HUGGINGGPT

JARVIS

By Microsoft

Task	Args
Text-cls	text
Token-cls	text
Text2text-generation	text
Summarization	text
Translation	text
Question-answering	text
Conversational	text
Text-generation	text
Tabular-cls	text

Table 1: NLP tasks.

Task	Args
Image-to-text	image
Text-to-image	image
VQA	text + image
Segmentation	image
DQA	text + image
Image-cls	image
Image-to-image	image
Object-detection	image
Controlnet-sd	image

Table 2: CV tasks.

Task	Args
Text-to-speech	text
Audio-cls	audio
ASR	audio
Audio-to-audio	audio

Table 3: Audio tasks.

Task	Args
Text-to-video	text
Video-cls	video

Table 4: Video tasks.

HUGGINGGPT

JARVIS

By Microsoft

Task Planning

Prompt

#1 Task Planning Stage - The AI assistant can parse user input to several tasks: [{"task": task, "id", task_id, "dep": dependency_task_ids, "args": {"text": text, "image": URL, "audio": URL, "video": URL}}]. The "dep" field denotes the id of the previous task which generates a new resource that the current task relies on. A special tag "<resource>-task_id" refers to the generated text image, audio and video in the dependency task with id as task_id. The task MUST be selected from the following options: {{ Available Task List }}. There is a logical relationship between tasks, please note their order. If the user input can't be parsed, you need to reply empty JSON. Here are several cases for your reference: {{ Demonstrations }}. The chat history is recorded as {{ Chat Logs }}. From the chat logs, you can find the path of the user-mentioned resources for your task planning.

Demonstrations

Look at /exp1.jpg, Can you tell me how many objects in the picture?	[{"task": "image-to-text", "id": 0, "dep": [-1], "args": {"image": "/exp1.jpg" }}, {"task": "object-detection", "id": 0, "dep": [-1], "args": {"image": "/exp1.jpg" }}]
In /exp2.jpg, what's the animal and what's it doing?	[{"task": "image-to-text", "id": 0, "dep":[-1], "args": {"image": "/exp2.jpg" }}, {"task":"image-classification", "id": 1, "dep": [-1], "args": {"image": "/exp2.jpg" }}, {"task":"object-detection", "id": 2, "dep": [-1], "args": {"image": "/exp2.jpg" }}, {"task": "visual-question-answering", "id": 3, "dep":[-1], "args": {"text": "What's the animal doing?", "image": "/exp2.jpg" }}]
Given an image /exp3.jpg, first generate a hed image, then based on the hed image and a prompt: a girl is reading a book, you need to reply with a new image.	[{"task": "image-to-text", "id": 0, "dep": [-1], "args": {"image": "/examples/boy.jpg" }}, {"task": "openpose-control", "id": 1, "dep": [-1], "args": {"image": "/examples/boy.jpg" }}, {"task": "openpose-text-to-image", "id": 2, "dep": [1], "args": {"text": "a girl is reading a book", "image": "<resource>-1" }}]

Model Selection

Prompt

#2 Model Selection Stage - Given the user request and the call command, the AI assistant helps the user to select a suitable model from a list of models to process the user request. The AI assistant merely outputs the model id of the most appropriate model. The output must be in a strict JSON format: "id": "id", "reason": "your detail reason for the choice". We have a list of models for you to choose from {{ Candidate Models }}. Please select one model from the list.

Candidate Models

{"model_id": model id #1, "metadata": meta-infos #1, "description": description of model #1}

{"model_id": model id #2, "metadata": meta-infos #2, "description": description of model #2}

[Select top-K candidate models with the same task type as the current task]

{"model_id": model id #K, "metadata": meta-infos #K, "description": description of model #K}

Response Generation

Prompt

#4 Response Generation Stage - With the input and the inference results, the AI assistant needs to describe the process and results. The previous stages can be formed as - User Input: {{ User Input }}, Task Planning: {{ Tasks }}, Model Selection: {{ Model Assignment }}, Task Execution: {{ Predictions }}. You must first answer the user's request in a straightforward manner. Then describe the task process and show your analysis and model inference results to the user in the first person. If inference results contain a file path, must tell the user the complete file path. If there is nothing in the results, please tell me you can't make it.

