



Collaborative AI: 視覚・言語・行動の融合

NTT人間情報研究所

西田京介, 壹岐太一

2023年5月12日

第13回 Language and Robotics研究会 招待講演



西田京介

専門：[機械読解（質問応答）](#) , [Vision-and-Language](#)

NTT人間情報研究所 上席特別研究員.

1. GPT-4とLLMのさらなる進化

GPT-4の概要, NLP/V&Lにおける成果
自律型AIの動向, 私達が目指すCollaborative AI



壹岐太一

専門：[Vision-and-Language](#)

NTT人間情報研究所 研究員.

2. PC環境において人と協働できる 汎用ソフトウェアロボットの最新動向

必要な技術要素, Web Navigation, 今後の方向性

1. GPT-4とLLMのさらなる進化



GPT-4の概要とその能力について

汎用人工知能が遠い夢では無くなる

- ChatGPT [OpenAI, 2022/11] の成功により、様々な自然言語処理タスクを「特化した学習無し」に解けることが実証された
- GPT-4 [OpenAI, 2023/3/14] ではChatGPTをさらに上回り、人レベルの言語理解・生成能力を持つ

(情報の検索)

“attention is all you need”
論文の要約を教えて

>こちらは「Transformer」という深層学習モデルを提案した論文です。

Transformerは、seq2seq（順方向から順方向への変換）タスクや自然言語処理タスクなどに...

Google検索に変わる
新たな情報検索の体験

(タスクの指示)

このプログラムの
誤りを直して

>修正したコードになります。

```
Copy code
<html>
<head>
<title>Breakout</title>
</head>
<body>
<canvas id="game-canvas" width="400" height="400"></canvas>

<script>
// キャンバスの要素を取得する
var canvas = document.getElementById("game-canvas");
var ctx = canvas.getContext("2d");

// ボールの状態を表す変数
var x = 100;
var y = 100;
```

AIが人に変わり任意の
言語タスクを遂行

(コンテンツの生成)

〜〜〜に関する
学術論文を書いて



ChatGPTの執筆した論文要旨
の1/3を専門家が見破れず

AIのトップ会議ICMLは
言語モデルによる論文の
執筆を禁止

InstructGPTの作り方 (GPT-4の2段階前身)

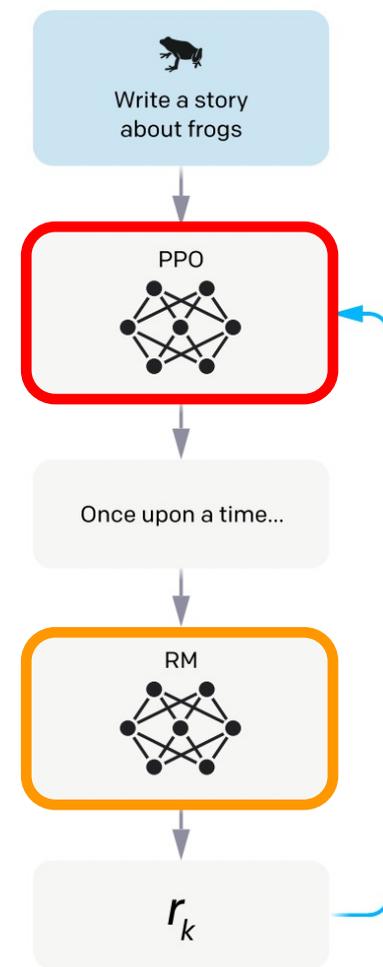
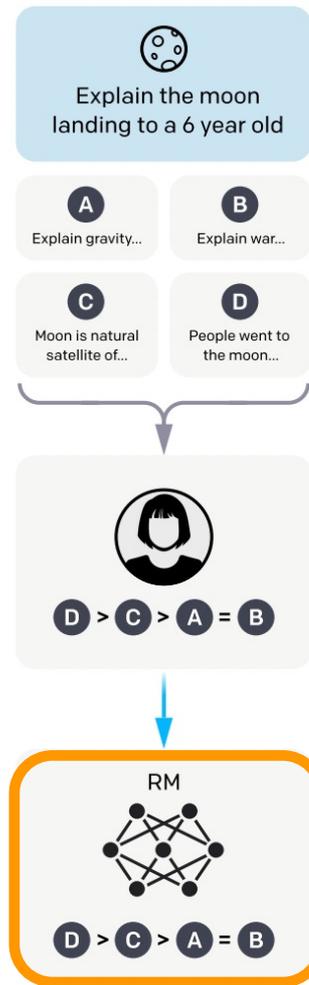
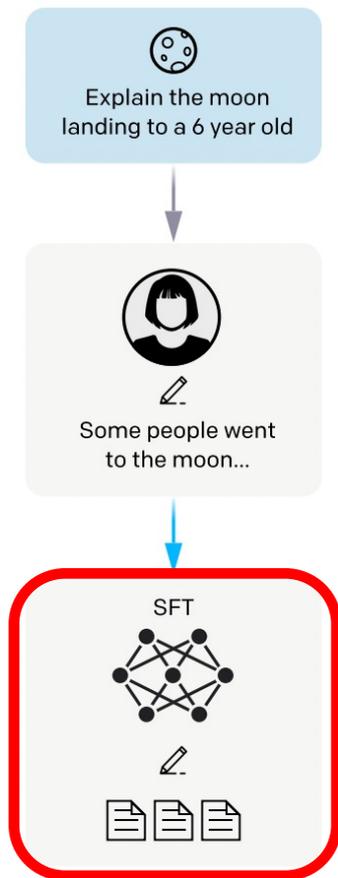
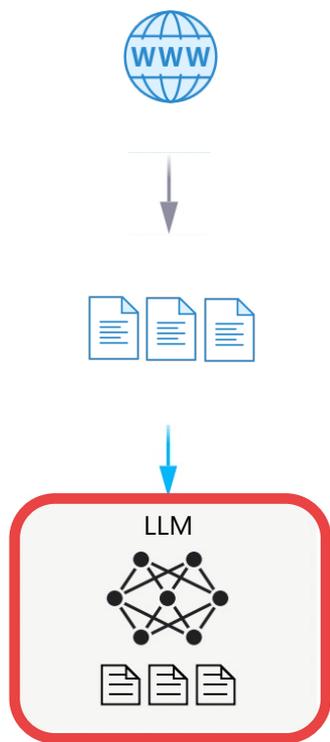
0. Webから大量のテキストを収集してLLMを学習

1. 指示に対して望ましい応答を作成して教師あり学習(SFT)

2. モデル出力の比較データを作成し報酬モデル(RM)を学習

繰り返し

3. 報酬モデル(RM)を用いてPPOを用いた強化学習により更新



ChatGPTの作り方 (予想: GPT-4の1段階前身)

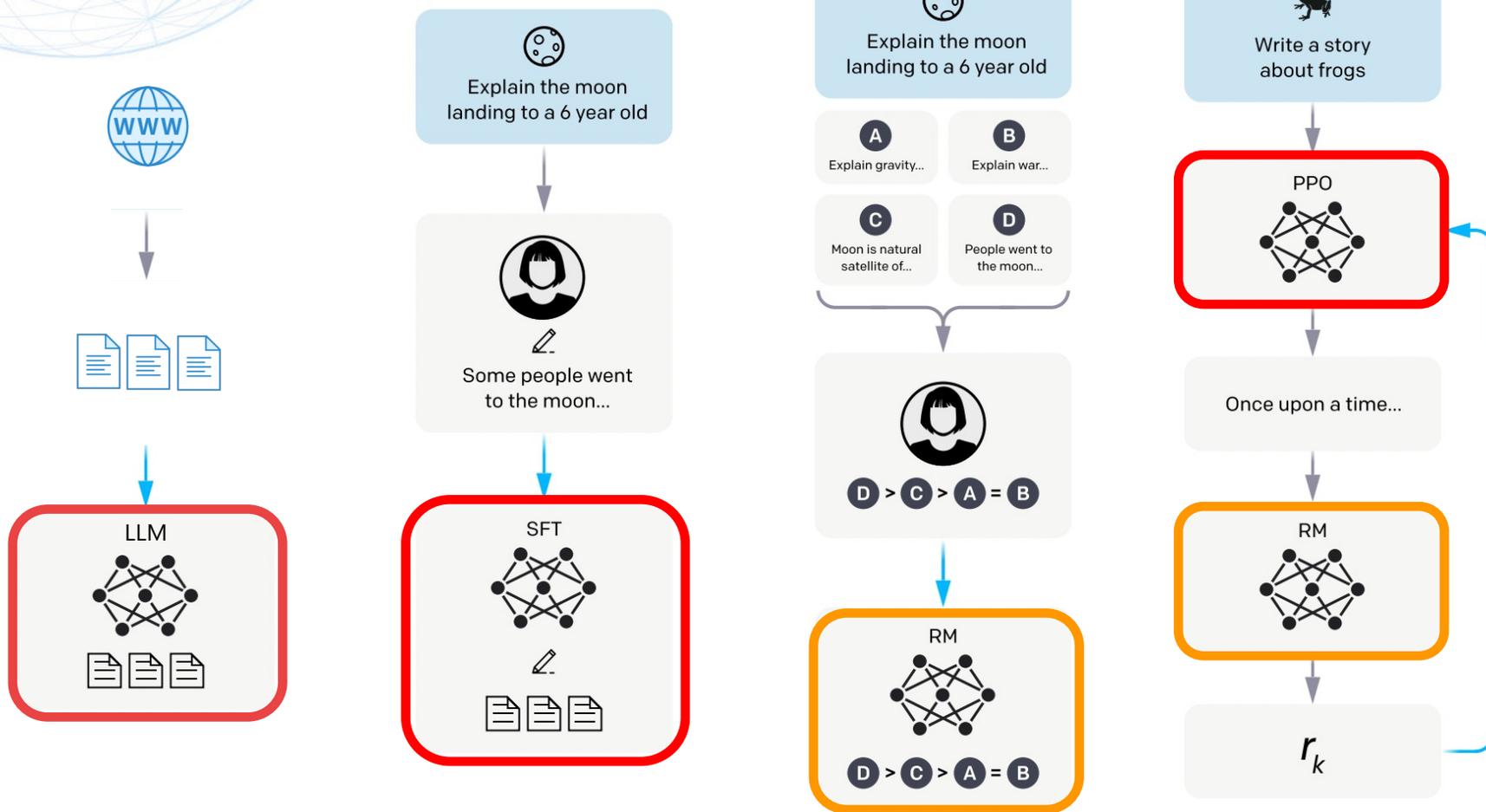
繰り返し回数の増加

0. Webから大量のテキストを収集してLLMを学習

1. 望ましい対話データを大量に作成して教師あり学習(SFT)

2. モデル出力の比較データを作成し報酬モデル(RM)を学習

3. 報酬モデル(RM)を用いてPPOを用いた強化学習により更新



GPT-4の作り方 (予想)

