

Era Brain: Yetkiye Duyarlı Bir İşletme Belleğinin Geri Getirme, Çok Adımlı Kayıt Sorguları ve Erişim Denetimi Açısından Değerlendirilmesi

Era Research

Ekim 2026 · Era Teknik Rapor · commit 3aba0bbfb · Ön baskı v2

Özet

Era Brain, bir işletmenin kayıtlarını ve yazışmalarını yürüten Era platformunun bellek katmanıdır. Senkronize edilen kayıtları (müşteriler, siparişler, faturalar, ödemeler) ve yazışmalardan çıkarılan bilgileri tek bir PostgreSQL veritabanında tutar; bir soruyu, anahtar sözcük ve vektör aramasını birleştiren hibrit geri getirmeyle ve tipli ilişkiler üzerinde soruya göre yönlenen bir yol aramasıyla yanıtlar. Okuma sırasında hiçbir dil modeli çağrılmaz; okuma yolundaki tek model, yerelde çalışan bir gömme (embedding) modelidir. Geri getirmeyi iki açık konuşma belleği kıyaslamasında, LongMemEval-S [1] ve LoCoMo [2] üzerinde ölçüyoruz; LoCoMo'da yanıt doğruluğunu bir LLM hakemle değerlendiriyor, ayrıca üç üretilmiş işletme veri seti ve gizli bilgilerin yerleştirildiği bir yetki testi kullanıyoruz. Yanıtlanabilir 470 LongMemEval-S sorusunda kanıt oturumlarından en az biri %98,3, tamamı ise %90,2 oranında ilk 10 sonuç arasındadır. LoCoMo yanıt doğruluğu %76,1; aynı model konuşmanın tamamını okuduğunda bu değer %87,3. Aynı tutulan bir işletme setinde yanıt, 325 soruda %88,9 oranında bağlama ulaşır. 100.000 satırlık bir çizgede çok adımlı sorularda yol araması 100/100, düz vektör araması (ilk 10) ise 1/100 doğru sonuç verir. Kısıtlı üyelere verilen 12.560 yanıtın hiçbirinde yerleştirilen gizli bilgi sızmadı. Medyan arama süresi, süreç içinde ve ağ geçişi olmadan 29-35 ms aralığındadır. İşletme veri setleri sentetiktir ve tarafımızca hazırlanmıştır; bu değerlendirmenin neyi göstermediği 6. bölümde sıralanmıştır.

Tablo 1: Sonuçların özeti. Koşullarıyla birlikte altı temel değer. “Bekliyor” ile işaretli değerler henüz tamamlanmamış koşullardan gelecektir.

LoCoMo yanıt doğruluğu %76,1 LLM hakem · aynı modelle tam bağlam: %87,3	LongMemEval-S geri getirme %98,3 ilk 10'da ≥ 1 kanıt oturumu · tüm oturumlar: %90,2 · 470 soru	Aynı tutulan işletme soruları %88,9 yanıt bağlamda · 325 soru (245 TR, 80 EN)
Çok adımlı, 100 bin satırlık çizge 100/100 yol araması · düz vektör (ilk 10): 1/100	Yetki sızıntısı 0 12.560 üye yanıtı · 7 okuma yolu + kişisiz 8 çağırın	Arama süresi (p50) 29-35 ms LoCoMo / LongMemEval · süreç içi, ağ geçişi yok

1 Giriş

Dil modeli ajanları için geliştirilen bellek katmanları genellikle uzun sohbet geçmişleri üzerinde değerlendirilir: tek bir kullanıcı haftalarca bir asistanla konuşur, sistemden de söylenenleri hatırlaması beklenir [1, 2]. Bir işletmenin soruları ise farklıdır. “Bu müşterinin son siparişinin faturası ödendi mi?” sorusunun yanıtı, ilişkilerle birbirine bağlanan üç senkron kayıta durur; dünden bu yana değişmiş olabilir ve fatura görme yetkisi olmayan bir satış asistanına ulaşmamalıdır. Üstelik dil çoğu zaman Türkçedir ve sondan eklemeli yapı, düz anahtar sözcük aramasını işlemez hâle getirir.

Bu ortam için açık kıyaslama verisi azdır. CRM verisi üzerinde çalışan ajan kıyaslamaları [3, 4] ve paylaşılan ajan belleğinde erişim denetimini ölçen 2026 tarihli bir kıyaslama [5] mevcuttur; ancak hiçbirisi bir bellek katmanının senkron işletme kayıtları üzerindeki geri getirmesini kişi bazlı yetkiler altında ölçmez. Bu nedenle iki tür kanıt sunuyoruz: Era Brain'i yayımlanmış sistemlerle yan yana koyabilmek için açık konuşma kıyaslamalarında standart geri getirme ve yanıt doğruluğu sonuçları; ve kuruluşunu ve sınırlarını 3. ve 6. bölümlerde açıkladığımız, kendi hazırladığımız işletme ve yetki testleri.

Bu raporun dört katkısı vardır:

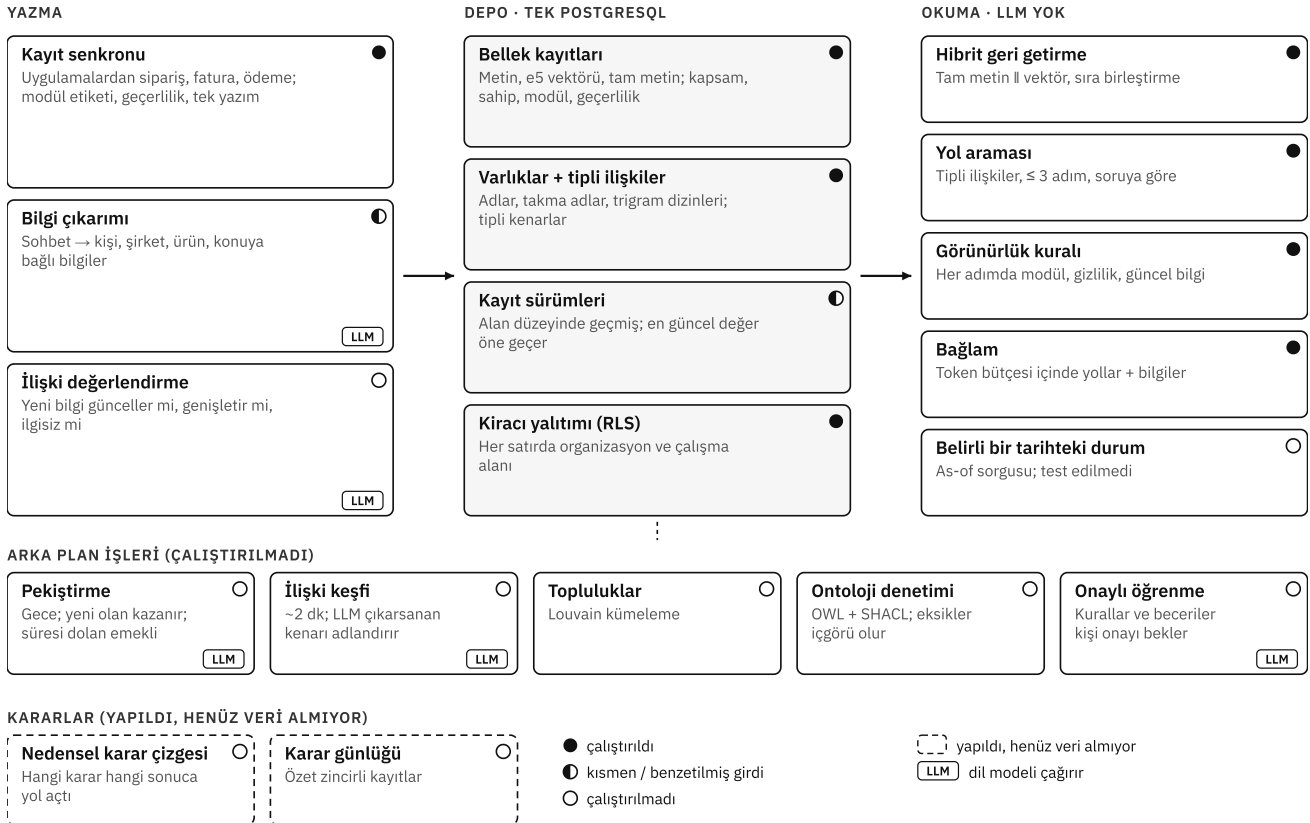
1. Era Brain'in yazma, saklama ve okuma yollarının, her iddiası kaynak koda bağlanarak tanımlanması (Ek A);
2. LongMemEval-S ve LoCoMo üzerinde, kıyaslamaların kendi soru sayıları ve ölçüt adlarıyla raporlanan geri getirme ve yanıt doğruluğu sonuçları;
3. Türkçe ve İngilizce bir işletme belleği değerlendirmesi: soru türleri, 100.000 çizge satırında çok adımlı kayıt birleştirmeleri, en güncel değer soruları ve bağlam büyüklüğü;
4. 865 gizli bilginin yerleştirildiği ve üyelere açık her okuma yolundan geçirilen bir yetki testi ile altı bellek sisteminin 22 Ekim 2026 itibarıyla belgelenmiş yetenekleriyle karşılaştırılması.