HUGGINGGPT

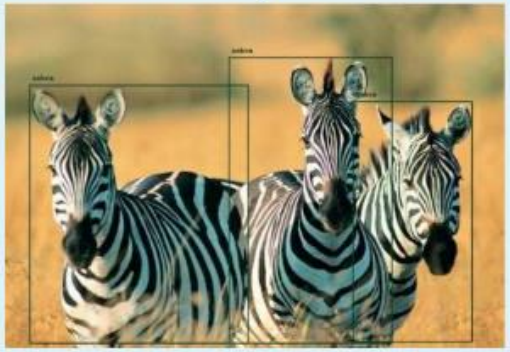
JARVIS

By Microsoft

Query: Given a collection of images A: /examples/a.jpg, B: /examples/b.jpg, C: /examples/c.jpg, please tell me how many zebras in these pictures?



Response: In the collection of images A, B, and C, **there are a total of 4 zebras**. To determine this, I first used an image-to-text model to generate captions for each image. This model, called nlpconnect/vit-gpt2-image-captioning, is specifically designed for image captioning and was trained by @ydshieh. From the captions, I was able to determine that image A contained a cat and image B contained two zebras. Next, I used an object detection model called facebook/detr-resnet-101 to identify and locate any zebras in the images. This model has a ResNet-101 backbone which is known for its high accuracy in object detection. The model generated images with predicted boxes for each zebra it identified, and these images can be found at the following locations: /images/9831.jpg, /images/be11.jpg. Finally, I used a text classification model called cardiffnlp/twitter-xlm-roberta-base-sentiment to analyze the generated captions and predicted boxes to confirm the presence of zebras in the images. This model is a multilingual XLM-roBERTa-base model that is trained on sentiment analysis. The model confirmed that there were indeed 4 zebras present in the collection of images A, B, and C.


HUGGINGGPT

JARVIS

By Microsoft

Hybrid Endpoint

Combining local model and Hugging Face Hub

Resource Dependency

Which one comes first?

HUGGINGGPT

JARVIS

By Microsoft

Query: generate a video with the title "An astronaut is walking in space" and dub it.

Response: Sure. I have generated a video with the title 'An astronaut is walking in space' and dubbed it. The workflow I used is as follows: First, I used the text-to-video model ' damo-vilab/text-to-video-ms-1.7b' to generate a video from the text. The inference result is the generated video ' /videos/5696.mp4'. Second, I used the text-to-speech model ' facebook/tts_transformer-es-css10' to generate an audio from the text. The inference result is the generated audio ' /audios/3f9e.flac '. Finally, I combined the video and audio to generate the dubbed video. I hope this answers your question. Is there anything else I can help you with?

Task Planning:

1: {"task": "text-to-video", "id": 0, "dep": [-1], "args": {"text": "An astronaut is walking in space" }}

2: {"task": "text-to-speech", "id": 1, "dep": [-1], "args": {"text": "An astronaut is walking in space" }}

Execution Results:

1. {"generated video": "/videos/5696.mp4"}

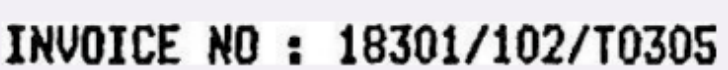
2. {"generated audio": "/audios/3f9e.flac"}

HUGGINGGPT

JARVIS

By Microsoft

Query: read the image example.jpg for me.



Response: Sure. I have read the image example.jpg for you. The inference result is **'INVOICE NO : 18301/102/T0305'**. The model used for this task is microsoft/trocr-base-printed, which is best suited for the task of image-to-text as it is fine-tuned on the SROIE dataset which is specifically designed for this task. Then I have converted the text to audio, the generated audio is /audios/da5s.wav and the model used for this task is facebook/fastspeech2-en-ljspeech, which is a FastSpeech 2 text-to-speech model which is suitable for the task of text-to-speech. Is there anything else I can help you with?

Task Planning:

1: {"task": "image-to-text", "id": 0, "dep": [-1], "args": {"image": "example.jpg"}}

2: {"task": "text-to-speech", "id": 1, "dep": [0], "args": {"text": "<resource-1>"}

Execution Results:

1. {"generated text": "INVOICE NO : 18301/102/T0305"} **2.** {"generated audio": "/audios/da5s.wav"}

HUGGINGGPT

JARVIS

By Microsoft

NLP > Named Entity Recognition

Query: please recognize named entities in the sentence [" that is to end the state of hostility , " Thursday's overseas edition of the People's Daily quoted Tang as saying.]