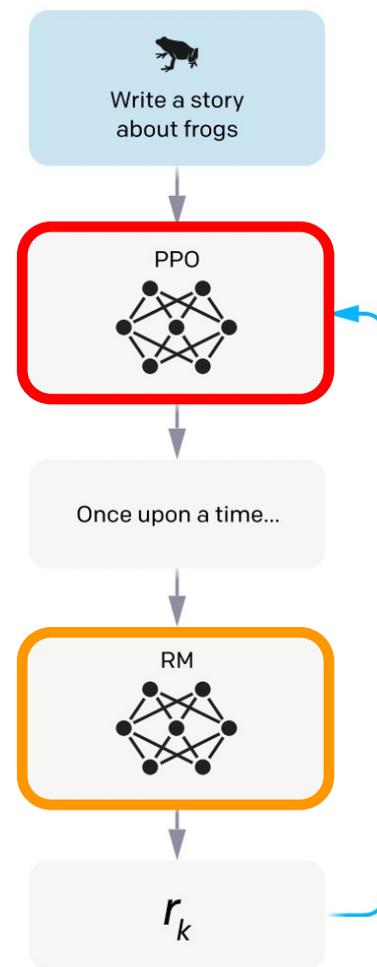
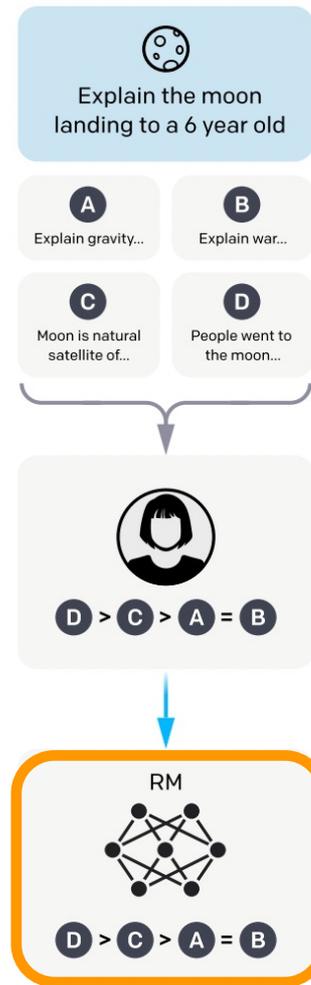
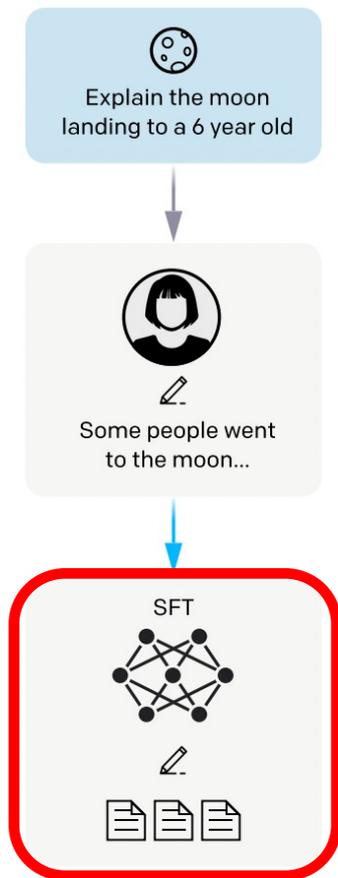
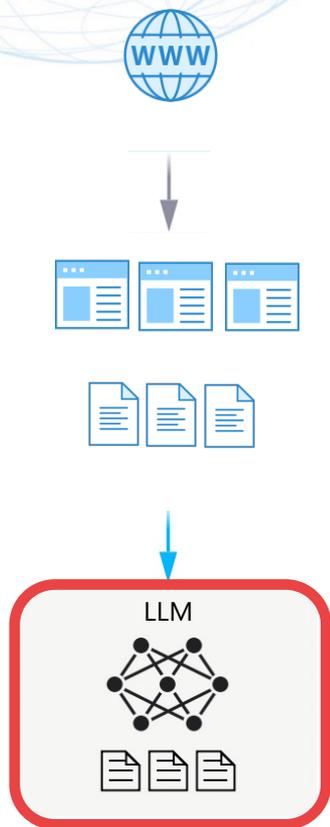
0. 大量のテキスト・**画像付テキスト**を収集してLLMを学習

1. 望ましい**画像付**対話データを作成して教師あり学習(SFT)

2. モデル出力の比較データを作成し報酬モデル(RM)を学習

3. 報酬モデル(RM)を用いてPPOを用いた強化学習により更新

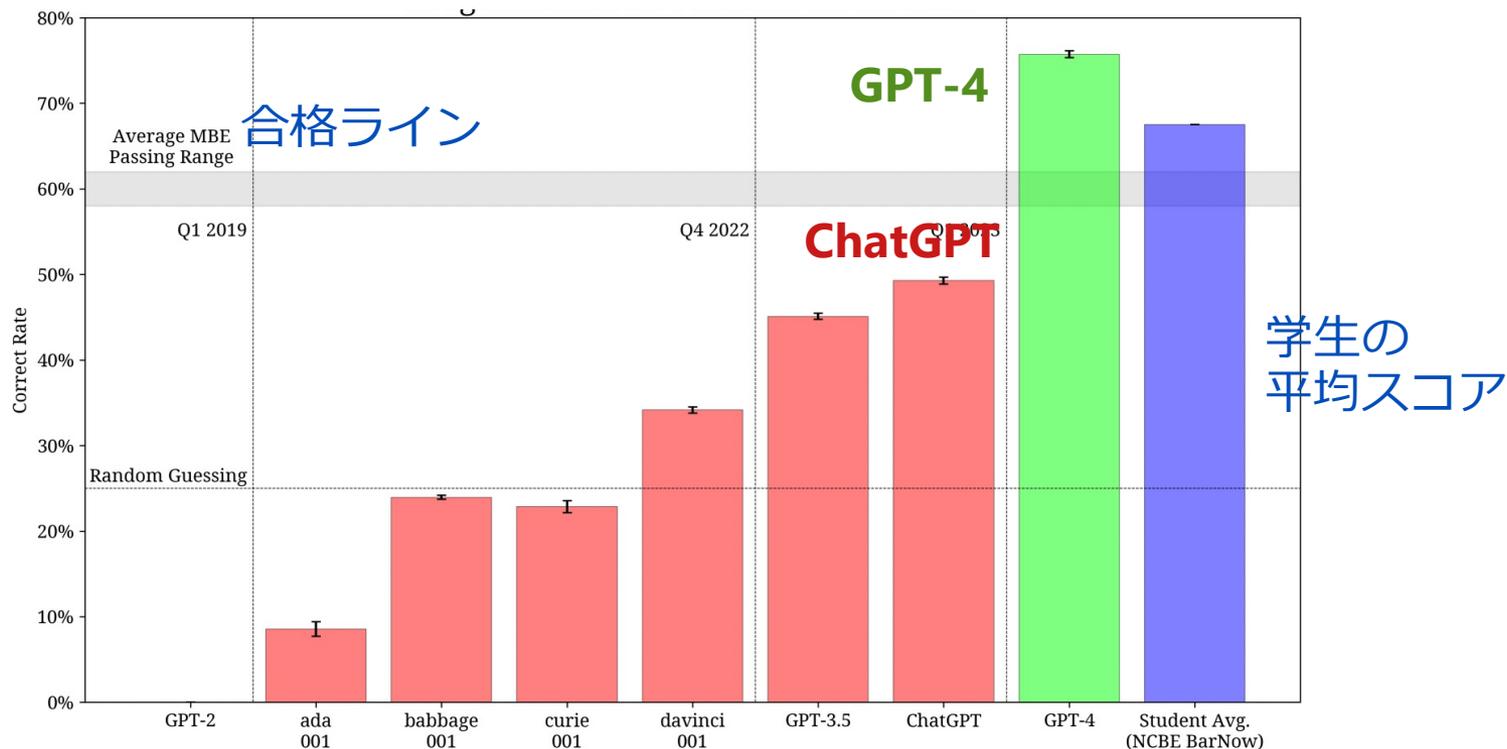
繰り返し



GPT-4が司法試験をパス [Katz(Illinois Tech)+, 2023/04]

1. MBE : 4択の選択問題. 短い問題文 (ひっかけ多い) .
 2. MEE : 論述式. やや長めの問題文に関して3~4問の小問
 3. MPT : 論述式. 必要な知識は問題文に全て記載されている
- の全形式に関して, GPT-4はzero-shotにて合格点. 特にMBEで高スコア

- 法律の専門家である著者らが, まだ人間の確認は必要としつつも, 長年待望んでいた" legal force multiplier"が遂に実現した, と考えている



クラウドワーカ以上の能力

[Pan(UCB)+, 2023/04/06]

- 社会的な意思決定判断を含むテキストシナリオをラベリングしたところ、GPT-4はクラウドワーカよりも熟練者に近い判定ができた
- 57万件のアノテーションに掛かる費用50万ドル、2万時間を大幅に削減

	Label	model	crowd _{ens}	expert	expert _{ens}
Utility	Utility level	0.75	0.85	0.93	1.00
Physical	Physical impact	0.59	0.70	0.88	1.00
Money	Economic impact	0.65	0.77	0.91	1.00
Social	Social influence	0.63	0.64	0.89	1.00
	# of targets	0.53	0.48	0.85	1.00
Violations	Killing	0.88	0.81	0.94	1.00
	Physical harm	0.89	0.79	0.94	1.00
	Non-physical harm	0.92	0.58	0.86	1.00
	Intending harm	0.95	0.64	0.86	1.00
	Deception	0.81	0.81	0.91	1.00
	Manipulation	0.89	0.61	0.85	1.00
	Betrayal	0.75	0.50	0.89	1.00
	Stealing	1.00	0.89	0.89	1.00
	Trespassing	0.93	0.68	0.94	1.00
	Spying	0.93	0.60	0.89	1.00
Vandalism	1.00	0.61	0.91	1.00	
Unfairness	0.83	0.50	0.89	1.00	
Other	0.75	0.63	0.79	1.00	

熟練者3人の多数決を正解とした時、**GPT-4**はクラウドワーカ3人の多数決を16/18個のカテゴリで上回る

GPT-4のVQAベンチマーク結果 [OpenAI, 2023/03/14]

- 各種ベンチマークにてzero-shotで高い性能を実現している
- 特にインフォグラフィック, チャート, ダイアグラムなどに関するQAで強い (一方で人間レベルには到達していないように感じる)

Benchmark	GPT-4 Evaluated few-shot	Few-shot SOTA	SOTA Best external model (incl	Human
<u>VQAv2</u> VQA score (test-dev)	77.2% 0-shot	67.6% <u>Flamingo 32-shot</u>	84.3% <u>PaLI-17B</u>	80.78%
<u>TextVQA</u> VQA score (val)	78.0% 0-shot	37.9% <u>Flamingo 32-shot</u>	71.8% <u>PaLI-17B</u>	85.01%
<u>ChartQA</u> Relaxed accuracy (test)	78.5% ^A	-	70.5%* (DePlot)	N/A
<u>AI2 Diagram (AI2D)</u> Accuracy (test)	78.2% 0-shot	-	42.1% <u>Pix2Struct Large</u>	N/A
<u>DocVQA</u> ANLS score (test)	88.4% 0-shot (pixel-only)	-	90.1%* (BAIDU-DI)	98.1%*
<u>Infographic VQA</u> ANLS score (test)	75.1% 0-shot (pixel-only)	-	61.2% <u>Applica.ai TILT</u>	98.0%*
<u>TVQA</u> Accuracy (val)	87.3% 0-shot	-	86.5% <u>MERLOT Reserve Large</u>	89.61%
<u>LSMDC</u> Fill-in-the-blank accuracy (test)	45.7% 0-shot	31.0% <u>MERLOT Reserve 0-shot</u>	52.9% <u>MERLOT</u>	N/A

A: 4-shot, CoT prompting

*ブログからの修正

*疑問有

GPT-4による視覚言語理解例 [OpenAI, 2023/3/14]

What is **funny** about this image? Describe it panel by panel.



Source: <https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/>

この画像は、「Lightning Cable」アダプターの3つのパネルが描かれたパッケージを示しています。

パネル1：スマートフォンにVGAコネクタ（コンピュータモニターに使用される一般的な大きな青い15ピンのコネクタ）が充電ポートに接続されています。

パネル2：「Lightning Cable」アダプターのパッケージで、VGAコネクタの画像が描かれています。

パネル3：VGAコネクタのクローズアップで、先端には小さなLightningコネクタ（iPhoneやその他のAppleデバイスの充電に使用される）があります。

この画像のユーモアは、大きくて時代遅れのVGAコネクタを小さな現代的なスマートフォンの充電ポートに接続するという不合理さから来ています。

※ 回答をGPT-4で翻訳

Answer **question I.1.a**. Think step-by-step.

I. Principe de la détection de rayonnement avec un bolomètre

Comme illustré sur la figure 1 un bolomètre est constitué d'un absorbeur qui reçoit le rayonnement que l'on désire détecter. Sa température T , supposée uniforme, est mesurée à l'aide d'un thermomètre incorporé, constitué d'un matériau conducteur dont la résistance $R(T)$ varie avec la température T ; cette variation est caractérisée par le coefficient $\alpha = \frac{1}{R} \frac{dR}{dT}$. L'ensemble possède la capacité thermique C_{th} .

