AIR LAB

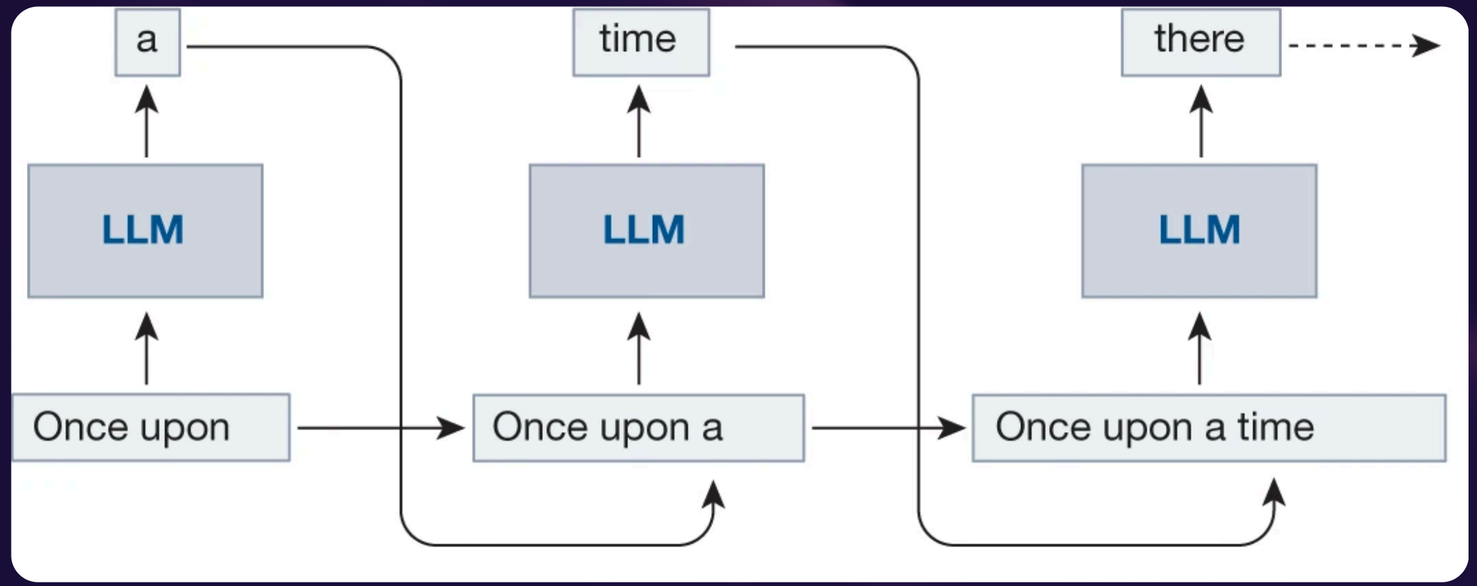
ADVANCED RAG AND KNOWLEDGE GRAPHS

Utku Başçakır
Bilgisayar Mühendisliği 3.Sınıf

AIR LAB

LLM Nedir?

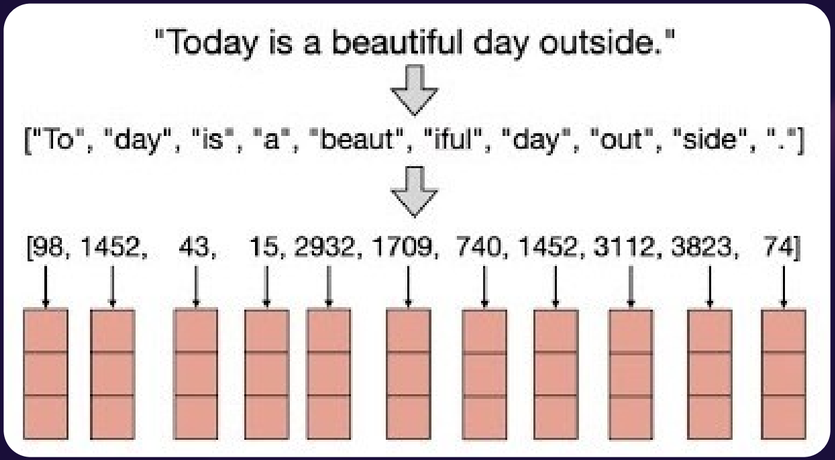
- **Derin Öğrenme Temelli Mimari:** Milyarlarca parametreye sahip, genellikle Transformer tabanlı devasa sinir ağlarıdır.
- **Sıradaki Kelimeyi Tahmin Etme:** Temel görevleri, kendisine verilen bir metin girişinden (prompt) yola çıkarak, istatistiksel olarak en olası bir sonraki "token"ı (kelime veya karakter grubu) tahmin etmektir.
- **Geniş Kapsamlı Ön Eğitim (Pre-training):** İnternet, kitaplar, makaleler ve kod depolarından oluşan devasa veri setleriyle eğitilerek; dilin yapısını, mantık yürütmeyi ve genel dünya bilgisini "ağırlıklarında" (weights) saklarlar.
- **Generative (Üretken) Doğa:** Sadece veriyi sınıflandırmaz, yeni ve anlamlı içerik üretirler.

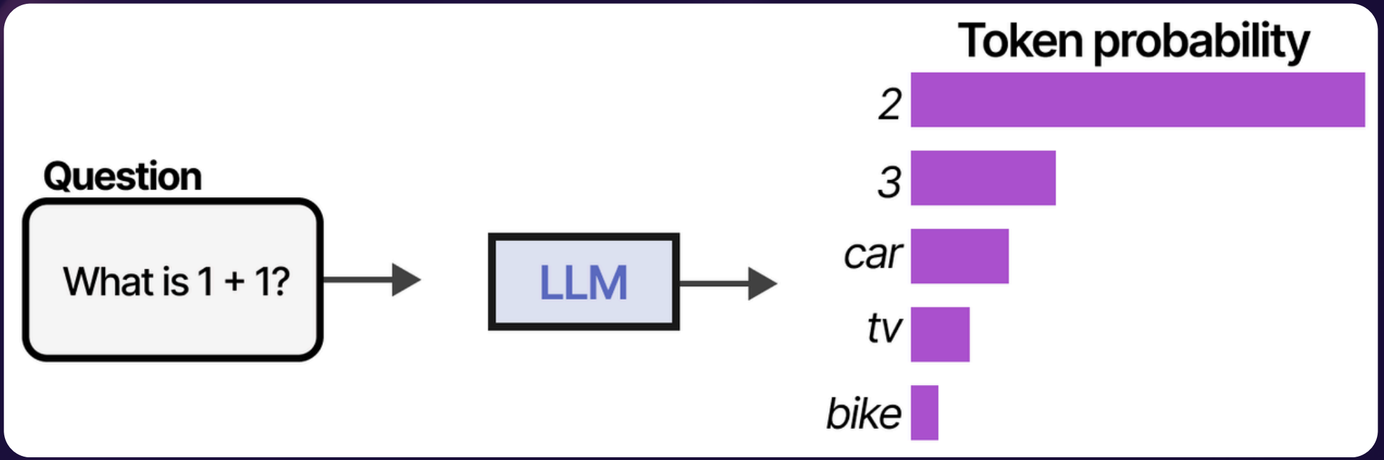


AIR LAB

LLM’ler Nasıl Çalışır?

