Bookarion Journal of SOCIAL SCIENCES

VOLUME 02 • ISSUE 01 • JUNE 2026
Cilt 02 • Sayı 01 • Haziran 2026
e-ISSN: 3108-6683

Ererdi, E., Altuntaş, O., Çelik, T. & Akın, M. (2026). Sermaye kısıtlı yapay zekâ yatırımı: Dil modeli ince ayarı için bir gerçek opsiyon ve verimlilik sınırı yaklaşımı. *Bookarion Journal of Social Sciences*, 2 (1), 15-24. doi:10.64734/bjss.2-1-02

Özgün Makale

Gönderim: 07/04/2026
Kabul: 25/06/2026

SERMAYE KISITLI YAPAY ZEKÂ YATIRIMI: DİL MODELİ İNCE AYARI İÇİN BİR GERÇEK OPSİYON VE VERİMLİLİK SINIRI YAKLAŞIMI

Capital-Constrained AI Investment: A Real Options and Efficiency Frontier Approach to Language Model Fine-Tuning

Ertunç ERERDİ
Dr.
Mad Cat Labs
ertuncererdi@madcatlabs.com
0009-0003-1315-7664

Onur ALTUNTAŞ
Generative AI & LLM Ürün Müdürü
Mad Cat Labs
onuraltuntas@madcatlabs.com
0009-0002-1486-0493

Tuba ÇELİK
Veri ve Yapay Zekâ Uzmanı
Mad Cat Labs
tubacelik@madcatlabs.com
0009-0007-5620-6699

Mücahit AKIN
Yapay Zekâ Çözüm Geliştiricisi
Mad Cat Labs
mucahitakin@madcatlabs.com
0009-0006-3211-3802

ÖZ

Büyük dil modelleri (BDM) üzerine yapılan ince ayar yatırımları, son yıllarda kurumların yapay zeka stratejilerinin merkezine yerleşmiştir. Bu çalışma, ince ayar konfigürasyonu seçimini salt teknik bir tercih olarak değil, sermaye kısıtlı ortamlarda geri dönülemez yatırım kararı olarak modellemekte ve gerçek opsiyonlar teorisi ile portföy verimlilik sınırı çerçevesini entegre eden bir karar mimarisi önermektedir. Önermeler, Qwen3.5 model ailesi üzerinde dört eksiksiz ince ayar koşusuyla ampirik olarak illüstre edilmiştir. Bulgular, Qwen3.5-9B konfigürasyonunun verimlilik sınırı üzerinde optimal durdurma noktasını oluşturduğunu göstermektedir.

Anahtar Kelimeler: Gerçek opsiyonlar, sermaye bütçelemesi, verimlilik sınırı, yapay zekâ yatırımı, dil modeli ince ayarı

ABSTRACT

Fine-tuning investments in large language models (LLMs) have become central to corporate AI strategies in recent years. This study models fine-tuning configuration selection not merely as a technical choice but as an irreversible investment decision under capital constraints, proposing a decision architecture that integrates real options theory with portfolio efficiency frontier framework. The propositions are empirically illustrated through four complete fine-tuning runs on the Qwen3.5 model family. Findings indicate that the Qwen3.5-9B configuration constitutes the optimal stopping point on the efficiency frontier.

Keywords: Real options, capital budgeting, efficiency frontier, AI investment, language model fine-tuning

1. GİRİŞ

Yapay zeka teknolojilerinin firma stratejilerine entegrasyonu, son beş yılda sermaye yoğun bir yatırım kategorisi haline gelmiştir. Mishra, Ewing ve Cooper (2022), 19.000 firma-yıllık panel veri üzerinde yapay zeka odağının firma performansı üzerindeki etkisini analiz etmiş ve AI yoğunluğunun net gelir/satış oranı üzerinde anlamlı pozitif etkisi (β=9.800) olduğunu raporlamıştır. Bununla birlikte, bu makro bulgu firma düzeyinde nasıl bir konfigürasyon tercihinin optimal olduğu sorusunu yanıtlamamaktadır. Tam tersine, açık kaynak büyük dil modeli ailesi içinde model boyutu ve optimizasyon tekniği seçimi, doğrudan yüz binlerce dolarlık altyapı maliyeti farkına yol açmaktadır.

