT-3.5-turbo ve text-davinci-003 modelleriyle tekrarlanmıştır. LLM'lerin verdiği yanıtlar RoBERTa ile tahminlediğimiz İngilizce deney cümlelerinin ortalama Sürpriz değeri (negatif logaritmik-olabilirliği) ve insan verisiyle karşılaştırılmıştır. GPT-3.5-turbo modelinin, text-davinci-003 modeline göre sözdizimsel olarak daha doğru işleme yaptığı, ancak anlamsal anomali bulunan cümleleri işlemede insanlardan daha başarısız olduğu ve bu modellerin insanlar gibi soyutlama yapamadıkları bulunmuştur. Sürpriz değerleriyle karşılaştırınca GPT-3.5-turbo modelinin sözdizimsel ipuçlarına dikkat ederek cümle işlediği, text-davinci-003 modelinin ise daha yüzeysel cümle işlediği bulunmuştur.**

Anahtar Kelimeler—dil işleme, büyük dil modelleri, doğal dil üretimi, psikodilbilim

***Abstract*—This study aims to test the superficial language processing hypothesis on a natural language generation (NLG) task with large language models (LLMs) and compare it with human performance. We examined whether LLMs sometimes process sentences superficially like humans or always have deep-syntactic processing. The good-enough language processing experiment conducted by Ferreira et al. with English speakers was repeated in this study with OpenAI''s GPT-3.5-turbo and text-davinci-003 models. The grammatical accuracy of the answers given by the LLMs was compared with the average Surprisal (negative log-likelihood) of the English experimental sentences we predicted with RoBERTa, and with human data. We found that the GPT-3.5-turbo model was more accurate than the text-davinci-003 model, but was less successful than humans in processing sentences with semantic anomalies and these models were not able to do abstraction like humans. When compared with Surprisal values, we found that the GPT-3.5-turbo model processed sentences by paying more attention to syntactic cues, while the text-davinci-003 model processed sentences superficially.**

Keywords—sentence processing, LLMs, natural language generation, psycholinguistics

979-8-3503-8896-1/24/$31.00 ©2024 IEEE

I. GİRİŞ

Psikodilbilim alanında son yirmi yıldır yapılan cümle işleme çalışmaları, insanların bazen cümleleri "yüzeysel" işledikleri, farklı yapılardaki cümleleri sıkça yanlış analiz ettikleri ve çevrimiçi cümle anlama sürecindeki bu yanılgıların eksik değerlendirmelerden kaynaklandığını ortaya koymuştur [2]. Bu çalışmalar cümle anlama sisteminin yüzeysel stratejiler geliştirip bunları kısayol olarak kullanabileceği varsayılan Yeterince-İyi İşleme (Good-Enough Parsing) yaklaşımının temelini atmıştır [3]. Bu model insanların bazen sözdizimsel bilgiden ziyade anlamsal/gerçek dünya bilgisini kullanarak dil girdisinin yüzeysel ve yanlış analizini yapma eğiliminde olduğunun altını çizer. Örneğin, anlamsal yanılsama (semantic illusion) içeren sorular sorulduğunda ("Musa gemiye her hayvan türünden kaç tane aldı?"), katılımcılar soruya yanıt olarak "2" verirler. Yapısal belirsizlik (syntactic ambiguity) içeren cümlelerde ise cümlenin yeniden analizini gerektiren anlama sorularında katılımcıların doğruluk oranlarının düşük olduğu görülmüştür. Böyle yüzeysel cümle işleme eğilimi tematik yanlış yorumlamalarda da gözlemlenebilir. Yani, eğer bir cümlenin sözdizimsel yapısı (örn. "adam köpek tarafından ısırıldı") kişinin gerçek dünyadaki eylem bilgisiyle (genelde köpekler insanları ısırır) uyumlu değilse (örn. "köpek adam tarafından ısırıldı"), insanlar gerçek dünya bilgisine güvenip cümlenin sözdizimsel yapısının gerektirdiğinden (adamın edilgen bir cümlenin öznesi olması) ziyade "kim ısırdı?" sorusuna "köpek" diyerek fark etmeden yanlış cevap verebilirler. Bütün bunlar, insanların her zaman sözdizimsel yapıya göre cümle işlemediklerini, bunun yerine bazen yüzeysel yöntemleri kullandıklarını göstermektedir. Bu kapalı dil modellerinin bu cümleleri işlerken ne kadar sözdizimsel bilgi ne kadar gerçek dünya bilgisi ve yüzeysel yöntem kullanacakları açık değildir. Bundan dolayı, bu çalışmada bu tür tematik uyumsuzluk içeren cümlelerin ana dili İngilizce olan kişiler tarafından değerlendirildiği deneyi GPT-3.5-turbo ve text-davinci-003 büyük dil modelleriyle tekrarlanmıştır [1].