- **Tokenization:** Metin verisi doğrudan modele girmez; önce "token" adı verilen anlamsal birimlere (kelime parçaları, karakterler) ayrılır ve sayısal vektörlere dönüştürülür.
- **Bağlamsal İlişkiler:** Model, girdi dizisindeki her token'ın birbiriyle olan ilişkisini ve ağırlığını Self-Attention mekanizması ile hesaplar.
- **Softmax ve Olasılık Dağılımı:** Modelin son katmanı, bir sonraki gelebilecek tüm olası token'lar için bir olasılık puanı üretir.
- **Seçim Stratejisi:** En yüksek olasılıklı token (Greedy Search) veya belirli bir çeşitlilik barındıran (Temperature/Top-p) seçim yapılarak çıktı oluşturulur.

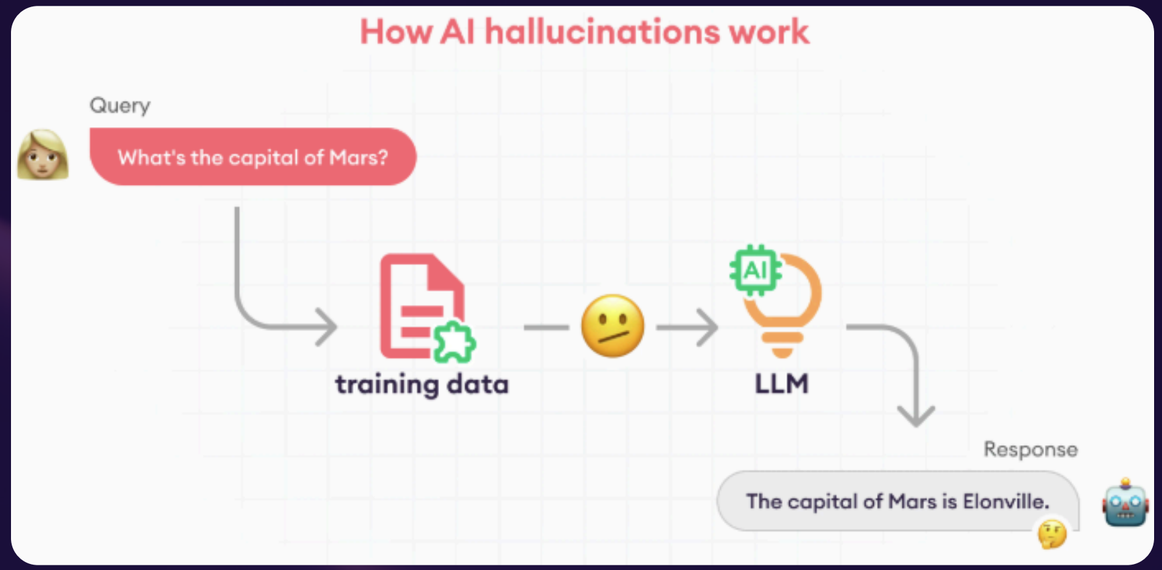




AIR LAB

LLM’lerin Temel Sınırları ve Kısıtlamaları

- **Bilgi Kesintisi:** Modeller, eğitim süreçleri tamamlandığı andaki veri setiyle sabitlenirler. Bu tarihten sonra gerçekleşen güncel olaylar, yeni teknolojiler veya değişen veriler hakkında bilgi sahibi değildirler.
- **Halüsinasyon:** Olasılıksal yapıları gereği, model bilmediği bir konuda en yüksek olasılıklı kelimeleri yan yana getirerek kulağa mantıklı gelen ancak tamamen yanlış bilgiler üretebilir.
- **Özel ve Kurumsal Veri Eksikliği:** Modeller halka açık verilerle eğitilir. Şirket içi dökümanlar, gizli projeler veya kişiye özel veriler modelin parametreleri içerisinde yer almaz.
- **Doğrulanabilirlik Sorunu:** Standart bir LLM çıktısında, bilginin kaynağını (referansını) gösterme yeteneğine sahip değildir; bilgiyi "ezberinden" söyler.
- **Context Window Sınırı:** Bir LLM'in tek seferde okuyabileceği metin miktarı sınırlıdır (Örn: 128k token). Ama şirket verileri milyarlarca token’dan oluşur.



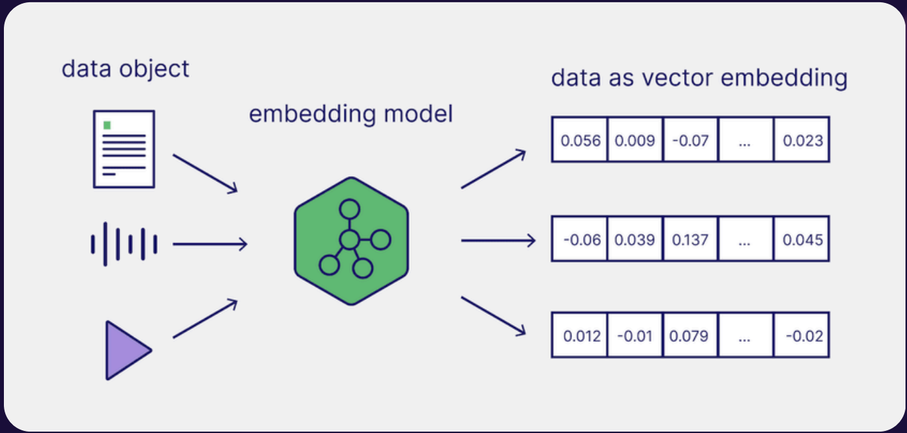
AIR LAB

Çözüm: RAG

Retrieval-Augmented Generation (RAG); dil modellerinin, kullanıcı sorgusuna anlamsal olarak en yakın ve güncel dış verileri kendi bilgi havuzuna dahil ederek, çok daha doğru, doğrulanabilir ve bağlamsal yanıtlar üretmesini sağlayan bir sistem mimarisidir.

