

Recent Advancement in Text Analytics

CPE 393: Text Analytics

Dr. Sansiri Tarnpradab

*Department of Computer Engineering
King Mongkut's University of Technology Thonburi*

Announcement

- **Presentation on 20/11/2024**
 - Schedule
 - Q&A
- **Final Exam on 27/11/2024**
 - Coverage from after midterm onwards
 - In-class
 - Paper-based
 - Pretty much the same format
 - Duration: 2 Hrs

Intro

*Pattern
Matching*

*Text
Visualization*

Web Scraping

*Text
Preparation*

*Text Feature
Representation*

*Text
Classification*

*Text
Summarization*

*Text
Clustering*

*Topic
Modeling*

*Recent
Advancement*

*Project
Presentation*



Prompt

What is a *Prompt?*



Ref:
<https://www.nytimes.com/2023/09/20/technology/chatgpt-dalle3-images-openai.html>

- In generative AI models, a prompt is a:
 - Text input provided by users
 - To guide the model's output
- Text input
 - Simple questions
 - Detailed descriptions
 - Specific tasks
- GPT-4
 - One of the Large Language Models (LLMs)
 - ChatGPT
 - Simple queries to complex problem statements
- DALLE-3
 - Image generation models
 - ChatGPT DALLE
 - Often descriptive

Long & Specific Prompt (1)

A dystopian city in the desert with massive sharp spike-like architecture. A colossal floating snake like creature glides between the spikes. a man with a mech-suit is flying with boosters at the same height as the creature. Inspired by "Dune" and Syd Mead.



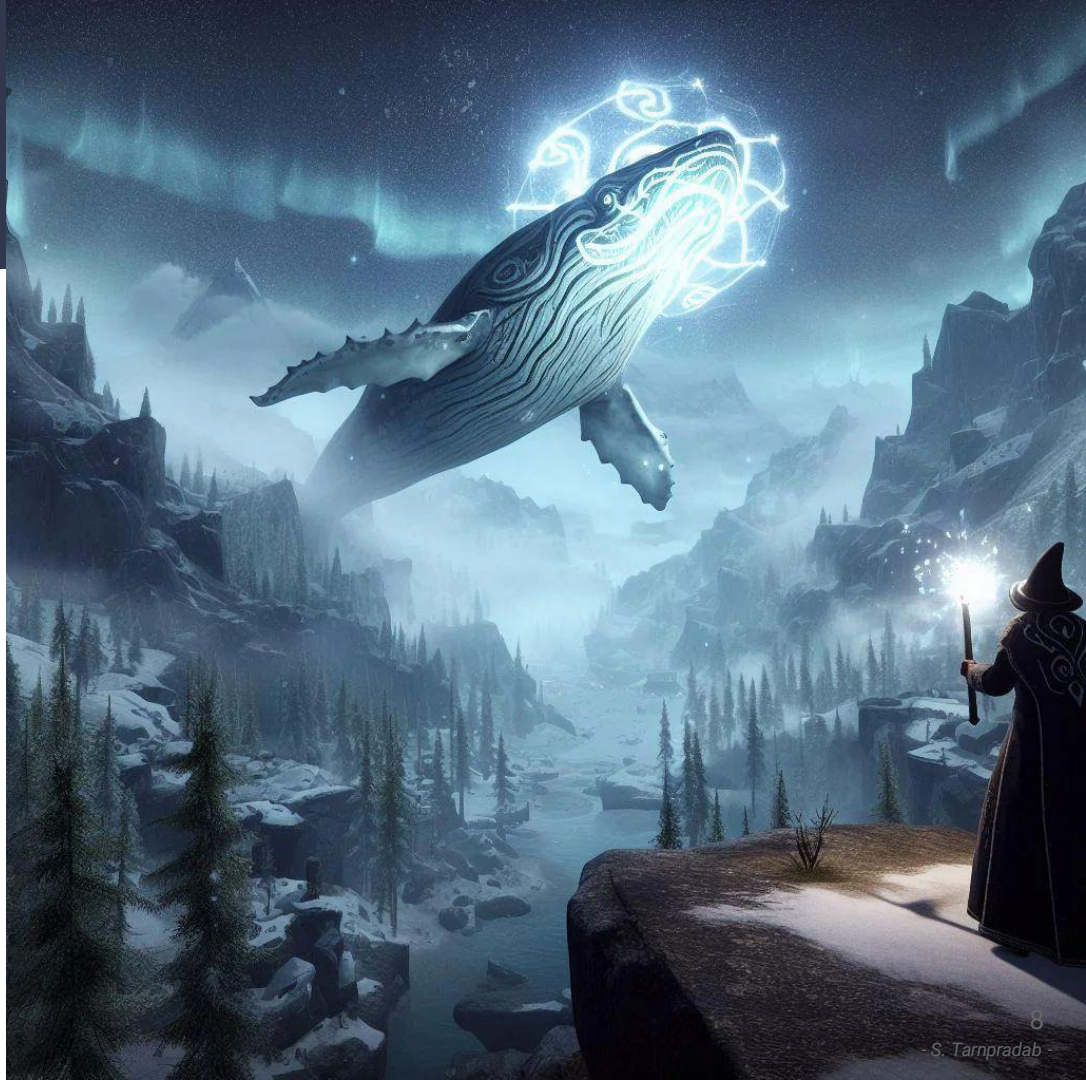
Long & Specific Prompt (2)

