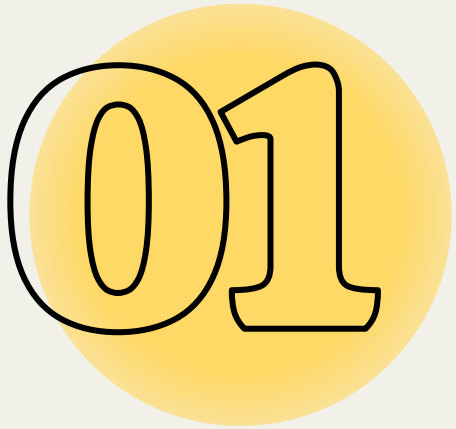
Running LLMs for Story Generation

WiM DRP CS1 Presentation

Presented by: Chelsea Fang, Kun Zhu
Mentor: Aisha Khatun

Table of Contents.

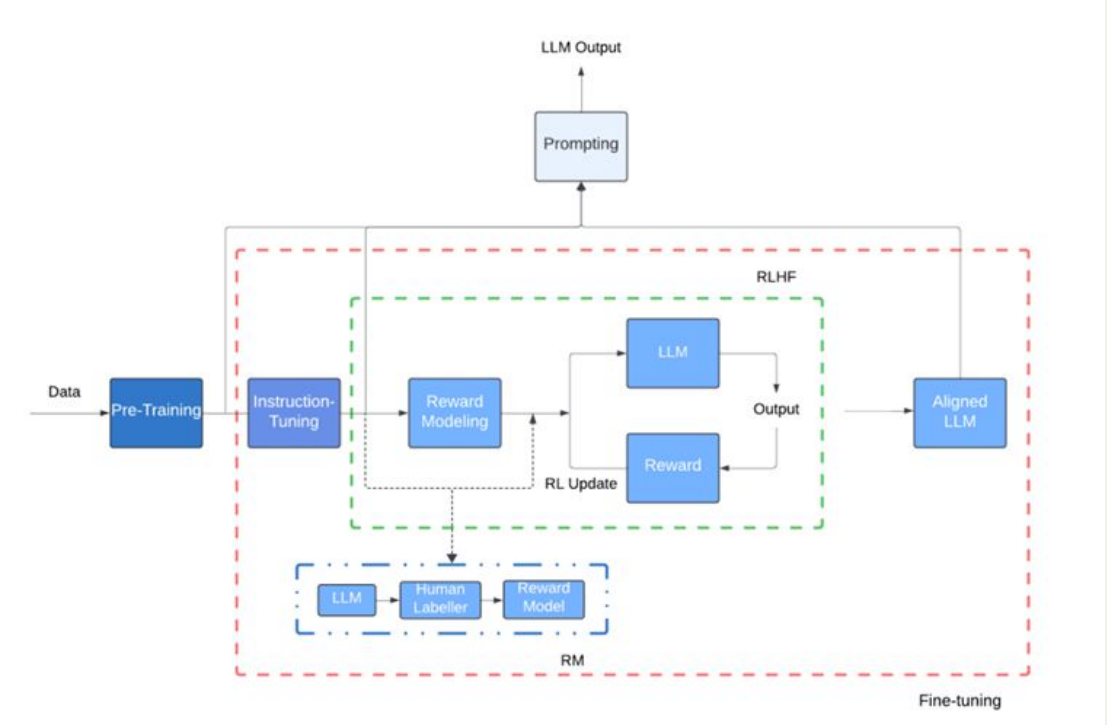


Introduction to LLM

What is LLM and how does it work?