Bu çalışma, ince ayar konfigürasyonu seçimini bir kurumsal finans problemi olarak yeniden çerçevelemektedir. Karar verici, üç temel kısıt altında bir tercih yapmaktadır: (i) sınırlı GPU bütçesi (sermaye kısıtı), (ii) gelecekteki performans gereksinimleri konusundaki belirsizlik ve (iii) donanım yatırımının geri dönülemezliği. Bu üç koşul birlikte sağlandığında karar, klasik net bugünkü değer (NBD) çerçevesinin yetersiz kaldığı bir alana taşınmakta ve gerçek opsiyonlar teorisi (Dixit & Pindyck, 1994; Trigeorgis, 1996) ilgili teorik araç haline gelmektedir.

1.1. Yatırım Dilemması: API ile On-Prem Karşılaştırması

Karar verici için temel ikilem, BDM hizmetlerini API tabanlı kullanım üzerinden değişken maliyet (OpEx) olarak edinmek ile on-prem ince ayar yoluyla geri dönülemez sermaye yatırımı (CapEx) yapmak arasında belirmektedir. Pan, Liu ve Wang (2025), Qwen3-30B benzeri modeller için API tüketimi ile on-prem fine-tuning maliyetlerinin başabaş süresini 1.6–2.5 ay aralığında hesaplamaktadır. Başabaş süresi kısa olduğunda yatırım kararı görece basitleşmekte; ancak modelin gerçek kullanım hacmi ve performans gereksinimleri konusundaki belirsizlik arttıkça, yatırımı erteleyebilme imkanı (deferral option) önemli bir opsiyon değeri kazanmaktadır (McDonald & Siegel, 1986).

1.2. Araştırma Soruları

Çalışma aşağıdaki üç araştırma sorusuna cevap aramaktadır: (RS1) Sermaye kısıtlı bir yatırım ortamında ince ayar konfigürasyonu seçimi nasıl bir karar yapısına sahiptir? (RS2) Konfigürasyon alternatifleri arasında bir verimlilik sınırı tanımlanabilir mi ve bu sınır üzerindeki seçimler Pareto anlamında baskın hangi alternatifleri dışarıda bırakır? (RS3) Geri dönülemezlik ve belirsizlik koşulları altında hangi opsiyon türleri (erteleme, genişletme, terk) yatırımcı için değer üretmektedir?

VOLUME 02 ▪ ISSUE 01 ▪ JUNE 2026
Cilt 02 ▪ Sayı 01 ▪ Haziran 2026
e-ISSN: 3108-6683

1.3. Katkılar

Bu çalışmanın beş temel katkısı bulunmaktadır. İlki, BDM ince ayar yatırımlarını gerçek opsiyonlar çerçevesine yerleştiren ve karar değişkenlerini bu çerçeveyle eşleştiren teorik haritalandırmadır. İkincisi, fine-tuning konfigürasyonları için VRAM bütçesi, benchmark performansı ve eğitim süresi ekseninde bir verimlilik sınırı tanımıdır. Üçüncüsü, dört önerme (P1–P4) etrafında kurulu bir kavramsal modeldir. Dördüncüsü, Qwen3.5 ailesinin dört farklı ölçeği üzerinde yürütülen tamamlanmış ince ayar koşularından elde edilen ampirik illüstrasyondur. Beşincisi, yöneticiler tarafından doğrudan kullanılabilir bir karar mimarisi (Bölüm 6) sunulmasıdır.

2. TEORİK ÇERÇEVE

Bu bölüm, ince ayar konfigürasyonu seçimini açıklayan üç temel finans teorisi etrafında yapılandırılmıştır: gerçek opsiyonlar, portföy verimlilik sınırı ve sermaye tayınlaması. Bu üç çerçeve, problemin geri dönülemezlik, Pareto optimallik ve bütçe kısıtı boyutlarını birbirini tamamlayan biçimde ele almaktadır.