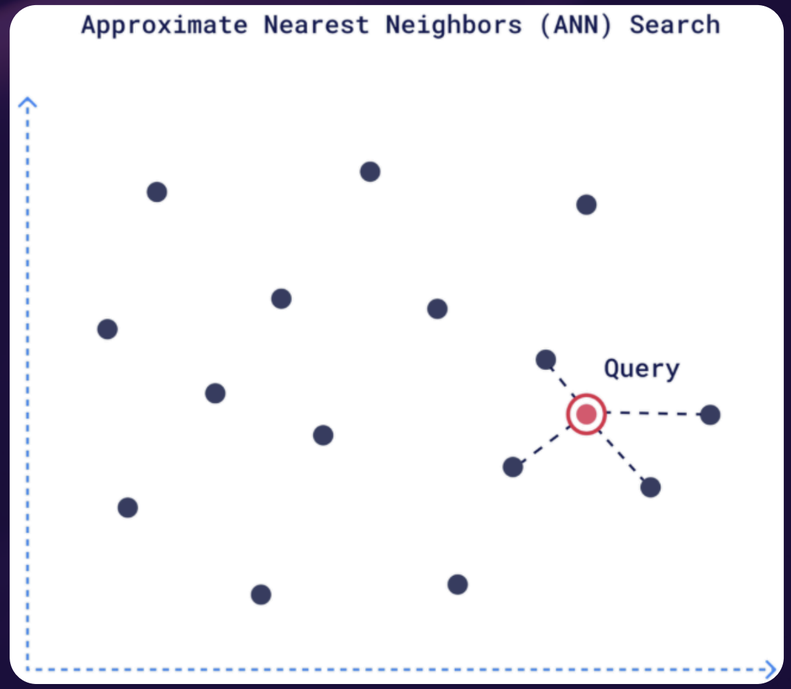
1) Verinin Sayısallaştırılması (Embedding)

Farklı türdeki verilerin (metin, görsel, ses), bu türlere özel eğitilmiş modeller aracılığıyla, anlamsal özünü koruyacak şekilde yüksek boyutlu sayısal vektörlere dönüştürülmesidir.



2) Vektör Veritabanları

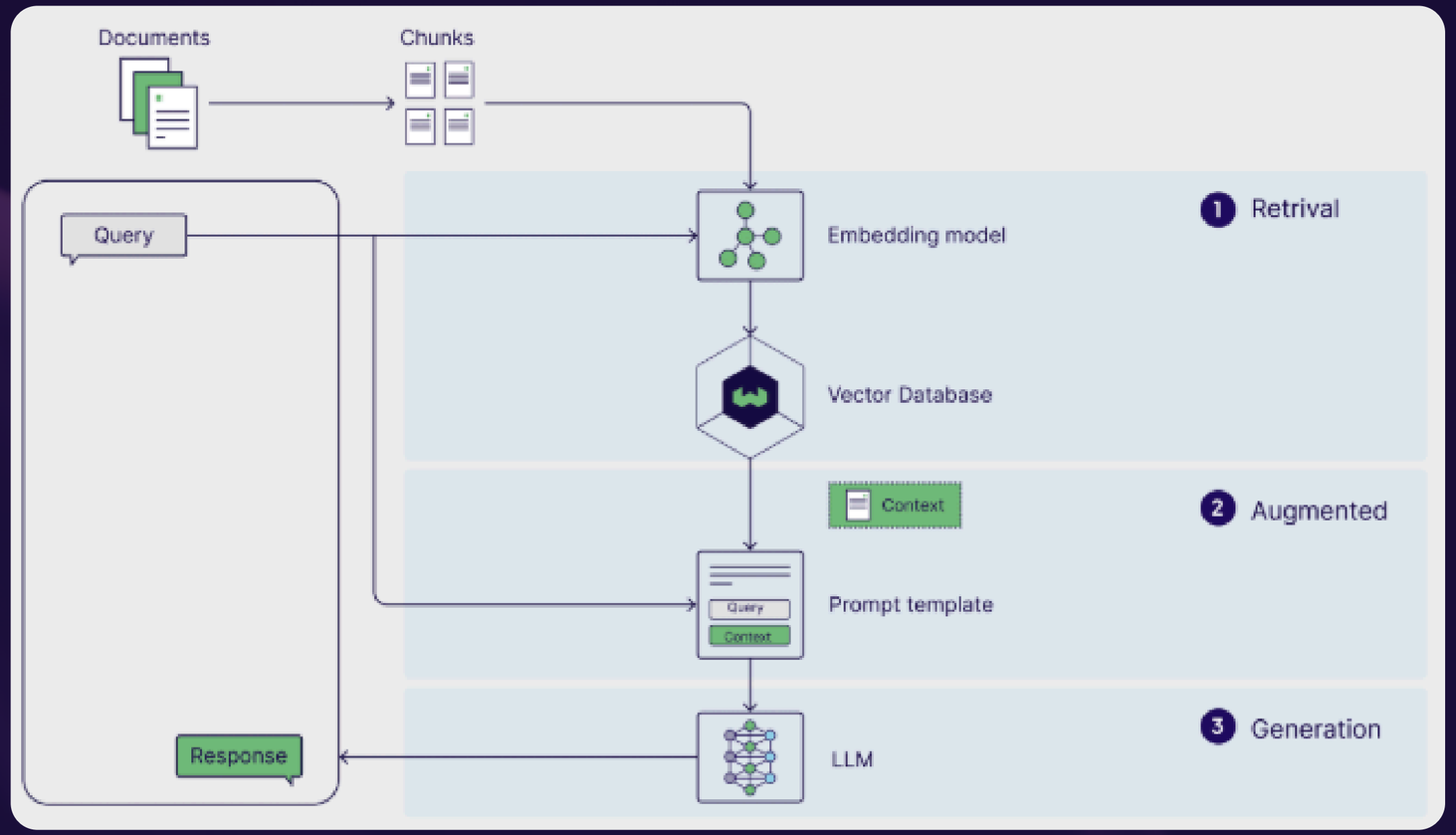
Embedding modelleri tarafından anlamsal özelliklerine göre sayısallaştırılmış milyonlarca veriyi, matematiksel vektörler olarak depolayan ve bu vektörler arasındaki mesafe ölçümleriyle hızlı arama yapılmasını sağlayan yüksek performanslı sistemlerdir.