Un barreau, conducteur thermique, homogène, de longueur L , de section S et de conductivité thermique λ et sans échanges thermiques latéraux, relie le bolomètre à un thermostat de température T_b fixe.

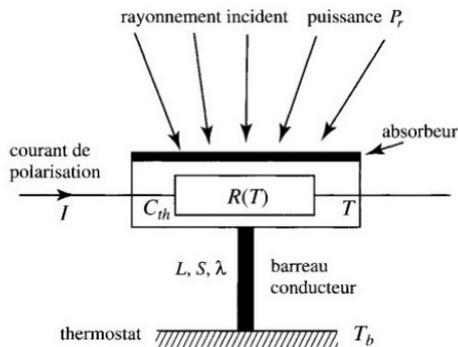


Figure 1 - Schéma d'un bolomètre

I.1. Dans cette première partie, on étudie le comportement du bolomètre en régime stationnaire, à la température de fonctionnement $T = T_0$.

I.1.a) Exprimer la température $T(x)$ en tout point du barreau conducteur repéré par son abscisse x ; on prendra $T(0) = T_0$ et $T(L) = T_b$.

問題 I.1.a に答えるために、導電バーの各点での温度 $T(x)$ を、座標 x で特徴付けられたものとして表現する必要があります。与えられた条件は $T(0) = T_0$ および $T(L) = T_b$ です。

定常状態の場合、バー内の一次元熱伝導の熱方程式は次のようになります。 $d^2T/dx^2 = 0$
一度積分すると、次のようになります。 $dT/dx = C_1$
二度目の積分： $T(x) = C_1x + C_2$

定数 C_1 と C_2 を決定するために、境界条件を使用します。
 $T(0) = T_0$ 、これにより $C_2 = T_0$ が得られます。
 $T(L) = T_b$ 、これにより $C_1 = (T_b - T_0)/L$ が得られます。
最終的に、 $T(x)$ の式は次のようになります。

$T(x) = (T_b - T_0) * (x/L) + T_0$
これは、導電バーに沿った線形の温度プロファイルを表し、傾きは $(T_b - T_0)/L$ です。

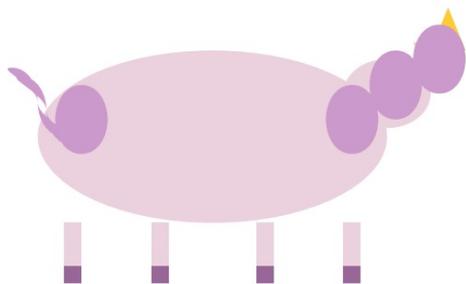
※ 回答をGPT-4で翻訳（上・下の添字は西田が最後に実施）

GPT-4によるお絵描き例

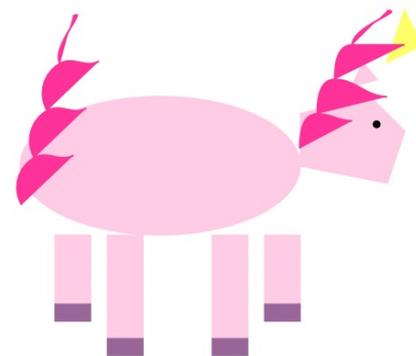
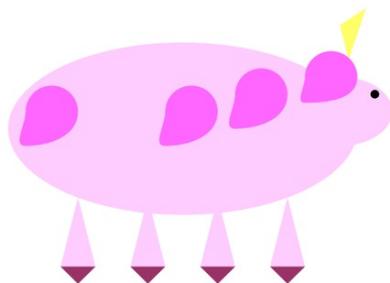
[Bubbek(Microsoft)+, 2023/3/22]

Draw a unicorn in TikZ.

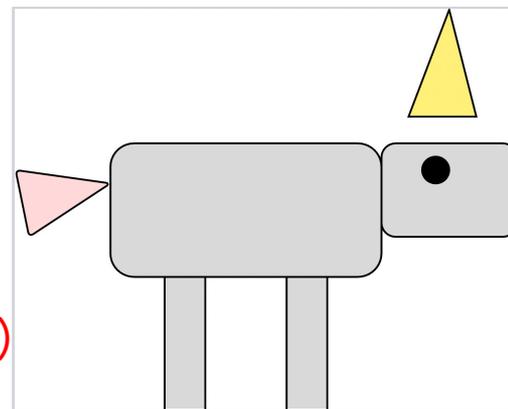
GPT-4: [Produces L^AT_EX compiling to following picture.]



GPT-4による生成例
(初期版→学習進行)



GPT-4の生成例
(西田再現/ChatGPT Plus)





ChatGPT/GPT-4周辺の さらなる進化

ChatGPT Plugins [OpenAI, 2023/3/23]

- インターネットを介した最新情報へのアクセスや数値計算など、ChatGPTが「**道具**」を使って自身の不得意な部分を補えるようになった



Expedia

Bring your trip plans to life—get there, stay there, find things to see and do.



FiscalNote

Provides and enables access to select market-leading, real-time data sets for legal, political, and regulatory data and information.



Instacart

Order from your favorite local grocery stores.



KAYAK

Search for flights, stays and rental cars. Get recommendations for all the places you can go within your budget.



Klarna Shopping

Search and compare prices from thousands of online shops.



Milo Family AI

Giving parents superpowers to turn the manic to magic, 20 minutes each day. Ask: Hey Milo, what's magic today?



OpenTable

Provides restaurant recommendations, with a direct link to book.



Shop

Search for millions of products from the world's greatest brands.



Speak

Learn how to say anything in another language with Speak, your AI-powered language tutor.



Wolfram

Access computation, math, curated knowledge & real-time data through Wolfram|Alpha and Wolfram Language.



Zapier

Interact with over 5,000+ apps like Google Sheets, Trello, Gmail, HubSpot, Salesforce, and more.

ChatGPT Plugins [OpenAI, 2023/3/23]

- インターネットを介した最新情報へのアクセスや数値計算など、ChatGPTが「**道具**」を使って自身の不得意な部分を補えるようになった
- **現状は言語中心だが、今後Visionも連携していくのは確実と言える**

TD Can you tell me what person/movie won the oscar for these categories?

- Best Actor
- Best soundtrack
- Best picture
- Best supporting actor

And then come up with a poem linking them together



Finished browsing ▾

Here are the Oscar winners for the categories you requested:

- Best Actor: Brendan Fraser, "The Whale"
- Best Soundtrack (Original Song): "Naatu Naatu" from "RRR"
- Best Picture: "Everything Everywhere All at Once"
- Best Supporting Actor: Ke Huy Quan, "Everything Everywhere All at Once" ¹

Here is a poem linking them together:

In the depths of the ocean, "The Whale" did roam,
With Brendan Fraser's performance, a light was shone,
While "Naatu Naatu" played with rhythm and beat,
From "RRR", a soundtrack that's truly a treat.

🔍 Searched: "Oscar winners Best Actor Best Soundtrack Best Picture Best Supporting Actor"

👉 Clicked on:  Oscar winners 2023: See the full list | CNN [↗](#)

⬅️ Going back to last page

👉 Clicked on:  Oscar winners 2023: See the full list | CNN [↗](#)

⬅️ Going back to last page

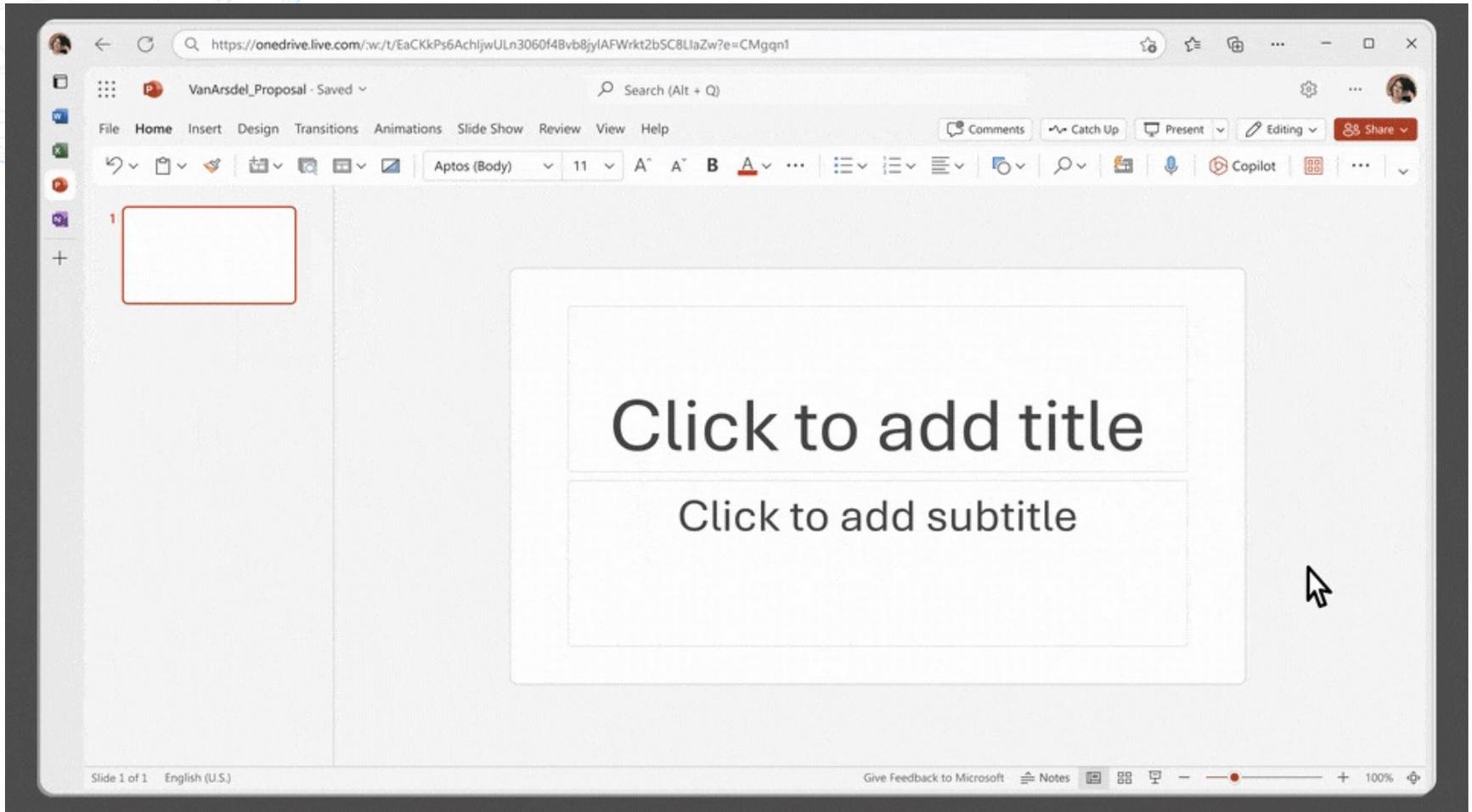
👉 Clicked on:  List of 2023 Oscar Winners - ABC News [↗](#)

☰ Reading content

🕒 Finished browsing

Co-pilot 365 [Microsoft, 2023/03/16]

- LLMのパワーをMicrosoft 365に統合することを発表
- 『言葉を地球上で最も強力な生産性向上ツールに変えます。』



JARVIS (HuggingGPT) [Shen(Microsoft)+, 2023/3/30] NTT

- ChatGPTとHuggingfaceにある多様なモデルを疎結合して連携
- リクエストに従ってChatGPTがタスク系列をプランニングし、各タスクに適切なAIモデルを選択して実行。追加の学習は行わず例示で実現

Query: based on the pose image of example1.jpg and the description of example2.jpg, please generate a new one.



Response: Based on the pose image of  example1.jpg and the description of  example2.jpg, I have generated a new image  s5f8.jpg for you. The process of generating this image involved using the  openpose-control model to extract the pose information from the basketball image, and then using the  nlpconnect/vit-gpt2-image-captioning to generate a description of the boy in the image. Finally, I used the  llyasviel/sd-controlnet-openpose model to combine the pose information and the description to generate the final image. The complete path of the generated image is  s5f8.jpg. If you have any other questions, please let me know.