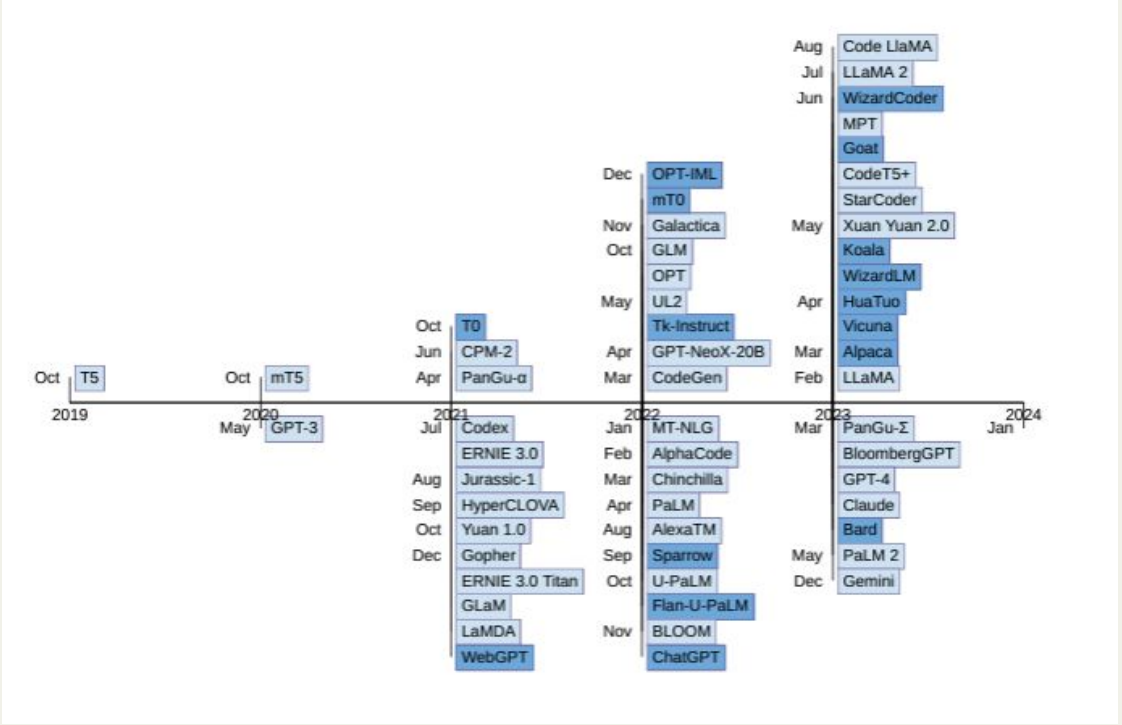
Basic Concepts

- **Tokenization** - The essential preprocessing step of LLMs in which texts are parsed into non-decomposing units of text called tokens (e.g. characters, subwords, symbols)
- **Pre-trained LLMs** - Trained with the objective of next token prediction
  - Has limited capacity to follow user intent; prone to generate bad responses
- **Fine-tuned LLMs** - Continual training on a more specific dataset
  - Resolves some of the issues of pre-trained LLMs



A Timeline of LLM Development

1. **Statistical Modelling** - Early approaches to NLP relied heavily on statistical models, which used probabilistic techniques to analyze and generate text
2. **Neural Language Modeling** -Trained in supervised settings for specific tasks
3. **Pre-trained Language Models (PLM)** - Trained in a self-supervised setting on a large dataset of text
4. **Large Language Models (LLM)** - Significantly increasing the model parameters & training datasets of PLMs !
    a. Capable of emergent abilities (e.g. reasoning, planning, decision-making, zero-shot learning, etc.)