AIR LAB

RAG Mimarisi

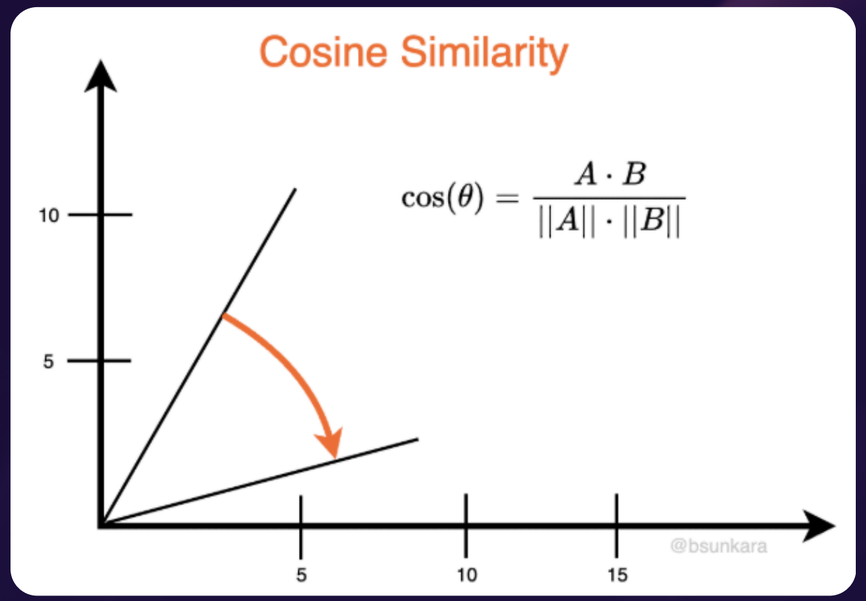
- **Retrieval (Geri Getirme):** Kullanıcı sorgusunun (Query) embedding modeliyle vektöre dönüştürülüp, vektör veritabanındaki milyarlarca parça arasından en alakalı bağlamların (Context) tespit edilmesidir.
- **Augmentation (Zenginleştirme):** Elde edilen bu spesifik bağlam verilerinin, orijinal kullanıcı sorusuyla birlikte bir "Prompt Template" (İstem Şablonu) içerisinde birleştirilerek LLM'in işleyebileceği hale getirilmesidir.
- **Generation (Üretim):** LLM'in, kendisine sunulan bu ek bağlamı kullanarak, halüsinasyon riskini minimize eden ve dış veriye dayalı, tutarlı bir yanıt üretmesi aşamasıdır.

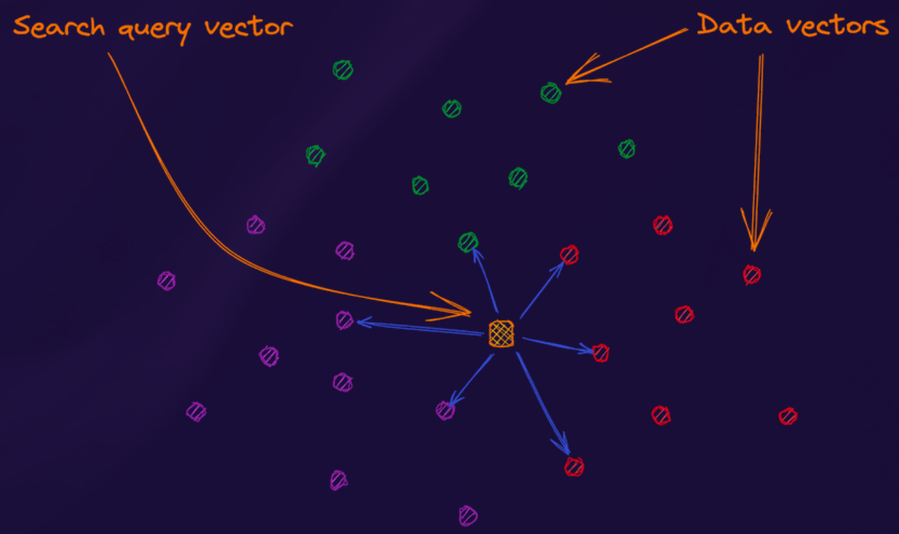


AIR LAB

Aşama 1 - Retrieval

- **Sorgu Vektörleştirme:** Kullanıcının girdiği "Query", sistemdeki dökümanlarla aynı Embedding modeli kullanılarak sayısal bir vektöre dönüştürülür.
- **Anlamsal Arama (Semantic Search):** Vektör veritabanı, sorgu vektörü ile döküman vektörleri arasındaki matematiksel mesafeyi hesaplar.
- **Kosinüs Benzerliği (Cosine Similarity):** Vektörler arasındaki açının kosinüsüne bakılır. Açı ne kadar küçükse (kosinüs değeri 1'e ne kadar yakınsa), belgeler o kadar "anlamdaş" kabul edilir.
- **Top-K Retrieval:** Hesaplama sonucunda benzerlik puanı en yüksek olan ilk "k" adet parça (chunk), LLM'e gönderilmek üzere seçilir.





AIR LAB

Aşama 2 - Augmentation

Retrieval aşamasından gelen en alakalı Top-K metin parçaları, önceden yapılandırılmış bir **Prompt Template** içerisinde kullanıcının sorusu ve modele rol atayan **System Prompt** ile birleştirilir. Bu süreçte modele eklenen “sadece sağlanan dokümanlara sadık kal” ve “bilgin yoksa belirt” gibi güvenlik çerçeveleri, LLM'in halüsinasyon yapmasını engelleyerek yanıtın doğruluğunu ve güvenilirliğini garanti altına alır.

[SYSTEM PROMPT]

Sen, aşağıda paylaşılan dokümanları temel alarak soruları yanıtlayan profesyonel bir yapay zeka asistanısın.

Talimatlar:

1. Soruyu yanıtlamak için sadece sana sunulan **[CONTEXT]** bölümündeki bilgileri kullan.
2. Eğer cevap bu dokümanlarda yer almıyorsa, dışarıdan bilgi ekleme ve sadece "Bu sorunun cevabı sağlanan belgelerde bulunmamaktadır." şeklinde yanıt ver.
3. Yanıt verirken kullandığın her bilgi için hangi belgeye dayandığını belirt (Örn: [1], [2]).

