Bölüm 4

Büyük Dil Modellerinin (LLM) Kuramsal Sınırları ve Varsayımları

Tevfik Erdal Baylav[1]

Atınç Yılmaz[2]

Özet

Bu bölüm, büyük dil modellerinin kuramsal sınırlarını ve varsayımlarını, sistem başarısızlığı göstergesi değil, kullanım bağlamının sınırlarını tanımlayan özellikler olarak ele alır. Benchmark skorlarının anlamı hangi varsayımların devrede olduğuna bağlıdır; sınır analizi bu varsayım kümesini ve dolayısıyla karar veya bilgi kaynağı olarak modelin hangi bağlamlarda konumlandırılabileceğini görünür kılar. Dilin olasılıksal temsili, sonlu bağlam penceresi ve token düzeyinde eğitim hedefi hesaplama ve temsil varsayımlarını oluşturur; genelleme eğitim–değerlendirme dağılımları ve görev formatına bağımlıdır. Çıktının epistemik güvenilirliği bağlama bağımlıdır; dağıtım koşulları ve doğrulama pratikleri tasarım varsayımlarıyla örtüştüğünde çıktı bilgi kaynağı işlevi görebilir, aksi durumda güvenilirlik ve belirsizlik uygulama katmanında yönetilir. Halüsinasyon tasarım özelliği olarak, hangi görevlerde retrieval, doğrulama modülleri veya insan-in-the-loop ile sınırlandırılması gerektiğinin gerekçesini oluşturur. Açıklanabilirlik ve agent mimarileri, sınırların pratikte nasıl yönetileceği ve hangi mimari güvencelerin devreye alınacağı sorusunu gündeme getirir. Bölüm, LLM etkinliğinin bağlama bağımlı olduğunu ve retrieval, insan denetimi ve modüler doğrulama katmanlarıyla artırılabileceğini vurgulayarak sonuçlandırır.

1 MSc, PhD Candidate, Department of Computer Engineering, Beykent University, Istanbul, Turkey, baylav.tevfik@gmail.com

2 Assoc. Prof. Dr., Faculty of Applied Sciences, Marmara University, Istanbul, Turkey, atinc.yilmaz@marmara.edu.tr

https://doi.org/10.58830/ozgur.pub1351.c5537

1. Kuramsal Çerçeve ve Kullanım Çerçevesi

Benchmark skorları anlam taşıyor—taşıdıkları anlam hangi varsayımların devrede olduğuna bağlı. LLM'lerin sınırlarına bakmak, sistemin başarısızlığını değil, kullanım bağlamının sınırlarını tanımlayan özellikleri öne çıkarır. Tasarımın getirdiği bu özellikler, hangi bağlamlarda modelin karar veya bilgi kaynağı olarak konumlandırılabileceğini belirler; dolayısıyla sınır analizi, epistemolojik bir çerçeve sunarak uygulama katmanında dağıtım ve tasarım kararlarının temelini oluşturur.

Model, verilen token dizisinden sonraki tokenın olasılığını tahmin eder. Dil burada iletişim aracı veya anlam taşıyıcı değil, ayrık sembollerin ardışık olasılık yapısı olarak temsil edilir; semantik ve pragmatik yalnızca bu yapıya yansıdığı ölçüde modele girer. Bu temsil hesaplanabilirlik için neredeyse zorunludur; "anlama" ile "yüksek olasılıklı çıktı üretme" ise aynı şey değildir (Bender ve ark., 2021). Bu ayrım, olasılıksal çıktının bilgi ve karar sistemlerinde hangi bağlamlarda kaynak olarak kullanılabileceğinin sınırını tanımlar. Chomsky'ci perspektif dilin yalnızca gözlemlenebilir dizilerin istatistiği olmadığını vurgular (Chomsky, 1965); LLM tarafı ise kuralların veri içinde örtük öğrenildiğini varsayar (Brown ve ark., 2020). Bu metinde vurgu, bu temsilin hangi kullanım sınırlarını belirlediği ve bu sınırların sistem tasarımında nasıl dikkate alınacağı üzerinedir.

"İyi çalışma hangi koşullarda tanımlanır?" sorusu, başarı metriklerinin anlamlı olduğu varsayım kümesini netleştirir. Bu varsayımlar ihlal edildiğinde metrikler yanıltıcı hale gelebilir; sınır analizi, tam da bu varsayım kümesini ve dolayısıyla dağıtım bağlamını görünür kılar.

