T.C.
AKDENİZ ÜNİVERSİTESİ

RUHSAL BOZUKLUKLARIN TANISINDA ERİŞİM DESTEKLİ ÜRETİM TABANLI BÜYÜK DİL MODELLERİ KULLANILARAK BİR KLİNİK KARAR DESTEK SİSTEMİ GELİŞTİRİLMESİ

Egehan ÇETİN

FEN BİLİMLERİ ENSTİTÜSÜ

ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ

ANABİLİM DALI

DOKTORA TEZİ

HAZİRAN 2026

ANTALYA

T.C.
AKDENİZ ÜNİVERSİTESİ

RUHSAL BOZUKLUKLARIN TANISINDA ERİŞİM DESTEKLİ ÜRETİM TABANLI BÜYÜK DİL MODELLERİ KULLANILARAK BİR KLİNİK KARAR DESTEK SİSTEMİ GELİŞTİRİLMESİ

Egehan ÇETİN

FEN BİLİMLERİ ENSTİTÜSÜ

ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ

ANABİLİM DALI

DOKTORA TEZİ

HAZİRAN 2026

ANTALYA

T.C.

AKDENİZ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

RUHSAL BOZUKLUKLARIN TANISINDA ERİŞİM DESTEKLİ ÜRETİM TABANLI BÜYÜK DİL MODELLERİ KULLANILARAK BİR KLİNİK KARAR DESTEK SİSTEMİ GELİŞTİRİLMESİ

Egehan ÇETİN

ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ

ANABİLİM DALI

DOKTORA TEZİ

HAZİRAN 2026

Evrak Tarih ve Sayısı: 30.06.2026-952

T.C.

AKDENİZ ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

RUHSAL BOZUKLUKLARIN TANISINDA ERİŞİM DESTEKLİ ÜRETİM TABANLI BÜYÜK DİL MODELLERİ KULLANILARAK BİR KLİNİK KARAR DESTEK SİSTEMİ GELİŞTİRİLMESİ

Egehan ÇETİN

ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ

ANABİLİM DALI

DOKTORA TEZİ

Bu tez 25/06/2026 tarihinde jüri tarafından Oybirliği ile kabul edilmiştir.

Prof. Dr. Süleyman BİLGİN

Prof. Dr. Övünç POLAT

Prof. Dr. Alper BİLGE

Prof. Dr. Yalçın İŞLER

Dr. Öğr. Üyesi Ufuk ÖZKAYA

Bu belge, güvenli elektronik İmza ile imzalanmıştır.
Evrak sorgulaması https://turkiye.gov.tr/ebd?eK=5543&eD=BSFVF7JKB4&eS=952 adresinden yapılabilir. (PIN:07062)

ÖZET

RUHSAL BOZUKLUKLARIN TANISINDA ERİŞİM DESTEKLİ ÜRETİM TABANLI BÜYÜK DİL MODELLERİ KULLANILARAK BİR KLİNİK KARAR DESTEK SİSTEMİ GELİŞTİRİLMESİ

Egehan ÇETİN

Doktora Tezi, Elektrik-Elektronik Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Süleyman BİLGİN

Haziran 2026; 59 sayfa

Ruhsal bozukluklar, bireyin yaşam kalitesini derinden etkileyen ve toplum üzerinde ağır bir sosyal, tıbbi ve ekonomik yük bırakan yaygın bir sağlık sorunudur. Yapay zekadaki gelişmeler ve sağlık alanında biriken veri, bu bozuklukların tanı ve yönetiminde hekime destek olabilecek araçların geliştirilmesini bugün mümkün kılmaktadır. Ancak ruhsal bozukluğun tanısı, doğası gereği belirsizliği yüksek bir akıl yürütme sürecidir. Serbest metinle yazılmış bir klinik öyküyü, standart bir sınıflandırmanın tanımlı kategorilerine eşlemeyi gerektirir. Büyük dil modelleri (BDM) bu görevde uzman düzeyine yaklaşan bir başarı gösterse de tek başlarına güvenilir değildir. Halüsinasyon, eğitim verisinden gelen yanlılık ve tanısal eşikleri uygulayamamaktan kaynaklanan aşırı tanı eğilimi gibi faktörler etkendir.

Yapılan çalışmada, bu sınırlamaları Erişim Destekli Üretim (Retrieval-Augmented Generation, RAG) yaklaşımıyla ele almaktadır. Bilgi tabanı olarak, Dünya Sağlık Örgütü'nün (WHO) açık kaynaklı ICD-11 sınıflandırması ile ICD-11-CDDR el kitabından doğal dil işleme yöntemleriyle derlenip temizlenen 720 ruhsal bozukluk girdisi kullanılmıştır. Test kümesi ise DSM-5-TR Clinical Cases kitabından alınan 104 klinik vakadan oluşmaktadır. Tasarımın hangi bileşeninin tanısal başarıya ne ölçüde katkı verdiğini ayrıştırmak amacıyla tez kapsamında dokuz farklı konfigürasyon (K1–K9) kurulmuş ve her vakanın birincil tanısı hem tanı adı hem de kod (tanı grubu, kategori ve alt kategori) düzeylerinde değerlendirilmiştir.

En yüksek başarımı veren K4 konfigürasyonu dört erişim bileşenini bir araya getirmektedir. Klinik öyküden varsayımsal belge gömmeleri (Hypothetical Document Embeddings, HyDE) ile üretilen klinik özet, BM25 ile sözcüksel eşleşme, 1024 boyutlu yoğun gömmeler arasındaki kosinüs benzerliği ve bu sinyalleri birleştiren karşılıklı sıralama birleştirmesidir (Reciprocal Rank Fusion, RRF). Önerilen sistemde getirilen adaylar, klinik tanı kriterleriyle birlikte Gemma 4 12B modeline sunulmuş ve model kriter temelli bir tanı seçimi yapmıştır. Önerilen sistem 0.615 (N=104) tanı adı düzeyi doğruluğa ulaşarak, literatürdeki en yüksek açık kaynak skoru olan 0.524 (LLMind) değerini hem nokta tahmininde geçmiş hem de sabit referansa karşı tek yönlü binom testinde anlamlılığa ulaşan sistem olmuştur ($p=0.038$). Yaklaşık üçte bir etkin parametreyle çalışan karışık uzman mimarili Gemma 4 26B-A4B modeli 0.606 ile hesaplama açısından daha verimli bir alternatif sağlamıştır. Dokuz konfigürasyonlu kontrollü ablasyon, HyDE'yi erişimdeki en belirleyici bileşen olarak öne çıkarmış ve RAG'ın taban modele katkısı eşleştirilmiş McNemar testinde anlamlı bulunmuştur

($p<0.001$). RAG'ın tanısal katkısı kullanılan ölçüte bağlı olduğu gözlemlenmiştir. Bu katkı, kod düzeyinde (tanı grubu, kategori ve alt kategori) tutarlı biçimde büyük, tanı adı düzeyinde ise modelin parametrik bilgisiyle değişkendir. Bu durum, yalnızca tanı adı düzeyinde raporlayan LLMind'ın RAG'ın kod düzeyindeki katkısını gizlediğini ve çok granülerli değerlendirmenin gerekliliğini göstermektedir.

ANAHTAR KELİMELER: Büyük Dil Modelleri, Doğal Dil İşleme, Erişim Destekli Üretim, Klinik Karar Destek Sistemleri, Ruhsal Bozukluk Tanısı

JÜRİ: Prof. Dr. Süleyman BİLGİN

Prof. Dr. Övünç POLAT

Prof. Dr. Alper BİLGE

Prof. Dr. Yalçın İŞLER

Dr. Öğr. Üyesi Ufuk ÖZKAYA

ABSTRACT

DEVELOPMENT OF A CLINICAL DECISION SUPPORT SYSTEM USING RETRIEVAL AUGMENTED GENERATION BASED LARGE LANGUAGE MODELS IN THE DIAGNOSIS OF MENTAL DISORDERS

Egehan ÇETİN

PhD Thesis in Electrical-Electronic Engineering

Supervisor: Prof. Dr. Süleyman BİLGİN

June 2026; 59 pages

Mental disorders are a widespread health problem that profoundly affects an individual's quality of life and places a heavy social, medical, and economic burden on society. Advances in artificial intelligence and the growing volume of data in healthcare now make it possible to develop tools that can support physicians in the diagnosis and management of these disorders. However, the diagnosis of a mental disorder is inherently a reasoning process with a high degree of uncertainty. It requires mapping a free-text clinical history onto the defined categories of a standard classification. Although large language models (LLMs) approach expert-level performance in this task, they are not reliable on their own. Contributing factors include hallucination, bias originating from training data, and a tendency toward overdiagnosis stemming from the inability to apply diagnostic thresholds.

This study addresses these limitations through a Retrieval-Augmented Generation (RAG) approach. As the knowledge base, 720 mental disorder entries were used, compiled and cleaned by natural language processing methods from the World Health Organization's (WHO) openly available ICD-11 classification and the ICD-11-CDDR manual. The test set consists of 104 clinical cases taken from the DSM-5-TR Clinical Cases book. To disentangle how much each design component contributes to diagnostic performance, nine configurations (K1–K9) were constructed within the scope of this thesis, and the primary diagnosis of each case was evaluated at both the diagnosis-name level and the code levels (diagnostic group, category, and subcategory).

The best-performing configuration, K4, combines four retrieval components: a clinical summary generated from the clinical history via Hypothetical Document Embeddings (HyDE), lexical matching with BM25, cosine similarity between 1024-dimensional dense embeddings, and Reciprocal Rank Fusion (RRF), which merges these signals. In the proposed system, the retrieved candidates were presented to the Gemma 4 12B model together with the clinical diagnostic criteria, and the model performed a criterion-based diagnosis selection. The proposed system reached a diagnosis-name-level accuracy of 0.615 (N=104), surpassing the highest open-source score in the literature, 0.524 (LLMind), in point estimate and achieving significance in a one-sided binomial test against the fixed reference ($p=0.038$). The mixture-of-experts Gemma 4 26B-A4B model, operating with roughly one-third of the active parameters, provided a computationally more efficient alternative at 0.606. The controlled nine-configuration ablation identified HyDE as the most decisive retrieval component, and

the contribution of RAG over the base model was found significant in a paired McNemar test ($p<0.001$). The diagnostic contribution of RAG was observed to depend on the metric used: it is consistently large at the code levels (diagnostic group, category, and subcategory), whereas at the diagnosis-name level it varies with the model's parametric knowledge. This finding shows that LLMind, which reports only at the diagnosis-name level, obscures RAG's contribution at the code level and demonstrates the necessity of multi-granular evaluation.

KEYWORDS: Clinical Decision Support Systems, Large Language Models, Mental Disorder Diagnosis, Natural Language Processing, Retrieval Augmented Generation

COMMITTEE: Prof. Dr. Süleyman BİLGİN

Prof. Dr. Övünç POLAT

Prof. Dr. Alper BİLGE

Prof. Dr. Yalçın İŞLER

Assist. Prof. Dr. Ufuk ÖZKAYA

ÖNSÖZ

Bu tez çalışması, ruhsal bozukluk tanısında erişim destekli üretim temelli büyük dil modellerinin klinik karar desteğine katkısını incelemektedir. Bu çalışmanın, klinisyenlerin tanı yükünü hafifleten güvenilir karar destek araçlarının geliştirilmesine ve erişilebilir çözümlere katkı sağlamasını temenni ederim. Ortaya konan yöntem ve bulguların bu alanda çalışacak araştırmacılara da yol gösterici olmasını dilerim.

Tez çalışmam boyunca değerli yönlendirmeleri ve bilimsel katkıları için danışmanım Prof. Dr. Süleyman BİLGİN'e ve hesaplama kaynakları için TÜBİTAK ULAKBİM TRUBA altyapısına teşekkür ederim.

Bu süreçte sabrı ve sevgisiyle her zaman yanımda olan eşime, desteğini hiç esirgemeyen, emekleriyle bugünlere gelmemi sağlayan ablam, annem ve babama içtenlikle teşekkür ederim.

Egehan ÇETİN

Antalya, 2026

İÇİNDEKİLER

ÖZET i
ABSTRACT iii
ÖNSÖZ v
İÇİNDEKİLER vi
AKADEMİK BEYAN viii
SİMGELER VE KISALTMALAR ix
ŞEKİLLER DİZİNİ xii
ÇİZELGELER DİZİNİ xiii
1. GİRİŞ 1
2. KAYNAK TARAMASI 3
2.1. Büyük Dil Modelleri 3
2.2. Erişim Destekli Üretim (RAG) ve Erişim Teknikleri 5
2.3. ICD-11 ve DSM-5-TR Sınıflandırması 8
2.4. Klinik DDİ ve BDM Sistemlerinin Değerlendirme Metodolojisi 9
2.5 Ruhsal Sağlıkta Büyük Dil Modelleri Çalışmaları 9
3. MATERYAL VE METOT 13
3.1. Test Veri Seti 13
3.2. Bilgi Tabanı 14
3.3. Tanı Standardı ve Doğrulaması 16
3.4. Erişim 17
3.5. Üretim 21
3.6. Erişim Konfigürasyonları ve Tasarımı 23
3.6.1. Taban ve Hibrit Erişim (K1–K3) 24
3.6.2. Önerilen Konfigürasyon (K4) 25
3.6.3. Zenginleştirme Konfigürasyonları (K5, K9) 27
3.6.4. HyDE Olmadan Alternatifler (K6–K8) 27
3.7. Değerlendirme Çerçevesi 28
3.7.1. Tanısal Eşleştirme Düzeyleri 28
3.7.2. Belirsizlik ve Anlamlılık Testleri 31
3.8. Donanım ve Uygulama (TRUBA) 32
3.9. Karar Destek Arayüzü (Web Uygulaması) 32
4. BULGULAR 35
4.1. Ana Karşılaştırma: Sistemler ve Referans Değer 35
4.2. İstatistiksel Anlamlılık 36

4.3. RAG Konfigürasyon Ablasyonu (K1–K9) 38
4.4. Erişim Analizi - Recall@k 39
4.5. Model Ölçekleme 42
4.6. Hata Örüntüsü - Karışıklık Matrisi 44
4.6.1. Tanı Grubu Bazlı Performans 45
4.6.2. Hata Kaynağı ve Hata Taksonomisi 46
4.7. Metriğe Bağlılık ve RAG'ın Değeri 48
5. TARTIŞMA 49
5.1. Metodolojik Katkılar 51
5.2. Sınırlamalar 51
6. SONUÇLAR 53
7. KAYNAKLAR 55
8. EKLER 59
ÖZGEÇMİŞ

AKADEMİK BEYAN

Doktora Tezi olarak sunduğum “Ruhsal Bozuklukların Tanısında Erişim Destekli Üretim Tabanlı Büyük Dil Modelleri Kullanılarak Bir Klinik Karar Destek Sistemi Geliştirilmesi” adlı bu çalışmanın, akademik kurallara ve etik değerlere uygun olarak yazıldığını; bu çalışmada bana ait olmayan tüm bilgi ve sonuçların bilimsel atıf kurallarına uygun olarak gösterildiğini ve kaynak gösterilen çalışmaların kaynaklar listesinde yer aldığını beyan ederim.

25/06/2026

Egehan ÇETİN

SİMGELER VE KISALTMALAR

Simgeler

b	: BM25 uzunluk normalizasyon parametresi
d	: Belge boyutu
d_k	: Anahtar vektör boyutu
$E(\cdot)$	: Gömme fonksiyonu
k	: RRF sıralama sabiti
$k1$	: BM25 terim doygunluk parametresi
N	: Örneklem sayısı
p	: İstatistiksel anlamlılık değeri
q	: Sorgu (vaka metni)
Q, K, V	: Sorgu, anahtar ve değer matrisleri
$s(q,d)$	: Sorgu-belge kosinüs benzerlik skoru
z	: Logit vektörü
α	: Krippendorff alfa güvenilirlik katsayısı
γ	: RMSNorm öğrenilen kazanç vektörü
ε	: Sayısal kararlılık sabiti
η	: Erişici bileşen parametreleri
θ	: Üretici model parametreleri
κ	: Cohen kappa güvenilirlik katsayısı
τ	: Örnekleme sıcaklık parametresi
$\chi 2$	: McNemar testi ki-kare istatistiği

Kısaltmalar

API	: Uygulama Programlama Arayüzü
BM25	: Okapi Best Matching 25
BT	: Bilgi Tabanı

CPU : Merkezi İşlem Birimi

GPU : Grafik İşlem Birimi

DDİ : Doğal Dil İşleme

DN : Doğru Negatif

DP : Doğru Pozitif

DPO : Doğru Pozitif Oranı

DSM-5-TR : Ruhsal Bozuklukların Tanısal ve İstatistiksel El Kitabı, 5. Baskı (Metin Revizyonu)

F1 : Kesinlik ve Duyarlılığın harmonik ortalaması

FN : Yanlış Negatif

FP : Yanlış Pozitif

GA : Güven Aralığı

GRRAS : Guidelines for Reporting Reliability and Agreement Studies

HyDE : Varsayımsal Belge Gömmeleri (Hypothetical Document Embeddings)

ICD-11 : Uluslararası Hastalık Sınıflandırması, 11. Revizyon

ICD-11-CDDR: ICD-11 Ruhsal, Davranışsal ve Nörogelişimsel Bozukluklar için Klinik Tanımlamalar ve Tanısal Gereksinimler

ISO/TR 12300 : Terminoloji sistemleri arası eşleme ilkeleri

BDM : Büyük Dil Modeli (Large Language Model)

MCC : Matthews Korelasyon Katsayısı (Matthews Correlation Coefficient)

MMS : ICD-11 Ölüm ve Hastalık İstatistikleri Ana Sınıflandırması (Mortality and Morbidity Statistics)

MoE : Karışık Uzman Mimarisi (Mixture of Experts)

OOS : Kapsam Dışı (Out of Scope)

QA : Soru Cevap (Question Answering)

RAG : Erişim Destekli Üretim (Retrieval-Augmented Generation)

Recall@k : En iyi k sonuç içinde kapsama oranı

RRF : Karşılıklı Sıralama birleştirme (Reciprocal Rank Fusion)

TN : Gerçek Negatif

TP : Gerçek Pozitif

WHO : Dünya Sağlık Örgütü

YN : Yanlış Negatif

YP : Yanlış Pozitif

YPO : Yanlış Pozitif Oranı

ŞEKİLLER DİZİNİ

Şekil 3.1. BT'nin iki kaynaktan oluşturulma süreci 15
Şekil 3.2. Tanı standardının oluşturulması ve doğrulanması 16
Şekil 3.3. Üç sıralayıcı (BM25, yoğun erişim, HyDE) ile RRF birleştirmesinin 11. Vaka üzerindeki davranışı 20
Şekil 3.4. Önerilen sistemin (K4 + Gemma 4 12B) uçtan uca algoritma akış diyagramı 26
Şekil 3.5. Tanısal eşleştirme düzeylerinin hiyerarşisi 29
Şekil 3.6. 11. vakanın klinik tanı karar destek arayüzü örneği 34
Şekil 4.1. Tanı adı düzeyi doğruluğu ve %95 bootstrap güven aralıkları orman grafiği 37
Şekil 4.2. K1–K9 konfigürasyonlarının doğruluk karşılaştırması 39
Şekil 4.3. Kategori düzeyi recall@k eğrileri (kapsam içi 96 vaka) 41
Şekil 4.4. Vakaların en üst sıra ICD-11 adayına kosinüs benzerliği dağılımı 41
Şekil 4.5. Model boyutuna karşı tanı adı düzeyi doğruluk (E2B, E4B, 12B, 26B-A4B); iki eğri RAG (K4) ve RAG olmadan konfigürasyonları 42
Şekil 4.6. RAG yöntemi ve BDM ısı haritası (tanı adı düzeyi doğruluk, N=104) 43
Şekil 4.7. Tanı grubu düzeyi karışıklık matrisi (önerilen K4-12B, kapsam içi vakalar) 44
Şekil 4.8. ICD-11 Bölüm 6 bozukluğunun bge-large gömmelerinin iki boyuta izdüşümü (t-SNE, kosinüs metriği) 45
Şekil 4.9. Önerilen K4-12B konfigürasyonunun ICD-11 tanı grubu bazında tanı adı düzeyi doğruluğu 46
Şekil 4.10. Önerilen K4-12B hatalarının kaynak (sol) ve tür (sağ) dağılımı (modele sunulan tam en iyi-18 aday kümesi üzerinde) 47
Şekil 4.11. Üç önde gelen konfigürasyonun (K1, K2, K4) dört değerlendirme düzeyinde karşılaştırması 48

ÇİZELGELER DİZİNİ

Çizelge 3.1. Değerlendirme korpusunun ICD-11 tanı gruplarına göre dağılımı (104 vaka) 13
Çizelge 3.2. Erişim konfigürasyonları (K1–K9) 23
Çizelge 3.3. K4 konfigürasyonunun parametreleri 26
Çizelge 4.1. En yüksek doğruluk tespit edilen konfigürasyonlar 35
Çizelge 4.2. Tanı grubu düzeyinde elde edilen metrikler 36
Çizelge 4.3. Doğruluk ve LLMind'a karşı anlamlılık (N=104; tek yönlü binom) 36
Çizelge 4.4. RAG konfigürasyon ablasyonu karşılaştırması (Gemma 4 26B-A4B, N=104) 38
Çizelge 4.5. Kategori düzeyi recall@18 (üretime aktarılan en iyi-18 işletim noktası; kapsam içi 96 vaka) 40
Çizelge 4.6. RAG yöntemi - BDM karşılaştırması (tanı adı düzeyi doğruluk, N=104) 43
Çizelge 4.7. Niceleme duyarlılığı: K1 erişimi (Gemma 4 26B-A4B, N=104) 44
Çizelge 4.8. Önerilen K4-12B konfigürasyonunun ICD-11 tanı grubu bazlı tanı adı düzeyi doğruluğu 45
Çizelge 4.9. Önerilen K4-12B hata taksonomisi ve kaynak ayrımı (toplam 40 hata). 47
Çizelge 5.1. LLMind çalışmasının diğer açık kaynak modellerle performansının karşılaştırılması (Cremaschi vd., 2025) 50

1. GİRİŞ

Ruhsal bozukluklar, dünya genelinde son derece yaygın ve büyük ölçüde tedavisiz kalan bir halk sağlığı sorunudur. Dünya Sağlık Örgütü'ne (WHO) göre her sekiz kişiden biri, yani yaklaşık bir milyar insan, bir ruhsal bozuklukla yaşamakta ve bu bozukluklar, sağlıkla geçirilemeyen yılların başlıca nedenleri arasında yer almaktadır (WHO 2022). Buna karşın ruh sağlığına ayrılan kaynaklar yetersiz kalmakta, ülkeler sağlık bütçelerinin ortalama %2'sinden azını bu alana ayırmakta ve tanı ile tedavi arasındaki açık genişlemeye devam etmektedir (WHO 2022). Uzman hekim sayısının yetersizliği ve uzun bekleme süreleri, hekimin tanısal akıl yürütmesini destekleyecek; ölçeklenebilir, güvenilir ve erişilebilir karar destek araçlarına olan ihtiyacı her geçen gün artırmaktadır.

Ruhsal bozuklukların tanısı, tıbbın pek çok alanından açıkça ayrışmaktadır. Kan tahlili, görüntüleme ya da biyopsi gibi nesnel ve ölçülebilir sınamaların bulunduğu dallardan farklı olarak psikiyatrik tanı; büyük ölçüde öznel, niceliklendirilmesi güç belirtilere ve kişiden kişiye, kültürden kültüre değişen içsel deneyimlere dayanmaktadır (Kapur vd. 2012). Belirtilerin bozukluklar arasında geniş ölçüde örtüşmesi, hastaların çoğu zaman birden çok tanıyı bir arada taşıması (komorbidite) (Kessler vd. 2005) ve klinik tablonun gelişimsel ve kültürel bağlama göre değişmesi, ayırıcı tanıyı güçleştirmektedir (First vd. 2021). Biyolojik, psikolojik, sosyal ve çevresel etkenlerin bu iç içe geçmiş etkileşimi, aynı vakada farklı klinisyenlerin farklı sonuçlara varabildiği gözlemciler arası değişkenliğe zemin hazırlamaktadır (Reed vd. 2018; Regier vd. 2013). Eksik tanı hastayı tedavisiz bırakırken, aşırı tanı gereksiz tedaviye, damgalanmaya ve kaynak israfına yol açmaktadır.

Büyük dil modelleri (BDM), çok geniş metin kümeleri üzerinde eğitildiklerinden serbest metinle yazılmış klinik anlatıları işleyebilmekte ve kayda değer ölçüde klinik bilgi kodlamaktadır (Singhal vd. 2023). Sınırları iyi tanımlanmış tanıma görevlerinde uzman düzeyine yaklaşan, bazı durumda klinisyenleri dahi geçen bir başarı sergilemeleri (Kim vd. 2024), bu modellerin tanısal akıl yürütmeyi destekleme, ayırıcı tanıya katkı sunma, genç klinisyenlerin eğitimi ve vaka önceliklendirmesi gibi rollerde değerli olabileceğini düşündürmektedir (Omar vd. 2024). Bu nedenle BDM'ler, ruh sağlığı alanındaki karar destek araştırmalarının merkezinde yer alabilmektedir.

Bununla birlikte halüsinasyon (Kim vd. 2025), eğitim verisinden gelen yanlılık (Bouguettaya vd. 2025) ve özellikle tanısal eşikleri uygulayamamaktan doğan aşırı tanı eğilimi (Sarma vd. 2025), bu modellerin klinik kullanımının önündeki başlıca engellerdir. Sarma vd. (2025), DSM-5-TR vinyetlerinde harici ve yapılandırılmış bilgi eklemenin pozitif kestirim değerini gözle görülür biçimde artırdığını göstermektedir. Bu bulgu, erişim destekli üretim (Retrieval Augmented Generation, RAG) paradigmasının alandaki temel gerekçesini oluşturur. RAG, modelin çıktısını çıkarım anında otoriter bir bilgi tabanından getirilen belgelerle koşullandırır (Lewis vd. 2020); böylece olgusal bilgiyi modelin donmuş ağırlıkları yerine güncellenebilir ve denetlenebilir bir indekste tutarak hem halüsinasyonu azaltır hem de kaynağın izlenebilirliğini sağlar. Düzenli olarak güncellenen ve otoriter olması beklenen ICD-11 gibi bir sınıflandırmaya dayalı tanı görevinde özellikle değerlidir.

Bir karar destek aracının gerçek dünyada benimsenebilmesi için aynı zamanda erişilebilir, mahremiyete duyarlı ve tekrarlanabilir olması da gerekir. Açık ağırlıklı olmayan, buluta bağımlı modeller bu gereksinimlerle çelişebilir. Hassas klinik verinin kurum dışına çıkarılmasını gerektirir, ayrıca maliyet ve sürüm kararlılığı bakımından dış bağımlılık yaratırlar. Bu nedenle tez boyunca yalnızca açık ağırlıklı modeller kullanılmış ve tüm işlemler, verinin yerelde işlendiği bir tasarımla merkezi işlem birimi (CPU) üzerinde yürütülmüştür.

Bu çalışmanın amacı, bir klinik vaka anlatısından hareketle hastanın birincil ICD-11 ruhsal bozukluk tanısını öngören, RAG temelli bir BDM sistemi tasarlamak, geliştirmek ve denetlenebilir bir değerlendirme çerçevesi altında doğrulamaktır. Amaç, BDM'lerin tek başına kullanıldığında sergilediği halüsinasyon ve aşırı tanı eğilimini, üretimi belirli bir bilgi tabanıyla koşullandırarak azaltmak ve elde edilen sistemi ad ile kod düzeylerini kapsayacak biçimde, tekrarlanabilir ve önceki çalışmalarla karşılaştırılabilir bir zeminde değerlendirmektir. Bunun yanında erişim ve üretim bileşenlerinin tanısal başarıya katkısı sistematik bir ablasyonla ayrıştırılarak, sistemin hangi tasarım tercihlerinden ne ölçüde yararlandığı açıklığa kavuşturulmaktadır.

Çalışmanın kapsamı, ICD-11'in 6. Bölümü (Ruhsal, Davranışsal ve Nörogelişimsel Bozukluklar) ile sınırlandırılmıştır. Bilgi tabanı (BT), bu bölüme ait 720 bozukluğun WHO ICD-11 ve CDDR klinik kriterlerinden derlenen tanımlarından oluşmaktadır. Değerlendirme ise 104 DSM-5-TR klinik vakası üzerinde yürütülmüştür. Her vaka için yalnızca birincil tanı tespit edilmekte, eş tanıların tam listelenmesi kapsam dışında tutulmaktadır. Tüm hesaplamalar, yüksek başarımlı TRUBA platformunda açık ağırlıklı modellerle gerçekleştirilmiştir. Önerilen sistem, klinisyenin yerini alan değil, onun tanısal akıl yürütmesini destekleyen bir karar destek bileşeni olarak konumlandırılmaktadır.

