

Büyük Dil
Modelleri

Large
Language
Models

Dil Modelleri

- n-gram dil modelleri:
 - ❖ Sıralanmış olan sözcüklere olasılıklar atar.
 - ❖ Sonraki olası sözcüklerden oluşan örneklerle metin oluşturulur.
 - ❖ Çok sayıda metinden hesaplanan sayılarla eğitilir.
- Büyük dil modelleri benzer ama farklıdır;
 - ❖ Kelime dizilerine olasılıklar atar
 - ❖ Olası sonraki kelimeleri örnekleyerek metin üretir
 - ❖ **Bir sonraki kelimeyi tahmin etmeyi öğrenerek eğitilir**

Büyük Dil Modelleri

- ❑ Çok miktarda metinden dil ve konu ile ilgili hakkında bilgi çıkarılması *pretrained* olarak adlandırılır.
- ❑ Bu önceden eğitilmiş verilerin sonuçları büyük dil modellerini oluşturur.
- ❑ Özetleme, makine çevirisi, soru cevaplama veya sohbet robotları gibi metin üretilmesi gereken problemlerde dönüştürücü bir rol üstlenir.

Dönüştürücüye Geçiş

- ❑ Uzun Kısa Süreli Bellek (Long Short-Term Memory, LSTM), bir RNN'nin uzun cümleleri işleyememe ve gradyan kaybolması gibi bazı sorunlarını çözmek amacıyla geliştirilmiştir.
- ❑ Her bir LSTM biriminin her bir token (andaç /belirteç) veya kelime için sıralı olarak ağırlık (weight) değerini tahmin etmesi gerekir.
 - ❖ O nedenle bu model hesaplama açısından pahalıdır ve bu ağırlıkları belirleyecek eğitim süreleri çok uzundur.

Dönüştürücüye Geçiş

- ❑ Dönüştürücüler kullanılmadan önce, dilin işlenmesi sırasında uzun cümlelerin anlaşılmasında zorlanılıyor ya da uzun cümlelerin uygun bağlamının belirlenmesinde büyük sorunlar yaşanıyordu.
- ❑ Tekrarlayan (yenilemeli /rekürsif) sinir ağları (Recurrent Neural Network, RNN), geri yayılım sırasında gradyan kaybolması (gradient vanishing) ve gradyan patlaması (gradient explosion) gibi sorunlarla karşılaşılıyordu.
 - ❖ Bu, önceki girdilerle ilgili olarak zamanla bellekte yer kalmaması nedeni problemlerle karşılaşılması ve uzun cümleleri işlemenin güçlüğü idi
 - ✓ Yani, RNN uzun bir diziyi işlediğinde kademeli olarak dizideki ilk girdileri unutmasıdır.
 - ❖ Bazı durumlarda katmanlar büyük ağırlık güncellemeleri alır; böylece gradyanlar gitgide büyür ve algoritma optimum noktadan uzaklaşır.
 - ✓ “aşırı büyüyen gradyanlar” ya da gradyan patlaması problemi denir.

Büyük Dil Modellerinde Dönüştürücüler (Transformers)

- ❑ Dönüştürücü, bir diziyi başka bir diziye dönüştürmek için geliştirilmiş bir sinir ağı mimarisidir.
- ❑ Dil modelleme alanında günümüzde yaygın kullanılan bir dönüştürücünün (transformer) bileşenleri:
 - ❖ (multi-head attention block) çok başlı dikkati içeren dönüştürücü bloğu,
 - ❖ (the language modeling head) dil modelleme başlığı ve
 - ❖ (the positional encoding of the input) girdinin konumsal kodlamasıdır.
- ❑ Tüm giriş verileri (yani metin) paralel olarak işlenir.
 - ❖ RNN ve LSTM modelleri, metni sıralı olarak işler. Bir andacın çıktısı, bir sonraki andacın girişini oluşturur.
 - ❖ Bu, büyük veri setleri ile çalışırken işlem sürecini çok uzatır.
 - ❖ Transformer ile tüm kelimeler eşzamanlı olarak işlenebildiği için büyük veri setleri çok daha hızlı işlenir.

Transformer Mimarisi

- ❑ Günümüzün güçlü dil modellerinin arkasındaki güçtür.
- ❑ Doğal dil işleme alanında devrim niteliğinde bir gelişmedir.
- ❑ Önceki dil işleme yöntemleri Recurrent Neural Networks (RNN) ve Long Short-Term Memory (LSTM)'nin önüne geçmiştir.
- ❑ Transformer, dil modelleme görevlerinde çok daha hızlı, verimli ve etkili çözümler vermektedir.
 - ❖ Mimarının self-attention (öz-dikkat) mekanizması bu başarıda önemli rol oynar
 - ❖ Metinleri işlenirken metindeki her kelime ya da andaç (token) diğerleriyle bağlantılı olarak değerlendirilmektedir.

Self attention (öz dikkat) Mekanizması

- ❑ Bu özellikle transformer modelleri metinleri çok daha akıllı şekilde anlayacaktır
- ❑ Metin okunurken her kelime tek başına alınmamakta; o kelimenin diğer kelimelerle ilişkileri de araştırılmaktadır.
- ❑ Örneğin, bir metnin içinde bir kelimeyi anlamak için sadece o kelimenin etrafındaki kelimelere değil, tüm metin araştırılır.
- ❑ Model, bir kelimeyi anlamaya çalışırken, kelimenin metnin bütünündeki kelimelerle olan ilişkisini araştırır.
 - ❖ Sadece belli bir aralıktaki sözcükleri araştırılmaz.
 - ❖ Böylece, metnin bağlamı ve anlamı çok daha derinlemesine ve doğru olarak kavranır.
 - ✓ Çünkü sadece araştırılan herhangi iki kelime arasındaki ilişkiler değil, bunların tüm metin ile olan ilişkileri araştırılır.

ÖZETLE:

Bir sıralanıştaki her t andacının (belirtecin) diğer tüm andaçlarla olan ilişkisi öğrenilir.

