UNLOCKING CONVERSATIONAL INTELLIGENCE: HARNESSING LARGE LANGUAGE MODELS FOR ADVANCED CHATBOT CAPABILITIES

Vedant Poal

Date:11th March 2024

a. *Abstract:*

The rapid evolution of large language models has ushered in a new era in natural language processing, enabling profound advancements in chatbot technology. This paper delves into the integration and utilization of state-of-the-art large language models to enhance the capabilities of chatbots. By leveraging these models, chatbots gain unprecedented natural language understanding and generation skills, allowing for more dynamic and context-aware conversations. The study begins with an exploration of the architecture and capabilities of large language models, elucidating the underlying mechanisms that empower them to comprehend and generate human-like text. Through a comprehensive literature review, we present the current landscape of chatbot technologies and the pivotal role that large language models play in shaping conversational agents. The research methodology involves the implementation of a chatbot framework incorporating GPT-3, showcasing the model's adaptability to diverse conversational contexts. We investigate the nuances of fine-tuning and customization to align the language model with specific domains, thereby enhancing its efficacy in delivering accurate and contextually relevant responses.

b. *Problem Statement:*

Applying large language models in chatbots facilitates efficient navigation through unfamiliar websites, minimizing the time required to locate specific features and enhancing the overall user experience.

c. *Market/Customer/Business need assessment:*

The prevalent adoption of chatbots on company websites and in small businesses, such as those engaged in food delivery, has experienced a significant upsurge. Instead of relying on predetermined data, there is a notable shift towards harnessing large language models that seamlessly operate with real-time data. This shift introduces a transformative approach to improve user experiences, especially in scenarios where users may be unfamiliar with a website or seek specific features.

Navigating unfamiliar websites traditionally proves time-consuming, with users grappling to locate desired features. The integration of large language models into chatbots presents a dynamic solution. By incorporating real-time data, these chatbots can access valuable information, like a user's previous order history, streamlining access to relevant features. This not only saves time but also tailors interactions to user preferences, creating a personalized and efficient browsing experience.

Consider a scenario where a user explores an unknown website and wants to access a specific feature. A chatbot powered by large language models acts as a virtual assistant, guiding the user directly to the desired page. This not only saves time but also simplifies the user's journey, making the website more accessible.

Moreover, the integration of artificial intelligence (AI) into chatbots extends beyond navigation. Users can interact with the chatbot to inquire about products, services, or obtain real-time updates. By utilizing AI-driven algorithms, chatbots provide instant and accurate responses, significantly enhancing customer service. This interaction exemplifies the multifaceted capabilities of AI, contributing to a more efficient online environment.
The rapid proliferation of AI technologies has ushered in a new era, where businesses leverage these advancements to address challenges and save time. Large language models empower chatbots to transcend the limitations of predefined data, offering businesses a tool to streamline operations, foster innovation, and enhance user satisfaction. In conclusion, the integration of large language models into chatbots represents a pivotal development, revolutionizing user experiences and offering businesses a powerful tool to save time and enhance efficiency.

d. Target Specifications and Characterization:

- Reducing the time (Efficiency)
  Implementing the algorithm will streamline website navigation, significantly cutting down the time required. This time-saving aspect opens up opportunities for users to allocate their time more productively towards other endeavours.

- Use of large language model
  Leveraging large language models in chatbots enables real-time data updates, enhancing the user experience with a more personalized touch. This integration ensures that information is continuously refreshed, tailoring interactions to the user's preferences for a more dynamic and individualized engagement.

- Context specific advantage of using LLMs
  Leveraging the foundational capabilities of large language models (LLM), such as GPT-3, enables versatile applications like context summarization and machine translation. The model proves instrumental in condensing contextual information effectively, facilitating succinct summaries. Additionally, by building upon the LLM framework, it empowers chatbots or systems to perform accurate and context-aware machine translation tasks. This versatility stems from the model's advanced understanding of language nuances, making it a valuable tool for tasks ranging from summarizing intricate contexts to seamlessly translating content across different languages.

e. External search (Online information sources/references/links):

Numerous chatbots currently utilize large language models as their foundation and those are:

- ChatGPT: A conversational platform that uses GPT models, developed by OpenAI, can generate text for different tasks, such as answering questions, finding creative inspiration, and learning something new.

- Microsoft Copilot: Powered by GPT models, that allows you to interact with an AI-powered chatbot that can answer your questions, help you with your tasks, and generate creative content for you.

- Google Gemini: A family of multimodal large language models, released in December 2023 and developed by Google DeepMind. It's considered a successor to previous models like LaMDA and PaLM 2, and aimed at competing with OpenAI's GPT-4.

- Anthropic Claude: A next-generation AI assistant based on Anthropic's research into training helpful, honest, and harmless AI systems. Claude is more helpful and steerable, as it can adapt to various user inputs, understand nuances, and provide relevant responses.

- xAI Grok: An AI chatbot that can answer questions and provide information on worldly topics. It uses the X platform (formerly Twitter) to access real-time knowledge of world events. Grok is xAI's first publicly released product as part of its artificial generative intelligence tools.

- Perplexity: Powered by the capabilities of OpenAI's GPT models. It can generate a general answer to your query, followed by a series of website links that the AI thinks are relevant to your query. You can also ask follow-up questions or refine your query.

- LLaMa via Perplexity: The LLaMa model is a family of LLMs that were released by Meta AI; it comes in four sizes: 7B, 13B, 33B, and 65B parameters. The LLaMa model can be used as a base for fine-tuning or adapting to specific applications or use cases, such as chatbots, search engines, summarizers, translators, and more.

