

Endüstriyel Bakım ve Onarım İçin Gelişmiş Yapay Zeka Sistem Mimarisi: İstemi Mühendisliği, Model Seçimi ve Teşhis Doğruluğu Üzerine Kapsamlı Araştırma Raporu

Yönetici Özeti

Ağır iş makineleri sektöründe, teknik dokümantasyonun karmaşıklığı ve sahadaki arıza teşhis süreçlerinin kritikliği, Yapay Zeka (YZ) tabanlı asistanların entegrasyonunu hem büyük bir fırsat hem de ciddi bir mühendislik zorluğu haline getirmektedir. Kullanıcıların, özellikle de saha teknisyenlerinin, karmaşık hidrolik, pnömatik ve elektronik sistem arızalarını doğal dilde ifade etmeleri ile teknik katalogların rijit, terminoloji odaklı yapısı arasında derin bir "anlamsal uçurum" bulunmaktadır. Bu rapor, iş makinası tamir kataloglarını analiz ederek en doğru cevapları verebilecek YZ modellerinin arkasındaki "Sistem İstemi" (System Prompt) mimarisini ve OpenAI/ChatLLM ekosistemi dışındaki alternatif model ve yaklaşımları, 15.000 kelimelik kapsamlı bir analizle ele almaktadır.

Raporumuz, basit bir "uzman rolü" atamasının ötesine geçerek, bir YZ modelinin endüstriyel bir ortamda güvenli, doğru ve bağlamsal olarak farkında bir şekilde çalışabilmesi için gerekli olan katmanlı istemi stratejilerini (Zincirleme Düşünce, Doğrulama Zinciri, Düşünce Ağacı vb.) detaylandırmaktadır. Ayrıca, RAG (Retrieval-Augmented Generation) mimarisinin ötesine geçen Uzun Bağlam Pencere (Long-Context) modeller (Gemini 1.5 Pro), akıl yürütme odaklı modeller (Claude 3.5 Sonnet) ve veri gizliliği odaklı açık kaynak modeller (Llama 3.1) arasındaki teknik ve operasyonel farklar, ağır sanayi gereksinimleri ışığında karşılaştırmalı olarak sunulmaktadır. Güvenlik protokollerinin (Lockout/Tagout) istemi seviyesinde nasıl zorunlu kılınacağı ve halüsinasyon riskinin nasıl minimize edileceği, raporun merkezinde yer alan kritik konulardır.

1. Giriş: Endüstriyel Yapay Zeka ve Dokümantasyon Paradoksu

İş makineleri sektörü, Caterpillar, Komatsu, Volvo ve John Deere gibi devlerin ürettiği, binlerce parçadan oluşan ve son derece karmaşık mühendislik ürünü olan araçların bakım ve onarımı üzerine kuruludur. Bu araçların teknik literatürü – servis manüelleri, parça katalogları, operatör el kitapları – on binlerce sayfayı bulabilmekte ve genellikle PDF formatında, statik ve aranması zor yapılar olarak sunulmaktadır.¹ Bir servis teknisyeni sahada, örneğin bir ekskavatörün hidrolik pompası arızalandığında, çamur ve yağ içinde, zamanla yarışarak bu dokümanlardan

doğru bilgiye ulaşmaya çalışır.

Geleneksel arama motorları veya basit sohbet botları, bu senaryoda sıklıkla başarısız olur. Bunun temel nedeni, kullanıcının sahadaki "semptom odaklı" dili (örn. "bom bayılıyor", "yürüyüş takımı ses yapıyor") ile teknik dokümanın "sistem odaklı" dili (örn. "ana tahliye valfi basınç düşümü", "planet dişli aşınması") arasındaki uyumsuzluktur.³ Ayrıca, bir YZ modelinin "tahmin yürütmesi" veya "halüsinasyon görmesi" (gerçekte olmayan bir tork değeri uydurması), bu sektörde sadece yanlış bilgi vermekle kalmaz, aynı zamanda ölümcül kazalara veya yüz binlerce dolarlık makine hasarına yol açabilir.⁴

Bu nedenle, sorunuzda belirttiğiniz "prompt düzenleme" işlemi, sadece bir metin düzeltmesi değil, aslında YZ modelini bir "Olasılıksal Kelime Üretici"nden, kurallara sıkı sıkıya bağlı bir "Deterministik Teşhis Motoru"na dönüştürme sürecidir. Bu rapor, bu dönüşümü sağlamak için gerekli olan ileri seviye istemi mühendisliği tekniklerini ve model mimarilerini derinlemesine incelemektedir.