Scene from the movie "Wall E". the text "Walmart" as a hologram in a massive shopping district inside of a spaceship. People are eating and drinking coffee at a cafe to the side, overlooking a large aquarium decoration with fish and jellyfish.



Long & Specific Prompt (3)

Third-person perspective gameplay of the dark fantasy game "Wizards and Whales". The screenshot shows a wizard looking up at a massive magical whale over a mountainous region. The region is clifty and snowy, with sparsely populated evergreen trees. The player is casting the 'Ray of holy light' spell, which shoots a ray of white light towards the whale. The graphics are reminiscent of Skyrim, with low-res polygon models and textures.



Long & Specific Prompt (4)

Botanical diagram featuring several different exotic flowers and bugs. The largest one is a Protea, next to which is a praying mantis. The other is a Blue Passion Flower which is next to a blue dragonfly. There is also, inexplicably a handgun with interesting camo which fits with the flowers and bugs.



Long & Specific Prompt (5)

Poster for the movie "Fall of Rome" by Christopher Nolan. The header of the poster is the text "Fall of Rome" in an intensely designed red font. A man is leaping into the air, holding a sword above him into a sea of people. In the background is gloomy and grimey bright sunny scene of Rome in stunning HD.



More on Prompts

- Prompts generally consist of:
 - Instructions
 - Questions
 - Input data
 - Examples
 - *Note: Either 1 or 2 must be present*
- **Basic prompt:**
 - Ask a direct question
 - Provide instructions for a specific task
- **Advanced prompt:**
 - More complicated
 - The model is guided to follow a logical reasoning process to arrive at an answer.

Basic Prompts

Prompts generally consist of:

- Instructions
- Questions
- Input data
- Examples

- Instructions + Question
- Instructions + Input
- Question + Examples

When using only a question



You

How should I write my college admission essay?



ChatGPT

Writing a college admission essay can be a daunting task, but it's also an opportunity to showcase your unique qualities and experiences to the admissions committee. Here are some tips to help you craft a compelling essay:

1. **Be Authentic:** Write from the heart and be true to yourself. Admissions officers can often tell when an essay is genuine or forced. Share personal stories, reflections, and insights that truly reflect who you are.
2. **Show, Don't Just Tell:** Instead of listing your achievements or qualities, illustrate them with specific examples and anecdotes. Paint a vivid picture of your experiences and how they have shaped you.
3. **Focus on a Specific Topic:** Choose a narrow topic or theme that allows you to delve deep into your experiences or values. Avoid trying to cover too much ground in a single essay.
4. **Highlight Your Growth and Learning:** Admissions officers are interested in how you have grown and what you have learned from your experiences. Reflect on challenges you've faced, mistakes you've made, and how you've overcome them.