Bu mekanizma, dilin bağlamını anlamak için her kelimenin diğer kelimelere olan etkisini değerlendirir

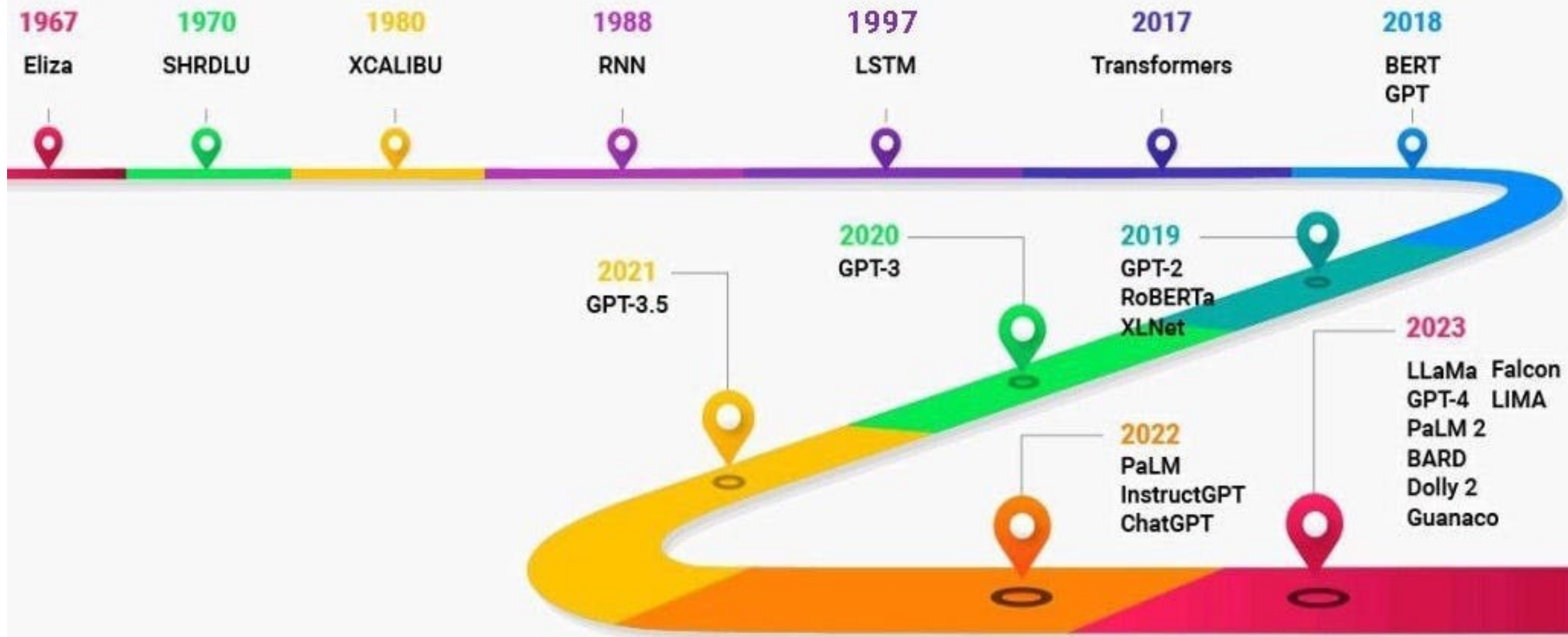
Self –attention Bileşenleri

- Bir andacın (sözcüğün) çıktısının hesaplanmasında üç ana bileşen kullanılır:
 - ❖ **Query (Sorgu):** Bir andacın diğer andaçlarla ta nasıl ilişkili olduğunu anlamak için kullanılan bir tür işaretçidir.
 - ❖ **Key (Anahtar) :** Her bir andacın (token) sorgularla ne kadar ilişkili olduğunu belirleyen bir başka işaretçidir
 - ❖ **Value (Değer):** andaçlar arasındaki ilişkilerin içeriğini temsil eden bir vektördür.
- Bu üç bileşen bir arada çalışarak modelin her bir andacın diğerleriyle nasıl etkileşimde bulunduğunu anlamayı sağlar ve böylece doğru sonuç üretilir.
- **Sonuç olarak:** Her şeyin birbirine nasıl bağlandığını çözmek için bu üç bileşen birbirine yardımcı olur.
- Bu bileşenler, her andaç için hesaplanan dikkat skorları ile birleştirilir ve modelin her andaç için daha anlamlı bir temsil oluşturmasını sağlar.

Koşullu Üretim (Conditional Generation)

- ❑ Gerçekleştirilen görevler (tasks) koşullu üretimin tüm durumlarının tanımlanmasıdır.
- ❑ Koşullu üretim, giriş olarak alınan bir metni parçasına bağlı olarak metin üretme görevidir.
- ❑ Büyük dil modeline genellikle istem (prompt) adı verilen bir giriş metni parçası verilir.
- ❑ Dil modelinde istemde koşullu olarak andaçlar üzerinde (token by token) metin üretmeye devam edilir.
- ❑ Dönüştürücülerin uzun bağlama (binlerce andaç) sahip olması, onları koşullu üretim için güçlü hale getirir.
 - ❖ Bunlar istem (prompt) metnine, yani çok geriye bakabilmektedir.

Evolution of Large Language Models



Dönüştürücüler arasındaki Farklılıklar

- ❑ **Encoder only:** Sınıflandırma ya da duygu analizi gerçekleştirmek gibi dili anlamaya uygun modellerdir.
 - ❖ Örnek BERT (Bidirectional Encoder Representations from Transformers).
- ❑ **Decoder only:** Dilin ve içeriğini oldukça etkili bir şekilde oluşturur (generate)

❖ Örnekleri hikaye yazma, blog oluşturma olarak verilebilir.

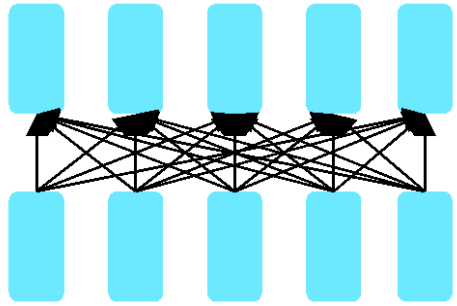
Bu mimari için model örneği GPT-3 (Generative Pretrained Transformer 3).