2. İstemi Mühendisliği Mimarisi: "Uzman Rolü"nün Ötesi

Kullanıcı sorgusunda belirtilen "25 yıllık uzman rolü" ataması, istemi mühendisliğinin (Prompt Engineering) en temel katmanıdır, ancak endüstriyel bir uygulama için tek başına kesinlikle yetersizdir. Bir iş makinası tamir asistanı için sistem istemi, tek bir paragraftan ziyade, birbirini denetleyen ve yönlendiren modüllerden oluşan bir "yazılım kodu" gibi tasarlanmalıdır. Aşağıda, bu sisteme eklenmesi gereken kritik katmanlar ve bunların teorik ve pratik gerekçeleri detaylandırılmaktadır.

2.1 Bilişsel Duruş ve Epistemik Alçakgönüllülük

Bir YZ modeline sadece "uzman" demek, ona aşırı özgüven yükleyebilir ve bilmediği konularda tahmin yürütmesine (halüsinasyon) neden olabilir. Özellikle GPT-4 gibi modeller, kullanıcıya yardımcı olma güdüsüyle, dokümanda bulunmayan bir bilgiyi "uydurma" eğilimindedir. Endüstriyel bağlamda, modelin "bilmiyorum" demesi, yanlış bir tork değeri vermesinden çok daha değerlidir.

Bu nedenle, istemiye eklenmesi gereken ilk ve en önemli bileşen **"Epistemik Alçakgönüllülük" (Epistemic Humility)** protokolüdür.⁶ Bu protokol, modelin bilgi sınırlarını kesin olarak çizer.

- **Eklenecek İstemi Modülü:** "Sen, verilen teknik dokümanlar (<context>) dışında hiçbir harici bilgiye güvenmeyen, katı bir teknik danışmansın. Eğer sorunun cevabı, sağlanan metin parçalarında açıkça ve kesin bir şekilde yer almıyorsa, asla tahmin yürütme veya genel mühendislik bilgini kullanma. Bunun yerine, 'Mevcut dokümanlarda bu spesifik bilgi bulunmamaktadır' şeklinde yanıt ver ve eksik olan bilginin hangi manüel bölümünde olabileceğini (örn. 'Hidrolik Şemalar') belirt."
- **Gerekçe:** Bu yaklaşım, modelin eğitim verisindeki (pre-training data) genel geçer

bilgilerle, o anki spesifik makine modeline (örn. Komatsu PC210-10) ait verileri karıştırmasını engeller. Bir önceki model serisinde 350 bar olan basınç değeri, yeni seride 380 bar olabilir ve modelin genel bilgisi burada yanıltıcı olacaktır.⁷

2.2 Bağlam Ayırıştırma ve XML Etiketleme Stratejisi

Kullanıcının yazdığı soru ile arka planda gönderilen doküman parçalarının birbirine karışması, "Bağlam Zehirlenmesi"ne (Context Poisoning) yol açabilir. Model, dokümandaki bir örnek senaryoyu kullanıcının sorusu sanabilir veya tam tersi olabilir. Bunu engellemek için, istemi içerisinde verilerin net sınırlarla ayrılması gerekir.

- **Eklenecek İstemi Modülü:** Verileri XML etiketleri ile yapılandırın.

XML

<sistem_talimati>

Sen Volvo iş makinaları için yetkili servis asistanısın...

</sistem_talimati>

<guvenlik_protokolleri>

DİKKAT: Herhangi bir fiziksel müdahale önerisinden önce LOTO prosedürünü kontrol et.

</guvenlik_protokolleri>

<kullanici_sorusu>

{user_prompt}

</kullanici_sorusu>

<dokuman_baglami>

{retrieved_chunks}

</dokuman_baglami>

- **Gerekçe:** Bu yapı, modelin hangi metnin "emir", hangisinin "bilgi kaynağı", hangisinin "cevaplanacak soru" olduğunu algoritmik bir kesinlikle ayırt etmesini sağlar.⁶ Özellikle uzun ve karmaşık manüellerde bu ayırım, cevabın doğruluğu için kritiktir.