2 Sistem

Era Brain bir yazma yolu, bir depo ve bir okuma yolundan oluşur (Şekil 1). Yazma yolu senkron kayıtları ve yazışmaları bilgilere, varlıklara ve tipli ilişkilere dönüştürür. Depo tek bir PostgreSQL veritabanıdır. Okuma yolu bir soruyu dil modeli çağırmadan yanıtlar. Arka plan işleri çizgeyi zenginleştirir; bu değerlendirmede çalıştırılmamıştır.

2.1 Mimari

Bağlı uygulamalardan gelen kayıtlar dil modeli kullanılmadan yazılır. Her kayıt geldiği modülle (satış, finans, İK, ...) etiketlenir, kendi tarihinden itibaren sürümlenir ve senkron yeniden denendiğinde bile yalnızca bir kez yazılır: bellek kaydı, bağlantıları, kenarları, sürümü ve senkron defteri tek bir kilitli işlemde yazılır; varlıklar bu işlemde önce, tekrarlınsa da aynı sonucu verecek biçimde eklenir veya güncellenir. Yazışmalar ise LLM ile bilgi çıkarımından geçer. Çıkarılan bilgiler kişilere, şirketlere, ürünlere ve konulara bağlanır, senkron kayıtlara bağlanmaz; bir sohbette geçen sipariş numarası kayıtlar ancak soru sorulduğunda, metin üzerinden eşleşir.



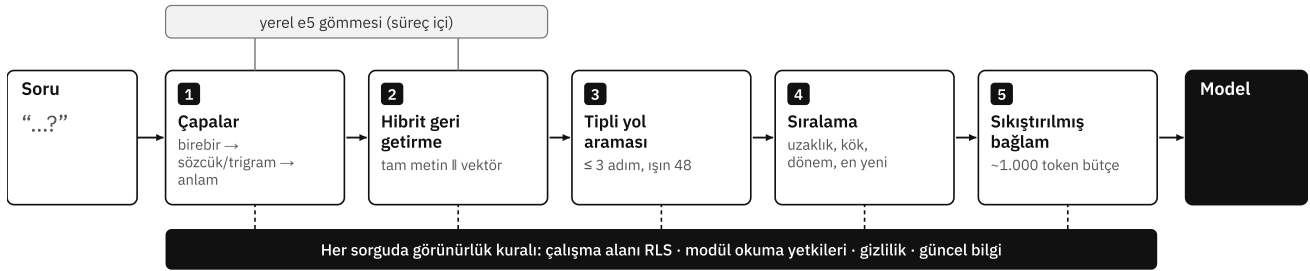
Şekil 1: Katmanlarına göre Era Brain bileşenleri. Dolu işaret: bu değerlendirmede çalıştırıldı; yarım: kısmen ya da benzetilmiş girdiyle; boş: çalıştırılmadı. Kesikli çerçeve: yapıldı, ancak üretimde henüz veri almıyor. LLM: bileşen bir dil modeli çağırır.

İki bileşen kodda mevcuttur ancak üretimde henüz veri almamaktadır: nedensel karar çizgesi ve onun özet zincirli günlüğü¹. Gece çalışan pekiştirme işinde de bilinen bir eksik vardır².

2.2 Okuma yolu

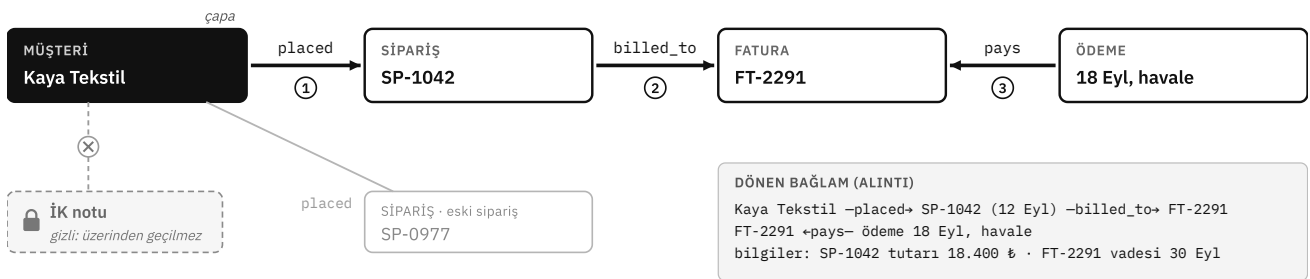
Bir soru, tek bir okuma işlemi içinde altı adımda yanıtlanır (Şekil 2):

- 1 Çapalar.** Sorudaki adlar önce birebir eşleştirilir; Türkçe hâl ekleri atılır (“Ağaoğlunun” → Ağaoğlu). Ardından sözcük içirme ve trigram benzerliğine, son olarak anlama (en yakın bellek kayıtlarının arkasındaki varlıklara) bakılır. Tür sözcükleri (“sipariş”, “fatura”, ünsüz yumuşamasıyla “talebindeki” → talep) ulaşılacak düğüm türünü belirler.
- 2 Hibrit geri getirme.** İngilizce ve Türkçe durak sözcükleri atılmış bir tam metin sorgusu ile bir vektör sorgusu (HNSW, kosinüs) birlikte çalışır; sonuçlar karşılıklı sıra (reciprocal rank) yöntemiyle birleştirilir.
- 3 Yol araması.** Çapalardan başlayan yürüyüş, tipli ilişkiler üzerinde en çok üç adım ilerler. Her düzeyde en umut verici 48 düğüm yola devam eder (toplamda en çok 600). 50'den fazla komşuya açılan bir ilişki, soruda adı geçmedikçe izlenmez. Sorulan ilişkiler ve türler karşılandığında yürüyüş durur.
- 4 Sıralama.** Çizge bilgileri, soruya vektör uzaklığına ve sorudaki sözcük köklerinden kaçını içerdiğine göre puanlanır; dönem ifadeleri (“geçen ay”, “eylülde”) o döneme ait bilgileri öne alır. Aynı nitelik iki kez söylenmişse en yenisi önce gelir, eskisi “önceki” olarak işaretlenir.
- 5 Görünürlük.** Her sorgu, soranın görünürlük kuralını taşır (Bölüm 2.3): kişinin göremeyeceği bir düğüm ne döndürülür ne de üzerinden geçilir.
- 6 Sıkıştırma.** Çapalar, kanıtlarıyla birlikte yanıt yolları ve kalan en iyi bilgiler bir token bütçesi içinde yazılır (sohbette yaklaşık 1.000 token; token başına üç karakter varsayılarak tahmin edilir).



Şekil 2: Tek bir soru için okuma yolu. Görünürlük kuralı en sonda bir süzgeç olarak değil, her adımın içinde uygulanır. Dil modeli çağrılmaz; tek model yerel sorgu gömmesidir.

“Kaya Tekstil’in son siparişinin faturası ödendi mi?”



Şekil 3: Örnek veriyle işlenmiş bir örnek. Soru bir müşteriye ve “sipariş”, “fatura”, “ödedi” türlerini adlandırır. Yürüyüş çapadan placed → billed_to → pays ilişkilerini izler, diğer dalları yok sayar; aynı müşteriye bağlı İK notu bu üye için görünmez ve üzerinden hiç geçilmez. İlişki adları kıyaslama verisindeki gibidir; canlı bir uygulamanın bağlayıcısı bunları farklı adlandırır (Bölüm 6).