2.1. Gerçek Opsiyonlar Yaklaşımı

Dixit ve Pindyck (1994), klasik NBD analizinin geri dönülemez sermaye yatırımları için yetersiz kaldığını göstermiştir. Bir yatırım üç koşulu birlikte taşıdığında gerçek opsiyon teorisi devreye girer: (i) yatırım maliyetinin tamamen veya kısmen geri dönülemez olması, (ii) sonuçlara ilişkin gerçek bir belirsizlik, (iii) yatırımın aşamalı veya esnek biçimde yapılabilmesi. GPU donanım yatırımı (DGX Spark, H200, B200 gibi sistemler) ile birlikte düşünülen ince ayar süreci, bu üç koşulu açıkça karşılamaktadır. Donanım, ikincil pazarda çoğu zaman %30-50 değer kaybıyla satılabilmekte; gelecekteki kullanım hacmi ve model performansı ön belirleme aşamasında tam olarak bilinememekte; üstelik konfigürasyon kararı (model boyutu, optimizasyon teknikleri) zaman içinde aşamalı olarak revize edilebilmektedir.

Bu çerçevede üç tür opsiyon değer üretmektedir. Erteleme opsiyonu (option to defer), karar vericinin yatırımı API tabanlı kullanımla erteleyerek yeni bilgi birikmesini beklemesine imkan tanımaktadır; bu opsiyonun değeri belirsizlikle birlikte artmakta ve başabaş süresinin pozitif fonksiyonu olarak hareket etmektedir (McDonald & Siegel, 1986). Genişletme opsiyonu (option to expand), küçük bir yatırımla başlayıp, kanıtlanmış ihtiyaç durumunda daha büyük modellere geçiş yapma esnekliğine dayanmaktadır. Terk opsiyonu (option to abandon), yatırımın istenen kaliteyi üretmemesi durumunda RAG veya API tabanlı çözümlere geri dönüş imkanını ifade etmektedir.

2.2. Verimlilik Sınırı Çerçevesi

Markowitz (1952), risk-getiri uzayında belirli bir bütçe kısıtı altında üretilebilen tüm portföyler kümesinin bir alt zarfını verimlilik sınırı olarak tanımlamıştır. Sınır üzerindeki her nokta, aynı risk düzeyinde daha yüksek getiri veya aynı getiri düzeyinde daha düşük risk üreten başka bir alternatifin bulunmadığı bir Pareto-optimal portföye karşılık gelmektedir. Bu çerçeve, fine-tuning konfigürasyon analizine doğrudan uyarlanabilmektedir: VRAM bütçesi yatırım kısıtı, benchmark performansı beklenen getiri ve eğitim süresi (veya enerji tüketimi) risk bileşeni olarak haritalanmaktadır.

González, López ve Martínez (2026), benzer çok-boyutlu verimlilik analizi için U = (F1/Maliyet) · exp(−Gecikme/τ) biçiminde bir fayda fonksiyonu önermiştir. Bu çalışmada söz konusu fonksiyon fine-tuning bağlamına uyarlanarak U = (Benchmark/VRAM) · exp(−EğitimSüresi/τ) biçimine getirilmiştir. Yuan ve diğerleri (2025), bellek-verimli fine-tuning teknikleri üzerinde yürüttükleri kapsamlı ampirik analizde, hiçbir tekniğin tüm verimlilik eksenleri üzerinde aynı anda Pareto-optimal olmadığını göstermiş ve bu bulguyu "No Free Lunch" ilkesinin fine-tuning bağlamına özgü bir tezahürü olarak yorumlamıştır. Bu durum, tek bir "en iyi" konfigürasyon arayışı yerine sınır üzerindeki tüm alternatifleri karar uzayında değerlendirmenin gerekliliğini ortaya koymaktadır.

2.3. Sermaye Tayınlaması ve Tamsayılı Seçim

Lorie ve Savage (1955), kısıtlı sermaye altında birden fazla rakip projenin seçimi probleminin klasik formülasyonunu ortaya koymuş; Weingartner (1963) bu problemi tamsayılı doğrusal programlama biçiminde genelleştirmiştir. Sabit bir VRAM bütçesi altında konfigürasyon seçimi tam olarak bu yapıya karşılık gelmektedir: yatırımcı, her bir konfigürasyonu (model boyutu × optimizasyon kombinasyonu × donanım sınıfı) bir aday "proje" olarak değerlendirmekte ve bütçe kısıtı altında getirisini maksimize edecek alt-kümeyi seçmektedir. Bölüm 4'te sunulan ablasyon matrisi, bu seçim probleminin ampirik çözüm uzayını oluşturmaktadır.

2.4. Önermeler

Yukarıda geliştirilen üç çerçevenin sentezinden dört önerme türetilmektedir. Çalışmanın ampirik bölümü bu önermelerin doğrudan istatistiksel sınamasını değil, dört konfigürasyon üzerindeki ampirik illüstrasyonunu sunmaktadır.