Historical LLM releases

Comprehensive overview of LLMs, 2023

A Timeline of LLM Development

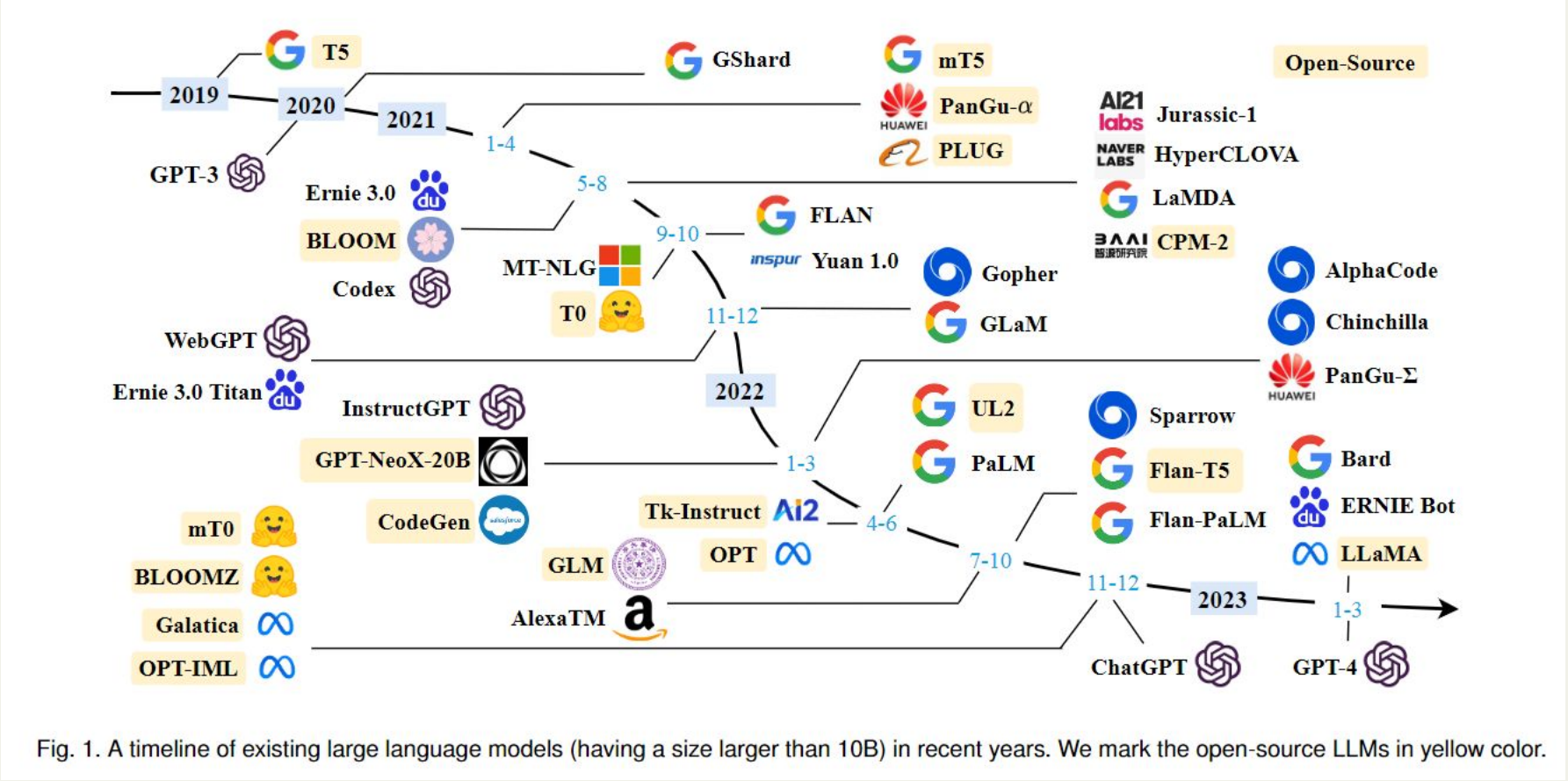


Fig. 1. A timeline of existing large language models (having a size larger than 10B) in recent years. We mark the open-source LLMs in yellow color.

Challenges & Future Directions

1. Computational Cost
   - *How can we improve performance under the constraint of computational resources?*
2. Bias & Fairness
   - *How do we make sure that LLM does not perpetuate harmful biases in their decision-making?*
3. Overfitting
   - *How do we balance between memorization and generalization?*
4. Hallucinations
   - *How can we ensure that LLMs generate realistic and factually accurate content?*
5. Explainability
   - *How can we understand why the LLM produces its specific outputs?*
6. Multi-modality
   - *How can we incorporate info from multiple modalities, such as text, images, and audio?*
7. Limited Knowledge
   - *How can we understand why the LLM produces its specific outputs?*
8. Prompt Engineering
   - *How can we effectively design prompts to elicit the desired responses?*

LLMs for Story Generation

How can we use LLMs to generate stories and how can we evaluate them?