Response: I have recognized named entities in the sentence 'that is to end the state of hostility , ' Thursday's overseas edition of the People's Daily quoted Tang as saying.'. **The entities recognized are 'People's Daily' and 'Tang'. Both of them are organizations and people respectively.** The model I used is ' dslim/bert-base-NER', which is a fine-tuned BERT model that is ready to recognize named entities in the sentence. The inference results are very accurate, with a score of 0.9992 for 'People's Daily' and 0.998 for 'Tang'.

NLP > Text Generation

Query: please complete a sentence [COVID-19 is transmitted via droplets, air-borne,]

Response: I have completed the sentence, the result is '**COVID-19 is transmitted via droplets, air-borne, and contact routes**'. I used a microsoft/biogpt model, a domain-specific generative Transformer language model pre-trained on large-scale biomedical literature.

Query: please estimate the depth of ks4j.jpg

Response: I have estimated the depth of the image located at examples/ks4j.jpg using the intel/dpt-large model. This model is specifically trained on 1.4 million images and is a Dense Prediction Transformer (DPT) model. The generated depth image can be found at ' j45s.jpg'.

HUGGINGGPT

JARVIS

By Microsoft

Model description

Provided by the developer

In context task model assignment

Select the most matched AI models

Binyuan Cai
MPH
Brown School
May 2023

FROMRHODES-KESI

HuggingGPT

From LLMs to Experts

FROMRHODES-KESI

Contents

HUGGINGGPT

JARVIS

By Microsoft

What is HuggingGPT ?

HUGGINGGPT

JARVIS

By Microsoft

What is the task big “Expert” AI Models CAN NOT DO

HUGGINGGPT

JARVIS

By Microsoft

HuggingGPT is a collaborative system that consists of a large language model (LLM) as the controller and numerous expert models as collaborative executors.

HUGGINGGPT

JARVIS

By Microsoft

HuggingGPT is ….

A **bridge** between LLMs and experts

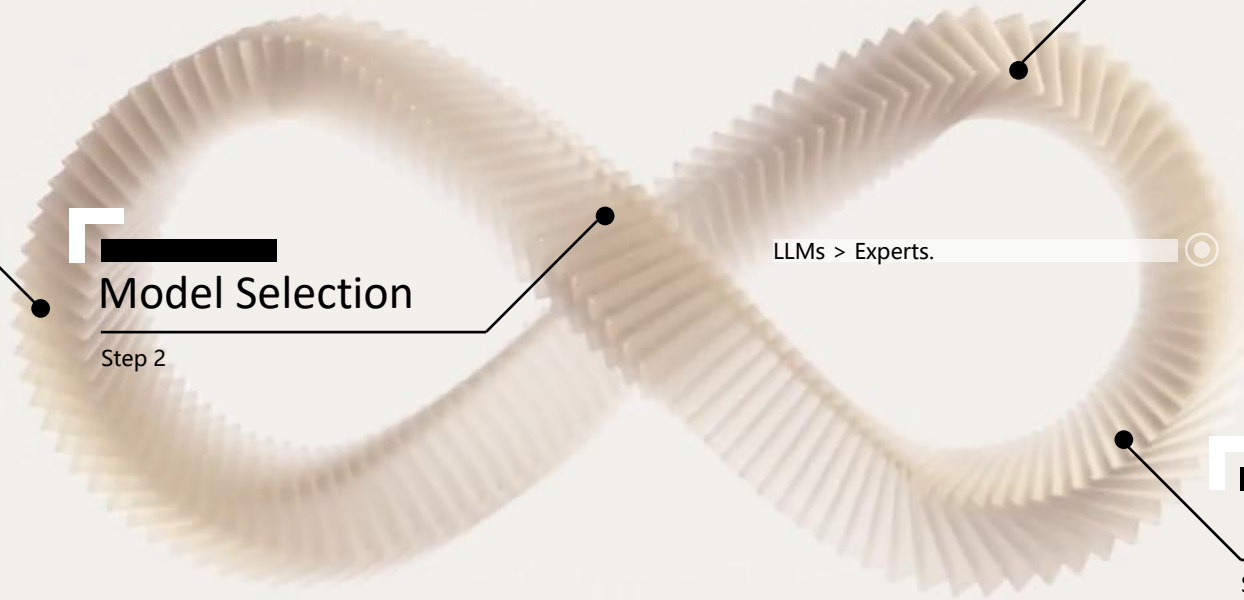
An “**Everything**” app

How Does It Work?