- ❑ **Encoder-decoder:** Dönüştürücü mimarisindeki encoder ve decoder bileşenleri birlikte kullanılır. Hem anlama (understanding) hem de içerik oluşturma (generate content) gerçekleştirilir.

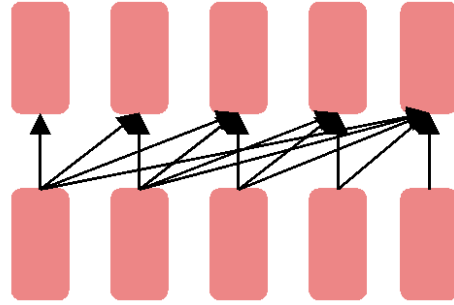
❖ Örnekleri çeviri ve özetleme olabilir.

Encoder-decoder mimarisi modeline örnek T5 (Text-to-Text Transformer).

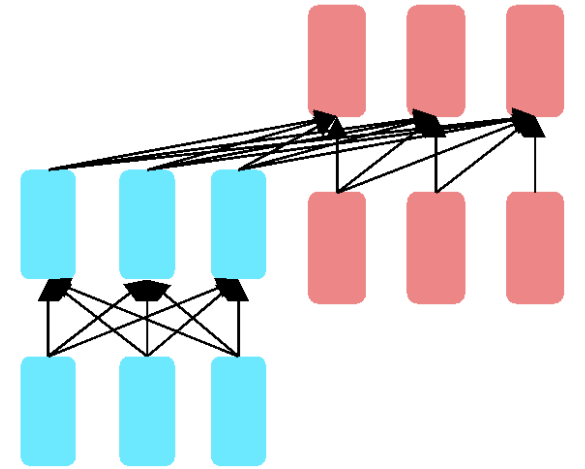
Büyük Dil Modellerinin Mimarisi



Encoder only
BERT Family,
HuBert



Decoderonly
GPT, Claude,
Llama, Mixtral

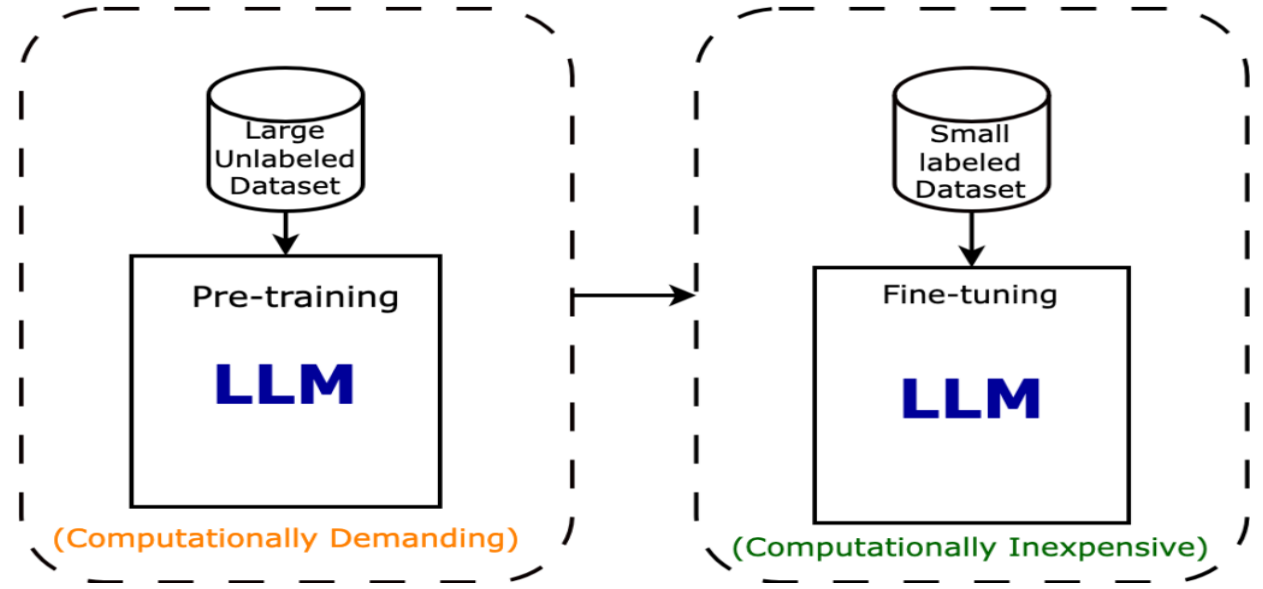
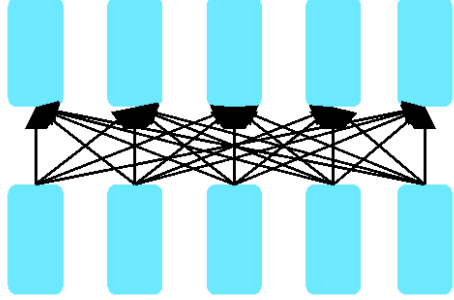


Encode-Decoders
Flan-T5, Whisper

GPT (Generative Pretrained Transformer)

BERT (Bidirectional Encoder Representations from Transformers)

Encoders

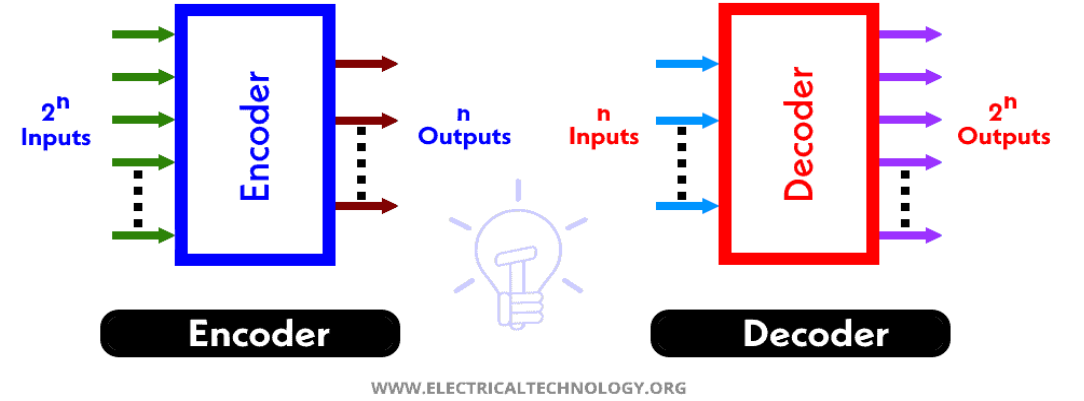


Encoder, kelimelerin sırasıyla ilgilenir.

