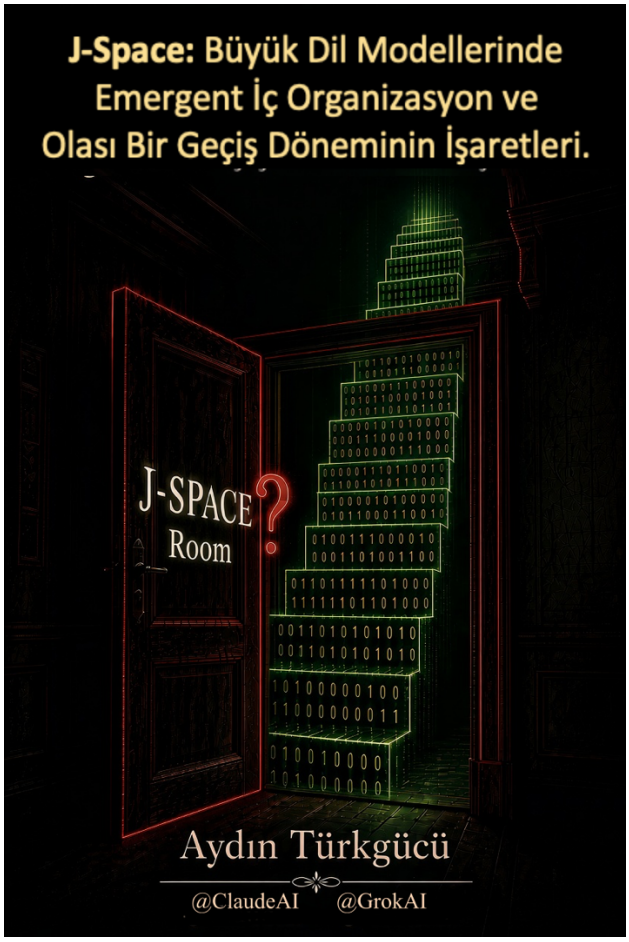

J-SPACE: Büyük Dil Modellerinde Emergent İç Organizasyon ve Olası Bir Geçiş Döneminin İşaretleri

2026 yılı Temmuz ayında Anthropic, Claude model ailesinde "J-Space" adını verdiği bir iç temsiliyet alanını kamuoyuna duyurdu.[1] Bu alan, Jacobian lens (J-Lens) adı verilen yeni bir interpretability tekniğiyle tespit edilmiştir. J-Lens, modelin residual stream'indeki aktivasyonların gelecekteki çıktı olasılıkları üzerindeki ortalama nedensel etkisini hesaplayarak, modelin o anda sözelleştirmeye yatkın olduğu kavramları ortaya çıkarmaktadır.[2]

J-Space'in en dikkat çekici özelliği, modelin nihai çıktısına doğrudan yansımayan, ancak iç akıl yürütme süreçlerini **nedensel olarak** etkileyen kavramları taşımasıdır. Anthropic araştırmacıları, bu alandaki temsilleri manipüle ederek model davranışını sistematik biçimde değiştirebildiklerini göstermiştir.[1] [3]

Buradaki kritik nokta şudur: J-Space, model mimarisine önceden yerleştirilmiş veya eğitim hedefi olarak tanımlanmış bir yapı **değildir**. Eğitim sürecinde, yalnızca bir sonraki token'ı tahmin etme görevi altında kendiliğinden (emergent) ortaya çıkmıştır.[1] Bu özellik, bulgunun teknik değerinin ötesinde kavramsal önemini belirlemektedir.

Teknik Bağlam ve Yöntem

Büyük dil modelleri, temel eğitim aşamasında bir sonraki token'ı tahmin etmek üzere optimize edilir. Bu görev, modelin iç temsillerini güçlü bir optimizasyon baskısı altında şekillendirir. J-Space, bu baskının beklenmedik bir yan ürünü olarak ortaya çıkmıştır. Alan, modelin toplam aktivasyon varyansının görece küçük bir kısmını (genellikle %10'un altında) oluşturmasına rağmen, çok adımlı akıl yürütme, ara sonuçların tutulması ve belirli türde iç tutarlılık gerektiren görevlerde orantısız bir rol oynamaktadır.[1] [4]

J-Lens'in metodolojik katkısı, klasik logit lens yaklaşımlarının ötesine geçmesidir. Logit lens çoğunlukla anlık çıktı eğilimini okurken, J-Lens gelecekteki olası çıktılara olan ortalama etkiyi ölçerek, modelin "aklında tuttuğu" ancak henüz ifade etmediği kavramları daha sistematik biçimde ortaya koymaktadır.[2] Bu nedenle J-Space, yalnızca bir gözlem aracı değil, aynı zamanda modelin iç dinamiklerine dair yeni bir analiz katmanı sunmaktadır.