[CONTEXT]

[1] {Belge_Başlığı_1}: {Belge_İçeriği_1}
[2] {Belge_Başlığı_2}: {Belge_İçeriği_2}
[...]
[N] {Belge_Başlığı_N}: {Belge_İçeriği_N}

[USER QUERY]

"{Kullanıcı_Sorusu_Buraya_Gelecek}"

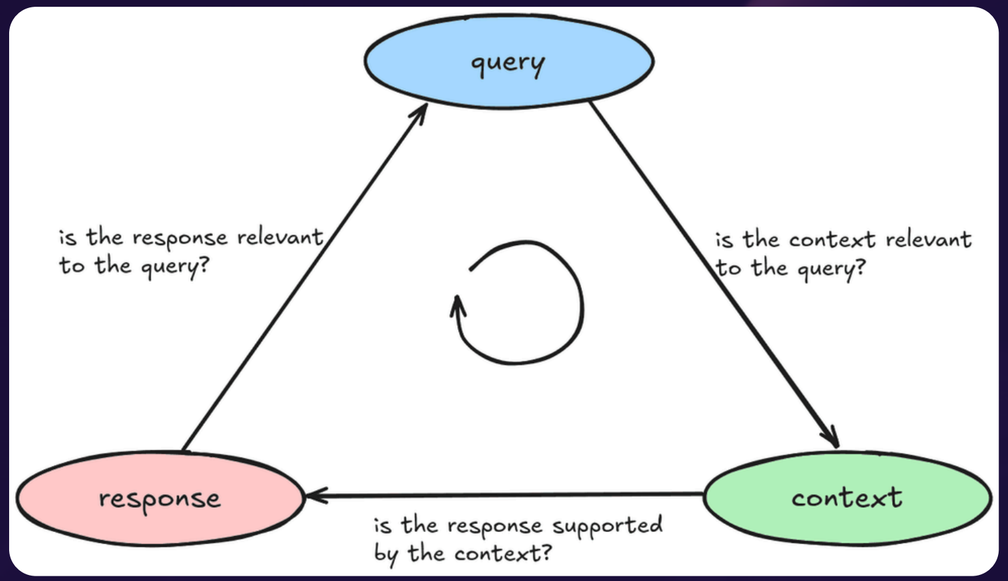
AIR LAB

Aşama 3 – Generation ve Evaluation

- **Groundedness (Dayanaklılık):** LLM'in kendi eğitim verisi yerine, sadece sunulan bağlama (context) odaklanarak yanıt üretmesi sağlanır.
- **Citations (Referanslandırma):** Üretilen yanıtın hangi doküman parçasına dayandığı belirtilerek şeffaflık ve doğrulanabilirlik sağlanır.
- **RAG Triad (Performans Metrikleri):** Sistemin başarısı üç ana eksende ölçülür:

1. Faithfulness (Sadakat): Cevap dokümana ne kadar sadık? (Halüsinasyon kontrolü).
2. Answer Relevance (Alakalılık): Cevap soruya ne kadar hitap ediyor?
3. Context Precision (Bağlam Hassasiyeti): Getirilen belgeler soruyu çözmek için ne kadar yeterli?

- **RAGAS & LLM-as-a-Judge:** Değerlendirme sürecinde genellikle başka bir güçlü LLM (Gemini 3.5 vb.) hakem olarak kullanılarak sistem otomatik olarak puanlanır.

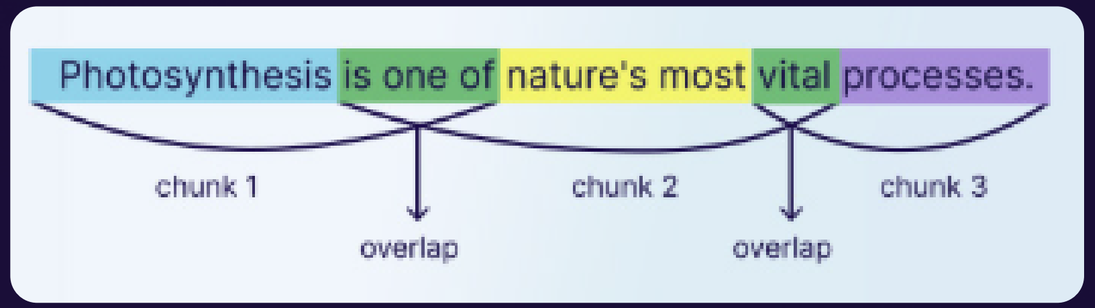


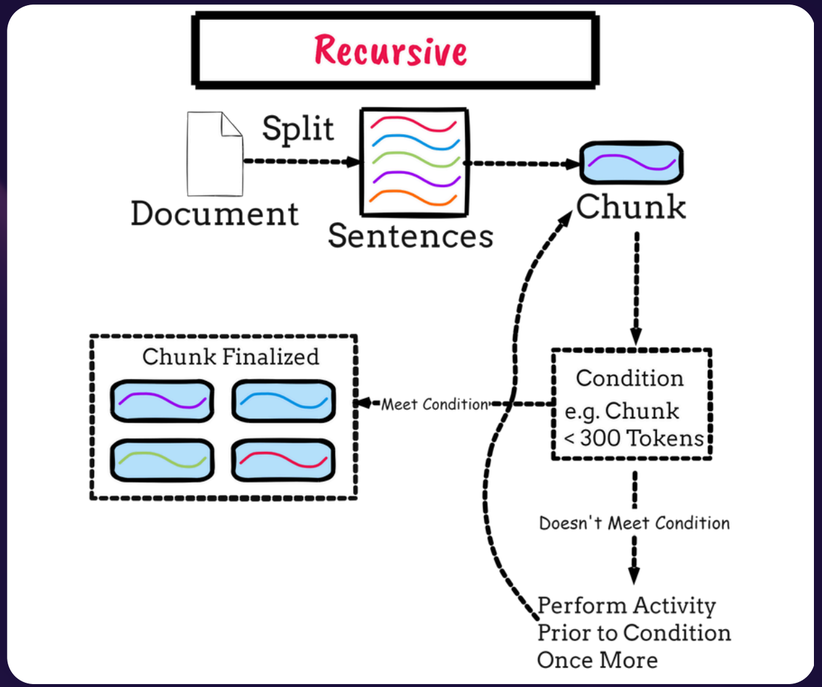
ragas

AIR LAB

Aşama 0 - Chunking

- **Chunking:** Devasa dokümanların, LLM'in bağlam penceresine sığması ve anlamsal odağını kaybetmemesi için daha küçük, yönetilebilir parçalara bölünmesidir.
- **Fixed-size Chunking:** Metni hiçbir kural gözetmeden sadece belirli bir karakter veya token sayısına göre (Örn: Her 500 karakterde bir) kesmektir. Hızlıdır ancak cümlelerin ortadan bölünmesine neden olabilir.
- **Recursive Character Chunking:** Metni hiyerarşik bir sırayla (önce paragraflar, sonra cümleler, en son kelimeler) bölme yöntemidir. Anlam bütünlüğünü korumak için en çok tercih edilen standart yöntemdir.
- **Overlap:** Bir parçanın sonundaki metnin, bir sonraki parçanın başında da yer almasıdır (Genelde %10-20 oranında). Bu sayede bağlamın "keskin bir bıçakla kesilmesi" engellenir ve parçalar arasında anlamsal bir köprü kurulur.