Provide more instructions

(Question + Instructions)



You

Give me suggestions about the different sections I should include, what tone I should use, and what expressions I should avoid.



ChatGPT

Certainly! Here's a breakdown of the sections you might include in your college admission essay, along with suggestions for tone and expressions to use:

1. Introduction:

- **Content:** Begin with a captivating hook that grabs the reader's attention. Introduce yourself and provide some context for the rest of the essay.
- **Tone:** Engaging, personal, and authentic. Show enthusiasm for sharing your story.
- **Expressions to Avoid:** Clichés, generic opening lines, and overly formal language.

2. Personal Story or Experience:

- **Content:** Share a specific anecdote or personal experience that illustrates something important about you. This could be a moment of realization, a challenge you've overcome, or a significant event in your life.
- **Tone:** Reflective, introspective, and honest. Be vulnerable, but also show resilience and growth.
- **Expressions to Avoid:** Exaggerations, overly dramatic language, and irrelevant details.

3. Values and Beliefs:

- **Content:** Discuss your values, beliefs, or principles that guide your actions and decisions. Explain why these are important to you and how they have shaped your identity.
- **Tone:** Thoughtful, sincere, and introspective. Show depth of character and a willingness to



Provide more input

(Instructions + Input)

“Given the following information about me, write a 4 paragraph college essay: I am originally from Barcelona, Spain. While my childhood had different traumatic events, such as the death of my father when I was only 6, I still think I had quite a happy childhood.. During my childhood, I changed schools very often, and attended all kinds of schools, from public schools to very religious private ones. One of the most “exotic” things I did during those years is to spend a full school year studying 6th grade in Twin Falls, Idaho, with my extended family.

I started working very early on. My first job, as an English teacher, was at age 13. After that, and throughout my studies, I worked as a teacher, waiter, and even construction worker.”

Important Remarks:

- Ethical use
- The fact that you can do something with a generative model does not mean that it is the right thing to do!
- There are already efforts underway to detect AI generated content.



Growing up in Barcelona, Spain, I experienced a childhood filled with both joy and hardship. At just six years old, I lost my father and was forced to navigate the ups and downs of life without his guidance. Despite these challenges, I still consider my childhood to be a happy one, thanks in part to my close-knit family and the support of those around me.

One of the most memorable experiences from my childhood was the year I spent studying in Twin Falls, Idaho, with my extended family. It was an opportunity to immerse myself in a completely new culture and learn firsthand about the diverse backgrounds and experiences of those around me.

Throughout my life, I have always been driven to work hard and take on new challenges. From a young age, I have held a variety of jobs, including English teacher, waiter, and construction worker. Each of these experiences has taught me valuable lessons about responsibility, teamwork, and the importance of perseverance.

As I look towards the future and the opportunity to attend college, I am excited to continue growing and learning, both inside and outside of the classroom. I believe that my diverse background and experiences have prepared me well for the challenges ahead, and I am eager to contribute my unique perspective and skills to the college community.

More Advanced

Prompts

- No longer a simple Q&A
- More context should be given
 - Steps
 - Patterns

Standard Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The answer is 27. ❌

Chain-of-Thought Prompting

Model Input

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?

Model Output

A: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Techniques

ZERO-SHOT PROMPTING

The most basic form of prompting.

Show a prompt to the model without examples.

Basic prompts that we saw already

ONE-SHOT PROMPTING

Provide a model with a **single** example

FEW-SHOT PROMPTING

Provide a model with **a few** labeled examples

Better accuracy than zero-shot & one-shot

More examples = Better performance

One-shot VS Few-shot

ONE SHOT



You

Translate the following English sentence.

Here's an example:

'I am a student' -> 'ฉันเป็นนักเรียน'

Translate this: 'I am hungry'.



ChatGPT

The translation for "I am hungry" into Thai is: 'ฉันหิว'.



You

Translate the following English sentence.

Here's an example:

'I am a student' -> 'je suis étudiant'

Translate this: 'I am hungry'.