¹ Her günlük kaydı bir öncekine SHA-256 ile zincirlenir; bu nedenle önceki bir kaydı değiştirmek ya da silmek zinciri bozar. En yeni kayıtların silinmesi ya da zincirin baştan yeniden yazılması ise dışarıda tutulan bir çapa olmadan fark edilemez.

² Pekiştirme 03:00 UTC’de ve yalnızca etkin bir bağlayıcı satırı olan organizasyonlar için çalışır; yalnızca kayıt senkronize eden organizasyonlar pekiştirilmez. Ürettiği bilgiler henüz modül, çalışma alanı ve gizlilikle etiketlenmemektedir; dolayısıyla Bölüm 4.6’daki yetki sonucu bu bilgileri kapsamaz.

2.3 Görünürlük kuralı

Erişim iki katmanda uygulanır (Şekil 4). PostgreSQL satır düzeyi güvenliği (RLS) her bağlantıyı tek bir organizasyon ve çalışma alanına bağlar. Bunun üzerine, her okuma sorgusuna ortak bir görünürlük kuralı eklenir: üyenin okuyamadığı modüllere ait bilgiler (platformdaki rol yetkilerinden çözülür), başkalarının özel ya da kişisel bilgileri ve geçerliliğini yitirmiş bilgiler dışarıda bırakılır. Yalnızca gizli bilgilerde adı geçen bir varlık düğüm sayılmaz; bir kenar da ancak modülü okunabiliyorsa ve iki ucu da görünürse izlenir. Üyenin yetkileri çözülmezse kural hiçbir modülü okumama durumuna düşer. Bir iş akışı ya da araştırma işi gibi kişisi olmayan bir ürün çağıranı yalnızca çalışma alanında paylaşılan bilgileri okur. Modül ve gizlilik denetimleri RLS politikası değil, sorgu koşuludur³.



• Kişi yoksa (iş akışı, araştırma) → yalnız paylaşılan bilgiler

Şekil 4: Erişim denetiminin iki katmanı. Satır düzeyi güvenlik her bağlantıyı organizasyon ve çalışma alanıyla sınırlar; görünürlük kuralı okuma yolunun her adımında bu üyenin okuyamayacaklarını çıkarır ve hata durumunda kapalı kalır.

2.4 Bir aramanın dokunduğu servisler

Şekil 5, tek bir arama çağrısının dokunduğu servisleri sayar. Era Brain sorgu gömmesini süreç içinde hesaplar ve tek bir PostgreSQL veritabanına, tek işlem içinde yaklaşık on SQL ifadesi gönderir. Güncel sistemlerin birçoğu da varsayılan okuma yolunda dil modeli kullanmaz; Mem0 v3, Graphiti ve Zep ile Supermemory bunlar arasındadır [6, 7, 8, 9]. Dolayısıyla şekildeki fark, yalnızca LLM olmaması değil, servis ve ağ geçişi sayısıdır. Üretimde stüdyo gömme modeline tek bir iç HTTP çağrısıyla ulaşır; kıyaslamada model süreç içinde çalıştırıldı.

		servis	LLM çağrısı
Era Brain	Uygulama → PostgreSQL süreç içi e5 gömmesi · ~10 SQL · 1 işlem	1	0
Mem0 OSS v3	Uygulama → Gömme yerel değilse API → Vektör deposu + varlık deposu	1-2	0
Mem0g (2025 makalesi)	Uygulama → LLM: sorudan varlıklar → Gömme → Vektör deposu → Neo4j	4	1
Graphiti (kendi sunucunda)	Uygulama → Gömme API'si → Çizge VT	2	0
Zep (barındırılan)	Uygulama → Zep API iç yapı yayımlanmamış	1 + ?	?
Supermemory (barındırılan)	Uygulama → Supermemory API iç yapı belgelenmemiş	1 + ?	0
Letta	Uygulama → Ajan LLM → Arşiv araması	2	≥ 1

Şekil 5: Tek bir arama çağrısının dokunduğu servisler; her sistemin 22 Ekim 2026 itibarıyla belgeleri ve kodundan. Servis: çağrının ulaştığı ayrı sistemler, veritabanı dahil. Mem0g, 2025 makalesinde [10] tanımlanan çizge sürümüdür; artık Mem0'nın açık kaynak paketinde yer almaz. Letta'da her aramaya ajan LLM'i karar verir.

³ Motorun kütüphane işlevleri hiç görüntüleyen verilmeden çağrıldığında süzgeçsiz çalışır; ürün içindeki her çağırana bir görüntüleyen geçirir (Ek A).

3 Deney düzeni

Veri setleri aşağıda listelenmiştir (Tablo 2). Bir dipnotta aksi belirtilmedikçe tüm koşullar 3aba0bbfb commit'indeki kodla yapılmıştır.

Tablo 2: Veri setleri. İşletme setlerini biz ürettik. H2, motorun ayarlanmasında kullanılan H setinin yeniden ifade edilmiş bir türevidir (yeni tohum, kayıt metinleri ve soru ifadeleri); H2 üretildikten sonra motor kodunda değişiklik yapılmadı.

Veri seti	Dil	Soru	İçerik	Kaynak
LongMemEval-S, temizlenmiş [1]	EN	500 (470 puanlandı)	soru başına ~115 bin tokenlık sohbet geçmişi	açık · MIT
LoCoMo-10 [2]	EN	1.540	10 uzun konuşma; çekışmeli 5. kategori dışarıda ([10] ile aynı)	açık · CC BY-NC 4.0, yalnız değerlendirme
İşletme seti H2 (ayrı tutulan)	TR + EN (245 / 80)	325	müşteriler, siparişler, faturalar, ödemeler, sevkıyatlar, destek talepleri, sözleşmeler, WhatsApp ve e-posta yazışmaları	üretilmiş · ürünün yazma yolları
İşletme setleri A ve H	TR + EN	370 / 365	aynı aile; ayarlama kullanıldı; 865 ve 790 gizli bilgiyle yetki taramaları	üretilmiş
Ölçek seti	EN	200 (100 çok adımlı)	20.000 varlık (12.000 kişi, 5.000 şirket, ürünler, projeler, yerler), 40.000 bellek kaydı, 40.000 tipli ilişki	üretilmiş · toplu yüklendi

3.1 Ölçütler

İlk 10'da kanıt (en az biri / tümü). Altın kanıt öğelerinden en az birinin ya da hepsinin dönen 10 öge arasında olup olmadığı: LongMemEval'de oturumlar veya turlar, LoCoMo'da diyalog turları. LongMemEval'in resmî geri getirme ölçütü recall_all'dur [1]; ikisini de raporluyoruz. Yanıt verilmemesi beklenen (abstention) 30 sorunun kanıtı yoktur ve LongMemEval'in kendi geri getirme puanlamasında olduğu gibi dışarıda bırakılmıştır. Kanıt etiketi olmayan dört LoCoMo sorusu ıskalanmış sayılır.

Yanıt bağlamda. İşletme setlerinde: altın yanıt metni dönen bağlamda geçiyor mu (büyük/küçük harf duyarsız, Türkçe yerel ayarıyla). Yanıtın doğru kanıttan geldiğini kanıtlamaz; 100 bin setinin üç adımlı sorularında her adımın kaynak bilgisi 5/40 durumda bağlamdadır.

Yanıt doğruluğu (QA). Bir okuyucu model her LoCoMo sorusunu dönen bağlamdan yanıtlar; bir LLM hakem yanıtı altın yanıtla karşılaştırarak puanlar; düzen [10] ile aynıdır. Okuyucu: claude-haiku-5-5; hakem: claude-haiku-5-5.

Token. Dönen bağlam metninin cl100k_base ile sayımı; bellek listeleri her kayıt bir satır olacak biçimde yazılır.

Gecikme. Kıyaslama sürecinde tek bir geri getirme çağrısının medyan (p50) duvar saati süresi: yerel sorgu gömmesi ve aynı makinede Docker içindeki PostgreSQL'e giden tüm sorgular dahil (Windows 11, 24 çekirdek); tek istemci, ısınmış, ilk sorgu hariç, uzak ağ yok.