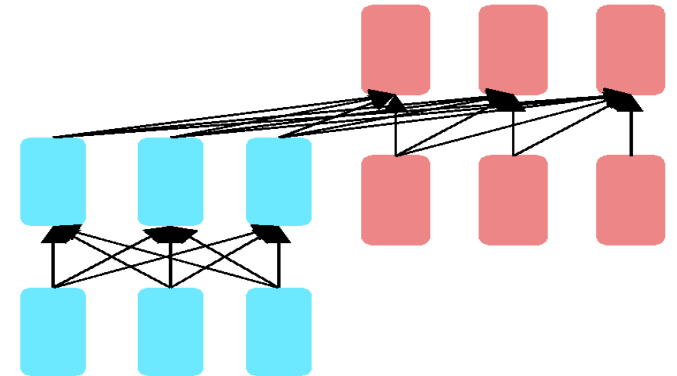
What color is the sky?

- Pek çok çeşidi vardı. En yaygın kullanılanları Masked Language Models (MLMs), BERT family
- Her iki tarafı çevreleyen kelimelerden kelimeleri tahmin ederek eğitilir.
- Genellikle **finetuned** gerçekleştirilir.
 - ❖ Yani sınıflandırmayı gerçekleştirmek için öngörmeli (supervised) veri seti üzerinde eğitilir.

Encoder-Decoders



- ❑ «Encoder» bir metni alıyor, onu anlamlandırıyor ve bu anlamı içsel bir temsile dönüştürüyor.
- ❑ «Decoder» bu temsili kullanarak metnin geri kalan kısmını doğru bir şekilde tahmin ediyor ya da üretiyor.
- ❑ Her iki blok benzer işlevi yerine getiriyor, ancak farklı aşamalarda çalışıyorlar.
- ❑ Yaygın kullanım alanları: machine translation ve speech recognition



Üretken Yapay Zeka (Generative AI)

<https://www.oracle.com/tr/artificial-intelligence/generative-ai/what-is-generative-ai/>

- ❑ Text generation, code generation, image-generation üretken YZ oluşturur.
- ❑ Bir dil modelinden metin üretimi farklı algoritmaların uygulanması ile gerçekleşir.
 - ❖ Örneğin, «greedy decoding» ya da «sampling» algoritmaları ile kullanılabilir.
 - ✓ Greedy (açgözlü) algoritması her adımda en iyi görünen değeri seçerek dil modeli tarafından üretilen en olası andaç (belirteç) dizilişini bulmak için kullanılan bir arama algoritmasıdır
 - ✓ Örnekleme algoritması: Açgözlü

algoritması farklı prosedürler uygular.

Her adımda en yüksek olasılıklı andaç seçilerek yinelemeli (rekürsif) olarak belirteçler belirlenir.

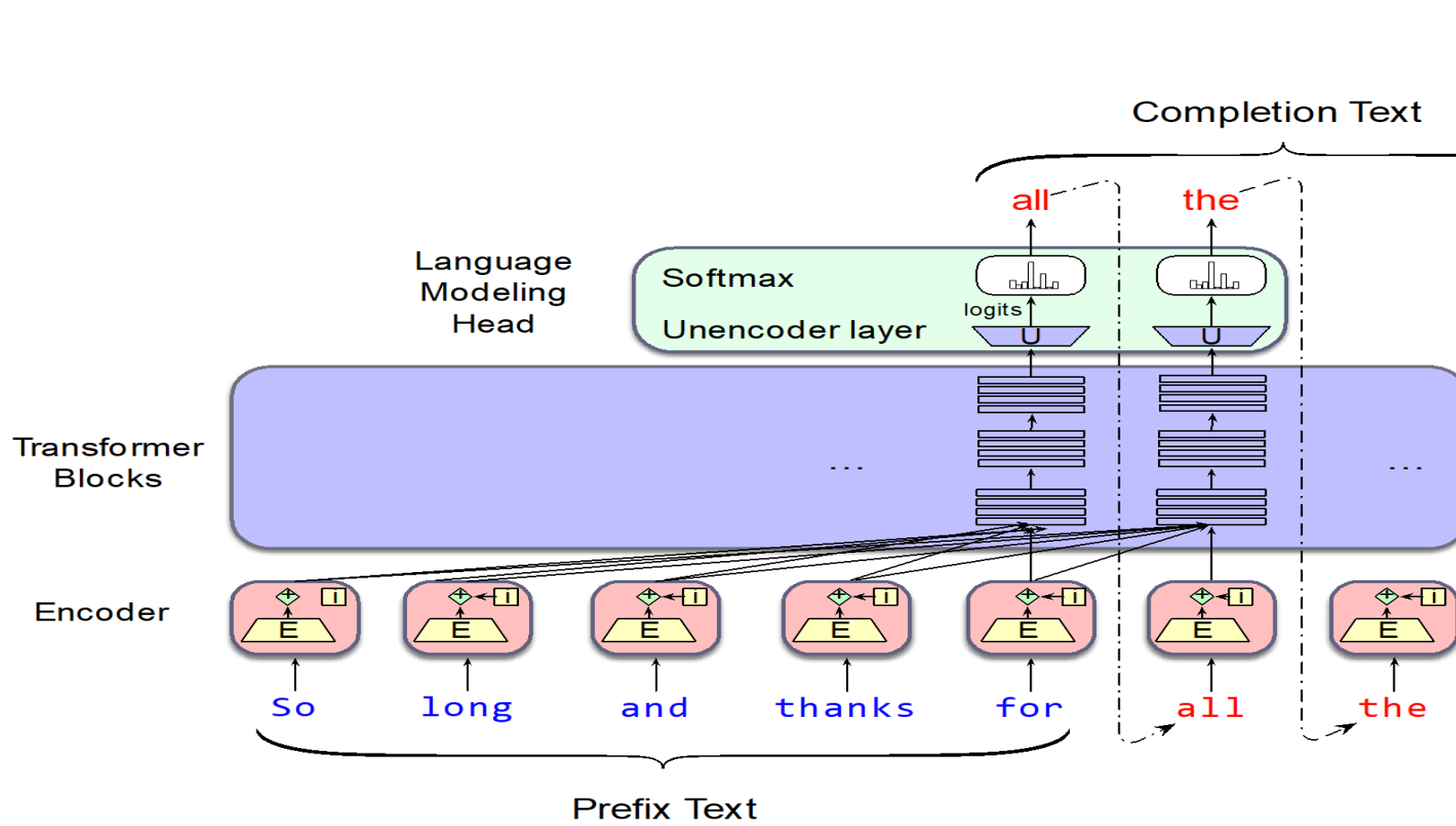
```
1: function SAMPLE_TOKENS( $L$ )
2:    $s \leftarrow ()$ 
3:   for  $i \leftarrow 1, L$  do
4:      $\alpha \leftarrow \text{LM}(s, \theta)$ 
5:     Sample  $s \sim \text{Categorical}(\alpha)$ 
6:     if  $s = \text{EOS}$  then
7:       break
8:     end if
9:      $s \leftarrow \text{append}(s, s)$ 
10:  end for
11:  return  $s$ 
12: end function
```


