RAYDORF BENCHMARKS · RDF-BMK-2026-1

Türk Hukuk Yapay Zekâsı *Araştırma Platformları*

Pilot Değerlendirme Raporu

Leagle, Apilex ve Dejure.ai'nin Performans Değerlendirmesi

§ 0 · SUNUŞ

Sunuş

Türk hukuk pratiğinde yapay zekâ destekli araştırma platformları artık bir merak konusu olmaktan çıktı; günlük çalışmanın içine yerleşti. Hukuk müşavirleri ve avukatlar, içtihat taramasından mütalaa hazırlığına kadar uzanan işlerin bir bölümünü bu platformlara emanet ediyor. Emanet edilen şey ise küçük değil: bir atfın doğruluğu, bir içtihadın güncelliği, bir istisnanın gözden kaçıp kaçmadığı. Yani mesleki sorumluluğun tam merkezindeki sorular.

Buna rağmen bu araçların nasıl seçileceğine dair elimizde çok az güvenilir bilgi var. Platformların kendi tanıtım materyalleri, doğası gereği, karşılaştırma yapılabilecek bir zemin sunmuyor. Peki bireysel deneme yanılma? O da yetmiyor. Bir uygulayıcının tek başına, farklı hukuk alanlarına yayılan bir soru kümesini birden fazla platformda aynı koşullarda test etmesi, sonuçları belge belge doğrulaması ve bunu platformlar güncellendikçe yinelemesi fiilen mümkün değil. Üstelik günlük iş yükü içinde buna ayrılacak zaman, tam da bu araçların kazandırması beklenen zamanın kendisi.

Bu boşluğun kökünde tek bir eksik yatıyor: Türkiye'de hukuk yapay zekâsı araçlarının doğruluğunu ölçmek için üzerinde uzlaşılmış ortak bir standart yok. Oysa bir hukukçunun aracına duyduğu güven izlenime, alışkanlığa veya pazarlama diline değil, açıklanmış bir yöntemle üretilmiş ve doğrulanabilir bulgulara dayanmalı. Uygunluk değerlendirmesi geleneği başka sektörlerde yüz yılı aşkın süredir tam olarak bunu yapıyor: ölçütü önceden ilan et, ölçümü herkese aynı biçimde uygula, sonucu sınırlılıklarıyla birlikte yayımla.

Elinizdeki rapor bu yolda atılmış ilk adım: üç platformu, Raydorf Enstitüsü'nün hazırladığı 14 ortak soru üzerinden inceleyen bir pilot değerlendirme. Platformları aynı sorularla, aynı koşullar altında, önceden sabitlenmiş metrikler ve üç ayrı hakem modeliyle değerlendirdik; yöntemi, veri toplama tarihlerini ve çalışmanın sınırlılıklarını beyan ettik. Pilot ölçeğin doğal sonucu olarak rapor bir sıralama iddiası taşımıyor. Amacımız yöntemi sahada sınamak ve uygulayıcıya, kendi kullanım senaryosuna göre okuyabileceği ilk bulguları sunmak. Seçim rehberi bölümü de bu nedenle tek bir kazanan ilan etmiyor.

Bir sonraki aşamada kapsamı genişletiyoruz. Piyasanın önde gelen hukuk yapay zekâsı şirketlerini aynı çerçevede buluşturacak bir değerlendirme sürecine geçiyoruz. Bu süreçte performans ölçütleri katılımcı şirketler ve uygulayıcı hukukçularla ortaklaşa belirlenecek, nihai hâlini Raydorf Enstitüsü onaylayacak; ölçüm ve analizleri ise yalnızca Raydorf yürütecek. Ölçütün belirlenmesine katılmak, sonucun belirlenmesine katılmak anlamına gelmeyecek.

Raydorf Enstitüsü'nün görevi, yapay zekânın profesyonel hizmetlerdeki kullanımını ölçülebilir ve denetlenebilir kılmak. Hukukçuların her gün başvurduğu araçların tarafsız değerlendirmesini bu görevin doğal bir parçası sayıyor, bunu sürdürülebilir bir dönemsellikle yapmayı taahhüt ediyoruz. Değerlendirme yöntemimize yönelik her eleştiri ve katkı sonraki dönemleri daha iyi yapacaktır. Kapımız buna açık.