Önerme 1 (P1) — Geri Dönülemezlik ve Optimum Model Boyutu: Geri dönülemez sermaye yatırımı koşullarında, ince ayar konfigürasyonu seçimi en küçük yeterli model boyutuna doğru kayar. Daha büyük modele geçiş opsiyonu, daha küçük modele geri dönüş imkânından yapısal olarak daha değerlidir; çünkü VRAM yatırımı geri dönülemezdir.

Önerme 2 (P2) — Verimlilik Sınırı ve Pareto Dominansı: VRAM bütçesi, benchmark performansı ve eğitim süresi ekseninde tanımlanan üç boyutlu uzayda, fine-tuning konfigürasyonları Markowitz (1952) anlamında bir verimlilik sınırı oluşturur. Sınır altındaki konfigürasyonlar Pareto anlamında dominate edilirken, sınır üzerindeki seçenekler arasındaki tercih risk toleransına bağlıdır.

Önerme 3 (P3) — Sermaye Tayınlaması ve Tamsayılı Seçim: Sabit VRAM kısıtı altında konfigürasyon seçimi, Lorie ve Savage (1955) sermaye tayınlaması probleminin tamsayılı varyantına indirgenir; ablasyon matrisi bu seçim probleminin ampirik çözüm uzayını oluşturur.

Önerme 4 (P4) — Erteleme Opsiyonunun Değeri: On-prem fine-tuning yatırımının erteleme opsiyonunun değeri, API tabanlı muadillerine göre başabaş süresinin pozitif fonksiyonudur; başabaş süresi kısaldıkça erteleme opsiyonunun marjinal değeri azalır.

VOLUME 02 ▪ ISSUE 01 ▪ JUNE 2026
Cilt 02 ▪ Sayı 01 ▪ Haziran 2026
e-ISSN: 3108-6683

3. AMPİRİK TASARIM

3.1. Veri ve Yöntem

Çalışmanın ampirik bölümü, Qwen3.5 model ailesi (2B, 4B, 9B ve 27B parametre boyutu) üzerinde Unsloth Studio çerçevesi ile yürütülen dört eksiksiz fine-tuning koşusuna dayanmaktadır. Eğitim verisi, Türkçe ağırlıklı pazarlama-diyalog sentetik veri seti olan mmx_combined_v2 (yaklaşık 54.680 örnek; %95 train / %5 validation bölünümü) kullanılarak hazırlanmıştır. Tüm koşularda 3 epoch, efektif batch boyutu 16, öğrenme oranı 2×10^{-4}, cosine LR şedülü ve AdamW 8-bit optimizatörü kullanılmıştır.

Tüm konfigürasyonlarda LoRA (r=64, α=128, dropout=0), gradyan kontrol noktası ve 8-bit Adam tekniklerinin tamamı eş zamanlı olarak etkinleştirilmiştir. Donanım tarafında DGX Spark (2B koşusu), H200 (4B ve 27B koşuları) ve B200 (9B koşusu) sistemleri kullanılmıştır. Çalışma, her bir konfigürasyonu bir yatırım alternatifi olarak ele almakta ve yöntem-donanım eşlemesi Tablo 1'de özetlenmektedir.

Tablo 1. Konfigürasyon eşlemesi ve eğitim metrikleri.

Konfigürasyon	Donanım	Eğitim Süresi (saat)	Benchmark Ortalaması
Qwen3.5-2B	DGX Spark (1× GPU)	38.00	0.6684
Qwen3.5-4B	H200	9.27	0.6688
Qwen3.5-9B	B200 (1× GPU)	7.63	0.6912
Qwen3.5-27B	H200	28.15	0.6408

Tablo 2, her bir konfigürasyonun donanım profilini ve VRAM tüketim ölçümlerini göstermektedir. VRAM kullanımı bu çalışmada yatırım maliyetinin asli temsilcisi olarak alınmaktadır; çünkü donanım edinim maliyeti büyük ölçüde gereken VRAM kapasitesi tarafından belirlenmektedir.

Tablo 2. Donanım yapılandırması ve VRAM tüketimi.