2. Hesaplama ve Temsil Varsayımları

2.1. Bağlam Penceresi ve Mimari Sınırları

Bağlam penceresi sonludur—bu bir arıza değil, mimarinin tanımlayıcı özelliğidir. Pencere dışındaki bilgi model için yapısal olarak erişilemez; temsilin sınırı burada çizilir. Transformer mimarisinde dikkat mekanizması $O(n^2)$ karmaşıklığa sahip olduğundan pencereyi büyütmek maliyet ve bellek getirir; sonsuz pencere de "tüm metni anlamak" değil, daha uzun sonlu dizi işlemek anlamına gelir (Vaswani ve ark., 2017). Bu özellik, modelin hangi soru ve görev türlerinde anlamlı yanıt üretebileceğinin sınırını belirler; sistem tasarımında bağlam gereksinimi ve retrieval-augmented generation (RAG) gibi tamamlayıcı mimariler bu sınırla uyumlu biçimde seçilebilir (Lewis ve ark., 2020).

2.2. Eğitim Hedefi, Ölçek ve Bağlam Bağımlılığı

Hesaplama varsayımları—sonlu parametre, sonlu bağlam, token düzeyinde çapraz entropi hedefi—modelin hangi fonksiyonları temsil edebileceğini yapısal olarak koşullandırır. Eğitim hedefi, T uzunluğundaki bir dizi için negatif log-olabilirlik minimize edilmesi olarak şu biçimde ifade edilebilir:

$$L(\theta) = -(1/T) \Sigma \log P(x_t \mid x_1, \ldots, x_{t-1}; \theta)$$

Bu hedef, modelin "en yüksek olasılıklı temsil"i öğrenmesini sağlar—"tüm geçerli doğrular" veya "bağlama göre en uygun" yanıtı değil. Çoklu geçerli yorumlar veya belirsizlik altında stratejik tercihler eğitim hedefine dolaylı yansır. Eğitim hedefi ile kullanım bağlamı arasındaki bu ilişki, hangi bağlamlarda modelin çıktısının doğrudan kullanılabileceğini, hangi bağlamlarda ise insan denetimi veya modüler doğrulama katmanlarıyla desteklenmesi gerektiğini belirler. Ölçek yasaları belirli metriklerde iyileşme vaat eder (Kaplan ve ark., 2020; Hoffmann ve ark., 2022); bu metrikler genelleme, tutarlılık veya güvenilirlikle özdeş değildir. Sınırlar sabit değil, kullanım bağlamına göre değişkendir—etkinlik bağlama bağımlıdır ve mimari seçimlerle kullanım bağlamı içinde kalınarak artırılabilir.

2.3. Formal Problem Statement: LLM as Bounded Stochastic Sequence Estimator

LLM, mühendislik perspektifinden sınırlı bir stokastik dizi tahmincisi olarak tanımlanabilir. Model, bir vokabüler V üzerinde tanımlı token dizilerinin koşullu olasılık dağılımını parametrize eder:

$$P\_\theta(x_1, \ldots, x_n) = \Pi\ P\_\theta(x_t \mid x_1, \ldots, x_{t-1})$$

Burada θ, eğitim verisi D üzerinden ampirik risk minimizasyonuyla öğrenilen parametre kümesidir. Model P_θ, gerçek veri dağılımı P_data'nın bir yaklaşımıdır; bu yaklaşımın kalitesi dağılım kayması (distribution shift) durumunda garanti edilemez. Bu formal çerçeve, modelin ne yaptığını—dağılıma uygun token dizisi üretmek—ve ne yapmadığını—gerçeği doğrulamak, nedensel çıkarım yapmak—açıkça tanımlar. Dağıtım kararları bu sınırlılık çerçevesinde alınmalıdır.

3. Genelleme ve Dağılımsal Sınırlar

3.1. Dağılım ve Görev Bağımlılığı

Genelleme, eğitim ve değerlendirme dağılımları arasındaki ilişkiye bağlıdır; bu ilişki görev-uyumlu kullanım alanının tanımında merkezi rol oynar. Eğitim dağılımı dışındaki girdilere çıktı kalitesi, test dağılımının eğitime yakınlığına

bağlıdır; görev genellemesi ise modelin belirli formatlarda eğitilip ince ayarlanması nedeniyle format ve görev tanımına duyarlıdır. Tıp alanında bu mesele somut biçimde gündeme gelmiştir: klinisyenlerin LLM çıktısını doğrudan klinik karar verme sürecine dahil ettiği durumlarda, modelin dağılım dışı sorgulara verdiği yanıtların güvenilirlik sorunları gözlemlenmiştir (Singhal ve ark., 2023). Bu özellikler başarısızlık göstergesi değil, dağıtım kararlarının dayandığı tasarım bilgisidir.