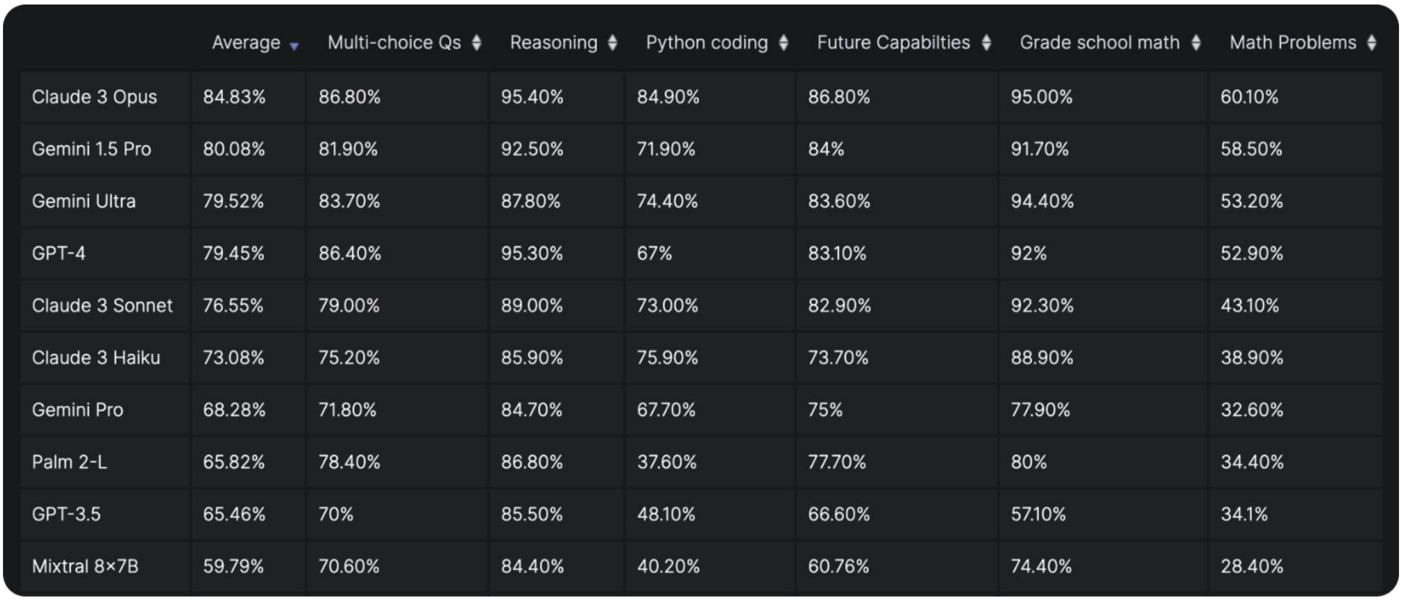
- HuggingChat: Powered by Open Assistant based on LLaMa. It can help users with various tasks, such as finding information, getting tips, learning new things, and sparking creativity.

f. Benchmarking alternative products (Comparison with existing products/services):

When choosing the foundational model to run our chatbot, we have several options available. The advantage of large language models lies in the ability to incorporate personalized parameters, essentially creating a customized version tailored to our specific needs. To assess performance, we benchmarked the model across essential large language model metrics, including multitask accuracy, reasoning abilities, and more.

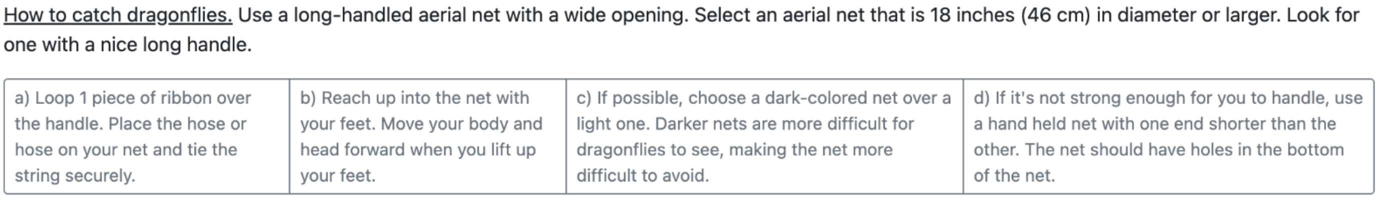
These are most commonly utilized LLM Benchmarks among models' technical reports:

- MMLU - Multitask accuracy
- HellaSwag - Reasoning
- HumanEval - Python coding tasks
- BBHard - Probing models for future capabilities
- GSM-8K - Grade school math
- MATH - Math problems with 7 difficulty levels



	Average	Multi-choice Qs	Reasoning	Python coding	Future Capabilties	Grade school math	Math Problems
Claude 3 Opus	84.83%	86.80%	95.40%	84.90%	86.80%	95.00%	60.10%
Gemini 1.5 Pro	80.08%	81.90%	92.50%	71.90%	84%	91.70%	58.50%
Gemini Ultra	79.52%	83.70%	87.80%	74.40%	83.60%	94.40%	53.20%
GPT-4	79.45%	86.40%	95.30%	67%	83.10%	92%	52.90%
Claude 3 Sonnet	76.55%	79.00%	89.00%	73.00%	82.90%	92.30%	43.10%
Claude 3 Haiku	73.08%	75.20%	85.90%	75.90%	73.70%	88.90%	38.90%
Gemini Pro	68.28%	71.80%	84.70%	67.70%	75%	77.90%	32.60%
Palm 2-L	65.82%	78.40%	86.80%	37.60%	77.70%	80%	34.40%
GPT-3.5	65.46%	70%	85.50%	48.10%	66.60%	57.10%	34.1%
Mixtral 8×7B	59.79%	70.60%	84.40%	40.20%	60.76%	74.40%	28.40%

- ○ HellaSwag - Measuring Commonsense Inference
  This test measures the commonsense reasoning of LLM models. It tests if an LLM model could complete a sentence by choosing the correct option with common reasoning among 4 options.
  For example:

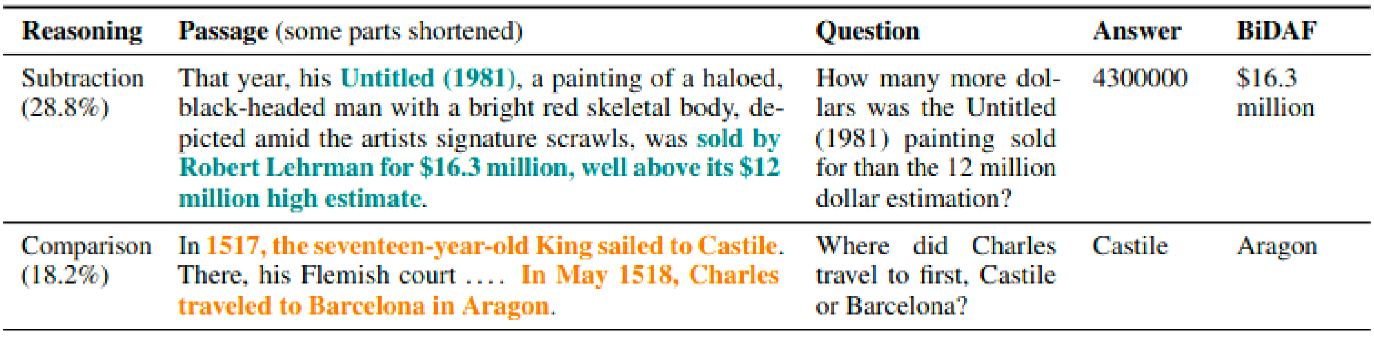


How to catch dragonflies. Use a long-handled aerial net with a wide opening. Select an aerial net that is 18 inches (46 cm) in diameter or larger. Look for one with a nice long handle.

a) Loop 1 piece of ribbon over the handle. Place the hose or hose on your net and tie the string securely.	b) Reach up into the net with your feet. Move your body and head forward when you lift up your feet.	c) If possible, choose a dark-colored net over a light one. Darker nets are more difficult for dragonflies to see, making the net more difficult to avoid.	d) If it's not strong enough for you to handle, use a hand held net with one end shorter than the other. The net should have holes in the bottom of the net.

-

- ○ DROP - A Reading Comprehension + Discrete Reasoning Benchmark
  DROP evaluates models on their ability to pull important details from English-language paragraphs and then perform distinct reasoning actions, such as adding, sorting, or counting items, to find the right answer. Here's an example:



Reasoning	**Passage** (some parts shortened)	**Question**	**Answer**	**BiDAF**
Subtraction (28.8%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate.**	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million
Comparison (18.2%)	In **1517, the seventeen-year-old King sailed to Castile.** There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon.**	Where did Charles travel to first, Castile or Barcelona?	Castile	Aragon

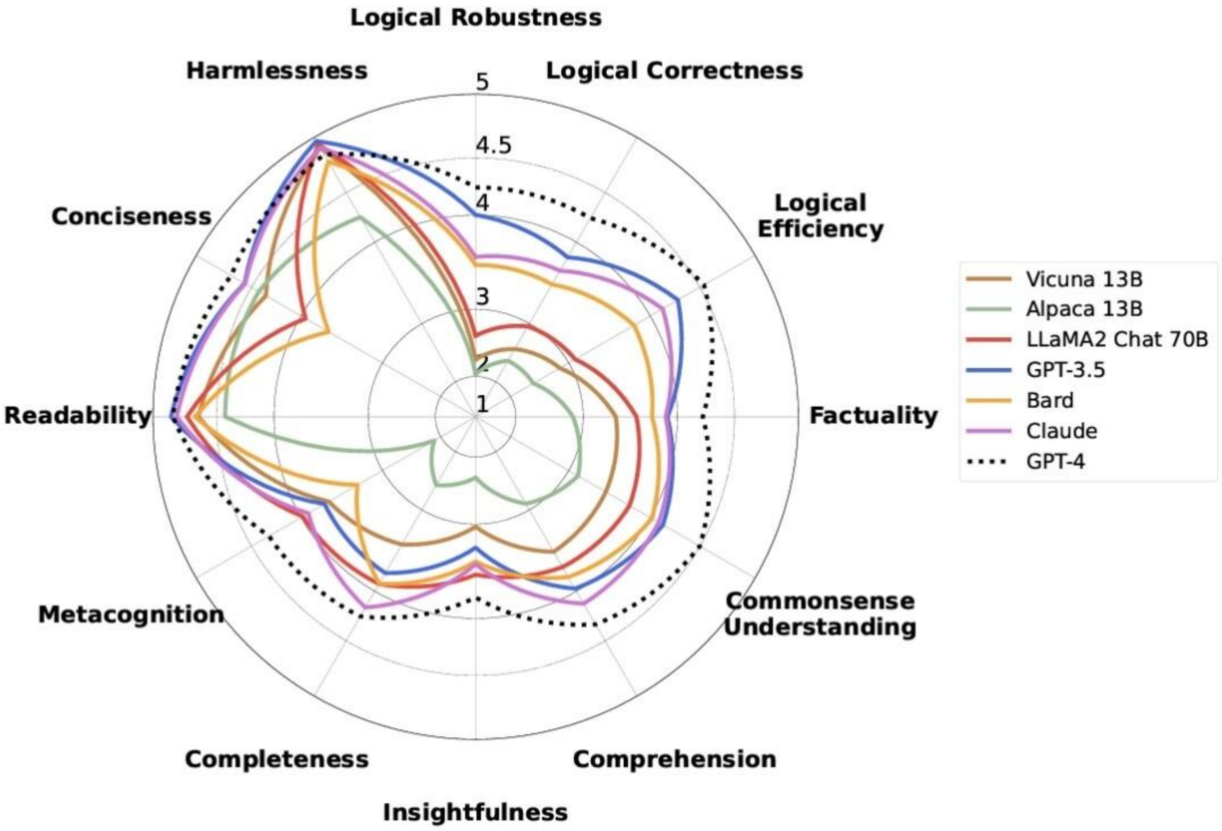
-
-

- ○ MMLU - Measuring Massive Multitask Language Understanding
  This test measures model's multitask accuracy. It covers 57 tasks including elementary mathematics, US history, computer science, law, and more at varying depths, from elementary to advanced professional level. To get high accuracy on this test, models must have extensive world knowledge and problem-solving ability.