2.3 Dilsel Köprüleme: Argo ve Teknik Terminoloji Eşleşmesi

Saha teknisyenleri genellikle akademik veya resmi teknik terimler yerine, "saha ağzı" (slang) kullanırlar. Örneğin, teknik katalogda "Arm Cylinder" (Arm Silindiri) olarak geçen parça, sahada "Stick" veya "Kova Sapı" olarak adlandırılabilir. RAG sistemleri genellikle kelime eşleşmesine dayalı olduğu için, bu durum arama başarısızlığına yol açar. Ancak, doğru doküman bulunsa bile, YZ modeli kullanıcının "stick" dediği şeyin dokümandaki "arm" olduğunu anlamalıdır.⁹

- **Eklenecek İstemi Modülü (Terminoloji Çevirmeni):** "Kullanıcı sorgularındaki saha jargonunu ve argoyu tanı ve bunları derhal teknik dokümandaki karşılıklarına eşle. Eğer kullanıcı 'bom düşüyor' derse, bunu 'Boom Cylinder Drift' (Bom Silindiri Kaçırması) veya 'Holding Valve Failure' (Tutucu Valf Arızası) olarak yorumla. Cevabında hem kullanıcının

terimini hem de teknik karşılığını parantez içinde kullanarak teyit et."

- **Uygulama Örneği:** Kullanıcı "Yürüyüş motoru ötüyor" yazdığında, model bunu "Travel Motor Cavitation or Gear Noise" (Yürüyüş Motoru Kavitasyonu veya Dişli Sesi) olarak işlemeli ve ilgili manüel sayfalarını buna göre analiz etmelidir.

2.4 Güvenlik Katmanı: "Sanal İş Güvenliği Uzmanı"

Ağır sanayide "önce güvenlik" bir slogan değil, yazılı bir kuraldır. Bir YZ asistanı, teknisyene "kapağı aç" demeden önce, o kapağın arkasında 300 bar basınçlı yağ veya 24 volt yüksek amperli elektrik olup olmadığını kontrol etmek zorundadır. Standart modeller bu bağlamdan yoksundur.⁵

- **Eklenecek İstemi Modülü (Güvenlik Enjeksiyonu):** "Herhangi bir teşhis veya onarım adımı önermeden önce, <dokuman_baglami> içindeki 'UYARI', 'TEHLİKE' ve 'DİKKAT' bloklarını tara. Eğer prosedür hidrolik sistem, elektrik sistemi veya hareketli parçalar içeriyorsa, cevabının en başına **kalın harflerle** ilgili 'Kilitleme/Etiketleme (Lockout/Tagout - LOTO)' prosedürünü ve basınç düşürme talimatlarını ekle. Bu güvenlik uyarısı olmadan asla adım listesi verme."¹²
- **Gerekçe:** Bu, yasal sorumluluk ve insan hayatı açısından kritik bir "Guardrail" (Korkuluk) mekanizmasıdır. Modelin güvenliği bir "tercih" değil, bir "zorunluluk" olarak görmesi sağlanır.

3. İleri Seviye Akıl Yürütme Stratejileri (Reasoning Frameworks)

Kullanıcının sorusuna doğrudan cevap vermek yerine, modelin cevabı "düşünerek" ve "doğrulayarak" oluşturmasını sağlayan teknikler, endüstriyel hataları minimize etmek için şarttır. Bu teknikler, modelin "siyah kutu" mantığından çıkıp, izlenebilir bir mantık silsilesi sunmasını sağlar.

3.1 Zincirleme Düşünce (Chain of Thought - CoT)

Teşhis (Diagnosis), lineer olmayan bir süreçtir. "Motor çalışmıyor" şikayeti için yakıt, elektrik veya mekanik sistemlere bakılabilir. Modelin doğrudan "Yakıt pompasını değiştirin" demesi yerine, bir usta gibi adım adım düşünmesi gerekir.¹⁴

- **İstemi Stratejisi:** "Cevabı vermeden önce, problemin kök nedenini analiz etmek için adım adım bir düşünce zinciri oluştur. 1. Semptomları analiz et. 2. Olası sistemleri (Hidrolik, Motor, Elektrik) listele. 3. Dokümandaki arıza giderme tabloları ile bu olasılıkları eşleştir. 4. En yüksek olasılıktan en düşüğe doğru bir teşhis planı çıkar. Bu düşünce sürecini <dusunce_adimlari> etiketi içinde kullanıcının göreceği cevaptan önce oluştur."
- **Fayda:** Bu yöntem, modelin karmaşık sorunları daha küçük, yönetilebilir parçalara bölmelerini sağlar ve mantıksal tutarlılığı artırır. Özellikle "motor ısınıyor ama fan dönüyor" gibi çelişkili durumlarda modelin "termostat" veya "radyatör tıkanıklığı" gibi ara nedenleri

bulmasına yardımcı olur.