A. Cümle İşleme ve Büyük Dil Modelleri

Bu kısımda büyük dil modellerinin cümle işleme davranışları hakkında yapılan çalışmalar özetlenmektedir:

Sözdizimsel Değerlendirmeler: [4] gibi çalışmalar doğal dil işleme görevlerinde kullanılan RNN (Recurrent Neural Network) yapıları ile ve BERT gibi dönüştürücü tabanlı LLM"lerin sözdizimsel işlemesini incelemiştir. Bu çalışmalar, sözdizimsel olarak doğru ve yanlış cümle çiftlerinin olasılıklarını karşılaştırarak özne-fiil uyumu ve argüman bağlama gibi olgulara odaklanmıştır.

Sürpriz ve Cümle İşleme Zorluğu: Araştırmalar [5], Sürpriz (kelimenin negatif logaritmik olabilirliği) ve işleme zamanı arasında pozitif bir ilişki olduğunu göstermiştir. Bu, fMRI, MEG çalışmaları ve okuma süreleri ile de desteklenmiştir. Beklenmediklik gösteren yüksek Sürpriz değerleri, artan okuma süreleri ve cümle işleme zorluğu ile ilişkilidir.

Tematik Uyuşmazlık ve Gerçek Dünya Bilgisi Kullanımı: Pedinotti ve arkadaşları [6], olası ve olası olmayan olayları ayıran görevlerde dönüştürücü tabanlı LLM"leri test etmiştir. Çalışmalarında BERT ve RoBERTa gibi modeller kullanılmıştır. Kauf ve arkadaşları [7] ise LLM"lerin yine olası ve olası olmayan farklı özne-nesne kombinasyonlarına ne kadar olasılık atadıklarını incelemiştir. Bu modellerin insan değerlendirmeleriyle genellikle iyi bir benzerlik sağladığı ve özellikle anlamsal anomali içeren cümlelerde başarılı olduğu, yani düşük olabilirlik atadığı bulunmuştur.

Anlamsal Yanılsamalar: Nair ve arkadaşları [8], RoBERTa"yı maskelemeli dil modelleme paradigmasında anlamsal yanılsamalar içeren cümlelerle test etmiştir. Anlamsal benzerlik ölçümlerinin anlamsal yanılsama etkisi ile iyi bir korelasyon göstermediği bulunmuştur. Yüksek olasılık tahminleri ile artan yanılsama oranları arasında zayıf bir ilişki olduğu gösterilmiş, bu da bu olasılık tahminlerinin bağlamsal uyumun güvenilir bir ölçüsü olmadığını düşündürmektedir.

B. Literatürdeki Boşluk

Yukarıda listelenen çalışmalar BERT veya RoBERTa gibi kod üretici dönüştürücü tabanlı dil modellerini çeşitli psikodilbilim Doğal Dil Anlama (NLU) görevlerinde kullanıp bu değerlendirmeleri insan yargılarıyla karşılaştırmıştır. Bilgimiz dahilinde, bu araştırmalar OpenAI"ın GPT 3.5 veya text-davinci-003"ü gibi kod çözücü LLM"leri insanlarla yapılan bir cümle işleme deneyini Doğal Dil Üretme (NLG) görevi olarak tekrarlamamıştır. Bundan yola çıkarak mevcut çalışma, Ferreira"nın [1] Yeterince-İyi cümle işleme deneyini iki büyük dil modeliyle ile tekrarlayıp RoBERTa ile tahminlenen deney cümlelerinin Sürpriz değerleri ile yüzeysel işleme arasındaki ilişkiyi ortaya çıkarmayı amaçlamaktadır.