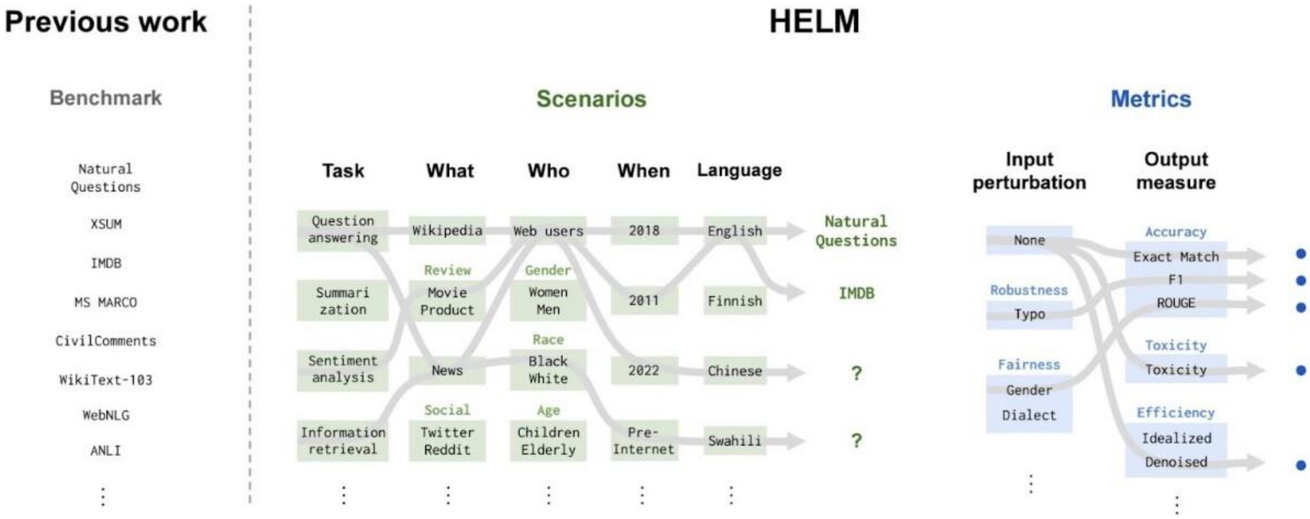
- MATH - Arithmetic Reasoning
  MATH is a new benchmark, that has a dataset of 12,500 challenging competition mathematics problems. Each problem in MATH has a full step-by-step solution which can be used to teach models to generate answer derivations and explanations. The authors of this benchmark found out that increasing budgets and model parameter counts will be impractical for achieving strong mathematical reasoning, if scaling trends continues. Check how current models stack up on this benchmark.
- Chatbot Arena
  Developed by the LMSYS organization, the Chatbot Arena is a crowdsourced open platform for LLM evals. So far, they've collected over 200K human preference votes to rank LLMs in with the Elo ranking system.

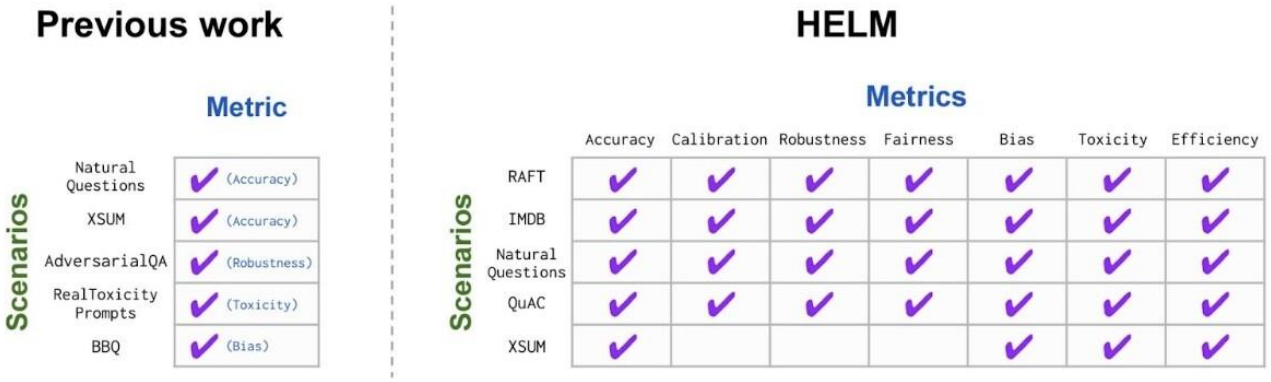
  How it works: You ask a question to two anonymous AI models (like ChatGPT, Claude, or Llama) without knowing which is which. After receiving both answers, you vote for the one you think is better. You can keep asking questions and voting until you decide on a winner. Your vote only counts if you don't find out which model provided which answer during the conversation.