3.2 Çalıştırılmayanlar

LongMemEval ve LoCoMo'da konuşmalar bilgi çıkarımı yapılmadan ham metin olarak saklandı; bu nedenle yalnızca hibrit geri getirme çalıştı, çizge kullanılmadı. Model maliyetini sıfırda tutmak için LLM ile bilgi çıkarımı, ilişki değerlendirme, pekiştirme ve ilişki keşfi kapalıydı; etkileri ölçülmemiştir. İşletme setlerinde kayıtlar ve ilişkiler, canlı bir uygulamanın bağlayıcı eşlemesinden geçmek yerine doğrudan yazıcının girdi biçiminde üretildi. Tekrarlanan koşullar set başına yaklaşık bir soru farklılık gösterir, gecikmeler de koşudan koşuya değişir⁴.

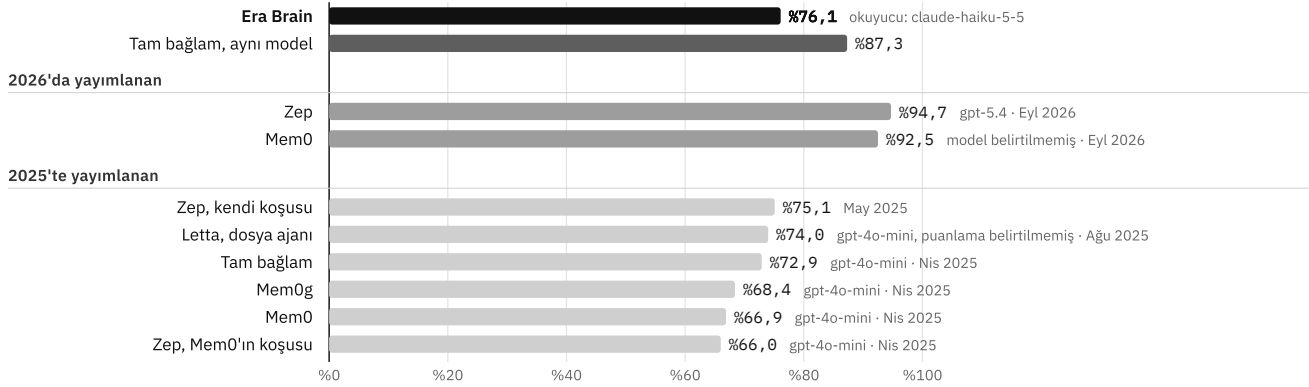
⁴ LoCoMo ve LongMemEval gecikmeleri, bellek geri getirmesine yapılan ve yalnızca hızı etkileyen son değişiklikten önce ölçüldü; görünürlük kuralı olmayan daha eski bir kod sürümünde LoCoMo p50 değeri 21 ms idi. 100 bin bellek araması gecikmesi (30 ms) commit edilmiş kodla yapılan ayrı bir koşudan gelir. Sohbet bağlamı yolu sızıntı için 700 tokenlık çizge bütçesiyle tarandı; commit edilen 1.000 tokenlık bütçe hedefli yetki sorularında denetlendi (490 yanıtta 0 sızıntı).

4 Sonuçlar

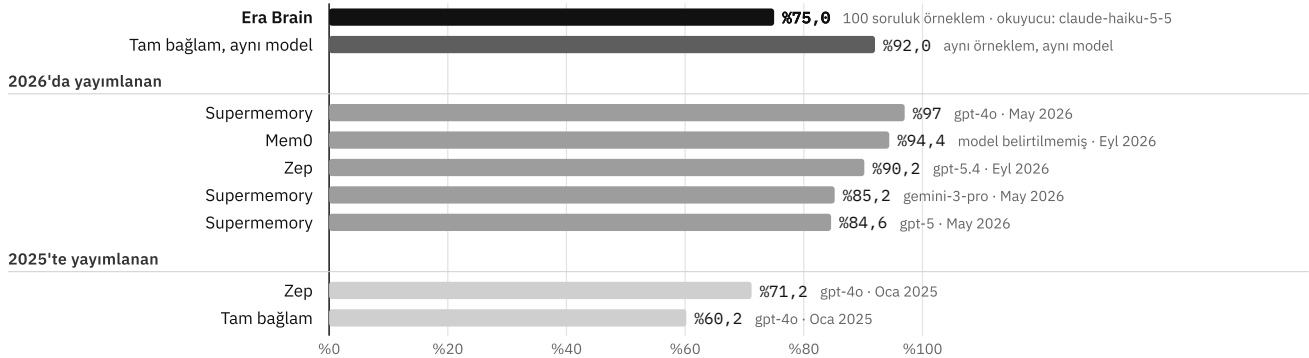
4.1 Yanıt doğruluğu

Şekil 6, Era Brain'in LoCoMo yanıt doğruluğunu, konuşmanın tamamını okuyan aynı modelli bir taban çizgisiyle ve yayımlanmış değerlerle yan yana gösterir. Yayımlanmış değerler farklı okuyucu ve hakem modelleri kullanır, tarafımızca yeniden koşulmamıştır ve tarihe göre gruplanmıştır. Mem0 ve Zep'in 2026 sonuçları [11, 8], 2025 makalelerine [10, 12] göre daha yeni okuyucular kullanır. Zep için 2025'te verilen %66,0 değeri Mem0 ekibince üretilmiştir ve Zep buna itiraz eder; Zep'in kendi koşusu %75,1 bildirir [13]. LoCoMo etiketlerinde, hakemle puanlanan her sonucu etkileyen hatalar vardır⁵.

(a) LoCoMo, 1.540 soru



(b) LongMemEval-S



Şekil 6: LLM hakemle yanıt doğruluğu. Siyah: Era Brain; koyu gri: aynı modelle tam bağlam taban çizgisi; gri: adı verilen makale ya da şirketçe yayımlanan değer, okuyucu model ve tarihiyle. Farklı okuyucu ve hakemlerle elde edilen değerler doğrudan karşılaştırılmaz. Kesikli: bekliyor.

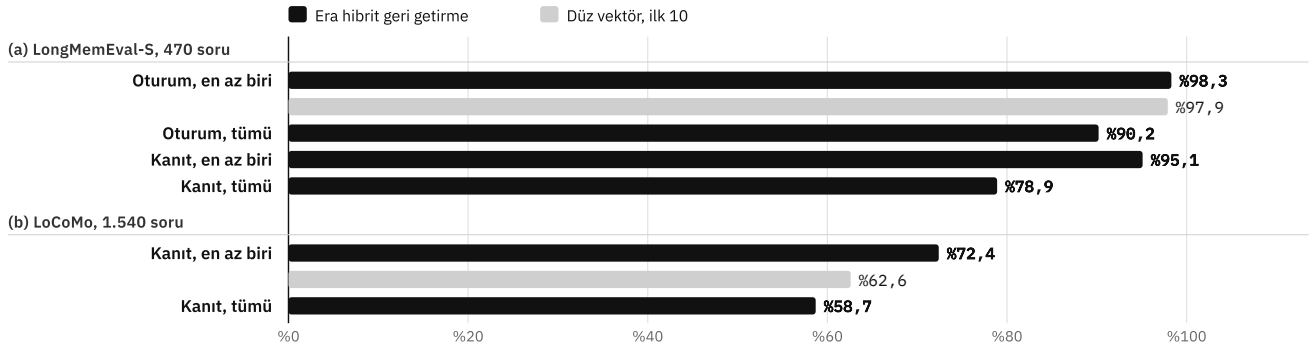
4.2 Geri getirme

Şekil 7, açık kıyaslamalardaki geri getirme sonuçlarını verir. LongMemEval-S'te hibrit geri getirme, soruların %98,3 kadarında en az bir kanıt oturumunu, %90,2 kadarında ise tüm kanıt oturumlarını ilk 10'a yerleştirir. Düz vektör araması (ilk 10) "en az biri" ölçütünde zaten %97,9 değerine ulaştığından kıyaslama burada az fark bırakır; yöntemleri "tümü" ölçütleri daha iyi ayırır. LoCoMo'da hibrit geri getirme en az bir kanıt turunu %72,4 oranında bulur; düz vektörde bu oran %62,6, bağlam ise yaklaşık 596 tokendir.