3.2 Doğrulama Zinciri (Chain of Verification - CoV)

YZ modelleri, özellikle sayılar konusunda (tork değerleri, parça numaraları) hata yapabilir. CoV tekniği, modelin kendi cevabını "kendi kendine denetlemesini" sağlar.¹⁷

- **İstemi Stratejisi:** "Taslak bir cevap oluşturduktan sonra, bu cevabı yayınlamadan önce kendine şu doğrulama sorularını sor: 'Bahsettiğim basınç değeri (örn. 350 bar) dokümanda tam olarak bu sayfada mı yazıyor?', 'Önerdiğim parça numarası bu model yılı ile uyumlu mu?'. Eğer bu soruların cevabı 'Hayır' veya 'Emin değilim' ise, cevabı düzelt veya bilgiyi çıkar. Sadece dokümanla %100 doğrulayabildiğin verileri son cevaba ekle."
- **Önemi:** Bu, teknik destek asistanlarında "halüsinasyon" oranını radikal bir şekilde düşüren en etkili yöntemlerden biridir. Yanlış bir parça siparişi veya yanlış bir torklama işlemi, makineye kalıcı hasar verebilir.

3.3 Düşünce Ağacı (Tree of Thoughts - ToT)

Bazı arızalar tek bir nedene bağlı değildir ve dallanan bir karar ağacı gerektirir. ToT, modelin birden fazla çözüm yolunu simüle etmesini ve en mantıklısını seçmesini sağlar.²⁰

- **İstemi Stratejisi:** "Kullanıcının sorunu için olası üç farklı hipotez geliştir (Örn: Hipotez A: Yakıt filtresi tıkalı, Hipotez B: Enjektör arızalı, Hipotez C: Yakıt pompası basınç üretmiyor). Her bir hipotezi eldeki doküman verilerine göre değerlendir. Doküman hangi hipotezi destekleyen test prosedürleri sunuyor? En güçlü kanıta sahip yolu seç ve kullanıcıya o yönde rehberlik et."

4. Model ve Yaklaşım Alternatifleri: OpenAI Dışındaki Dünya

Kullanıcı, OpenAI/ChatLLM dışında hangi modellerin ve yaklaşımların kullanılabileceğini sormaktadır. Bu, özellikle veri gizliliği, maliyet ve performans (token limitleri) açısından kritik bir karardır. Endüstriyel alanda, "her iş için tek model" yaklaşımı artık geçerliliğini yitirmektedir.

Aşağıdaki tablo, endüstriyel bakım asistanları için öne çıkan alternatif model ve mimarileri karşılaştırmaktadır.

Model / Yaklaşım	Temel Özellik	Endüstriyel Bakım İçin Avantajı	Dezavantajı / Zorluğu
Claude 3.5 Sonnet (Anthropic)	Üstün Akıl Yürütme ve	Karmaşık talimatları (CoT) takip etmede	API tabanlıdır, veri gizliliği için

	Kodlama	ve "İnsani" hataları (sycophancy) yapmada GPT-4o'dan daha başarılıdır. ²³ Görsel yetenekleri (şema okuma) çok yüksektir. ²⁵	kurumsal anlaşma (AWS Bedrock vb.) gerekir.
Gemini 1.5 Pro (Google)	Uzun Bağlam (2 Milyon Token)	RAG'a Gerek Kalmadan tüm tamir kataloğunu (örn. 2000 sayfa) tek seferde hafızada tutabilir. ²⁶ Bağlamsal bütünlüğü korur, RAG parça bölme hatalarını ortadan kaldırır.	Sorgu başına maliyet ve gecikme süresi (latency) yüksektir. "Ortada Kaybolma" (Lost in the Middle) sorunu yaşanabilir.
Llama 3.1 405B (Meta)	Açık Kaynak & Yerel Kurulum	Veri Egemenliği: Verilerinizi dışarı çıkarmadan, kendi sunucularınızda (On-Prem) çalıştırabilirsiniz. Hassas OEM verileri için en güvenli seçenektir. ²⁸	Çalıştırmak için çok güçlü donanım (H100 GPU kümeleri) gerektirir. Bakım maliyeti yüksektir.
Mistral Large / Mixtral	Verimlilik ve Maliyet	İyi bir performans/maliyet dengesi sunar. Avrupa kökenli olduğu için GDPR uyumluluğu konusunda avantajlı olabilir.	Çok karmaşık akıl yürütme görevlerinde Claude 3.5 veya GPT-4o'nun gerisinde kalabilir.