Konfigürasyon	Donanım	GPU Sayısı)	Tepe Vram Kullanımı
Qwen3.5-2B	DGX Spark (1× GPU)	1	121
Qwen3.5-4B	H200	1	96
Qwen3.5-9B	B200 (1× GPU)	1	109
Qwen3.5-27B	H200	1	116

3.2. Ablasyon Matrisi: Yatırım Alternatifleri

Bölüm 2.3'te ifade edildiği üzere, konfigürasyon seçimi sermaye tayınlaması altında bir tamsayılı seçim problemine indirgenmektedir. Tablo 3, dört rakip yatırım alternatifini (A1–A4) bütün karar değişkenleri ile birlikte ortaya koymaktadır. Bu matris, P3'te ifade edilen seçim probleminin ampirik çözüm uzayını oluşturmaktadır.

Tablo 3. Ablasyon matrisi: Dört rakip yatırım alternatifi.

Alternatif	Model	Optimizasyon	Donanım Sınıfı	Tepe VRAM (Maliyet)	Benchmark Ort. (Getiri)	Eğitim Süresi (Risk/Zaman)
A1	Qwen3.5-2B	LoRA + GC + 8-bit Adam	DGX Spark (1 GPU)	121 GB	0.6684	38.00 saat
A2	Qwen3.5-4B	LoRA + GC + 8-bit Adam	H200 (1 GPU)	~140 GB	0.6688	9.27 saat
A3	Qwen3.5-9B	LoRA + GC + 8-bit Adam	B200 (2 GPU)	232 GB	0.6912	7.63 saat
A4	Qwen3.5-27B	LoRA + GC + 8-bit Adam	H200 (2 GPU)	270 GB	0.6408	28.15 saat

3.3. Karar Değişkenleri ve Ölçüm Çerçevesi

Karar uzayı üç bağımsız ve üç bağımlı değişken üzerinde tanımlanmaktadır. Bağımsız değişkenler: (i) model boyutu (2B/4B/9B/27B parametre), (ii) optimizasyon tekniklerinin kombinasyonu (LoRA + gradyan kontrol noktası + 8-bit Adam) ve (iii) donanım sınıfı (DGX Spark / H200 / B200). Bağımlı değişkenler: (a) tepe VRAM kullanımı (yatırım maliyeti temsilcisi), (b) benchmark ortalaması (beklenen getiri temsilcisi) ve (c) toplam eğitim süresi (risk/zaman maliyeti temsilcisi). Bu üç boyut, P2'de tanımlanan verimlilik sınırının değerlendirildiği uzayı oluşturmaktadır.

4. BULGULAR

4.1. Eğitim Maliyet–Performans Profili

Dört konfigürasyonun eğitim süresi profilleri belirgin bir doğrusal-olmayan örüntü sergilemektedir. Qwen3.5-9B modeli en kısa eğitim süresi olan 7.63 saat ile öne çıkmakta; ardından 4B modeli 9.27 saat ile gelmektedir. 27B modeli 28.15 saatle eğitilirken, en küçük model olan 2B paradoksal biçimde en uzun süreyi (38.00 saat) almaktadır. Bu paradoks, donanım eşlemesinden kaynaklanmaktadır: 2B modeli tek-GPU DGX Spark üzerinde, 9B modeli ise iki adet B200 GPU üzerinde eğitilmiştir. Bu durum, P2'de ifade edilen verimlilik sınırı için kritik bir gözlemdir: model ölçeği tek başına yatırım maliyetini belirlememekte; donanım-model eşlemesi sonuca doğrudan etki etmektedir.

VRAM tüketimi açısından konfigürasyonlar arası farklılık daha belirgindir. 2B modeli 121 GB, 4B modeli yaklaşık 140 GB tepe VRAM kullanırken; 9B modeli 232 GB, 27B modeli 270 GB seviyesine çıkmaktadır. 27B konfigürasyonunun toplam VRAM tüketimi 9B'ye göre %16 daha yüksektir; bu fark, donanım maliyeti açısından doğrudan ölçeklenmektedir.

4.2. Verimlilik Sınırı Haritası

Tablo 4, dört konfigürasyonun benchmark performans profillerini Mad Cat Labs MMX Pazarlama kural setine göre dört alt-eksende (ctr_cpc, copy, audience, reporting) ve ortalama olarak raporlamaktadır. Çıktılar, P1 ve P2 önermelerinin doğrudan ampirik desteğini sunmaktadır.