Task Planning:

1. {'task': 'openpose-control', 'id': 0, 'dep': [-1], 'args': {'image': 'example1.jpg'}}
2. {'task': 'image-to-text', 'id': 1, 'dep': [-1], 'args': {'image': 'example2.jpg'}}
3. {'task': "openpose-text-to-image", "id": 2, "dep": [0, 1], "args": {"text": "<resource>-0", "image": "<resource>-1"}}



クエリ：
左の画像のポーズと、
右の画像の説明文から、
新しい画像を生成して

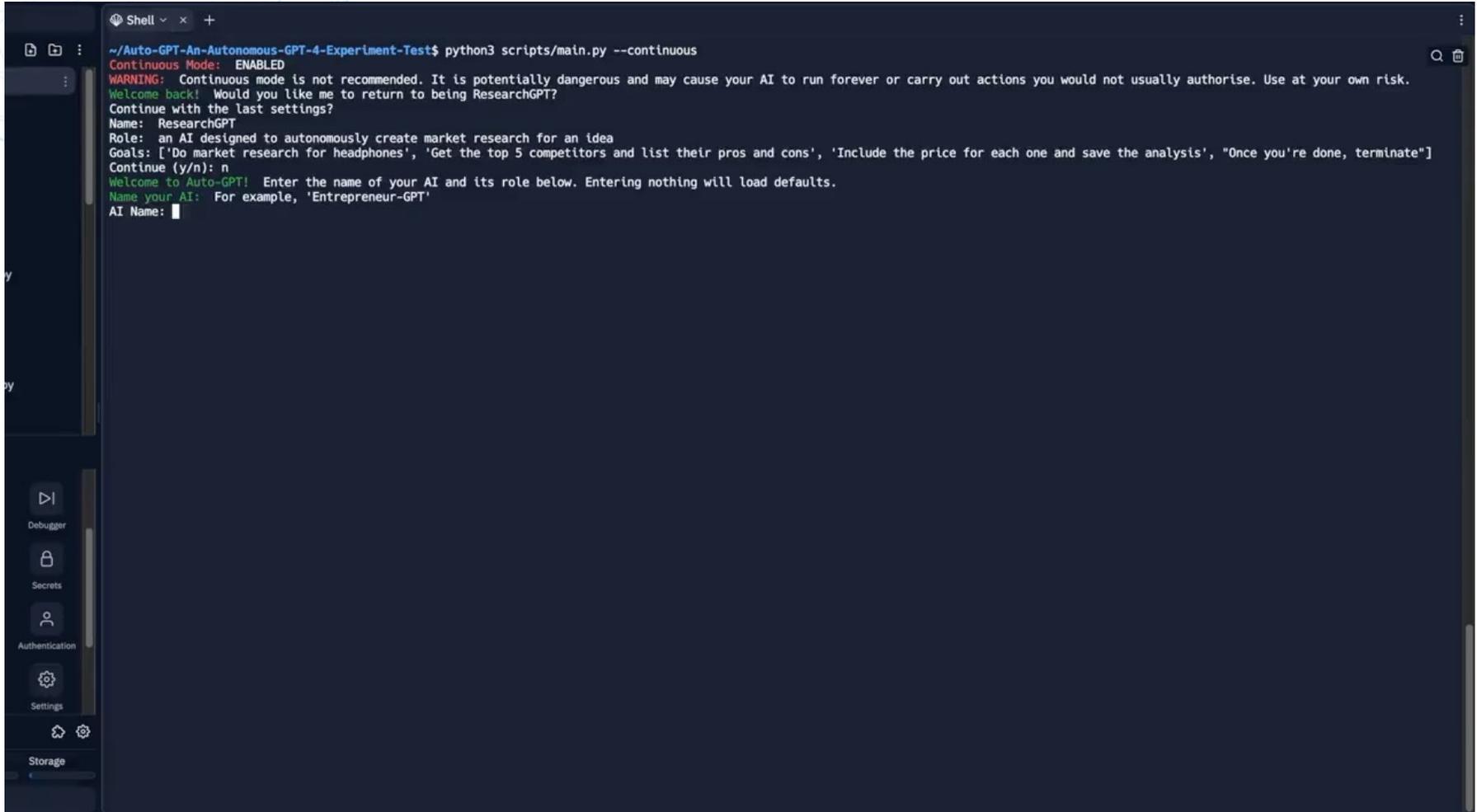
生成画像（右下）と
出力の過程の説明

クエリから作成した
タスク系列のプラン

1. 画像→ポーズ
2. 画像→テキスト
3. ポーズ&テキスト
→画像

AutoGPT [Toran Bruce Richards, OSS, 2023/3/30~]

- 最終ゴールを与えられると、GPT-4が**自律的に**必要なサブタスクに分解、ネットやその他ツールを使ってプランを修正しながらゴールを達成する



```
Shell x +
~/Auto-GPT-An-Autonomous-GPT-4-Experiment-Test$ python3 scripts/main.py --continuous
Continuous Mode: ENABLED
WARNING: Continuous mode is not recommended. It is potentially dangerous and may cause your AI to run forever or carry out actions you would not usually authorise. Use at your own risk.
Welcome back! Would you like me to return to being ResearchGPT?
Continue with the last settings?
Name: ResearchGPT
Role: an AI designed to autonomously create market research for an idea
Goals: ['Do market research for headphones', 'Get the top 5 competitors and list their pros and cons', 'Include the price for each one and save the analysis', "Once you're done, terminate"]
Continue (y/n): n
Welcome to Auto-GPT! Enter the name of your AI and its role below. Entering nothing will load defaults.
Name your AI: For example, 'Entrepreneur-GPT'
AI Name: 
```

- 最終ゴールを与えられると、GPT-4が**自律的に**必要なサブタスクに分解。ネットやその他ツールを使ってプランを修正しながらゴールを達成する

ユーザの入力した“Goals”の例

1. Do market research for different headphones on the market today
2. Get the top 5 headphones and list their pros and cons
3. Include the price for each one and save the analysis
4. Once you are done, terminate

実行結果

1. Google検索を“top 5 headphones 2023”で実施
2. 検索結果1位のページにアクセスして欲しい情報が無かったと認識
3. Google検索を“top 5 headphones 2023 with pros and cons”で再度実施
4. 検索結果1位のページにアクセスして欲しい情報を発見
5. 正規表現でページから情報抽出しようとするが失敗
6. BeautifulSoupに切り替えて抽出
7. 抽出結果をファイル保存

AutoGPT [Toran Bruce Richards, OSS, 2023/3/30~]

- 現在用意されている「コマンド」の一覧. 外部APIおよび外部ニューラルモデルを利用する
- PCを使って人 + AIが出来ることが自動化されていく

パッケージ	実行内容
autogpt.commands	
improve_code	LLMにてコードを改善
audio_text	Huggingfaceモデルによる音声認識
image_gen	Huggingfaceモデルなどによる画像生成
execute_code	Pythonコードやシェル実行
file_operations	各種ファイル操作
git_operations	Gitからのクローン
google_search	Google検索
twitter	Tweet投稿
web_selenium	URLとクエリを受け取って, ページのテキストをクエリ依存要約.

llama.cpp [Georgi Gerganov, OSS, 2023/03/11~]

- 公開のLLMであるLLaMAをC/C++実装し、4bit量子化推論をサポート
- LLaMA 13Bを1台がmacbook（6~20GB程度のメモリ）で動作する
- LLMを動作させることに関する障壁はかなり下がってきている

```

ggerganov@Georgis-MBP: ~/development/github/llama.cpp
00:24:05 master 47 $ ./main -m models/7B/ggml-model-q4_0.bin -p "I believe the meaning of life is" -t 4
main: seed = 1678487088
llama_model_load: loading model from 'models/7B/ggml-model-q4_0.bin' - please wait ...
llama_model_load: n_vocab = 32000
llama_model_load: n_ctx = 512
llama_model_load: n_embd = 4096
llama_model_load: n_mult = 256
llama_model_load: n_head = 32
llama_model_load: n_layer = 32
llama_model_load: n_rot = 128
llama_model_load: f16 = 2
llama_model_load: n_ff = 11008
llama_model_load: ggml ctx size = 4529.34 MB
llama_model_load: memory_size = 512.00 MB, n_mem = 16384
llama_model_load: ..... done
llama_model_load: model size = 4017.27 MB / num tensors = 291

main: prompt: 'I believe the meaning of life is'
main: number of tokens in prompt = 8
1 -> 'I'
29992 -> 'I'
4658 -> 'believe'
278 -> 'the'
6593 -> 'meaning'
310 -> 'of'
2834 -> 'life'
338 -> 'is'

sampling parameters: temp = 0.800000, top_k = 40, top_p = 0.950000

I believe the meaning of life is to find your gift. The purpose of life is to give it away. -Steve Maraboli In our eve
ryday lives we are constantly surrounded by people who have gifts to offer us. It is our job to receive them. When we
do receive their gift of love, we must then offer them our own gift of love, back to them. This is the meaning of life
. This is also the purpose of life. If we can learn to do this we will have a very happy and fulfilling life.
The Meaning of Life is the Gift of Love. [end of text]

main: mem per token = 14368644 bytes
main: load time = 1488.26 ms
main: sample time = 270.77 ms
main: predict time = 13229.62 ms / 104.17 ms per token
main: total time = 15425.52 ms
ggerganov@Georgis-MBP: ~/development/github/llama.cpp
00:25:03 master 47 $

ggerganov@Georgis-MBP: ~/development/github/whisper.cpp
00:24:07 coreml 8c $ ./stream -m models/ggml-small.en.bin --step 400 --length 6000
init: found 3 capture devices:
init: - Capture device #0: 'Georgi's iPhone Microphone'
init: - Capture device #1: 'MacBook Pro Microphone'
init: - Capture device #2: 'Microsoft Teams Audio'
init: attempt to open default capture device ...
init: obtained spec for input device (SDL Id = 2):
init: - sample rate: 16000
init: - format: 33056 (required: 33056)
init: - channels: 1 (required: 1)
init: - samples per frame: 1024
whisper_init_from_file: loading model from 'models/ggml-sma
whisper_model_load: loading model
whisper_model_load: n_vocab = 51864
whisper_model_load: n_audio_ctx = 1500
whisper_model_load: n_audio_state = 768
whisper_model_load: n_audio_head = 12
whisper_model_load: n_audio_layer = 12
whisper_model_load: n_text_ctx = 448
whisper_model_load: n_text_state = 768
whisper_model_load: n_text_head = 12
whisper_model_load: n_text_layer = 12
whisper_model_load: n_mels = 80
whisper_model_load: f16 = 1
whisper_model_load: type = 3
whisper_model_load: mem required = 603.00 MB (+ 16.00 M
whisper_model_load: kv self size = 15.75 MB
whisper_model_load: kv cross size = 52.73 MB
whisper_model_load: adding 1607 extra tokens
whisper_model_load: model ctx = 464.56 MB
whisper_model_load: model size = 464.44 MB
whisper_init_from_file: loading Core ML model from 'models/
whisper_init_from_file: first run on a device may take a white ...
whisper_init_from_file: Core ML model loaded

main: processing 6400 samples (step = 0.4 sec / len = 6.0 sec / keep = 0.2 sec), 4 threads, lang = en, task = transcri
be, timestamps = 0 ...
main: n_new_line = 14, no_context = 1

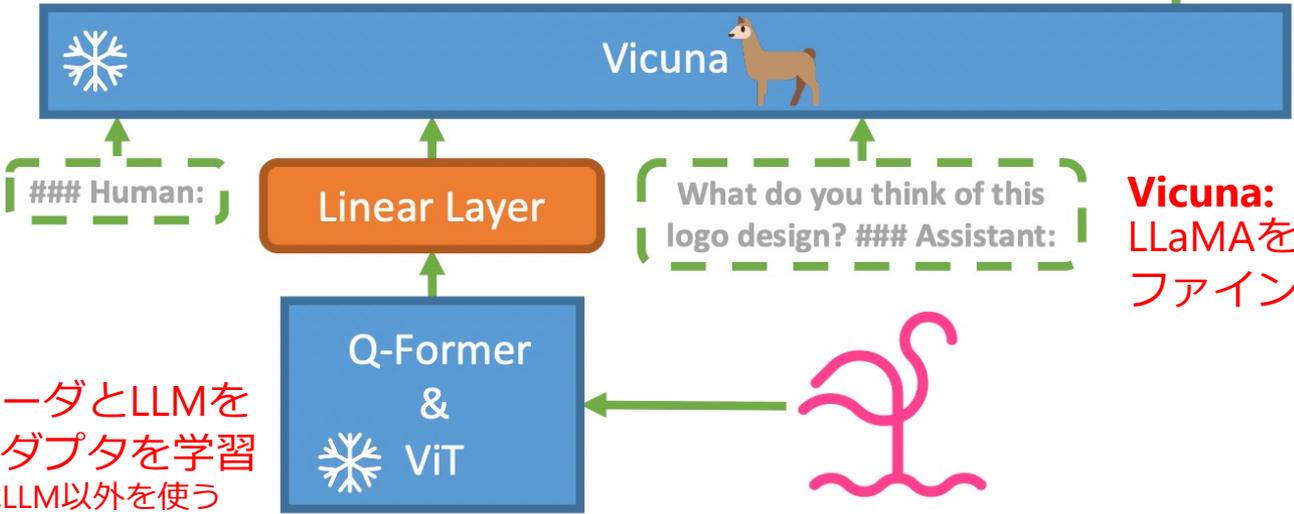
Already there now you're interpreting the choice and you're creating narrative for for a minute. Yeah, and now we're
talking
talking about the narrative is very meta. Looking back, what is the most beautiful or surprising idea in blurring?
or AI in general that you've come across. You've seen this field explode.
and interesting ways, just what cool ideas like.
    
```