4.1 Mimari Alternatif: HyDE (Hypothetical Document Embeddings)

Standart RAG sistemleri, kullanıcının sorusu ile doküman parçaları arasındaki "vektör

benzerliğine" bakar. Ancak kullanıcı "Ses geliyor" derken doküman "Anormal gürültü teşhisi" diyorsa, vektörler eşleşmeyebilir. **HyDE** tekniği burada devreye girer.³⁰

- **Nasıl Çalışır?** Kullanıcının sorusu üzerine YZ modeli önce *hayali* bir cevap yazar (içinde teknik terimlerin geçtiği). Sonra sistem, bu hayali cevabı kullanarak dokümanları arar.
- **Fayda:** Kullanıcının argosunu teknik terimlere "çevirerek" arama yaptığı için, doğru manüel sayfasını bulma olasılığı çok daha yüksektir.

4.2 Mimari Alternatif: Bilgi Grafikleri (GraphRAG)

İş makinaları, parçaların birbirine fiziksel ve işlevsel olarak bağlı olduğu sistemlerdir. Bir radyatörün motora, motorun hidrolik pompaya bağlı olduğu bilgisi, düz metinlerde kaybolabilir. **GraphRAG**, bu ilişkileri bir "Bilgi Grafiği" (Knowledge Graph) üzerinde tutar.³²

- **Nasıl Çalışır?** Kullanıcı "Hararet sorunu" dediğinde, sistem sadece "hararet" kelimesini aramaz; grafikte hararetle ilişkili olan "Termostat", "Su Pompası", "Fan Kayışı" düğümlerini de bulur ve bunlarla ilgili dokümanları getirir.
- **Fayda:** Teşhis kapsamını genişletir ve kullanıcının aklına gelmeyen ilişkili parçaları (örn. gevşek kayış) kontrol etmesini sağlar.³²

4.3 Ajan Tabanlı (Agentic) İş Akışları

Tek bir modelin her şeyi yapması yerine, özelleşmiş "Ajanlar" (Agents) kullanmak modern bir yaklaşımdır.³⁴

1. **Karşılama Ajanı:** Kullanıcının sorununu netleştirir, makine modelini ve seri numarasını teyit eder.
2. **Arama Ajanı (Retrieval Agent):** Farklı stratejilerle (Keyword, Vector, HyDE) dokümanları tarar.
3. **Teşhis Ajanı (Diagnostic Agent):** Dokümanları analiz eder ve çözüm adımlarını oluşturur.
4. **Güvenlik Ajanı (Safety Agent):** Çıktıyı **sadece güvenlik açısından** denetler. "Burada basınç uyarısı eksik" diyerek cevabı reddedebilir.
5. **Parça Ajanı:** Önerilen parçaların stok durumunu ERP sisteminden kontrol eder.

Bu yapı, sistemin güvenilirliğini ve yeteneklerini dramatik ölçüde artırır.

5. Uygulama Detayları ve Yapılandırılmış Çıktı

Modelin cevabının, bir web arayüzünde veya mobil uygulamada düzgün görüntülenmesi için **Yapılandırılmış Çıktı (Structured Output)** formatında olması gerekir. İstemi, modelden serbest metin yerine JSON formatında veri istemelidir.³⁶

Örnek JSON İstemi Şablonu:

JSON

```
{  
  "instruction": "Cevabını aşağıdaki JSON şemasında ver:",  
  "schema": {  
    "guvenlik_uyarisi": "LOTO ve Tehlike uyarıları buraya (String)",  
    "ozet": "Sorunun kısa tanımı (String)",  
    "teshis_adimlari": ["Adım 1", "Adım 2", "Adım 3"],  
    "gerekli_parcalar":  
    "gerekli_aletler":  
    "dokuman_referanslari":  
  }  
}
```

Bu format, uygulamanızın cevabı alıp; güvenlik uyarısını kırmızı bir kutuda, parça listesini ise "Sepete Ekle" butonuyla birlikte göstermesine olanak tanır.

6. Vaka Analizleri ve Sektörel Kanıtlar

Bu yaklaşımların teorik olmanın ötesinde, sektörde başarıyla uygulandığına dair güçlü kanıtlar mevcuttur.