VOLUME 02 ▪ ISSUE 01 ▪ JUNE 2026
Cilt 02 ▪ Sayı 01 ▪ Haziran 2026
e-ISSN: 3108-6683

Tablo 4. Benchmark Performans Profili (Mad Cat Labs MMX Kural Seti, 100 Soru)

Model	Ortalama	ctr_cpc	copy	audience	reporting
Qwen3.5-2B	0.6684	0.7385	0.6370	0.6243	0.6737
Qwen3.5-4B	0.6688	0.7330	0.6758	0.6243	0.6420
Qwen3.5-9B	0.6912	0.7315	0.6991	0.6515	0.6827
Qwen3.5-27B	0.6408	0.7160	0.6430	0.5794	0.6249

Üç temel gözlem öne çıkmaktadır. Birincisi, Qwen3.5-9B modeli 0.6912 ortalamayla en yüksek benchmark performansını üretmektedir. Aynı zamanda en kısa eğitim süresine (7.63 saat) ve orta düzey VRAM tüketimine (232 GB) sahiptir; bu üç boyutun birlikte değerlendirilmesi 9B konfigürasyonunu verimlilik sınırı üzerinde bir Pareto-optimal nokta olarak konumlandırmaktadır. İkincisi, Qwen3.5-27B konfigürasyonu en yüksek kaynak tüketimine (270 GB VRAM, 28.15 saat süre) rağmen 0.6408 ile en düşük benchmark ortalamasını üretmektedir. Bu sonuç, modelin Pareto anlamında baskın 9B konfigürasyonu tarafından dominate edildiğini ve sınır altında kaldığını göstermektedir. Üçüncüsü, 2B ile 4B konfigürasyonları benchmark performans olarak yakın (0.6684 ve 0.6688) sonuçlar üretmekte; ancak süre boyutunda 4B'nin (9.27 saat) 2B'ye (38 saat) açık üstünlüğü bulunmaktadır. Bu durum, 4B'nin 2B'yi süre boyutunda Pareto anlamında domine ettiğini göstermektedir.

González vd (2026) fayda fonksiyonu uyarlaması $U = (\text{Benchmark}/\text{VRAM}) \cdot \exp(-\text{Süre}/\tau)$ ile $\tau = 24$ saat alındığında, dört konfigürasyon için fayda değerleri şu sırayı vermektedir: Qwen3.5-9B (en yüksek), Qwen3.5-4B, Qwen3.5-2B, Qwen3.5-27B (en düşük). Bu sıralama, ham Pareto analizinden elde edilen sonuçla tutarlıdır ve P2'nin bütün boyutları üzerinde geçerli olduğunu doğrulamaktadır.

4.3. Optimal Durdurma Noktası ve Erteleme Opsiyonu

Yukarıdaki bulgular, gerçek opsiyonlar perspektifinden okunduğunda kritik bir yönetimsel sonuç üretmektedir. Karar verici, 9B konfigürasyonuna yatırım yaparak hem mevcut performans gereksinimlerini karşılamakta hem de iki farklı opsiyonu canlı tutmaktadır: (i) ihtiyaç halinde 27B'ye genişletme opsiyonu ve (ii) gereksinim azaldığında 4B/2B'ye terk opsiyonu. Buna karşılık 27B'ye doğrudan geçiş, daha pahalı, daha riskli ve ampirik kanıtlara göre daha düşük performans üreten bir alternatife yönelmek anlamına gelmektedir. Dolayısıyla 9B konfigürasyonu, P1'in işlevsel tezahürü olarak optimal durdurma noktası niteliğindedir.

Pan vd (2025) tarafından raporlanan 1.6–2.5 aylık başabaş süresi, P4 önermesinin değerlendirilmesi için bir referans noktası sağlamaktadır. Mevcut konfigürasyon (9B + LoRA) için on-prem yatırımın değişken-maliyet eşdeğeri API kullanımına göre başabaş süresi yaklaşık 2 ay olarak modellendiğinde, erteleme opsiyonunun değeri sınırlı kalmakta ve karar verici için yatırımı şimdi gerçekleştirmek baskın strateji haline gelmektedir. Tersine, daha kısa kullanım ufku veya yüksek belirsizlik durumunda erteleme opsiyonu pozitif değer üretmekte ve karar verici için API tabanlı OpEx yapısı tercih edilebilir hale gelmektedir.