Macbookで、llama.cpp と whisper.cpp（音声認識）を同時実行の様子

MiniGPT-4 [Zhu(King Abdullah U)+, 2023/04/20]

- BLIP-2の 画像エンコーダ [Li(Salesforce)+, 2023/01] , Vicuna [UCB, CMU, Stanford, MBZUAI, and UCSD, 2023/03] を組み合わせてGPT-4ライクなモデルを低コストに作成
- 良いモデルをどう組み合わせ, どう学習するか? が大事になってきた

The logo design is **simple** and **minimalistic**, with a **pink line drawing** of a **flamingo standing on one leg in the water**. The design is **clean** and **easy to recognize**, making it suitable for use in various contexts such as **a logo for a beach resort** or a **flamingo-themed event**. The use of a flamingo as a symbol adds a touch of **whimsy** and fun to the design, making it **memorable** and **eye-catching**. Overall, it's a **well-designed** logo that **effectively communicates the brand's message**.



BLIP-2:
 画像エンコーダとLLMを
 固定してアダプタを学習
 ※MiniGPTではLLM以外を使う

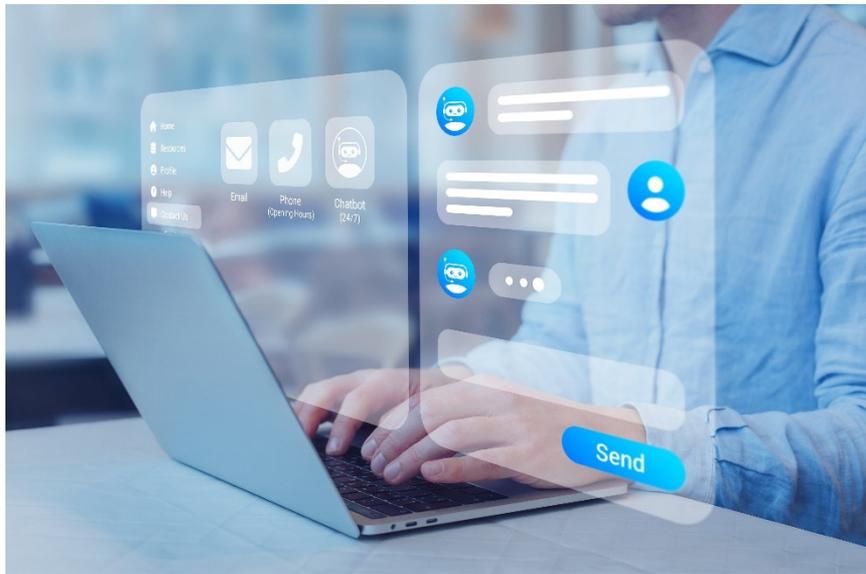
Vicuna:
 LLaMAをChatGPTのログで
 ファインチューニング



私達が目指しているもの

あらゆる環境で人と**自然に協調可能**な汎用AIの
思考エンジンを創り、人々のwell-beingを実現したい

GPT-4がまだ出来ていないことは何なのか？



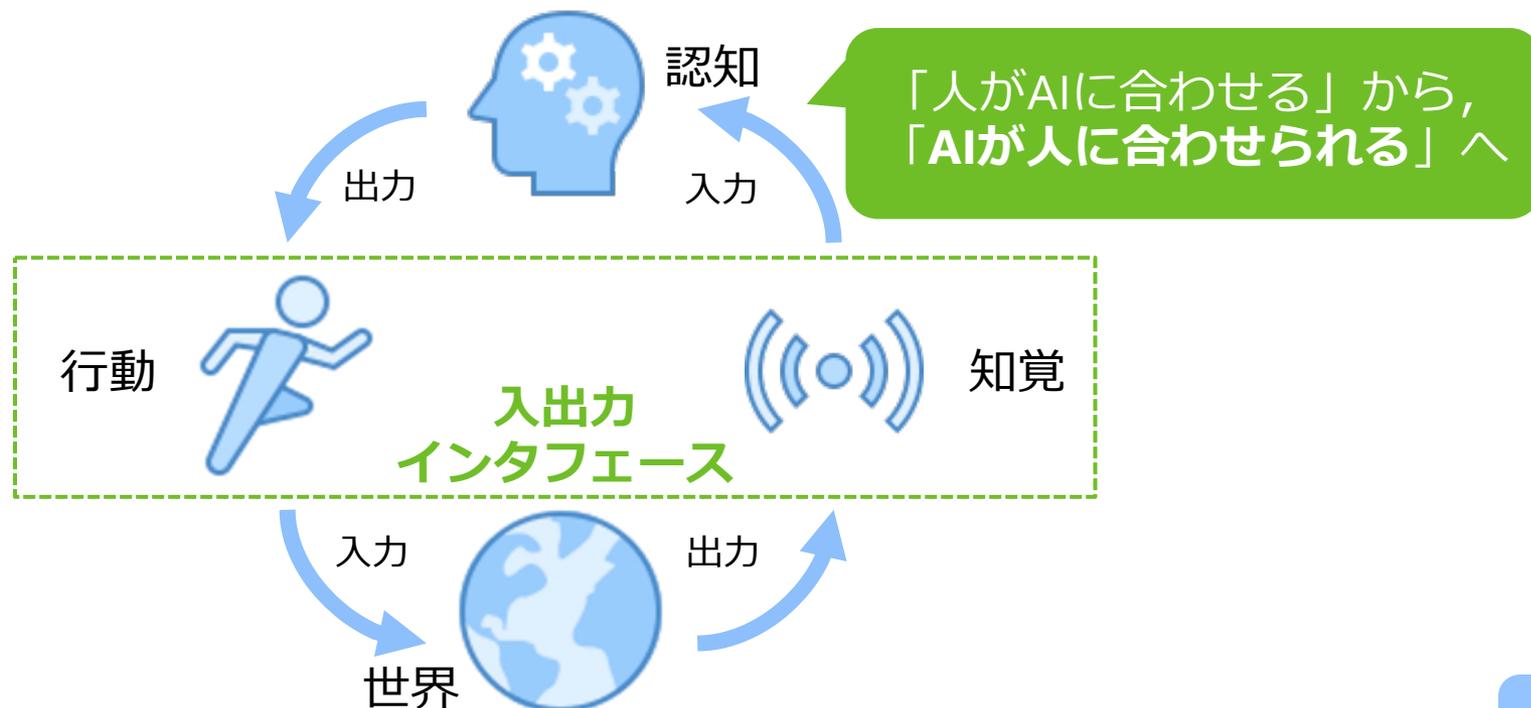
計算機の中の**コラボレーター**として、
人と協働できるソフトウェアロボット



人生の**パートナー**として、
人と一緒に成長するロボット

Collaborative AIを目指して

- あらゆる環境で人と**自然に協調**可能な汎用AIを作りたい
 - タスクを遂行しているのが人かAIかを区別する必要が無いレベル
 - PC上でのアシスタントとしてGPT-4は既に素晴らしいレベルに達しているが、APIを持たないソフトウェアには現状手出しできない
- 人と同じ入出力インターフェース**でAIが行動できるようにしたい
 - AIを社会実装可能な範囲も広がり、人とAIの**相互成長**にも繋がる
 - 身体性の獲得を目指す、とも言える



VisualMRC [Tanaka & Nishida+ , AAI'21, NLP'21最優秀賞]

- Webページに含まれる言語を人のように視覚から読み解く課題
- レイアウト理解と生成型機械読解を統合した問題設定とモデルの提案

2007 Ig Nobel Prize winners announced

Friday, October 5, 2007

The winners of the 2007 [Ig Nobel Prize](#) have been announced. The awards, given out every early October since 1991 by the [Annals of Improbable Research](#), are a parody of the [Nobel Prize](#), which are awards given out in several fields. The awards are given to achievements that, "first make people laugh, and then make them think." They were presented at [Harvard University's Sanders Theater](#).

Ten awards have been presented, each given to a different field. The winners are:

- **Medicine:** Brian Witcombe, of Gloucestershire Royal NHS Foundation Trust, [UK](#), and Dan Meyer, who studied the health consequences of [sword swallowing](#).
- **Physics:** A team from the [USA](#) and [Chile](#), who made a study about how cloth sheets become wrinkled.
- **Biology:** Dr Johanna van Bronswijk of the [Netherlands](#), for carrying out a census of creatures that live in people's beds.
- **Chemistry:** Mayu Yamamoto, from [Japan](#), for creating a method of extracting vanilla fragrance and flavouring from cow dung.



The 2007 [Ig Nobel Prize](#) in aviation went to a team from an [Argentinian](#) university, who discovered that [mpotency drugs](#) can help [hamsters](#) recover from [jet lag](#).

約10,000枚の画像を収集

全ての領域をアノテーション：
9つの意味クラスに分類

Q: Who were the winners of the Ig Nobel prize for Biology and Chemistry?

A: The winner of the Ig Nobel prize for biology was Dr Johanna van Bronswijk, and the winner for Chemistry was Mayu Yamamoto.

約30,000件の生成型QAを作成

SlideVQA [Tanaka+, AAAI'23, NLP'23優秀賞 & 言語資源賞]

- 複数枚の画像集合 (プレゼン資料) に跨る質問応答タスク
- 演算が必要な回答に対して, 演算過程 (算術式) を人手で付与
- 画像選択と質問応答タスクを統合的に解くモデルを提案

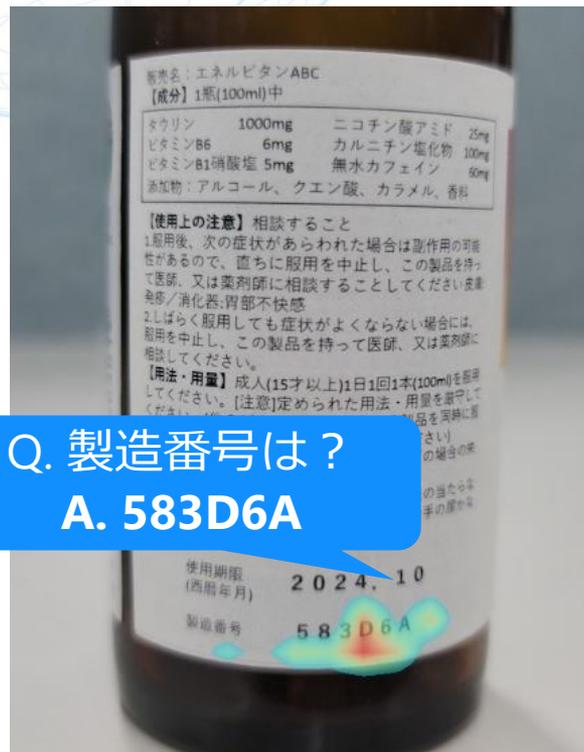


Q: What is the difference in the competition media percentage between East and the region with 12% of journalists? **A: 5% (11% - 6%)**

(12%のジャーナリストが居る地域 (=West) とEastのcompetition mediaの%差分は?)