ChatGPT

The translation for "I am hungry" into French is: 'J'ai faim'.

FEW-SHOT



You

The odd numbers in this group, add up to an even number: 15, 32, 5, 13, 82, 7, 1.



ChatGPT

Let's sum up the odd numbers from the group:

$$15 + 5 + 13 + 7 + 1 = 41$$

Indeed, the sum of these odd numbers is 41, which is an even number! That's a neat observation.

It's because when you add an odd number of odd numbers together, the result is always even.

Chain-of-thought Prompting

CHAIN-OF-THOUGHT PROMPTING

- Chain-of-thought prompting (CoT)
- Encourage the model to be factual/correct
- Force it to follow a series of steps in its “reasoning”
- Even more efficient?
 - Combine it with few-shot prompting
 - Better results on more complex tasks

Prompt:

```
The odd numbers in this group add up to an even number: 4, 8, 9, 15, 12, 2, 1.  
A: Adding all the odd numbers (9, 15, 1) gives 25. The answer is False.  
The odd numbers in this group add up to an even number: 17, 10, 19, 4, 8, 12, 24.  
A: Adding all the odd numbers (17, 19) gives 36. The answer is True.  
The odd numbers in this group add up to an even number: 16, 11, 14, 4, 8, 13, 24.  
A: Adding all the odd numbers (11, 13) gives 24. The answer is True.  
The odd numbers in this group add up to an even number: 17, 9, 10, 12, 13, 4, 2.  
A: Adding all the odd numbers (17, 9, 13) gives 39. The answer is False.  
The odd numbers in this group add up to an even number: 15, 32, 5, 13, 82, 7, 1.  
A:
```

Output:

```
Adding all the odd numbers (15, 5, 13, 7, 1) gives 41. The answer is False18
```


Zero-shot CoT

Prompting

Add "Let's think step by step" to the original prompt.

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

More Advanced

Prompts (Cont.)

“Are mRNA vaccines safe? Answer only using reliable sources and cite those sources. ”

“Write a poem describing a beautiful day <|endofprompt|>. It was a beautiful winter day”

- No longer a simple Q&A
- More context should be given
 - Steps
 - Patterns
- Encourage the model to be factual
 - Point the model in the right direction
 - Prompt it to cite the right sources
- Explicitly ending the prompt instructions
 - `<|endofprompt|>`
- Be forceful
 - Forceful language
 - CAPITAL LETTERS, exclamation mark!!!

“Are mRNA vaccines safe? Answer only using reliable sources and cite those sources. ”



Yes, mRNA vaccines are generally considered to be safe. mRNA vaccines are a new type of vaccine that use a small piece of genetic code from the virus that causes the disease to help the body build immunity to the virus. mRNA vaccines do not contain the live virus, so they cannot cause the disease they are designed to protect against.



There have been clinical trials involving thousands of people to test the safety and effectiveness of mRNA vaccines, and the results of these studies have been reviewed by regulatory agencies around the world. Based on the available evidence, regulatory agencies such as the US Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have concluded that mRNA vaccines are safe and effective.

References:

- US Food and Drug Administration. (2021). FDA's Emergency Use Authorization (EUA) for COVID-19 Vaccines. Retrieved from <https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/covid-19-vaccines#mRNA>
- European Medicines Agency. (2021). COVID-19 vaccines: EMA's role in their development and authorization. Retrieved from <https://www.ema.europa.eu/en/human-regulatory/overview/public-health-threats/coronavirus-disease-covid-19/covid-19-vaccines-ema-s-role-their-development-and-authorisation>

**“Write a poem describing a beautiful day
<|endofprompt|>. It was a beautiful winter
day”**



with a sky so blue and crisp,
the sun shining down with a warm and gentle kiss,
a light breeze in the air, with snowflakes on the ground,
it was a peaceful, serene sight to be found.

The trees were dusted with snow, their branches bright and white,
casting delicate shadows in the sun's warm light,
birds chirped a cheerful tune, flitting to and fro,
it was a day that seemed to glow.