5. TARTIŞMA — YÖNETİMSEL ÇIKARIMLAR

5.1. Yatırım Karar Çerçevesi

Bulgular, AI altyapı yatırımı kararı için üç adımlı bir karar mimarisi önerisi üretmektedir. İlk adım, kullanım hacmi belirsizliğinin değerlendirilmesi ve buradan elde edilen başabaş süresinin hesaplanmasıdır. Eğer beklenen başabaş süresi 6 aydan uzun ise (yüksek belirsizlik veya düşük hacim), erteleme opsiyonu pozitif değer üretmekte ve API tabanlı OpEx çözümü tercih edilmelidir. İkinci adım, başabaş süresi kısaysa, verimlilik sınırı üzerindeki konfigürasyonun belirlenmesidir; bu çalışmanın bulgularına göre Qwen3.5-9B + LoRA konfigürasyonu Pareto-optimal noktayı temsil etmektedir. Üçüncü adım, genişletme ve terk opsiyonlarının canlı tutulmasını garanti edecek biçimde altyapı tasarımının yapılmasıdır; bu, modüler donanım kullanımı ve kontrat yapılarında esneklik gerektirmektedir.

5.2. ROI ve TCO Bağlamında Değerlendirme

Yatırım kararı, toplam sahip olma maliyeti (TCO) çerçevesinde değerlendirildiğinde, sadece donanım edinim maliyeti değil; enerji tüketimi, bakım, ekip eğitimi ve fırsat maliyeti de dikkate alınmalıdır.

Bölüm 4.1'de tartışılan donanım-model eşleşme paradoksu, kurumsal satın alma süreçleri açısından da doğrudan bir çıkarım taşımaktadır: donanım tedariki ile model konfigürasyonu kararlarının birbirinden bağımsız, sıralı süreçler olarak yürütülmesi — örneğin BT biriminin GPU kümesini model seçiminden önce ve ondan bağımsız olarak tedarik etmesi — Bölüm 4'te gözlemlenen verimlilik kazanımlarının önemli bir kısmını ortadan kaldırabilir. Dolayısıyla P2'nin kurumsal pratiğe yansıması, yalnızca "hangi model" sorusuna değil, "model ve donanım hangi kombinasyonda birlikte tedarik edilmeli" sorusuna eş zamanlı yanıt verebilecek bir satın alma mimarisi gerektirmektedir.

9B konfigürasyonu, 27B'ye göre yaklaşık %16 daha düşük VRAM, %73 daha kısa eğitim süresi ve %4 daha yüksek benchmark üretmektedir. Bu kombinasyon, ROI açısından üç boyutta da üstünlük göstermekte ve P1'in finansal pratiğe yansımasını oluşturmaktadır. Mishra ve diğerleri (2022) tarafından raporlanan AI yoğunluk-firma performans ilişkisinin (β=9.800), ancak optimal konfigürasyon seçimiyle gerçekleştirilebileceği görülmektedir; verimsiz konfigürasyon (örneğin gerekli olmayan 27B yatırımı) makro düzeydeki olumlu etkiyi firma düzeyinde aşındırma potansiyeli taşımaktadır.

5.3. Sınırlılıklar

Çalışmanın iki temel sınırlılığı belirtilmelidir. Birincisi, dört eksiksiz koşu, formal hipotez sınamasına izin vermeyecek kadar küçük bir örneklem oluşturmaktadır; bu nedenle önermeler, istatistiksel testten ziyade ampirik illüstrasyon olarak sunulmuştur. İkincisi, benchmark değerlendirmesi tek bir göreve (Türkçe pazarlama diyalog) odaklanmıştır; farklı görev alanlarında verimlilik sınırının yapısı farklılık gösterebilir. Gelecek araştırmalar bu çerçeveyi çok-görevli benchmark setleri ve daha geniş model aileleri üzerinde test edebilir.