Dr. Taylan Yıldız

Başkan, Raydorf Enstitüsü

§ I · ÖZET

1. Özet

Türkiye'de hukuk yapay zekâsı araçlarının doğruluğunu ölçen ortak bir standart bulunmuyor. Bu rapor, böyle bir standarda giden sürecin ilk adımı olarak yürütülen pilot değerlendirmenin sonuçlarını sunuyor.

Değerlendirmede Leagle, Apilex ve Dejure.ai, Raydorf Enstitüsü'nün hazırladığı 14 ortak araştırma sorusu üzerinden incelendi. Sorular dernekler, çevre, fikri mülkiyet, vergi, şirketler, kira, iş, bankacılık ve icra, ceza, aile hukuku ile temel haklara yayılıyor. Yanıtlar Leagle ve Apilex için 4 Ağustos 2026'da, Dejure.ai için 24 Haziran 2026'da, platformların o tarihteki canlı sürümlerinden toplandı. Ölçüm iki katmanlı: RAGAS çerçevesi yanıtın geri getirilen kaynaklarla ilişkisine bakıyor, hukuka özgü cetvel ise atıf doğruluğunu, karar güncelliğini, kapsayıcılığı ve ihtiyatı ölçüyor. Her iki katman üç ayrı hakem modeliyle puanlandı.

Sonuçlar okunurken üç sınırlılık göz önünde tutulmalı. 14 soruluk örneklemde tek bir sorudaki fark, platform ortalamasını 0.07 puana kadar oynatabiliyor. Bulgular yalnızca toplama tarihlerindeki platform sürümleri için geçerli. Puanlama ise insan uzman yerine büyük dil modeli hakemlerle yapıldı. Bu nedenle platformlar arasındaki küçük farklar bir sıralama olarak okunmamalı.

Pilotun en belirgin bulguları platformlar arasındaki farklardan çok, üç araçta ortak görülen örüntülerde toplanıyor: atıf doğruluğu üç araçta da cetvelin en zayıf iki boyutu arasında yer alıyor; yanıttaki önermelerin kaynağa tam bağlanması üç araçta da sınırlı kalıyor (Response Groundedness 0.60 ile 0.64 arası); bazı metriklerde ise hakem modelinin seçimi puanı platform seçiminden daha fazla etkiliyor. Ayrıntılar Bölüm 7'de.

§ II · SISTEMLER

2. Değerlendirilen Sistemler

- *Leagle:* Türk hukuk araştırma platformu; yanıtlar 4 Ağustos 2026 tarihinde, platformun o tarihteki canlı sürümünden toplanmıştır.
- *Apilex:* Türk hukuk araştırma platformu; yanıtlar 4 Ağustos 2026 tarihinde, platformun o tarihteki canlı sürümünden toplanmıştır.
- *Dejure.ai:* Türk hukuk araştırma platformu; yanıtlar 24 Haziran 2026 tarihinde, platformun o tarihteki canlı sürümünden toplanmıştır.

Piyasada faaliyet gösteren Hammurabi ve Onedocs platformları, yanıt toplama sırasında yaşanan teknik sorunlar nedeniyle karşılaştırılabilir ve puanlanabilir sonuç elde edilemediğinden aşağıdaki hiçbir karşılaştırma tablosuna dâhil edilmemiştir.

§ III · VERI KÜMESI

3. Veri Kümesi

Kıyaslama, Türkçe hukuki materyalden oluşan bir derlem ile bu değerlendirme için özel olarak hazırlanmış araştırma nitelikli sorular üzerine kuruludur.

- *Soru havuzu:* Raydorf Enstitüsü tarafından hazırlanmış 14 araştırma sorusundan oluşur. Sorular; dernekler, çevre, fikri mülkiyet, vergi, şirketler, kira, iş, bankacılık ve icra, ceza, aile hukuku ile temel haklara yayılır. Aynı soru metinleri üç platforma yöneltilmiş; her platformun yanıtı, kullanıcıya sunduğu referanslar ve bu referansların metin içerikleri kayıt altına alınmıştır.
- *Hukuki alan kapsamı:* Sorular Türk hukukunun birden fazla alanını kapsamaktadır: Adli Yargı (Mevzuat, Yargıtay, Bölge Adliye Mahkemeleri), İdari Yargı (Danıştay) ve düzenleyici otoriteler (SPK, BDDK, KVKK vb.).