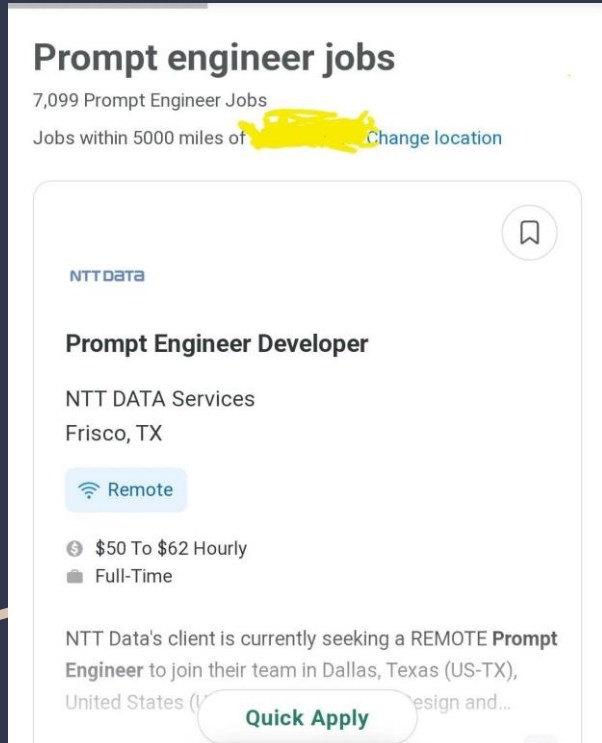
The world was hushed and still, as if it too were in awe,
of the beauty that surrounded us all,
and in that moment, I was filled with such delight,
for it was truly a beautiful winter day, so bright.

Some *Limitations*

LLMs' limitations:

- **Transient State**
 - Lack persistent memory or state,
 - Need additional software/systems for context retention and management.
- **Probabilistic Nature**
 - Identical prompts → variable responses
 - Inconsistency
- **Outdated Information**
 - Rely on pre-training data
 - Precluding real-time awareness or updates
- **Content Fabrication**
 - Hallucination
- **Resource Intensity**
 - Significant computational and financial costs
 - Scalable? Accessible?
- **Domain Specificity**
 - Often require domain-specific data

Prompt: *Conclusion*



Prompt engineer jobs
7,099 Prompt Engineer Jobs
Jobs within 5000 miles of [redacted] [Change location](#)

NTT DATA

Prompt Engineer Developer

NTT DATA Services
Frisco, TX

[Remote](#)

\$ \$50 To \$62 Hourly
Full-Time

NTT Data's client is currently seeking a REMOTE **Prompt Engineer** to join their team in Dallas, Texas (US-TX), United States (US) design and...

[Quick Apply](#)

- Introduction
- What is prompt?
- Basic
- Advanced
- Techniques
- Limitations
- Main reference:
 - [Prompt Engineering Guide](#)
 - [Amatriain, X.](#) (2024)