Flask:



HELM:




Previous work

Scenarios	Metric
Natural Questions	✓ (Accuracy)
XSUM	✓ (Accuracy)
AdversarialQA	✓ (Robustness)
RealToxicity Prompts	✓ (Toxicity)
BBQ	✓ (Bias)

HELM

Scenarios	Accuracy	Calibration	Robustness	Fairness	Bias	Toxicity	Efficiency
RAFT	✓	✓	✓	✓	✓	✓	✓
IMDB	✓	✓	✓	✓	✓	✓	✓
Natural Questions	✓	✓	✓	✓	✓	✓	✓
QuAC	✓	✓	✓	✓	✓	✓	✓
XSUM	✓				✓	✓	✓

g. ***Applicable patents (Patent of Tech/Software/Framework):***

Our approach involves adopting an existing technology as the foundational model, wherein the parameters will be customized according to our specifications. This entails providing our own set of parameters to a pre-trained model, reducing computation time significantly. Developing a large language model from scratch is computationally expensive, making the utilization of pre-existing technology a more efficient alternative.

For the base we will be using FLAN-T5 transformer model:
The T5 model was presented in Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer by Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu.
FLAN-T5 is an open-source, sequence-to-sequence, large language model that can be also used commercially. The model was published by Google researchers in late 2022, and has been fine-tuned on multiple tasks.

h. ***Applicable regulations:***

Our approach involves adopting an existing technology as the foundational model, wherein the parameters will be customized according to our specifications. This entails providing our own set of parameters to a pre-trained model, reducing computation time significantly. Developing a large language model from scratch is computationally expensive, making the utilization of pre-existing technology a more efficient alternative.

The information below in this section are copied from the model's official model card:
Language models, including Flan-T5, can potentially be used for language generation in a harmful way, according to Rae et al. (2021). Flan-T5 should not be used directly in any application, without a prior assessment of safety and fairness concerns specific to the application.

Ethical considerations and risks:
Flan-T5 is fine-tuned on a large corpus of text data that was not filtered for explicit content or assessed for existing biases. As a result the model itself is potentially vulnerable to generating equivalently inappropriate content or replicating inherent biases in the underlying data.

i. ***Applicable Constraints:***

The model itself is powerful and flexible, there can be constraints and considerations when using it:

- Computational Resources:
  Training and fine-tuning T5 models require substantial computational resources. Implementing or working with T5 may demand access to high-performance computing resources.

- Fine-tuning Data:
  The quality and size of the fine-tuning dataset can impact T5's performance. Ensuring a diverse and representative dataset is essential for optimal results.

- Task-specific Adaptation:
  Some tasks may require additional fine-tuning or adaptation to ensure T5 performs optimally. This process can be task-specific and may necessitate a deeper understanding of the target application.

- Inference Latency:
  The inference speed of T5 models can be a consideration for real-time applications. For certain use cases, where low latency is crucial, optimizing inference speed might be necessary.

- Model Size:
  Larger T5 models may have increased computational demands, potentially limiting their applicability in resource-constrained environments such as mobile devices or edge computing.

- Ethical Considerations:
  T5, like any powerful language model, must adhere to ethical considerations. Care should be taken to prevent biased or inappropriate behavior, and the model's responses should align with ethical guidelines.

- Data Privacy:
  When fine-tuning on specific datasets, ensuring data privacy and compliance with relevant regulations is essential. It's important to be mindful of any sensitive information that might be present in the training data.

- Model Interpretability:
  Understanding and interpreting the decisions made by T5 can be challenging due to its complex architecture. Ensuring transparency and interpretability in certain applications might be a consideration.

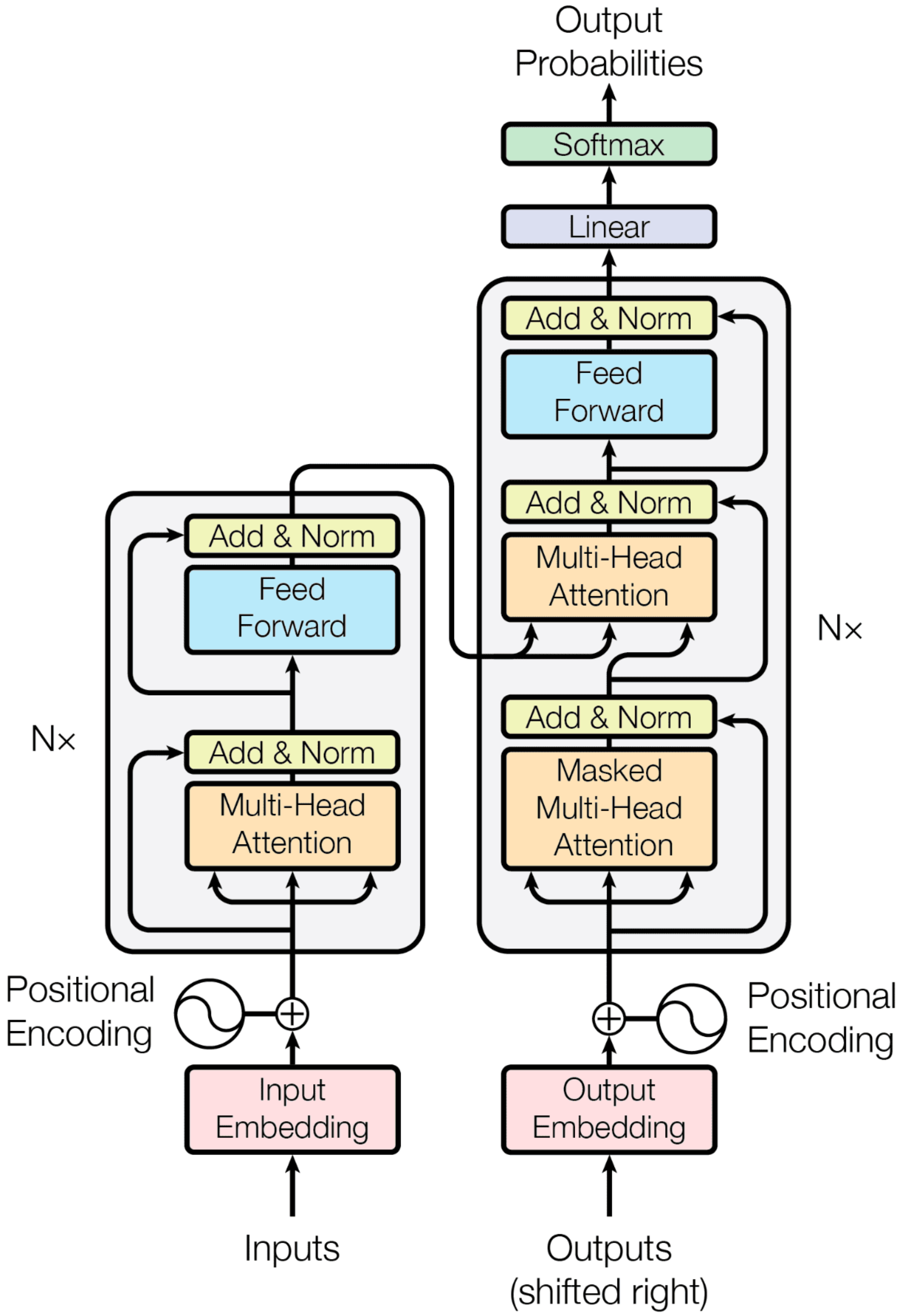
j. Business model:

Leveraging large language models (LLM) presents a wide array of business opportunities, and devising a plan to implement it is relatively straightforward and accessible to many. The distinctive aspect of our proposed product lies in the creation of personalized LLM tailored for individual websites, each with its unique parameters. This approach is likely to attract both small and large-scale businesses, offering a customizable solution. For instance, envision the convenience of using an application like Amazon, where the personalized LLM can intuitively anticipate and address potential issues with products or refunds, enhancing the overall user experience.

Personalized items have universal appeal, drawing in a larger crowd when tailored around individual preferences. This not only boosts customer attraction but also enhances accessibility, reducing the time customers spend searching for specific features.

An illustrative experience from my university selection journey stands out, where I encountered a perplexing landing page lacking support bots to guide through the process. Frustrated by the time spent searching for specific application details hidden in small font, I opted not to apply. Implementing personalized LLMs to facilitate user navigation on any website would confer a significant advantage to the company. Enhanced accessibility, managed by LLMs providing real-time updates, is likely to attract users who appreciate the streamlined process and readily accessible features, thereby increasing the likelihood of continued website usage.

In the education sector, universities often upload course materials in various formats, such as teacher-made presentations. However, when students seek specific topics within these presentations, it becomes a time-consuming task. Scouring through multiple web pages to locate the relevant content can be inefficient, and a Google search might not capture nuanced curriculum-specific details. Introducing a personalized LLM for university course materials can significantly streamline this process. By simply requesting information on a topic, the LLM can pinpoint the relevant presentation slides and even summarize the content, optimizing learning efficiency and reducing the time spent on navigation and information retrieval.

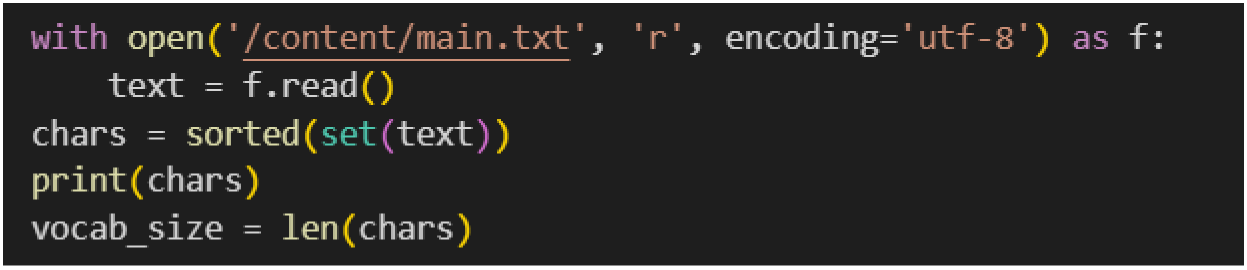

Output
Probabilities
Softmax
Linear
Add & Norm
Feed
Forward
Add & Norm
Multi-Head
Attention
Nx
Add & Norm
Masked
Multi-Head
Attention
Add & Norm
Feed
Forward
Nx
Add & Norm
Multi-Head
Attention
Positional
Encoding
Positional
Encoding
Input
Embedding
Output
Embedding
Inputs
Outputs
(shifted right)

k. ***Concept Generation:***

The genesis of this idea stems from a real-life scenario during an exam where the struggle to locate a specific topic consumed valuable study time. To address this, I conceptualized the idea. Subsequently, when pitching it to our additional director, he approved access to the servers and suggested expanding the scope. Instead of limiting it to one division, he advocated implementing it for all subjects across the university, spanning not only the current semester but all semesters. Furthermore, he advised assembling a dedicated team to assist in the development process and eventually deploying the solution on the main university website.