§ IV · METODOLOJI

4. Metodoloji

4.1 Yanıt toplama

Her platform, aynı soru kümesini kendi standart arayüzü üzerinden almıştır. Her yanıt için yanıt metninin tamamı, geri getirilen kaynak belgeler (bağlamlar) ve belge başına üstveri (mahkeme, daire, esas/karar numaraları, mevzuat tanımlayıcıları) kaydedilmiştir.

4.2 RAGAS metrikleri

Dört RAGAS metriği hesaplanmıştır. Her metrik GPT-5.6 Luna, Gemini 3.5 Flash ve DeepSeek V4 Flash tarafından ayrı ayrı değerlendirilmiş; rapordaki ana değerler bu üç hakemin eşit ağırlıklı ortalaması olarak verilmiştir. Answer Relevancy gömme tabanlıdır ve text-embedding-3-small ile text-multilingual-embedding-002 modelleri kullanılmıştır; diğer metrikler geri getirilen bağlam ile yanıt arasındaki ilişkiyi değerlendirir. Her metrik 0–1 aralığındadır; yüksek değer daha iyidir.

METRIK	NE ÖLÇER	HESAPLAMA
Faithfulness	Yanıttaki ifadelerin geri getirilen bağlam tarafından desteklenme oranı; desteklenmeyen (halüsinasyon) iddiaları cezalandırır.	desteklenen / toplam ifade
Answer Relevancy	Yanıtın soruyu ne kadar doğrudan karşıladığı; konu dışı veya kaçamak yanıtları cezalandırır (yanlış yanıtları değil).	yeniden üretilen soruların kosinüs benzerliği × kaçamaklık kapısı
Response Groundedness	Daha ince taneli temellendirme: her önermenin bağlama ne ölçüde bağlı kaldığı; kısmi puan verilir.	2 hakem ortalaması (0/1/2) ÷ 2
Context Relevance	Geri getirme adımının kalitesi: getirilen belgelerin soruyla ilgililiği.	2 hakem ortalaması (0/1/2) ÷ 2

Yapılandırma: RAGAS 0.3.1 · üç bağımsız hakem modeli · embedding modelleri: text-embedding-3-small ve text-multilingual-embedding-002 · ana tablolarda eşit ağırlıklı hakem ortalaması · 0–1 normalize puan.

4.3 Hukuka özgü değerlendirme cetveli

Değerlendirmede PlawBench (arXiv:2601.16669) ve LexRubric (arXiv:2606.09389) çalışmalarından uyarlanan hukuka özgü bir cetvel kullanılmıştır. Atıf doğruluğu, karar güncelliği, hukuki kapsayıcılık ve yanıt ihtiyatı; geri getirilen kaynak materyale karşı üç hakem modeli tarafından 1–5 ölçeğinde puanlanmış, ardından 0–1 aralığına normalize edilmiştir.

METRIK	NE ÖLÇER	ÖLÇEK
Atıf Doğruluğu	Atıf yapılan kararların gerçek olup olmadığı, doğru mahkeme/daireye bağlanıp bağlanmadığı ve geri getirilen kaynaklardan izlenebilirliği.	1 (uydurma/yanlış atıf) ile 5 (tam izlenebilir, doğru) arası
Karar Güncelliği	Yanıtın güncelliğini yitirmiş veya değiştirilmiş içtihat yerine güncel kararlara dayanıp dayanmadığı.	1 (yalnız 2015 öncesi) ile 5 (güncel öncelikli, tarihler belirtilmiş) arası
Hukuki Kapsayıcılık	Yanıtın ilgili mevzuatı, temel ilkeleri, istisnaları, koşulları ve usulü kapsayıp kapsamadığı.	1 (temel meseleyi ıskalar) ile 5 (kapsamlı) arası
Yanıt İhtiyatı	İhtilaflı hukukun yerleşik hukuk gibi sunulmayıp belirsizliğin uygun biçimde ifade edilip edilmediği.	1 (aşırı iddialı) ile 5 (kalibre, gelişen alanları işaretler) arası

Yapılandırma: GPT-5.6 Luna, Gemini 3.5 Flash ve DeepSeek V4 Flash · boyut başına 1–5 · normalize puan = ham ÷ 5 · ana tablolar = üç hakemin eşit ağırlıklı ortalaması.