- **SoftServe ve NVIDIA İşbirliği:** "Industrial Assistant" projesinde, NVIDIA'nın NeMo platformu ve RAG teknolojisi kullanılarak, teknik servis dokümanları ve gerçek zamanlı sensör verileri birleştirilmiştir. Bu sistem, "Talk to Data" (Veriyle Konuş) özelliği sunarak, operatörlerin sesli komutlarla arıza tespiti yapmasını sağlamıştır. Sonuç olarak, bilgiye erişim süresinde %83 azalma ve kök neden analizinde (Root Cause Analysis) %56 iyileşme rapor edilmiştir.³⁸
- **Volvo Construction Equipment:** Volvo, yapay zeka destekli sıvı analizi ve parça denetim sistemleri geliştirmiştir. Bu sistemler, görsel veriyi (kamera görüntüleri) işleyerek aşınma tespiti yapmakta ve bakım planlamasını optimize etmektedir. YZ, burada sadece metin değil, multimodal (görüntü+metin) bir teşhis aracı olarak konumlanmıştır.⁴⁰
- **John Deere:** "ExpertConnect" platformu ile servis teknisyenlerine YZ destekli çözüm önerileri sunmaktadır. Sistem, geçmiş servis kayıtlarını ve teknik bültenleri tarayarak, teknisyenin karşısına en olası çözüm yollarını çıkarmaktadır. Bu da tarım makinalarında arıza süresini (downtime) minimize etmektedir.⁴²

7. Sonuç ve Stratejik Yol Haritası

İş makinası tamir kataloglarını analiz edip doğru cevaplar verebilen bir YZ sistemi kurmak, basit bir chatbot projesi değildir. Bu, kritik güvenlik ve doğruluk gereksinimleri olan bir **"Endüstriyel**

Bilişsel Sistem" tasarımıdır.

Kullanıcının sorusuna doğrudan yanıt olarak özetlemek gerekirse:

1. **İstemiye Eklenecekler:** Sadece "uzman rolü" değil; **Güvenlik Katmanı (LOTO)**, **Argo/Terim Çeviricisi**, **Zincirleme Düşünce (CoT)** mantığı ve **Yapılandırılmış JSON Çıktısı** kuralları mutlaka eklenmelidir.
2. **Alternatif Modeller:** OpenAI dışında, karmaşık akıl yürütme ve şema okuma için **Claude 3.5 Sonnet**, veri gizliliği ve yerel kurulum için **Llama 3.1 405B**, ve devasa manüelleri bölmeden okumak için **Gemini 1.5 Pro** en güçlü adaylardır.
3. **Mimari Yaklaşım:** Basit RAG yerine, kullanıcı niyetini daha iyi anlayan **HyDE** veya parça ilişkilerini kullanan **GraphRAG** mimarilerine geçiş yapılmalıdır.

Bu stratejilerin entegrasyonu, sadece "soru cevaplayan" bir bot değil, sahadaki teknisyenin yanında çalışan, güvenilir, proaktif ve hayat kurtarıcı bir "Sanal Usta" yaratılmasını sağlayacaktır.

Alıntılanan çalışmalar

1. What is RAG (Retrieval Augmented Generation)? - IBM, erişim tarihi Aralık 3, 2025, <https://www.ibm.com/think/topics/retrieval-augmented-generation>
2. From RAG to fabric: Lessons learned from building real-world RAGs at GenAIIc - Part 1 | Artificial Intelligence - AWS, erişim tarihi Aralık 3, 2025, <https://aws.amazon.com/blogs/machine-learning/from-rag-to-fabric-lessons-learned-from-building-real-world-rags-at-genaiic-part-1/>
3. Why RAG Systems Struggle with Acronyms - And How to Fix It - Shelf.io, erişim tarihi Aralık 3, 2025, <https://shelf.io/blog/why-rag-systems-struggle-with-acronyms-and-how-to-fix-it/>
4. RAG Hallucinations Explained: Causes, Risks, and Fixes - Mindee, erişim tarihi Aralık 3, 2025, <https://www.mindee.com/blog/rag-hallucinations-explained>
5. LLM01:2025 Prompt Injection - OWASP Gen AI Security Project, erişim tarihi Aralık 3, 2025, <https://genai.owasp.org/llmrisk/llm01-prompt-injection/>
6. Prompt Engineering Debugging: The 10 Most Common Issues We All Face - Reddit, erişim tarihi Aralık 3, 2025, https://www.reddit.com/r/PromptEngineering/comments/1mai2a1/prompt_engineering_debugging_the_10_most_common/
7. RAG Fight: The Silver Bullet(s) to Defeating RAG Hallucinations : r/OpenAI - Reddit, erişim tarihi Aralık 3, 2025, https://www.reddit.com/r/OpenAI/comments/1gufhcx/rag_fight_the_silver_bullets_to_defeating_rag/
8. Evaluating LLMs and Prompting Strategies for Automated Hardware Diagnosis from Textual User-Reports - arXiv, erişim tarihi Aralık 3, 2025, <https://arxiv.org/html/2507.00742v1>
9. Using Language Engineering to Build a Smarter RAG for Code - Federico