3.2. Dünya Bilgisi, Zamansal Sürüklenme ve Format Duyarlılığı

"Dünya bilgisi" örtük temsilde kodlanır; açık ontoloji veya nedensel grafik yoktur. Aynı bilgi farklı ifadeyle sorulduğunda tutarsızlık veya güncelliğini yitirmiş bilginin sunulması bu temsilin doğal sonucudur. Güncelleme etkisi izlenebilir değildir; ince ayar veya müdahale yan etkileri öngörülemez (Zhu ve ark., 2020). Bu özellikler, modelin hangi bilgi türleri ve zaman dilimleri için uygun olduğunu, hangi durumlarda retrieval veya dış veri kaynaklarıyla desteklenmesi gerektiğini belirler. Zamansal sürüklenme ve format duyarlılığı, genellemenin sunum ve bağlama bağımlı olduğunu gösterir; bu da sistem tasarımında bağlam eşlemesi ve periyodik güncelleme stratejilerinin gerekçesidir.

4. Epistemolojik Sınırlar ve Bilgi Statüsü

4.1. Olasılıksal Çıktı, Atıf ve İçerik-Form Ayrımı

Model doğruluk iddiasında bulunmaz; yalnızca veri dağılımına uygunluk gösterir. Kullanıcı çıktıyı "bilgi" olarak aldığında bu bir atıftır—model bu atfı doğrulayacak mekanizma sunmaz. Epistemoloji literatüründe bilgi için genellikle üç koşul aranır: doğruluk (truth), gerekçelendirilmiş inanç (justified belief) ve güvenilirlik (reliability) (Goldman, 1979). LLM çıktısı bu koşulların hiçbirini içsel olarak karşılamaz; çıktının bilgi statüsü kazanması, dış doğrulama mekanizmalarına bağımlıdır. İçerik ile yüzey formu ayrımı da bu bağlamda önemlidir: model anlamsal içeriği doğrudan işlemez, token dizilerinin olasılık yapısını işler.

Aşağıdaki tablo epistemolojik katmanlar ile bunlara karşılık gelen sistem düzeyi telafi mekanizmalarını özetlemektedir:

Tablo 1. Epistemolojik Katmanlar ve Sistem Düzeyi Telafi Mekanizmaları

Katman	Model İçsel Özellik	Sistem Düzeyi Telafi
Olasılık	Cross-entropy eğitim hedefi	Retrieval mekanizmaları (RAG)
Belirsizlik	Aleatorik/epistemik ayrımı yapılamaz	Kalibrasyon katmanı (temperature scaling)
Halüsinasyon	Distributional fit, gerçeklik değil	Harici doğrulayıcı / insan denetimi
Açıklanabilirlik	Nedensel zincir kurulamaz	Modüler doğrulama, yapısal çıktı

4.2. Belirsizlik, Gerekçelendirme ve Açıklanabilirlik Tasarımı

Olasılık dağılımı aleatorik ile epistemik belirsizliği ayırmaz; çıktı olasılıkları "bilmiyorum" sinyali olarak güvenilir biçimde kullanılamaz (Kadavath ve ark., 2022). Kalibrasyon araştırmaları, modelin token başına ürettiği olasılık değerlerinin gerçek doğruluk olasılığına ne ölçüde karşılık geldiğini inceler; bu değerlerin doğrudan karar girişi olarak kullanılması yanıltıcı olabilir (Guo ve ark., 2017). Bu özellikler, otonom veya karar-destek ortamlarında LLM çıktısının nasıl konumlandırılacağını—ham veri mi, yoksa insan veya üst sistem tarafından olasılık atamasıyla birleştirilecek girdi mi—belirleyen tasarım ölçütleridir. Gerekçelendirme ve chain-of-thought çıktısının nedensel ya da mantıksal rolü belirsizdir; tutarlılık doğruluk için gerekli ama yeterli değildir (Wei ve ark., 2022). Bu da açıklanabilirlik ve hesap verebilirlik gereksinimlerinin uygulama katmanında nasıl karşılanacağının tasarım konusu olduğunu gösterir.