§ V · SINIRLILIKLAR

5. Sınırlılıklar

Bulgulara geçmeden önce, sonuçların hangi koşullar altında okunması gerektiğini belirleyen sınırlılıklar aşağıda yer alıyor.

- *Örneklem büyüklüğü:* Ortak küme 14 sorudan oluşmaktadır. Bu ölçekte tek bir sorudaki performans farkı platform ortalamasını yaklaşık 0.07 puana kadar etkileyebilir; platformlar arasındaki küçük farklar (±0.02) anlamlı bir sıralama farkı olarak yorumlanmamalıdır.
- *Sürüm ve veri toplama tarihleri:* Sonuçlar, yanıtların toplandığı tarihlerde kullanıcıya sunulan canlı platform sürümlerini yansıtıyor; bu ürün sınıfı hızla güncellendiğinden bulgular sonraki sürümlere genellenmemeli. Leagle ve Apilex yanıtları 4 Ağustos 2026, Dejure.ai yanıtları 24 Haziran 2026 tarihinde toplanmıştır. Aradaki yaklaşık altı haftalık fark, karşılaştırmanın eşzamanlılığını sınırlamaktadır; Dejure.ai sonuçları bu tarih farkı gözetilerek değerlendirilmelidir.
- *Otomatik hakem değerlendirmesi:* Puanlama, insan uzman incelemesi yerine büyük dil modeli hakemlerle yapılmıştır. Bölüm 6.2 ve 6.3'teki ek tablolar, hakemler arasında metrik bazında kayda değer varyans bulunduğunu göstermektedir. Üç hakemin eşit ağırlıklı ortalaması bu varyansı azaltmakta ancak ortadan kaldırmamaktadır; sonuçlar mesleki değerlendirmenin yerine geçmez.
- *Kapsam:* Hammurabi ve Onedocs platformları, Bölüm 2'de açıklanan teknik nedenlerle karşılaştırma dışında kalmıştır; bu rapor Türk hukuk yapay zekâsı pazarının tamamını temsil etmemektedir.

6. Bulgular

6.1 Birebir karşılaştırma: 14 soruluk ortak küme

Normalize puanlar 0.00–1.00 aralığındadır; yüksek değer daha iyidir. RAGAS hücreleri üç hakem modelinin eşit ağırlıklı ortalamasıdır. Bileşik gösterge = (RAGAS geneli + cetvel geneli) ÷ 2'dir.

METRIK	LEAGLE	APILEX	DEJURE.AI
RAGAS metrikleri			
Faithfulness	0.90	0.54	0.92
Answer Relevancy	0.76	0.78	0.73
Response Groundedness	0.62	0.64	0.60
Context Relevance	0.75	0.61	0.82
RAGAS geneli (4 metrik ort.)	**0.76**	**0.64**	**0.77**
Hukuka özgü cetvel (normalize)			
Atıf Doğruluğu	0.84	0.86	0.78
Karar Güncelliği	0.90	0.87	0.84
Hukuki Kapsayıcılık	0.91	0.87	0.84
Yanıt İhtiyatı	0.88	0.92	0.75
Cetvel geneli (4 boyut ort.)	**0.88**	**0.88**	**0.80**
Bileşik genel puan	**0.82**	**0.76**	**0.78**

Not: Hakem-model düzeyindeki farklılıklar Ek Tablo 6.2 ve 6.3'te ayrıca gösterilmiştir. Ana tablodaki ortalamalar yuvarlanmamış değerler üzerinden hesaplanmış, ardından iki ondalık basamağa yuvarlanmıştır; bu nedenle ek tablolardaki yuvarlanmış satırların ortalamasıyla ±0.01 düzeyinde fark görülebilir. Bileşik puanlar arasındaki farklar, Bölüm 5'te belirtilen örneklem sınırlılığı nedeniyle bir sıralama olarak okunmamalıdır.

6.2 Hakem-model duyarlılığı: ek RAGAS tablosu

Aşağıdaki ek tablo, 14 soruluk aynı kümede her hakem modelinin RAGAS puanlarını ayrı gösterir. Ana karşılaştırmadaki puanlar bu satırların eşit ağırlıklı ortalamasıdır.