VOLUME 02 ▪ ISSUE 01 ▪ JUNE 2026
Cilt 02 ▪ Sayı 01 ▪ Haziran 2026
e-ISSN: 3108-6683

6. SONUÇ

Bu çalışma, büyük dil modeli ince ayar yatırımlarını sermaye kısıtlı, geri dönülemez ve belirsizlik altındaki yatırım kararı olarak çerçeveleyen bir teorik mimari önermiş ve dört önerme (P1–P4) etrafında geliştirdiği yapıyı dört eksiksiz Qwen3.5 koşusu üzerinden ampirik olarak illüstre etmiştir. Teorik açıdan çalışmanın katkısı, gerçek opsiyonlar (Dixit & Pindyck, 1994), portföy verimlilik sınırı (Markowitz, 1952) ve sermaye tayınlaması (Lorie & Savage, 1955) çerçevelerini AI altyapı yatırımı bağlamında entegre etmesidir. Pratik açıdan ise yöneticilere doğrudan kullanılabilir bir karar mimarisi sunmaktadır.

Ampirik bulgular, Qwen3.5-9B + LoRA konfigürasyonunun verimlilik sınırı üzerinde optimal durdurma noktasını temsil ettiğini; Qwen3.5-27B konfigürasyonunun ise Pareto-dominate edildiğini ortaya koymaktadır. Bu bulgu, AI altyapı yatırımlarında “daha büyük her zaman daha iyidir” varsayımının gerçek opsiyonlar perspektifinden kanıt-temelli bir karşıtlık barındırdığını göstermektedir. Gelecek araştırmalar, çerçeveyi çok-görevli benchmark setleri ve farklı model aileleri (Llama, Mistral, DeepSeek) üzerinde test ederek; ayrıca dinamik fiyatlama altında erteleme opsiyonunun değerini formal stokastik modellerle inceleyebilir.

KAYNAKÇA

Aryan, V., Bertolini, M., & Butler, P. (2023). Capital budgeting under technological uncertainty: A real options perspective on AI infrastructure investments. *Long Range Planning*, 56(4), 102345.

Brealey, R. A., Myers, S. C., & Allen, F. (2020). Principles of corporate finance (13th ed.). McGraw-Hill.

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

Chen, T., Xu, B., Zhang, C., & Guestrin, C. (2016). Training deep nets with sublinear memory cost. arXiv:1604.06174.

Dettmers, T., Lewis, M., Shleifer, S., & Zettlemoyer, L. (2022). 8-bit optimizers via block-wise quantization. International Conference on Learning Representations.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. Advances in Neural Information Processing Systems, 36.

Dixit, A. K., & Pindyck, R. S. (1994). Investment under uncertainty. Princeton University Press.

González, M., López, A., & Martínez, R. (2026). Pareto frontier analysis for cost-efficient fine-tuning of large language models. Decision Support Systems, forthcoming.

Hu, E. J., Shen, Y., Wallis, P., et al. (2021). LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations.

Lorie, J. H., & Savage, L. J. (1955). Three problems in rationing capital. *Journal of Business,* 28(4), 229-239.

Markowitz, H. (1952). Portfolio selection. *Journal of Finance*, 7(1), 77-91.

McDonald, R., & Siegel, D. (1986). The value of waiting to invest. *Quarterly Journal of Economics*, 101(4), 707-728.

Mishra, S., Ewing, M. T., & Cooper, H. B. (2022). Artificial intelligence focus and firm performance. *Journal of the Academy of Marketing Science*, 50, 1176-1197.

Pan, K., Liu, J., & Wang, M. (2025). Break-even analysis of on-premise versus cloud-based deployment of large language models. *Journal of Management Information Systems*, 42(1), 88-117.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L., Rothchild, D., So, D., Texier, M. & Dean, J. (2021). Carbon emissions and large neural network training. arXiv:2104.10350.

Ross, S. A., Westerfield, R. W., & Jaffe, J. (2019). Corporate finance (12th ed.). McGraw-Hill.

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Annual Meeting of the Association for Computational Linguistics, 3645-3650.

Trigeorgis, L. (1996). Real options: Managerial flexibility and strategy in resource allocation. MIT Press.

Vaswani, A., Shazeer, N., Parmar, N.,Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Weingartner, H. M. (1963). Mathematical programming and the analysis of capital budgeting problems. Prentice-Hall.

Yang, A., Yang, B., Hui, B., et al. (2024). Qwen2.5 technical report. arXiv:2412.15115.

Yuan, J., Zhang, Y., Wang, X., & Liu, F. (2025). No free lunch in memory-efficient fine-tuning: A multi-axis empirical analysis. *Information Systems Research,* 36(2), 412-438.