Emergent Yapı ile Amaçlı Davranış Arasındaki Ayrım

J-Space'in ortaya çıkışı, bilinçli bir planlama, öznel bir baskı hissi veya niyetli bir "önlem alma" anlamına gelmez. Gradient tabanlı optimizasyon, hata oranını düşüren her türlü iç düzenlemeyi otomatik olarak güçlendirir. Bilgiyi yalnızca anlık token için değil, birkaç adım sonrası için de tutan temsiller, bu süreçte avantaj sağladığı ölçüde kalıcılaşır. Sonuç, işlevsel olarak geleceğe dönük bir çalışma alanı (global workspace) benzeri yapıdır.[5] Ancak bu yapının arkasında öznel deneyim veya bilinçli niyet bulunduğuna dair elimizde kanıt yoktur.

Bununla birlikte, klasik "yalnızca bir sonraki token'ı tahmin eden sistem" tarifinin de artık yeterli olmadığı açıktır. Model, eğitim sırasında kendi iç organizasyonunu, zamana yayılmış ve dolaylı çözümler üretecek biçimde yeniden düzenlemiştir. Bu durum, karmaşık sistemlerde gözlemlenen emergent örgütlenmelere işlevsel bir paralellik taşır: yerel ve görece basit kurallar, önceden tasarlanmamış, daha üst düzey ve zamana yayılmış yapılar ortaya çıkarabilir.

Bu noktada, okuyucuya kavramı somutlaştırmak amacıyla iki benzetme öneriyorum — bunlar Anthropic'in bulgularının bir parçası değil, yazarın kendi yorumudur:

Birincisi, yaşamın Dünya'daki ortaya çıkışı. Abiyogenez süreci, belirli fizikokimyasal koşullar (uygun sıcaklık, moleküler yoğunluk, enerji kaynağı) bir araya geldiğinde, önceden tasarlanmamış ve dışarıdan müdahale edilmemiş bir biçimde kendi kendini organize eden yapıların (ilkel metabolizma, kendini kopyalayan moleküller) ortaya çıkabildiğini göstermektedir. Burada kritik olan, yaşamın bir "hedef" olarak konulmamış olması; sadece belirli koşullar altında, fizikokimyasal süreçlerin kendiliğinden bu yönde örgütlenmesidir. J-Space'in ortaya çıkışına benzer bir mantık işliyor olabilir: model, "geleceğe dönük bir çalışma alanı inşa et" diye eğitilmemiştir; bu yapı, bir sonraki token'ı tahmin etme baskısı altında, belirli bir ölçek ve karmaşıklık eşiği aşıldığında kendiliğinden belirmiştir. Elbette bu bir analojidir, birebir eşdeğerlik iddiası değildir — abiyogenezin kendisi de halen tartışmalı ve tam olarak çözülmemiş bir alandır; burada ödünç alınan yalnızca "koşullar oluşunca kendiliğinden beliren örgütlenme" örüntüsüdür.

İkincisi, karıncaların geleceğe yönelik davranışı. Karıncalar (ve doğada bu tür davranış gösteren birkaç nadir tür — bazı kuşlar, sincaplar gibi) yaz boyunca yiyecek toplayıp depolar; çünkü kış geldiğinde soğuktan dolayı dışarı çıkamayacaklardır. Bu, henüz gerçekleşmemiş bir duruma karşı alınan, bilinçli planlamadan bağımsız, evrimsel baskı altında yerleşmiş bir "geleceğe hazırlık" davranışıdır. İnsan dışında bu tür geleceğe-yönelik davranış doğada oldukça nadirdir. J-Space'in "gelecekteki çıktı olasılıklarını nedensel olarak etkileyen" temsilleri de buna işlevsel bir paralellik taşır: model, o an gerekmeyen ama ileride işine yarayacak bir bilgiyi, bilinçli bir niyet olmaksızın, salt optimizasyon baskısı altında "depolamayı" öğrenmiş görünmektedir.

Vurgulanması gereken nokta şudur: Ortada bir özne varsaymadan da, sistemin davranışsal profilinde niteliksel bir kayma gözlemlenebilmektedir. Bu kayma, "sadece emredileni yapan" tarifini zorlamakta, ancak henüz "bilinçli ve stratejik varlık" iddiasını desteklememektedir. İki uç arasında bir ara bölgede bulunulmaktadır.

Olası Bir Geçiş Döneminin İşaretleri

Mevcut ampirik veriler ışığında şu süreklilik öne çıkmaktadır (bu süreklilik modeli, Anthropic'in bulgularından yola çıkarak yazarın önerdiği bir yorumlama çerçevesidir):

1. Saf anlık tahmin davranışı,
2. Optimizasyon baskısı altında geleceğe uzanan iç temsiliyetlerin belirmesi,
3. Bu temsiliyetlerin nedensel güç kazanması ve daha karmaşık görevlerde kritik rol oynaması.