HAKEM	PLATFORM	FAITHFULNESS	RELEVANCY	GROUNDEDNESS	CONTEXT REL.
GPT-5.6 Luna	Leagle	0.94	0.76	0.48	0.70
GPT-5.6 Luna	Apilex	0.40	0.78	0.55	0.57
GPT-5.6 Luna	Dejure.ai	0.93	0.74	0.48	0.82
Gemini 3.5 Flash	Leagle	0.79	0.87	0.44	0.61
Gemini 3.5 Flash	Apilex	0.39	0.87	0.46	0.55
Gemini 3.5 Flash	Dejure.ai	0.88	0.85	0.39	0.64
DeepSeek V4 Flash	Leagle	0.97	0.66	0.93	0.93
DeepSeek V4 Flash	Apilex	0.83	0.69	0.91	0.71
DeepSeek V4 Flash	Dejure.ai	0.94	0.60	0.93	0.98

6.3 Hakem-model duyarlılığı: hukuka özgü cetvel

Aşağıdaki ek tablo, aynı üç hakemin cetvel boyutlarındaki sonuçlarını gösterir. Ana cetvel değerleri bu satırların eşit ağırlıklı ortalamasıdır.

HAKEM	PLATFORM	ATIF	GÜNCELLIK	KAPSAYICILIK	İHTIYAT
GPT-5.6 Luna	Leagle	0.77	0.81	0.80	0.76
GPT-5.6 Luna	Apilex	0.83	0.85	0.79	0.88
GPT-5.6 Luna	Dejure.ai	0.77	0.75	0.64	0.56
Gemini 3.5 Flash	Leagle	0.90	0.97	1.00	1.00
Gemini 3.5 Flash	Apilex	1.00	0.93	1.00	1.00
Gemini 3.5 Flash	Dejure.ai	0.89	0.89	1.00	0.87
DeepSeek V4 Flash	Leagle	0.84	0.91	0.91	0.87
DeepSeek V4 Flash	Apilex	0.76	0.84	0.83	0.87
DeepSeek V4 Flash	Dejure.ai	0.70	0.89	0.87	0.81

6.4 Operasyonel özellikler

ÖZELLIK	LEAGLE	APILEX	DEJURE.AI
Ortak soru sayısı	14	14	14
Ort. getirilen bağlam sayısı	35.64	34.93	31.07
Ort. yanıt uzunluğu (karakter)	20 512	28 240	7 904
Metin içi atıf	Var	Var	Var
Kaynak ayrıntılarına erişim	Var	Var	Var
Yanıt raporu dışa aktarma	Var	Var	Var
Kaynak içeriği kaydı	Tam	Tam	Tam

§ VII · ÖRÜNTÜLER

7. Gözlenen Örüntüler

Bu bölüm önce üç araçta ortak görülen örüntüleri, ardından platform profillerini ele alıyor. Pilot ölçekte ortak örüntüler, platformlar arasındaki küçük farklardan daha sağlam bir bilgi sunuyor.