Closed vs. Open-Source LLMs

Closed-Sourced

- **Examples: ChatGPT, AI21**
- Access through API,with usage fees; source code not available.
- Limited to API parameters; cannot modify the underlying model.

Open-Sourced

- **Examples: Mistral, Neuralbeagle**
- Source code and pre-trained models freely available through huggingface
- Full freedom to modify, improve, or adapt the model.

Key Comparisons

- **Performance**:
  - Closed-source models may offer cutting-edge performance but at a cost.
- **Cost**:
  - Open-source models are free to use, but deploying at scale can be expensive.
- **Ethics and Bias**:
  - Companies behind closed-source projects may face regulatory scrutiny based on their industry, pushing them to address ethical issues and bias mitigation.

Choices of Language Models

Chat-GPT 4

- Closed source
- 1.7 trillion parameters

AI21

- Closed source
- 178 billion parameters

Mistral

- Open source
- 7 billion parameters

NeuralBeagel

- Open source
- 7 billion parameters

7B models

- One of the most important AI trends in 2024.
- More efficient and sustainable.
- Lower cost and hardware requirements.
- More accessible.
- More transparent.

Story Generation Prompts

Fantasy

In the hidden folds of modern society lies the ancient werewolf tribe, guardians of humanity, surrounded in mystery and lore. Ellie, the daughter of the tribe's esteemed leader, is about to become an adult. Legend foretold her to rise as the formidable Wolf King, yet fate weaves a different path when she transforms into a little puppy under the full moon's gaze.

Sci-Fi

In the shadowed corridors of the Orion Space Hub, where peace hangs by a thread after an android uprising, Alex and her android companion, Jordan, find solace in their unchanged bond. This solace is shattered when Jordan disappears into the void, leaving behind a trail of questions and a heart full of unspoken love.

Horror

Five friends - Mark, Sarah, Jake, Lily, and Ryan - head off to the infamous Camp Crystal Lake, a place with a terrifying reputation. One by one, a mysterious killer picks them off in the dead of night. Soon, it becomes clear that one of their own is the killer, leaving the survivors unsure who to trust.

Realistic

Sarah lived in a small and quiet town with her daughter Lily. One night, Lily vanished without a trace. As Sarah frantically searched for her child, she found out Lily voluntarily joined a cult for the charismatic leader Christopher. Driven by maternal instincts, Sarah is determined to rescue her daughter.

Prompt example

Sarah lived in a small and quiet town with her daughter Lily. One night, Lily vanished without a trace. As Sarah frantically searched for her child, she found out Lily voluntarily joined a cult for the charismatic leader Christopher. Driven by maternal instincts, Sarah is determined to rescue her daughter.

Characters: Sarah and her daughter Lily, cult leader Christopher
Setting: quiet and small town
Core action: Lily is missing and found in a cult.
Challenges: How Sarah will save her daughter from the cult

Story Evaluation

We generated stories using different LLMs and evaluated them based on some selected criteria

Evaluation Criteria for Rubric

Fluency

The story should be easily readable, with sentences that are free of grammatical errors.

Logicality

The story should adhere to commonsense reasoning and contain events that are plausible

Coherence

The story should maintain a logical flow, with sentences that connect well.