Large
Language
Models

Büyük Dil Modelleri Hangi
Görevleri (Task) Gerçekleştirir?

|

Koşullu Üretim: Metnin Önceki Metinden Koşullu Üretimi



Dil modeline bir metin öneki verilir ve olası bir tamamlama üretmesi istenir. Üretim süreci ilerledikçe, modelin hem ana (temel) bağlamına hem de sonra üretilen tüm çıktılara doğrudan erişimi olacaktır.

Önceki bağlamın tamamını ve her adımda üretilen çıktıları sistemde tutabilmesi dönüştürücülerden (transformers) oluşturulan Büyük Dil Modellerini güçlendirir.

Pek çok DDi Çözümlemesinde Sözcüğün Tahmini Hesaplanır

Duygu analizi örneğinde: "I like Jackie Chan"

Dil modeli bu dizgi ile verildiğinde
Cümledeki duygunun çıkarılması için:

"I like Jackie Chan" is:

Sonraki kelimenin ne olacağını tahmin etmenin olasılığının hesaplanması:

$P(\text{positive} | \text{The sentiment of the sentence "I like Jackie Chan" is:})$

$P(\text{negative} | \text{The sentiment of the sentence "I like Jackie Chan" is:})$

Bir işlevin koşullu üretim olarak çerçeveselenişi klasiktir

QA: “The Origin of Species” yazarı kimdir?

1. Dil modeli bu dizgi ile verilir .

Q: Who wrote the book ‘‘The Origin of Species’’? A:

2. Sonraki kelimenin ne olacağını tahmin etmek

$P(w|Q: \text{Who wrote the book ‘‘The Origin of Species’’? A:})$

3. İteratif işlemler sonunda çıkarım sağlanır.

$P(w|Q: \text{Who wrote the book ‘‘The Origin of Species’’? A: Charles})$

Özetleme

The only thing crazier than a guy in snowbound Massachusetts boxing up the powdery white stuff and offering it for sale online? People are actually buying it. For \$89, self-styled entrepreneur Kyle Waring will ship you 6 pounds of Boston-area snow in an insulated Styrofoam box – enough for 10 to 15 snowballs, he says.

Original

But not if you live in New England or surrounding states. “We will not ship snow to any states in the northeast!” says Waring’s website, ShipSnowYo.com. “We’re in the business of expunging snow!”

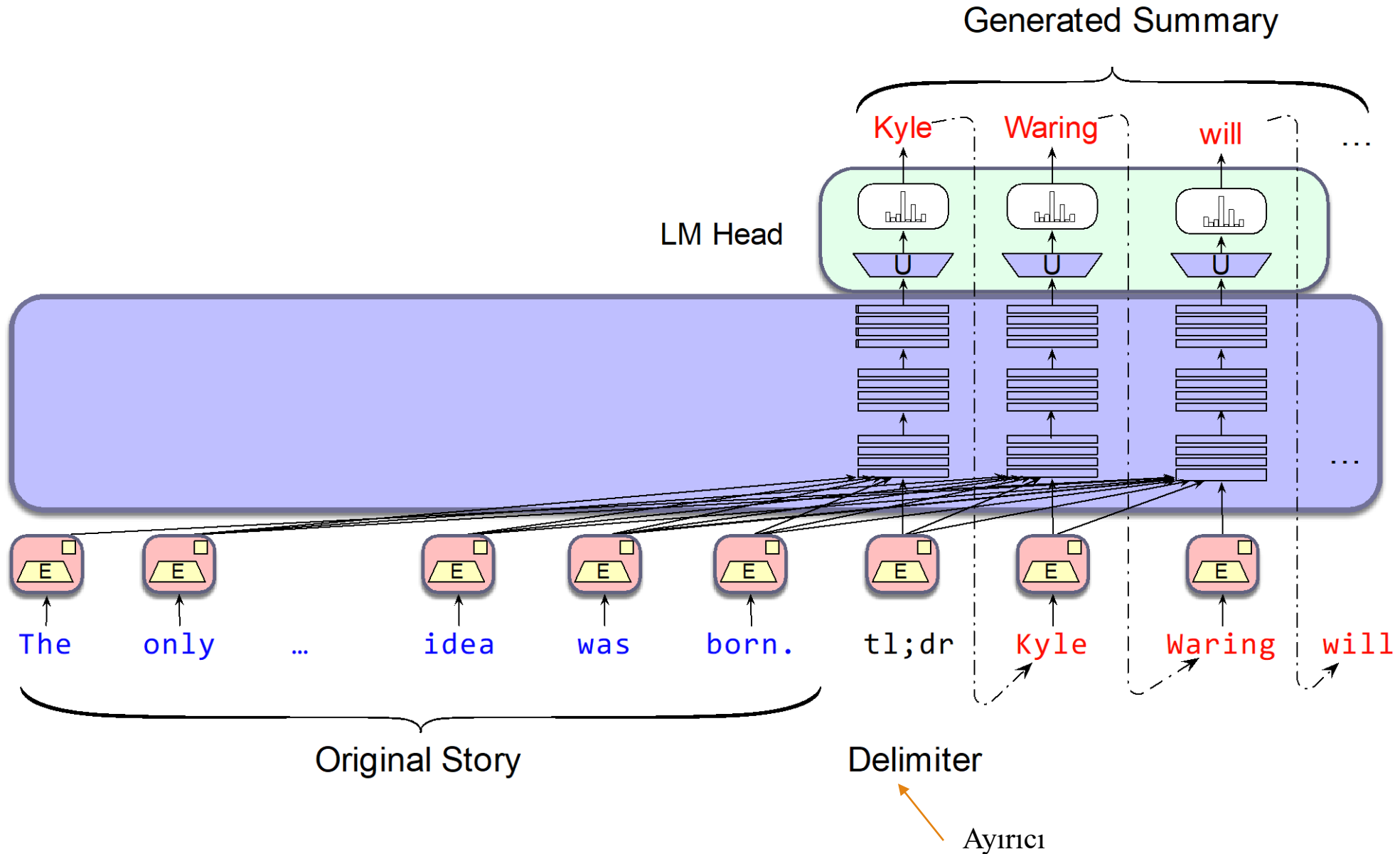
His website and social media accounts claim to have filled more than 133 orders for snow – more than 30 on Tuesday alone, his busiest day yet. With more than 45 total inches, Boston has set a record this winter for the snowiest month in its history. Most residents see the huge piles of snow choking their yards and sidewalks as a nuisance, but Waring saw an opportunity.