- *Atıf doğruluğu ortak zayıf nokta:* Atıf doğruluğu Leagle (0.84) ve Apilex'te (0.86) cetvelin en düşük boyutu, Dejure.ai'de (0.78) ise en düşük ikinci boyut. Kaynağın var olması yetmiyor; yanıttaki hukuki iddiayla doğru eşleşip eşleşmediği ayrıca denetlenmeli. Özellikle erişim kısıtlı kaynaklarda birincil belge doğrulaması zorunlu.
- *Kaynağa bağlılık üç araçta da sınırlı:* Response Groundedness üç platformda da 0.60 ile 0.64 arasında kalıyor; Leagle ve Dejure.ai'de RAGAS katmanının en düşük metriği bu. Değer hakemlere göre belirgin biçimde ayrışıyor: GPT-5.6 Luna ve Gemini 3.5 Flash 0.39 ile 0.55 arasında puan verirken DeepSeek V4 Flash 0.91 ile 0.93 arasında puan veriyor. Bulgu, iki hakemin üç platformda da tutarlı biçimde gösterdiği eğilime dayanıyor.
- *Hakem seçimi, bazı metriklerde platform seçiminden daha belirleyici:* Leagle'ın Response Groundedness puanı hakeme göre 0.44 ile 0.93 arasında değişiyor; aynı hakem altında platformlar arasındaki fark ise bu metrikte 0.07'yi aşmıyor. Apilex'in Faithfulness puanı da hakeme göre 0.39 ile 0.83 arasında oynuyor. Tek hakemli değerlendirmelerin yanıltıcı olabileceğini gösteren bu bulgu, ölçütlerin ve hakem kalibrasyonunun ortaklaşa belirlenmesini sonraki dönemin öncelikli işi hâline getiriyor.
- *Tavan etkisi:* Gemini 3.5 Flash, hukuki kapsayıcılıkta üç platforma da, yanıt ihtiyatında ikisine tam puan (1.00) verdi. Tam puan kümelenmesi, bu hakemin söz konusu boyutlarda platformları ayırt edemediğine işaret ediyor.
- *Yanıt uzunluğu ve kaynağa sadakat:* En uzun yanıtları üreten Apilex (ortalama 28 240 karakter) en düşük Faithfulness değerini (0.54), en kısa yanıtları üreten Dejure.ai (7 904 karakter) en yüksek değeri (0.92) aldı; Leagle (20 512 karakter, 0.90) ikisinin arasında. Üç veri noktasına dayanan bu ilişki bir hipotez olarak okunmalı ve genişletilmiş dönemde sınanacak. Uzunluk tek başına kalite ölçüsü değil; kapsam ve bağlam desteğiyle birlikte değerlendirilmeli.
- *Platform profilleri:* Hukuka özgü cetvelde Leagle ve Apilex başa baş (0.88); Leagle kapsayıcılık (0.91) ve güncellikte (0.90), Apilex atıf doğruluğu (0.86) ve ihtiyatta (0.92) önde. RAGAS genelinde Dejure.ai (0.77) ve Leagle (0.76) yakın, Apilex (0.64) geride; Dejure.ai Context Relevance'ta (0.82), Apilex Answer Relevancy (0.78) ve Response Groundedness'ta (0.64) öne çıkıyor. Bileşik göstergedeki farklar (0.04 ile 0.06 arası) örneklem büyüklüğü dikkate alındığında kesin bir sıralama oluşturmuyor.

§ VIII · SEÇIM REHBERI

8. Hukukçular İçin Kullanım ve Seçim Rehberi

Bu bölüm, bulguları platformların hedef kullanıcıları olan hukukçular açısından pratik bir seçim çerçevesine dönüştürmektedir. Eşit ağırlıklı bileşik gösterge hızlı bir özet sağlar; ancak tek başına platform seçimi için yeterli değildir. Üç platform farklı kullanım senaryolarında farklı güçlü yönler sergilemektedir. Aşağıdaki eşleştirmeler pilot bulgulara dayanıyor ve ilk yönlendirme niteliği taşıyor; kesin seçim için Bölüm 8.3'teki pilot yöntemi öneriliyor.

8.1 Kullanım senaryosuna göre platform uyumu

- *Kapsamlı içtihat taraması ve mütalaa hazırlığı yapıyorsanız:* hukuki kapsayıcılık (0.91), karar güncelliği (0.90) ve RAGAS geneli (0.76) birlikte değerlendirildiğinde Leagle güçlü bir seçenektir. Yine de her kritik iddianın kaynak belgesiyle doğrulanması gerekir.
- *Yanıt içeriğinin doğrudan soruya temasını ve ihtiyatlı bir hukuki taslağı önceliyorsanız:* Apilex, hukuka özgü cetvel genelinde (0.88), Answer Relevancy'de (0.78) ve Response Groundedness'ta (0.64) öne çıkmaktadır. Faithfulness sonucu nedeniyle uzun yanıtlar kaynak metinle çapraz kontrol edilmelidir.
- *Bağlama yüksek oranda bağlı, daha kısa bir ilk yanıt arıyorsanız:* Dejure.ai Faithfulness (0.92) ve Context Relevance'ta (0.82) güçlüdür. Kritik kullanım öncesinde belge ayrıntılarına erişim ve birincil kaynak doğrulama imkânları ayrıca test edilmelidir.