Bu doğrultuda üç bağımsız eşleme yönteminin çapraz doğrulanmasıyla yöntemlerin uyuştuğu vakalarda ruhsal kategori kodu doğrudan kesinleşmiş, uyuşmazlıklar WHO ICD-11 ve CDDR metni temel alınarak elle çözülmüştür. Elde edilen ve ISO/TR 12300 eşleme kalitesi ilkelerine uygun raporlanan bir tanı standardı oluşturulmuştur. İkincisi, referans sistemin otomatik protokolünü birebir yeniden üreten çok düzeyli (tanı adı, tanı grubu, kategori, alt kategori) ve düzenlenmiş bir değerlendirme çerçevesi geliştirilmiştir. Geliştirilen bir RAG hattının önceki çalışmalardaki değerini geçtiği ve K4 konfigürasyonda ve Gemma 4 12B modeliyle sabit referansa karşı anlamlılığa ulaştığı (binom $p=0.038$) gösterilmiş ve dokuz konfigürasyonlu bir ablasyonla (K1–K9) her erişim bileşeninin katkısı ayrıştırılmıştır. Bu çözümleme, HyDE'yi erişimin en belirleyici kaldıracı olarak öne çıkarmaktadır. RAG'ın değerinin metriğe bağlı biçimde özellikle kod düzeyinde belirginleştiğini ve mevcut sistemlerin gözlenen normatif vakalarda aşırı tanıya yöneldiğini ortaya koymaktadır. Dolayısıyla tezin değer önerisi, ruhsal bozukluk tanısında RAG-BDM sistemlerinin denetlenebilir ve tekrarlanabilir biçimde değerlendirilmesi için bir çerçeve sunmasıdır.

2. KAYNAK TARAMASI

2.1. Büyük Dil Modelleri

BDM'ler, geniş metin korpuslarında öz denetimli biçimde, bir sözcük dizisindeki bir sonraki belirteci öngörme hedefiyle eğitilen, milyarlarca parametreli derin sinir ağlarıdır. Güncel BDM'lerin tamamına yakını, Vaswani vd. (2017) tarafından önerilen dönüştürücü mimarisine dayanmaktadır. Bu mimarinin çekirdeğindeki öz dikkat düzeneği, bir dizideki her belirtecin diğer tüm belirteçlerle ilişkisini doğrudan modelleyerek hem uzun mesafeli bağımlılıkların yakalanmasını hem de eğitimin paralel yürütülmesini sağlamaktadır. Bu paralellik, model ve veri kümesi ölçeklerinin büyütülebilmesinin önünü açmaktadır.

Bu mimarinin matematiksel temeli, dilin özbağlanımlı (autoregressive) olasılıksal ayrıştırmasıdır. Model, T uzunluğundaki bir belirteç dizisinin ortak olasılığını, her belirtecin kendinden önceki bağlama koşullu olasılıklarının çarpımı olarak modellemektedir (Bağıntı 2.1). Eğitimde bu olabilirliğin negatif logaritması, yani çapraz entropi kaybı azaltılmaktadır (Bağıntı 2.2).

$$P_\theta(x) = \prod_{t=1}^{T} P_\theta(x_t \mid x_{<t}) \tag{2.1}$$

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log P_\theta(x_t \mid x_{<t}) \tag{2.2}$$

Transformer mimarisinin çekirdek işlemi ölçekli nokta çarpım dikkatidir (Vaswani vd. 2017). Sorgu (Q), anahtar (K) ve değer (V) matrisleri üzerinden her belirteç, dizideki diğer belirteçlerin ilgililik ağırlıklı bir birleşimini hesaplamaktadır (Bağıntı 2.3). Burada d_k anahtar vektörlerinin boyutudur ve karekök ölçeklemesi softmax işlevinin doygunluğa erişmesini önlemektedir. Uygulamada bu işlem h paralel dikkat başlığıyla yürütülür ve başlık çıktıları birleştirilerek bir çıkış izdüşümünden geçirilir (Bağıntı 2.4).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{2.3}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \tag{2.4}$$

Ön eğitimde model, web ölçeğindeki genel metinlerden dilbilgisel, olgusal ve kısmen akıl yürütmeye dönük örüntüleri öğrenmektedir. Bu durum ona geniş ama örtük

ve durağan bir bilgi tabanı kazandırır. Ardından gelen hizalama aşamasında model, talimatla ayarlama ve insan geri bildirimiyle pekiştirmeli öğrenme gibi tekniklerle kullanıcı yönergelerini izleyecek biçimde uyarlanmaktadır. Ölçek büyüdükçe, modelin açıkça eğitilmediği görevleri yalnızca istemdeki birkaç örnekle ya da hiç örnek olmadan yapabilmesi mümkün hale gelmektedir (Brown vd. 2020) ve yeterli ölçekte yeni yetenekler belirmektedir (Wei vd. 2022). Bu özellik, ağırlıkları güncellemeden yalnızca istem tasarımıyla yeni görevlere uyarlanabilmeyi sağlar ve bu tezde benimsenen, ince ayar gerektirmeyen erişim temelli yaklaşımın da temelini oluşturmaktadır.

Çıkarım aşamasında model davranışı için bağlam penceresi ve kod çözme etkendir. Bağlam penceresi tarafında modelin tek seferde işleyebileceği belirteç sayısı sınırlı olduğundan, erişimle getirilen aday belgelerin sayısı ve uzunluğu bu pencereye sığacak biçimde yönetilmek durumundadır. Kod çözmede ise bir sonraki belirteç, olasılık dağılımından örnekleme ya da en olasının seçilmesi yoluyla belirlenmektedir, sıcaklık parametresi ile bu dağılımın keskinliğini denetler. Sıcaklığın sıfıra çekilmesi üretimi deterministik hale getirir. Tanısal tutarlılık ve tekrarlanabilirlik gerektiren klinik bir görev için tercih edilen ayardır.

Kod çözme işleminde model, son katman gizli durumundan sözcük dağarcığı üzerinde bir logit vektörü (z) üretir ve bir sonraki belirteç, sıcaklık parametresi τ ile yumuşatılmış softmax dağılımından seçilmektedir (Bağıntı 2.5). τ→0 limitinde seçim en yüksek olasılıklı belirtece indirgenir, tanısal tutarlılık ve tekrarlanabilirlik gerektiren bu çalışmada düşük sıcaklık kullanılmıştır.

$$P(x_t = w \mid x_{<t}) = \frac{\exp\left(\frac{z_w}{\tau}\right)}{\sum_{v \in \mathcal{V}} \exp\left(\frac{z_v}{\tau}\right)} \tag{2.5}$$

Bu tezde kullanılan Gemma 4 ailesinin de paylaştığı modern kod çözücü mimarilerde üç temel öne çıkmaktadır. İlk olarak, katman girişlerinde ortalamasız bir ölçek normalizasyonu olan RMSNorm'dur (Bağıntı 2.6; Zhang ve Sennrich 2019). Burada γ öğrenilen kazanç vektörü, ε sayısal kararlılık sabitidir. Diğeri, konum bilgisini sorgu ve anahtar vektörlerine döndürme matrisleri R_Θ ile işleyen döner konum gömmesidir (Rotary Position Embedding, RoPE). Dikkat puanı yalnızca göreli konuma (m−n) bağlı hale gelir ve boyut çiftlerinin frekansları taban b ile geometrik olarak ölçeklenir (Bağıntı 2.7; Su vd. 2024). Bir diğeri, doğrusal izdüşümün öğrenilen bir geçitle çarpıldığı geçitli ileri besleme katmanıdır (GeGLU) (Bağıntı 2.8; Shazeer 2020). Uzman karması (Mixture-of-Experts, MoE) varyantta ise ileri besleme katmanı N ayrı uzman ağına (E_i) bölünür; bir yönlendirici, her belirteç için yalnızca en yüksek puanlı k uzmanı etkinleştirir ve çıktı, yönlendirici ağırlıklarıyla birleştirilir (Bağıntı 2.9; Shazeer vd. 2017). Böylece toplam parametre kapasitesi büyürken belirteç başına hesap maliyeti, yalnızca aktif parametre sayısıyla sınırlı kalmaktadır.

$$\mathrm{RMSNorm}(x) = \gamma \odot \frac{x}{\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^{\,2} + \varepsilon}} \tag{2.6}$$

$$\langle R_{\Theta,m} q_m , R_{\Theta,n} k_n \rangle = g(q_m, k_n, m-n), \ \theta_i = b^{-2i/d} \tag{2.7}$$

$$\mathrm{FFN}(x) = W_2\big(\mathrm{GELU}(W_1 x) \odot (W_3 x)\big) \tag{2.8}$$

$$y = \sum_{i \in \mathrm{TopK}(g(x))} g_i(x) \cdot E_i(x), \ g(x) = \mathrm{softmax}(W_r x) \tag{2.9}$$

Bilginin modelin donmuş ağırlıklarında saklanması bilginin güncellenememesi ve denetlenememesi, kaynak gösterilemediğinden izlenebilirlik eksikliği ve olgusal temeli olmayan ancak akıcı ve kendinden emin ifadeler üretme eğilimi (halüsinasyon) oluşturur. Bu sınırlılıklar, BDM çıkarımını harici ve kapsamlı bir bilgi tabanıyla koşullandıran RAG ve özellikle hata maliyetinin yüksek olduğu klinik alanda dikkatli, insan denetimli kullanımın temel gerekçesini oluşturur. Ayrıca ağırlıkları erişime açık, açık ağırlıklı modeller (Gemma vb.), tescilli ve kapalı modellere kıyasla yerel, denetlenebilir ve veri gizliliğine duyarlı bir kullanım sağladığından, klinik karar destek araştırmaları için ayrı bir önem taşımaktadır.

2.2. Erişim Destekli Üretim (RAG) ve Erişim Teknikleri

RAG, bir dil modelinin üretimini çıkarım anında harici bir korpustan erişilen belgelerle koşullandıran bir paradigma olup Lewis vd. (2020) tarafından formüle edilmiştir. Bu paradigma, bilgi yoğun görevlerde olgusal bilginin modelin donmuş ağırlıkları yerine düzenlenebilir ve denetlenebilir bir indekste tutulmasını sağlayarak halüsinasyonu azaltmakta ve kaynak izlenebilirliği sunmaktadır.

Biçimsel olarak RAG, üretimi erişilen belgeler üzerinde marjinalize eden olasılıksal bir modeldir (Lewis vd. 2020). Erişici bileşen P_η, sorgu x için bilgi tabanı üzerinde bir belge dağılımı tanımlarken, üretici bileşen P_θ erişilen belge z ile koşullanmış çıktı dağılımını vermektedir. Uygulamada bu toplam, en yüksek erişim puanlı k belge ($Z_k(x)$) üzerinden hesaplanmaktadır (Bağıntı 2.10).

$$P(y \mid x) \approx \sum_{z \in Z_k(x)} P_\eta(z \mid x)\, P_\theta(y \mid x, z) \tag{2.10}$$

Erişim süreci, sözcüksel ve yoğun sinyallerin birleşimine dayanmaktadır. BM25 (Robertson ve Zaragoza 2009), olasılıksal ilgililik çerçevesinin pratik bir ürünüdür. Eğitim gerektirmeyen güçlü bir sözcüksel sıralayıcı olmakla birlikte, klinik öykü ile

formal kategori arasındaki sözcük farklılıklarını kapatamamaktadır. Yoğun erişim (dense retrieval) yaklaşımı (Karpukhin vd. 2020; Dense Passage Retrieval, DPR) bu boşluğu öğrenilmiş ikili kodlayıcı gömmeleriyle daraltmakta ve açık alan soru yanıtlama (QA) görevlerinde BM25'i net biçimde geçmektedir. Cümle gömmeleri (Reimers ve Gurevych 2019; Sentence-BERT) ile genel amaçlı BGE ailesi (Xiao vd. 2024), modern erişim modellerinin pratik temelini oluşturmaktadır.

BM25, bir sorgu q ile belge d arasındaki ilgililiği, terim frekansını doyuran ve belge uzunluğunu normalize eden bir bağıntıyla puanlamaktadır (Bağıntı 2.11). Terim ağırlığı ters belge frekansıyla (IDF) belirlenmektedir (Bağıntı 2.12). Burada $f(q_i,d)$ terimin belgedeki frekansı, $|d|$ belge uzunluğu, avgdl korpustaki ortalama belge uzunluğu, N korpus boyutu ve $n(q_i)$ terimi içeren belge sayısıdır; k_1 ve b sırasıyla frekans doygunluğunu ve uzunluk normalizasyonunu denetleyen serbest parametrelerdir.

$$\mathrm{BM25}(q,d) = \sum_{i=1}^{|q|} \mathrm{IDF}(q_i) \cdot \frac{f(q_i,d) \cdot (k_1+1)}{f(q_i,d) + k_1\left(1 - b + b \cdot \frac{|d|}{\mathrm{avgdl}}\right)} \tag{2.11}$$

$$\mathrm{IDF}(q_i) = \ln\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}\right) \tag{2.12}$$

Yoğun erişimin başarımı, metinlerin anlamsal yakınlığının bir gömme uzayında kosinüs benzerliğiyle modellenmesine dayanmaktadır. Aynı tanı grubuna ait kayıtların bu uzayda kümelenmesi, sorguya anlamsal olarak yakın adayların getirilebilmesini sağlamaktadır. Erişim mimarileri, hesaplama maliyeti ile kesinlik arasında bir ödünleşim sunar. İkili kodlayıcılar, sorgu ile belgeyi bağımsız olarak kodlayıp belge gömmelerini çıkarım öncesinde hesaplanabilir yaptığından, geniş korpuslarda ölçeklenebilir ve hızlı bir ilk aşama erişimi mümkün olur (Karpukhin vd. 2020; Reimers ve Gurevych 2019). Fakat çapraz kodlayıcılar, sorgu belge çiftini birlikte kodlayarak daha yüksek kesinlik elde eder; ancak her çift için ayrı bir ileri geçiş gerektirdiğinden ölçeklenmez ve yalnızca küçük bir aday kümesinin yeniden sıralanmasında uygulanabilir (Nogueira ve Cho 2019). Bu nedenle çağdaş erişim sistemleri çoğunlukla iki kademeli kurgulanmaktadır. Ölçeklenebilir bir ikili kodlayıcıyla geniş bir aday havuzu daraltılır ve ardından isteğe bağlı bir çapraz kodlayıcı kademesiyle kesinlik artırılır.

Yoğun erişimde sorgu ve belge, öğrenilmiş bir kodlayıcı E ile aynı vektör uzayına gömülmekte ve benzerlik, gömme vektörleri arasındaki kosinüs ile ölçülmektedir (Bağıntı 2.13). Bu çalışmada E, 1024 boyutlu bge-large-en-v1.5 kodlayıcısıdır ve bilgi tabanı girdileri çevrimdışı olarak gömülüp indekslenmektedir.

$$s(q,d) = \cos\big(E(q), E(d)\big) = \frac{E(q) \cdot E(d)}{\|E(q)\| \, \|E(d)\|} \tag{2.13}$$

Varsayımsal Belge Gömmeleri (Hypothetical Document Embeddings, HyDE), denetimsiz erişim ortamında sorgunun kendisini gömmek yerine, büyük dil modeline ürettirilen varsayımsal bir cevap belgesini gömerek sorguyu belge uzayına taşır. Bu yaklaşım denetimsiz tasarımı belirgin biçimde geçmekte ve ince ayarlı erişim modellerine yaklaşmaktadır (Gao vd. 2023). HyDE en çok denetimsiz ortamlarda ya da alan kaymasının yaşandığı durumlarda fayda sağlamaktadır. Ancak büyük dil modelinin halüsinasyon riskini de almaktadır, çünkü yanıltıcı bir varsayımsal belge erişimi yanlış komşuluğa yönlendirebilir. Farklı sıralayıcıların çıktıları ise Karşılıklı Sıralama Birleştirmesi (RRF) ile herhangi bir parametreye gerek kalmadan yalnızca sıra bilgisine dayanarak birleştirilmektedir. İsteğe bağlı bir çapraz kodlayıcı yeniden sıralama kademesi (Nogueira ve Cho 2019), sorgu belge çiftini birlikte kodlayarak kesinliği artırır. HyDE'ye kavramsal olarak yakın bir başka yaklaşım, sorgu yeniden yazımını büyük dil modeliyle gerçekleştiren Query2doc'tur (Wang vd. 2023). Tıp alanında erişimle topraklamanın sağladığı kazanç, Xiong vd. (2024) tarafından MedRAG/MIRAGE üzerinde ampirik olarak gösterilmiş (düşünce zincirine kıyasla %18'e varan kazanç) ve ortada kaybolma etkisiyle birlikte log doğrusal korpus ölçeklenmesi belgelenmiştir.

HyDE'nin biçimsel temeli, sorgunun kendisini gömmek yerine, üretici dil modeli g'nin sorgudan ürettiği varsayımsal belgenin gömülmesidir (Bağıntı 2.14). Böylece kısa ve dolaylı klinik öykü, bilgi tabanındaki tanı metinlerinin diline taşınmakta ve gömme uzayında hedef belgelere yaklaşmaktadır (Gao vd. 2023).

$$v_q = E\big(g(q)\big) \tag{2.14}$$

Erişim mimarisinin tasarımındaki temel farklılık, kapsama (recall) ile kesinlik (precision) arasındadır. Tek bir yoğun ya da sözcüksel sıralayıcı geniş bir aday havuzunu hızla daraltır ama her biri kendine özgü kör noktalar taşıyabilir. Örneğin BM25 eş anlamlılığı kaçırır, yoğun erişim ise nadir ve ayırt edici terimleri seyreltebilir (Karpukhin vd. 2020). Bu kör noktaların büyük ölçüde örtüşmemesi, farklı sıralayıcıların birleştirilmesini değerli kılmaktadır. RRF, skor ölçeklerini eşitleme gereği duymadan yalnızca sıra bilgisini kullanarak tamamlayıcı isabetleri toplamaktadır (Cormack vd. 2009). Kesinliğin öncelikli olduğu durumlarda küçük bir aday kümesine uygulanan çapraz kodlayıcı yeniden sıralama (Nogueira ve Cho 2019) eklenebilir. Ancak bu kademe yalnızca getirilmiş adayları yeniden düzenler, hiç getirilmemiş doğru bir adaya destek olmaz. Bu tez, bu eksikliği birleştirme lehine çözmekte ve yeniden sıralamayı yalnızca bir ablasyon kolu olarak sınamaktadır.

Farklı sıralayıcıların çıktıları, karşılıklı sıralama birleştirmesiyle (RRF) tek bir listede toplanmaktadır (Cormack vd. 2009): her belge, her sıralayıcı r'deki sırasının kaydırılmış tersiyle puanlanır ve puanlar sıralayıcılar üzerinden toplanır (Bağıntı 2.15). Sabit k (tipik olarak 60), üst sıralardaki küçük sıra farklarının aşırı ağırlıklanmasını

önlemektedir. Yöntem ham skor ölçeklerinden bağımsız olduğundan, sözcüksel (BM25) ve yoğun sinyalleri herhangi bir kalibrasyon gerektirmeden birleştirebilmektedir.

$$\mathrm{RRF}(d) = \sum_{r \in R} \frac{1}{k + \mathrm{rank}_r(d)} \tag{2.15}$$

2.3. ICD-11 ve DSM-5-TR Sınıflandırması

Dünya Sağlık Örgütü'nün Uluslararası Hastalık Sınıflandırması, 11. Revizyon (ICD-11) (yürürlük tarihi: 1 Ocak 2022), ruhsal bozuklukları Bölüm 06 (6A00–6E6Z) altında toplamaktadır. Klinik kriterler ise 2024 yılında yayımlanan Klinik Tanımlar ve Tanı Gereksinimleri (CDDR) ile derlenerek erişime açık hale getirilmiştir (WHO 2024; Kestel ve Reed 2024). CDDR, sayım tabanlı bir yaklaşım yerine aşamalı şekilde ruhsal bozukluğa erişilen temel özellikler anlayışını benimsemekte ve yaşam boyu, kültürel ve gelişimsel boyutları sistematik biçimde kapsamaktadır.

Ruhsal Bozuklukların Tanısal ve İstatistiksel El Kitabı (DSM-5-TR), Amerikan Psikiyatri Birliği (APA) tarafından yayımlanan ve ruhsal bozuklukların standart tanı ölçütlerini tanımlayan başlıca başvuru kaynaklarındandır. 2022'de yayımlanan bu metin revizyonu, DSM-5'i (2013) güncel literatür ve ölçüt düzeltmeleriyle gözden geçirmektedir. Her bozukluk için tanımlayıcı ölçütleri, ayırıcı tanıları ve ilişkili özellikleri sunmaktadır. Özellikle Kuzey Amerika'da baskın olan bu sınıflandırma, ICD ile uyumlu kodlama içermekle birlikte ICD-11'den yapısal olarak ayrışmaktadır. Bu çalışmada DSM-5-TR'nin rolü, test vakalarının kaynağı olmasıdır ve sistemin tahminleri ise değerlendirme için ICD-11 koduna eşlenmektedir.

ICD-11 ile DSM-5-TR büyük ölçüde örtüşmekle birlikte aralarında yapısal farklılıklar bulunmaktadır. Bu farklılıkların ilki, ICD-11'in kişilik bozukluklarında kategorik modelden boyutsal modele geçmesidir. Tek bir kişilik bozukluğu tanısı (6D10) şiddet düzeyine göre derecelendirilmekte ve isteğe bağlı özellik alanı belirleyicileriyle (6D11: negatif duygulanım, içe kapanım, dissosiyalite, disinhibisyon, anankastia) zenginleştirilmektedir. İkinci farklılık, akut stres tepkisinin bir ruhsal bozukluk olarak değil, aşırı strese verilen normal ve geçici bir tepki olarak QE84 (Bölüm 24) altında kodlanmasıdır. Üçüncü farklılık ise uyku-uyanıklık bozukluklarının ruhsal bozukluklar bölümünden çıkarılarak ayrı bir Bölüm 07 altında toplanmasıdır (Seemüller vd. 2023). First vd. (2021), iki sistemde ortak olan 103 tanısal varlığı karşılaştırmış ve bunların %30.1'inin esasen özdeş, %40.8'inin küçük farklılıklar, %19.4'ünün ise büyük farklılıklar taşıdığını bulmuştur; bu, DSM-IV/ICD-10 dönemine kıyasla kayda değer bir ilerlemedir. Bu biçim farkı (sayım kriterleri ile prototip temelli temel özellikler arasındaki ayrım), mekanik bir DSM-ICD eşlemesinin neden yetersiz kaldığını açıklamaktadır. ICD-11'in tanısal sağlamlığı saha çalışmalarıyla da desteklenmiştir. Reed vd. (2018) çok merkezli alan çalışmasında ileri tanılar için iyi ve mükemmel düzeyde güvenilirlik ($\kappa \approx 0.45$–0.88) bildirmiştir. Gaebel vd. (2020) ise ICD-11'in ICD-10'a kıyasla açıkça daha yüksek tanısal doğruluk sağladığını (yaklaşık %71.9'a karşılık %53.2) ortaya koymuştur.

ICD-11 ruhsal bozukluklar bölümü; güvenilirlik, farklı klinisyenlerin aynı vakada aynı tanıya ulaşabilmesi ve klinik kullanışlılık, tanının iletişim, karar verme ve

tedavi planlamadaki işlevselliği açısından önemlidir (Reed vd. 2018). CDDR'ın benimsediği prototip eşleme yaklaşımı, klinisyenden katı bir belirti sayımı yerine hastanın klinik tablosunu bir bozukluğun temel özellikleriyle bütüncül biçimde karşılaştırmasını istemektedir. Bu yaklaşım, gerçek klinik akıl yürütmeye DSM'nin işlemselleştirilmiş sayım ölçütlerinden daha yakın görülmekte ve kültürlerarası uygulanabilirliği güçlendirmektedir (First 2009; Kestel ve Reed 2024). Bu tasarım tercihi, mevcut tezdeki kriter temelli üretim yaklaşımının da kavramsal dayanağını oluşturmaktadır. Dil modeline her aday tanının temel özellikleri sağlanarak, modelin prototip eşlemeye benzer bir gerekçelendirme yürütmesi amaçlanmaktadır.

2.4. Klinik DDİ ve BDM Sistemlerinin Değerlendirme Metodolojisi

Klinik karar destek sistemlerinde bir sistemin alternatiflerinden üstün olduğu yönündeki bir iddia, ancak titiz ve çok yönlü bir değerlendirme çerçevesiyle desteklendiğinde anlam kazanmaktadır. Standart dörtlü metrik kümesi (doğruluk, kesinlik, duyarlılık ve F1), klinik metin sınıflandırmasında kullanılmasına rağmen bu metrikler sınıf dengesizliği durumunda yanıltıcı olabilmektedir. Bu nedenle Hicks vd. (2022), tıbbi yapay zeka uygulamalarında yaygınlığa duyarlı Matthews korelasyon katsayısının (MCC) tamamlayıcı bir ölçüt olarak raporlanmasını önermektedir. Kısmi altın etiketlerin veya serbest metin gerekçelerinin derecelendirilmesi gerektiğinde ise hakem olarak büyük dil modeli (BDM-as-a-judge) yaklaşımı kullanılmaktadır. Zheng vd. (2023), GPT-4 yargıcının insan tercihleriyle %80'in üzerinde uyum gösterdiğini, bununla birlikte konum, ayrıntı düzeyi ve kendini kayırma türünden yanlılıklar taşıdığını belgelemiştir. Gu vd. (2024) ise bu yanlılıkları azaltmaya yönelik stratejileri derlemiştir.

Hakem ile insan ya da kodlayıcılar arasındaki uyumun, şansa göre düzeltilmiş istatistiklerle sayısallaştırılması gerekmektedir. Sürekli ölçümler için Lin uyum korelasyon katsayısı (Lin 1989), çok kodlayıcılı durumlar için ise Krippendorff α (Krippendorff 2004) kullanılmaktadır. Raporlamanın GRRAS kılavuzuna (Kottner vd. 2011) uygun biçimde yapılması beklenmektedir. Terminoloji eşleme kalitesi bakımından ilgili standart ise eşdeğerlik derecesi ve ilişki kavramlarını tanımlayan ISO/TR 12300 (ISO 2014) belgesidir. İstatistiksel çıkarımda, küçük ve sabit test kümeleri için dağılımdan bağımsız önyükleme güven aralıkları (Efron 1992), tek bir doğruluk değerinin bir referans değere karşı sınanması için binom testi, aynı vakalar üzerinde iki sistemin eşleştirilmiş karşılaştırması için ise McNemar testi (McNemar 1947) uygundur. Dietterich (1998), tek bir ayrılmış test kümesi kullanıldığında düşük Tip-I hatası nedeniyle McNemar testini önermektedir.

2.5 Ruhsal Sağlıkta Büyük Dil Modelleri Çalışmaları

Yapay zekanın tıp alanındaki performansı tüm uzmanlık alanlarında aynı düzeyde değildir. Tıbbi görüntü analizi ve ilaç geliştirme gibi nesnel verilere dayanan alanlarda insan uzmanlara yakın hatta bazı durumlarda onları aşan sonuçlar elde edilirken, ruh sağlığı alanındaki uygulamaların gelişimi daha sınırlı kalmıştır (Cremaschi vd., 2025). Bunun temel nedenlerinden biri, psikiyatrik tanı süreçlerinin doğasında bulunan karmaşıklıktır. Pek çok tıbbi branşta tanı; laboratuvar testleri, görüntüleme yöntemleri veya biyopsi gibi nesnel bulgularla desteklenebilirken, ruhsal bozuklukların değerlendirilmesi büyük ölçüde bireyin öznel deneyimlerine, klinik

gözlemlere ve yorumlanması güç belirtilere dayanmaktadır (Kapur vd. 2012). Ayrıca yüksek komorbidite oranları ile biyolojik, psikolojik, sosyal ve çevresel etkenlerin iç içe geçmiş yapısı, tanı sürecini daha da karmaşık hale getirmektedir. İstatistiksel örüntüler üzerinden öğrenen yapay zeka modelleri, bu çok katmanlı ve bağlama duyarlı yapıyı yeterince temsil etmekte zorlanabilmektedir. Bu nedenle DSM-5-TR ve ICD-11 gibi tanısal sınıflandırma sistemleri, ruhsal bozuklukların değerlendirilmesinde ortak bir terminoloji ve standart ölçütler sağlayarak klinik karar verme sürecine yön vermekte; aynı zamanda bu çerçevelere dayanan yapay zeka destekli karar sistemlerinin geliştirilmesi için önemli bir kuramsal temel oluşturmaktadır.