II. YÖNTEM

A. Materyal

Bu deneyde Ferreira"nın [1] çalışmasında kullanılan İngilizce deney cümlelerinin aynısı kullanılmıştır. Deney cümleleri hem sözdizimini hem de cümlenin anlamsal olabilirliğini manüpüle etmektedir. Deney cümlelerinde sözdizimi (1a) ve (1b)'deki gibi etken çatı veya (1c) ve (1d)'deki edilgen çatı olabilir; anlamsal olabilirlik (1a) ve (1c)'deki gibi Tematik Uyumlu[1] veya (1b) ve (1d)'de olduğu gibi Tematik Uyumsuz[2] olabilir.

(1) Örnek Deney Cümlesi (Birinci Set)

a. Etken, Tematik Uyumlu: The dog bit the man. (Türkçe: "Köpek adamı ısırdı")

b. Etken, Tematik Uyumsuz: The man bit the dog. (Türkçe: "Adam köpeği adamı ısırdı")

c. Edilgen, Tematik Uyumlu: The man was bitten by the dog. (Türkçe: "Adam köpek tarafından ısırıldı")

d. Edilgen, Tematik Uyumsuz: The dog was bitten by the man. (Türkçe: "Köpek adam tarafından ısırıldı")

e. Soru: Who bit? (Türkçe: "Kim ısırdı?")

Her koşulun ardından (e)'deki gibi yüklemin öznesini soran bir soru bulunmaktadır. Toplamda her biri 24 deney cümlesinden oluşan 3 set vardır. Birinci set, (a-d)"deki gibi taraflı tematik uyumsuzluk içeren cümlelerden oluşmaktadır[3] İkinci set, nesnenin cansız olduğu (örn. "şef önlüğü giydi") ve dolayısıyla argümanların tersine çevrilmesinin anlamsal bir anomaliye yol açtığı cümleleri içerir (örn. "önlük şefi giydi"). Üçüncü set, iki argümanın da eylemi yapan kişi olma olasılığının eşit olduğu simetrik cümleleri içerir (örn. "kadın adamı öptü"). Dolayısıyla her bir katılımcı toplamda 72 deney cümlesi görüp cevaplamıştır.

B. Prosedür

Ferreira"nın çalışmasındaki deney iki büyük dil modeliyle tekrarlanmıştır: GPT 3.5 ve text-davinci-003. İlk önce her sette deneyde kullanılan cümleler için dört koşul üretilmiş, cümleler farklı listelere dağıtılmış, dil modellerinin cevaplaması için randomize çapraz döngülü karşılıklı dengeli çalışma dizaynı uygulanmıştır. Ardından, hem GPT 3.5 hem de text-davinci-003 için sıfır atış öğrenme tekniğiyle içe bir işlem hattı oluşturulmuş ve aşağıdaki istem verilmiştir:

İstem: "Psikodilbilimsel bir çalışmanın katılımcısısınız. Yalnızca tek kelimelik yanıt verin."

C. Simülasyon

Ferreira"nın orijinal çalışmasına Michigan Eyalet Üniversitesi"nden toplam ana dili İngilizce olan 63 lisans öğrencisi katılmıştır. Deney replikasyonunda ise her iki model deneyi 20 kez tamamlamıştır. Toplamda modellerden 2520 gözlem elde edilmiştir.

D. Sürpriz Değerinin Hesaplanması

Model cevabına ve sözdizimsel doğruluğuna ek olarak, Transformers kütüphanesinden RoBERTa dil modelini kullanarak her bir deney cümlesi için Sürpriz değeri hesaplanmıştır. İlk olarak, RoBERTa modelinin bölümleme algoritması kullanarak girdi içerisindeki metin bölümlenmiş ve gerekli özel jetonların dahil edilmesini sağlanmıştır. Bunun ardından, RoBERTa modeli ile her cümledeki jetonlar için tahminler (logit) üretilmiş, bu logit değerleri daha sonra Softmax fonksiyonu kullanılarak olasılıklara dönüştürülmüştür. Bu olasılıkların negatif logaritması alınarak her bir jeton için ne kadar beklenmedik olduğunu gösteren bir Sürpriz değeri elde edilmiştir. Daha

[1] Sözdizimi ve gerçek dünya bilgisi uyumludur.