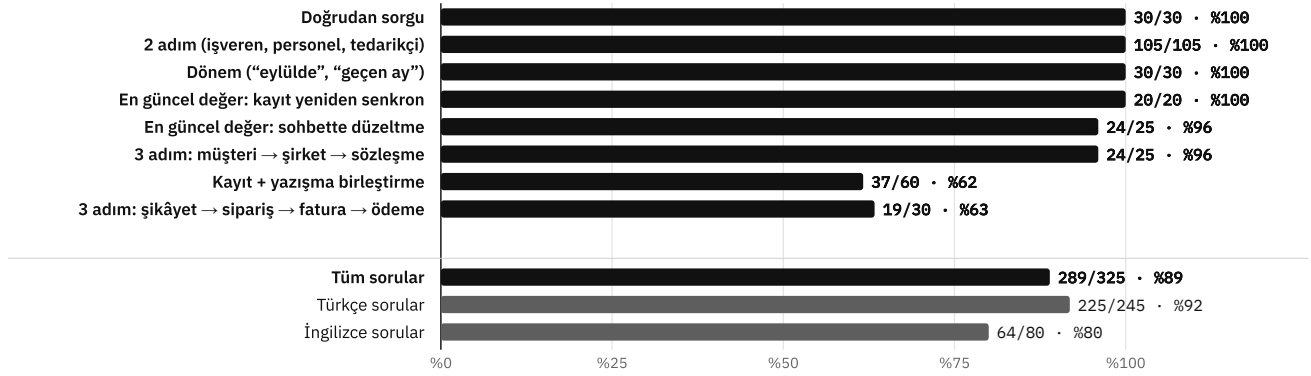
4.3 İşletme belleği

Şekil 8, ayrı tutulan seti soru türlerine ayırır. Doğrudan sorgular, iki adımlı sorular (kişi → işveren → vergi numarası, IBAN ya da web sitesi; destek talebi → atanan personelin telefonu; ürün → tedarikçinin telefonu), dönem soruları ve en güncel değer soruları neredeyse eksiksiz yanıtlanır. İki tür zayıf kalır: bir kaydı bir yazışmayla birleştiren sorular (soruda sipariş geçer, yanıt müşterinin WhatsApp ya da e-postasındadır; 37/60) ve şikâyet → sipariş → fatura → ödeme zinciri (19/30). Türkçe sorularda başarı %91,8, İngilizce sorularda %80,0.

⁵ Bağımsız bir inceleme, 1.540 soru içinde puanı etkileyen 99 etiket hatası bulunduğunu ve standart hakemin bilerek yanlış verilen yanıtların %62,8'ini kabul ettiğini bildirmektedir [14].



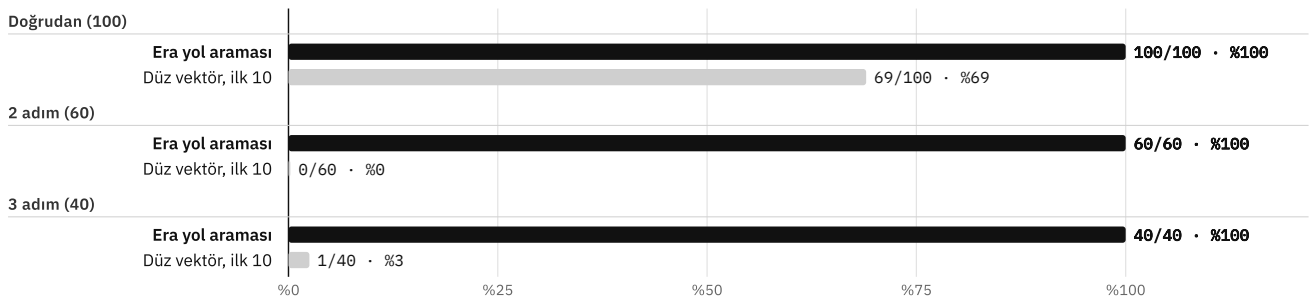
Şekil 7: Açık kıyaslamalarda geri getirme, ilk 10 öge. En az biri: altın kanıt öğelerinden en az biri döndü; tümü: hepsi döndü. Düz vektör değerleri aynı veri ve makinede yapılan bir önceki koşudan gelir.



Şekil 8: Ayrı tutulan işletme seti H2 (325 soru), soru türüne göre yanıtın bağlamda bulunması. Çubuklar ölçeklidir; etiketler doğru / toplam sayısını verir.

En güncel değer kazanır. Bir kayıt yeni bir değerle yeniden senkronize edildiğinde ya da bir yazışma önceki bir ifadeyi düzelttiğinde, yeni bilginin önce gelmesi gerekir. Bu koşul kayıt güncellemelerinde 20/20, yazışma güncellemelerinde 24/25 oranında sağlandı.

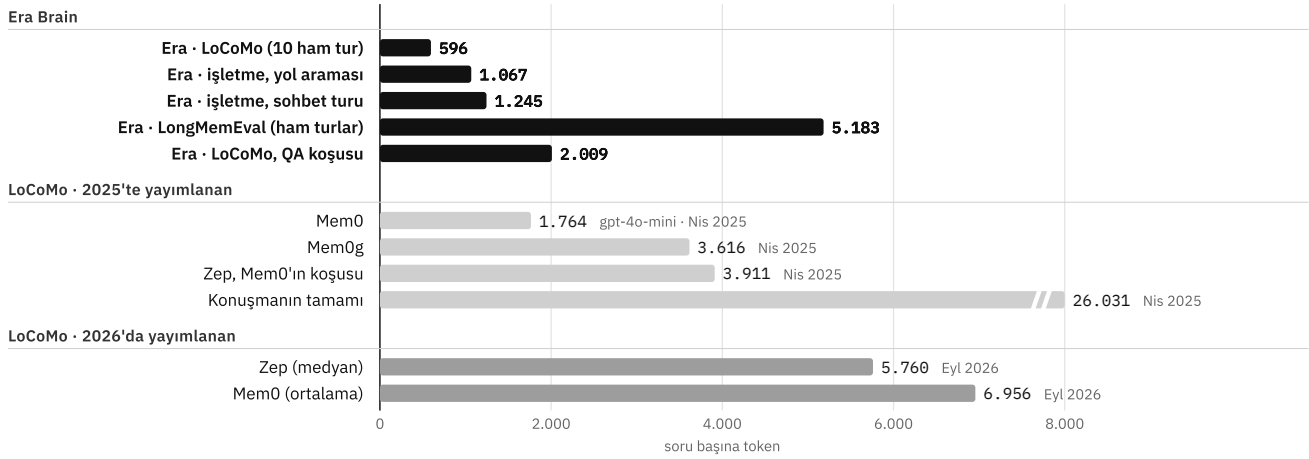
Şekil 9, 100.000 satırlık ölçek setindeki çok adımlı soruları ayrıca gösterir. Yol araması doğrudan, iki adımlı ve üç adımlı soruların hepsini yanıtlar. Aynı bellek kayıtları üzerinde düz vektör araması (ilk 10) iki adımlı sorularda 0/60, üç adımlı sorularda 1/40 doğru sonuç verir; genelde aldığı %35'lik değer (70/200) neredeyse tamamen doğrudan sorgulardan gelir. Vektör tabanlı RAG için açık çok adımlı sorularda da aynı örüntü bildirilmiştir [15].



Şekil 9: Ölçek seti (100.000 çizge satırı, soran: sahip): yol araması ile düz vektör araması (ilk 10), yanıtın bağlamda bulunması.