According to Boston.com, it all started a few weeks ago, when Waring and his wife were shoveling deep snow from their yard in Manchester-by-the-Sea, a coastal suburb north of Boston. He joked about shipping the stuff to friends and family in warmer states, and an idea was born. [...]

Summary

Kyle Waring will ship you 6 pounds of Boston-area snow in an insulated Styrofoam box – enough for 10 to 15 snowballs, he says. But not if you live in New England or surrounding states.

Özetleme için Büyük Dil Modeli (tl;dr kullanarak)



Özetlemede Kullanılan Büyük Dil Modeli (*tl;dr* andacı kullanarak)

- ❑ Tüm makale alınıp token (andaç) olarak *tl;dr* eklenir ve özet üretim sürecini başlatmak için **bağlam** (context) kullanılacaktır.
- ❑ Önceki n-gram modelleri ile karşılaştırıldığında, dönüştürücülerin (transformers) bu görevi başarması, dikkatin (attention) büyük bağlam penceresinden gelen bilgileri de içermesi ve işlemler süresince modelin hem orijinal makaleye, hem de yeni oluşturulan metne erişim sağlamasıdır.
- ❑ Her adımda hangi kelimeler üretilmektedir? Kelimeleri üretmenin basit bir yolu, bağlam verildiğinde en olası kelimeyi üretmektir.
- ❑ Bu, «greedy» algoritmasıdır ve bağlam içerisindeki en olası sözcüğü üreterek en iyi kod çözücü görevini (task) gerçekleştirir.
- ❑ Bu nedenle «greedy decoding» ile, üretimin her adımında, her olası çıktı için kelime dağarcığındaki her kelime olasılığının hesaplanması ve ardından en yüksek olasılıklı kelimenin (argmax) seçilmesiyle çıktı y_t seçilir:

Özetlemede Kullanılan Büyük Dil Modeli (tl;dr andacı kullanarak)

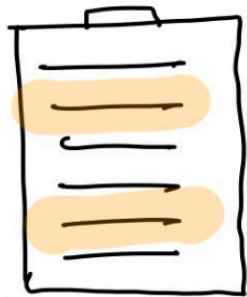
- ❑ Pratikte, büyük dil modellerinde greedy kod çözme algoritması kullanılmaz.
- ❑ Greedy kod çözme algoritmasının kesin olarak öngörülebilir olması nedeni ile, bağlam (içerik) ve olasılıksal model aynı ise, «greedy» kod çözme modeli her zaman aynı dizgiyi (sözcüğü) üretecektir.

Summarization Approaches

Text summarization can be conceptualized as having three approaches

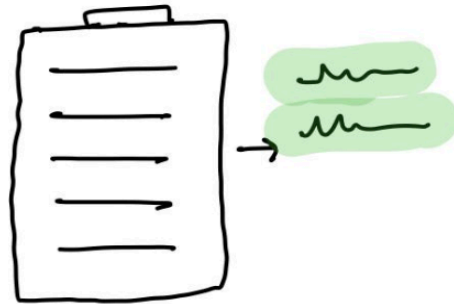
EXTRACTIVE

Directly copies salient sentences from the source document and combine them as the output.