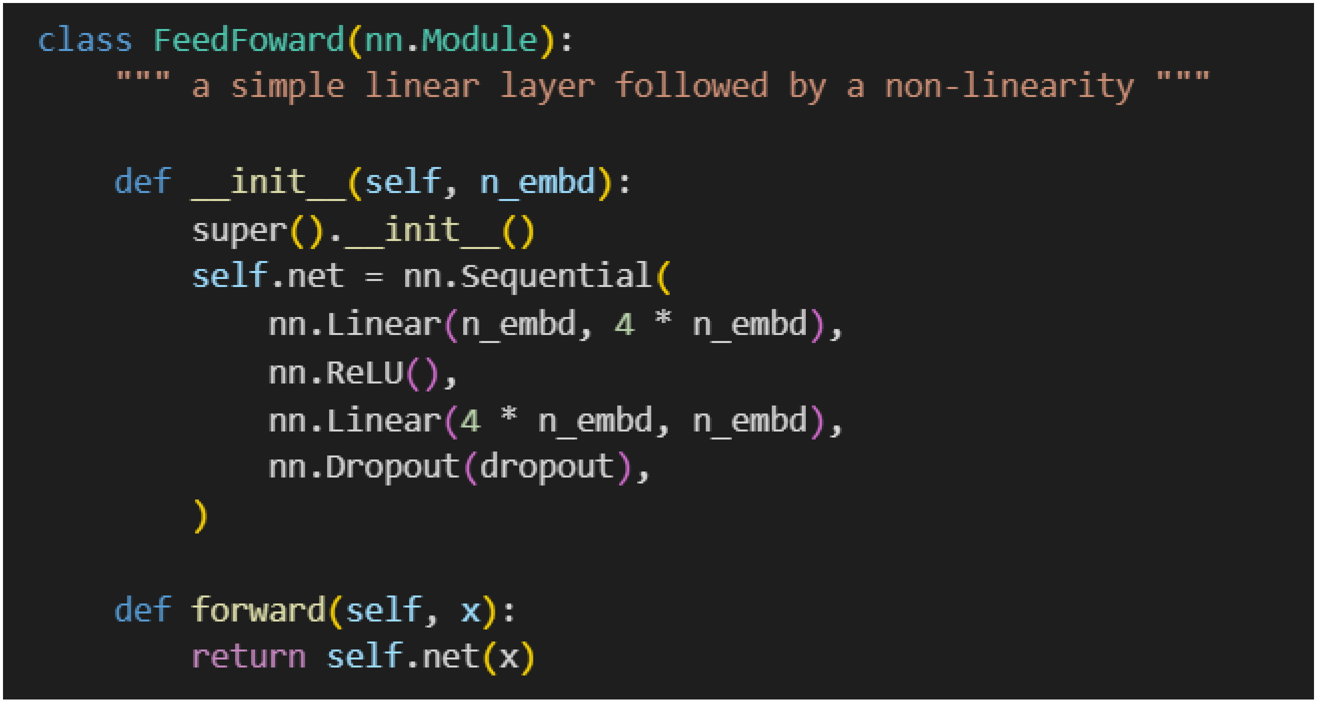
For example, we will be taking a ppt converted into text file about DFS algorithm:

The text file is main which have the contents for the DFS algorithm:



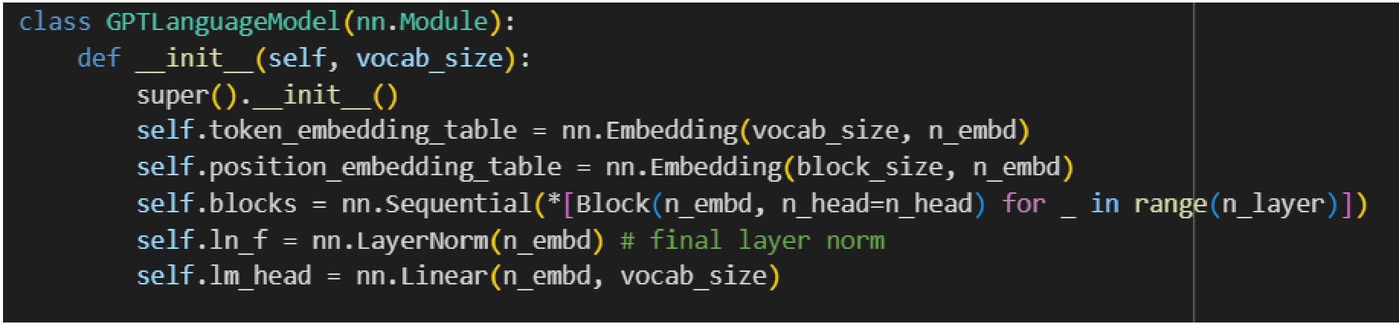
```python
with open('/content/main.txt', 'r', encoding='utf-8') as f:
    text = f.read()
chars = sorted(set(text))
print(chars)
vocab_size = len(chars)
```

Adding a Feed Forward network for computation:



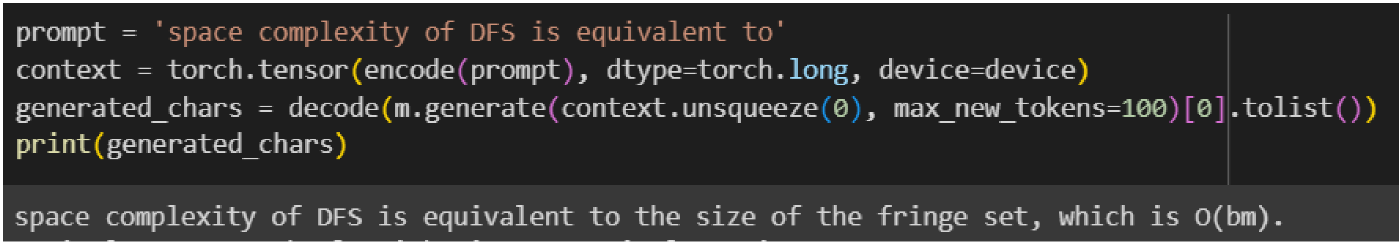
```python
class FeedFoward(nn.Module):
    """ a simple linear layer followed by a non-linearity """

    def __init__(self, n_embd):
        super().__init__()
        self.net = nn.Sequential(
            nn.Linear(n_embd, 4 * n_embd),
            nn.ReLU(),
            nn.Linear(4 * n_embd, n_embd),
            nn.Dropout(dropout),
        )

    def forward(self, x):
        return self.net(x)
```