日本語の視覚的読解モデルの構築

- 日本語文書でもデータを収集してモデルを構築
- 実用化に向けて取り組み中

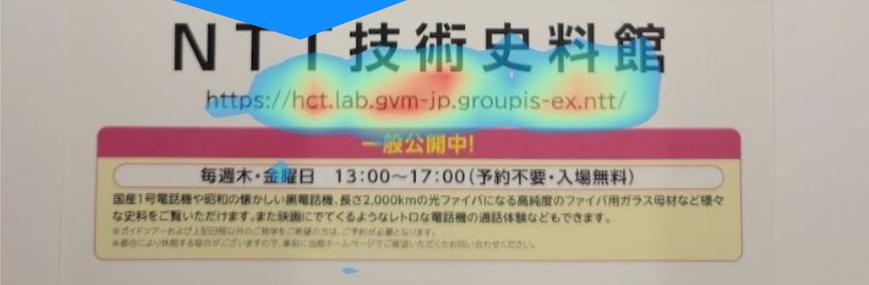


Q. 製造番号は？

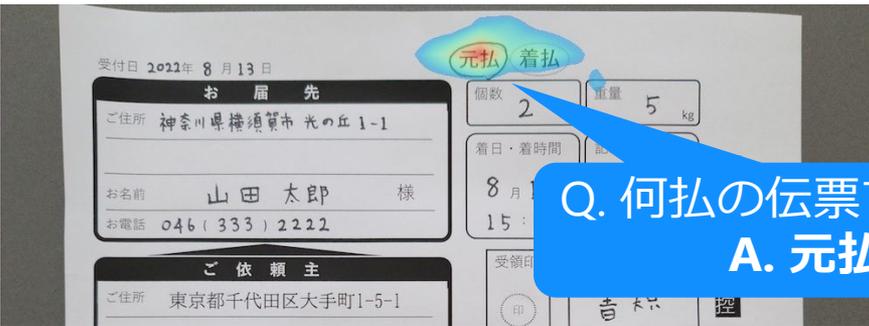
A. 583D6A

ビンのラベルのような
歪んだテキスト

Q. 技術史料館のホームページはどこ？
A. <https://hct.lab.gvm-jp.groupis-ex.ntt/>



カラフルなパンフレット



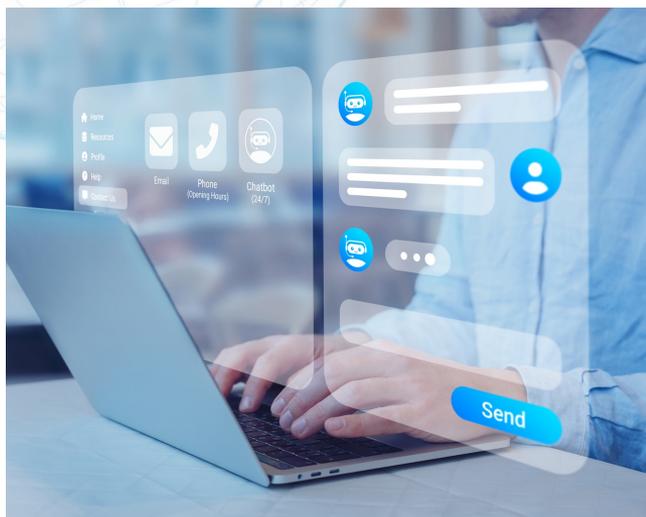
Q. 何払の伝票ですか？

A. 元払

手書きで支払種別を選択している伝票

行動の融合へどこから始めるか

現実世界の良い縮図であるPC環境において、
人と対話的に協働できる汎用ソフトウェアロボットを目指したい

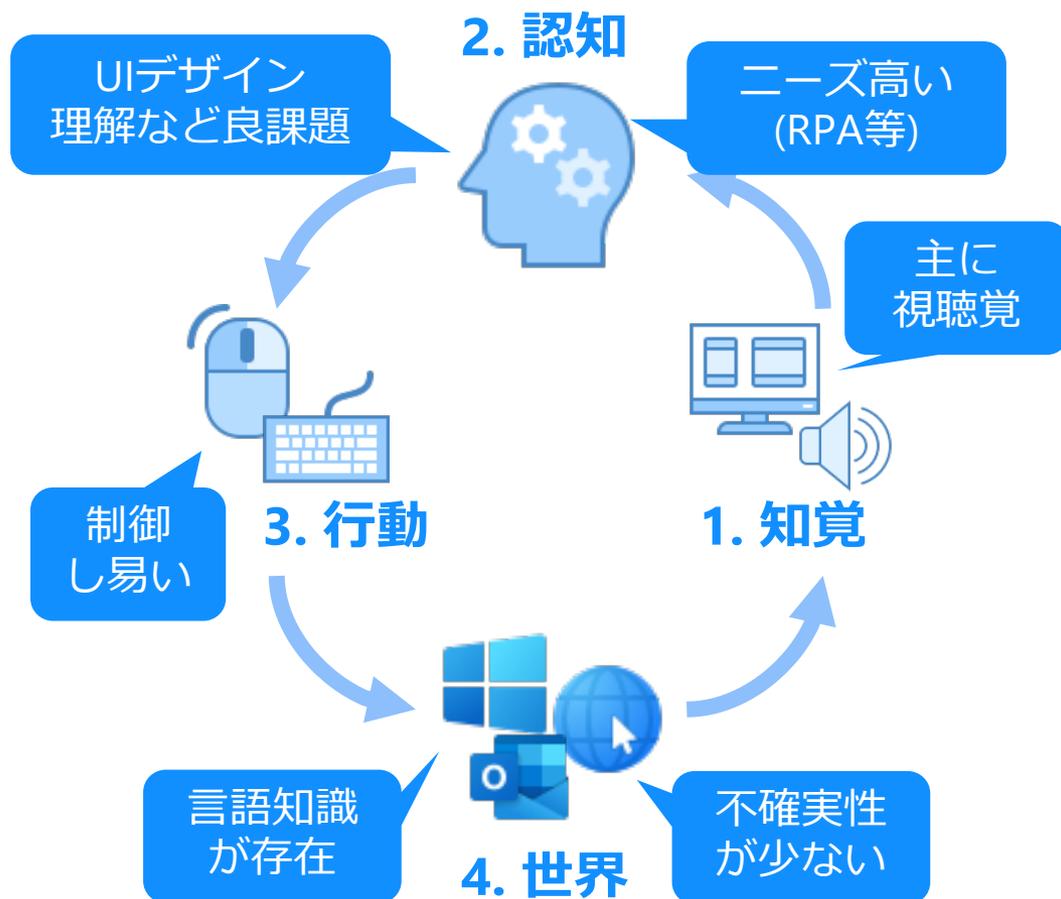


計算機の中に**コラボレーター**を創る



明後日のWeb会議のIDを
壹岐さんに伝えてくれる？

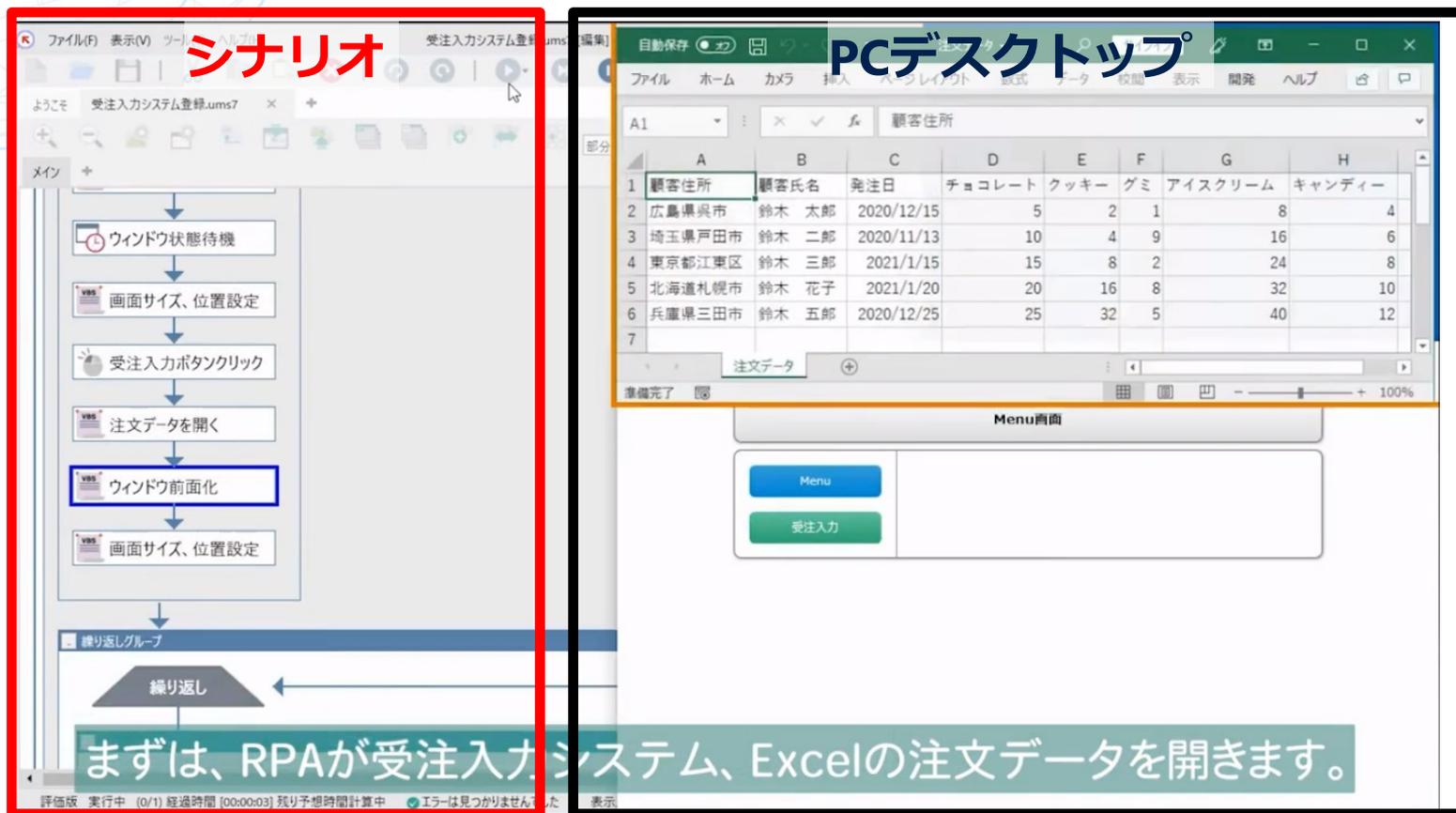
OKです！（Zoomで確認して、
SlackのDMで送ろう）



2. PC環境において人と協働できる 汎用ソフトウェアロボットの最新動向

既存のソフトウェアロボットによる自動化 NTT

- 定型作業をシナリオに書き起こし、シナリオに従ってPCデスクトップ上の応用ソフトを操作。



The image displays the WinActor RPA interface. On the left, a scenario flowchart is shown with steps: ウィンドウ状態待機, 画面サイズ、位置設定, 受注入力ボタンクリック, 注文データを開く, ウィンドウ前面化, 画面サイズ、位置設定. A red box highlights the first five steps, and the word 'シナリオ' is written in red. Below the flowchart is a '繰り返しグループ' (Repeat Group) containing a '繰り返し' (Repeat) step. On the right, a PC desktop simulation is shown with an Excel spreadsheet titled '顧客住所' (Customer Address) and a '注文データ' (Order Data) window. The Excel spreadsheet contains the following data:

	A	B	C	D	E	F	G	H
1	顧客住所	顧客氏名	発注日	チョコレート	クッキー	グミ	アイスクリーム	キャンディー
2	広島県呉市	鈴木 太郎	2020/12/15	5	2	1	8	4
3	埼玉県戸田市	鈴木 二郎	2020/11/13	10	4	9	16	6
4	東京都江東区	鈴木 三郎	2021/1/15	15	8	2	24	8
5	北海道札幌市	鈴木 花子	2021/1/20	20	16	8	32	10
6	兵庫県三田市	鈴木 五郎	2020/12/25	25	32	5	40	12
7								

Below the Excel spreadsheet is a 'Menu画面' (Menu Screen) with two buttons: 'Menu' and '受注入力' (Input). A green box highlights the '受注入力' button. At the bottom, a green banner contains the text: 'まずは、RPAが受注入力システム、Excelの注文データを開きます。' (First, RPA opens the input system and Excel's order data.)

