DS8008
TMU

HuggingGPT:

Kunal Gurnani, Murad Taher

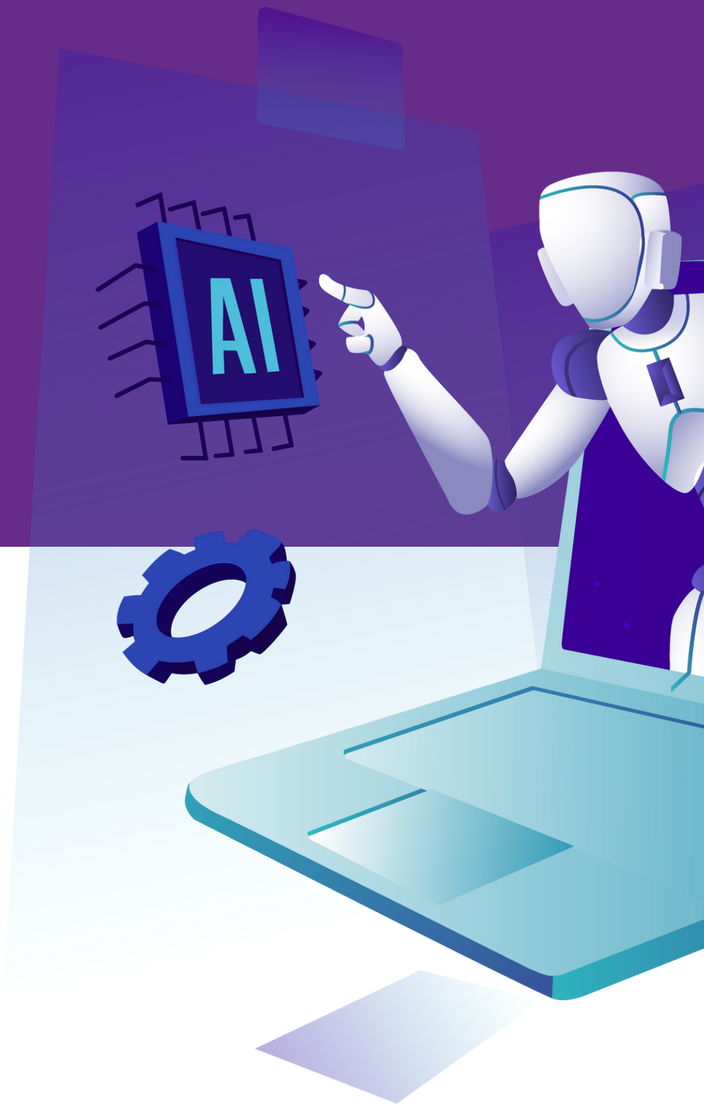
“Solving AI Tasks in the new age of artificial technology.”

mtaher@torontomu.ca, kunal.gurnani@torontomu.ca

Natural Language Processing

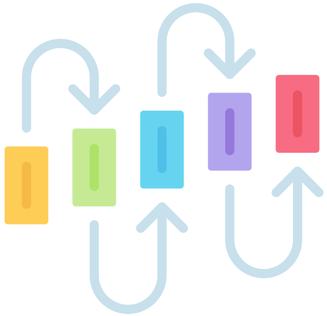
PROBLEM

Current LLMs lack the ability to process complex information and coordinate multiple models to solve complex tasks.



Vision and Speech

Modern models to some extent still have an inability to process complicated modalities of info like vision and speech.



Multi-model Complexity

Coordinating multiple models to solve a task that has a number of sub-objectives first is not easily accessible in the present.