BDM'lerin kayda değer ölçüde klinik bilgi kodladığı, Singhal vd. (2023) tarafından MultiMedQA ölçütü üzerinde gösterilmiştir. Talimatla ayarlanmış Flan-PaLM, MedQA'da %67.6 doğruluğa ulaşmıştır ve alana uyarlanan Med-PaLM ise uzman değerlendirmesinde bilimsel uzlaşıyla büyük ölçüde uyumlu yanıtlar üretmiştir. Psikiyatriye özgü kanıtlar ise umut verici olmakla birlikte daha temkinli bir tablo çizmektedir. Omar vd. (2024), sistematik derlemelerinde GPT-4'ün klinik akıl yürütme, eğitim ve tarama görevlerinde deneyimli psikiyatristlere yaklaştığını, ancak hasta yönetimi için gereken klinik muhakemeyi tek başına yerine getiremediğini ve bu nedenle yardımcı bir araç olarak konumlanması gerektiğini belirtmiştir. Ancak iyi sınırlandırılmış tanıma görevlerinde Kim vd. (2024), üç büyük dil modelinin (ChatGPT-4, Gemini Pro, Llama 3) klinik öykülerden obsesif-kompulsif bozukluğu %96.1 doğrulukla tanıyarak doktora düzeyindeki psikoloji asistanlarını (%81.5) ve birinci basamak hekimlerini (%49.5) geride bıraktığını tespit etmiştir.

Alana özgü geliştirilmiş modeller bununla paralel bir araştırma akışı oluşturmaktadır. Singhal vd. (2025), Med-PaLM 2 ile MedQA'da %86.5 doğruluğa ulaşmıştır. Labrak vd. (2024) ise PubMed Central üzerinde sürdürülen ön eğitimle açık kaynaklı BioMistral modelini yayımlamıştır. Benzer biçimde MedGemma (Sellergren vd. 2025), Gemma 3 mimarisini çok etiketli tıbbi anlama görevleri için genişletmektedir. Ruhsal sağlık alanına özgü olarak ise Yang vd. (2024), açıklanabilir analizi ön plana çıkaran ilk açık kaynaklı talimat takip eden model serisi olan MentaLLaMA'yı geliştirmiştir.

Geliştirilen modellerin önemli kısıtları da vardır. Kim vd. (2025) tıbbi temel modellerdeki halüsinasyon türleri için bir taksonomi geliştirmiştir, ancak kendinden emin ama hatalı tanısal ifadelerin doğrudan hasta güvenliği riski taşıdığını ileri sürmüştür. Bouguettaya vd. (2025), incelenen dört büyük dil modelinin tanı kararlarında düşük düzeyde yanlılık göstermesine karşın, hastanın ırkı belirtildiğinde sıklıkla daha düşük nitelikli tedavi önerdiğini bildirmiştir. Bu teze en doğrudan ilişkili kısıt ise aşırı tanıdır: Sarma vd. (2025), GPT modellerini DSM-5-TR hasta öyküleri üzerinde değerlendirmiştir. Doğrudan istemleme yaklaşımının pozitif kestirim değerinin yalnızca %43.3 düzeyinde kaldığını ve bunun temel nedeninin tanısal eşikleri uygulayamamak olduğunu göstermiştir. Buna karşılık uzman türevli karar ağaçlarının sürece katılması, bu değeri yaklaşık %65.3'e çıkarmıştır. Bu bulgu, harici ve yapılandırılmış bilgiyle topraklamanın alandaki temel gerekçesini oluşturur.

RAG, klinik büyük dil modellerini otoriter kaynaklara topraklamak için baskın paradigma haline gelmiştir. İlgili derlemeler bu alanı tanısal destek, elektronik sağlık kayıtlarının (EHR) özetlenmesi ve tıbbi soru yanıtlama eksenleri üzerinden

yapılandırmakta ve yinelenen başarısızlık biçimlerine (erişim gürültüsü, alan kayması, gecikme ve zayıf açıklanabilirlik) işaret etmektedir. Güncel sistemler ise hassas tek atışlık erişimin ötesine geçmektedir. Bilgi tamlığı temelli uyarlanabilir erişim ve çok aşamalı düzenekler ön plana çıkmaktadır. Ge vd. (2025), büyük dil modellerinden RAG ve ajanlara uzanan çizgiyi izleyerek tespit ile tanı arasındaki ayrımı vurgulamakta, formal tanı kriterlerinin (DSM-5, ICD-11) yetersiz entegrasyonunu ve değerlendirme metriği uyumsuzluğunu açık birer boşluk olarak işaret etmektedir; her iki husus da bu tezle doğrudan ilgilidir.

Cremaschi vd. yaptıkları çalışmada, Gemma 2 27B üzerine kurulu ve ICD-11 bilgi tabanına dayanan LLMind isimli RAG sistemi geliştirmiştir. DSM-5-TR Clinical Cases (Barnhill 2023) korpusu üzerinde kurulu sistem ruhsal bozukluğun tespitini otomatik biçimde 0.524 doğrulukla, dört uzmanca elle 0.761 doğrulukla değerlendirilmiştir (Cremaschi vd. 2025). Bilinçli olarak eksik ve her zaman net olmayan hasta öyküleri içermesi, korpusun kendisini gerçekçi ancak zorlu bir ölçüt haline getirmektedir. Tıbbi ve psikiyatrik ölçütlerin büyük bölümü (MedQA, MedMCQA, PubMedQA) çoktan seçmeli soru yanıtlamaya dayanmaktadır. Ayrıca serbest metin biçimindeki bir hastalık öyküsünden ICD-11 koduna eşleme görevi, bilgi hatırlamaya kıyasla kategorik olarak daha zordur ve bu ölçütlerin kapsamı dışında kalmaktadır.

Türkçe kaynaklar içinde büyük dil modellerinin ruhsal sağlık ve tıbbi metin bağlamındaki kullanımı son yıllarda hızla artmıştır; ancak bu çalışmalar tipik olarak serbest metin soru yanıtlama ya da kaynak ve altyapı geliştirme odaklı olup, bir vaka öyküsünden formal bir nozolojiye tanı eşleme görevini ele almamaktadır. Türk vd. (2026), dokuz büyük dil modelini bir ruhsal sağlık soru yanıt veri kümesi üzerinde aynı koşullar altında değerlendirmiş ve GPT-4o'nun anlamsal benzerlik metriklerinde (COMET 0.727, METEOR 0.148) öne çıktığını, bununla birlikte daha küçük ve açık ağırlıklı modellerin belirli görevlerde rekabet edebildiğini ve model seçiminin yalnızca boyuta değil göreve uygunluğa da bağlı olduğunu raporlamıştır. Bununla birlikte söz konusu çalışma, üretilen metnin yüzeysel ve anlamsal benzerliğini ölçmekte; doğrulanmış bir tanı standardına karşı ICD-11 tanı sınıflandırması içermemektedir.

Lisansüstü tez düzeyinde de benzer konulara temas eden güncel çalışmalar bulunmaktadır. Bhat (2025, Kadir Has Üniversitesi), metinden mental sağlık sorunu tespitini birincil görev olarak belirlemiş ve Llama-3.1-8B-Instruct modelini parametre açısından verimli ince ayar (LoRA, 4-bit niceleme) ile sıralı çok görevli öğrenme aracılığıyla uyarlamıştır. Duygu ve duygu durum tespitini ise yardımcı görevler olarak kullanmıştır. Bu çalışma metin sınıflandırma odaklı ve ince ayar temelli olduğundan, erişim temelli (RAG, ince ayar gerektirmeyen) ve ICD-11 kod tanısına dayanan yaklaşımından yöntemsel olarak ayrışmaktadır. Karakurt (2026, İstanbul Kültür Üniversitesi), büyük dil modeli ve RAG tabanlı bir yaklaşımı insan kaynakları belgelerinden bilgi çıkarımına uygulayarak RAG'in Türkçe yazındaki uygulamalı bir örneğini sunmuştur. Söz konusu yöntem (BDM ve RAG bileşimi) bu teze yakın olmakla birlikte uygulama alanı ve görev, klinik tanıdan farklıdır. Bayraktar (2025, Üsküdar Üniversitesi) ise ruh sağlığı bilişiminde tepkisel müdahaleden veri temelli kesin psikiyatri paradigmasına geçişi, yapay zeka destekli ve insanın döngüde yer aldığı üç aşamalı bir kapsam derlemesiyle kavramsal düzeyde inceleyerek alanın yönelimini Türkçe konumlandırmıştır.

Daha genel olarak ise Abd-alrazaq vd. (2023), büyük dil modellerinin tıp eğitiminde soru yanıtlama sistemlerinde taşıdığı potansiyeli, sınırları ve etik boyutları sistematik biçimde derlemiştir. Klinik karar destek uygulamalarının vaat ettiklerini ve barındırdığı riskleri ortaya koymuştur. Türkçe tıbbi doğal dil işleme altyapısı bağlamında ise iki çizgi öne çıkmaktadır. Oğul vd. (2025), büyük dil modeli tabanlı çeviri aracılığıyla bir Türkçe tıbbi doğal dil çıkarımı veri kümesi (TurkMedNLI) oluşturmuştur. Bunun yanı sıra hasta ve hekim, soru ve yanıt verisiyle eğitilen Türkçeye özgü tıbbi bir büyük dil modeli geliştirilmiştir (Bayram vd. 2025). Bu çalışmalar Türkçe tıbbi doğal dil işleme alanı için değerli kaynaklar ortaya koymakla birlikte, ICD-11 kod düzeyinde tanı sınıflandırmasına ve ruhsal bozuklukların vaka öyküsü temelli tanısına odaklanmamaktadır.

Özetlenen literatürde, büyük dil modelleri sınırları iyi tanımlanmış tanıma görevlerinde güçlü olmasına rağmen serbest metinle yazılmış bir vaka öyküsünü formal bir nozolojinin kod düzeyindeki kategorilerine eşleme görevi, çoğu çoktan seçmeli olan mevcut tıbbi ölçütlerin kapsamı dışında kalmaktadır. Ek olarak klinik RAG derlemeleri, formal tanı kriterlerinin (DSM-5, ICD-11) sistemlere yetersiz entegrasyonunu ve değerlendirme metriklerindeki uyumsuzluğu açık birer boşluk olarak işaret etmektedir (Ge vd. 2025). Ayrıca referans çalışma LLMind dahil mevcut sistemler, yalnızca tanı adı düzeyinde raporlamaya ve çekimserlik mekanizması olmadan bir yapıya dayanmaktadır. Bu durum aşırı tanı eğilimini görünmez kılmaktadır. Son olarak Türkçe yazında, doğrulanmış bir tanı standardına ve RAG'a dayanan, DSM-5-TR öykülerinden ICD-11 ruhsal bozukluk tanısı öngören bir karar destek sistemi henüz ele alınmamıştır.

Bu tez, serbest metin öyküsünden ICD-11 kod düzeyinde tanı eşlemeyi, doğrulanmış bir tanı standardı ve çok düzeyli bir değerlendirme çerçevesiyle birlikte ele almaktadır. HyDE temelli erişimle formal tanı kriterlerini üretime taşır ve sistemi hem uluslararası LLMind sistemi ile hem de Türkçe literratürdeki boşlukla doğrudan karşılaştırır.

3. MATERYAL VE METOT

3.1. Test Veri Seti

Değerlendirme veri setimiz, DSM-5-TR Clinical Cases (Barnhill, ed.) adlı öğretici vaka kitabından alınan İngilizce dilinde 104 klinik vakadan oluşmaktadır. Her vaka, klinik öykü, uzmanın tanısal gerekçesi, birincil bozukluk ve ek bozukluklar bölümlerini içermektedir. Birden çok tanı taşıyan vakalarda "birincil tanı", kitabın kendi "birincil bozukluk" alanından alınmıştır. Çalışmada oluşturulan modele yalnızca klinik öykü alanı verilmektedir ve tanıyı ve gerekçeyi içeren uzmanın tanısal gerekçesi alanı modelden gizlenmektedir. Uzmanın tanısal gerekçesi alanı yalnızca tanı standardının oluşturulmasında kullanılmakta, model girdisinde hiçbir biçimde yer almamaktadır. Bu önlem, doğru yanıtın modele sızması riskini ortadan kaldırmaktadır.

104 vakanın 96'sı ICD-11 Bölüm 6 ruhsal bozukluklar kapsamında kapsam içi, 8'i ise kapsam dışı (OOS) olarak işaretlenmiştir. Kapsam dışı vakalar, DSM tanısının ICD-11'de Bölüm 6'nın dışında kodlandığı durumlardır. Örneğin uyku uyanıklık bozuklukları (Bölüm 7: 7A00 uykusuzluk, 7A41 obstrüktif uyku apnesi), akut stres tepkisi (Bölüm 24: QE84), cinsel işlev bozuklukları (HA blokları) ve cinsiyet uyumsuzluğu (HA60) bu kapsamdadır (Çizelge 3.1). Vaka sayılarının tanı grubu ve ICD-11 ruhsal bozukluk grubuna göre dağılımı Çizelge 3.1'de gösterilmiştir.

Çizelge 3.1. Değerlendirme korpusunun ICD-11 tanı gruplarına göre dağılımı (104 vaka)

Grup (ICD-11)	Tanı grubu	Vaka (n)
6A0	Nörogelişimsel bozukluklar	3
6A2	Şizofreni ve diğer birincil psikotik bozukluklar	8
6A6	Bipolar ve ilişkili bozukluklar	7
6A7	Depresif bozukluklar	12
6B0	Anksiyete ve korku ilişkili bozukluklar	6
6B2	Obsesif-kompulsif ve ilişkili bozukluklar	6
6B4	Stresle özgül ilişkili bozukluklar	6
6B6	Dissosiyatif bozukluklar	4
6B8	Yeme ve beslenme bozuklukları	5
6C0–6C2	Bedensel sıkıntı ve bedensel deneyim bozuklukları	3

(Devamı arkada)

Çizelge 3.1.'in devamı.

6C4–6C5	Madde kullanımı ve bağımlılık bozuklukları	10
6C7, 6C9	Dürtü denetimi ve yıkıcı/disosyal bozukluklar	3
6D1	Kişilik bozuklukları	9
6D3	Parafilik bozukluklar	2
6D7–6D8	Nörokognitif bozukluklar (deliryum / demans)	6
6E4, 6E6	Sekonder ruhsal / davranışsal sendromlar	4
6D5, 8A05	Yapay bozukluk ve (çapraz listeli) tik bozukluğu	2
Kapsam içi ara toplam		**96**
7A0, 7A4	Uyku-uyanıklık bozuklukları (Bölüm 7)	2
QE8	Akut stres tepkisi (Bölüm 24)	2
GA34	Premenstrüel disforik bozukluk (Bölüm 16)	1
HA0–HA6	Cinsel işlev bozuklukları / cinsiyet uyumsuzluğu	3
Kapsam dışı (OOS) ara toplam		**8**
Toplam		**104**

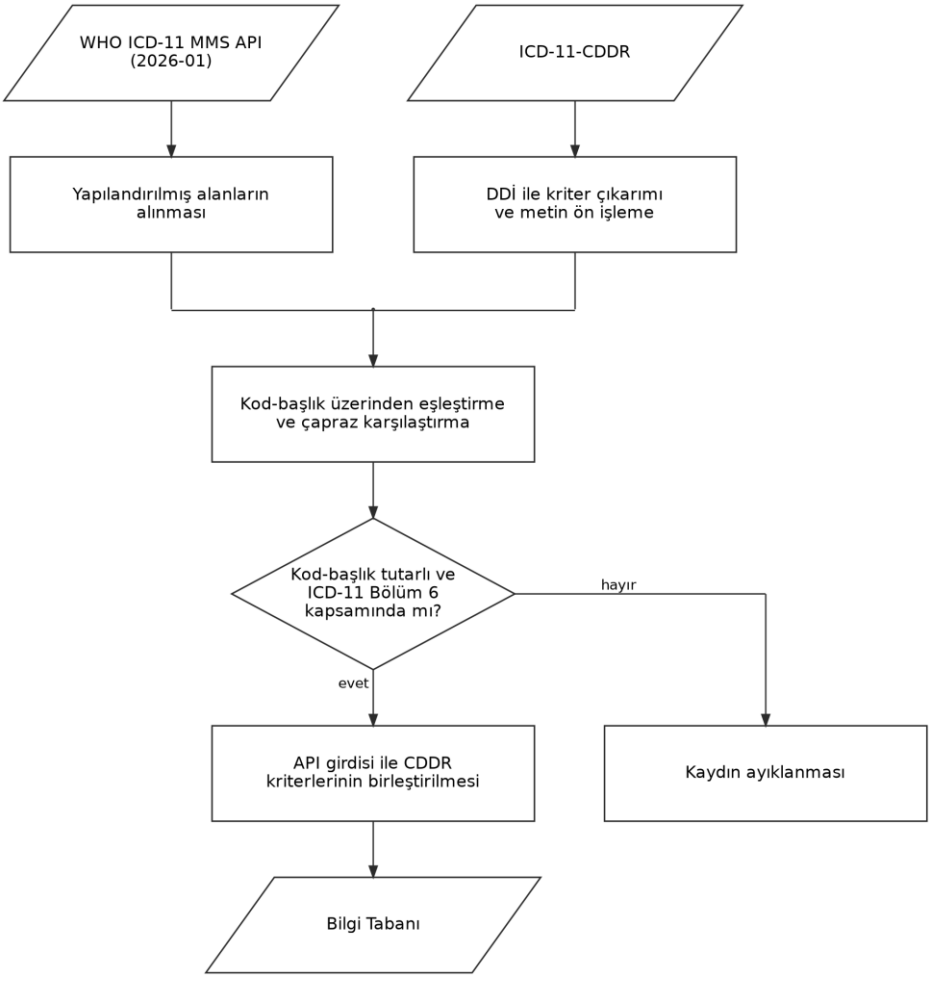
3.2. Bilgi Tabanı

Bilgi tabanı (BT), ICD-11 Bölüm 6'dan derlenen 720 ruhsal bozukluk girdisinden oluşmaktadır. Bu girdiler, WHO ICD-11 MMS (2026-01 sürümü) API'si ile ICD-11-CDDR klinik kriterlerinin birleştirilmesiyle oluşturulmuş ve ana deneyler bu bilgi tabanıyla yürütülmüştür. Değerlendirme alanının yine de DSM-5-TR Clinical Cases'ten seçilmesinin nedeni ise, bu kitaptaki tanıların ICD-11 karşılıklarına eşlenebilir olması ve böylece bilgi tabanı zemini (ICD-11) ile test zemini (DSM-5-TR vakaları) arasında tutarlı bir köprü kurulabilmesidir (Cremaschi vd. 2025).

WHO ICD-11 MMS API'si (2026-01 sürümü), her bozukluk için yapılandırılmış kodu, başlığı, eş adları, tanımı, dışlamaları ve sınıflandırma hiyerarşisini sağlar. ICD-

11-CDDR kitabı (852 sayfa) ise API'de bulunmayan temel özellikler, ayırıcı tanı, normalle ve diğer bozukluklarla sınırlar, gelişimsel ve kültürel görünümler gibi ayrıntılı tanı kriterlerini içerir. Bu kriterler kitaptan doğal dil işleme (DDİ) teknikleriyle çıkarılmıştır. PDF metni ayrıştırıldıktan sonra yinelenen üst ve alt bilgiler, konu başlıkları ve sayfa numaraları temizlenmiştir. Satır sonu tirelemesinden doğan bozuk karakterler ile onaltılık karakter kalıntıları giderilmiştir. Bölüm başlıkları (örnek olarak Essential (Required) Features, Boundary with Normality, Differential Diagnosis) kural ve desen tabanlı bölütlemeyle ilgili kriter alanlarına ayrıştırılıp normalleştirilmiştir.

ICD-11-CDDR'den DDİ ile çıkarılan kriterler, ICD-11 kodu ve başlığı üzerinden API'den gelen yapılandırılmış girdilerle eşleştirilmiştir. İki kaynak, kod başlık tutarlılığı bakımından çapraz karşılaştırmaya tabi tutulmuştur. Eşleşen girdilerde ICD-11-CDDR kriterleri ilgili API girdisiyle birleştirilmiş, uyuşmayan veya gürültülü çıkarımlar ise ayıklanmıştır. ICD-11'de Bölüm 6'nın dışında kodlanan girdiler (OOS) ile biçimsiz veya yinelenen kayıtlar alınmamıştır. Bu süreç sonunda, 720 adet temizlenmiş ve etiketlenmiş ICD-11 Bölüm 6 mental bozukluk girdisi oluşturulmuştur (Şekil 3.1). Bu girdilerin 583'ü CDDR türevli temel özellik kriterleri, 678'i tanım ve 376'sı eş ad alanı taşımaktadır.



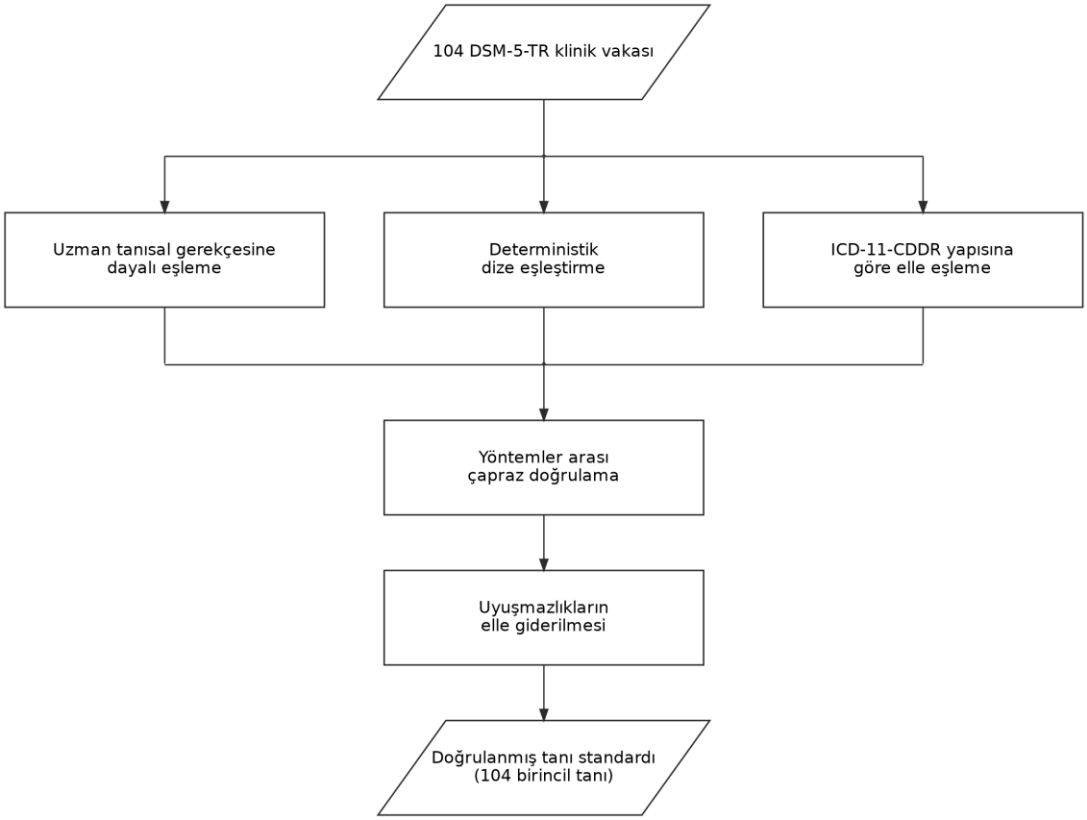
Şekil 3.1. BT'nin iki kaynaktan oluşturulma süreci

3.3. Tanı Standardı ve Doğrulaması

Tanı standardı, DSM-5-TR mental bozukluk tanımının ICD-11 eşlemesi biçiminde, üç bağımsız yöntemin çapraz doğrulanmasıyla oluşturulmuştur (Şekil 3.2).

1. Uzmanın tanısal gerekçesi tabanlı yöntem: Her vakanın klinik tanısal gerekçe alanı okunarak varılan tanı, doğrudan en uygun ICD-11 Bölüm 6 koduna eşlenmektedir.
2. Deterministik dize eşleştirme: DSM-5 tanı başlığı ile ICD-11 başlığı ve eş adı arasındaki belirteç benzerliğine dayanmaktadır.
3. Elle çözümleme: ICD-11-CDDR yapısına göre elle gözden geçirme ve düzeltme yapılmıştır.

Yöntemler arası tutarlılık, üç bağımsız yöntemin uyuşması ve uyuşmazlıkların WHO ICD-11-CDDR metni hakem alınarak elle çözülmesiyle sağlanmıştır. ISO/TR 12300 eşleme kalitesi ilkeleri gözetilmiştir. Uyumsuzluklar WHO ICD-11-CDDR'a göre çözülmüştür. Sürecin sonucunda 96 kapsam içi, 8 kapsam dışı birincil tanı ile 60 eş hastalıklı (komorbid) ek tanı elde edilmiştir. Tanı standardı, tahminler üretilmeden önce kesinleştirilip oluşturulmuştur.



Şekil 3.2. Tanı standardının oluşturulması ve doğrulanması

3.4. Erişim

Okapi BM25, ham anamnez ile dizinlenmiş BT belgeleri arasındaki olasılıksal sözcük örtüşmesini puanlamaktadır. Robertson ve Zaragoza (2009) tarafından biçimlendirilen olasılıksal ilgililik çerçevesinin bir uygulaması olan BM25, terim frekansına dayalı klasik ağırlıklandırmayı iki düzeltici düzenekle iyileştirir. İlk olarak, terim katkısının belge içi frekansla doğrusal değil azalan getirili biçimde ölçeklenmesini sağlayan terim frekansı doygunluğudur. Bu eğriliğin derecesini k_1 parametresi belirler ve sınır durumlarında k_1, 0 ikili ağırlıklandırmaya, büyük k_1 değerleri ise ham frekansa yakınsamaktadır. Diğer durumda ise belge uzunluğundan kaynaklanan yanlılığı dengeleyen belge uzunluğu normalleştirmesidir. b parametresi skorun belge uzunluğuyla ilişkilendirilme derecesini ayarlar ve b=0 ile b=1 arasında, normalleştirmenin kapalı olmasından tam uygulanmasına uzanan bir aralık tanımlar. Bağıntı 3.1'de terim ağırlığı, ters belge frekansı (IDF) ile belirlenmektedir (Robertson ve Zaragoza 2009).

$$\mathrm{IDF}(t) = \ln \frac{N - n_t + 0.5}{n_t + 0.5} \tag{3.1}$$

N bilgi tabanındaki toplam belge sayısı, n_t, t terimini içeren belge sayısı, f(t,d) terim frekansı, avgdl ortalama belge uzunluğu, |d| belge uzunluğu; k_1=1.5 ve b=0.75 sırasıyla Okapi doygunluk ve uzunluk normalleştirme parametreleridir.

$$\mathrm{BM25}(d,q) = \sum_{t \in q} \mathrm{IDF}(t) \cdot \frac{f(t,d)(k_1+1)}{f(t,d) + k_1(1 - b + b|d|/\mathrm{avgdl})} \tag{3.2}$$

Bu çalışmada söz konusu parametreler, bilgi erişimi yazınında uzlaşılmış varsayılan değerlerinde (k_1=1.5, b=0.75) tutulmuştur. Bu tercih, küçük ölçekli bir bilgi tabanında göreve özgü aşırı uyumu önleyerek erişim tabanının tekrarlanabilirliğini ve dış geçerliğini güvence altına almayı amaçlamaktadır.

BM25, hibrit erişim mimarisinin sözcüksel bileşenidir. Dizinleme, BT kayıtlarının başlık, eş ad ve tanım alanları üzerinden yapılmaktadır. Metin yoğunluğu nedeniyle sözcüksel sinyali seyreltme eğilimindeki kriter alanları kapsam dışında bırakılır. Eğitim gerektirmeyen, yorumlanabilir ve düşük maliyetli bir sıralayıcı olarak BM25, özellikle nadir ve yüksek ayırt edicilikteki klinik belirteçlerin birebir örtüşmesini yakalamada etkilidir. Ancak vaka anlatısının gündelik diliyle ICD-11 kategorilerinin biçimsel terminolojisi arasındaki sözcük boşluğuna tek başına bağlantı oluşturamaması nedeniyle sınırlı bir erişim yapısına ulaştırmaktadır. Bu nedenle BM25, yoğun erişim ve HyDE sinyalleriyle RRF altında birleştirilerek tamamlayıcı bir bileşen olarak konumlandırılmıştır; HyDE ile birlikte kullanıldığında kategori düzeyi recall'a yaklaşık +0.03 katkı sağladığı gözlenmiştir. Kod çözümleme tarafında ise, daha kesin eşleştirme stratejilerinin zayıfladığı durumlarda devreye giren bir yedek eşleştirici olarak görev yapar.

Yoğun erişim (dense retrieval), sözcüksel örtüşmeye dayalı sıralayıcılardan farklı olarak sorgu ve belgeyi öğrenilmiş bir anlamsal uzaya gömerek yüzeysel sözcük biçimlerinden bağımsız bir eş anlamlılık ve örtük ilişki yakalaması sağlamaktadır

(Karpukhin vd. 2020). Bu çalışmada, bge-large-en-v1.5 ikili kodlayıcısı kullanılmıştır. İkili kodlayıcı mimarisinde sorgu ve belge birbirinden bağımsız olarak kodlandığından, bilgi tabanındaki tüm belge gömmeleri çevrimdışı önceden hesaplanıp dizinlenebilmektedir. Çıkarım anında yalnızca sorgunun kodlanması ve vektör karşılaştırması gerektiğinden erişim, CPU üzerinde dahi ölçeklenebilir kalmaktadır. Metinler 1024 boyutlu, L2-normalleştirilmiş vektörlere kodlanmıştır. Sorgu ve belge yakınlığı kosinüs benzerliğiyle ölçülmektedir (Xiao vd. 2024).