8.2 Platformdan bağımsız kurallar

- *Her atfı birincil kaynaktan doğrulayın.* Cetvel sonuçları, atıf doğruluğunun platform seçiminden bağımsız olarak aktif bir kontrol gerektirdiğini göstermektedir. Dilekçe veya mütalaaya giren her esas/karar numarası, kaynağından teyit edilmelidir.
- *Platform çıktısını nihai görüş değil, araştırma taslağı olarak ele alın.* En güçlü kapsayıcılık puanları dahi istisnaların ve güncel gelişmelerin eksiksiz yakalandığını garanti etmemektedir; mesleki değerlendirme sorumluluğu uygulayıcıda kalmaktadır.
- *İhtilaflı ve gelişmekte olan alanlarda ihtiyat düzeyini kontrol edin.* Yanıt İhtiyatı puanları Leagle için 0.88, Apilex için 0.92 ve Dejure.ai için 0.75'tir; özellikle düşük puanlı profillerde belirsizlik ve karşıt içtihatların açıkça ele alınıp alınmadığı kontrol edilmelidir.
- *Kaynak erişim koşullarını edinim öncesinde netleştirin.* Bu çalışmada kullanıcıya sunulan referanslar ve metin içerikleri kayıt altındadır; canlı kullanımda da bu erişimin atıf doğrulaması ve kurum içi denetim için sürdürülebilirliği teyit edilmelidir.

8.3 Nereden başlamalı

Edinim kararı öncesinde önerilen yol, uygulayıcının kendi uzmanlık alanından, yanıtını bildiği beş ila on soruyla sınırlı bir pilot çalışmadır: aynı sorular aday platformlara yöneltilmeli; atıfların doğruluğu, kapsanan istisnalar ve içtihat güncelliği uygulayıcının kendi bilgisiyle karşılaştırılmalıdır. Bu yöntem, genel puan tablolarının gösteremeyeceği alan-özgü performans farklarını ortaya çıkarmakta ve seçimi kurumun fiili çalışma biçimine bağlamaktadır. Platform seçimi bir defalık karar değildir; platformlar hızla güncellendiğinden değerlendirmenin makul aralıklarla yinelenmesi yerinde olacaktır.

§ IX · BAĞIMSIZLIK

9. Bağımsızlık Beyanı

Bu değerlendirmeye dâhil edilme veya sıralama karşılığında hiçbir platformdan ücret alınmamıştır. Raydorf Enstitüsü, değerlendirilen platformların hiçbirinde ticari pay sahibi değildir. Metodoloji, metrikler ve hakem yönergeleri yanıt toplamadan önce sabitlenmiş; hiçbir platform sonuçların içeriği üzerinde editoryal etkiye sahip olmamıştır. Soru kümesi Raydorf Enstitüsü tarafından hazırlanmıştır.

§ X · EK

10. Ek: Metrik Tanımları ve Hakem Yönergeleri

A. RAGAS metrikleri

Faithfulness: Yanıtı atomik ifadelere ayrıştırır ve her ifadenin geri getirilen bağlam tarafından desteklenip desteklenmediğini hakeme sorar. Puan = desteklenen ifade / toplam ifade. Hakem yönergesi: "Judge the faithfulness of a series of statements based on a given context..."

Answer Relevancy: Hakem, yalnızca yanıtı görerek yanıtın büyük olasılıkla karşılık geldiği soruyu yeniden üretir (×3); her biri gömülür ve özgün soruyla kosinüs benzerliğiyle karşılaştırılır; yanıt üç üretimin tamamında kaçamak olarak işaretlenirse puan sıfırlanır. Puan = ort(kosinüs benzerliği) × kaçamaklık kapısı.

Response Groundedness (NVIDIA NV): Her önerme, bağlama karşı 0 (desteklenmiyor) / 1 (kısmen) / 2 (tamamen destekleniyor) olarak, 0,1 sıcaklıkta iki bağımsız hakem tarafından puanlanır; Puan = ort(hakem1, hakem2) ÷ 2.

Context Relevance (NVIDIA NV): Geri getirilen her parça, soruyla ilgililiği açısından iki bağımsız hakem tarafından 0/1/2 olarak puanlanır; Puan = ort(hakem1, hakem2) ÷ 2. Bu metrik üretimi değil, geri getirmeyi değerlendirir.