5. Karar Desteği ve Otonom Sistemlerde Tasarım

5.1. Halüsinasyon, Görev-Uyumlu Kullanım Alanı ve Mimari Güvenceler

Halüsinasyon, modelin hedefinin "gerçeğe uygunluk" değil "veri dağılımına uygunluk" olmasının doğal sonucudur—tasarım özelliği, arıza değil (Ji ve ark., 2023). Bu özellik, hangi görevlerde LLM çıktısının doğrudan eyleme dönüştürülebileceğini, hangi görevlerde ise retrieval, doğrulama modülleri veya insan-in-the-loop ile sınırlandırılması gerektiğini belirler. Hukuki belge üretiminde bu sınır özellikle kritiktir: LLM'lerin mahkeme içtihatlarında var olmayan atıflar ürettiği belgelenmiştir (Magesh ve ark., 2024). Bu tür vakalar, yüksek riskli alanlarda harici doğrulayıcı katmanının tasarım zorunluluğu olduğunu somutlaştırmaktadır. Otonom sistemlerde model çıktısı eyleme dönüşebilir; otomasyon piramidi—insanın loop'ta, üzerinde veya dışında olması—tasarım seçimidir.

5.2. Tekrarlanabilirlik, Kalibrasyon ve Sistem Düzeyi Etkinlik

Tekrarlanabilirlik ve doğruluk bağlama bağımlıdır; görev-spesifik risk profili ve kabul edilebilir eşikler tasarımın parçası olarak tanımlandığında görev-uyumlu kullanım alanı netleşir. LLM çıktısı olasılık tahminlerini açık sunmuyorsa, karar verici çıktıyı ham veri alıp kendi olasılık atamasıyla birleştirir—bu birleştirme, uygun mimari seçimlerle yönetilir. Retrieval, doğrulama modülleri ve insan-in-the-loop, kuramsal özellikleri değiştirmez; ancak bu özelliklerin tanımladığı sınırlar içinde etkinliği artırır ve kullanım bağlamını genişletir. Otonom araç sistemlerinde de benzer bir yaklaşım gözlemlenmektedir: LLM, algı ve planlama bileşenlerinin çıktısını yorumlamak için kullanılmakta, ancak kritik güvenlik kararları kural tabanlı doğrulama katmanlarıyla denetlenmektedir (Wen ve ark., 2023).

6. Açıklanabilirlik ve Agent Mimarileri

6.1. Açıklama, Doğrulama ve Hesap Verebilirlik

"Neden bu cevap?" sorusu LLM'de nedensel zincir veya kural tabanlı gerekçe ile yanıtlanmaz; karar milyarlarca parametrenin etkileşimidir. Global açıklama pratikte ulaşılamaz; lokal açıklama kısmen sağlanabilir (attention ağırlıkları, integrated gradients vb.) ancak nedensel yanıt vermez (Ribeiro ve ark., 2016; Lipton, 2018). Çıktının formal veya mantıksal doğrulanması birçok karar alanında istenir; doğal dil olduğu için otomatik doğrulama ek bileşen (yapısal forma çevirme, modüler doğrulama katmanı) gerektirir. "Model ne dedi?" yanıtlanabilir; "modelin dediği geçerli mi?" sorusu ise sistem tasarımında insan veya doğrulama modülüyle konumlandırılır. Bu ayrım, tasarım güvencelerinin seçiminde merkezi rol oynar.

6.2. Agent Tasarımı ve Sınırları Sarmalayan Katmanlar

Agent mimarilerinde LLM planlama, araç kullanımı ve çok adımlı görevlerde merkezi bileşen olarak kullanılır (Yao ve ark., 2023; Shinn ve ark., 2023). Uzun zincirlerde tutarlılık ve dış dünya geri bildirimine tepki, tasarımın dikkate aldığı özelliklerdir; hata yayılımı ve araç yan etkilerinin temsili bu özelliklerle sınırlıdır. Dilsel temsil ile dünya durumları arasındaki eşleme model tarafından garanti edilmez; bu eşleme, plan doğrulayıcı, insan denetimi, modüler kontrol gibi tasarım ve doğrulama katmanlarıyla kısmen telafi edilir. Agent mimarisi, LLM'in tanımladığı sınırları sarmalayan katman olarak tasarlanır; LLM sınırsız karar verici olarak değil, görev-uyumlu kullanım alanı içinde konumlandırılarak etkinlik artırılır.