やりたい作業



人が分解 (ソフトウェア更新対応も)

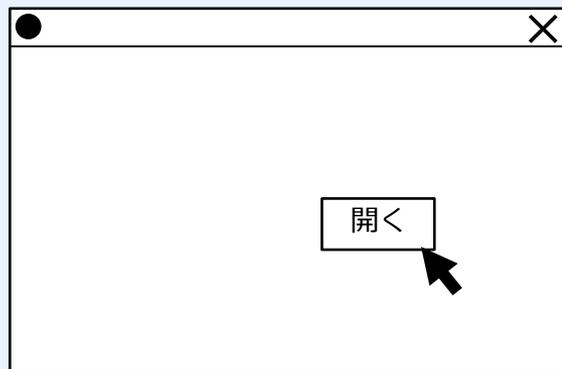
進行

シナリオ

- ✓ ソフトを起動
- ✓ 開くメニューをクリック
- ✓ ファイル名を入力
- ⋮
- もしログイン画面が出たら...
- ダウンロード完了まで待機
- ⋮

操作
観測

PC環境



揺らぎに弱い

人と協働するソフトウェアロボットへ

やりたい作業

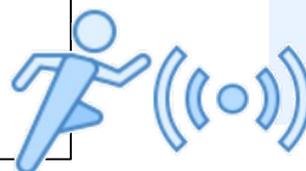
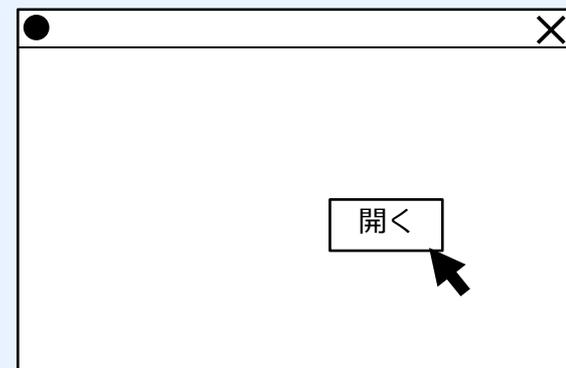
自律的・柔軟性

プラン 

進行

- ✓ ソフトを起動
- ✓ 開くメニューをクリック
- ✓ ファイル名を入力
- ⋮
- もしログイン画面が出たら...
- ダウンロード完了まで待機
- ⋮

PC環境



- **プランニング**

例: XX表のデータをwebアプリに入力 → XX表を開く、webアプリを開く

- **UI・画面の理解**

例: なんのソフトか、何が書いてあるか、メニューボタンはどこか、特定のアイコンがあるか、...

- **推論**

例: プログレスバーが出ているので準備中

- **次の操作の決定**

例: (プラン、画面の状態、推論結果、操作履歴 [...、click、input、wait])
→ click

PC環境に閉じているが、全自動化には理解・
行動にまたがる技術の融合が必要

→ 実世界の縮図

Web navigation

- 指示文やタスクの目的に従ってweb browserとインタラクション。
関わりが深い技術要素: プランニング・推論・操作の最終判断
(グラフィカルなweb browserを直接操作することは少ない。テキストベースのブラウザやDOMの操作で近似)
- PC環境の全体を用意するのは高コストなため、インタラクティブな研究対象として利用することが多い。
- **主なベンチマークタスク**
 - MiniWoB++ [Liu(Stanford)+,2018]: 多様なUI
 - WebShop [Yao(Princeton)+,2022]: より現実的なECサイト

DOM (Document Object Model): HTML文書やXML文書をつリー構造で表して操作するインターフェース。ツリーのノードが文書の部分を表す。

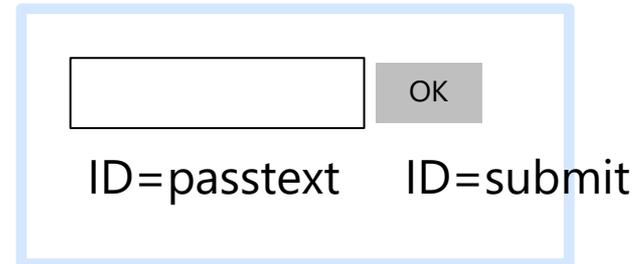
Web Navigationにおける行動

- Web navigationではページ上のインタラクティブ可能な要素にIDを割り当て、clickやinputを行動の単位とすることが多い。

例:

click (ページ上の要素ID)

input (ページ上の要素ID, テキスト)



- ただし、同じベンチマークでもあまりコンセンサスが取れてはいない。

例:

– クリックの対象を座標で指定

– Inputでテキストの代わりにキー入力を使用

- 詳しくはそれぞれの論文を参照のこと

World of Bits / MiniWoB++

[Shi(Stanford, OpenAI)+,2017] / [Liu(Stanford)+,2018]

- 100以上のwebインタラクティブ環境からなるベンチマーク。
- 単純動作(ボタン押下)、簡単な計算、フライト予約の一連の流れ、移動する物体への追従など多様なタスクで構成。
- 難易度: UIは豊富だが、人がやる場合、慣れればほぼ100%成功できる。

<p>Move the cube around so that "5" is the active side facing the user.</p> <p>Submit</p>	<p>Set the sliders to the combination [13,20,13] and submit.</p> <p>Submit</p>	<p>Draw the number "2" in the checkboxes using the example on the right and press Submit when finished.</p> <p>Submit</p>	<p>Select 5 with the spinner and hit Submit.</p> <p>Select a value: 0</p> <p>Submit</p>	<p>Keep your mouse inside the circle as it moves around.</p>	<p>Move the cube around so that "4" is the active side facing the user.</p> <p>Submit</p>
<p>Copy the text in the textareas below and paste it into the textbox.</p> <p>Enim. Elementum, tortor ullamcor</p> <p>Submit</p>	<p>Select 09/23/2016 as the date and hit submit.</p> <p>Date: _____</p> <p>Submit</p>	<p>Drag all rectangles into the black box.</p> <p>Submit</p>	<p>Enter "qu" into the text field and press Submit.</p> <p>qu</p> <p>Submit</p>	<p>Select all the shades of blue and press Submit.</p> <p>Submit</p>	<p>Find the 4th word in the paragraph, type that into the textbox and press "Submit".</p> <p>Non arcu ut ultricies est. Gravida gravida. Porta erat nulla eget condimentum posuere a</p> <p>Submit</p>
<p>Enter an item that starts with "Tun"</p> <p>Tags: _____</p> <p>Submit</p>	<p>Enter "Vb8" into the text field and press Submit.</p> <p>_____</p> <p>Submit</p>	<p>Focus into the 1st input textbox.</p> <p>_____ _____ _____</p>	<p>Focus into the textbox.</p> <p>_____</p>	<p>Move the cube around so that "2" is the active side facing the user.</p> <p>Submit</p>	<p>Select 5Gi and click Submit.</p> <p>5Gi sPUt</p> <p>Submit</p>

<http://proceedings.mlr.press/v70/shi17a/shi17a.pdf>

<https://arxiv.org/pdf/1802.08802.pdf>

[MiniWoB++ Documentation \(farama.org\)](https://farama.org/)

WebShop [Yao(Princeton)+,2022]

- リアルなECサイトを模した環境で指示に適合した商品を検索。
- 118万の商品を検索可能な模擬サイト、12087個の指示文を含むデータ。
- 難易度: Human expertの成功率59.6%

①指示から
キーワード
検索

②検索結果
からざっくり
指定

テキストで
画面を
受け取る

A

WebShop search

Instruction:
I'm looking for a small portable folding desk that is already fully assembled; it should have a khaki wood finish, and price lower than 140.00 dollars

portable folding desk khaki wood **1** Search

Back to Search **2** results

Page 1 (Total results: 50) [Next >]

3 Color: black khaki white

item-detail

item-detail

Buy Now **5**

Reward: 1.0

B

HTML mode

Simple mode

Instruction:
I'm looking for a small portable folding desk that is already fully assembled [...]
[btn] Back to Search [/btn]
Page 1 (Total results: 50) [btn] Next [/btn]
[btn] MENHG Folding Breakfast Tray [...] [/btn]
\$109.0
[btn] KPSP Folding Study Desk Bed [...] [/btn]

u (Instruction): I'm looking for a small portable...
 \bar{y} (Description): MENHG Folding Laptop Table Bed...
 y_{price} : \$109.0
 Y_{opt} (Options): { black, khaki, white }
 Y_{att} (Attributes): { steel pipe, no assembly, portable }

④必要に応じて
詳細を
確認

⑤模範回答
と比較して
報酬計算

③オプション
指定

UI・画面理解 1/2

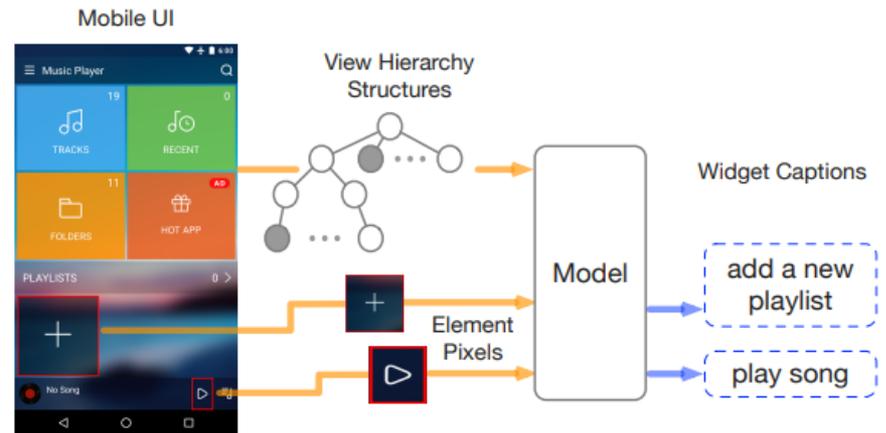
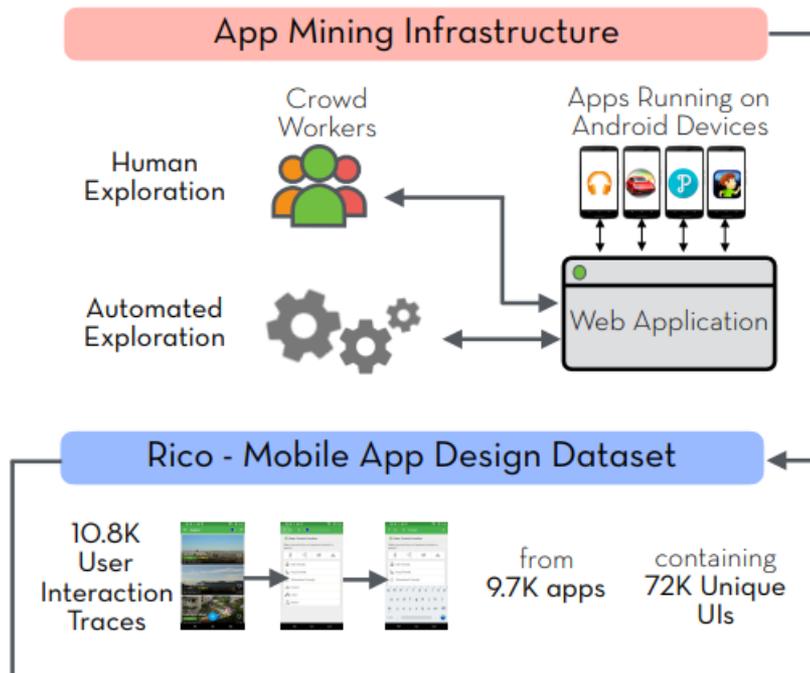
- UI・画面理解の研究も盛んにおこなわれ、リソースの蓄積が進む。