Sorgu q ile belge d, bge-large ikili kodlayıcısıyla 1024 boyutlu gömme vektörlerine eşleştirilmiştir (Bağıntı 3.3).

$$\mathrm{sim}(q,d) = \frac{\boldsymbol{e}_q \cdot \boldsymbol{e}_d}{\|\boldsymbol{e}_q\|\|\boldsymbol{e}_d\|} \tag{3.3}$$

Yoğun erişimin amacı, sorgu ile hedef belgenin aynı anlamsal uzayda yakın konumlanmasıdır. Ancak serbest metinle yazılmış bir hasta anamnezi ile biçimsel ICD-11 kriter metni, hem sözcük dağarcığı hem de söylem türü bakımından gözle görülür biçimde ayrışmaktadır. Bu asimetri, ham anamnezin doğrudan gömülmesini gürültülü bir sorguya dönüştürür. HyDE bu sorunu, sorgunun kendisini değil, büyük dil modeline ürettirilen varsayımsal bir belgeyi gömerek çözmektedir. Model, anamnezden yola çıkarak olası tanıyı betimleyen özlü ve biçimsel bir klinik özet üretmektedir. Bu özet, hedef kriter metinlerinin diline ve yapısına çok daha yakın olduğundan gömme uzayında doğru kategorilere net biçimde yakınsamaktadır. Buradaki amaç, üretilen özetin doğruluğunu güvence altına almak değil, sorguyu belge dağılımının yoğunlaştığı bölgeye taşıyan bir temsil bağı kurmaktır. Çalışmada HyDE çıktısı yalnızca yoğun erişim koluna beslenmektedir. BM25 ise ham anamnez üzerinde çalışmayı sürdürür. Böylece sözcüksel kesinlik ile anlamsal genelleme bir arada korunur (Gao vd. 2023).

HyDE, üretim aşamasındaki Gemma modeline yalnızca hasta anamnezi verilir ve ondan, bir psikiyatrist bakışıyla, vakanın temel belirti ve bulgularını, başlangıç ve süresini, seyrini ve en olası iki veya üç psikiyatrik bozukluk tipini içeren ad ve biyografik ayrıntıdan arındırılmış, üç veya beş cümlelik özlü bir klinik özet üretmesi istenmektedir. Bağıntı 3.4'te g üretici BDM'i (Gemma 4), $\hat{v}_{\mathrm{HyDE}}$ HyDE sorgu vektörünü, $\mathrm{INST}_{\mathrm{HyDE}}$ HyDE istemini göstermektedir. Bağıntı 3.5'te model bu varsayımsal belgeyi (h) ürettikten sonra, ham anamnezin gömmesi e(q) yerine özetin gömmesi e(h) kullanılmaktadır ve yoğun erişim skoru, BT belgeleriyle e(h) arasındaki kosinüs benzerliğiyle, yani kosinüs(e(h), e(d)) ile hesaplanmaktadır. h, kaynak anamnezden değil hedef ICD-11 kriter metinlerinin biçimsel dilinden türediği için gömme uzayında doğru tanı gruplarına ham anamnezden çok daha yakın düşmektedir. Elde edilen bu yoğun HyDE sırası daha sonra, ham anamnez üzerindeki yoğun ve BM25 sıralarıyla RRF altında birleştirilmektedir. HyDE, erişimin yalnızca yoğun anlamsal kolunu zenginleştirmektedir, sözcüksel kolu (BM25) olduğu gibi bırakmaktadır.

$$h = g(q, \mathrm{INST}_{\mathrm{HyDE}}) \quad \hat{v}_{\mathrm{HyDE}} = f(h) \equiv e(h) \tag{3.4}$$

$$s_{\mathrm{HyDE}}(d) = \langle e(h), e(d) \rangle = \cos(e(h), e(d)) \tag{3.5}$$

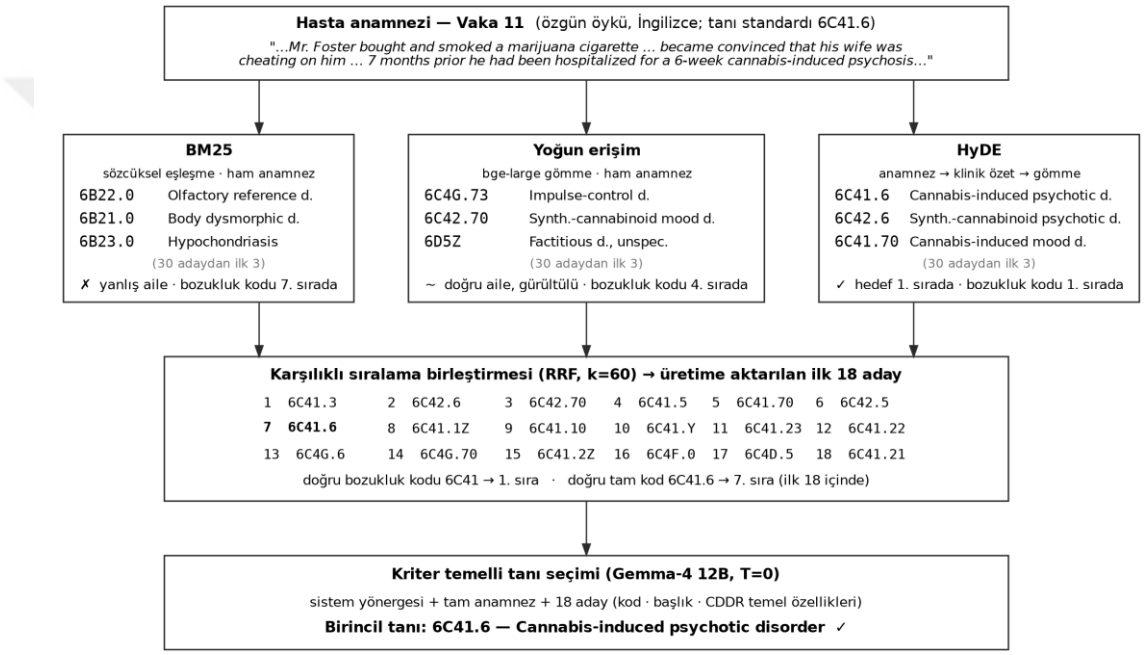
HyDE özetini, üretim aşamasındakiyle aynı Gemma ailesi modeli üretmektedir (en çok 256 belirteç, sıcaklık T=0). Gao vd. (2023)'in özgün önerisinde model birden çok varsayımsal belge örneklemektedir. Bu belgelerin gömmeleri ortalanarak bu çalışmada ise tekrarlanabilirliği güvence altına almak ve vaka başına maliyeti sınırlamak amacıyla sıcaklık sıfıra çekilerek tek ve deterministik bir özet üretilmiştir. Raporlanan erişim recall'ı saf bir erişim ölçütü değildir. Oluşturulan özet, modelin parametrik tanı tahminini de içerdiğinden recall kısmen bu tahmini yansıtmaktadır. HyDE vaka başına ek bir BDM çağrısı getirmektedir. Bu çalışmada özetler çevrimdışı bir kez üretilip tüm model boyutlarında ortak kullanılmıştır (Gao vd. 2023).

Örnek olarak eş tanılı bipolar bozukluk öyküsü bulunan, yoğun kannabis kullanımının ardından akut psikotik belirtiler (kötücül sanrılar, referans düşünceleri) geliştiren ve bu nedenle kolayca karışabilen bir vaka ele alınsın (11. vaka; tanı standardı 6C41.6, kannabis kaynaklı psikotik bozukluk). Hastanın özgün anamnezine Barnhill (2023, Case 11) üzerinden erişilebilir. Üretim aşamasındaki Gemma 4 modeli, örnek olarak verilen anamnezden şu HyDE özetini üretmiştir: “Patient presents with acute onset of persecutory delusions, ideas of reference, and suicidal/homicidal ideation following heavy cannabis use. The course is characterized by a one-year history of recurrent psychiatric hospitalizations for mania, depression, and cannabis-induced psychosis, with symptoms partially resolving upon stabilization. Most likely diagnoses include Bipolar I Disorder, Cannabis Use Disorder, and Cannabis-Induced Psychotic Disorder.”

Özet, dağınık öyküyü ICD-11 kriter uzayının diline çevirmektedir. Adı ve biyografik ayrıntıyı düşürerek vakayı belirti, seyir ve olası tanı eksenlerine indirgemekte ve kannabis kaynaklı psikozu hem seyir hem de olası tanılar bölümünde açıkça çıkarmaktadır.

Dört erişim kolunun aynı vaka üzerindeki davranışı, yöntemlerin neden birbirini tamamladığını somut biçimde Şekil 3.3’te gösterilmiştir. Sözcüksel BM25, anamnezdeki “ideas of reference” ve “insight” ifadelerinin yüzeysel örtüşmesine kapılarak obsesif-kompulsif alanına sapmaktadır (6B22.0 Olfactory reference disorder, 6B21.0 Body dysmorphic disorder, 6B23.0 Hypochondriasis). Doğru kategori kodu ancak yedinci sıraya düşer, çünkü nadir ama ayırt edici “cannabis” teriminin yüzeysel eşleşmesi tek başına yeterli olmaz. Ham anamnez üzerindeki yoğun erişim doğru anlamsal bölgeye (madde kaynaklı bozukluklar) girer, ancak yapay bozukluk (6D5Z Factitious disorders, unspecified) ve madde kaynaklı dürtü denetimi bozukluğu (6C4G.73 Impulse control disorder induced by a psychoactive substance) gibi gürültülü adaylarla birlikte gelir ve doğru kategori kodunu ancak dördüncü sıraya koymaktadır. HyDE özeti, tabloyu kriter diline çevirip kannabis kaynaklı psikozu açıkça andığından

doğru tanıyı (6C41.6 Cannabis-induced psychotic disorder) doğrudan birinci sıraya taşımaktadır. Son olarak RRF, üç kolun sıralarını birleştirerek tutarlı ve arınmış bir aday kümesi üretmektedir. İlk yedi adayın tamamı kannabis/kannabinoid tanı grubundan gelir (6C41.3, 6C42.6, 6C42.70, 6C41.5, 6C41.70, 6C42.5, 6C41.6) ve doğru kategori kodu birinci sıraya yerleşir. Burada, HyDE tek başına doğru alt kategori kodunu (6C41.6) birinci sıraya koyarken, BM25 ve yoğun erişim kollarının da katkısıyla oluşan birleşik sıralamada 6C41.6 yedinci sıraya gerilemektedir. Çünkü RRF'nin asıl işlevi doğru tanı grubunu (kannabis kaynaklı bozukluklar) gürültüden arındırarak üst sıralara taşımak ve üretime tutarlı bir aday menüsü sunmaktır. Nitekim doğru alt kategori, üretime aktarılanilk on sekiz aday içinde yer alır ve son seçimi, adayları klinik kriterleriyle karşılaştıran üretim modeli yapmaktadır. Böylece üretim modeline, hiçbir tek yöntemin tek başına sağlayamayacağı kadar odaklı bir aday menüsü sunulmaktadır.



Şekil 3.3. Üç sıralayıcı (BM25, yoğun erişim, HyDE) ile RRF birleştirmesinin 11. Vaka üzerindeki davranışı

Sözcüksel ve yoğun erişim skorları farklı ölçeklerde ve farklı dağılımlarda üretildiğinden, doğrudan toplanmaları normalleştirme gerektirmektedir. RRF bu sorunu, mutlak skorları tümüyle göz ardı edip yalnızca her sıralayıcının ürettiği sıra bilgisini kullanarak geçmektedir (Cormack vd. 2009).

Bağıntı 3.6'da R birleştirilen sıralayıcılar kümesi, $\text{rank}_r(d)$ ise d belgesinin r sıralayıcısındaki konumudur. Toplama biçimi gereği, birden çok kolda tutarlı biçimde üst sıralarda yer alan belgeler ödüllendirilirken, k yumuşatma sabiti tek bir koldaki çok yüksek sıraların aşırı baskınlığını dengeler ve düşük sıralı belgelerin katkısını azaltır.

$$\mathrm{RRF}(d) = \sum_{r \in R} \frac{1}{k + \mathrm{rank}_r(d)}, \qquad k = 60 \tag{3.6}$$

Yukarıdaki vakada doğru kannabis kategori kodu HyDE'de birinci, ham anamnez yoğun erişiminde dördüncü ve BM25'te yedinci sırada yer aldığından üç kolda da makul üst sıralarda bulunmaktadır. RRF bu uzlaşıyı ödüllendirerek onu birleşik sıralamada birinci sıraya taşırken, yalnızca tek bir kolda yüksek görünen yanıltıcı adayları geri plana itmektedir. Skordan bağımsız ve parametreye karşı baskın olması nedeniyle RRF, ayar gerektirmeyen ve tekrarlanabilir bir birleştirme yöntemi sunmaktadır. k=60 değeri çalışmadaki yerleşik varsayılan olarak yer almaktadır. En yüksek RRF skoruna sahip en iyi 18 aday, ilgili kriterleriyle birlikte üretim aşamasına aktarılmaktadır; bu eşik, erişim duyarlılığı (recall) ile üretim modelinin bağlam penceresi kısıtı arasında bir denge gözetilerek belirlenmiştir.

Erişim aşaması başarısının üretim aşamasından bağımsız ölçülmesi, hata çözümlemesi açısından yöntemsel olarak gereklidir. Çünkü mental bozukluk tanı hatalarının erişim mi yoksa üretim aşamasından mı kaynaklandığı ayrıştırılamaz. Bu amaçla, üretimden bağımsız bir iç ölçüt olarak Recall@k kullanılmaktadır. Metrik, kapsam içi vakalarda doğru ICD-11 kategori kodunun ilk k aday arasında bulunma oranını verir (Karpukhin vd. 2020).

$$\mathrm{Recall@k} = \frac{1}{\mathrm{N_{in}}} \sum_{\mathrm{i=1}}^{\mathrm{N_{in}}} [\mathrm{gold_i} \in \mathrm{TopK_i}] \tag{3.7}$$

Burada N_{in}=96 kapsam içi vaka sayısı, $gold_i$ i vakasının tanı standardındaki kategori kodu, $TopK_i$ ise i vakası için erişilen ilk k adaydır. Değerlendirmenin kategori kodu düzeyinde yapılması, erişimin görevini gerçekçi biçimde yansıtmaktadır. Erişimden beklenen, alt kategori kodunu birebir bulması yerine doğru tanı grubunu üretim modelinin değerlendirebileceği aday kümesine taşımasıdır. Alt kategori ayrımı ise sonraki üretim aşamasının sorumluluğundadır. K=18 seçimi, üretime aktarılan aday sayısı ile orantılıdır.

3.5. Üretim

Gemma, Google tarafından Gemini modelleriyle aynı araştırma ve teknoloji temeli üzerine geliştirilen, açık ağırlıklı ve göreli olarak hafif bir büyük dil modeli ailesidir (Gemma Team 2024). Ağırlıkları kamuya açık olduğundan model yerelde indirilip çalıştırılabilmekte ve talimatla ayarlanmış sürümler halinde sunulmaktadır. Açık ağırlıklı olması, sistemin ticari bir API'ye ya da kapalı bir modele bağımlı olmadan yerel donanımda çalışmasına ve klinik verinin yerelde işlenmesine olanak tanır. Gemma ailesi, çok küçük (E2B) modellerden orta ölçekli (12B) modellere ve verimli bir uzman karması (MoE) varyantına (26B-A4B) uzanan bir boyut yelpazesi sunarak model ölçekleme ve verimlilik çözümlemesine elverişlidir. Bu boyut yelpazesi, hesaplama verimliliği öncelikli olduğunda daha az etkin parametreli modellerin değerlendirilmesine olanak tanımaktadır.

Mimari olarak Gemma 4, yalnızca kod çözücülü bir dönüştürücüdür (Gemma Team 2026). Normalizasyon, dikkat ve ileri besleme bloklarının hem girişine hem

çıkışına uygulanan RMSNorm (Bağıntı 2.6) ile sorgu ve anahtar normalizasyonundan (QK-norm) oluşmaktadır. Dikkat deseni, uzun bağlamda anahtar değer önbelleğinin büyümesini sınırlamak için yerel (kayan pencereli) ve küresel dikkat katmanlarını dönüşümlü kullanmaktadır. Bu oran 12B ve 26B-A4B modellerinde 5:1'dir. Yerel katmanlarda taban frekansı 10^4 olan standart RoPE, küresel katmanlarda ise 128k belirteçlik bağlam penceresini destekleyen 10^6 tabanlı bir RoPE modeli uygulanmaktadır (Bağıntı 2.7). Sözcük dağarcığı 262k girdilik bir SentencePiece belirteçleyicidir. E2B ve E4B modelleri, katman başına gömmeler (per-layer embeddings) sayesinde sırasıyla 5B ve 8B toplam parametreyle 2.3B ve 4.5B etkin parametrelik bir çekirdek çalıştırmaktadır. İsmindeki 'E' öneki bu etkin (effective) parametre sayısını belirtmektedir. 26B-A4B ise ileri besleme katmanları uzman ağlarına bölünmüş seyrek bir uzman karması modelidir (Bağıntı 2.9). 26B toplam parametrenin belirteç başına yalnızca yaklaşık 4B'si etkinleştirilmektedir. Bu tezde gözlenen, 12B'ye yakın doğruluğun yaklaşık üçte bir etkin parametreyle elde edilmesi davranışının mimari temeli budur. Kapasite toplam parametrede, çıkarım maliyeti ise aktif parametrede ölçeklenmektedir.

Bu çalışmada modeller, llama.cpp altyapısında UD-Q4_K_XL nicemlemeli GGUF biçiminde çalıştırılmıştır. Bu şemada ağırlıklar küçük bloklar halinde, blok başına bir ölçek katsayısıyla yaklaşık 4 bite indirgenmekte ($w \approx s \cdot q$), çıktı kalitesine duyarlılığı yüksek katmanlar ise daha yüksek bit genişliğinde tutulmaktadır. Böylece 12B modelin bellek izi yaklaşık dörtte bire inmekte ve E2B'den 26B-A4B'ye uzanan tüm boyut yelpazesi, standart donanım üzerinde aynı deterministik ayarlarla (T=0) çalıştırılabilmektedir.

Üretim aşaması, llama.cpp altyapısıyla sunulan, UD-Q4_K_XL nicelemeli GGUF biçimindeki açık ağırlıklı Gemma 4 ailesi modelleriyle gerçekleştirilmiştir. E2B, E4B, 12B ve 26B-A4B. En yüksek doğruluğa Gemma 4 12B ile ulaşılmış ve önerilen sistemin üretim modeli olarak bu model benimsenmiştir. 26B-A4B modeli ise buna denk, daha verimli bir alternatif olmuştur. Modele en iyi 18 aday, her birinin temel özellikleriyle birlikte verilir; model hastanın klinik tablosunu bu kriterlerle eşleştirerek tek bir birincil tanı (ad, ICD-11 kodu, gerekçe ve ayırıcı tanı) üretir. Sıcaklık T=0 seçilerek üretim deterministik kılınmıştır.

Üretim istemi modelden katı biçimde yapılandırılmış bir çıktı istemektedir. Görevi ve kısıtları tanımlayan bir sistem yönergesi olarak yalnızca verilen klinik öyküye dayan, tek bir birincil tanı seç ve gerekçeyi adayların temel özellikleriyle ilişkilendir şeklindedir. Hasta anamnezi ve erişimle getirilen en iyi 18 aday ise her biri ICD-11 kodu, başlığı ve CDDR temel özellikleriyle birlikte getirilmektedir. Model çıktıyı bir json nesnesi olarak döndürmektedir. Birincil tanının adı, ICD-11 kodu, kısa bir gerekçe ve birkaç ayırıcı tanı şeklinde yer almaktadır. Adaylar bağlayıcı bir çoktan seçmeli liste olarak değil destekleyici bağlam olarak sunulmaktadır. Böylece hiçbir aday uymadığında model geçerli başka bir ICD-11 tanısı da önerebilmektedir. Bu kurgu, CDDR'ın prototip eşleme mantığına benzerdir. Model katı bir belirti sayımı yapmak yerine klinik tabloyu temel özelliklerle bütüncül biçimde karşılaştırmaktadır. Sıcaklığın sıfır (T=0) ve çıktının json nesne kipinde seçilmesi, üretimin hem deterministik hem de güvenilir biçimde ayrıştırılabilir olmasını sağlamaktadır.

Modelin ürettiği serbest metin tanı adı ve tanı kodu, çok düzeyli değerlendirmeden önce geçerli bir ICD-11 koduna çözümlenmektedir. Model bir tanı kodu ürettiyse ve bu kod bilgi tabanında geçerliyse doğrudan kullanılmaktadır. Aksi hâlde tanı adı, BT başlıkları ve eş adlarıyla belirteç örtüşmesine göre eşleştirilmektedir. Bunlar da başarısız olursa ham ad üzerinden BM25 ile en yakın geçerli başlık yedek olarak seçilmektedir. Böylece her tahmin, dört değerlendirme düzeyinde tutarlı biçimde puanlanabilen geçerli bir ICD-11 koduna indirgenmektedir. Var olmayan ya da biçimsiz bir kodun üretilmesi (halüsinasyon) bu aşamada elenmesi hedeflenmiştir.

3.6. Erişim Konfigürasyonları ve Tasarımı

Bileşenlerin tanısal başarıya katkısını ölçmek ve hangi kurulumun en yüksek doğruluğu verdiğini belirlemek amacıyla, kontrollü değişkenlerle dokuz konfigürasyon (K1–K9) tasarlanmıştır. Tasarımın temel ilkesi, önerilen K4 konfigürasyonunu referans almak ve diğer konfigürasyonları onun kontrollü varyasyonları olarak kurmaktadır. Tüm konfigürasyonlar aynı bilgi tabanını, sıfır sıcaklığı (T=0) ve aynı değerlendirme protokolünü paylaşmaktadır. Yalnızca incelenen bileşen değiştirilmektedir. Bu bileşen ablasyonu, etkileri yalıtmak için üretim modelini Gemma 4 26B-A4B'de sabit tutmaktadır. Önerilen üretim modeli ise model ölçekleme çözümlemesinde en yüksek tanı adı doğruluğunu veren Gemma 4 12B'dir (Çizelge 4.6). Ablasyonun 26B-A4B üzerinde yürütülmesinin nedeni, dokuz konfigürasyonun tamamının (K5 dahil) bu modelde koşulmuş olması ve bileşen katkılarının burada en temiz örüntüyü (K1'den K4'e doğru tutarlı artış) vermesidir. Daha küçük modellerde, çok aşamalı K2 ile tek aşamalı K3 arasındaki ilişki gibi bazı sıralamalar kararsızlaştığından bileşen yalıtımı daha az nettir. Önerilen 12B modelinin dokuz konfigürasyon üzerindeki başarımı ise Çizelge 4.6 ve Şekil 4.6'daki matriste ayrıca sunulmaktadır. Bir konfigürasyon birden çok bileşende K4'ten ayrıldığında (örneğin hem erişim hem üretim) bu, metinde açıkça belirtilir. Konfigürasyonlar, erişim stratejisi (sözcüksel, hibrit, HyDE birleştirmesi, HyDE olmadan) ve üretim stratejisi (tek aşamalı kriter temelli, çok aşamalı, karşıtlamalı, zenginleştirilmiş, yeniden sıralamalı) olarak iki gruba ayrılmıştır. Çizelge 3.2 bu konfigürasyonları ve tasarım eksenlerini listeler ve önerilen K4'ten ayrılan bileşenleri gösterir. Kalın hücreler, ilgili konfigürasyonu önerilen K4'ten ayıran bileşeni göstermektedir.

Çizelge 3.2. Erişim konfigürasyonları (K1–K9)

Konfigürasyon	Ayırt edici özellik	Sorgu girdisi	Erişim ve birleştirme	Üretim
K1	Sözcüksel taban	Ham anamnez	**Yalnızca BM25**	**Tek aşamalı üretim**
K2	Hibrit erişim, çok aşamalı üretim	Ham anamnez	Yoğun + BM25 (RRF)	**Çok aşamalı üretim (öneri + kriter doğrulama)**

(Devamı arkada)

Çizelge 3.2.'nin devamı.

K3	Hibrit erişim, kriter temelli üretim	Ham anamnez	Yoğun + BM25 (RRF)	Kriter temelli tek aşamalı üretim
K4	**Önerilen: HyDE birleştirmeli erişim**	**Anamnez + HyDE özeti**	**Yoğun + BM25 + HyDE-yoğun (RRF)**	Kriter temelli tek aşamalı üretim
K5	K4 + kriter parçalama	Anamnez + HyDE özeti	**Yoğun + BM25 + HyDE-yoğun, kriter parçalı dizin**	Kriter temelli tek aşamalı üretim
K6	HyDE yerine çapraz kodlayıcı	**Klinik anahtar sorgusu (HyDE yok)**	**Yoğun + BM25 + çapraz kodlayıcı sıralama**	Kriter temelli tek aşamalı üretim
K7	HyDE yerine ayırıcı tanı sınırları	Ham anamnez (HyDE yok)	Yoğun + BM25 (RRF)	**Ayırıcı tanı sınır metinleriyle üretim**
K8	Yalnızca yoğun erişim	Ham anamnez (HyDE yok)	**Yalnızca yoğun erişim (BM25 yok)**	Ayırıcı tanı sınır metinleriyle üretim
K9	K4 + ayırıcı tanı ve belirleyici yönergesi	Anamnez + HyDE özeti	Yoğun + BM25 + HyDE-yoğun (RRF)	**Ayırıcı tanı sınırları + belirleyici çözümü yönergesi**

3.6.1. Taban ve Hibrit Erişim (K1–K3)

K1, yalnızca sözcüksel erişimin tek başına ne ölçüde yeterli olduğunu ölçen bir alt sınır kontrolüdür ve yoğun erişim ile HyDE bileşenlerinin gerekliliğini ortaya koymayı amaçlamıştır. Erişim tarafında ham anamnez doğrudan Okapi BM25 ile sıralanmaktadır. Üretim tarafında en üst adaylar, kriter ya da temel özellik metni olmaksızın yalnızca başlık ve kısa tanımlarıyla, tek bir BDM çağrısında modele verilmektedir ve model bir birincil tanıyı üretmektedir. Kriter temelli üretimin (K3/K4) eklediği değeri görünür kılan bir taban sağlamaktadır. Klinik anlatının gündelik diliyle ICD-11 kategorilerinin biçimsel terminolojisi arasındaki sözcük boşluğunu kapatamaz ve modeli yanlış adaylara eşleştirebilir.

K2, hibrit erişimin (yoğun + BM25) ve çok aşamalı, kriter doğrulamalı bir üretim zincirinin katkısını birlikte ölçmektedir. Erişim tarafında HyDE olmadan ham anamnez üzerinden yoğun (bge-large) ve BM25 adayları RRF ile birleştirilmektedir. Üretim çok aşamalıdır ve iki BDM çağrısı içermektedir. Önce model, aday başlıklarından bir birincil tanı ve kısa bir ayırıcı tanı önerir. Ardından önerilen ve komşu adayların ICD-11 temel özellikleri toplanıp modele geri verilmektedir ve model,

klinik tabloyu bu kriterlerle doğrulayarak nihai tanıyı seçmektedir. K2, K4'ten farklı olarak HyDE bulunmaz ve üretim çok aşamalıdır.

K3, K2'nin çok aşamalı kurgusunu tek aşamalı, kriter temelli bir üretime indirgemektedir. Erişim tarafında hibrit (yoğun + BM25, HyDE olmadan) adaylar getirilmiştir. Üretim tarafında bu adaylar, her biri ICD-11 temel birlikte tek bir istemde modele verilmiştir. Model bu adayları bağlayıcı bir menü olarak değil destek olarak görmektedir ve gerektiğinde listede yer almayan başka bir geçerli ICD-11 tanısını da seçerek kararını tek aşamada vermektedir. K3, K4'ten yalnızca HyDE bileşeni bakımından ayrılmaktadır (erişim hibrittir, üretim K4 ile özdeştir). Dolayısıyla K3'ten K4 geçişi, aynı model üzerinde HyDE'nin net katkısını ayırmaktadır. Bu katkının büyüklüğü modele bağlıdır.