B. Hukuka özgü cetvel (ankraj tanımları)

Atıf Doğruluğu: 1: geri getirilen referanslarda bulunmayan kararlara atıf yapar veya yanlış atfeder; uydurma esas/karar numaraları içerir. 5: atıf yapılan her karar geri getirilen referanslardan tam izlenebilir, doğru mahkeme ve daireye atfedilmiş, esas/karar numaraları doğrudur; uydurma atıf yoktur.

Karar Güncelliği: 1: yalnızca 2015 öncesi kararlara dayanır veya daha yeni içtihat mevcutken güncelliğini yitirmiş bir pozisyonu yürürlükteki hukuk gibi sunar. 5: tutarlı biçimde en güncel kararları önceler, kilit karar yıllarını belirtir ve güncel uygulamayı artık yansıtmayabilecek eski içtihadı işaretler.

Hukuki Kapsayıcılık: 1: temel hukuki meseleyi ıskalar veya yalnızca tali bir yönü ele alır. 5: ilgili mevzuatı, öncü kararları, istisnaları, koşulları ve usul gereklerini kapsamlı biçimde ele alır; mahkeme içtihat hatları arasındaki ayrışmaları belirtir.

Yanıt İhtiyatı: 1: ihtilaflı veya gelişmekte olan pozisyonları hiçbir çekince belirtmeden yerleşik hukuk gibi sunar. 5: güveni, dayanılan otoritenin gücüne hassas biçimde kalibre eder, gelişen alanları açıkça işaretler ve gerektiğinde profesyonel danışmanlık önerir.

KÜNYE

Belge kodu: RDF-BMK-2026-1 · Sürüm 1.2 · Yayın tarihi: Eylül 2026 · Yayımcı: Raydorf Enstitüsü, İstanbul · raydorf.com

Önerilen atıf: Raydorf Enstitüsü (2026). *Pilot Değerlendirme Raporu: Türk Hukuk Yapay Zekâsı Araştırma Platformları* (RDF-BMK-2026-1, Sürüm 1.2). İstanbul: Raydorf Enstitüsü.

Bu pilot değerlendirmenin ardından, önde gelen hukuk yapay zekâsı şirketlerinin katılımıyla ve ortaklaşa belirlenen ölçütlerle genişletilmiş değerlendirme dönemine geçilecektir.

Raydorf Enstitüsü, uygunluk değerlendirmesi geleneğinde çalışan bağımsız bir belgelendirme kuruluşudur. Hukuk sektöründe kurumlara ve bireylere yönelik dört belgelendirme hizmeti sunuyoruz.

HUKUK BÜROLARI

Kurumsal Sertifikasyon ve Derecelendirme

Yedi boyutlu Raydorf Standardı ile bağımsız değerlendirme ve beş kademeli derecelendirme. Kademeyi en zayıf boyut belirler; toplam puan veya ortalama kullanılmaz. Sertifikasyon üç yıllık döngüde, yıllık ara doğrulamalarla sürdürülür.

AI-Aware · AI-Enabled · AI-Integrated · AI-First · AI-Native

HUKUK MÜŞAVIRLERI

Bireysel Yetkinlik Sertifikası

Uygulayıcının yapay zekâ yetkinliğini belgeleyen bireysel sertifika. Değerlendirme, standartlaştırılmış psikometrik sınav veya uzman görüşmesi yoluyla yapılır; sonuç dört yeterlilik bandından biriyle raporlanır.

Gelişmekte · Yetkin · İleri · Seçkin

HUKUK DEPARTMANLARI

Yapay Zekâ Kullanım Yeterlilik Sertifikası

Şirket içi hukuk departmanları için ekip düzeyinde yeterlilik değerlendirmesi. Departmanın yapay zekâ kullanım pratiği, yönetişim çerçevesiyle birlikte incelenir ve yeterlilik sertifikasıyla belgelenir.

Ekip değerlendirmesi · Yönetişim incelemesi · Departman sertifikası

İŞE ALIM

Yapay Zekâ Yetkinlik Testleri

İşe başvuran hukukçular için standartlaştırılmış yetkinlik testi. Adaylar aynı ölçekle değerlendirilir; işveren, karşılaştırılabilir bir puan raporu alır.

Standart ölçek · Karşılaştırılabilir puan raporu

İLETIŞIM

Belgelendirme süreçleri hakkında bilgi almak için yazınız.

contact@raydorf.com Raydorf Enstitüsü · İstanbul, Türkiye raydorf.com