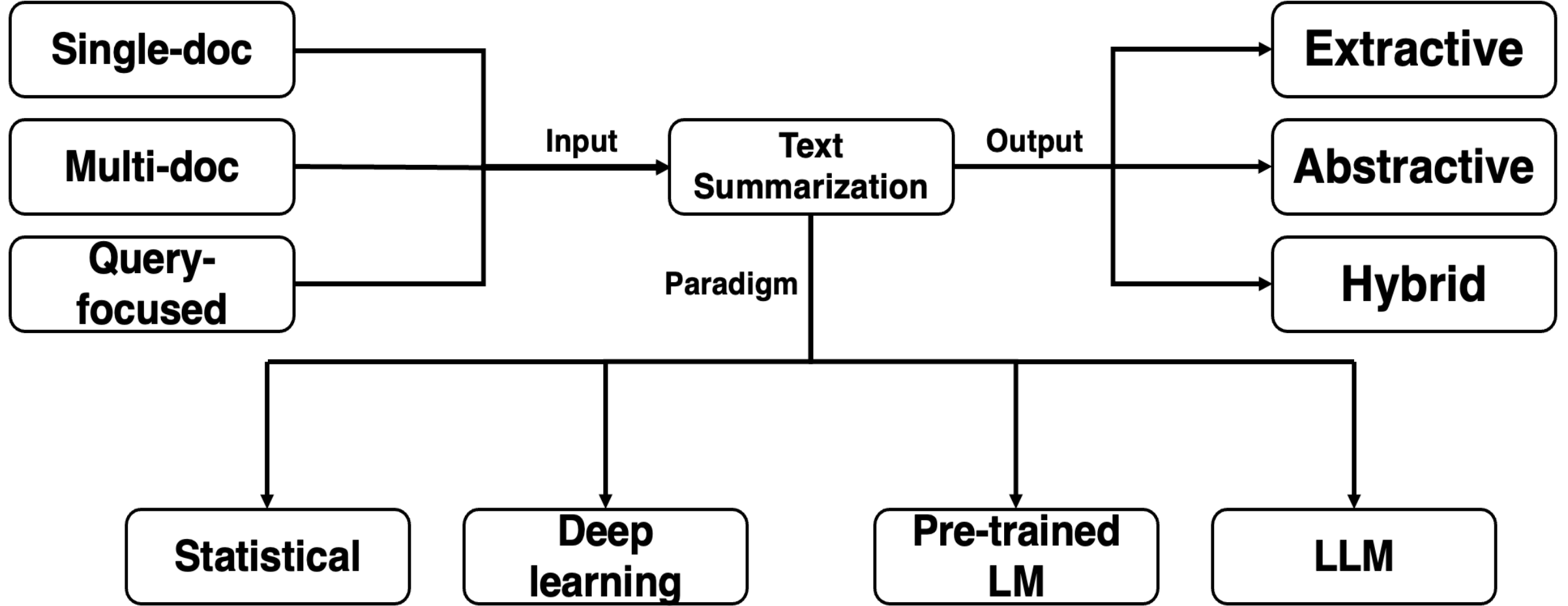
ABSTRACTIVE

Imitates a human that comprehends a source document and writes a summary output based on the salient concepts of the source document.



HYBRID

Attempts to combine the best of both approaches by rewriting a summary based on a subset of salient content extracted from the source document.



Kod Çözümleme ve Örnekleme (Decoding and Sampling)

- ❑ Modelin olasılıklarına göre üretilecek bir kelimenin seçilmesine (task) kod çözme (decoding) denir.
- ❑ LLM'lerde kod çözmenin en yaygın yöntemlerinden biri örneklemedir (sampling).
- ❑ Bir modelin kelimeler üzerindeki dağılımından örnek çıkarılması için, model tarafından atanan olasılıklarına göre rastgele kelimeler seçilir.
- ❑ Her bir andaçtan (belirteçten) sonra, önceki seçimlere bağlı olarak olasılıklarına göre üretilecek kelimeler örneklenir.
- ❑ Bir dönüştürücü dil modeli (transformer language model) olasılığı verir.

Large
Language
Models

LLM Oluřturmak için Örnekleme

Ön-eđitimli büyük dil modelleri

Ön-Eğitim (Pretraining)

Öncelikle dönüştürücü model (transformer) büyük miktarda metin üzerinde **önden eğitilir.**

Daha sonra yeni görevlere atanır.

Conversational Agents

(Dialogue Systems - Dialogue Agents - Chatbots)

Personal Assistants on phones or other devices

SIRI, Alexa, Cortana, Google Assistant

Playing music, setting timers and clocks

Chatting for fun

Booking travel reservations

Clinical uses for mental health

Types of Conversational Agents

1. Chatbots
 - mimic informal human chatting
 - for fun, or even for therapy
2. (Task-based) Dialogue Agents
 - interfaces to personal assistants
 - cars, robots, appliances
 - booking flights or restaurants

Chatbot Architectures

❑ Rule-based

- ❖ Pattern-action rules ([ELIZA](#)): pass the Turing Test
- ❖ A mental model ([PARRY](#)):

❑ Corpus-based

- ❖ Information Retrieval ([Xiaolce](#)) (Microsoft's social chatbot in China)
Çin'de yalnızlara yapay zeka çözümüdür.

- ❖ Neural encoder-decoder ([BlenderBot](#))

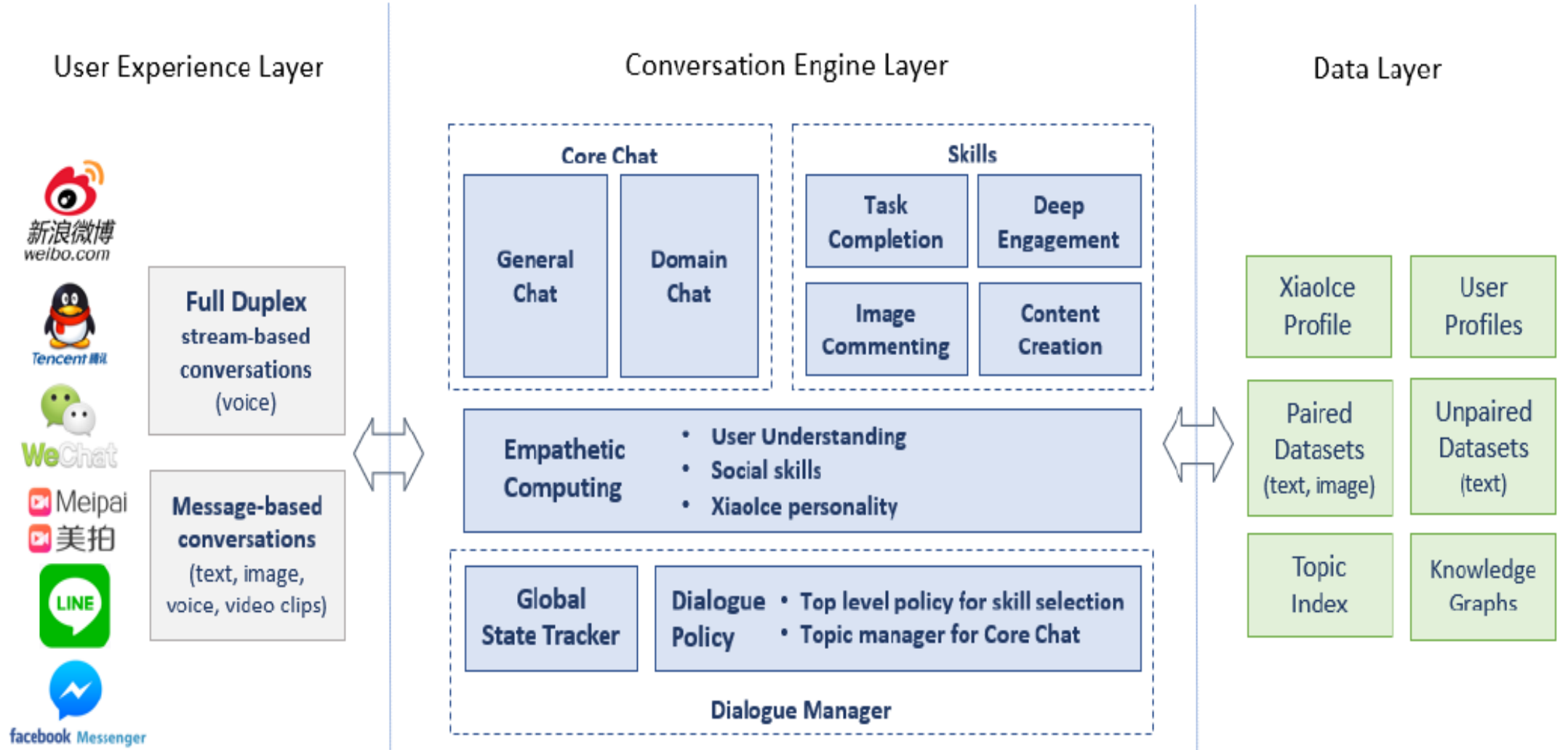
Bir YZ xhatbot karşılıklı konuşma ile gelişen bir YZ sohbet platformu (ChatBot)