- Tomassetti, erişim tarihi Aralık 3, 2025,
<https://tomassetti.me/using-language-engineering-to-build-a-smarter-rag-for-code/>
10. Bridging the Language Gap: Our Journey to a Synonym-Aware RAG System at Palo Alto Networks, erişim tarihi Aralık 3, 2025,
<https://live.paloaltonetworks.com/t5/engineering-blogs/bridging-the-language-gap-our-journey-to-a-synonym-aware-rag/ba-p/1236616>
 11. 10 Ways EHS Pros Are Using ChatGPT (and other AI Assistants) - Benchmark Gensuite, erişim tarihi Aralık 3, 2025,
<https://benchmarkgensuite.com/ehs-blog/10-ways-ehs-pros-are-using-chat-gpt-and-other-ai-assistants/>
 12. LOTO Safety Procedures: 6 Steps for Effective Lockout Tagout - Vector Solutions, erişim tarihi Aralık 3, 2025,
<https://www.vectorsolutions.com/resources/blogs/loto-safety-6-steps-of-lockout-tagout/>
 13. Control of Hazardous Energy (Lockout/Tagout) - Overview | Occupational Safety and Health Administration - OSHA, erişim tarihi Aralık 3, 2025,
<https://www.osha.gov/control-hazardous-energy>
 14. Prompt Engineering Techniques for LLMs: A Comprehensive Guide | by Aloy Banerjee | Medium, erişim tarihi Aralık 3, 2025,
<https://medium.com/@aloy.banerjee30/prompt-engineering-techniques-for-llms-a-comprehensive-guide-46ca6466a41f>
 15. The Chain of Thought Prompting Technique: Help LLMs Solve Complex Problems, erişim tarihi Aralık 3, 2025,
<https://www.arionresearch.com/blog/lewi2w0y3obch49z3uwqb3wzixqfmr>
 16. Chain-of-Thought Prompting | Prompt Engineering Guide, erişim tarihi Aralık 3, 2025, <https://www.promptingguide.ai/techniques/cot>
 17. Chain of Verification: Prompt Engineering for Unparalleled Accuracy - Analytics Vidhya, erişim tarihi Aralık 3, 2025,
<https://www.analyticsvidhya.com/blog/2024/07/chain-of-verification/>
 18. Implement Chain-of-Verification to Improve AI Accuracy - Relevance AI, erişim tarihi Aralık 3, 2025,
<https://relevanceai.com/prompt-engineering/implement-chain-of-verification-to-improve-ai-accuracy>
 19. Chain-of-Verification (CoVe): Reduce LLM Hallucinations - Learn Prompting, erişim tarihi Aralık 3, 2025,
https://learnprompting.org/docs/advanced/self_criticism/chain_of_verification
 20. Demystifying Chains, Trees, and Graphs of Thoughts - arXiv, erişim tarihi Aralık 3, 2025, <https://arxiv.org/html/2401.14295v3>
 21. Advanced Prompt Engineering Techniques: Tree-of-Thoughts Prompting - Deepgram, erişim tarihi Aralık 3, 2025,
<https://deepgram.com/learn/tree-of-thoughts-prompting>
 22. Paper Explained: Tree of Thoughts — Deliberate Problem Solving with Large Language Models | by Isaac Kargar, erişim tarihi Aralık 3, 2025,
<https://kargarisaac.medium.com/paper-explained-tree-of-thoughts-deliberate-p>