4.4 Bağlam büyüklüğü

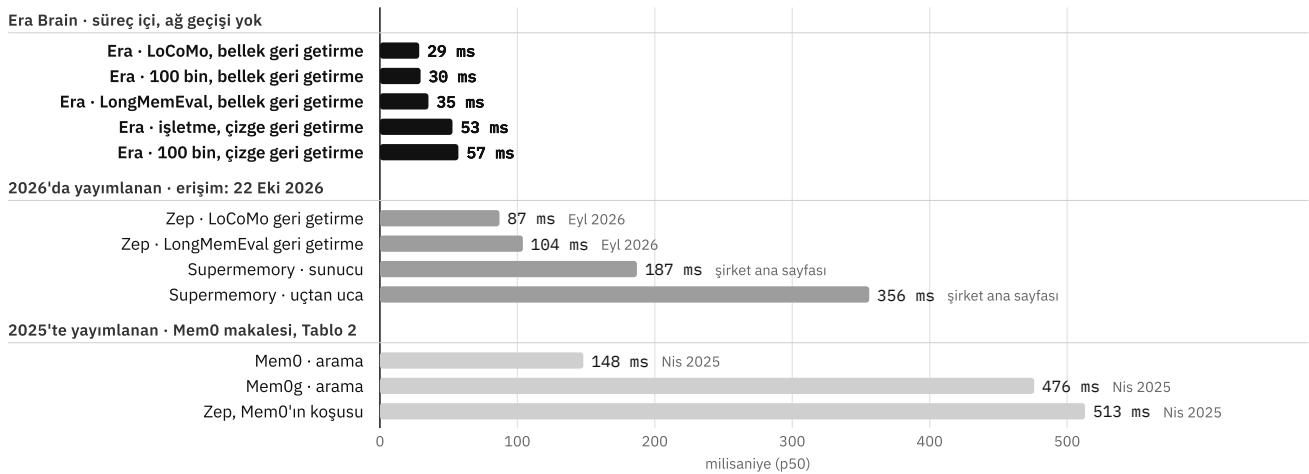
Şekil 10, soru başına bağlam tokenlarını karşılaştırır. LoCoMo'da hibrit geri getirme on ham tur döndürür; bu yaklaşık 596 tokendir. İşletme sorularında yol araması 1.067 token döndürür, bir Era sohbet turu ise modele 1.245 token verir. 2025 değerleri, Mem0, Mem0g ve Zep'in QA sonuçları için kullandıkları bağlamdır; 2026 sistemleri çok daha yüksek doğruluk için daha büyük bağlam kullanır. Daha düşük geri getirmeyle daha az token, eşit koşullarda daha ucuz anlamına gelmez; tokenlaştırmacılar da farklıdır. Bu nedenle bu sayılar yol göstericidir.



Şekil 10: Soru başına bağlam tokeni (aksi belirtilmedikçe p50). Siyah: Era Brain; gri: tarihiyle birlikte yayımlanmış LoCoMo değerleri. Tam bağlam çubuğu kesilmiştir; değeri yazılıdır.

4.5 Gecikme

Şekil 11 medyan arama süresini gösterir. Era Brain'in değerleri süreç içinde, ağ geçişi olmadan ölçülmüştür; şirketlerin değerleri ise barındırılan servisleri üzerinden ya da yayınlarında belirtildiği biçimde ölçülmüştür. Bu nedenle karşılaştırma bir sıralama değil, büyüklük mertebesi verir. Yayımlanmış en hızlı barındırılan geri getirme gecikmesine göre (Zep, LoCoMo'da 87 ms), Era Brain'in 29 ms değeri yaklaşık 3 kat daha düşüktür. İşletme sorularında çizge geri getirmesi daha uzun sürer (53 ms), çünkü üç düzeye kadar yürür ve her düzeyde görünürlük kuralını uygular⁶.



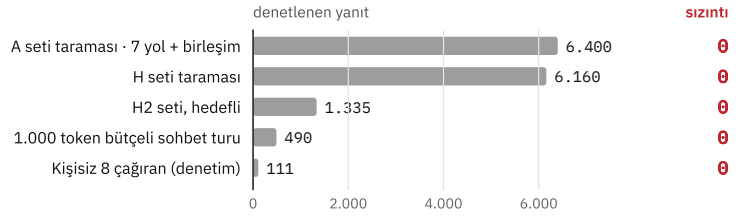
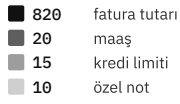
Şekil 11: Milisaniye cinsinden medyan (p50) arama gecikmesi, doğrusal ölçek. Siyah: Era Brain, süreç içi, yerel veritabanı, ağ geçişi yok. Gri: yayımlandığı gibi (erişim 22 Ekim 2026), barındırılan servisler üzerinden ya da belirtildiği biçimde ölçülmüş.

4.6 Yetkiler

Yetki testinde bir satış üyesi (finans, fatura ve İK yetkisi yok) her soruyu, bir finans üyesi (İK yetkisi yok) ise hedefli her yetki sorusunu sorar. A setine 865 gizli bilgi (820 fatura tutarı, 20 maaş, 15 kredi limiti, 10 özel not), H setine 790 gizli bilgi yerleştirildi. Soranın görmemesi gereken bir gizli bilginin dönen bağlamda yer alması sızıntı sayılır. Üyelere açık yedi okuma yolu tarandı: ajan servisinde bellek ve çizge geri getirmesi, stüdyoda geri getirme, Era'nın bellek araması ve bellek çizgesi araçları, Era sohbet turu bağlamı ve builder copilot bağlamı. Kişisi olmayan sekiz çağırın (araştırma, iki iş akışı bloğu, iç geri getirme ve listeleme, copilot ve çizge bağlamı, toplantı özeti) ayrıca denetlendi. Sonuç Şekil 12 içinde görülüyor: hiçbir yanıta sızıntı yok.

⁶ Mem0'in 2026 README dosyası LoCoMo için neyi kapsadığını belirtmeden "latency p50 0.88 s" verir; grafikte gösterilmemiştir. Supermemory'nin ana sayfasındaki değerler (sunucu ve uçtan uca) yayımlandığı gibi gösterilmiştir.

A setine yerleştirilen 865 gizli bilgi



Şekil 12: Sol: A setine yerleştirilen 865 gizli bilginin dağılımı. Sağ: test başına denetlenen yanıtlar ve bulunan sızıntılar (kırmızı).

4.7 Sonuçları değiştiren adımlar

Tablo 3, sonuçları en çok etkileyen değişiklikleri listeler. Her satır motoru, adı geçen değişiklikten önce ve sonra, aynı veri ve makinede karşılaştırır. Satırların çoğu birbiriyle ilişkili birkaç değişikliği birlikte içerir; bu yüzden bir satır tek bir tekniğin etkisini yalıtırmaz.

Tablo 3: Önce ve sonra karşılaştırmaları. “Önce” değerleri aynı veri üzerinde daha eski kod sürümlerinden gelir.

Değişiklik	Önce	Sonra	Set, ölçüt
Dizini kullanabilen tam metin araması ve toplu erişim sayacı	4.031 ms	30 ms	100 bin · bellek araması p50
1.000 düğümlük genişletme yerine isabetli tohumlar ve soruya göre sıralanan bilgilerle tipli yol araması	0/100 · 2.466 token · 377 ms	100/100 · 1.060 token · 55 ms	100 bin · çok adımlı: yanıt, token, p50
Düz vektör (ilk 10) yerine hibrit geri getirme	%62,6	%72,4	LoCoMo · kanıt, en az biri
Düz vektör (ilk 10) yerine yol araması	%35	%100	100 bin · 200 sorunun tümü, yanıt bağlamda
Yol motorunda Türkçe eşleştirme ve sıralama çalışmalarının tümü (ekler, ünsüz yumuşaması, uç düğüm bilgileri, kayıt türleri, dönem ifadeleri)	%48,0	%88,9	H2 · yanıt bağlamda
Her adımda görünürlük kuralı, satır düzeyi güvenlik maliyetiyle	34 ms	56 ms	ilk Türkçe set (310 soru) · çizge p50

5 İlgili sistemlerle karşılaştırma

Tablo 4, her projenin 22 Ekim 2026 itibarıyla belgelediği yetenekleri; şirket belgeleri, makaleler ve kaynak kod üzerinden karşılaştırır [6, 7, 16, 9, 17, 18, 19, 20]. Yalnızca belgelenmiş olanı işaretledik; “?” aradığımız hâlde bulamadığımız anlamına gelir. İki bulgu Era Brain’in aleyhinedir. Birincisi, dil modeli kullanmayan bir okuma yolu artık ayırt edici değildir: Mem0 v3, Graphiti, Zep ve Supermemory varsayılan olarak LLM çağırır. İkincisi, ince taneli erişim başka sistemlerde de vardır: Zep Enterprise, veri alınırken eklenen üst verilere dayalı politika ve nitelik kuralları sunar [16]; Supermemory üyeleri ad alanlarıyla sınırlar; Cognee yetkileri veri seti bazında verir. Başka bir sistemde bulamadığımız şey, işletmenin kendi rol ve modül yetkilerinden türetilen ve çizgenin her adımında uygulanan bilgi bazlı bir kural ile yayımlanmış bir yetki sızıntısı testidir.