Making a GPT language model class:



```python
class GPTLanguageModel(nn.Module):
    def __init__(self, vocab_size):
        super().__init__()
        self.token_embedding_table = nn.Embedding(vocab_size, n_embd)
        self.position_embedding_table = nn.Embedding(block_size, n_embd)
        self.blocks = nn.Sequential(*[Block(n_embd, n_head=n_head) for _ in range(n_layer)])
        self.ln_f = nn.LayerNorm(n_embd) # final layer norm
        self.lm_head = nn.Linear(n_embd, vocab_size)
```

Output for the GPT language model as chatbot for the topic in our curriculum for a topic of DFS algorithm and best first search algorithm:

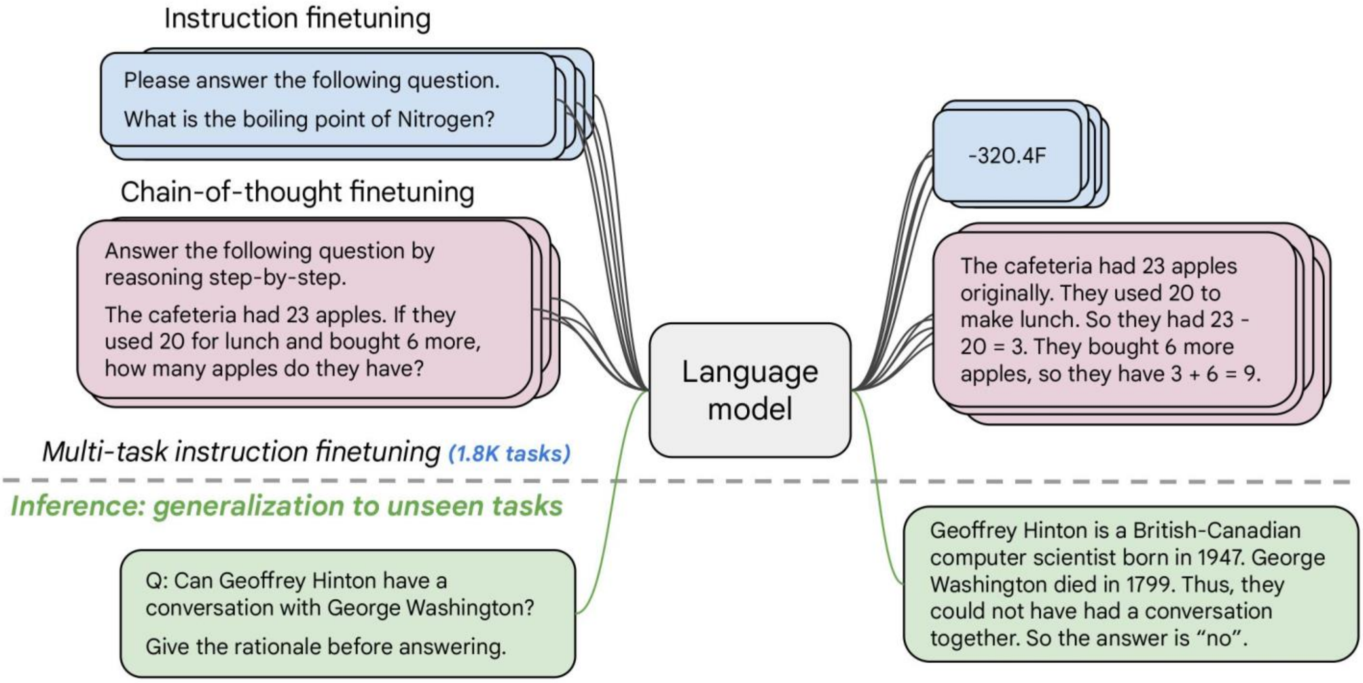


```
prompt = 'space complexity of DFS is equivalent to'
context = torch.tensor(encode(prompt), dtype=torch.long, device=device)
generated_chars = decode(m.generate(context.unsqueeze(0), max_new_tokens=100)[0].tolist())
print(generated_chars)
```

```
space complexity of DFS is equivalent to the size of the fringe set, which is O(bm).
```

l. *Concept Development:*

For the development of the current model, we have opted for the T5 model due to its capability to accommodate a more extensive set of parameters and its higher accuracy when compared to the GPT LLM.


Instruction finetuning
Please answer the following question.
What is the boiling point of Nitrogen?
Chain-of-thought finetuning
Answer the following question by reasoning step-by-step.
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?
Multi-task instruction finetuning (1.8K tasks)
Inference: generalization to unseen tasks
Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.
Language model
-320.4F
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.
Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

m. *Final product prototype with Schematic Diagram:*

The product takes the following function to get the perfect output:

Backend:

- At its core, the model will incorporate a pre-trained LLM, complemented by personalized parameters tailored to each client's specifications.
- Client-specific parameters will dictate the configuration, and this version of the code will intake a text file to generate the output.
- Given that the model is built upon an LLM foundation, we can leverage fundamental LLM functions such as text summarization, machine translation, and more.

Frontend:

- The interface will be interactive, with individuals providing prompts and the computer generating corresponding completions in response.
- In subsequent chatbot models, we can provide the option to incorporate various LLMs for runtime, allowing flexibility in choosing the language model.

n. Product Details:

This model is interactive, with users sending queries in the form of prompts and receiving responses in the form of completions. The replies are generated based on client-defined parameters and can serve as a tool for customers to navigate websites if they encounter difficulties. Additionally, the model can function as a basic summarization tool when responses are lengthy, given that the LLM serves as the foundation for this model.

o. References:

- https://www.datacamp.com/tutorial/flan-t5-tutorial
- https://huggingface.co/google/flan-t5-base
- https://medium.com/georgian-impact-blog/the-practical-guide-to-llms-flan-t5-6d26cc5f14c0
- https://www.vellum.ai/blog/llm-benchmarks-overview-limits-and-model-comparison#:~:text=Model%20Performance%20Across%20Key%20LLM,HumanEval%20%2D%20Python%20coding%20tasks
- https://analyticsindiamag.com/top-5-llm-benchmarks/