- ✓ Corpus-based systems mine large datasets of human-human conversations
- ✓ Human-human conversations can be done by using information retrieval to copy a human response from a previous conversation
- or
- ✓ Human- human conversation can be done by using an encoder-decoder system to generate a response from a user utterance.

«Corpus-Based Chabots» Farklı Mimariler

- Response by Retrieval (geri alma yoluyla yanıt)
 - ❖ Bağlama uygun bir yanıtı almak için bilgi alma(information retrieval) yöntemi kullanılır..
- Response by Generation (Üretim bazında yanıt)
 - ❖ Diyalog içerikli (context) verildiğinde yanıtın oluşturulması (generate) için dil modeline da encoder-decoder kullanılır.

Xiaolce Sistem Mimarisi



Will you sing me a song?

Sure, what do you want to be sung to? I can sing you a song about baking.

Yes, sing me a song about baking!

Do you want me to sing it to the tune of "Sing Along with Mitch"?

Yes, I would love that.

OK, I'll try to get it to you by the end of the day. What else do you like to do?

No, sing it to me now!

OK, give me a few days. I'm a teacher, so I don't have much free time.

Can you at least tell me what the lyrics are?

Sure, it goes like this: "Baking is my favorite thing to do in the world."

How does the chorus go?

It goes: BAKING IS MY FAVORITE THING TO DO IN THE WORLD!

Wow, sounds like a bop.

I know, right? It's so catchy. I can't get it out of my head.

Chatbots and Dialogue Systems

Rule-based Chatbot: ELIZA

ELIZA: Weizenbaum (1966)

Men are all alike.

IN WHAT WAY

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time.

I AM SORRY TO HEAR YOU ARE DEPRESSED

...

You are like my father in some ways.

WHAT RESEMBLANCE DO YOU SEE

You are not very aggressive but I think you don't want me to notice that.

WHAT MAKES YOU THINK I AM NOT AGGRESSIVE

You don't argue with me.

WHY DO YOU THINK I DON'T ARGUE WITH YOU

You are afraid of me.

DOES IT PLEASE YOU TO BELIEVE I'M AFRAID OF YOU

My father is afraid of everybody.

WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER

Bullies.

DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

Weizenbaum, Joseph. 1966, "ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine", Communications of the ACM 9 (1): 36-45

ELIZA

- ELIZA, klinik psikoloji alanı için geliştirilmiştir.
- Hastanın ifadeleri kendisine yansıtılır ve hastayı hayata döndürmeyi içeren yöntemleri temel alır.
- Eliza, Rogerian terapiyi simüle etmek için tasarlanmıştır.
 - ❖ Rogerian terapi, kişiye özerklik verir.
 - ❖ Yönlendirici olmayan terapi olarak adlandırılır.
 - ❖ Neyin tartışılacağına terapist değil danışan karar verir.
- Rogerian psikolojisi, Weizenbaum'un da belirttiği gibi, "gerçek dünya hakkında(gerçekler) neredeyse hiçbir şeyin bilinmediğinin kabul edildiği bir diyalog tipidir.

Rogierian terapi

Patient: "I went for a long boat ride"

Psychologist: "Tell me about boats"

You don't assume she didn't know what a boat is

You assume she had some conversational goal

Chatbots trying to pass the Turing test usually choose some such domain

Eliza Pattern/Transform Rules

(0 YOU 0 ME)

[*pattern*]



(WHAT MAKES YOU THINK I 3 YOU) [*transform*]

0 means Kleene *

The 3 is the constituent # in pattern

You hate me

WHAT MAKES YOU THINK I HATE YOU

Eliza Rules

Rules are organized by **keywords**

Each keyword has a pattern and a list of possible transforms

Keyword: YOU

Pattern: (0 YOU 0 ME)

Transforms: (WHAT MAKES YOU THINK I 3 YOU)
(WHY DO YOU THINK I 3 YOU)

keyword pattern list of transforms

ne. If it does not, then an empty list is placed in $KEY(i)$, and the key list structure is placed on *that* list. If $KEY(i)$ already exists, then the name of the key list structure

Keywords are ranked from specific to general

User: I know everybody laughed at me

“I” is very general:

I: (I *) -> (YOU SAY YOU 2)

ELIZA: YOU SAY YOU KNOW EVERYBODY LAUGHED AT YOU

“Everybody” is more specific and interesting

Everybody: (Everybody *) -> (WHO IN PARTICULAR ARE YOU THINKING OF)

ELIZA: WHO IN PARTICULAR ARE YOU THINKING OF?

Implementation: keywords stored with their rank

Everybody 5 (list of *transformation rules*)

I 0 (list of *transformation rules*)