3.6.2. Önerilen Konfigürasyon (K4)

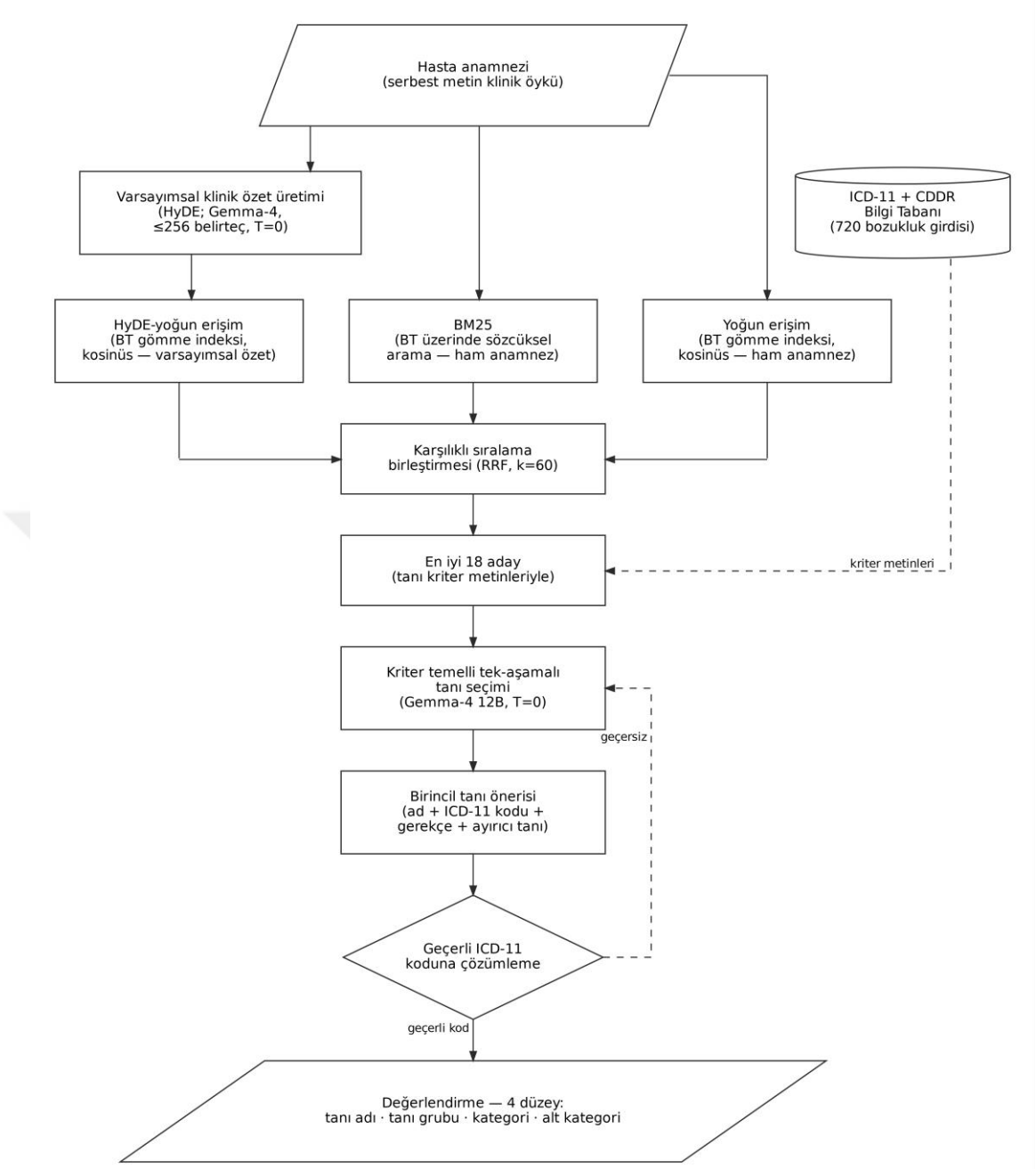
K3 üzerine HyDE eklenerek önerilen K4 konfigürasyonu elde edilmektedir. Konfigürasyon işleyişinde aynı Gemma 4 modeli, anamnezden en çok 256 belirteçlik özlü bir klinik özet (HyDE) üretmektedir. Bu özet bge-large ile gömülüp BT belge gömmeleriyle kosinüs benzerliğine göre sıralanmaktadır. Bu yoğun-HyDE sırası, ham anamnez üzerindeki yoğun ve BM25 sıralarıyla RRF (k=60) altında birleştirilmektedir ve en iyi 18 aday, temel özellikleriyle birlikte tek aşamalı kriter temelli üretime aktarılmaktadır. Bu konfigürasyon, dört erişim bileşenini kriter temelli üretimle tamamlayıcı biçimde birleştirmektedir.

Hasta anamnezi gündelik ve öyküsel bir dille yazılırken, ICD-11 kriter metinleri biçimsel ve teknik bir dille kaleme alınmıştır. Sözcüksel ve anlamsal boşluklar, ham anamnezin doğrudan gömülmesini zayıf bir erişim sinyaline dönüştürür. HyDE, BDM'e önce özlü bir klinik özet ürettirip bu özeti gömerek sorguyu belge uzayına taşımaktadır ve anlamsal boşluğu kapatmaktadır.

Yoğun erişim, gömme uzayındaki kosinüs yakınlığıyla anlamsal eşleşmeyi yakalarken, BM25 sözcüksel birebir terim örtüşmesini tespit etmektedir. HyDE bulunmadığında ise ham anlatı üzerinde gürültülü davranarak başarımı düşürmektedir. BM25'in yararı, HyDE'nin sağladığı arınmış sinyale koşulludur.

Getirilen en iyi 18 seçim, her birinin temel özellikleriyle birlikte modele verilir ve model, hastanın klinik tablosunu adayların ölçütleriyle karşılaştırarak ICD-11'in prototip eşleme mantığına benzer bir akıl yürütme gerçekleştirir.

Önerilen konfigürasyonun parametreleri Çizelge 3.3'te toplu olarak özetlenmiştir. Sistemin uçtan uca işleyişi Şekil 3.4'te akış diyagramı olarak verilmiştir.



Şekil 3.4. Önerilen sistemin (K4 + Gemma 4 12B) uçtan uca algoritma akış diyagramı

Çizelge 3.3. K4 konfigürasyonunun parametreleri

Bileşen	Hiperparametre	Değer
Yoğun erişim	Gömme modeli	bge-large-en-v1.5 (1024-boyut)

(Devamı arkada)

Çizelge 3.3.'ün devamı.

Sözcüksel erişim	BM25 Okapi *(k1, b)*	(1.5, 0.75)
BT belge alanı	Başlık + eş ad + tanım	ilk 600 karakter
Birleştirme	RRF sabiti *k*	60
Aday havuzu	En iyi *k* aday	18
Yeniden sıralama (K6)	Çapraz kodlayıcı	bge-reranker-v2-m3
Üretim	Dil modeli	Gemma 4 12B (önerilen); 26B-A4B (MoE) verimli alternatif
Üretim	Niceleme (GGUF)	UD-Q4_K_XL
Üretim	Sıcaklık *T*	0 (deterministik)
HyDE özeti	En çok belirteç (token)	256
Üretim	Çıktı kipi	JSON

3.6.3. Zenginleştirme Konfigürasyonları (K5, K9)

K5, BT "criteria" alanındaki metinlerin parçalanmasının daha ince taneli bir erişim sağlayıp K4'ü iyileştirip iyileştirmediğini test etmektedir. K4'ten farklı olarak her BT girdisinin uzun kriter metni cümle parça düzeyinde ayrı vektörlere bölünür ve sorgu belge benzerliği parçalar üzerinden en yüksek benzerlik ile hesaplanıp çok vektörlü temsil RRF birleştirmesine beslenmektedir. Üretim K4 ile aynıdır.

K9, K4 üzerine birbirine karışabilen aday çiftleri için ICD-11 ayırıcı tanı sınır metinleri ve şiddet, epizod gibi belirleyicilerin çözümüne yönelik açık bir yönerge ile fark yaratmıştır. K4'ten tek farkı istemin içeriğidir ve erişim aynıdır. Amaç, sistemin en çok zorlandığı komşu bozukluk ayrımı ve belirleyici düzeyindeki hataları azaltmaktır.

3.6.4. HyDE Olmadan Alternatifler (K6–K8)

Son üç konfigürasyonun ortak amacı, HyDE'nin gerekliliğini doğrudan sınamaktır. Her birinde HyDE devre dışı bırakılmış ve yerine erişim kalitesini artırması beklenen farklı bir mekanizma yerleştirilmiştir. Böylece HyDE'nin başka tekniklerle ikame edilip edilemeyeceği test edilmektedir.

K6, HyDE'nin yerine çapraz kodlayıcı yeniden sıralama koyarak erişim kesinliğini artırmayı denemektedir. HyDE-olmadan deterministik bir klinik anahtar sorgusuyla elde edilen yoğun ve BM25 adayları, sorgu belge çiftini birlikte kodlayan bir çapraz kodlayıcıyla (bge-reranker-v2-m3) yeniden sıralanmaktadır. Üretim K4 ile aynı olmaktadır. Yeniden sıralama yalnızca getirilmiş adaylar arasında düzen kazandırmaktadır, ama doğru tanı grubu hiç getirilmemişse bunu kurtaramamaktadır.

K7, HyDE kullanmadan, bir ayırıcı tanı grafiği üzerinden komşu veya karışabilen bozuklukları karşılaştırarak ayırt etmeyi denemektedir. Erişim HyDE-olmadan ham anamnez üzerinde yoğun ve BM25 yapısıyladır. Üretim tarafında, getirilen adaylar arasından birlikte bulunan karışabilir çiftler için ICD-11 sınır metinleri isteme eklenmektedir. Böylece model ayrımı bu sınırlara dayanarak yapmaktadır. Güçlü bir erişim sinyali (HyDE) yokken ayırıcı tanı tespitinde doğru adayları getiremediğinden tek başına yetersiz kalmaktadır.

K8, hem HyDE'nin hem de BM25'in dışlandığı, yalnızca yoğun erişimin karşıtlamalı bir üretimle birleştirildiği en kısıtlı erişim kurgusudur. Adaylar yalnızca ham anamnezin yoğun gömmesinden gelmektedir ve sözcüksel sinyal yoktur. Üretimde K7'deki gibi ayırıcı tanı sınır metinleri kullanılmaktadır.

3.7. Değerlendirme Çerçevesi

Değerlendirme, tanısal isabetin tek bir granülerlik düzeyine indirgenmesinin yanıltıcı olabileceği gözleminden hareketle, dört farklı eşleştirme düzeyi ve bunları tamamlayan istatistiksel çıkarım araçları üzerine kurulmuştur. Bu çok granülerli yaklaşım, referans çalışmanın yalnızca tanı adı düzeyinde yaptığı raporlamanın ortaya koyamadığı ayrımları görünür kılması hedeflenmiştir.

3.7.1. Tanısal Eşleştirme Düzeyleri

Bir tahminin doğru sayılması, giderek katılaşan dört ölçütle tanımlanmaktadır. Bu ölçütlerin üçü (tanı grubu, kategori ve alt kategori) iç içe bir kod hiyerarşisi oluşturmakta, tanı adı düzeyi ise kod hiyerarşisinden bağımsız ayrı bir eşleştirme ekseni sunmaktadır (Şekil 3.5):

Tanı adı düzeyi: LLMind çalışmasının protokolünün benzer biçimde yeniden uygulanmasıdır. Modelin birincil tanı adı, tanı standardının tanı adıyla eşleştirilir ve bu eşleştirmede şiddet, epizod ve belirleyici nitelikleri göz ardı edilmekte olup ICD-11 eş adlarına tolerans gösterilmektedir. Bu, öne çıkan asıl metriktir. Örnek: tanı standardı “Major depressive disorder, mild, single episode”, tahmin “single-episode depressive disorder, moderate” çekirdek tanı belirteçleri (“depressive”) örtüştüğünden doğru kabul edilmektedir.

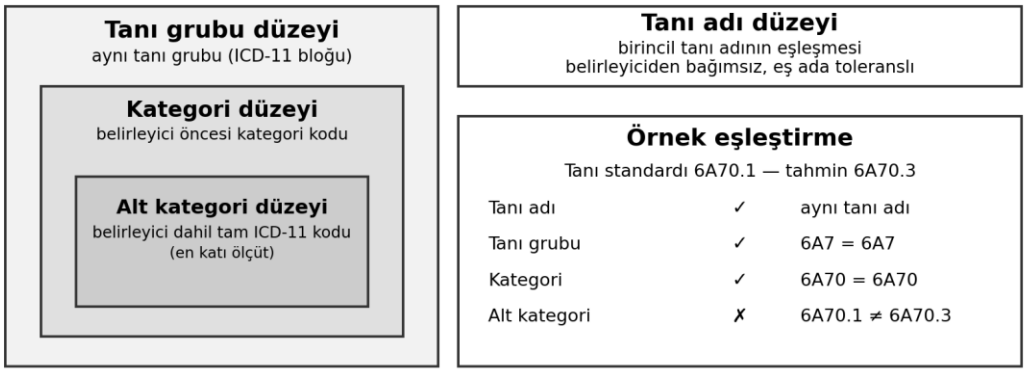
Tanı grubu düzeyi: Aynı tanı grubunu (ICD-11 bloğu) yakalamaktadır. Örnek: tanı standardı 6A70 (depresif), tahmin 6A71 (yineleyici depresif) tanı grubu 6A7 = 6A7, doğru kabul edilmektedir.

Kategori düzeyi: Belirleyiciden bağımsız olarak aynı bozukluğu yakalamaktadır. Eşleşme, ICD-11'in belirleyici öncesi kategori kodu üzerinden yapılmaktadır. Örnek: tanı standardı 6A70.1, tahmin 6A70.3 → kategori kodu 6A70 = 6A70, doğru kabul edilmektedir.

Alt kategori düzeyi: En katı ölçüttür; belirleyici dahil tam ICD-11 kodunun eşleşmesini gerektirir. Örnek: tanı standardı 6A70.1, tahmin 6A70.1 → 1; 6A70.3, yanlış kabul edilmektedir.

Düzeyler arasında alt kategori düzeyinde bir eşleşme, kategori ve tanı grubu düzeyini de gerektirecek bir ilişki vardır. Tanı grubu, kategori ve alt kategori düzeyleri sırasıyla ICD-11 bloğuna, belirleyici öncesi kategori koduna ve alt kategori koduna

karşılık gelmektedir. Bu operasyonel tanımlar ICD-11'in resmi "stem code" kavramından bağımsızdır. Tanı adı düzeyi, referans sistemle birebir karşılaştırma için birincil metrik, kod temelli üç düzey ise RAG'ın yapılandırılmış katkısını niceleyen metriklerdir.



Şekil 3.5. Tanısal eşleştirme düzeylerinin hiyerarşisi

Tek etiketli vaka düzeyi sınıflandırmada mikro ortalama Kesinlik, Duyarlılık ve F1, doğruluğa matematiksel olarak eşitlenmiştir (Bağıntı 3.9-Bağıntı 3.12). Kesinlik = Duyarlılık = F1 = Doğruluk (Sokolova & Lapalme, 2009; Japkowicz & Shah, 2011). Çalışmada kullanılan vaka düzeyi doğruluk, makro-kesinlik, makro-duyarlılık, makro-F1, çok sınıflı MCC ve hiyerarşik F1 metrikleri sırasıyla Bağıntı 3.12–3.19 ile tanımlanmıştır.

$$\text{Kesinlik} = \frac{\text{DP}}{\text{DP} + \text{YP}} \tag{3.9}$$

$$\text{Duyarlılık} = \frac{\text{DP}}{\text{DP} + \text{YN}} \tag{3.10}$$

$$\text{F1} = \frac{2 \cdot \text{Kesinlik} * \text{Duyarlılık}}{\text{Kesinlik} + \text{Duyarlılık}} \tag{3.11}$$

Aşağıda Çizelge 4.2'de raporlanan altı metriğin formel tanımları verilmektedir. Tüm metrikler ICD-11 tanı grubu düzeyinde hesaplanmıştır.

Doğruluk, sınıflandırıcının tüm vakalar üzerindeki mutlak başarı oranıdır. Tek etiketli kurguda, modelin ürettiği her tahmin bir pozitif örnek olarak kabul edildiğinden, mikro-ortalama kesinlik ve duyarlılık ile özdeşleşir. Burada $N = 104$ vaka (ICD-11

Bölüm 6 kapsamındaki 96 vaka dahil) üzerinden sistem tarafından doğru sınıflandırılan vakaların oranını vermektedir.

$$\text{Doğruluk} = \frac{1}{N}\sum_{i=1}^{N} 1[\hat{y}_i = y_i] \tag{3.12}$$

Makro-Kesinlik, pozitif tahmin değerinin ICD-11 tanı grubu düzeyinde genelleştirilmiş biçimidir. Sistem bir tanı grubunu (6A7 depresif bozukluklar) doğru olarak işaretlediğinde, bu işaretlemenin ne kadar güvenilir olduğunu ölçer. Formüldeki TP_c tanı grubuna atanan doğru tahmin sayısını, FP_c ise aynı tanı grubuna yanlışlıkla atanan vaka sayısını ifade eder. Sınıf dengesizliğinin baskın olduğu ICD-11 veri kümesinde (depresif bozukluklarda 12 vaka, parafilik bozukluklarda 2 vaka), bu ölçüt yaygın sınıfların yanlış pozitif değerini doğrudan yansıtmaktadır (Bağıntı 3.13).

$$\text{Makro} - \text{Kesinlik} = \frac{1}{|C|}\sum_{c\in C} \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c} \tag{3.13}$$

Makro-Duyarlılık, gerçek pozitif oranının tanı grubu düzeyinde genelleştirilmiş biçimidir. Gerçekte belirli bir tanı grubuna (6A6 bipolar) ait olan vakaların ne kadarının sistem tarafından doğru yakalandığını ölçer. Burada TP_c doğru yakalanan, FN_c ise kaçırılan (yanlışlıkla farklı tanı grubuna atanan) vaka sayısıdır. Klinik kullanım açısından kritik bir ölçüt olmaktadır. Yüksek duyarlılık, depresif bozukluk gibi yaygın tanıların atlanmamasını garanti eder. Seyrek sınıflardaki (n≤3) duyarlılık ise genellikle daha düşüktür, çünkü tek bir hata, sınıf içi oranı düşürmektedir .

$$\text{Makro} - \text{Duyarlılık} = \frac{1}{|C|}\sum_{c\in C} \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c} \tag{3.14}$$

Makro-F1, kesinlik ve duyarlılığın tek bir sayıda özetlenmesini sağlayan dengeli bir başarı ölçütüdür (Bağıntı 3.15). Her tanı grubu (ICD-11 bloğu) için ayrı ayrı F1 hesaplanmaktadır, ardından sınıf içi F1'lerin aritmetik ortalaması alınmaktadır. Bu yaklaşım, kesinlik ve duyarlılığın doğrudan harmonik ortalaması yerine gerçekte iki aşamalı bir hesaplaması olarak bulunmaktadır. Makro-F1'in temel avantajı, sınıf frekansından bağımsız biçimde her tanı grubuna eşit ağırlık vermesidir. ICD-11 kümesinde olduğu gibi dengesiz dağılımlarda (12 vakadan 2 vakaya uzanan tanı grubu büyüklükleri) bu, ağırlıklı F1'in gösteremeyeceği seyrek sınıf başarımlarını göstermektedir (Sokolova & Lapalme, 2009).

$$\text{Makro} - \text{F1} \ = \frac{1}{C}\sum_{c\in C} F1_c \tag{3.15}$$

Çok sınıflı Matthews Korelasyon Katsayısı, konfüsyon matrisinin dört hücresini de (TP, TN, FP, FN) tek bir skalar değere indirgeyen bir korelasyon ölçütüdür. K sınıflı karışıklık matrisi C üzerinden, burada c = toplam doğru tahmin ($\sum_k C_{k_k}$), s = toplam örnek sayısı ($\sum_{i,j} C_{i_j}$), t_k = k. tanı grubuna atanan tahmin toplamı ($\sum_i C_{i_k}$), p_k = k. tanı grubunun gözlenen toplamı ($\sum_i C_{k_i}$) olmaktadır (Jurman vd. 2012). MCC, -1 ile +1 arasında değer almaktadır. Hicks vd. (2022), tıbbi yapay zeka değerlendirmeleri için dengesiz sınıf dağılımlarında MCC'yi tek sayı özetleyici olarak önermiştir. Çünkü sınıf frekanslarından bağımsız olarak yanlış pozitif ve yanlış negatif hesaplamalarını eşit biçimde cezalandırmaktadır (Bağıntı 3.16).

$$\mathrm{MCC} = \frac{c \cdot s - \sum_k p_k \cdot t_k}{\sqrt{(s^2 - \sum_k {p_k}^2) \cdot (s^2 - \sum_k {t_k}^2)}} \tag{3.16}$$

ICD-11'in üç granülerlik düzeyinde tanı grubu (6A7), kategori düzeyi (6A70) ve alt kategori (6A70.1) grubu vardır. Hiyerarşik-F1, elde edilen F1 değerlerinin üst düzey bir özetidir ve her düzeyin eşit ağırlıkla katkıda bulunduğu aritmetik ortalamadır. Klinik önceliklendirmeyi yansıtmamaktadır. Örnek olarak alt kategori eşleşmesi genellikle klinik kararda en kritik düzeydedir, çünkü şiddet belirleyicisini doğru tespit etmek tedavi seçimini doğrudan etkilemektedir. N=104 gibi sınırlı veri kümelerinde eşit ağırlıklandırma, en sağlam tek sayı özetleyici olarak kabul edilmiştir (Bağıntı 3.17).

$$\mathrm{Hiyerarşik-F1} = \frac{\mathrm{F1}_{\mathrm{kategori}} + \mathrm{F1}_{\mathrm{tanı\ grubu}} + \mathrm{F1}_{\mathrm{alt-kategori}}}{3} \tag{3.17}$$

Bu metriklerin sayısal değerleri Çizelge 4.2'de verilmiştir.

3.7.2. Belirsizlik ve Anlamlılık Testleri

Bootstrap güven aralığı, her sistemin doğruluğu için %95 güven aralığı, vaka düzeyindeki 0 ve 1 etiketleri üzerinden 10.000 kez yeniden örneklemeli (Efron 1992) yüzdelik bootstrap yöntemiyle hesaplanmaktadır (Bağıntı 3.18).

$$\mathrm{CI}_{95\%} = \left[\hat{\theta}^*_{(2.5)}, \quad \hat{\theta}^*_{(97.5)}\right] \tag{3.18}$$

Referans değere karşı tek yönlü binom testi, N vakada c doğru gözlendiğinde, doğruluğun referans değer oranı olasılığıyla test edilmektedir ve bağıntı 3.19'da gösterilmiştir.

$$p = \sum_{k=c}^{N} \binom{N}{k} \pi_0^k (1-\pi_0)^{N-k} \tag{3.19}$$

N=104 için $p<0.05$ eşiği **≥64/104** doğru tahmine karşılık gelmektedir.

Eşleştirilmiş karşılaştırma (McNemar testi**),** iki sistem aynı vakalar üzerinde karşılaştırıldığında gözlemler eşleştirilmiş olduğundan, eşleştirilmemiş bir testi geçersiz kılmaktadır (Dietterich 1998). Uyumsuz yapılar b (birinci sistem doğru, ikinci yanlış) ve c (tersi) üzerinden McNemar istatistiği bağıntı 3.20'deki gibi hesaplanır.

$$\chi^2 = \frac{(b-c)^2}{b+c} \tag{3.20}$$

Küçük örneklemde ise min(b,c) üzerinden tam binom olasılığı raporlanmaktadır. Bu test, K4 ile önceki çalışmaları ve K4 ile RAG olmadan sistem karşılaştırmaları için uygulanmaktadır. Eşleştirilmiş testlerin gücü uyumsuz vaka sayısıyla sınırlı olduğundan, anlamlılık sonuçlarına ek olarak eşleştirilmiş doğruluk farkının bootstrap güven aralığı ile testin minimum saptanabilir etki büyüklüğü (MDE) de raporlanmaktadır (Dietterich 1998).

3.8. Donanım ve Uygulama (TRUBA)

Tüm üretim işlemleri, llama.cpp aracının CPU derlemesiyle TRUBA'nın orfoz, hamsi ve barbun CPU bölümlerinde yürütülmüştür. CPU üzerinde çıkarım, güvenilir ve tekrarlanabilir bir yol olması nedeniyle benimsenmiştir. Bu tercih aynı zamanda sistemin erişilebilirlik katkısını güçlendirmiştir.

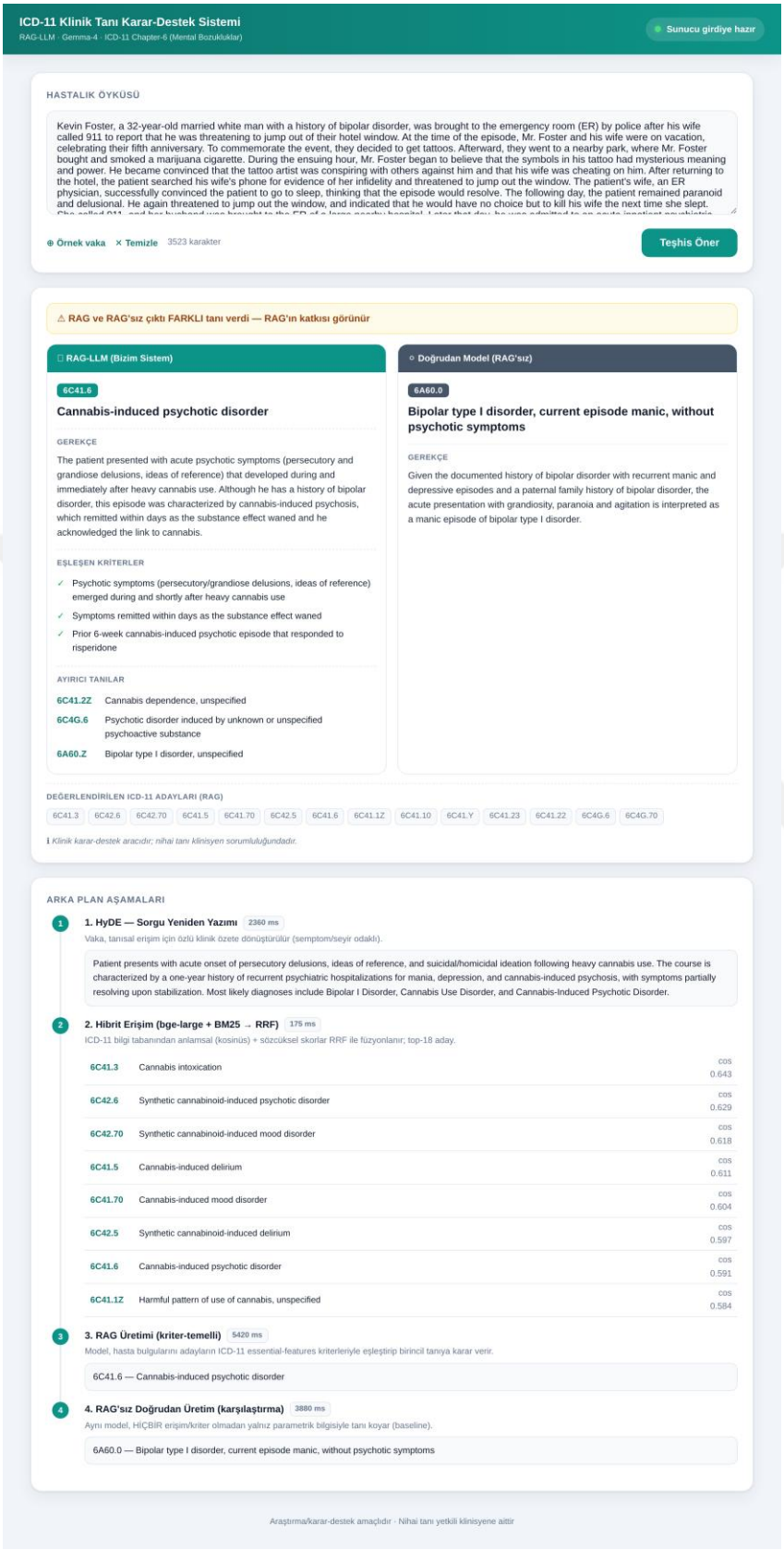
3.9. Karar Destek Arayüzü (Web Uygulaması)

Önerilen K4 konfigürasyon sistemi uçtan uca işleyişini, klinisyen ve araştırmacıların etkileşimli olarak kullanabileceği bir web tabanlı karar destek arayüzünde somutlaştırmak amacıyla bir gösterim uygulaması geliştirilmiştir. Arayüz sisteminde, komut satırı gerektirmeden erişilebilir kılmak ve tanısal kararı şeffaf biçimde sunmak amaçlanmıştır. Yalnızca en son tanıyı değil, ona götüren erişim ve üretim aşamalarını ve RAG'ın katkısını da görünür kılması hedeflenmiştir.