BACKGROUND

Reference	Explanation	Dataset/Input	Weakness
Alayrac, Jean-Baptiste, et al. [1]	They propose a Vision Language Model named Flamingo that can perform open-ended vision and language tasks	M3W, ALIGN, LTIP, and VTP datasets	Performance lags behind on classification tasks, trade-offs of few-shot learning methods, hallucination
Huang, Shaohan, et al. [2]	They propose Kosmos-1, a multimodal large language model that can perceive general modalities, learn in context and follow instructions	The Pile, Common Crawl, English LAION-2B, LAION-400M, and COYO-700M datasets	Worse performance in zero-shot and one-shot language tasks compared to a baseline LLM
Shen, Yongliang, et al. [3]	They propose an LLM-powered agent that disassembles tasks based on requests and assigns suitable models to the tasks	Requests submitted by annotators	Requires multiple interactions with LLMs thus increasing time costs and monetary costs

METHODOLOGY

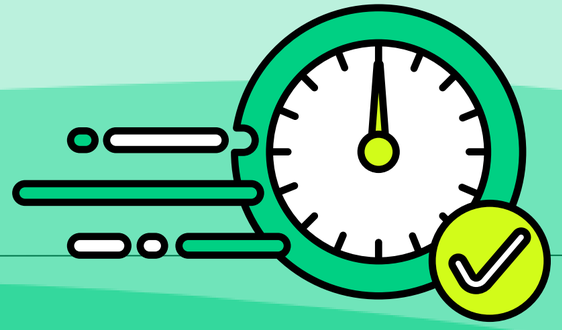
The whole process of HuggingGPT can be divided into four stages.