7. Sonuç ve Tasarım Ölçütleri

Belirlenen özellikler tek sonuca indirgenemez; farklı katmanlar farklı tür kullanım sınırları tanımlar. Bu sınırlar aşılacak kısıtlar değil, seçilen temsil ve öğrenme paradigmasının tanımlayıcı sonuçlarıdır; dağıtım bağlamına uygun tasarım kararları bu sınırlara göre alınır. Hesaplama varsayımları modelin neyi temsil edebileceğini yapısal olarak koşullandırır; genelleme ve dağılımsal özellikler başarının hangi eğitim–test ve görev bağlamlarında geçerli olduğunu belirler; epistemolojik özellikler çıktının bilgi statüsünün atıftan ibaret olduğunu ve belirsizliğin uygulama katmanında nasıl yönetileceğini gösterir.

Otonom karar ve karar-destek ortamlarında güvenilirlik, bu özelliklerin tanımladığı sınırlar içinde bir tasarım meselesidir; halüsinasyon ve tutarsızlık arıza değil, hangi mimari önlemlerin devreye alınacağının gerekçesidir. Açıklanabilirlik ve agent sistemleri, sınırların pratikte nasıl yönetileceği ve hangi bağlamlarda LLM'in uygun bileşen olarak konumlandırılacağı sorusunu gündeme getirir.

"LLM'ler güvenilir mi?" sorusu, "hangi varsayımlar altında, hangi görevlerde, hangi doğrulama ve denetim katmanlarıyla görev-uyumlu kullanım alanı içinde etkindir?" biçiminde çerçevelendiğinde anlamlı yanıt verir. Teknik özellikler yasaklama veya sınırsız kabul ikilemine indirgenemez; sınırların açık ifadesi, risk tabanlı düzenleme ve sorumlu dağıtım için ön koşuldur ve bağlama bağımlı etkinliğin retrieval, insan denetimi ve modüler doğrulama gibi mimari önlemlerle nasıl desteklenebileceğinin temelini oluşturur.

Kaynakça

Bender, E. M., Gebru, T., McMillan-Major, A., ve Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of FAccT 2021, 610–623.

Brown, T., Mann, B., Ryder, N., ve ark. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Chomsky, N. (1965). Aspects of the Theory of Syntax. MIT Press.

Goldman, A. (1979). What is justified belief? In G. Pappas (Ed.), Justification and Knowledge. Reidel.

Guo, C., Pleiss, G., Sun, Y., ve Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of ICML 2017, 1321–1330.

Hoffmann, J., Borgeaud, S., Mensch, A., ve ark. (2022). Training compute-optimal large language models. arXiv:2203.15556.

Ji, Z., Lee, N., Frieske, R., ve ark. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1–38.

Kadavath, S., Conerly, T., Askell, A., ve ark. (2022). Language models (mostly) know what they know. arXiv:2207.05221.

Kaplan, J., McCandlish, S., Henighan, T., ve ark. (2020). Scaling laws for neural language models. arXiv:2001.08361.

Lewis, P., Perez, E., Piktus, A., ve ark. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31–57.

Magesh, V., Surani, F., Dahl, M., ve ark. (2024). Hallucination-free? Assessing the reliability of leading AI legal research tools. arXiv:2405.20362.

Ribeiro, M. T., Singh, S., ve Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of KDD 2016, 1135–1144.

Shinn, N., Cassano, F., Berman, E., ve ark. (2023). Reflexion: Language agents with verbal reinforcement learning. arXiv:2303.11366.

Singhal, K., Azizi, S., Tu, T., ve ark. (2023). Large language models encode clinical knowledge. Nature, 620, 172–180.

Vaswani, A., Shazeer, N., Parmar, N., ve ark. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Wei, J., Wang, X., Schuurmans, D., ve ark. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35.

Wen, L., Yang, X., Fu, D., ve ark. (2023). Road GPT: Unifying large language model and autonomous driving. arXiv:2311.10813.

Yao, S., Zhao, J., Yu, D., ve ark. (2023). ReAct: Synergizing reasoning and acting in language models. Proceedings of ICLR 2023.

Zhu, C., Chen, A., Shen, T., ve ark. (2020). Modifying memories in transformer models. arXiv:2012.00363.