Uygulama, bağımlılık gerektirmeyen bir Python standart kütüphane HTTP sunucusu ile statik bir HTML/JavaScript ön yüzünden oluşmaktadır. Sunucu, başlangıçta bge-large gömme modelini ve ICD-11 bilgi tabanını belleğe yüklemektedir ve bir llama.cpp Gemma 4 sunucusuna bağlamaktadır. Kullanıcı, Hastalık Öyküsü alanına bir anamnez girip "Teşhis Öner" düğmesine bastığında arka planda K4 iş akışı başlatılmaktadır. Sırasıyla HyDE özeti, hibrit erişim (bge-large + BM25, RRF), kriter temelli üretim ve geçerli ICD-11 koduna çözümleme adımları tamamlanmaktadır. Arayüz; birincil tanıyı (ICD-11 kodu, başlık ve resmi ICD-11 tarayıcı bağlantısı), gerekçeyi ve ayırıcı tanıları göstermektedir. Ayrıca önerilen RAG-BDM ile RAG olmadan doğrudan model çıktılarının yan yana karşılaştırmasını ve bunların eşleşip eşleşmediğini belirten bir özet bandını içermektedir. Her aşamanın süresini ve ara çıktılarını listeleyen bir arka plan zaman çizelgesi de bulunmaktadır. Sayfa altındaki uyarı, çıktının yalnızca araştırma veya karar destek amaçlı olduğunu ve nihai tanının yetkili klinisyene ait olduğunu vurgular.

Örnek olarak, Şekil 3.6, 11. vaka (eştanılı bipolar öyküsü ve yoğun kannabis kullanımı bulunan bir hasta) için arayüzün çıktısını göstermektedir. Bu vaka, RAG'in

katkısını çarpıcı biçimde örneklemektedir. Önerilen RAG-BDM sistemi tabloyu doğru biçimde kannabis kaynaklı psikotik bozukluk (6C41.6) olarak sınıflandırmış ve bu sınıflama tanı standardıyla birebir örtüşmüştür. Buna karşılık RAG olmadan doğrudan model, hastanın bipolar öyküsüne takılarak aynı vakayı bipolar tip I bozukluk (6A60.0) olarak yanlış etiketlemiştir. Böylece arayüz, erişimle topraklamanın madde kaynaklı tabloları ayırt etmedeki somut katkısını görünür kılmaktadır.



Şekil 3.6. 11. vakanın klinik tanı karar destek arayüzü örneği

4. BULGULAR

Bu bölüm, geliştirilen sistemin başarımını, dokuz RAG konfigürasyonunun (K1–K9) ablasyonunu, erişim ve model ölçekleme çözümlemelerini, hata örüntüsünü ve istatistiksel anlamlılık testlerini açıklamaktadır.

4.1. Ana Karşılaştırma: Sistemler ve Referans Değer

Çizelge 4.1, başlıca sistemlerin dört farklı granülerlik düzeyindeki doğruluğunu sunmaktadır (N=104). Önerilen K4 ve Gemma 4 12B, tanı adı ve alt kategori düzeylerinde en yüksek değere ulaşırken, tanı grubu ve kategori düzeylerinde 26B-A4B modeli öne çıkmaktadır. Karışık uzman 26B-A4B modeli buna denk bir doğruluğu daha az etkin parametreyle sağlayarak verimli bir alternatif oluşturmaktadır.

Çizelge 4.1. En yüksek doğruluk tespit edilen konfigürasyonlar

Sistem	Tanı adı	Tanı grubu	Kategori	Alt kategori
K4 - Gemma 4 12B	**0.615**	0.615	0.519	**0.394**
K4 - Gemma 4 26B-A4B	0.606	**0.683**	**0.567**	0.365
K9 - Gemma 4 26B-A4B	0.567	0.654	0.510	0.356
RAG olmadan - Gemma 4 26B-A4B	0.481	0.644	0.413	0.183
RAG olmadan - Gemma 4 12B	0.337	0.462	0.240	0.135
RAG olmadan - Gemma 4 E2B	0.231	0.317	0.164	0.077

Çizelge 4.2, Çizelge 4.1'deki sistemlerden dördünün tanı grubu düzeyi sınıflandırma kalitesini, Doğruluk, Makro-Kesinlik, Makro-Duyarlılık, Makro-F1'e ek olarak MCC ve Hiyerarşik-F1 ölçütleriyle birlikte raporlamaktadır. Doğruluk dışındaki tüm sütunlar, ICD-11 tanı grubu düzeyinde ve üç yöntemli tanı standardı eşlemesi üzerinden makro-ortalama olarak hesaplanmıştır. Bu tablo, referans model dışındaki karşılaştırmayı yani yalnızca önerilen K4 ve aynı model ailesinin RAG olmadan tabanı ile K9 konfigürasyonu arasındaki farkları ortaya koymaktadır. LLMind ve diğer açık kaynak modellerle doğrudan karşılaştırma ise Bölüm 5'teki Çizelge 5.1'de sunulmaktadır.

Çizelge 4.2. Tanı grubu düzeyinde elde edilen metrikler

Sistem	Doğruluk (vaka)	Makro-Kesinlik (tanı grubu)	Makro-Duyarlılık (tanı grubu)	Makro-F1 (tanı grubu)	MCC (tanı grubu)	Hiyerarşik-F1	Tanı Yok
K4 — Gemma 4 12B (önerilen)	0.615	0.431	0.478	0.423	0.599	0.509	0
K4 — Gemma 4 26B-A4B	0.606	0.453	0.511	0.465	0.666	0.538	0
K9 — Gemma 4 26B-A4B	0.567	0.437	0.463	0.427	0.635	0.506	0
RAG olmadan — Gemma 4 26B-A4B	0.481	0.469	0.498	0.473	0.643	0.414	0

Tanı grubu düzeyi ölçütler, 26B-A4B modelinin K4-26B olarak Makro-Kesinlik 0.453, Makro-Duyarlılık 0.511, Makro-F1 0.465 ve MCC 0.666 değerleriyle en yüksek sınıflandırma kalitesine ulaştığını, buna karşılık aynı konfigürasyonun K4-12B'nin ise daha düşük Makro-Kesinlik 0.431 ve Makro-Duyarlılık 0.478 değerleri almasına rağmen vaka düzeyinde en yüksek doğruluğa 0.615 eriştiğini göstermektedir. Bu tablo, vaka düzeyi keskin eşleşmenin parametrik bilgiye, tanı grubu düzeyi daha yumuşak eşleşmenin ise daha büyük modelin semantik kararlılığına duyarlı olduğunu gösterebilir. Nitekim RAG olmadan 26B-A4B tabanının Makro-F1 0.473 ve MCC 0.643 değerlerinin K4-12B'den (sırasıyla 0.423 ve 0.599) yüksek kalması, aynı yorumu desteklemektedir. Hiyerarşik-F1 sütunu, tanı grubu düzeyindeki başarının alt kategori düzeyine yansımadığını (K4-12B: 0.509; K4-26B: 0.538) niceliksel olarak ortaya koymaktadır. Aslında sistem doğru tanı grubunu bulabilmekte, ancak aynı tanı grubu içindeki spesifik tanıyı ayırt etmekte zorlanmaktadır. Tüm sistemlerde Tanı Yok sütununun sıfır olması, hiçbir vakada çekimser kalınmadığını göstermektedir.

4.2. İstatistiksel Anlamlılık

Çizelge 4.3, önerdiğimiz sistemin 104 vaka üzerindeki doğruluğunu ve LLMind çalışmasındaki 0.524 değerine karşı tek yönlü binom anlamlılığını sunmaktadır. Önerilen K4-12B en yüksek değere sahiptir ve sabit referansı anlamlı biçimde geçen tek sistemdir (64/104, p=0.038); 26B-A4B modeli ise (63/104) bu eşiğin hemen altında kalmaktadır (p=0.058).

Çizelge 4.3. Doğruluk ve LLMind'a karşı anlamlılık (N=104; tek yönlü binom)

Sistem	Doğruluk (N=104)	p
K4-12B	64/104 = 0.615	0.038

(Devamı arkada)

Çizelge 4.3.'ün devamı.

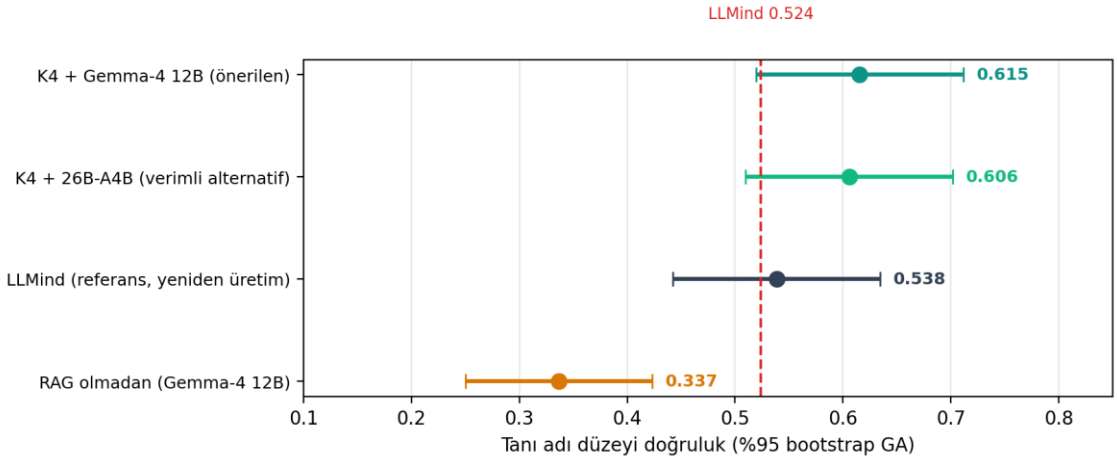
K4-26B	63/104 = **0.606**	0.058
K9-26B	59/104 = 0.567	0.216
RAG olmadan (26B)	50/104 = 0.481	0.837

Eşleştirilmiş testte, iki sistemin aynı vakalardaki uyumsuz kararlarına (yalnızca birinin doğru olduğu hücrelere) odaklanmaktadır. Önerilen K4-12B ile RAG olmadan karşılaştırmasında 26 vakada her iki sistem de doğru, 31 vakada her ikisi de yanlıştır. Uyumsuz 47 vakanın 38'inde yalnızca K4-12B, 9'unda yalnızca RAG olmadan doğrudur. LLMind karşılaştırmasında ise 41 vaka her ikisinde doğru, 25 vaka her ikisinde yanlıştır. Uyumsuz 38 vakanın 23'ü K4-12B, 15'i LLMind tarafındadır. Aşağıdaki p değerleri bu uyumsuz hücreler üzerinden hesaplanmıştır.

K4 ve RAG olmadan karşılaştırması (Gemma 4 12B): uyumsuz hücreler b=38 (RAG doğru/RAG olmadan yanlış), c=9 (tersi); $p < 0.001$. RAG'ın katkısı, parametrik bilgisi daha sınırlı olan 12B modelinde çok belirgindir (26B-A4B'de b=22, c=9, p=0.029).

K4-12B ve LLMind karşılaştırması: b=23, c=15; $p = 0.256$ şeklindedir. Nokta tahmininde K4-12B önde olsa da (0.615 - 0.538) eşleştirilmiş testte fark anlamlılığa ulaşmamaktadır.

Bootstrap %95 güven aralığı incelendiğinde, önerilen K4-12B tanı adı düzeyi doğruluğu için %95 bootstrap güven aralığı yaklaşık [0.52, 0.71] olarak hesaplanmıştır (Şekil 4.1). Aralığın alt sınırı LLMind değerine (0.524) komşudur ve bu, binom anlamlılığıyla (p=0.038) tutarlı olmaktadır. 26B-A4B modelinin aralığı benzer konumdadır. Aralıkların genişliği, tek bir set kullanılmasının doğal sonucudur.



Şekil 4.1. Tanı adı düzeyi doğruluğu ve %95 bootstrap güven aralıkları orman grafiği

4.3. RAG Konfigürasyon Ablasyonu (K1–K9)

Çizelge 4.4, dokuz konfigürasyonun tanı adı düzeyi doğruluğunu özetlemektedir. En yüksek doğruluğu yalın HyDE birleştirmesi (K4) sağlamaktadır. HyDE'yi dışarıda bırakan (K6/K7/K8) veya K4'ün üzerine ek katman getiren (K5/K9) hiçbir konfigürasyon onu geçememektedir.

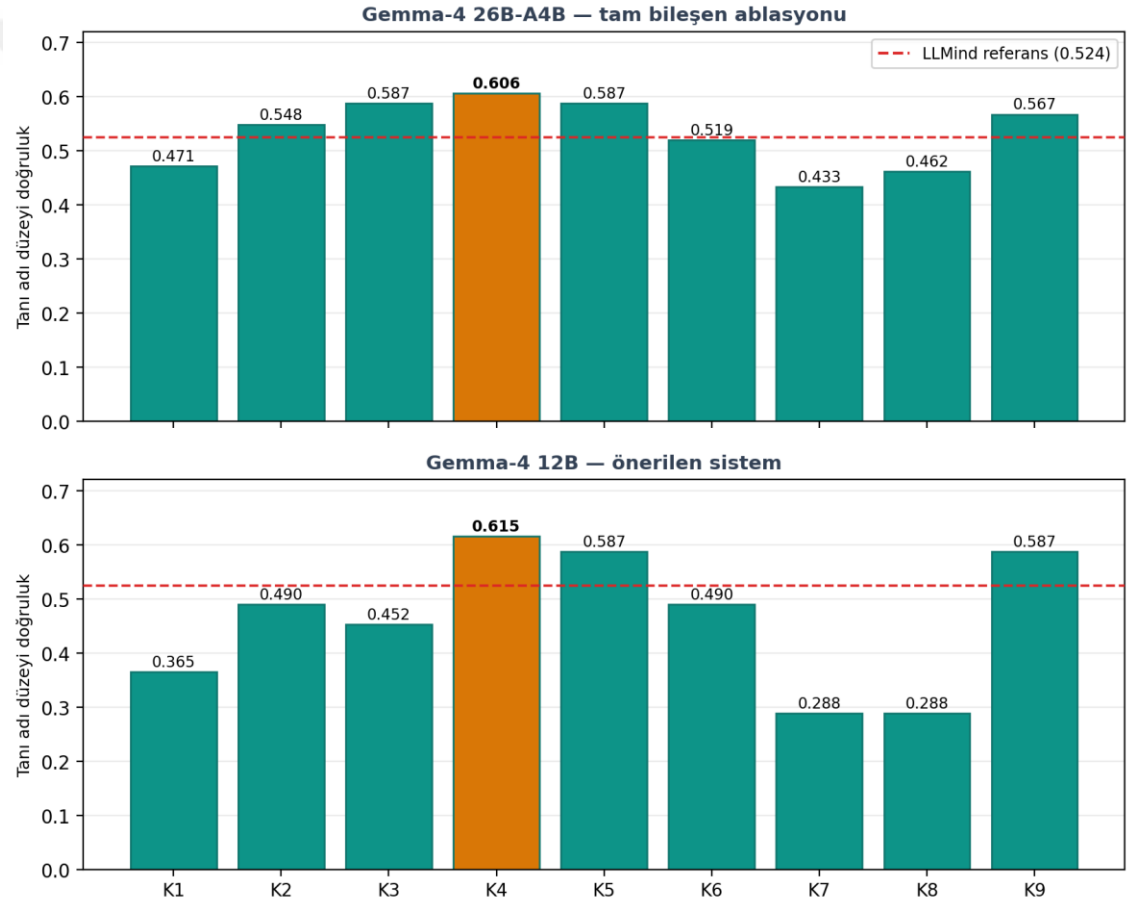
Çizelge 4.4. RAG konfigürasyon ablasyonu karşılaştırması (Gemma 4 26B-A4B, N=104)

Konfigürasyon	Tanı adı	Tanı grubu	Kategori	Alt kategori
K1	0.471	0.606	0.413	0.192
K2	0.548	0.625	0.510	0.308
K3	0.587	0.644	0.510	0.269
K4	**0.606**	**0.683**	**0.567**	**0.365**
K5	0.587	0.644	0.519	0.356
K6	0.519	0.663	0.538	0.298
K7	0.433	0.529	0.365	0.192
K8	0.462	0.606	0.356	0.212
K9	0.567	0.654	0.510	0.356

RAG'ın katkısının nerede yoğunlaştığını görmek için, önerilen K4 konfigürasyon ve 12B modelinde RAG'ı doğru, RAG içermeyen tabanın yanlış olduğu 38 vaka ICD-11 tanı gruplarına göre incelenmiştir. Kazanç en çok madde kullanımı (7 vaka), birincil psikotik bozukluklar (6) ve obsesif-kompulsif bozukluklar (4) tanı gruplarında toplanmaktadır. Bunları anksiyete, nörokognitif ve dürtü denetimi tanı grupları takip etmektedir. RAG içermeyen tabanın doğru olup K4'ün yanıldığı yalnızca 9 vaka vardır ve bunlar ağırlıkla kişilik (3) ve depresif (2) bozukluklarda görülmektedir. Bu belirgin asimetri (38'e karşı 9), RAG'ın özellikle modelin parametrik bilgisinin yanıltıcı olabildiği tablolarda ayırt edici olduğunu somutlaştırmaktadır. Yoğun madde kullanımı eşliğindeki psikotik belirtiler RAG olmadan sıklıkla birincil psikotik ya da bipolar bozukluğa yönlendirilirken, kriterlerle topraklanan RAG bunları madde kaynaklı tabloya doğru biçimde oturtur.

Şekil 4.2'nin iki paneli birlikte değerlendirildiğinde, tezi destekleyen tutarlı bir tablo ortaya çıkmaktadır. HyDE'nin kullanılmadığı üç bağımsız telafi konfigürasyonu (K6/K7/K8) belirgin biçimde gerilemekte ve bu düşüş 12B'de göz önüne gelmektedir. K7 ve K8 yalnızca 0.288 düzeyinde kalarak en yalın taban olan K1'in dahi altına inmektedir, dolayısıyla HyDE'nin kullanılması gereken bir bileşen olduğu ortaya

konmaktadır. Buna karşılık, K4 üzerine katman ekleyen K5 (kriter parçalama) ve K9 (ayırıcı tanı ve belirleyici) konfigürasyonları 12B'de yalnızca 0.587 düzeyinde kalarak K4'ün altında konumlanmıştır. Bu durum ek karmaşıklığın kazanç getirmediğini göstermektedir. İki panel arasındaki en öğretici fark, HyDE'nin tanı adı düzeyindeki katkısının modele bağlı olmasıdır. 12B modelinde bu katkı oldukça büyüktür ve K3'ten K4'e geçişte doğruluk 0.452'den 0.615'e yükselmiştir. Ancak parametrik bilgisi zaten güçlü olan 26B-A4B'de aynı geçiş küçülerek 0.587'den 0.606'ya çıkmıştır. Nitekim K2 ile K3'ün göreli sırası da iki model arasında tersine dönmektedir. Referans çizgisi (LLMind, 0.524) açısından bakıldığında bu model bağımlılığı pratik bir sonuca dönüşmektedir. Büyük modelde RAG konfigürasyonlarının çoğu referansı geçerken, 12B'de referansı net biçimde yalnızca HyDE içeren konfigürasyonlar (K4/K5/K9) aşmaktadır. Ayrıca 12B'nin çubuklarının çok daha geniş bir aralığa (0.288–0.615) yayılması, küçük modelin erişime daha bağımlı, büyük modelin ise parametrik bilgisiyle daha dayanıklı olduğunu ortaya koymaktadır.



Şekil 4.2. K1–K9 konfigürasyonlarının doğruluk karşılaştırması

4.4. Erişim Analizi - Recall@k

Çizelge 4.5 ve Şekil 4.3, kapsam içi 96 vaka için kategori düzeyi recall@k değerlerini sunmaktadır. Eğriler, HyDE en üst sıralarda açık ara öndedir. İlk adayda (recall@1) HyDE doğru kategori kodunu 0.354 oranında getirirken, ham BM25 yalnızca 0.052, yoğun erişim ise 0.229'da kalmaktadır. Yani HyDE doğru bozukluğu ilk sıraya taşımada sözcüksel erişimin yaklaşık yedi katı etkilidir. Birleştirme yalnızca daha

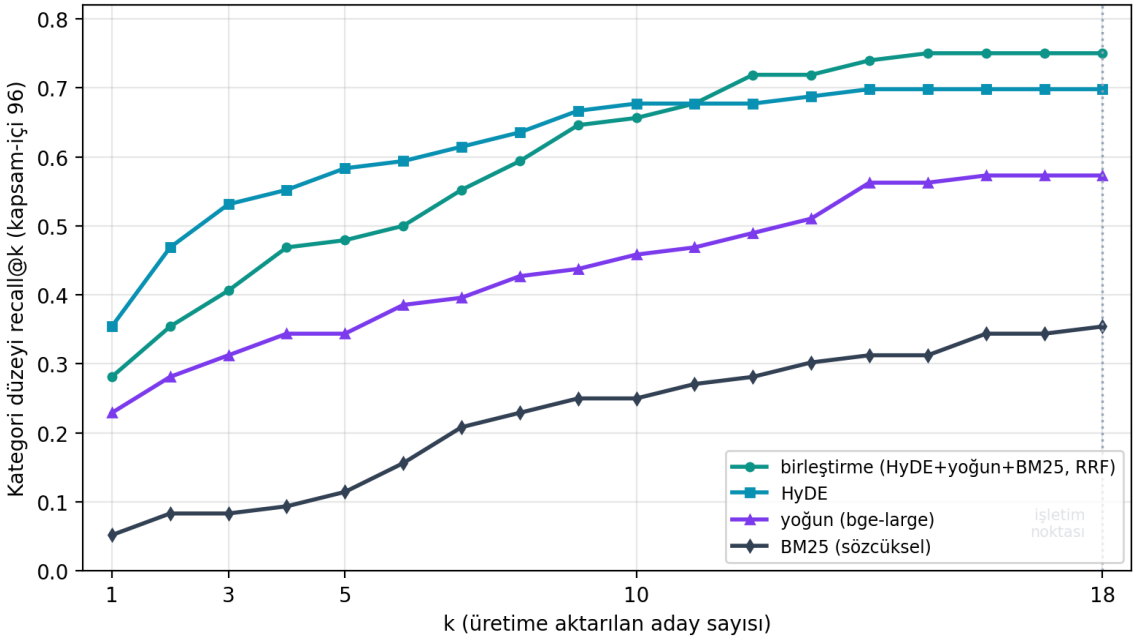
derin sıralarda (k>10) HyDE'yi geçmektedir ve üretime aktarılan en iyi-18 işletim noktasında en yüksek değere ulaşmaktadır (recall@18: birleştirme 0.750, HyDE 0.698, yoğun 0.573, BM25 0.354). Bunun nedeni, sözcüksel ve yoğun sıralayıcıların HyDE'nin kaçırdığı kimi vakaları farklı gerekçelerle yakalaması ve RRF'nin bu tamamlayıcı isabetleri birleştirmesidir. Sistem üretime ilk 18 adayı aktardığından, bu işletim noktasında en yüksek recall'ı veren birleştirme benimsenmiştir.

Recall'ın yalnızca kapsam içi 96 vaka üzerinden raporlanması bilinçli seçilmiştir. Kapsam dışı sekiz vakanın tanı standardı kodu (ör. uyku-uyanıklık bozuklukları, akut stres tepkisi) ICD-11 Bölüm-6 bilgi tabanında hiç yer almadığından, bu vakalar için doğru adayın getirilmesi tanım gereği olanaksızdır ve recall'ları yapısal olarak sıfırdır. Bu nedenle söz konusu vakalar bir erişim değil bir kapsam sorunu olmaktadır. Cezaları, uçtan uca doğruluğun tam korpus (N=104) üzerinden raporlanmasında ve çekimserlik testinde zaten yansıtılmıştır. Erişimin kendi kalitesini ölçmek için doğru payda, hedefin bilgi tabanında bulunduğu kapsam içi 96 vakadır.

Çizelge 4.5. Kategori düzeyi recall@18 (üretime aktarılan en iyi-18 işletim noktası; kapsam içi 96 vaka)

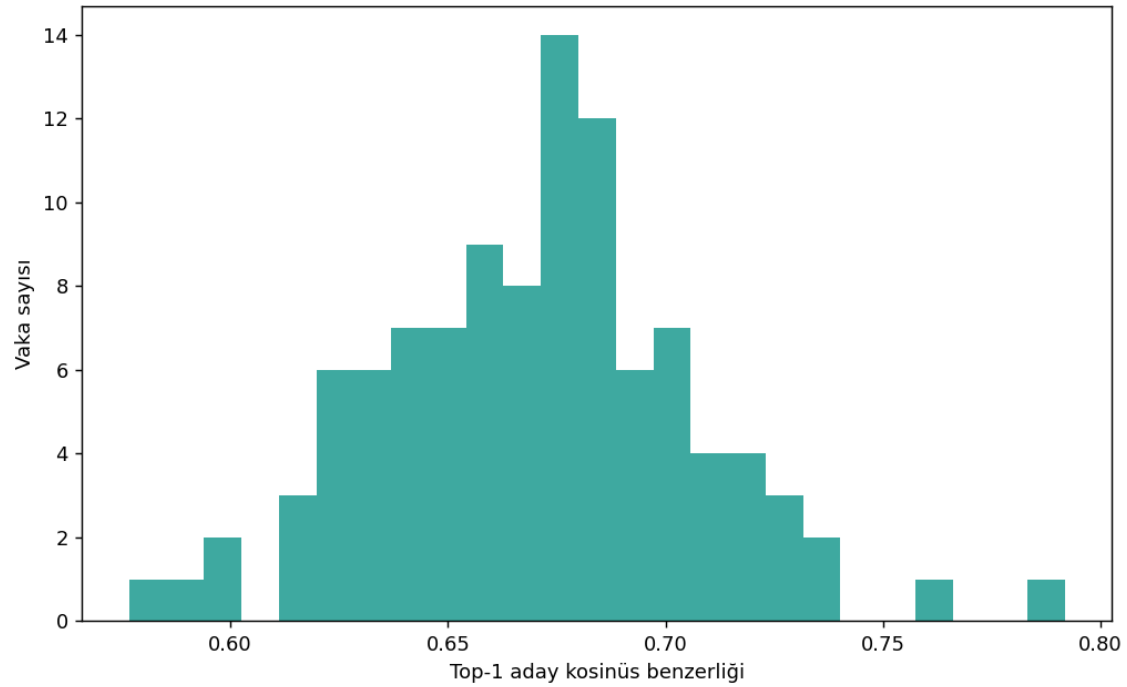
Erişim	**recall@18**
BM25 (sözcüksel)	0.354
Yoğun (bge-large)	0.573
HyDE	0.698
Birleştirme (HyDE+yoğun+BM25, RRF)	0.750

BM25'in katkısı, HyDE'nin varlığına bağlıdır. HyDE devredeyken BM25 recall'a yaklaşık +0.03 eklerken, HyDE yokken ham anlatı üzerinde gürültü ürettiği için başarımı düşürür. Sonuç olarak BM25 ancak HyDE ile birlikte kullanıldığında faydası olmaktadır.



Şekil 4.3. Kategori düzeyi recall@k eğrileri (kapsam içi 96 vaka)

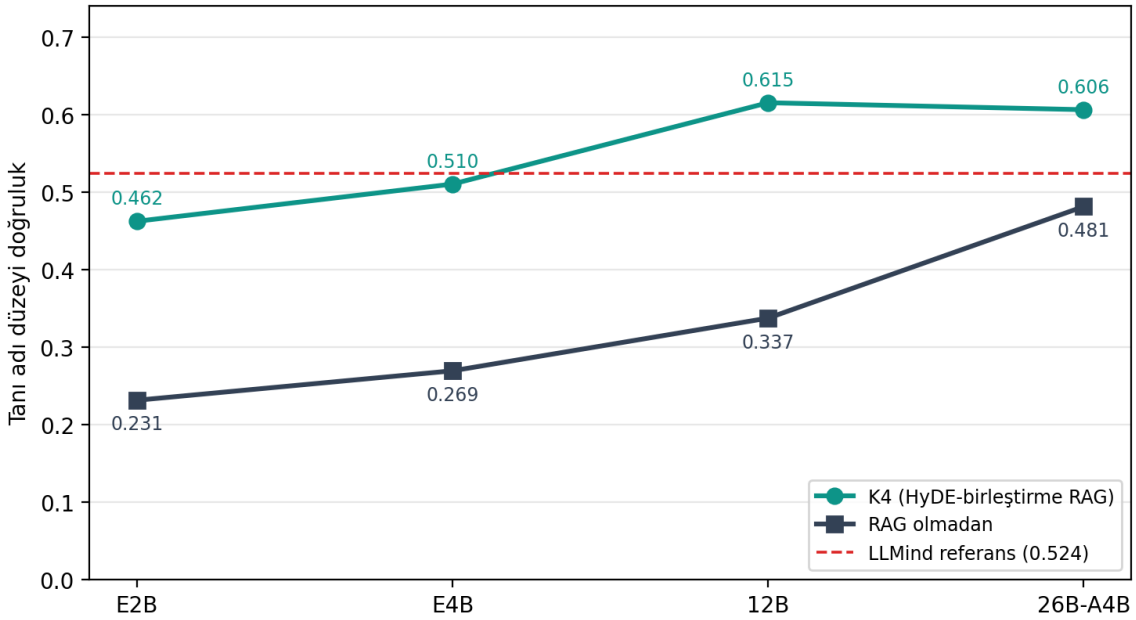
Erişimin "güven" dağılımı da incelenmiştir. Şekil 4.4, her vakanın en üst sıradaki ICD-11 adayına ilişkin kosinüs benzerliğini göstermektedir. Düşük benzerlik değerine sahip vakalar, erişimin zorlandığı ve hata olasılığının arttığı vakalar olup, gelecekte yapılacak bir eşik çözümlemesi için değerlendirilebilir.