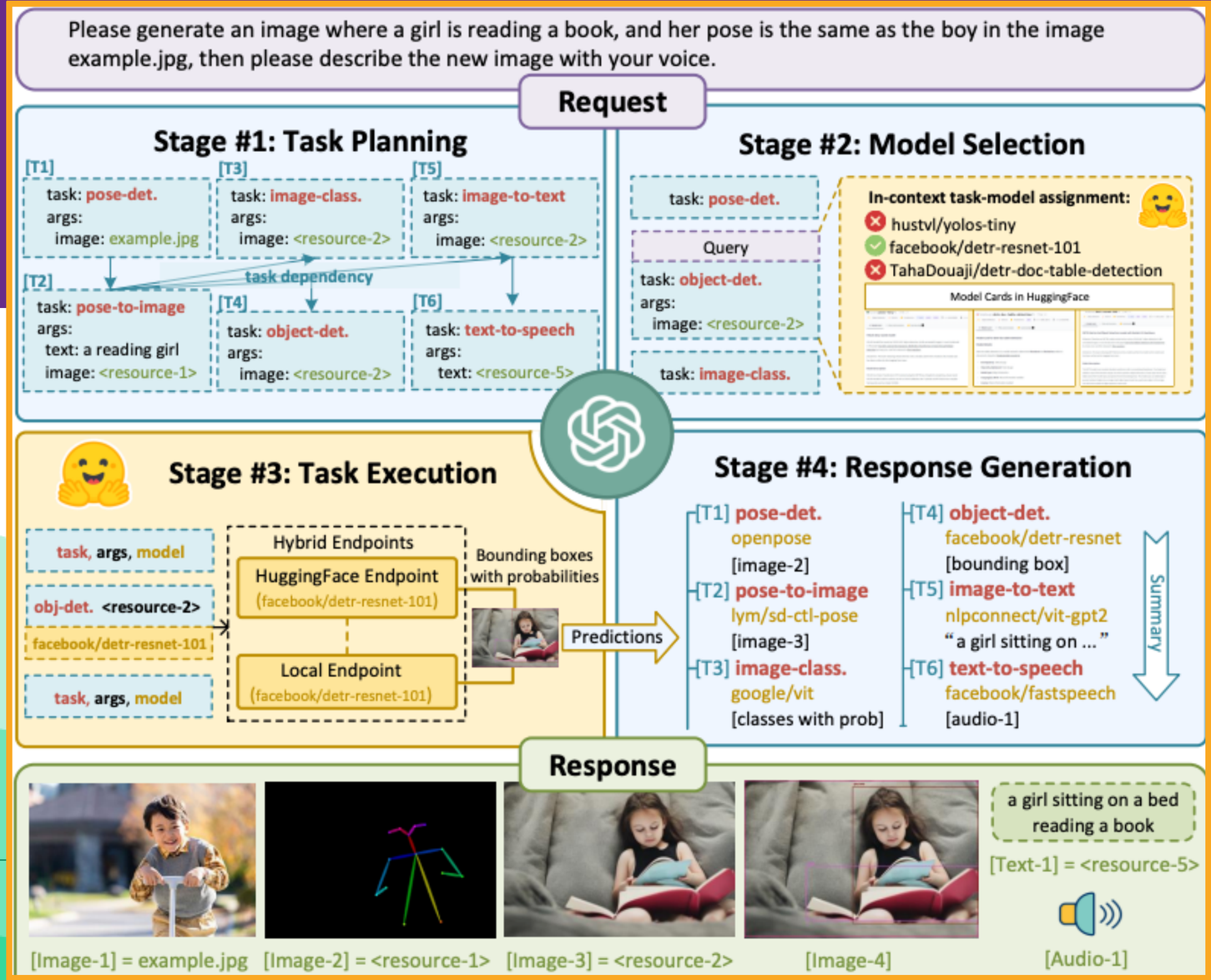
1.Task Planning

2.Model Selection

3.Task Execution

4.Response Gen





IMPLEMENTATION

HuggingGPT

A system to connect LLMs with ML community. See our Project and Paper.

Duplicate Space

Duplicate the Space and run securely with your OpenAI API Key and Hugging Face Token

Note: Only a few models are deployed in the local inference endpoint due to hardware limitations. In addition, online HuggingFace inference endpoints may sometimes not be available. Thus the capability of HuggingGPT is limited.

Submit

Submit

Chatbot

Here's a detailed explanation of the process:

1. For image A (/examples/a.jpg), I used the 'ydshieh/vit-gpt2-coco-en' model for image-to-text conversion and 'facebook/detr-resnet-101' for object detection. The visual-question-answering task was performed using 'dandelin/vilt-b32-finetuned-vqa'. The inference results indicated that there were no zebras in this image.
2. For image B (/examples/b.jpg), the same models were used. The object detection model identified a zebra in the image, and the visual-question-answering task confirmed that there was one zebra in the image.
3. For image C (/examples/c.jpg), again the same models were used. The object detection model identified two zebras in the image, and the visual-question-answering task confirmed that there were two zebras in the image.

Please note that the generated images with predicted boxes are stored at /images/9dce.jpg for image A, /images/3c17.jpg for image B, and /images/2a9e.jpg for image C.

I hope this answers your question. If you have any other requests or need further clarification, feel free to ask!


```
{
 0:  {
   task:  {
     task:  "image-to-text",
     id:  0,
     dep:  [
       0:  -1
     ],
     args:  {
       image:  "public//examples/a.jpg"
     }
   },
   inference result:  {
     generated text:  "a cat sitting on a window sill looking out "
   },
   choose model result:  {
     id:  "ydshieh/vit-gpt2-coco-en",
     reason:  "Only one model available."
   }
 },
 1:  {
   task:  {
     task:  "object-detection",
     id:  1,
     dep:  [
       0:  -1
```


Enter text and press enter. The url must contain the media type. e.g, https://example.com/example.jpg