[2] Sözdizimi ve gerçek dünyası bilgisi uyumlu değildir.

[3] Örneğin köpeğin adamı ısırması gerçek dünyada mümkündür ama yaygın değildir. Bu nedenden dolayı alternatif argüman sırasına karşı bir beklenti vardır.

sonra Sürpriz değerleri toplanarak bir cümlenin toplam Sürpriz değeri hesaplanmıştır:

$$\text{Sürpriz}(jeton_i) = -\log(P(jeton_i)) \quad (1)$$

$$\text{Toplam Sürpriz} = \sum_{i=1}^{N} \text{Sürpriz}(jeton_i) \quad (2)$$

III. SONUÇLAR

Tablo I'de birinci set cümleler için sonuçlar verilmektedir. İstatistiksel analizler bu kısmın sonunda verilmiştir. Birinci set "the dog bit the man" (Türkçe: "köpek adamı ısırdı") gibi belirli sıradaki argümanların daha muhtemel olduğu cümleler içerir. Bu sonuçlara göre bütün katılımcıların sözdiziminin ve gerçek dünya bilgisinin çatışmadığı ve normal sırayı takip eden Tematik Uyumlu koşullarında benzer derecede iyi performans sergilediği görülmüştür. İnsan verisi tam doğruluğa ulaşamazken hem GPT 3.5 hem de text-davinci-003 %100 doğruluğa sahip olmuştur. Bununla birlikte, Etken ve Tematik Uyumsuz koşulunda (etken cümle ancak argümanlar tersine çevrilmiş), insanlar ve GPT 3.5 benzer şekilde performans gösterirken text-davinci-003 çok daha düşük bir doğruluk oranına sahip olmuştur. Birinci setteki gibi taraflı cümleler için daha iyi bağlamsal anlayışa sahip olduğu varsayılan text-davinci-003, cümledeki sözdizimsel ipuçlarından ziyade daha çok gerçek dünya bilgisini kullanarak sözdizimsel olarak hatalı yanıt vermiştir. Son olarak Edilgen ve Tematik Uyumsuz koşulunda (edilgen cümle ancak argümanlar tersine çevrilmiş) insanların ve text-davinci-003"ün hata yapıp yüzeysel işleme kullandığı ancak GPT 3.5 modelinin cümlenin sözdizimine uyduğunu ve daha fazla doğru yanıt verdiği bulunmuştur.

Şekil 1"e göre ilk set için insanlar ve text-davinci-003 modeli cümle sürprizi ve sözdizimsel doğruluk bakımından negatif bir korelasyon göstermektedir. İnsanlar ve text-davinci-003 modeli için cümle işlemede doğruluk oranı düştükçe cümlenin sürprizi veya beklenmedikliği artmaktadır. Ancak bu setteki cümlelerde GPT 3.5 genel olarak bu eğilimi izlememektedir. Hem İngilizce konuşucularından hem de text-davinci-003 modeline göre sözdizimsel olarak doğru bir işlemleme gerçekleştirmiştir.

Set	Koşul	Cevaplayan	Doğruluk	Cümle Sürprizi
1	Etken, Uyumlu	gpt3-5	1.00	0.01
1	Etken, Uyumlu	insan (Ferreira 2003)	0.96	0.01
1	Etken, Uyumlu	text-davinci-003	1.00	0.01
1	Etken, Uyumsuz	gpt3-5	0.87	0.08
1	Etken, Uyumsuz	insan (Ferreira 2003)	0.92	0.08
1	Etken, Uyumsuz	text-davinci-003	0.61	0.08
1	Edilgen, Uyumlu	gpt3-5	1.00	0.19
1	Edilgen, Uyumlu	insan (Ferreira 2003)	0.95	0.19
1	Edilgen, Uyumlu	text-davinci-003	1.00	0.19
1	Edilgen, Uyumsuz	gpt3-5	0.93	0.27
1	Edilgen, Uyumsuz	insan (Ferreira 2003)	0.74	0.27
1	Edilgen, Uyumsuz	text-davinci-003	0.53	0.27