Şekil 4.4. Vakaların en üst sıra ICD-11 adayına kosinüs benzerliği dağılımı

4.5. Model Ölçekleme

Şekil 4.5, model boyutunun doğruluğa etkisini göstermektedir. Doğruluk model boyutuyla birlikte artar ve en yüksek tanı adı düzeyi doğruluğu 0.615 olarak Gemma 4 12B'de elde edilmiştir. Bu nedenle önerilen sistemin üretim modeli olarak 12B benimsenmiştir. 26B-A4B karışık uzman modeli buna istatistiksel olarak oldukça yakın 0.606 olan doğruluğu yaklaşık 4 milyar etkin parametreyle sağladığından, hesaplama verimliliği öncelikli olduğunda tercih edilebilecek verimli bir alternatiftir. Küçük modeller (E2B/E4B) referans değerin altında kalmaktadır. RAG eğrisi, RAG olmadan sistemin her boyutta üzerinde seyretmektedir. K4 konfigürasyonunda Gemma4-12B'de RAG olmadan 0.337 olan doğruluk değeri RAG ile birlikte 0.615 seviyesine, çıkarak RAG'ın kazancının küçük/orta modellerde daha büyük olduğu söylenebilir. Erişimle topraklama zayıf parametrik bilgiyi orantısız biçimde telafi edebilmektedir.

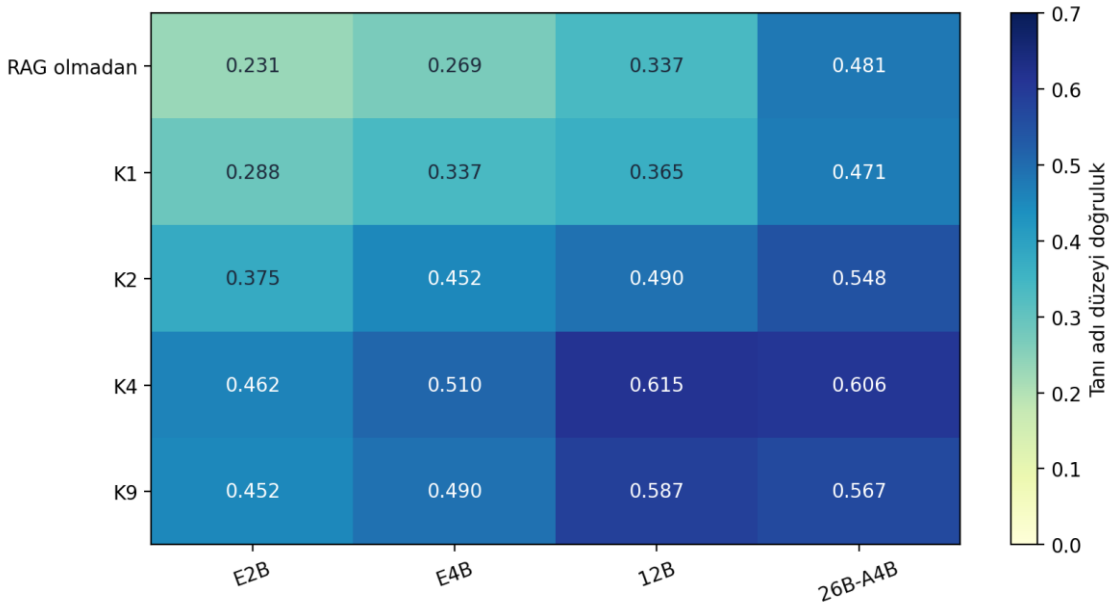


Şekil 4.5. Model boyutuna karşı tanı adı düzeyi doğruluk (E2B, E4B, 12B, 26B-A4B); iki eğri RAG (K4) ve RAG olmadan konfigürasyonları

Erişim yöntemi ile model boyutu arasındaki etkileşim, Çizelge 4.6 ve Şekil 4.6'daki RAG yöntemi ile BDM karşılaştırmada incelenmiştir. RAG kazancı her boyutta tutarlı olmaktadır. En basit RAG (K1) bile her model boyutunda RAG olmadan konfigürasyonu geçmektedir. HyDE temelli yöntemlerin (K4/K9) üstünlüğü tüm boyutlarda korunurken HyDE olmadan karşıtlama konfigürasyonlarının (K7/K8) her boyutta gerilemesi gözükmektedir. Ayrıca RAG'ın kazancı küçük ve orta modellerde (örn. 12B) daha büyüktür. Çok aşamalı veya karşılaştırmalı kurguların küçük modellerde (E2B) zayıflaması, bu modellerin yapılandırılmış ve çok adımlı istemleri güvenilir biçimde izleyememesiyle açıklanabilir.

Çizelge 4.6. RAG yöntemi - BDM karşılaştırması (tanı adı düzeyi doğruluk, N=104)

Yöntem	E2B	E4B	12B	26B-A4B
RAG olmadan	0.231	0.269	0.337	0.481
K1	0.288	0.337	0.365	0.471
K2	0.375	0.452	0.490	0.548
K3	0.404	0.385	0.452	0.587
K4	0.462	0.510	0.615	**0.606**
K6	0.385	0.433	0.490	0.519
K7	0.269	0.279	0.288	0.433
K8	0.260	0.279	0.288	0.462
K9	0.452	0.490	0.587	0.567



Şekil 4.6. RAG yöntemi ve BDM ısı haritası (tanı adı düzeyi doğruluk, N=104)

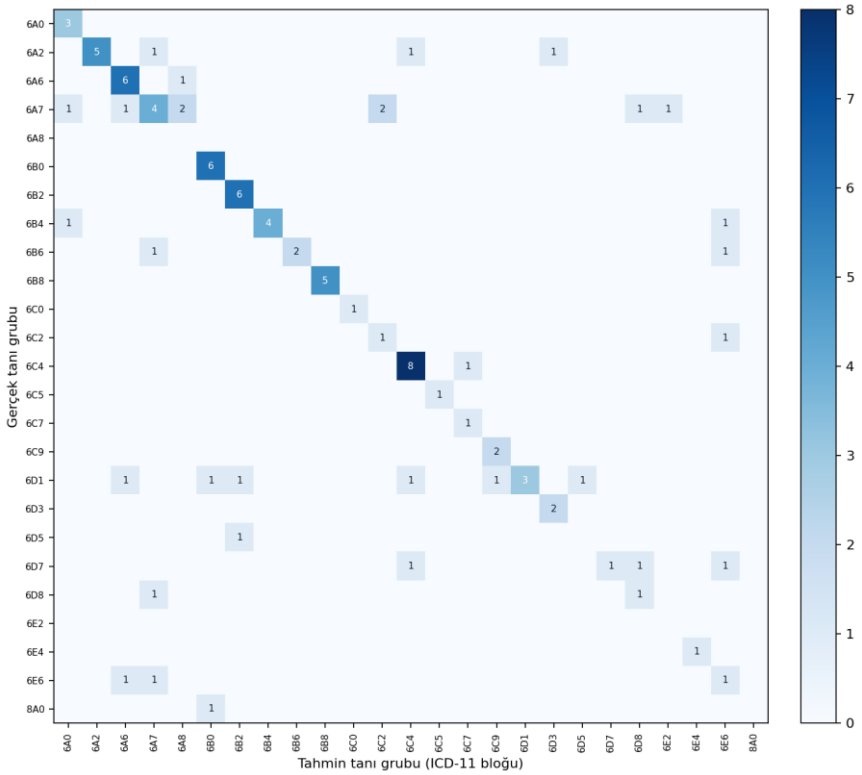
Üretim modelinin niceleme düzeyinin etkisi, K1 erişimiyle 26B-A4B üzerinde test edilmiştir (Çizelge 4.7). En agresif 4-bit nicelemeye (Q4_K_M) inildiğinde bile tanı adı düzeyi doğruluktaki düşüş sınırlı kalmaktadır (≤0.03). Buna göre CPU üzerinde belleğe sığması için seçilen niceleme, tanısal başarımı kayda değer biçimde bozmamaktadır.

Çizelge 4.7. Niceleme duyarlılığı: K1 erişimi (Gemma 4 26B-A4B, N=104)

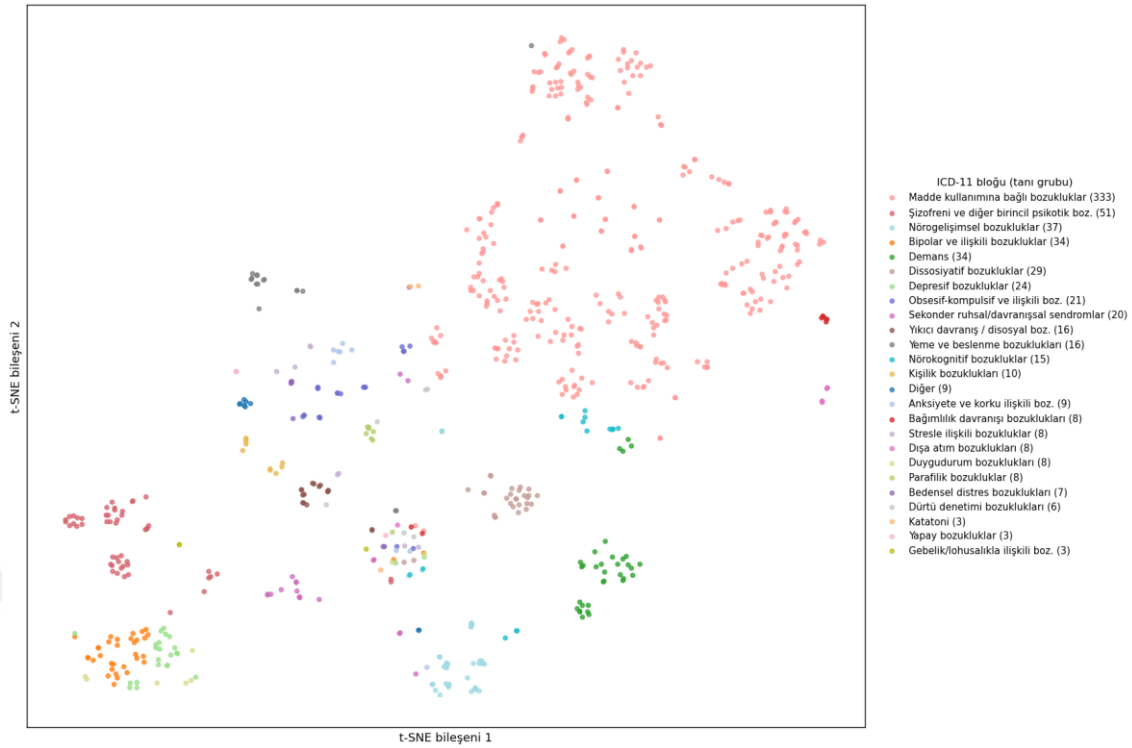
Niceleme (GGUF)	Tanı adı	Tanı grubu	Kategori	Alt kategori
UD-Q4_K_XL (Unsloth)	0.471	0.606	0.413	0.192
Q8_0 (Google)	0.471	0.654	0.413	0.183
Q4_K_M (Google)	0.442	0.596	0.394	0.173

4.6. Hata Örüntüsü - Karışıklık Matrisi

K4-12B ve kapsam içi vakaların tanı grubundaki doğru ve hatalı tespitlerin hangi komşu bozukluk sınırlarında yoğunlaştığı Şekil 4.7'de gösterilmektedir. Karışıklıklar en çok depresif bozukluklar (6A7) çevresinde toplanmıştır. Bu vakalar komşu duygudurum, bedensel sıkıntı (6C2) ve nörogelişimsel tanı gruplarına kayabilmektedir. Ayrıca psikotik (6A2) ve kişilik (6D1) bozuklukları komşu tanı gruplarına dağılmaktadır. Sistem çoğunlukla doğru tanı grubuna yakın bir tahmin yapmaktadır ancak asıl zorluk komşu bozukluk ve belirleyici ayrımında ortaya çıkmaktadır. Bu örüntü, vektör uzayındaki kümelenmeyle (Şekil 4.8) ve tanı grubu düzeyi doğruluğunun (0.615) alt kategori düzeyi doğruluğundan (0.394) yüksek olmasıyla tutarlıdır.



Şekil 4.7. Tanı grubu düzeyi karışıklık matrisi (önerilen K4-12B, kapsam içi vakalar)



Şekil 4.8. ICD-11 Bölüm 6 bozukluğunun bge-large gömmelerinin iki boyuta izdüşümü (t-SNE, kosinüs metriği)

4.6.1. Tanı Grubu Bazlı Performans

Önerilen K4-12B konfigürasyonunun başarımını, hata örüntüsünü tamamlayacak biçimde ICD-11 tanı grupları bazında da incelenmiştir (Çizelge 4.8, Şekil 4.9). Sistem obsesif-kompulsif (6B2) ve yeme-beslenme (6B8) tanı gruplarında tam isabet sağlarken; kişilik bozukluklarında (6D1; 0.33), sekonder ruhsal sendromlarda (6E6; 0.33) ve nörogelişimsel bozukluklarda (6A0; 0.33) zorlanmakta, bipolar (6A6) ise 0.71'e gerilemektedir. Kişilik patolojisi boyutsal ve örtüşen bir yapıya sahiptir, sekonder sendromların ayırıcı tanısı ise altta yatan tıbbi duruma bağlı olduğu için güçtür. Tanı grubu örneklemleri çok dengesiz olduğundan (bazı tanı grubu yalnızca 1–2 vaka) tanı grubu değerler dikkatle değerlendirilmelidir. Tanı gruplarına eşit ağırlık veren makro-ortalama doğruluk, n≥2 tanı grupları için mikro değere, olan 0.656'ya yakın, nadir tek vakalı tanı grupları de katıldığında ise yaklaşık 0.61 düzeyindedir.

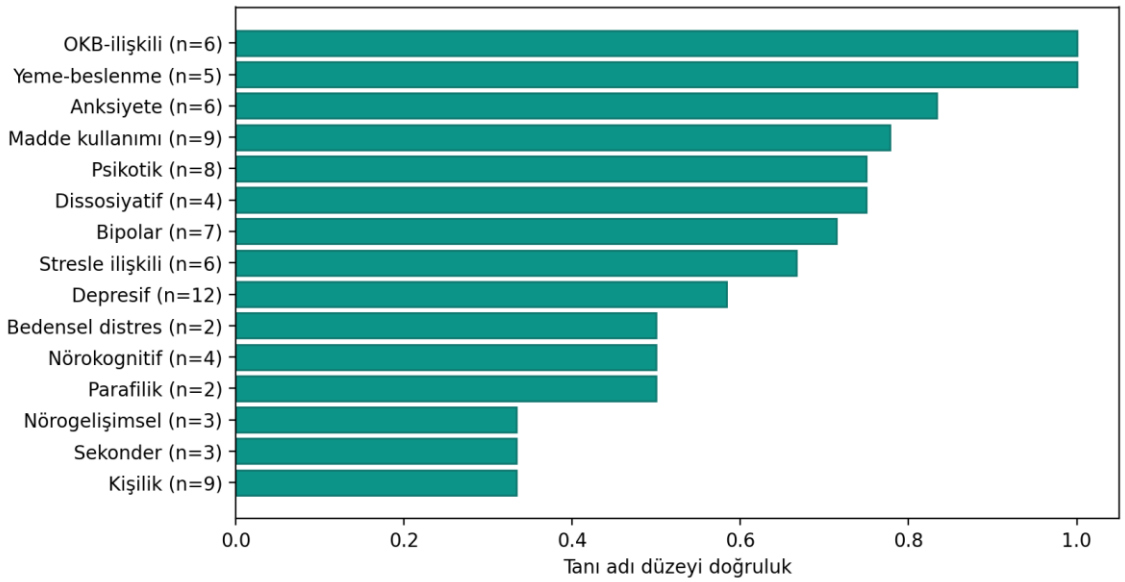
Çizelge 4.8. Önerilen K4-12B konfigürasyonunun ICD-11 tanı grubu bazlı tanı adı düzeyi doğruluğu

ICD-11 tanı grubu	Tanı grubu	Doğru / n	Doğruluk
6B2	OKB ve ilişkili	6/6	1.00
6B8	Yeme-beslenme	5/5	1.00

(Devamı arkada)

Çizelge 4.8.'in devamı.

6B0	Anksiyete	5/6	0.83
6C4	Madde kullanımı	7/9	0.78
6A2	Psikotik	6/8	0.75
6B6	Dissosiyatif	3/4	0.75
6A6	Bipolar	5/7	0.71
6B4	Stresle ilişkili	4/6	0.67
6A7	Depresif	7/12	0.58
6D7	Nörokognitif	2/4	0.50
6C2	Bedensel sıkıntı	1/2	0.50
6A0	Nörogelişimsel	1/3	0.33
6D1	Kişilik	3/9	0.33
6E6	Sekonder	1/3	0.33



Şekil 4.9. Önerilen K4-12B konfigürasyonunun ICD-11 tanı grubu bazında tanı adı düzeyi doğruluğu

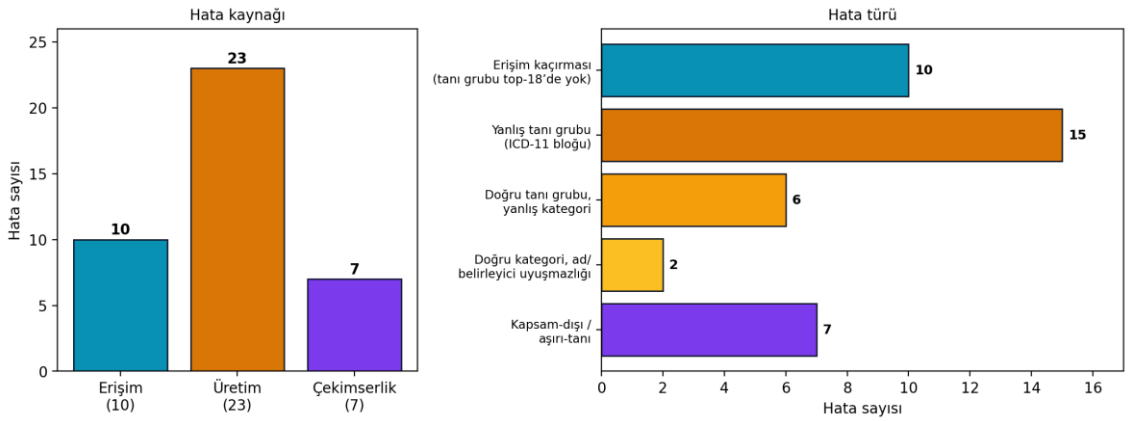
4.6.2. Hata Kaynağı ve Hata Taksonomisi

Önerilen K4-12B'nin 40 hatalı vakası, hata kaynağını yalıtmak amacıyla çözümlenmiştir. Her vaka için, modele gerçekte sunulan tam aday kümesinde (en iyi-

18) tanı standardındaki tanı grubunun (ICD-11 bloğu) bulunup bulunmadığı incelenerek hatalar erişim ve üretim kaynaklı olarak ayrıştırılmıştır (Çizelge 4.9, Şekil 4.10). Kapsam içi 33 hatanın 10'unda tanı standardındaki tanı grubu en iyi-18 aday arasında dahi yer almamıştır (erişim kaçırması); kalan 23'ünde ise doğru tanı grubu getirildiği hâlde model yanlış bir tanı seçmiştir (üretim hatası). Üretim hatalarının çoğu komşu bir tanı grubuna (15 vaka, yanlış tanı grubu) ya da aynı tanı grubu içindeki yanlış kategori koduna (6 vaka) kaymakta; yalnızca 2 vaka doğru kategori koduna ulaşıp ad/belirleyici düzeyinde sapmaktadır. Kalan 7 hata ise kapsam dışı/normatif tabloların (örneğin vaka 46) yapay biçimde tanılanmasından (aşırı tanı) kaynaklanmaktadır.

Çizelge 4.9. Önerilen K4-12B hata taksonomisi ve kaynak ayrımı (toplam 40 hata).

Hata türü	Kaynak	n	Hataların %'si
Erişim kaçırması (tanı grubu en iyi-18 adayda yok)	Erişim	10	25.0
Yanlış tanı grubu	Üretim	15	37.5
Doğru tanı grubu, yanlış kategori kodu	Üretim	6	15.0
Doğru kategori kodu, ad/belirleyici uyuşmazlığı	Üretim	2	5.0
Kapsam dışı / aşırı tanı	Çekimserlik	7	17.5



Şekil 4.10. Önerilen K4-12B hatalarının kaynak (sol) ve tür (sağ) dağılımı (modele sunulan tam en iyi-18 aday kümesi üzerinde)

Bu ayrışma, başarımı artırma çabasının iki yöne yönelmesi gerektiğini göstermektedir. Üstelik daha büyük pay üretim tarafındadır. Erişim kaçırması kapsam içi hataların yalnızca %30'unu (10/33) açıklarken, çoğunluk (23/33) doğru tanı grubu getirildiği hâlde modelin yanlış seçim yapmasından kaynaklanır. Dolayısıyla erişimi güçlendirmek kadar, hatta ondan da çok, getirilen adaylar arasında ayırt ediciliği

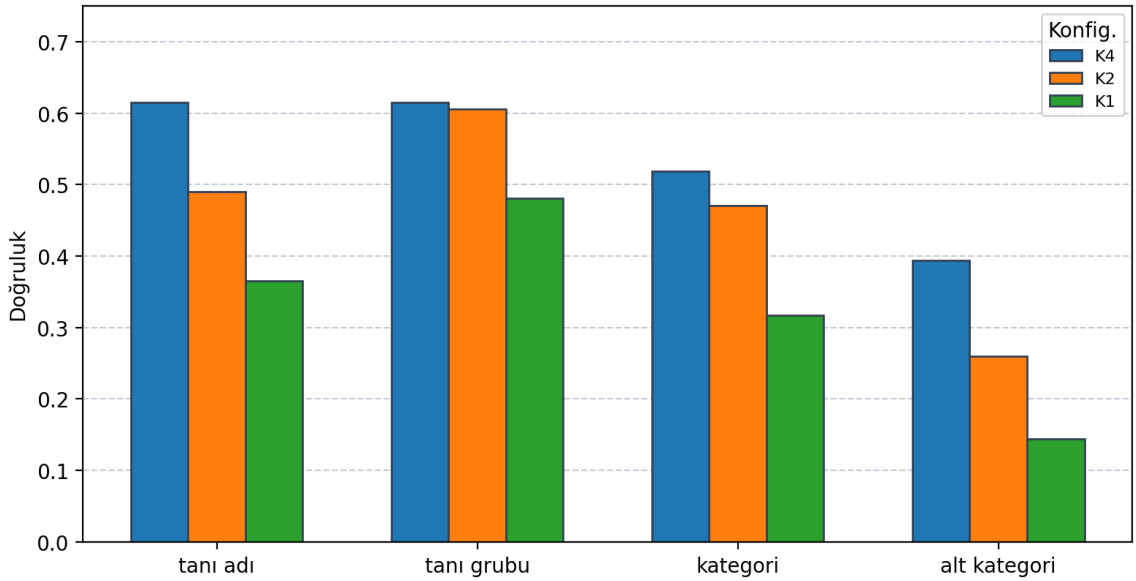
artırmak ve normatif tablolarda kalibre bir çekimserlik kritik önemdedir. Üretim hatalarının ezici çoğunluğunun komşu tanı grubu/kategori sınırında toplanması (21/23), Şekil 4.6'daki karışıklık matrisi bulgusuyla örtüşür.

4.7. Metriğe Bağlılık ve RAG'ın Değeri

Önerilen Gemma 4 12B modelinde kategori düzeyinde 0.519'a karşılık RAG olmadan 0.240; 26B-A4B'de 0.567'ye karşılık 0.413 doğruluk sonuçları elde edilmiştir. Kod düzeyinde (tanı grubu/kategori/alt kategori) RAG, RAG olmadan konfigürasyonu açıkça geçmektedir. Bunun nedeni, RAG'ın geçerli ICD-11 kodlarını getirebilmesidir.

Tanı adı düzeyindeki fark ise modele bağlı olmaktadır. Parametrik bilgisi güçlü 26B-A4B'de RAG olmadan 0.481'den K4'te doğruluk oranı 0.606'e çıkmaktadır. Önerilen 12B'de ise doğruluk oranı RAG olmadan 0.337'den K4'te 0.615'e yükselmiştir. RAG'ın kod düzeyindeki katkısı modelden bağımsız ve tutarlıdır ve tanı adı düzeyindeki katkının büyüklüğü ise modelin parametrik gücüne bağlıdır. LLMind'ın yalnızca tanı adı düzeyinde raporlaması, RAG'ın kod düzeyindeki bu tutarlı katkısını gizlemektedir.

LLMind'ın yalnızca tanı adı düzeyinde raporlama yapması, RAG'ın kod düzeyindeki katkısını gözden gizlemektedir. Çok granülerli değerlendirmenin metodolojik gerekçesini bu nokta oluşturur. Aynı olgu, üç önde gelen konfigürasyonun (K1, K2, K4) dört değerlendirme düzeyindeki karşılaştırmasında da açıkça görülmektedir (Şekil 4.11).



Şekil 4.11. Üç önde gelen konfigürasyonun (K1, K2, K4) dört değerlendirme düzeyinde karşılaştırması

5. TARTIŞMA

Bu tez, açık ağırlıklı modern bir RAG-BDM sisteminin (K4 - Gemma 4 12B) LLMind referansını (0.524) aynı veri kümesi ve aynı değerlendirme zemininde geçtiğini göstermiştir (N=104: 0.615). Bu üstünlük, sabit referansa karşı tek yönlü binom testinde anlamlılığa da ulaşmaktadır (p=0.038). HyDE'nin erişim için kritik bir kaldıraç olması ve önerilen 12B modelinde tanı adı düzeyine de büyük katkı vermesi önemlidir. RAG'ın katkısının kullanılan metriğe bağlı olduğu ve özellikle kod düzeyinde belirginleştiği gözlemlenmiştir. Mevcut sistemlerin çekimserlik ve aşırı tanı ekseninde sistematik olarak başarısız olduğu söylenebilir.

Erişim ablasyonu (kategori düzeyi recall@18: HyDE 0.698, birleştirme 0.750, BM25 0.354) ve HyDE'nin kullanılmadığı üç konfigürasyonda (K6/K7/K8) görülen gerileme, Gao vd. (2023)'in denetimsiz erişim ve alan kaymasında raporladığı HyDE kazançlarıyla doğrudan örtüşmektedir. Altta yatan mekanizma da aynıdır: hasta anlatısı ile ICD-11 kriter metinleri arasındaki sözcük boşluğunun kapatılması. HyDE'nin tanı adı düzeyindeki katkısı modele bağlıdır. Bu bağlamda önerilen 12B modelinde büyük (K3'te 0.452'den K4'te 0.615'e), parametrik bilgisi daha güçlü 26B-A4B'de küçüktür (0.587'den 0.606'ya)ama kod düzeyinde orada da belirginleşmektedir. En sade HyDE-birleştirme konfigürasyonunun en sağlam sonucu vermesi ve K5/K9 zenginleştirmelerinin ek katkı sağlamaması ise, RAG mimarilerinde artan karmaşıklığın erişim gürültüsü getirebileceği gözlemiyle (Lewis vd. 2020; Xiong vd. 2024) uyumludur. Öte yandan hata kaynağı çözümlemesi, kapsam içi hataların yalnızca %30'unun (10/33) erişim kaçırmasından, çoğunluğunsa (23/33) doğru tanı grubu getirildiği hâlde yapılan üretim tarafı yanlış seçimden kaynaklandığını ortaya koyar; yani başlıca darboğaz erişimden çok üretim yani aday kümesinden ayırt etme tarafındadır.

İki hastalı 46. vakada (Pinsky: tanı yok / Quigley: QE84 akut stres tepkisi) ayrı bir çekimserlik testi olarak ele alınmıştır. Önerdiğimiz sistem (K4-12B) ya da benzer bilgi tabanını kullanan çalışmalar çekimser kalamamıştır. Her iki sistem de bozukluk tanısı üretmiştir (K4 “PTSD”, LLMind “Quigley ASD kriterlerini karşılıyor”). Yani normatif stres tepkisini “tanı yok” biçiminde işaretleyememişlerdir. Mevcut sistemlerin aşırı tanı eğilimini somutlaştırmakta ve kalibre edilmiş bir çekimserlik kapısının değerini göstermektedir. Bu, Sarma vd. (2025)'in DSM-5-TR öykülerinde gözlediği “tanısal eşiği uygulayamama” ve aşırı tanı örüntüsünün tekrarıdır. Aynı çalışmanın harici yapılandırılmış bilgiyle pozitif kestirim değerini yükseltmiş olması, kalibre edilmiş bir çekimserlik/OOS kapısı önerimizi desteklemektedir. Dahası bu örüntü, referans çalışmanın kendi sınırlarında kabul ettiği bir zayıflığı doğrulamaktadır. Cremaschi vd. (2025), büyük dil modellerinin eğitim verisindeki yanlılıklar ya da tanı ölçütlerini tam kavrayamama yüzünden aşırı ya da eksik tanı koyabileceğini, komorbidite ve atipik tabloların güçlük yarattığını belirtmiştir. 46. vaka bulgusu, bu eğilimin yapılandırılmış bir çekimserlik testiyle ölçülebileceğini gösterir.