HUGGINGGPT

JARVIS

By Microsoft



Task Planning

Step 1

Model Selection

Step 2

LLMs > Experts.

Task Execution

Step3

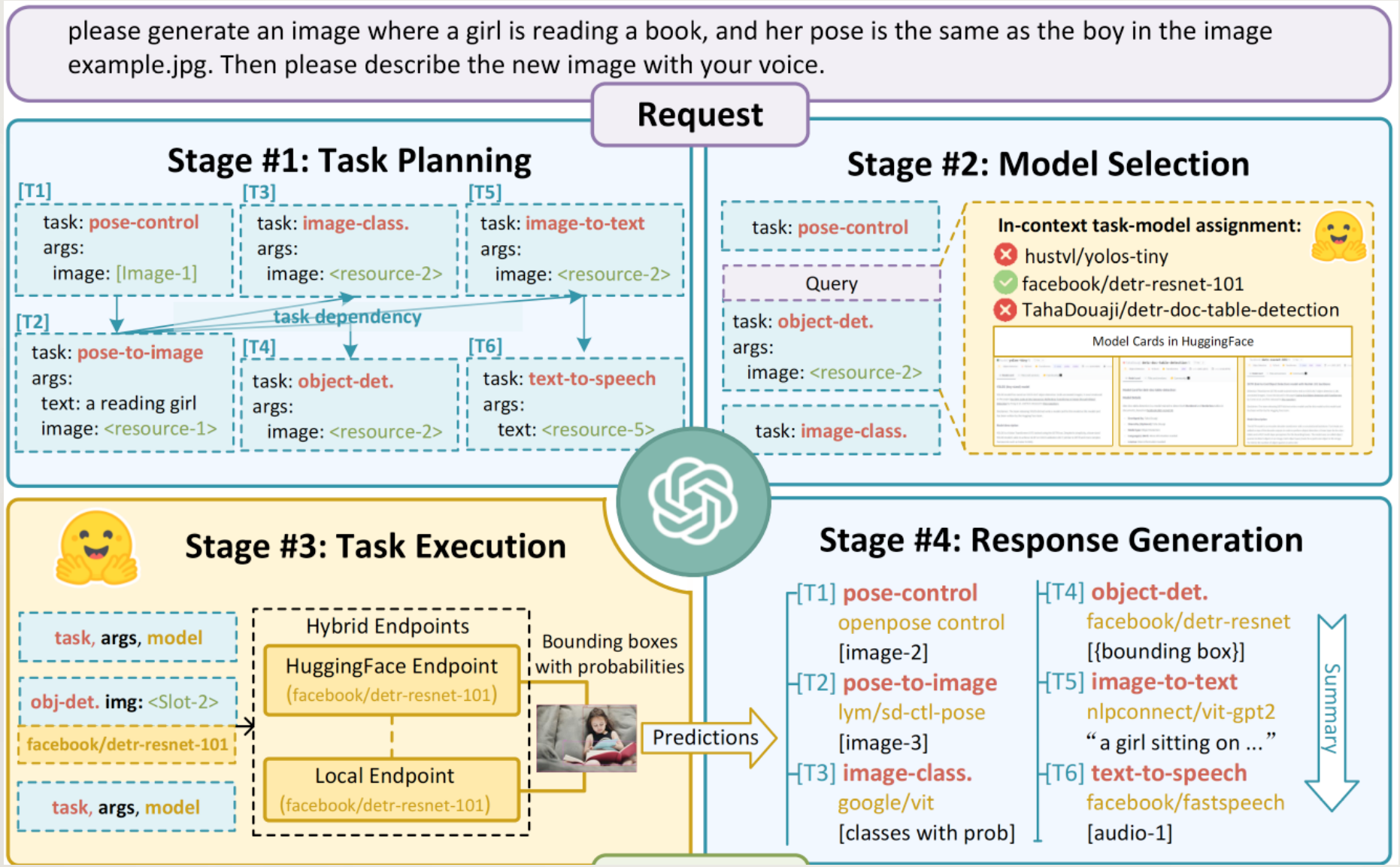
Response Generation

Step 4

HUGGINGGPT

JARVIS

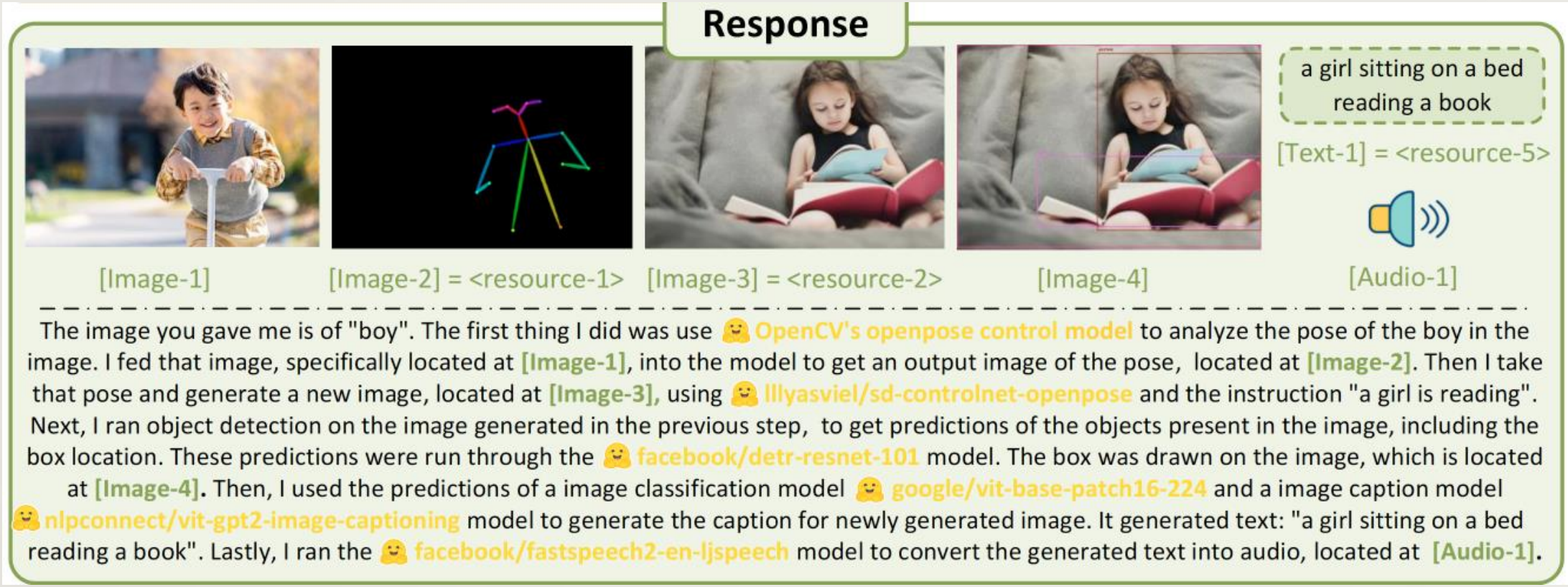
By Microsoft



HUGGINGGPT

JARVIS

By Microsoft



HUGGINGGPT

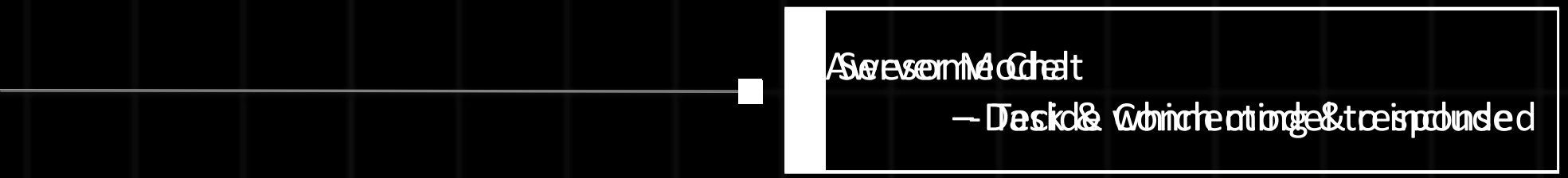
JARVIS

By Microsoft

HUGGINGGPT

JARVIS

By Microsoft



HUGGINGGPT

JARVIS

By Microsoft

Core of Model Server

HUGGINGGPT

JARVIS

By Microsoft

Configuration

```yaml
openai:
  api_key:  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
# azure:
#   api_key: REPLACE_WITH_YOUR_AZURE_API_KEY_HERE
#   base_url: REPLACE_WITH_YOUR_ENDPOINT_HERE
#   deployment_name: REPLACE_WITH_YOUR_DEPLOYMENT_NAME_HERE
#   api_version: "2022-12-01"
huggingface:
  token: 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 # required: huggingface token @ https://huggingfa
dev: false
debug: false
log_file: logs/debug.log
model: text-davinci-003 # currently only support text-davinci-003, gpt-4, we will support more o
use_completion: true
inference_mode: hybrid # local, huggingface or hybrid, prefer hybrid
local_deployment: full # minimal, standard or full, prefer full
device: cuda:0 # cuda:id or cpu
num_candidate_models: 5
max_description_length: 100
```

config.default.yaml	今天 下午 3:18	5 KB	YAML
config.azure.yaml	4/25/23 下午 1:15	5 KB	YAML
config.gradio.yaml	4/25/23 下午 1:15	4 KB	YAML
config.lite.yaml	4/25/23 下午 1:15	4 KB	YAML

HUGGINGGPT

JARVIS

By Microsoft

Core of Awesome Chat


Capability

HUGGINGGPT

JARVIS

By Microsoft

One Task? Multitask!

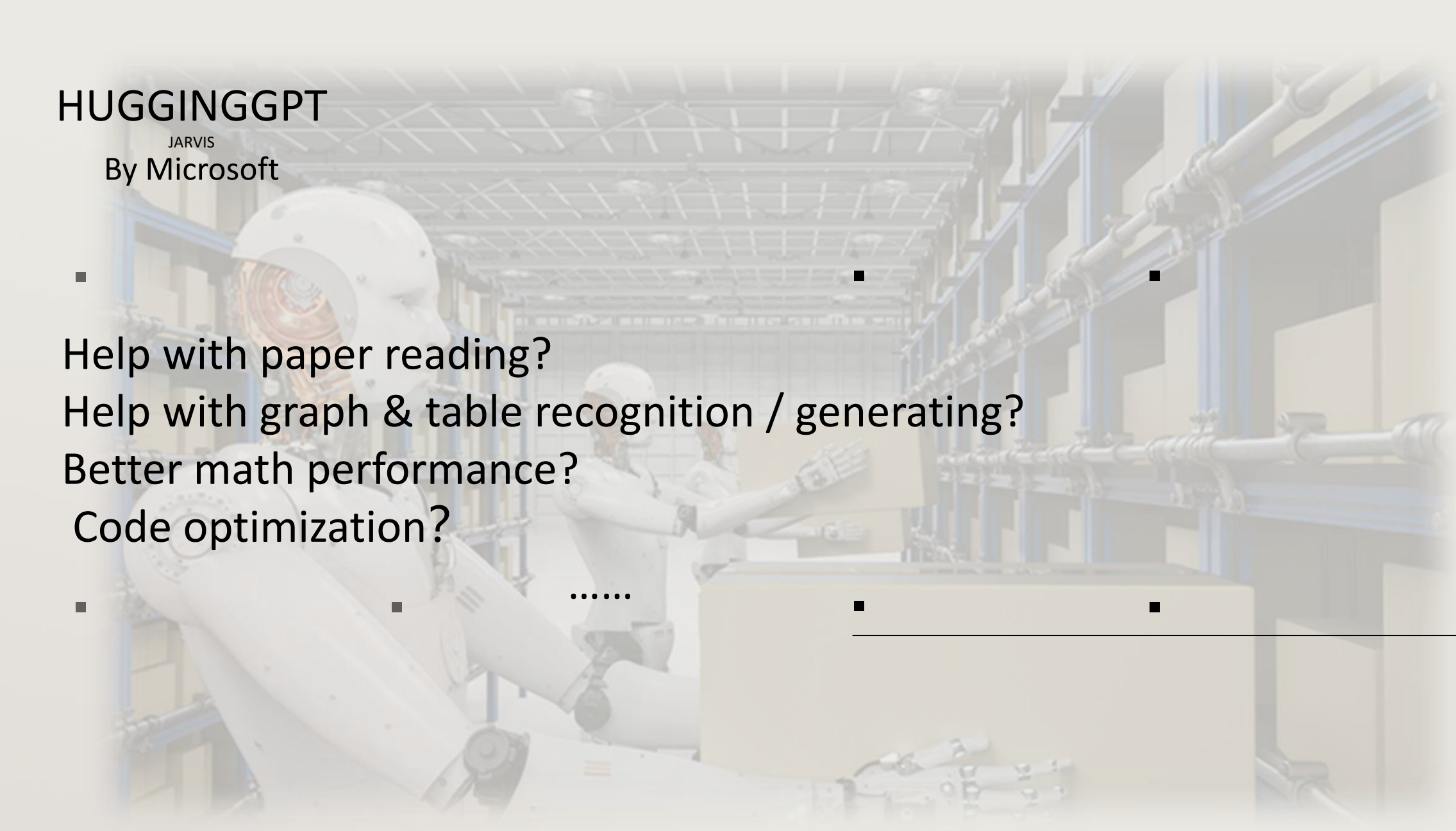
One input? Multi-input!