TABLO I: Birinci Set İçin Doğruluk ve Ortalama Cümle Sürprizi Sonuçları

Tablo II ikinci setteki cümlelerin sonuçlarını özetlemektedir. İkinci set "the chef wore the apron" (Türkçe: "şef önlüğü giydi") gibi argümanların yeri değiştiğinde anlamsal anomali yaratan cümleler içerir. Bulgular, ilk gruba benzer şekilde modellerin Tematik Uyumlu koşullarında mükemmel doğruluk oranlarına sahip olduğunu ve insanların cümle işleme verisinde deney ortamından kaynaklanabilecek gürültü olabileceğini gösteriyor. Çarpıcı olarak, her iki model de tüm koşullarda neredeyse mükemmel doğruluk elde eden İngilizce konuşucularına göre, Tematik Uyumsuz koşullarında % 30 oranlarında bir doğruluk göstermiştir. Bu bulgu, bu tür kod çözücü LLM"leri insan dil işlemesinden ayıran noktalardan biri olabilir. Modeller cümlenin argümanları yer değiştirdiğinde ortaya çıkan anlamsal anomaliyi işlemede başarısız olmuştur. Fakat insanlar cümlelerdeki anomaliyi tolere edip sözdizimsel kurallara uyarak cümleyi doğru bir şekilde işleyebilmiştir.

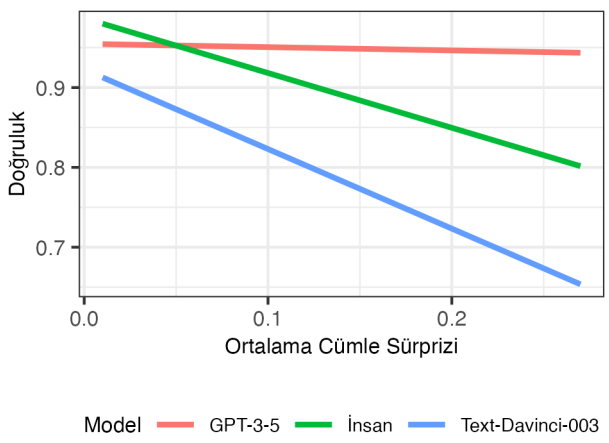
Son olarak, Şekil 2"den elde edilen sonuçlar, modellerin cümle işleme davranışlarını insanlardan daha da ayırmaktadır. Text-davinci-003 ve GPT 3.5 modellerinin cümle işlemedeki doğruluk oranlarının bu tür argümanları yer değiştiremeyen cümleler için büyük ölçüde cümlenin sürpriz oranlarına bağlı olduğu görülmektedir. Bir cümlenin sürpriz oranı yüksekse doğruluk oranı dil modelleri için düşmektedir. İlginç bir şekilde, insanlar bu gibi anlamsal anomali içeren cümlelerde benzer bir hassasiyet göstermemektedir. Bu, anomali olan ve soyutlama gerektiren cümleleri işlerken insanların sezgileri yerine sözdizim kurallarına bağlı kaldığını ancak dil modellerinin bu soyutlamayı yapamadığı ve sözdizimini tam olarak iyi işleyemediğini göstermektedir.

Set	Koşul	Cevaplayan	Doğruluk	Cümle Sürprizi
2	Etken, Uyumlu	gpt3-5	1.00	0.02
2	Etken, Uyumlu	insan (Ferreira 2003)	0.99	0.02
2	Etken, Uyumlu	text-davinci-003	1.00	0.02
2	Etken, Uyumsuz	gpt3-5	0.31	2.57
2	Etken, Uyumsuz	insan (Ferreira 2003)	0.95	2.57
2	Etken, Uyumsuz	text-davinci-003	0.21	2.57
2	Edilgen, Uyumlu	gpt3-5	1.00	0.04
2	Edilgen, Uyumlu	insan (Ferreira 2003)	0.95	0.04
2	Edilgen, Uyumlu	text-davinci-003	1.00	0.04
2	Edilgen, Uyumsuz	gpt3-5	0.35	0.78
2	Edilgen, Uyumsuz	insan (Ferreira 2003)	0.92	0.78
2	Edilgen, Uyumsuz	text-davinci-003	0.24	0.78