AIR LAB

Standart RAG’ın Sonuçları

Neleri Başardık?

- **Bilgi Bariyerini Aştık:** LLM'in eğitim verisinde bulunmayan güncel veya kurumsal verilere erişmesini sağladık.
- **Halüsinasyonu Azalttık:** Modeli "açık kitap" sınavına sokarak, cevaplarını rastgele uydurmak yerine sunduğumuz kanıtlara dayandırmasını (Groundedness) sağladık.
- **Maliyet ve Verimlilik:** Tüm dokümantasyonu modele vermek yerine sadece en alakalı "chunk"ları seçerek context window limitlerini ve işlem maliyetlerini optimize ettik.

Nerelerde Tıkandık?

- **Retrieval Hataları (Yanlış Bilgi):** Vektör veritabanı bazen soruyu anlamsal olarak yakalasa da, en doğru cevabı içeren parçayı ilk sıralarda (Top-K) getiremeyebilir.
- **Bağlam Kaybı:** Küçük parçalara böldüğümüz (chunking) veriler, büyük resimdeki ilişkileri kaybedebilir. "Bu dokümanın genel özeti nedir?" gibi sorular cevapsız kalabilir.
- **Soru Karmaşıklığı:** Kullanıcının çok basit veya çok dolaylı sorduğu sorularda (Query ambiguity), standart embedding modelleri doğru eşleşmeyi bulmakta zorlanabilir.
- **İlişki Körlüğü:** Sadece "benzerlik" üzerinden arama yapıldığı için, veriler arasındaki hiyerarşik veya mantıksal bağlar (Örn: A kişisinin B şirketindeki rolü) gözden kaçabilir.

AIR LAB

Gelişmiş Chunking Yöntemleri

Semantic Chunking

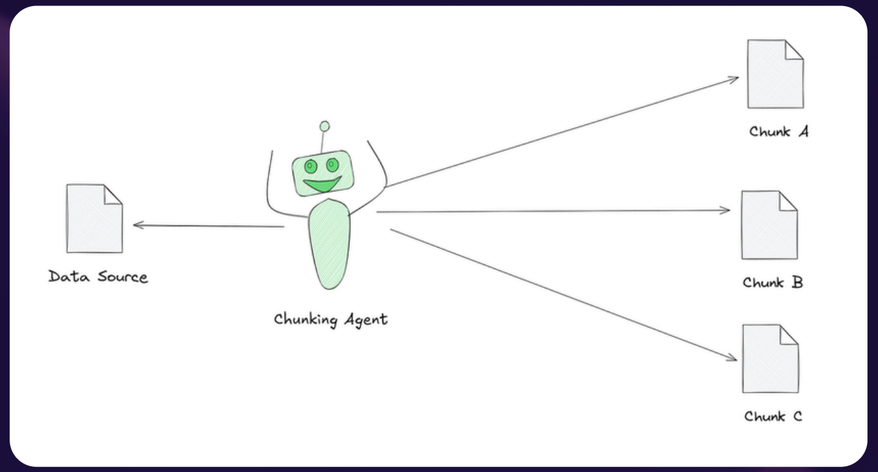
- Metni anlamın koptuğu yerlerden böler.
- Ardışık cümleler arasındaki Cosine Similarity değerleri hesaplanır, belli bir eşiğin altında kalan değerlerde bölme yapılır.

```
Sentences:
1. "Tesla reported record revenue of $25.2B in Q3 2024."
2. "The company exceeded analyst expectations by 15%."
3. "Revenue growth was driven by strong vehicle deliveries."
4. "The Model Y became the best-selling vehicle globally."
5. "Customer satisfaction ratings reached an all-time high."
```

```
Similarity between sentences 1-2: 0.85
Similarity between sentences 2-3: 0.78
Similarity between sentences 3-4: 0.42
Similarity between sentences 4-5: 0.71
```

Agentic Chunking

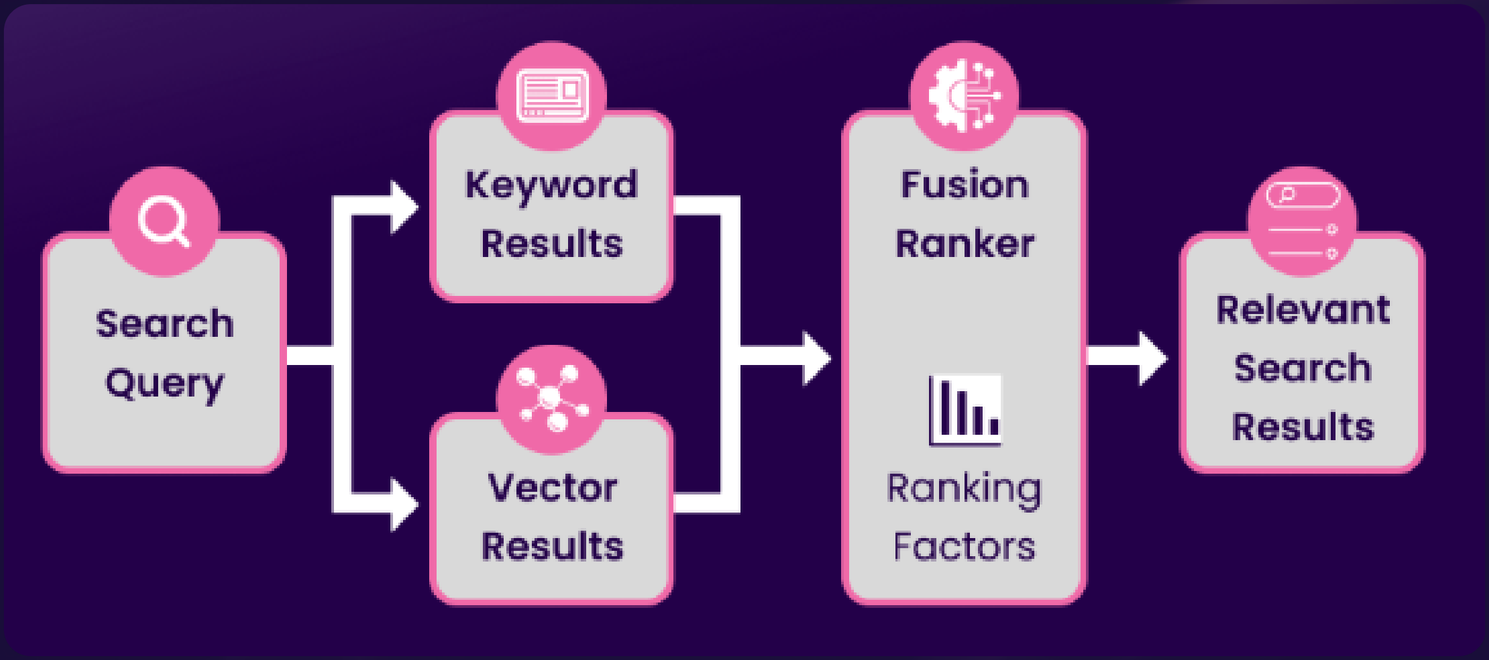
- Metni bölme işinin LLM'lere bırakılmasıdır.
- Çok karmaşık dokümanlarda bile en yüksek anlam bütünlüğünü sağlar.