Relevance

The content of the story should be relevant to the given condition or prompt

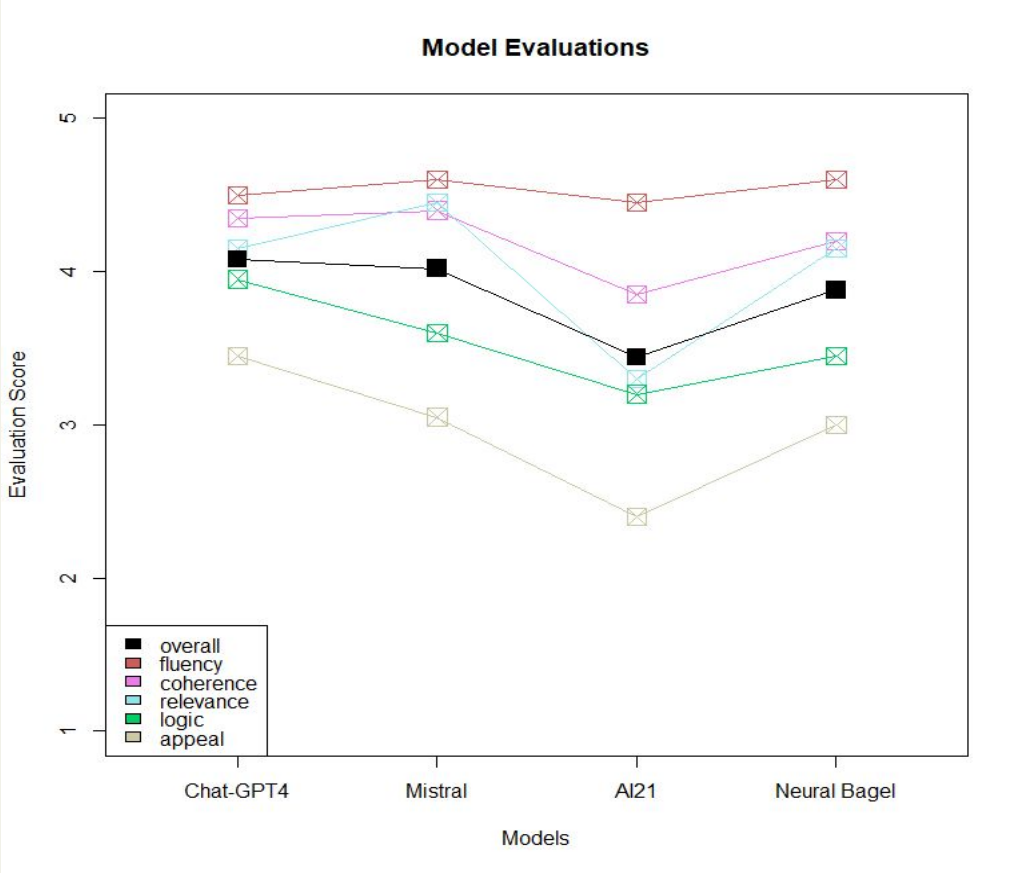
Appeal

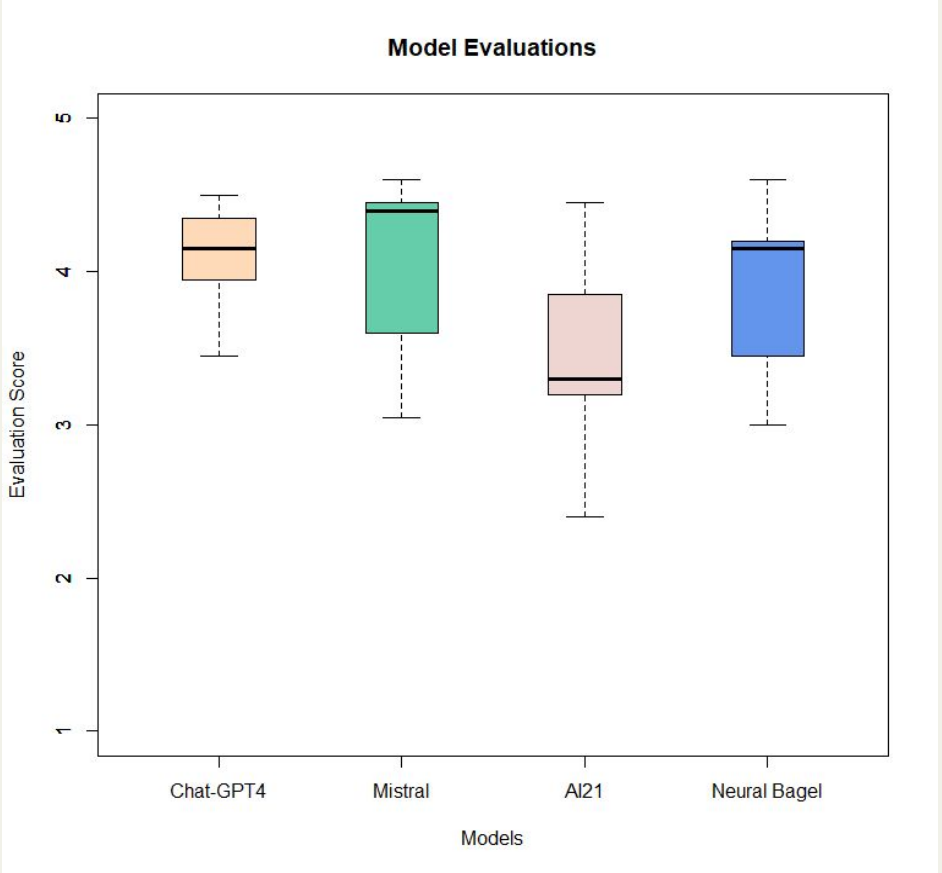
The story should captivate the reader by offering intriguing plot and character developments.