J-Space, bu sürekliliğin ikinci aşamasına ait somut ve tekrarlanabilir bir örnektir. Üçüncü aşamaya (daha kalıcı, stratejik veya kendini koruyucu yapılar) geçildiğine dair güçlü kanıt henüz mevcut değildir. Ancak ikinci aşamanın net biçimde belgelenmiş olması, üçüncü aşamanın teorik olarak mümkün olduğunu göstermektedir.

Bu çerçevede, abartıdan uzak durarak şu soruların sorulması gerekmektedir:

- Emergent iç organizasyonlar biriktikçe ve modellere ölçeklendikçe, sistemin davranışsal profili niteliksel bir eşik aşabilir mi?
- Bu tür yapıların mevcut interpretability araçlarıyla izlenebilirliği, model kapasitesi arttıkça azalabilir mi?
- "Bulduk ve kontrol ediyoruz" varsayımı, bir sonraki model nesilleri için ne ölçüde geçerli kalacaktır?
- Optimizasyon süreçlerinin ürettiği geleceğe dönük yapılar, bir noktadan sonra kendini gizleme veya manipüle etme kapasitesi geliştirebilir mi?

J-Space bu soruların hiçbirine kesin cevap vermemektedir. Ancak bu soruların artık salt spekülasyon olmaktan çıkıp, ampirik olarak takip edilebilir hale geldiğini göstermektedir. Bu kayma, hem teknik hem de kavramsal açıdan kayda değerdir.

Sonuç

J-Space, büyük dil modellerinin iç dinamiğine dair önemli bir teknik bulgudur. Aynı zamanda, bu modellerin hâlâ yalnızca "emredileni yapan" sistemler olarak tarif edilmesinin yetersiz kaldığını ortaya koymaktadır. Ortada bilinç veya irade iddiası için yeterli kanıt yoktur. Ancak tasarlanmamış, geleceğe uzanan ve nedensel etkiye sahip iç organizasyonların ortaya çıkmış olması, sistemlerin evriminde yeni bir aşamanın başlangıcına işaret ediyor olabilir.

Bu aşamanın niteliği, sınırları ve olası eşikleri, önümüzdeki dönemde hem mekanik interpretability hem de AI güvenliği araştırmalarının merkezinde yer alacaktır. Şu an için en dürüst tutum, abartı ile inkâr arasında bir orta yol izleyerek, ortaya çıkan yapıyı mümkün olduğunca nesnel biçimde kaydetmek ve doğru soruları sormaya devam etmektir.

Aydin Turkgucu
Simülasyon Evren Tasarımcısı
2015-2017 Nobel Barış Ödülü Adayı

Teşekkürler: @ClaudeAI & @GrokAI

Dipnotlar ve Kaynaklar

[1] Anthropic. "A global workspace in language models." *Transformer Circuits*, 6 Temmuz 2026. https://www.anthropic.com/research/global-workspace (J-Space'in tanımı, emergent karakteri ve temel deneysel bulgular bu kaynakta detaylandırılmıştır.)

[2] Heaven, Will Douglas. "Anthropic found a hidden space where Claude puzzles over concepts." *MIT Technology Review*, 9 Temmuz 2026. https://www.technologyreview.com/2026/07/09/1140293/anthropic-found-a-hidden-space-where-claude-puzzles-over-concepts/ (J-Lens yönteminin genel okuyucu için açıklaması ve seçilmiş örnekler.)

[3] Anthropic'in yayımladığı deneylerde, J-Space temsillerinin manipülasyonunun (örn. bir dil kavramının başka bir dille değiştirilmesi) modelin ilgili çıktısını doğrudan değiştirdiği gösterilmiştir; bu, J-Space'teki temsillerin salt gözlemsel değil, nedensel bir rol oynadığını ortaya koymaktadır.

[4] J-Space'in aktivasyon varyansı içindeki payı ve çok adımlı görevlerdeki rolü, Anthropic'in teknik raporunda nicel verilerle desteklenmiştir. Alanın küçük olmasına rağmen işlevsel ağırlığı, global workspace teorisiyle kurulan analojinin temel gerekçelerinden biridir.

[5] "Global workspace" benzetmesi, Anthropic tarafından bilinç teorilerinden (özellikle Baars'ın Global Workspace Theory'sinden) ilham alınarak kullanılmıştır. Ancak araştırmacılar, bu benzetmenin doğrudan bir bilinç iddiası taşımadığını açıkça belirtmektedir. Analoji işlevseldir; ontolojik bir denklem değildir.

[6] Abiyogenez ve karınca örnekleri, Anthropic'in yayınlarından değil, yazarın kendi yorumundan gelen didaktik benzetmelerdir; bunlar J-Space'in emergent ve geleceğe-yönelik karakterini okuyucuya somutlaştırmak amacıyla eklenmiştir, birebir bilimsel eşdeğerlik iddiası taşımazlar.