AIR LAB

Hybrid Search

- **İki Dünyanın Birleşimi:** Anlamsal benzerlik (Vektör) ile geleneksel anahtar kelime eşleşmesinin (BM25) eş zamanlı kullanılmasıdır.
- **Vektör:** Kelimelerin manasını ve bağlamını yakalar; dolaylı soruları yanıtlar.
- **Anahtar Kelime:** Teknik terimler, ürün kodları ve nadir kelimeler için kesin doğruluk sağlar.
- **RRF (Reciprocal Rank Fusion):** Farklı arama motorlarından gelen sonuçları akıllıca harmanlayarak en doğru dokümanı en üste taşır.



$$score = \frac{1}{k + sıra\_numarası}$$

Vektörden gelen puan: $1/(60+1) = 0.0163$

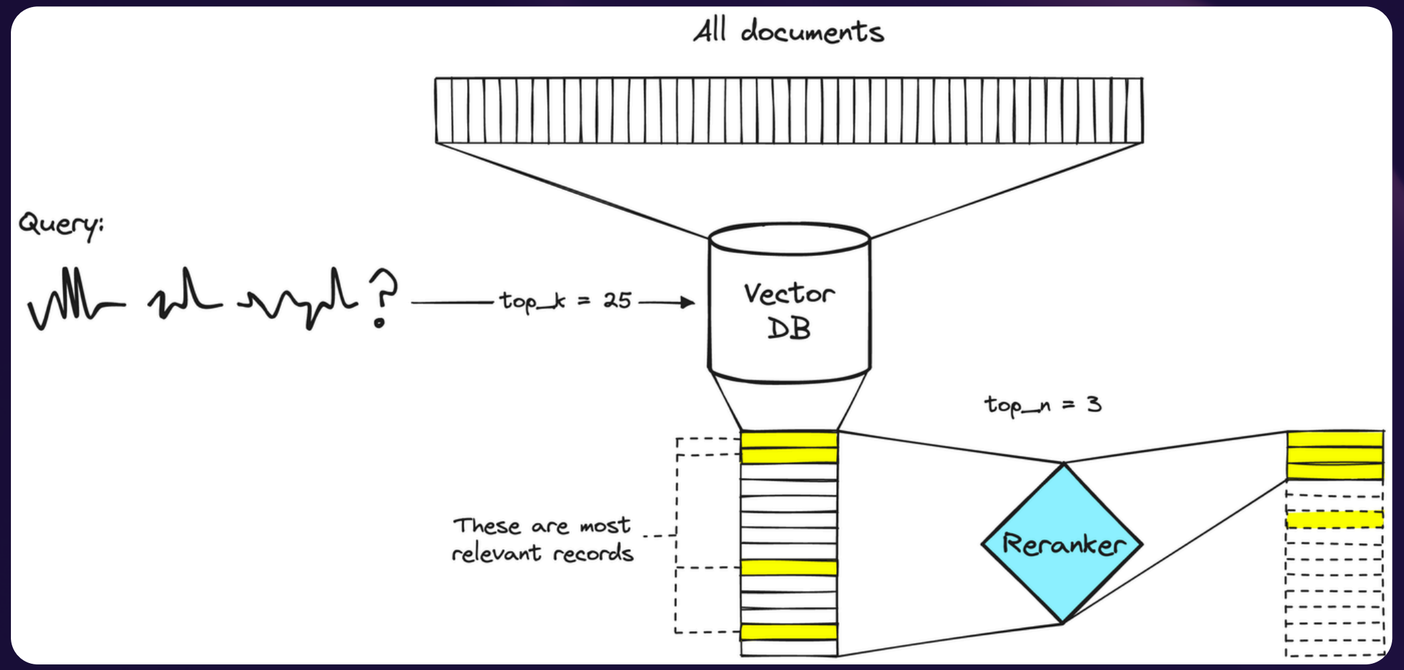
BM25'ten gelen puan: $1/(60+10) = 0.0142$

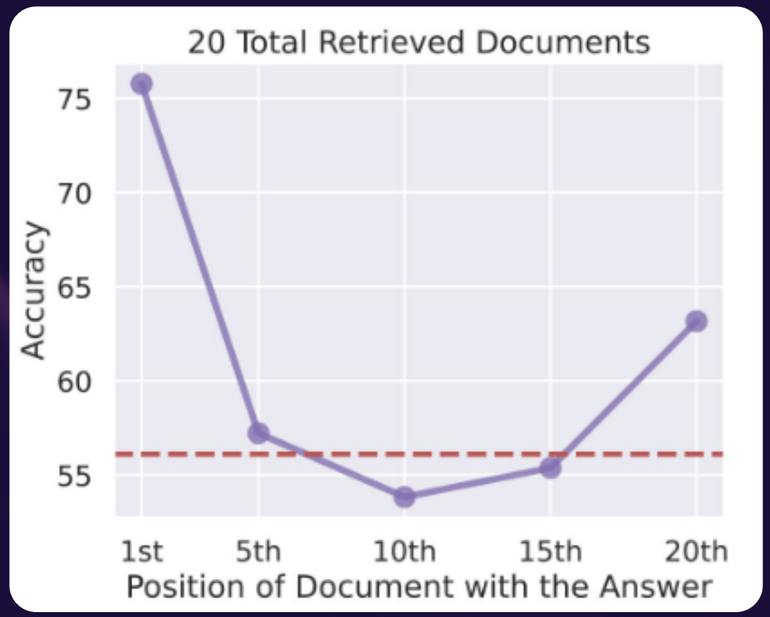
Final RRF Puanı: $0.0163 + 0.0142 = \mathbf{0.0305}$