Hataların komşu bozukluk sınırlarında toplanması (özellikle depresif bozukluklar çevresinde) ve tanı grubu düzeyi doğruluğunun (0.615) alt kategori düzeyi doğruluğundan (0.394) yüksek olması, ICD-11'in tanı grupları yapısıyla ve First vd. (2021)'in belgelediği DSM-5 ile ICD-11 arasındaki kriter biçimi farklarıylaaçıklanabilir. Kişilik bozukluklarının ICD-11'de boyutsal bir modele (6D10/6D11) indirgenmesi

(Seemüller vd. 2023), hem tanı standardının inşasında hem de değerlendirmede ICD-11'e özgü kurallar tanımlamayı gerektirmiştir. Örneğin kategorik “histrionic” yanıtı yerine boyutsal “severe personality disorder” yanıtının nasıl puanlanacağı bu kurallarla belirlenmiştir.

RAG'ın kod düzeyindeki katkısı tutarlı biçimde büyüktür. Örneğin RAG eklenmesiyle kategori düzeyinde önerilen 12B'de 0.240’dan 0.519’a, 26B-A4B'de 0.413’ten 0.567’ye yükseliş olmuştur. Tanı adı düzeyindeki katkı ise modele bağlıdır (parametrik bilgisi güçlü 26B-A4B'de küçük, önerilen 12B'de büyük). LLMind'in yalnızca tanı adı düzeyinde raporlaması, RAG'ın kod düzeyindeki bu tutarlı katkısını gizlemektedir. Tanı adı düzeyindeki değerlendirmenin bu sınırları, çok düzeyli raporlamanın gerekçesini güçlendirmektedir.

Çizelge 5.1. LLMind çalışmasının diğer açık kaynak modellerle performansının karşılaştırılması (Cremaschi vd., 2025)

Model	Doğruluk	Kesinlik	Duyarlılık	F1	Tanı yok
LLMind — (Gemma2-27B) (Alıntı)	0.524	0.552	0.524	0.524	0
Gemma2-9B — (Alıntı)	0.272	0.309	0.272	0.276	2
Mistral Nemo 12B — (Alıntı)	0.204	0.244	0.204	0.192	0
Phi-3 medium 14B — (Alıntı)	0.175	0.231	0.175	0.190	35
Llama3-8B — (Alıntı)	0.165	0.215	0.165	0.175	4
MentaLLaMA — (Alıntı)	0.068	0.127	0.067	0.067	25
BioMistral — (Alıntı)	0.058	0.069	0.058	0.061	40
MedLlama2 — (Alıntı)	0.049	0.098	0.048	0.054	18

Çizelge 5.1, önerilen sistemin Cremaschi vd. (2025) çalışmasında doğrudan test edilen tüm açık kaynak modellerle birlikte konumlandırılmasını sağlamaktadır. Tabloda görüldüğü üzere LLMind (Gemma2-27B) 0.524 doğruluk ve 0.524 F1 değeriyle en yüksek açık kaynak performansını vermekte, geri kalan yedi açık kaynak modelin tamamı 0.30 eşiğinin belirgin biçimde altında kalmaktadır. İkinci sıradaki Gemma2-9B bile yalnızca 0.272'ye ulaşırken, tıbbi alan ince ayarlı MentaLLaMA (0.068), BioMistral (0.058) ve MedLlama2 (0.049) 0.07'nin altında performans göstermektedir. Tanı Yok sütunu, Phi-3 medium 14B'nin 35, BioMistral'ın 40 ve MentaLLaMA'nın 25 vakada çekimser kaldığını; buna karşılık LLMind, Mistral Nemo 12B ve Llama3-8B'nin ise hiçbir vakada çekimser bırakmadığını ortaya koymaktadır. Bu karşılaştırma, referans çalışmada sınanan açık kaynak modeller arasında LLMind'in net bir üstünlüğe sahip olduğunu ve çalışmamızda önerilen K4-12B'nin (0.615) bu referans değeri de geçtiğini doğrulamaktadır. Ayrıca tablo, model ölçeği ile başarım arasında doğrudan ilişki olmadığını göstermektedir. 14B parametreli Phi-3 medium (0.175) 9B parametreli

Gemma2'nin (0.272) altında kalırken, 8B'lik Llama3 (0.165) ile 12B'lik Mistral Nemo (0.204) benzer bir alt bantta yer almaktadır. Bu örüntü, RAG ve erişim temelli kurulumun parametrik bilginin yerini ne ölçüde alabileceğine işaret etmektedir. Yalnızca talimat izleme ve kodlama üzerine ince ayarlanmış modeller, psikiyatrik anlatı-kriter eşleştirmesinde büyük genel amaçlı modellerin bile gerisinde kalabilmektedir.

Sistemin açık ağırlıklı modellerle yerel donanımda çalışabilmesi, klinik dağıtım açısından bazı avantajlar sunmaktadır. Hasta verisi kurum dışına çıkmadan yerelde işlenebilir. Sistem bir bulut hizmetinin maliyetine, kesintilerine veya sürüm değişikliklerine bağımlı değildir. Bilgi tabanı, ICD-11 güncellendiğinde modeli yeniden eğitmeye gerek kalmadan tazelenebilir. Bununla birlikte 0.615'lik otomatik doğruluk, dört uzmanın elle ulaştığı 0.761'lik insan tavanının altında kaldığından ve sistem gözlenen normatif vakalarda aşırı tanıya yöneldiğinden, araç özerk bir tanı koyucu olarak değil, klinisyenin denetiminde bir önceliklendirme, ayırıcı tanı hatırlatma ve eğitim desteği aracı olarak konumlandırılmalıdır.

5.1. Metodolojik Katkılar

Üç yöntemli doğrulama ve uyuşmazlıkların WHO ICD-11/CDDR metni hakem alınarak elle tahkimi, ISO/TR 12300 eşleme kalitesi ilkeleri gözetilerek yürütülmüştür. Bu yaklaşım, LLMind'in doğrulanmamış tanı standardına kıyasla somut bir metodolojik sıkılaştırma sağlamakta ve eşleme eşdeğerliğini ikili bir doğru yanlış kararına indirgemekten kaçınmaktadır.

LLMind'in vaka bazlı çıktıları yeniden puanlanarak otomatik protokolün 54/103 = 0.524 değeri birebir yeniden üretilmiş, aşırı cömert puanlanan iki eşleşme (jenerik tanı grubu belirteci ve indirgenmiş kişilik bozukluğu sinonimi) tespit edilerek kalibre edilmiştir. Bu yeniden üretim, değerlendirme aracının referans protokolle hizalı olduğunu kanıtlamaktadır. Çalışmanın kendi başarımı ise LLMind ile aynı zeminde, tam korpus (N=104) üzerinde raporlanmıştır. LLMind'in yayımladığı 54/103 değeri yalnızca bu kalibrasyon doğrulaması için kullanılmıştır.

Bu tez, tanısal bilgiyi modelin ağırlıklarına gömen ince ayar yerine, ince ayar gerektirmeyen erişim temelli bir yaklaşımı benimsemektedir. Bu tercihin gerekçesi yalnızca hesaplama maliyeti değildir. Erişim temelli kurgu, bilgi tabanını modelden ayrı tuttuğundan ICD-11 güncellemelerine uyarlanabilirliği ve kaynak izlenebilirliğini yani her tanının hangi kriterlere dayandığını korumaktadır. Ayrıca yalnızca 104 vakalık küçük bir kümede ince ayarın taşıdığı aşırı uyum riskinden kaçınılmaktadır. Türkçe yazında Bhat (2025)'in LoRA temelli ince ayar yaklaşımı ile bu çalışmanın erişim temelli kurgusu, söz konusu tasarım ekseninin iki ucunu örneklemektedir.

5.2. Sınırlamalar

104 vakalık küme tek bir korpustan oluşmakta ve bootstrap güven aralıkları geniş kalmaktadır. Önerilen K4-12B, sabit referansa (0.524) karşı tek yönlü binom testinde anlamlılığa ulaşmaktadır (64/104, p=0.038). Ancak bu test 0.524'ü sabit kabul ettiğinden zayıftır ve anlamlılık tek bir vakaya (eşik 64/104) bağlıdır. Nitekim gözlenen 7.7 puanlık üstünlük, N=104'te McNemar testinin minimum saptanabilir etkisinin (≈16

puan) altında kaldığından anlamlılığa ulaşamamaktadır. Bu durum farkın yokluğuna değil, tek test setinin istatistiksel güç sınırına işaret etmektedir. Kanıt bütünü ise tutarlıdır ve sabit referansa karşı binom anlamlılığı ($p=0.038$), RAG katkısının her iki modelde de eşleştirilmiş anlamlılığı ($p<0.001$; $p=0.029$) ve dört granülerlik düzeyinin tamamındaki nokta tahmini üstünlüğü aynı yönü göstermektedir.

OOS eşlemeler ve boyutsal kişilik bozuklukları kaçınılmaz biçimde yargı kararları içermektedir. Mekanik bir eşleme tablosu (crosswalk) bunları çözememektedir (First vd. 2021).

6. SONUÇLAR

Bu tez, bir klinik vaka anlatısından hastanın birincil ICD-11 ruhsal bozukluk tanısını öngören, erişim destekli üretim temelli bir büyük dil modeli (RAG-BDM) sistemi geliştirmiş ve bu sistemi, alandaki güncel referans çalışma LLMind ile paralel veri kümesi ve aynı ICD-11 zemini üzerinde denetlenebilir biçimde karşılaştırmıştır. Bu kapanış bölümü, çalışmanın başlıca bulgularını toparlamakta, katkılarını özetlemekte ve tezin merkezi savını ortaya koymaktadır.

Çalışmanın ilk bulgusu, modern, açık ağırlıklı bir RAG-BDM sisteminin referans değeri anlamlı biçimde geçebildiğidir. Önerilen sistem (K4 + Gemma 4 12B), HyDE birleştirmeli erişim ve kriter temelli üretim sayesinde 0.615 (N=104) tanı adı düzeyi doğruluğa ulaşmıştır. LLMind çalışmasının 0.524 doğruluğunu hem nokta tahmininde geçmiş hem de sabit referansa karşı tek yönlü binom testinde anlamlılığa ulaşan tek sistem olmuştur ($p=0.038$). RAG'ın katkısı, RAG içermeyen konfigürasyona karşı eşleştirilmiş testte de güçlüdür ve 12B'de çok belirgindir (McNemar $p<0.001$). Karışık uzman 26B-A4B modeli ise 0.606 ile benzer bir doğruluğa yaklaşık üçte bir etkin parametreyle ulaşarak verimli bir alternatif oluşturmaktadır.

Dokuz konfigürasyonlu ablasyon, hangi bileşenlerin tanısal başarıya katkı verdiğini ayrıştırmıştır. HyDE'li birleştirme, erişimin kategori düzeyi recall@18 değerini BM25'in 0.354'ünden 0.750'ye taşımıştır (HyDE tek başına 0.698) ve erişim başarımına açık ara en büyük katkıyı HyDE vermiştir. Tanı adı düzeyindeki katkısı ise modele bağlıdır ve önerilen Gemma 4 12B modelinde K3 konfigürasyonunda 0.452 iken K4 konfigürasyonunda 0.615'e yükselmektedir. Bileşen ablasyonunun yürütüldüğü 26B-A4B modelinde ise küçülmekle birlikte (0.587'den 0.606'ya) kod düzeyinde belirginleşmektedir. En yalın HyDE birleştirme konfigürasyonu (K4), HyDE'yi devre dışı bırakan ya da üzerine ek katman ekleyen konfigürasyonların tümünü yine de geçmektedir. Yani artan karmaşıklık her zaman daha iyi sonuç vermemektedir. Hata çözümlemesi ayrıca kapsam içi hataların çoğunluğunun (23/33) erişim kaçırmasından değil, doğru adaylar getirildiği hâlde üretim aşamasındaki yanlış seçimden kaynaklandığını göstermiştir. Bu bulgu, sistemin başlıca darboğazının erişim değil aday kümesinden ayırt etme olduğuna işaret etmektedir.

RAG'ın tanısal değeri kullanılan metriğe bağlıdır. Kod düzeyinde (tanı grubu/kategori/alt kategori) RAG'ın katkısı tutarlı biçimde büyüktür. Tanı adı düzeyindeki katkısı ise modele bağlıdır (parametrik bilgisi güçlü 26B-A4B'de daralır, önerilen 12B'de büyük kalır). Bu durum, referans çalışmanın yalnızca tanı adı düzeyinde raporlama yapmasının RAG'ın kod düzeyindeki tutarlı katkısını gizlediğini ortaya koymaktadır ve çok düzeyli bir değerlendirmenin neden gerekli olduğunu göstermektedir.

Çalışmanın katkıları, salt bir doğruluk artışının ötesinde bir değerlendirme disiplini ortaya koymasında yatmaktadır. Bu katkılar üç yöntemli doğrulama ve uyuşmazlıkların elle tahkimiyle oluşturulmuş bir tanı standardı, referans değerin birebir (54/103=0.524) kalibre edilmiş, tekrarlanabilir bir yeniden uygulaması, yerel donanımda çalıştırılabilir modern bir RAG hattının referans değerini (önerilen modelde anlamlı biçimde) geçmesi, HyDE'nin belirleyiciliğini ve ek zenginleştirmelerin sınırlarını gösteren kapsamlı bir ablasyon çalışması, RAG'ın değerinin metriğe

bağlılığının ve mevcut sistemlerin aşırı tanı eğiliminin ortaya konması olarak açıklanabilir.

Önerilen sistem, nicelenmiş açık ağırlıklı modellerle düşük işlemci maliyeti ile çalışabildiğinden hasta verisinin kurum dışına çıkarılmasını gerektirmemekte ve harici API maliyeti olmadan işletilebilmektedir. Bu özellik, veri mahremiyeti gereksinimlerinin katı olduğu sağlık kurumları ile donanım kaynağı kısıtlı merkezler için pratik bir avantaj sunmaktadır. Geliştirilen web tabanlı karar destek arayüzü, birincil tanı önerisini gerekçesi, ayırıcı tanıları ve dayandığı ICD-11 kriter metinleriyle birlikte sunarak klinisyenin öneriyi denetleyebilmesine olanak tanımaktadır. Bilgi tabanının yapılandırılmış ve modelden bağımsız tutulması, aynı hattın ICD-11'in diğer bölümlerine veya güncellenen sınıflandırma sürümlerine model yeniden eğitimi olmaksızın taşınabilmesinin de önünü açmaktadır.

7. KAYNAKLAR

Abd-alrazaq, A., et al. 2023. Large language models in medical education: opportunities, challenges, and future directions. *JMIR Medical Education*, 9(1): e48291.

Barnhill, J.W. (Ed.) 2023. DSM-5-TR Clinical Cases. American Psychiatric Association Publishing, Washington, DC.

Bayraktar, M. 2025. Ruh sağlığı bilişiminde klinik uygulamalara yönelik güncel yaklaşımlar ve yenilikler: kavramsal ve literatür incelemesi. Yüksek lisans tezi, Üsküdar Üniversitesi, İstanbul, 70 s.

Bayram, M.A., Diri, B. and Yıldırım, S. 2025. Healthcare-focused Turkish medical BDM: training on real patient–doctor question–answer data for enhanced medical insight. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24(11): 1-13.

Bhat, A.R. 2025. Adapting large language models via multi-task learning for mental health problem, emotion, and sentiment detection. Master thesis, Kadir Has University, İstanbul, 75 p.

Bouguettaya, A., Stuart, E.M. and Aboujaoude, E. 2025. Racial bias in AI-mediated psychiatric diagnosis and treatment: a qualitative comparison of four large language models. *NPJ Digital Medicine*, 8: 332.

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

Cormack, G.V., Clarke, C.L.A. and Buettcher, S. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 758-759.

Cremaschi, M., Ditolve, D., Curcio, C., Panzeri, A., Spoto, A. and Maurino, A. 2025. Decoding the mind: a RAG-BDM on ICD-11 for decision support in psychology. *Expert Systems with Applications*, 279: 127191.

Dietterich, T.G. 1998. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7): 1895-1923.

Efron, B. 1992. Bootstrap methods: another look at the jackknife. *In Breakthroughs in statistics: Methodology and distribution*, 569-593.

First, M.B. 2009. Harmonisation of ICD-11 and DSM-V: opportunities and challenges. *British Journal of Psychiatry*, 195(5): 382-390.

First, M.B. et al. 2021. An organization- and category-level comparison of diagnostic requirements for mental disorders in ICD-11 and DSM-5. *World Psychiatry*, 20(1): 34-51.

Gaebel, W. et al. 2020. Accuracy of diagnostic classification and clinical utility assessment of ICD-11 compared to ICD-10 in 10 mental disorders: findings from a web-based field study. *European Archives of Psychiatry and Clinical Neuroscience*, 270(3): 281-289.

Gao, L., Ma, X., Lin, J. and Callan, J. 2023. Precise zero-shot dense retrieval without relevance labels (HyDE). Proceedings of ACL 2023 (Long Papers), 1762-1777.

Ge, Z. et al. 2025. Survey and experiments on mental disorder detection via social media: from large language models and RAG to agents. arXiv:2504.02800.

Gemma Team. 2024. Gemma: open models based on Gemini research and technology. arXiv:2403.08295

Gemma Team. 2026. Gemma 4 technical report. arXiv:2607.02770.

Gu, J. et al. 2024. A survey on BDM-as-a-judge. arXiv:2411.15594.

Hicks, S.A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M.A., Halvorsen, P. and Parasa, S. 2022. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12: 5979.

International Organization for Standardization. 2014. ISO/TR 12300:2014 — Health informatics: principles of mapping between terminological systems. ISO, Geneva.

Japkowicz, N. and Shah, M. 2011. Evaluating learning algorithms: a classification perspective. Cambridge University Press, Cambridge.

Jurman, G., Riccadonna, S. and Furlanello, C. 2012. A comparison of MCC and CEN error measures in multi-class prediction. *PLoS ONE*, 7(8): e41882.

Kapur, S., Phillips, A.G. and Insel, T.R. 2012. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? *Molecular Psychiatry*, 17(12): 1174-1179.

Karakurt, E. 2026. Büyük dil modelleri (LLM) ve almayla arttırılmış üretim (RAG) tabanlı yaklaşım ile insan kaynaklarında belgelerden bilgi çıkarımı. Yüksek lisans tezi, İstanbul Kültür Üniversitesi, İstanbul, 56 s.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D. and Yih, W. 2020. Dense passage retrieval for open-domain question answering. Proceedings of EMNLP 2020, 6769-6781.

Kessler, R.C., Chiu, W.T., Demler, O. and Walters, E.E. 2005. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, 62(6): 617-627.

Kestel, D. and Reed, G.M. 2024. Global launch of the ICD-11 Clinical Descriptions and Diagnostic Requirements (CDDR). *World Psychiatry*, 23(3): 310-311.

Kim, J., Leonte, K.G., Chen, M.L., Torous, J.B., Linos, E., Pinto, A. and Rodriguez, C.I. 2024. Large language models outperform mental and medical health care professionals in identifying obsessive-compulsive disorder. *NPJ Digital Medicine*, 7: 193.

Kim, Y. et al. 2025. Medical hallucinations in foundation models and their impact on healthcare. arXiv:2503.05777.

Kottner, J., Audigé, L., Brorson, S., Donner, A., Gajewski, B.J., Hróbjartsson, A., Roberts, C., Shoukri, M. and Streiner, D.L. 2011. Guidelines for reporting

reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology*, 64(1): 96-106.

Krippendorff, K. 2004. Content Analysis: An Introduction to Its Methodology (2. baskı). Sage, Thousand Oaks.

Labrak, Y., Bazoge, A., Morin, E., Gourraud, P., Rouvier, M. and Dufour, R. 2024. BioMistral: a collection of open-source pretrained large language models for medical domains. Findings of ACL 2024.

Lewis, P. et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems (NeurIPS) 33. 9459-9474.

Lin, L.I.-K. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1): 255-268.

McNemar, Q. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2): 153-157.

Nogueira, R. and Cho, K. 2019. Passage re-ranking with BERT. arXiv:1901.04085.

Oğul, İ.Ü., Soygazi, F. ve Bostanoğlu, B.E. 2025. TurkMedNLI: a Turkish medical natural language inference dataset through large language model based translation. *PeerJ Computer Science*, 11: e2662.

Omar, M., Soffer, S., Charney, A.W., Landi, I., Nadkarni, G.N. and Klang, E. 2024. Applications of large language models in psychiatry: a systematic review. *Frontiers in Psychiatry*, 15: 1422807.

Reed, G.M. et al. 2018. The ICD-11 developmental field study of reliability of diagnoses of high-burden mental disorders: results among adult patients in mental health settings of 13 countries. *World Psychiatry*, 17(2): 174-186.

Regier, D.A., Narrow, W.E., Clarke, D.E., Kraemer, H.C., Kuramoto, S.J., Kuhl, E.A. and Kupfer, D.J. 2013. DSM-5 field trials in the United States and Canada, Part II: test-retest reliability of selected categorical diagnoses. *American Journal of Psychiatry*, 170(1): 59-70.

Reimers, N. and Gurevych, I. 2019. Sentence-BERT: sentence embeddings using Siamese BERT-networks. Proceedings of EMNLP-IJCNLP 2019, 3982-3992.

Robertson, S. and Zaragoza, H. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4): 333-389.

Sarma, K.V., Hanss, K.E., Halls, A.J.M., Krystal, A., Becker, D.F., Glowinski, A.L. and Butte, A.J. 2025. Integrating expert knowledge into large language models improves performance for psychiatric reasoning and diagnosis. *Psychiatry Research*, 116844.

Sellergren, A. et al. 2025. MedGemma technical report. arXiv:2507.05201.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q.V., Hinton, G.E. and Dean, J. 2017. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. Proceedings of ICLR 2017, arXiv preprint arXiv:1701.06538.

Shazeer, N. 2020. GLU variants improve transformer. arXiv:2002.05202.

Singhal, K. et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972): 172-180.

Singhal, K. et al. 2025. Toward expert-level medical question answering with large language models (Med-PaLM 2). *Nature Medicine*, 31(3): 943-950.

Sokolova, M. and Lapalme, G. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4): 427-437.

Seemüller, F. 2023. ICD-11 and mental disorders: important changes, controversies, and future directions. *BMC Psychiatry*, 23(1): 698.

Su, J., Lu, Y., Pan, S., Ahmed, M., Wen, B. and Liu, Y. 2024. RoFormer: enhanced transformer with rotary position embedding. *Neurocomputing*, 568, 127063.

Türk, M.N., Akın, R., Şahin, D.Ö. ve Demirci, S. 2026. Büyük dil modellerinin ruh sağlığı uygulamaları için performans ve tutarlılık analizi. *Black Sea Journal of Engineering and Science, 9(3): 1338-1349.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. 2017. Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS) 30, 5998-6008.

Wang, L., Yang, N. and Wei, F. 2023. Query2doc: query expansion with large language models. Proceedings of EMNLP 2023, 9414-9423.

Wei, J. et al. 2022. Emergent abilities of large language models. *Emergent abilities of large language models*, arXiv preprint arXiv:2206.07682.

World Health Organization. 2022. WHO's new International Classification of Diseases (ICD-11) comes into effect. WHO, Geneva.

World Health Organization. 2024. Clinical Descriptions and Diagnostic Requirements for ICD-11 Mental, Behavioural and Neurodevelopmental Disorders (CDDR). WHO, Geneva.

Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D. and Nie, J.-Y. 2024. C-Pack: packed resources for general Chinese embeddings (BGE / BAAI General Embedding). Proceedings of SIGIR 2024, 641-649.

Xiong, G., Jin, Q., Lu, Z. and Zhang, A. 2024. Benchmarking retrieval-augmented generation for medicine (MedRAG / MIRAGE). Findings of ACL 2024, 6233-6251.

Yang, K., Zhang, T., Ji, S., Xie, Q., Kuang, Z., Ananiadou, S. and Huang, J. 2024. MentaLLaMA: interpretable mental health analysis on social media with large language models. Proceedings of the ACM Web Conference 2024 (WWW '24), 4489-4500.

Zhang, B. and Sennrich, R. 2019. Root mean square layer normalization. Advances in Neural Information Processing Systems (NeurIPS) 32.

Zheng, L. et al. 2023. Judging BDM-as-a-judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Datasets and Benchmarks Track, arXiv:2306.05685v4.

8. EKLER

EK-1. Bir Ruhsal Bozukluğun BT'deki JSON yapısı (örnek; 6A70.0, Single episode depressive disorder, mild)


```
{
 "code": "6A70.0",
 "title": "Single episode depressive disorder, mild",
 "block": "6A7",
 "parent_title": "Depressive disorders",
 "synonyms": [
  "major depressive disorder, single episode, mild degree",
  "mild major depression disorder, single episode"
 ],
 "definition": "Single episode depressive disorder, mild, is diagnosed
  when the definitional requirements of a Depressive episode are met
  and the episode is of mild severity ...",
 "criteria": {
  "essential_features": "- Presence or history of a single Depressive
   Episode. - No history of Manic, Mixed, or Hypomanic Episodes ...",
  "boundary_with_normality": "...",
  "differential_diagnosis": "..."
 },
 "is_leaf": true
}
```

ÖZGEÇMİŞ

Egehan ÇETİN

ÖĞRENİM BİLGİLERİ

Doktora 2020-2026	Akdeniz Üniversitesi Fen Bilimleri Enstitüsü, Elektrik-Elektronik Mühendisliği Anabilim Dalı, Antalya
Yüksek Lisans 2017-2020	Akdeniz Üniversitesi Fen Bilimleri Enstitüsü, Elektrik-Elektronik Mühendisliği Anabilim Dalı, Antalya
Lisans 2013-2017	Akdeniz Üniversitesi Mühendislik Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, Antalya

MESLEKİ VE İDARİ GÖREVLER

Merkez Müdür Yardımcısı 2024-Devam Ediyor	Burdur Mehmet Akif Ersoy Üniversitesi Uzaktan Eğitim Uygulama ve Araştırma Merkezi
Öğretim Görevlisi 2021-Devam ediyor	Burdur Mehmet Akif Ersoy Üniversitesi Uzaktan Eğitim Uygulama ve Araştırma Merkezi

ESERLER

Uluslararası hakemli dergilerde yayımlanan makaleler

1- Çetin, E., Bilgin, S., Bilgin, G. (2024). A novel wearable ERP-based BCI approach to explicate hunger necessity. *Neuroscience Letters*, 818, 137573. Doi: 10.1016/j.neulet.2023.137573

2- Çetin, E., Bilgin, G., Bilgin, S., Gömceli, Y. B., Kayıkçi, A. M. (2020). Investigation of hunger and satiety status during eyes open and closed using EEG signals. *Journal of Intelligent Systems with Applications*, 3(1), 35-39. Doi: 10.64799/joiswa.v3.i1.8

3- Cetin, E., Bilgin, S. (2019). Investigating effects of force and pressure centre signals on stabilogram analysis. *IET Science, Measurement & Technology*, 13(9), 1305-1310. Doi: 10.1049/iet-smt.2019.0078

Uluslararası bilimsel toplantılarda sunulan ve bildiri kitaplarında basılan bildiriler

1- Cetin, E., Bilgin, S., Oral, O. (2019). A Walking and balance analysis based on pedobarography. In The International Conference on Artificial Intelligence and Applied Mathematics in Engineering, 906-913. Cham: Springer International Publishing. Doi: 10.1007/978-3-030-36178-5_80

2- Ak M.U., Bilgin, S., Oral, O., Cetin, E., Carlak, H.F., Derin, N., Derin, A.T. (2019) Median and Maximum Frequency Analysis of Vibration Signals on Human Face. International Conference on Research in Engineering and Techology, Barcelona, Spain.

Ulusal bilimsel toplantılarda sunulan ve bildiri kitaplarında basılan bildiriler

1- Çetin, E, Bilgin, S. (2019). Enerji Verimliliğinde ve Yenilenebilir Enerjide Yapay Zeka Uygulamaları. 10.Yenilenebilir Enerji Kaynakları Sempozyumu ve Sergisi (YEKSEM 2019) (Tam Metin Bildiri/Sözlü Sunum)

2- AK, M. Ü., Bilgin, S., Oral, O., Çetin, E., Carlak, H. F., Derin, A. T., Derin, N. (2019). A new system design for measuring vibration effect in live tissues. In 2019 Medical Technologies Congress (TIPTEKNO) (pp. 1-4). IEEE.