Send

HUMAN EVALUATION

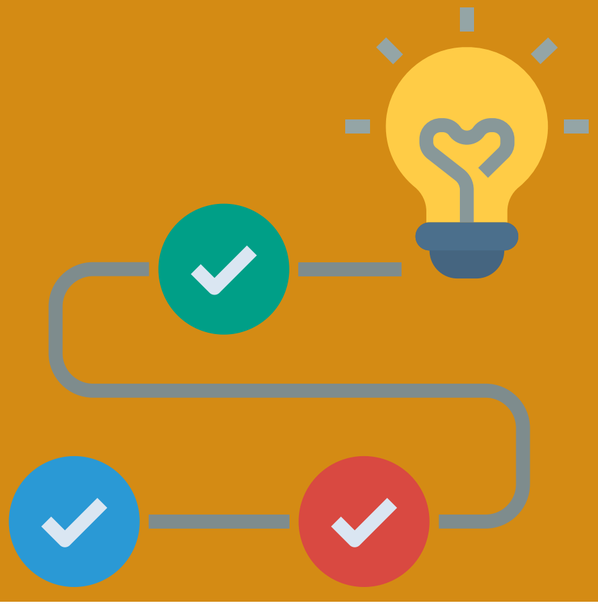
LLM	Task Planning		Model Selection		Response
	Passing Rate ↑	Rationality ↑	Passing Rate ↑	Rationality ↑	Success Rate↑
Alpaca-13b	51.04	32.17	-	-	6.92
Vicuna-13b	79.41	58.41	-	-	15.64
GPT-3.5	91.22	78.47	93.89	84.29	63.08

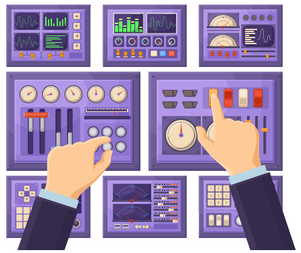
Table 8: Human Evaluation on different LLMs. We report two metrics, passing rate (%) and rationality (%), in the task planning and model selection stages and report a straightforward success rate (%) to evaluate whether the request raised by the user is finally resolved.

- **Passing Rate**: to determine whether the planned task graph or selected model can be successfully executed;
- **Rationality**: to assess whether the generated task sequence or selected tools align with user requests in a rational manner;
- **Success Rate**: to verify if the final results satisfy the user's request.

CONCLUSIONS

What insights and final things can we say about HuggingGPT?



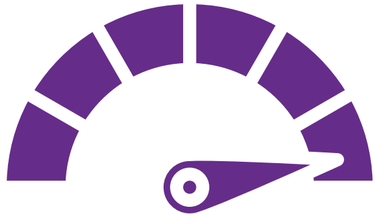


Successful Controller

HuggingGPT has established itself as a successful controller with results that are promising. Truly, a hope for autonomy.

Increased the Ceiling of AI

The use of Hugging Face and HuggingGPT as the respective source of models and controller pushes forward potential.

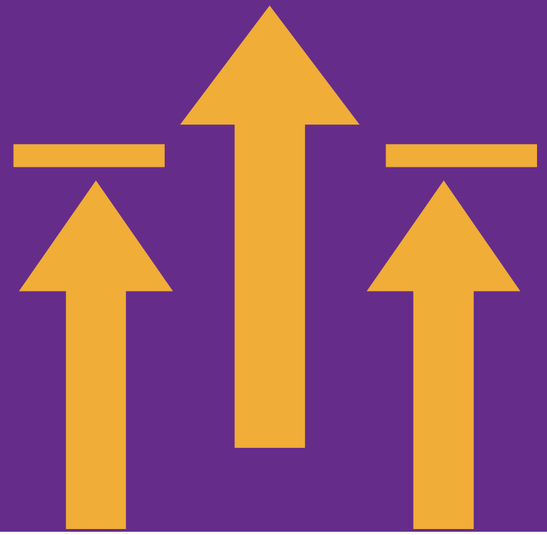


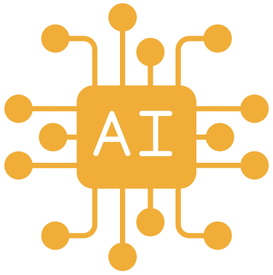
Limits Are Apparent

Despite the promising results that seem to be put forward by the study and implementation, there is some left to be done.

LIMITATIONS FOR NOW

What are the general future steps that can be taken with regards to further study and an understanding of the limitations that exist?





LLM Based Ceilings

Large Language Models that exist today naturally have limits attached such as token limit length and efficiency maximizations that exist.

Inherent Complexity

The structure that this model employs is naturally quite complex and so there are going to be inefficient interactions that exist.