Rubrics

	1	2	3	4	5
Fluency The story should be easily readable, with sentences that are free of grammatical errors.	The story is filled with grammar errors to the point that the content is incomprehensible	The story has many grammar errors that makes comprehension difficult	The story has some grammar errors but do not interfere too much with comprehension	The story has negligible grammar errors that do not at all interfere with comprehension	The story has no grammar issues
Coherence The story should maintain a logical flow of events and ideas, sentences and paragraphs that connect well.	Almost all the sentences in the story are irrelevant or contradictory with each other	The story has many irrelevant or contradictory parts that makes comprehension difficult	The story has some irrelevant or contradictory parts, but comprehension is still possible	Most sentences fit well within the story, with only a few contradictory sentences that are out of place	All the sentences in the story fit together with no logical irrelevance nor contradiction
Relatedness The content of the story should be relevant to the given condition or prompt, maintaining a clear connection to the main theme or title.	The story has no relation with the prompt	The story has a weak relationship the prompt	The story roughly matches the prompt	The story mostly matches the prompt except for one or two small details	The story perfectly matches the prompt
Logicality The story should adhere to commonsense reasoning, with events and actions that are plausible and consistent within the story's universe.	The story is completely absurd and does not match common sense reasoning	The story is mostly absurd with a few things that match common sense	The story basically makes sense	The story mostly makes sense with a few parts that seem absurd	The story completely complies with commonsense and logical reasoning
Interestingness The story should captivate the reader, offering intriguing plot and character developments.	The story is completely boring and read like an instruction manual	There are maybe one or two interesting parts in the story	The story is mildly interesting	The story is interesting and kept your engagement until the end	The story is so exhilaratingly interesting that you wish there could be a sequel

Comparison of the Stories Generated by the Different Models





Learnings

- Gain deeper understanding of Open-source Models
  - Realized there are much more options than ChatGPT
- Accessing UW WatGPU to run open-source models
- Prompt engineering
- A systematic evaluation of stories generated by AI
- 7B models have comparative performance
  - More parameters does not necessarily imply better story-creation abilities
- The current study might be biased due to small sample size
  - A larger experiment could reveal more about the creativity of LLMs of various sizes.
- Teamwork and Time-management

Thank you!

References:

- Bergmann, D. (2024). The most important AI trends in 2024. *IBM Blog*. https://www.ibm.com/blog/artificial-intelligence-trends/
- Fan et al. (2019). Strategies for Structuring Story Generation. *Association for Computational Linguistics*. https://aclanthology.org/P19-1254/
- Naveed et al. (2023). Comprehensive overview of LLMs.
- Xie et al. (2023). Fine-Grained Story Evaluation with Perturbations. *DeltaScore*. https://arxiv.org/pdf/2303.08991.pdf