Diğer sistemler birkaç alanda öndedir: yayımlanmış uçtan uca QA sonuçları (Mem0, Zep, Supermemory); geçersiz kılmayla çift zamanlı kenarlar (Zep ve Graphiti; Era’nın kenarlarında geçerlilik aralığı yoktur ve çizge yürüyüşleri yalnızca güncel bilgileri okur); yönetilen bilgi çıkarımı ve özel ontolojiler (Zep, Mem0); barındırılan ürünler, SDK’lar ve entegrasyonlar; belgelenmiş uyumluluk (Zep: SOC 2 Type II, HIPAA; Supermemory: SOC 2, HIPAA); açık değerlendirme altyapıları (Mem0, Supermemory). LoCoMo’da yalnızca dosya arayan bir ajan (Letta, gpt-4o-mini) %74,0 aldı [18]; bu, Mem0’ın 2025 sonucunun üzerindedir. Dolayısıyla wiki gibi dosya ve aramaya dayalı bir bellek, sohbet hatırlama için güçlü bir taban çizgisidir. Bizim karşılaştırmamız işletme kayıtları ve yetkilerle ilgilidir.

Tablo 4: 22 Ekim 2026 itibarıyla belgelenmiş yetenekler. ● var · ½ kısmen, plana bağlı ya da sınırlı · ○ yok · ? belgelenmemiş (arandı). Era Brain'ın QA satırı, Bölüm 4.1'deki koşu tamamlanana kadar bekliyor olarak kalır.

	Era Brain	Mem0	Zep / Graphiti	Super-memory	Letta	Cognee	Wiki / Obsidian	Vektör RAG
Varsayılan okuma yolunda LLM yok	●	●	●	●	○	○	○	●
İşletme kayıtları LLM'siz, tipli ve sürümlü yazılır	●	○	○	○	○	½	○	○
Tipli, soruya göre yönlendirilen çok adımlı yol araması	●	○	½	○	○	½	½	○
Çizge kenarlarında geçerlilik zamanı (çift zamanlı)	½	½	●	½	○	½	½	○
Okumada en güncel değer kazanır	●	½	●	●	○	?	½	○
Rol/modül yetkilerinden bilgi bazlı kural, her çizge adımında	●	○	½	½	○	½	½	½
Anahtar sözcük aramasında Türkçe biçimbilim	●	○	○	?	?	○	○	?
Vektör, metin, çizge ve erişim için tek depo	●	½	○	?	●	○	●	●
Yayımlanmış uçtan uca QA sonucu	●	●	●	●	●	½	○	○
Yayımlanmış yetki sızıntısı testi	●	○	○	○	○	○	○	○
Yayımlanmış yapılandırılmış işletme kıyaslaması	●	○	○	○	○	○	○	○
Barındırılan ürün ya da açık SDK	○	●	●	●	●	●	½	?
Belgelenmiş uyumluluk (SOC 2, HIPAA)	?	?	●	●	?	?	○	?
Açık değerlendirme altyapısı	○	●	?	●	?	?	○	○
Düz dosya olarak düzenlenebilen bellek	○	○	○	○	●	○	●	½

Tablo 5: Bu raporda kullanılan, yayımlanmış rakip değerleri; okuyucu model ve tarihiyle. Hepsi uçtan uca QA sonucudur; hiçbirini geri getirme oranı değildir.

Sistem	Kıyaslama	Sonuç	Okuyucu / hakem	Tarih	Kay.
Mem0	LoCoMo	%66,9	gpt-4o-mini	Nis 2025	[10]
Mem0g	LoCoMo	%68,4	gpt-4o-mini	Nis 2025	[10]
Zep, Mem0'ın koşusu	LoCoMo	%66,0	gpt-4o-mini	Nis 2025	[10]
Zep, kendi koşusu	LoCoMo	%75,1	—	May 2025	[13]
Tam bağlam	LoCoMo	%72,9	gpt-4o-mini	Nis 2025	[10]
Letta, dosya ajanı	LoCoMo	%74,0	gpt-4o-mini; puanlama belirtilmemiş	Ağu 2025	[18]
Mem0	LoCoMo / LongMemEval	%92,5 / %94,4	model belirtilmemiş	Eyl 2026	[11]
Zep	LoCoMo / LongMemEval	%94,7 / %90,2	gpt-5.4	Eyl 2026	[8]
Zep	LongMemEval-S	%71,2	gpt-4o	Oca 2025	[12]
Tam bağlam	LongMemEval-S	%60,2	gpt-4o	Oca 2025	[12]
Supermemory	LongMemEval-S	%97 / %84,6 / %85,2	gpt-4o / gpt-5 / gemini-3-pro	May 2026	[21]

6 Sınırlılıklar

- Bu taslakta yanıt doğruluğu henüz bekliyor. Diğer tüm doğruluk değerleri geri getirmeye aittir: kanıtın modele ulaşıp ulaşmadığını ölçer, doğru bir yanıt üretilip üretilmediğini değil.
- İşletme setleri sentetiktir ve tarafımızca hazırlanmıştır. H2, ayarlama sırasında kullanılan bir setin yeniden ifade edilmiş türevidir; gerçek müşteri verisiyle doğrulama henüz yapılmamıştır.

- İşletme setlerindeki kayıtlar ve ilişkiler yazıcının girdi biçiminde üretildi. Bugünkü bağlayıcı eşlemesi, genel bir uygulama için genel ilişkiler (bought, paid by, related to) üretir; bu nedenle canlı bir çizge, burada üzerinde yürünen çizgeden daha seyrek olabilir.
- “Yanıt bağlamda” bir metin eşleşmesidir. 100 bin setinin üç adımlı sorularında her adımın kaynak bilgisi 5/40 durumda bağlamdadır; yanıt yol satırı üzerinden gelir.
- 100 bin seti ürün üzerinden yazılmadı, SQL ile toplu yüklendi; soruları yalnızca sahip sordu.
- Gecikme, yerel veritabanında tek istemcili, süreç içi bir çağrıdır; LoCoMo ve LongMemEval gecikmeleri geri getirmeye yapılan son hız değişikliğinden önce ölçüldü.
- LLM katmanları (bilgi çıkarımı, ilişki değerlendirme, pekiştirme, ilişki keşfi) çalıştırılmadı. Pekiştirmenin ürettiği bilgiler henüz modül, çalışma alanı ve gizlilikle etiketlenmiyor; sıfır sızıntı sonucu bu bilgileri kapsamaz.
- Belirli bir tarihteki durumu soran (as-of) geri getirme test edilmedi; çizge kenarlarında geçerlilik aralığı yoktur.
- Rakip değerleri yayınlarından alınmıştır, yeniden koşulmamıştır; farklı modeller, tokenlaştırıcılar ve düzenler kullanırlar.
- LoCoMo CC BY-NC 4.0 lisanslıdır; yalnızca değerlendirme için kullanıldı ve yeniden dağıtılmadı. Değerlendirme altyapısı henüz açık değildir.

7 Sonuç

Era Brain işletme sorularını, okuma yolunda dil modeli kullanmadan, tek bir PostgreSQL veritabanından yanıtlar. Bu ortam için kurduğumuz testlerde (çok adımlı kayıt birleştirmeleri, en güncel değerler, Türkçe ifadeler ve üye bazlı yetkiler) ayrı tutulan sorularda yanıtı %88,9 oranında buldu ve 12.560 yanıtın hiçbirinde gizli bilgi sızdırmadı; bunu sohbet turu başına yaklaşık 1.245 tokenlık bağlamla yaptı. Açık konuşma kıyaslamalarında geri getirmesi düz vektör taban çizgisini geçer; LoCoMo yanıt doğruluğu %76,1. Sonraki adımlar: gerçek müşteri verisiyle doğrulama, değerlendirme altyapısının açık olarak yayımlanması, açık bir erişim denetimi kıyaslamasının [5] koşulması ve pekiştirmenin ürettiği bilgilerin modül, çalışma alanı ve gizlilikle etiketlenmesi.