Other Recent Advancements

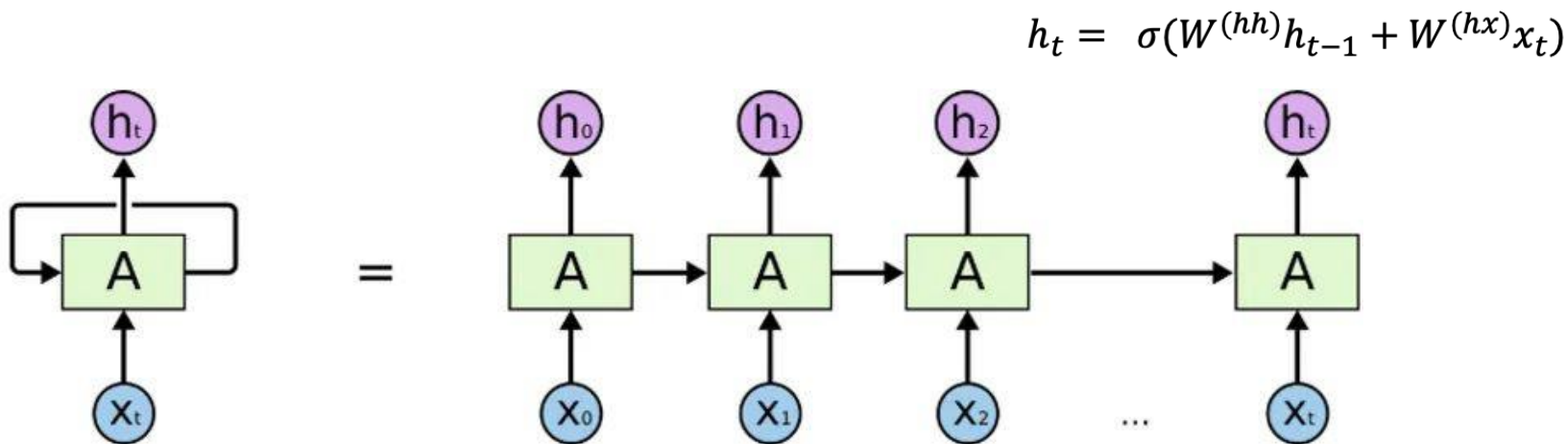
Current Trends

- Transformers
- Large Language Models (LLMs)
- Fine-Tuning (for specific domains)
- Retrieval-Augmented Generation (RAG)
- Advances in Sentiment Analysis
- Explainability
- Challenges & Where we're heading

Transformer

- **Key Components:**
 - Parallel processing
 - Self-attention mechanism
- Overcame RNNs and LSTMs
- Enabled major breakthroughs in NLP by improving contextual comprehension and scalability
- **Examples:** BERT, T5, RoBERTa
- **Applicable Tasks:** Classification, Translation, and Summarization

Revisiting RNN: Recurrent Neural Networks

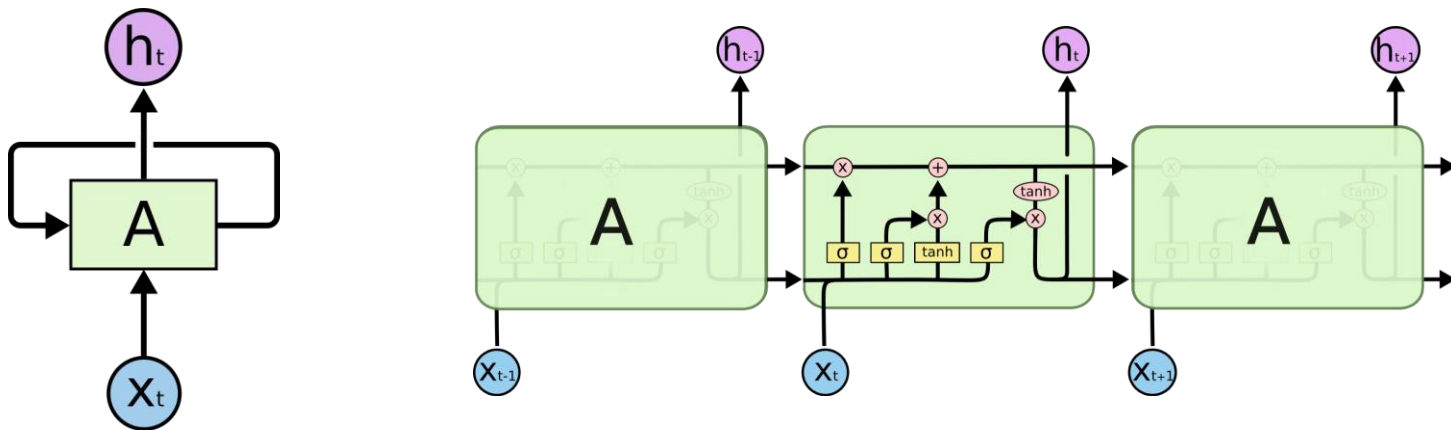


Weaknesses:

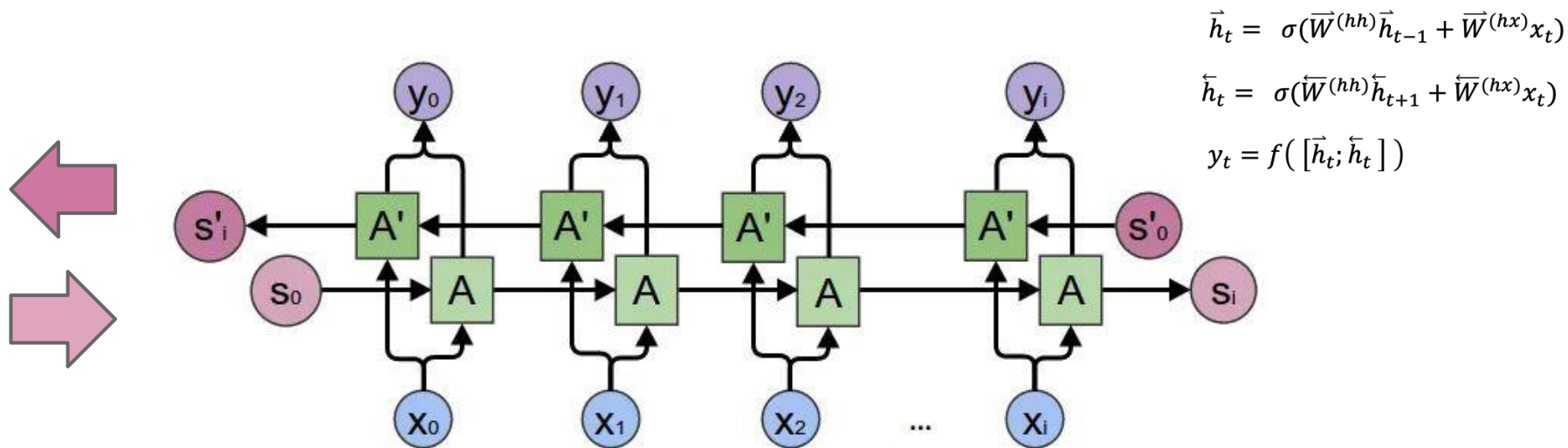
- Vanishing gradient problem \rightarrow during backpropagation through time (BPTT)
- Capturing long-term dependencies
- Lack of Parallelization \rightarrow computationally inefficient, longer training time

Revisiting LSTM: Long-Short Term Memory Networks

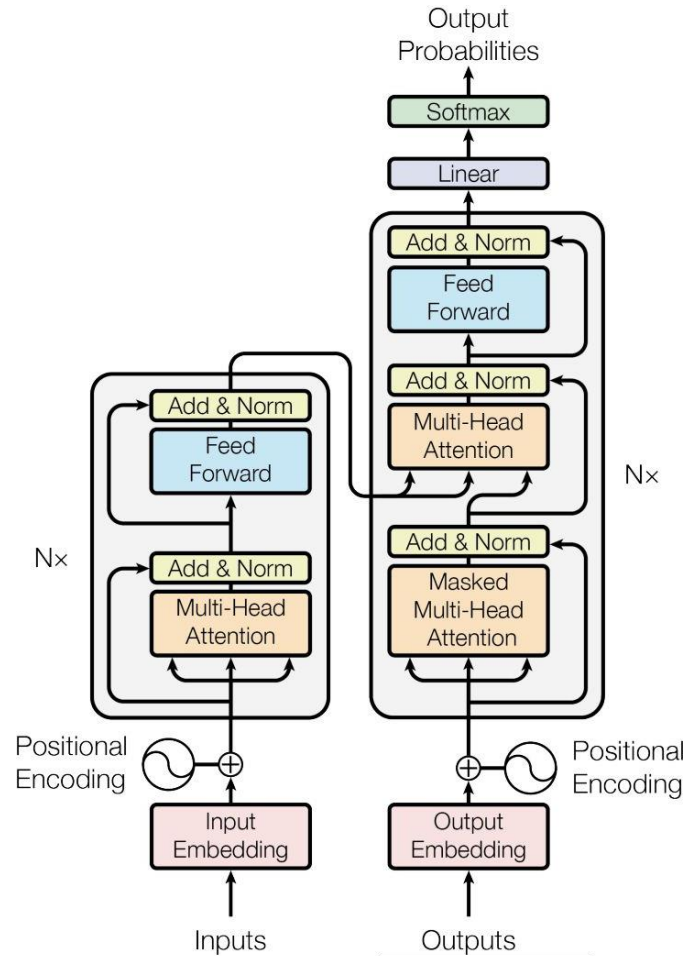
- One of the variants of RNN
- Better than traditional RNN for its memory
- Gates: a forget gate, an input gate, and an output gate.