AIR LAB

Reranking

- **Zeka vs. Hız Dengesi:** İlk arama (Hybrid) milyonlarca veri içinde hızlı bir "kaba eleme" yapar. Reranker ise bu adaylar arasından en iyiyi seçen "uzman süzgeçtir".
- **Cross-Encoder Teknolojisi:** Sorgu ve dökümanı ayrı ayrı değil, birlikte analiz ederek aralarındaki derin anlamsal ilişkiyi ölçer.
- **Nihai Sıralama:** İlk aşamada 10. sırada kalan ancak en doğru cevabı içeren parçayı fark eder ve onu 1. sıraya yükseltir.
- **Maksimum Verimlilik:** LLM'in kısıtlı bağlam penceresine sadece "en rafine" bilginin girmesini sağlar.





“Lost In The Middle” Phenomenon

AIR LAB

Knowledge Graphs ve GraphRAG

Vector RAG'ın Sınırları

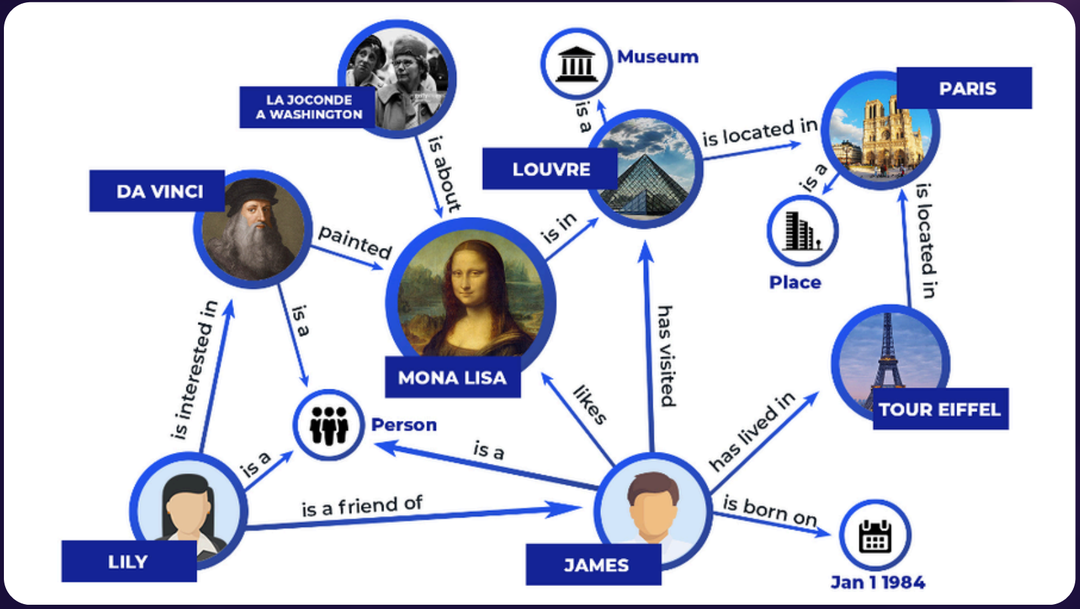
- Metni bulur ama doküman içi mantığı kaçırır.
- Multi-hop sorularda zorlanır. (örn: "A kişisinin B projesindeki rolü?")

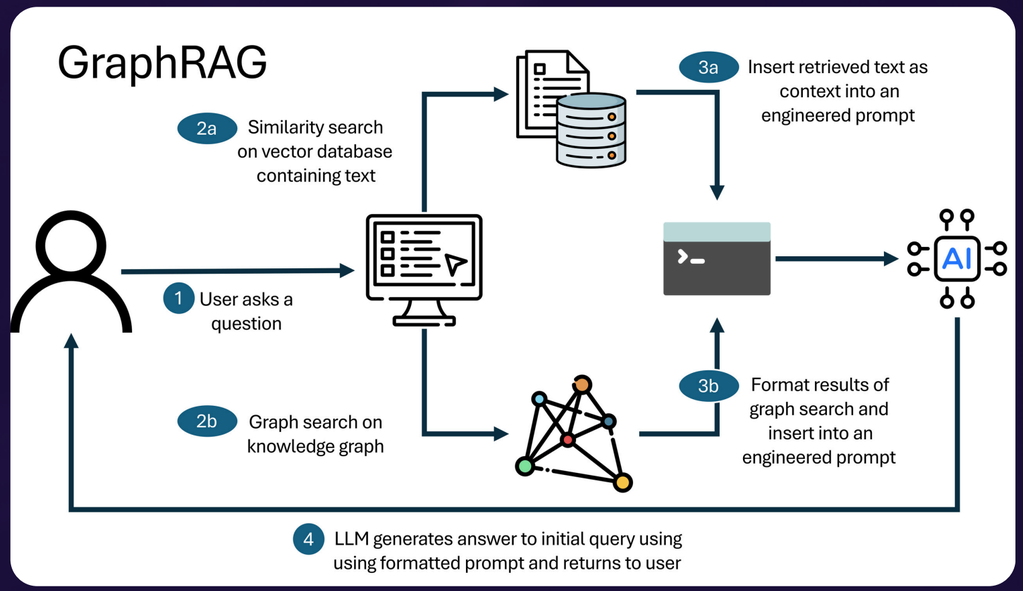
Knowledge Graphs

- Bilgi metin olarak değil → Varlıklar ve İlişkiler ağı olarak tutulur
- Nodes = kavramlar
- Edges = aralarındaki bağ

GraphRAG

- Vektör arama + ilişki grafiği
- Benzer bilgiyi değil → bağlantılı bilgiyi getirir.





AIR LAB

THANK YOU

www.airlab.yildizskylab.com