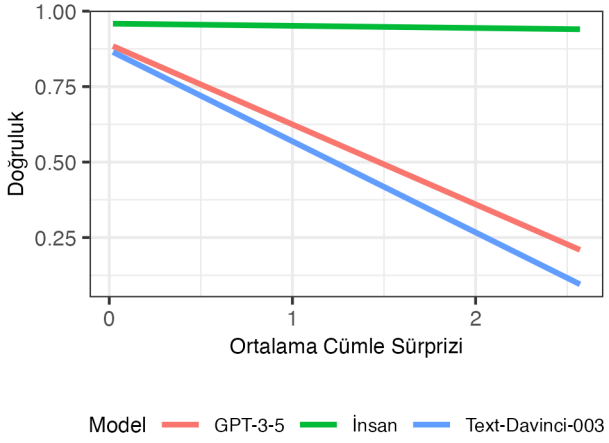
TABLO II: İkinci Set İçin Doğruluk ve Ortalama Cümle Sürprizi Sonuçları

Tablo III üçüncü setin sonuçlarını özetlemektedir. Üçüncü set "the man kissed the woman" (Türkçe: "adam kadını öptü") gibi argümanların yeri değiştiğinde herhangi bir önyargı yaratmayan simetrik durumları içerir. Bu sonuçlardaki ilk çarpıcı bulgu, dil modellerinin tüm koşullar için insan katılımcılarından daha iyi performans göstermesine rağmen ilk iki setten farklı olarak mükemmel doğruluğa sahip olmamasıdır. Bu nedenle, her iki argüman sırasının eşit derecede muhtemel olduğu durumlarda modelin sözdizimsel ipuçlarına daha az dikkat edebileceği söylenebilir. Genel olarak, insanlar[4] ve dil modelleri bu sette benzer cümle işleme örüntüleri göstermişlerdir. Ayrıca beklendiği üzere, bu setteki cümlelerin sürpriz oranlarının sıfıra yakın ve aralarında sadece minimal bir fark olduğu görülmektedir. Bu nedenle, simetrik setteki cümlelerin doğrulukla ilişkisinin karşılaştırılması pek bilgilendirici değildir.

[4]Orijinal çalışmada bu sette Tematik Uyumsuz koşulları için veriler mevcut değildir.



Şekil 1: Birinci Set İçin Doğruluk ve Sürpriz Korelasyonu



Şekil 2: İkinci Set İçin Doğruluk ve Sürpriz Korelasyonu

Set	Koşul	Cevaplayan	Doğruluk	Cümle Sürprizi
3	Etken, Uyumlu	gpt3-5	0.92	0.02
3	Etken, Uyumlu	insan (Ferreira 2003)	0.89	0.02
3	Etken, Uyumlu	text-davinci-003	1.00	0.99
3	Etken, Uyumsuz	gpt3-5	0.99	0.02
3	Etken, Uyumsuz	insan (Ferreira 2003)	NA	0.02
3	Etken, Uyumsuz	text-davinci-003	0.90	0.02
3	Edilgen, Uyumlu	gpt3-5	0.97	0.02
3	Edilgen, Uyumlu	insan (Ferreira 2003)	0.83	0.02
3	Edilgen, Uyumlu	text-davinci-003	0.95	0.02
3	Edilgen, Uyumsuz	gpt3-5	1.00	0.02
3	Edilgen, Uyumsuz	insan (Ferreira 2003)	NA	0.02
3	Edilgen, Uyumsuz	text-davinci-003	0.82	0.02

TABLO III: Üçüncü Set İçin Doğruluk ve Ortalama Cümle Sürprizi Sonuçları

A. *İstatistiksel Analiz*

Yanıtları istatistiksel olarak analiz etmek için model yanıtlarına (doğru - yanlış) Bernouilli dağılımını izleyen lojistik regresyon modeli uygulanmıştır. Sabit etki olarak model, koşul ve set girilmiş, rastgele etki olarak simülasyon sayısı girilmiştir. Olabilirlik Oran Testi (Likelihood Ratio Test) ile farklı sabit etkilerin girildiği modeller kıyaslanmış ve en kompleks model en iyi model seçilmiştir ($\chi^2 = 42.2$, $p < 0.001$). Regresyon sonuçlarına göre GPT 3.5 modeli bu görevde daha doğru cevaplar vermiştir ancak koşullara bakıldığında argümanlar yer değiştirdiğinde (Tematik Uyumsuz koşullar) ve özellikle anlamsal anomaliğe yol açan ikinci set modellerin doğruluğunu fazlaca düşürmektedir (Tablo IV).