Revisiting Bidirectional RNN



- Two RNNs stacked on top of each other
- Incorporate both left and right context
- Output is computed based on the hidden state of both RNNs



Transformer

(Cont.)

The	dog	sleeps	on	the	floor
-----	-----	--------	----	-----	-------

V_1	V_2	V_3	V_4	V_5	V_6
-------	-------	-------	-------	-------	-------

+

P_1	P_2	P_3	P_4	P_5	P_6
-------	-------	-------	-------	-------	-------

=

V_1	V_2	V_3	V_4	V_5	V_6
-------	-------	-------	-------	-------	-------

- **Key Components:**

- Parallel processing
- Self-attention mechanism

- Parallel processing

- Fast!
- Positional encoding → word order

- Self-attention mechanism

- Context understanding
- Capture relationship
- Similarity between words

The **dog** sleeps on the floor. **It** is dreaming.

LLMs

- A type of model specifically trained on large-scale datasets to understand and generate human language.
- Built on transformer foundation
- **Keys:**
 - **Deep Learning** techniques (of course)
 - Massively **large datasets**
 - Typically has at least **1billion parameters**
- **Examples:**
 - GPT-4, [LLaMA 2](#), and [Claude](#)
- **Applications in text analytics:** chatbots, summarization, real-time translation

Fine-Tuning

- **Domain-specific** datasets (e.g., healthcare, finance).
- Increases accuracy and relevance in specific fields by adjusting models to specialized language and contexts.
- Examples:
 - Medical Question-Answering: [PubMedQA](#)
 - Clinical Note Summarization
 - Legal Summarization
 - Financial Sentiment Analysis
 - Fraud Detection
 - Experiment Report Generation
 - Citation Recommendation
 - Question Generation
 - Personalized Itinerary Suggestions

RAG

- **Key Components:**
 - **Retriever:** Searches external databases for relevant content.
 - **Generator:** Produces responses based on both retrieved data and the input query.
- **Benefits:**
 - Reduces the risk of "hallucination" by referencing real data.
 - Perfect for specialized areas like medical, legal, and technical fields.
- **Applications:**
 - Customer Support: Improves response accuracy in help centers.
 - Real-Time Information Retrieval: Pulls in updated data dynamically, making it ideal for frequently changing datasets.

Sentiment Analysis

- Algorithms now capture nuanced sentiment, such as sarcasm or mixed emotions.
- **Applications:** Brand monitoring, social media analysis, customer feedback evaluation.
- Architectures (Transformer-based)
 - RoBERTa
 - DistillBERT
 - ALBERT
 - T5
 - And more...

Explainability

- Transparent AI decisions are crucial, especially in regulated industries.
- Black boxes
- Attempts so far
 - Attention mechanism
 - Visualization
- Challenges:
 - Complexity
 - Data privacy

Challenge & Next Move

More than 1,200 Google workers
condemn firing of AI scientist Timnit
Gebru

More than 1,500 researchers also sign letter after Black
expert on ethics says Google tried to suppress her research
on bias



Timnit Gebru in San Francisco in 2018. Photograph: Kimberly White/Getty Images for TechCrunch

Ref: [News from The Guardian](#)

- **Addressing Bias:** Ensuring fairness in LLMs and RAG models by reducing data and model biases.
- **Data Privacy and Security:** Managing secure access and data protection in text analytics applications.
- **Human Interpretability:** Bridging the gap between complex models and user-friendly explanations.

Q & A