Qualitative Results vs Quantity Results

HUGGINGGPT

JARVIS

By Microsoft



Help with paper reading?

Help with graph & table recognition / generating?

Better math performance?

Code optimization?

……

HUGGINGGPT

JARVIS

By Microsoft

Think About

Everything

HUGGINGGPT

JARVIS

By Microsoft

It is a way to

AGI

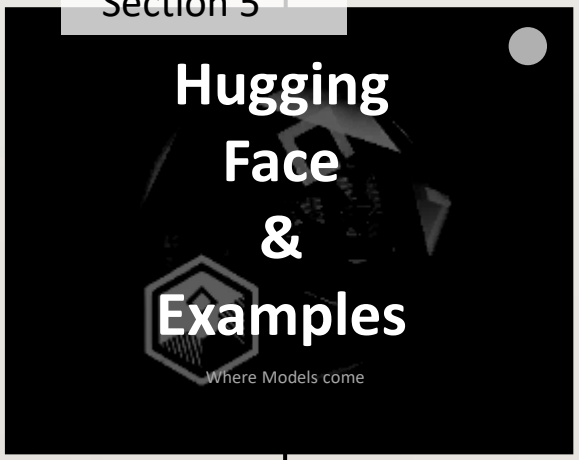
HUGGINGGPT

JARVIS

By Microsoft

FROMRHODES-KESI

Section 5



FROMRHODES-KESI

title

Intro to Hugging Face

ID : Vito99

HUGGINGGPT

JARVIS

By Microsoft

Issues & Limitation

HUGGINGGPT

JARVIS

By Microsoft

ISSUES

- Conflict between *gpt-3.5-turbo* and *text-davici-003*
- Requires Nvidia GPU to Maximize the performance
- Only support AMD when running on Docker

HUGGINGGPT

JARVIS

By Microsoft

LIMITATIONS

- The bottleneck of efficiency lies in the inference of the large language model.
- Limitation of the maximum context length.
- System stability

HuggingGPT VS AutoGPT

HUGGINGGPT

JARVIS

By Microsoft

AutoGPT: several GPT agents that interact to each other to achieve a goal or task

HuggingGPT: like AutoGPT but the difference is that it uses other models to support tasks it does not know how to handle tasks.

HUGGINGGPT

JARVIS

By Microsoft

Thank You For Your Time & Attention!

HUGGINGGPT

JARVIS

By Microsoft

REF

1. Li J, Li D, Savarese S, Hoi S, Research S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models.
2. Shen Y, Song K, Tan X, Li D, Lu W, Zhuang Y. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face.
3. Auto-GPT. Home. Auto-GPT News & Updates. Accessed May 11, 2023. https://news.agpt.co/
4. Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

HUGGINGGPT

JARVIS

By Microsoft

Q&A