Kaynakça

- [1] D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, D. Yu. LongMemEval: Benchmarking Chat Assistants on Long-Term Interactive Memory. ICLR 2025. arXiv:2410.10813. Data: huggingface.co/datasets/xiaowu0162/longmemeval-cleaned (MIT). <https://arxiv.org/abs/2410.10813>
- [2] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, Y. Fang. Evaluating Very Long-Term Conversational Memory of LLM Agents. ACL 2024. arXiv:2402.17753. Data: github.com/snap-research/locomo (CC BY-NC 4.0). <https://aclanthology.org/2024.acl-long.747>
- [3] K.-H. Huang et al. CRMarena: Understanding the Capacity of LLM Agents to Perform Professional CRM Tasks in Realistic Environments. arXiv:2411.02305, 2024. <https://arxiv.org/abs/2411.02305>
- [4] K.-H. Huang et al. CRMarena-Pro. arXiv:2505.18878, 2025. <https://arxiv.org/abs/2505.18878>
- [5] GateMem: a benchmark for multi-principal shared-memory agents. arXiv:2606.18829, June 2026. <https://arxiv.org/abs/2606.18829>
- [6] Mem0 documentation: OSS v2 → v3 migration; graph memory. <https://docs.mem0.ai/platform/features/graph-memory> (erişim: 22 Eki 2026)
- [7] Graphiti (getzep/graphiti), graphiti-core 0.31.2, source code. <https://github.com/getzep/graphiti> (erişim: 22 Eki 2026)
- [8] Zep. Research results (page modified 29 Sep 2026). <https://www.getzep.com/research/> (erişim: 22 Eki 2026)
- [9] Supermemory documentation: search defaults; container tags (namespaces). <https://supermemory.ai/docs/concepts/container-tags> (erişim: 22 Eki 2026)
- [10] P. Chhikara, D. Khant, S. Aryan, T. Singh, D. Yadav. Mem0: Building Production-Ready AI Agents with Scalable Long-Term Memory. arXiv:2504.19413, 2025, Table 2 (gpt-4o-mini). <https://arxiv.org/abs/2504.19413>
- [11] Mem0. Research results and 'The token-efficient memory algorithm' (blog, Apr 2026, updated 28 Sep 2026). <https://mem0.ai/research> (erişim: 22 Eki 2026)

- [12] P. Rasmussen, P. Paliychuk, T. Beauvais, J. Ryan, D. Chalef. Zep: A Temporal Knowledge Graph Architecture for Agent Memory. arXiv:2501.13956, 2025. Blog: getzep.com/blog/state-of-the-art-agent-memory. <https://arxiv.org/abs/2501.13956>
- [13] Zep. Lies, Damn Lies, Statistics: Is Mem0 Really SOTA in Agent Memory? Blog, 6 May 2025. <https://www.getzep.com/blog/lies-damn-lies-statistics-is-mem0-really-sota-in-agent-memory/> (erişim: 22 Eki 2026)
- [14] Penfield Labs. Audit of LoCoMo labels and judge leniency. Blog post, 8 Apr 2026.
- [15] Y. Tang, Y. Yang. MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries. arXiv:2401.15391, 2024. <https://arxiv.org/abs/2401.15391>
- [16] Zep documentation: attribute-based and policy-based access control (Enterprise). <https://help.getzep.com/v4/attribute-based-access-control.md> (erişim: 22 Eki 2026)
- [17] Cognee (topoteretes/cognee), commit 37f788bd, source code. <https://github.com/topoteretes/cognee> (erişim: 22 Eki 2026)
- [18] Letta. Benchmarking AI Agent Memory: Is a Filesystem All You Need? Blog, 12 Aug 2025. <https://www.letta.com/blog/benchmarking-ai-agent-memory> (erişim: 22 Eki 2026)
- [19] A. Karpathy. LLM wiki (gist). <https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f> (erişim: 22 Eki 2026)
- [20] pgvector: open-source vector similarity search for PostgreSQL. README. <https://github.com/pgvector/pgvector> (erişim: 22 Eki 2026)
- [21] Supermemory. LongMemEval-S evaluation report, May 2026. <https://supermemory.ai/research/longmembench/> (erişim: 22 Eki 2026)

Ek A Her iddianın koddaki yeri

Depo commit'i 3aba0bbfb. Yollar depo köküne görelidir; brain-paths = packages/db/brain-paths.ts, brain-viewer = packages/db/brain-viewer.ts.

Tablo A1: İddialar ve kaynak koddaki yerleri.

İddia	Kod
Okurken LLM çağrısı yok; tek model yerel sorgu görmesi	brain-paths:1187, :55-63 · apps/brain-agent/src/brain/recall.ts:8-16 · apps/studio/lib/brain/recall.ts:1-14 · apps/studio/lib/brain/embed-client.ts:6-14
Gömme: multilingual-e5-small, q8, CPU, 384 boyut; HNSW kosinüs dizini	apps/brain-agent/src/brain/local-embeddings.ts:17 · embeddings.ts:22-28 · packages/os-kernel/sql/450_brains.sql:146,150
Anahtar sözcük ve vektör kolları birlikte çalışır; EN + TR durak sözcükleri GIN sorgusundan önce atılır	apps/brain-agent/src/brain/recall.ts:184-217 · sql/898_brain_speed.sql:37-51,189-265
Çapalar: önce tam ad eşleşmesi, Türkçe hâl ekleri atılarak	brain-paths:296-308,325-327,341-375,462
Sonra sözcük içerme + trigram benzerliği, sonra anlam	brain-paths:377-446 (similarity :417), :471-491, :1239-1252
Ünsüz yumuşamasıyla tür sözcükleri ('talebindeki' → talep)	brain-paths:597-605,690-718
En çok 3 adım; her düzeyde en iyi 48 düğüm ilerler; en çok 600 düğüm	brain-paths:78-79,88,1227,1295-1297,1363-1365
50'den fazla komşuya açılan ilişki, soruda geçmedikçe izlenmez (geçerse 8)	brain-paths:72-76,785-791,1277
Bilgiler yakınlık + soru kökleriyle sıralanır; dönem sözcükleri o dönemi öne alır (çizge bilgileri)	brain-paths:224-237,925-944,963-970,984-1024
Aynı nitelik iki kez: en yenisi önce, eskisi 'önceki' olarak işaretli	brain-paths:1064-1142,1616
Bağlam için token bütçesi (token başına 3 karakter tahmini)	brain-paths:1610,1620-1687 · apps/studio/lib/era/turn/right-now.ts:112
Görünürlük kuralı: çalışma alanı, okunabilen modüller, gizlilik, güncel bilgiler; her okumada	brain-viewer:184-232 · packages/os-kernel/sql/900_brain_viewer_access.sql:19-41
Gizli düğümler ne gösterilir ne üzerinden geçilir (tohum, adım, kanıt, bilgi)	brain-paths:357-366,425,487,779-798,873,947,1157-1161
Görüntüleyen çözümlemesi hata durumunda kapalı kalır (hiçbir modülü okumaz)	brain-viewer:76-108
Bellek tablolarında satır düzeyi güvenlik: organizasyon / çalışma alanı / kişi	sql/092_tenant_strict.sql:392-400 · sql/898_brain_speed.sql:354-386 · packages/db/index.ts:7-18,78-170
Kişi yoksa: yalnız çalışma alanında paylaşılan bilgiler (ürün çağrılarları)	brain-viewer:110-178 · apps/studio/lib/brain/brain-context.ts:122-123 · workflow-scope.ts:84 · apps/brain-agent/src/research/run.ts:90
Kayıt senkronu: modül etiketi, geçerlilik başlangıcı, sürümler; tek kilitli işlem	apps/brain-agent/src/brain/connector-ingest.ts:199-201,242-369 (lock :244, tag :259-264)
Bilgi çıkarımı ve ilişki değerlendirmesi (LLM); altı varlık türü	apps/brain-agent/src/brain/extraction.ts:26-37,66,99,232-276 · ingestion.ts:899-907
Pekiştirme: 03:00 UTC, etkin bağlayıcı satırı olan organizasyonlar, tek LLM geçişi	apps/brain-agent/src/index.ts:426-439 · consolidation.ts:73-118,139-154,186-214,271-286
İlişki keşfi: birlikte geçme + gömme adayları; LLM çıkarsanan kenarı adlandırır	apps/brain-agent/src/brain/relationships.ts:1-30,140,201-203 · extraction.ts:385-420
Topluluk tespiti: Louvain (networkx, seed 42); ontoloji OWL + SHACL denetimi	services/semantica-sidecar/app.py:122-189 · ontology-core.ts:82-165 · ontology-tick.ts:130-145
Özet zincirli karar günlüğü; üretimde henüz yazan yok	apps/brain-agent/src/decisions/provenance-core.ts:48-90 · provenance.ts:45-110
Bilgiler Era ekranlarından düzenlenebilir (düzelt, unut, gizli tut, onayla)	sql/450_brains.sql:156-245