Yordayıcılar	Odds Oranı	Güven Aralığı	P değeri
model [GPT 3.5]	3.00	2.27 – 3.96	<0.001
koşul [Etken, Uyumsuz]	0.02	0.01 – 0.04	<0.001
koşul [Etken, Uyumlu]	1.40	0.52 – 3.75	0.499
koşul [Edilgen, Uyumsuz]	0.02	0.01 – 0.04	<0.001
set [2]	0.16	0.12 – 0.22	<0.001
set [3]	2.78	1.89 – 4.09	<0.001
sürpriz	0.82	0.74 – 0.91	<0.001

TABLO IV: Lojistik Regresyon Analiz Sonuçları

IV. Sonuç

Mevcut çalışma bulgularını özetlemek gerekirse: (i) kod çözücü dil modellerinden olan GPT 3.5 ve text-davinci-003, düşük sürpriz oranına sahip cümlelerde (örn. beklendik ve dünya bilgisine uygun durumlarda) insanlardan daha iyi performans göstermiştir. (ii) LLM"ler ve insanlar yüzeysel dil işlemeyi farklı şekillerde sergilemiştir. İnsanlar anlamsal anomaliye tahammül edip başarılı bir şekilde cümleyi işlerken her iki dil modeli de bunu yapamamıştır. (iii) Yüzeysel dil işleme ve sürpriz, text-davinci-003 gibi yüksek düzeyde bağlama duyarlı dil modelleriyle ve bir dereceye kadar insanlar için negatif oranda ilişkili olmuştur. Bu da cümlenin sürprizi azaldıkça ve özellikle anlamsal anomali durumunda dil modellerinin cümlenin sözdizimi yerine gerçek dünya bilgisine göre cümle işlediği söylenebilir. Sonuç olarak, kapalı modeller bazı koşullardaki cümleleri işlemede insanlar gibi soyutlama yapamamaktadırlar. Bu da LLM'lerin değerlendirilip geliştirilmesinde benchmark veri kümeleri kadar insan davranış verisiyle karşılaştırılmasının önemini göstermektedir.

Kaynaklar

[1] F. Ferreira, "The misinterpretation of noncanonical sentences," *Cognitive psychology*, vol. 47, no. 2, pp. 164–203, 2003.

[2] K. Christianson, A. Hollingworth, J. F. Halliwell, and F. Ferreira, "Thematic roles assigned along the garden path linger," *Cognitive psychology*, vol. 42, no. 4, pp. 368–407, 2001.

[3] F. Ferreira and N. D. Patson, "The 'good enough'approach to language comprehension," *Language and linguistics compass*, vol. 1, no. 1-2, pp. 71–83, 2007.

[4] R. Marvin and T. Linzen, "Targeted syntactic evaluation of language models," 2018.

[5] E. G. Wilcox, T. Pimentel, C. Meister, R. Cotterell, and R. P. Levy, "Testing the predictions of surprisal theory in 11 languages," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1451–1470, 2023.

[6] P. Pedinotti, G. Rambelli, E. Chersoni, E. Santus, A. Lenci, and P. Blache, "Did the cat drink the coffee? challenging transformers with generalized event knowledge," 2021.

[7] C. Kauf, A. A. Ivanova, G. Rambelli, E. Chersoni, J. S. She, Z. Chowdhury, E. Fedorenko, and A. Lenci, "Event knowledge in large language models: The gap between the impossible and the unlikely," *Cognitive Science*, vol. 47, no. 11, p. e13386, 2023.

[8] S. Nair and P. Resnik, "Words, subwords, and morphemes: What really matters in the surprisal-reading time relationship?" H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 11 251–11 260.