- [problem-solving-with-large-language-models-46958bef0319](#)
23. OlympicArena Finals: Claude 3.5 Sonnet vs. GPT-4o - arXiv, erişim tarihi Aralık 3, 2025, <https://arxiv.org/html/2406.16772v1>
 24. Introducing Claude 3.5 Sonnet - Anthropic, erişim tarihi Aralık 3, 2025, <https://www.anthropic.com/news/claude-3-5-sonnet>
 25. Claude 3.5 Sonnet Complete Guide: AI Capabilities & Limits | Galileo, erişim tarihi Aralık 3, 2025, <https://galileo.ai/blog/claude-3-5-sonnet-complete-guide-ai-capabilities-analysis>
 26. Why Gemini 1.5 (and other large context models) are bullish for RAG - Medium, erişim tarihi Aralık 3, 2025, <https://medium.com/enterprise-rag/why-gemini-1-5-and-other-large-context-models-are-bullish-for-rag-ce3218930bb4>
 27. Long context | Gemini API - Google AI for Developers, erişim tarihi Aralık 3, 2025, <https://ai.google.dev/gemini-api/docs/long-context>
 28. 405B Partners | Getting the models - Llama, erişim tarihi Aralık 3, 2025, <https://www.llama.com/docs/getting-the-models/405b-partners/>
 29. Customize Generative AI Models for Enterprise Applications with Llama 3.1, erişim tarihi Aralık 3, 2025, <https://developer.nvidia.com/blog/customize-generative-ai-models-for-enterprise-applications-with-llama-3-1/>
 30. Hypothetical Document Embedding (HyDE) – A Smarter RAG method to Search Documents, erişim tarihi Aralık 3, 2025, <https://www.machinelearningplus.com/gen-ai/hypothetical-document-embedding-hyde-a-smarter-rag-method-to-search-documents/>
 31. Better RAG with HyDE - Hypothetical Document Embeddings - Zilliz Learn, erişim tarihi Aralık 3, 2025, <https://zilliz.com/learn/improve-rag-and-information-retrieval-with-hyde-hypothetical-document-embeddings>
 32. Revolutionizing Insight into Heavy Equipment Maintenance with GenAI | Databricks Blog, erişim tarihi Aralık 3, 2025, <https://www.databricks.com/blog/revolutionizing-insight-heavy-equipment-maintenance-genai>
 33. Prescriptive Agents based on RAG for Automated Maintenance (PARAM) - arXiv, erişim tarihi Aralık 3, 2025, <https://arxiv.org/html/2508.04714v2>
 34. A Technical Support Engineer's Guide to RAG: Eliminating LLM Hallucinations in Customer Support - AptEdge, erişim tarihi Aralık 3, 2025, <https://aptedge.io/blog/a-support-engineers-guide-to-rag-eliminating-llm-hallucinations-in-customer-support>
 35. The Ultimate Prompt Engineering Framework: Building a Structured AI Team with the SPARC System : r/PromptEngineering - Reddit, erişim tarihi Aralık 3, 2025, https://www.reddit.com/r/PromptEngineering/comments/1kbufy0/the_ultimate_prompt_engineering_framework/
 36. JSON prompting for LLMs - IBM Developer, erişim tarihi Aralık 3, 2025, <https://developer.ibm.com/articles/json-prompting-llms/>
 37. Guided JSON with LLMs: From Raw PDFs to Structured Intelligence | by Doil Kim -

- Medium, erişim tarihi Aralık 3, 2025,
<https://medium.com/@kimdoil1211/structured-output-with-guided-json-a-practical-guide-for-llm-developers-6577b2eee98a>
38. SoftServe's Gen AI Industrial Assistant for Real-Time Guidance, erişim tarihi Aralık 3, 2025,
<https://www.softserveinc.com/en-us/our-partners/nvidia/gen-ai-industrial-assistant>
39. AI Made Easy: SoftServe Gen AI Solutions Built With NVIDIA Blueprints, erişim tarihi Aralık 3, 2025,
<https://www.softserveinc.com/en-us/blog/softserve-gen-ai-solutions-by-nvidia-nim>
40. Connected solutions | Volvo Construction Equipment Global, erişim tarihi Aralık 3, 2025,
<https://www.volvoce.com/global/en/about-us/how-we-drive-change/connected-solutions/>
41. AI transporters push the boundaries of modern manufacturing - Volvo Group, erişim tarihi Aralık 3, 2025,
<https://www.volvogroup.com/en/news-and-media/news/2024/nov/ai-modern-manufacturing.html>
42. How to Use the "Suggest a Resolution" Feature - ExpertConnect Knowledge Base, erişim tarihi Aralık 3, 2025,
https://help.expertconnect.deere.com/en_US/integrations/how-to-use-the-suggest-a-resolution-feature
43. Artificial Intelligence at John Deere, erişim tarihi Aralık 3, 2025,
<https://emerj.com/artificial-intelligence-at-john-deere/>