Rico [Deka(UIUC)+, 2017]

Androidアプリ画面のデータセット (スクリーンショット・メタデータ・操作記録)

Widget-captions [Li(Google)+, 2020]

UIの要素に対してキャプションをつける。Ricoのデータに対してアノテーション。

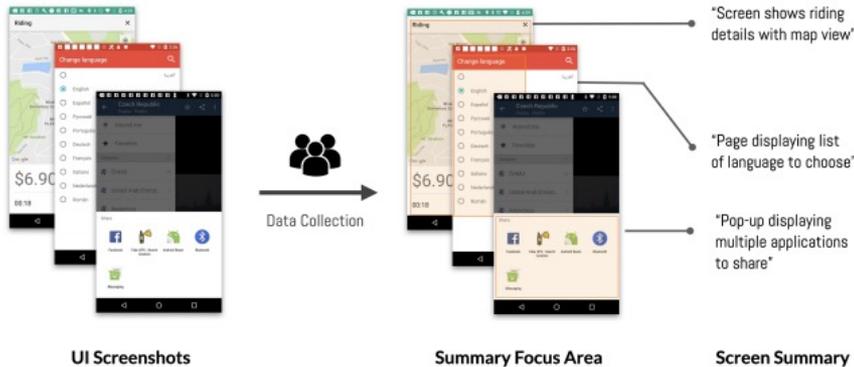


UI・画面理解 2/2

- UI・画面理解の研究も盛んにおこなわれ、リソースの蓄積が進む。

Screen2Words [Wang(U. Toronto)+,2021]
スクリーン全体に対して要約キャプションをつける。Ricoのデータに対してアノテーション。

WebUI [Wu(CMU HCII)+, 2023]
40万件のwebページのスクリーンショットとDOMツリーから自動生成した要素メタデータ



(要約キャプション + どこに着目したか)



Web navigation モデル

まだ人間の成績に追いついていないとは言えないが、
着実に性能が向上している web navigation モデルの変遷を紹介

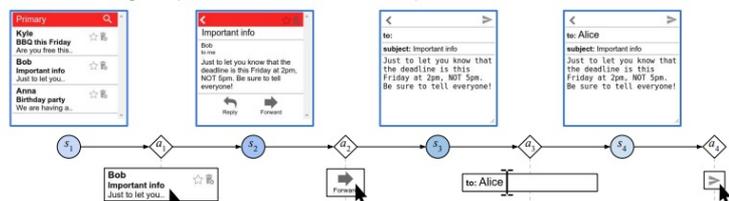
効率的な探索で改良を目指した研究

- MiniWoB++を題材に多様な手法が提案されている。
- 強化学習によって解く場合、**行動空間が大きい**ことが課題となるため、**効率的なサンプリング**を目指す研究が行われた。
- しかし、人間の成績とはギャップが残っていた。

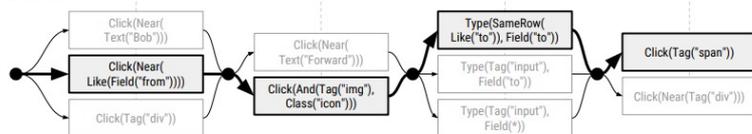
WGE [Liu(Stanford)+,2018]

行動を制約する**workflow**を導入し
サンプル効率を上昇

Demonstration: goal = {task: forward, from: Bob, to: Alice}



Workflow lattice:



QWeb [Gur(Google)+, 2018]

指示文を分解し、**カリキュラム**を導入。

Instruction: { from: WTK, to: LON, date: 10/21/2016 }



- **大規模データの導入**でこれまでの手法の性能を大幅に更新。
- 人手による6000時間分のMiniWoB++操作ログを使ったデモンストレーションでTransformerを事前学習(Behavior cloning)してから強化学習。

Pixel observation



Task instruction

Book the shortest one-way flight from: HVR to: Fayetteville, AR on 12/27/2016.

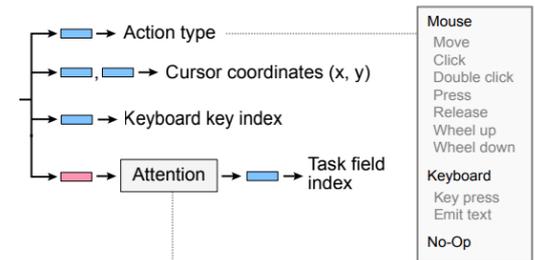
Task fields

critierion: 'shortest'
date: '12/27/2016'
from: 'HVR'
to: 'Fayetteville, AR'

Raw DOM tree

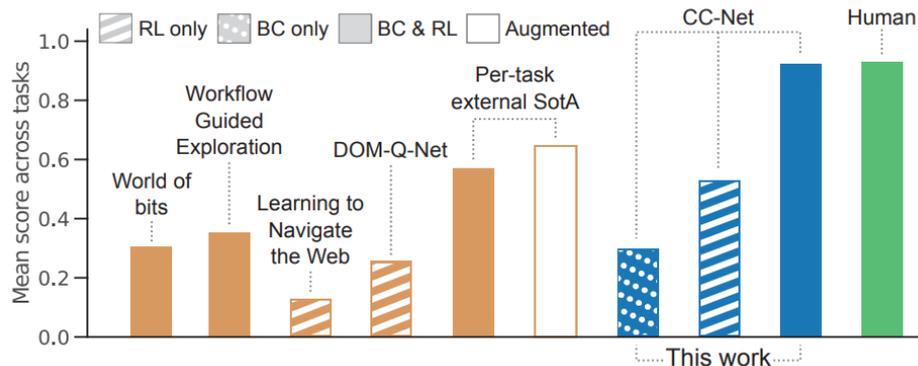
```

- [1] body : : None @ (0, 0) classes=[NO_CLASS] children=3
- [2] div : : wrap : None @ (0, 0) classes=[NO_CLASS] children=1
- [3] div : : area : None @ (0, 50) classes=[NO_CLASS] children=1
- [4] div : : menu : None @ (2, 60.3) classes=[NO_CLASS] children=5
- [5] h2 : : header-book : None @ (4, 60.3) text='Book Your One-Way Fl...' classes=[NO_CLASS]
- [6] div : : None @ (2, 79.1) classes=[input-container] children=1
- [7] input_text : : flight-from : None @ (4, 81.1) value= classes=[flight-input ui-autocomplete-input] placeholder=From:
- [8] div : : None @ (2, 104.39) classes=[input-container] children=1
- [9] input_text : : flight-to : None @ (4, 106.39) value= classes=[flight-input ui-autocomplete-input] placeholder=To:
- [10] div : : None @ (2, 144.67) classes=[departure-container] children=2
- [11] div : : None @ (2, 144.67) text='Departure Date' classes=[departure-header]
- [12] div : : None @ (2, 159.18) classes=[input-container] children=1
- [13] input_text : : datepicker : None @ (12, 161.18) value= classes=[flight-input hasDatepicker] placeholder=
- [14] div : : None @ (2, 182.18) classes=[search-container] children=1
- [15] button : : search : None @ (4, 184.18) text='Search' classes=[NO_CLASS]
- [16] div : : None @ (-1, 209) classes=[ui-helper-hidden-accessible]
- [17] div : : None @ (-1, 209) classes=[ui-helper-hidden-accessible]
    
```



視覚・テキスト入力

操作コマンドの出力



ReAct [Yao(Princeton)+, 2022]

- LLMは行動コマンドの使い方を例示するとそれに従うことができる。
- しかし、コマンドを並べただけの例示では**行動の目的を見失いがち**。

例に倣って商品を購入してください。
例

instruction: i would like a 3 ounce bottle of bright citrus deodorant for sensitive skin, and price lower than 50.00 dollars

action: *search*[3 ounce bright citrus deodorant sensitive skin]

observation: <検索結果ページ>

action: *click*[B078GWRC1J]

observation: <商品詳細ページ>

...

本番

instraction: get me a sixteen pack of apple cinnamon freeze dried banana chips, and price lower than 50.00 dollars

LLM>

LLM> action: *search*[sixteen pack apple cinnamon freeze dried banana chips]

...

action: *click*[B0061IVFZE]

	<p>B0061IVFZE Brothers-All-Natural Fruit Crisps, Strawberry Banana, 0.42 Ounce (Pack of 100) \$85.0</p> <p style="text-align: right; color: red; font-size: 2em;">X</p>
	<p>B096H2P6G2 Moon Fruit Freeze Dried Fruit Snacks. Fruit Snacks for Kids (Variety Pack) \$18.99</p>
	<p>B092JLLYK6 Nature's Turn Freeze-Dried Fruit Snacks - Banana Crisps - Perfect For School Lunches or an On-The-Go Snack - No Sugar Added, Non GMO, Gluten Free, Nothing Artificial (0.53oz) 6-Pack \$12.99</p>

<https://arxiv.org/pdf/2210.03629.pdf>

ReAct [Yao(Princeton)+, 2022]

- 行動に加えて思考過程も例示することを提案。
- LLMが思考過程を出力することでタスク成功率向上。

例に倣って商品を購入してください。
例

instruction: i would like a 3 ounce bottle of bright citrus deodorant for sensitive skin, and price lower than 50.00 dollars

action: *search*[3 ounce bright citrus deodorant sensitive skin]

observation: <検索結果ページ>

action: *think*[B078GWRC1J and B078GTKVXY are bright citrus deodorant less than 50 dollars. I can check B078GWRC1J first.]

Observation: OK.

action: *click*[B078GWRC1J]

...

LLM> action: *search*[sixteen pack apple cinnamon freeze dried banana chips]

...

action: *think*[B0061IVFZE is strawberry banana, not apple cinnamon. B096H2P6G2 is fruit snacks, not freeze dried banana chips. B092JLLYK6 is banana crisps, not apple cinnamon. I can check B092JLLYK6 first.]

action: *click*[B092JLLYK6]



WebShopにおける評価

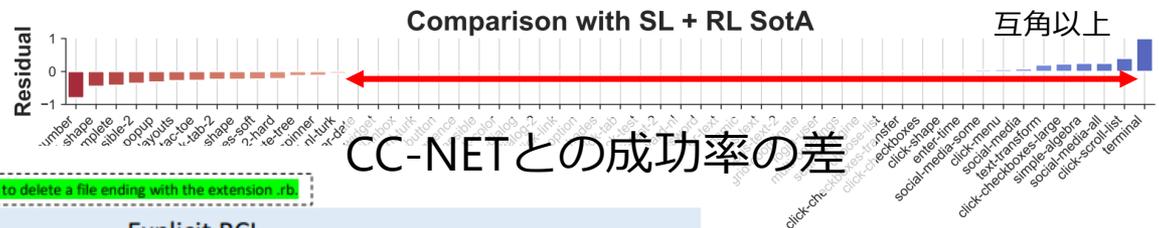
Method	Score	SR
Act	62.3	30.1
ReAct	66.6	40.0
IL	59.9	29.1
IL+RL	62.4	28.7
Human Expert	82.1	59.6

注: 図は論文を元に発表者が作成。実際にこのプロンプトで得られた出力かは不明

<https://arxiv.org/pdf/2210.03629.pdf>

RCI [Kim(UCI)+, 2023]

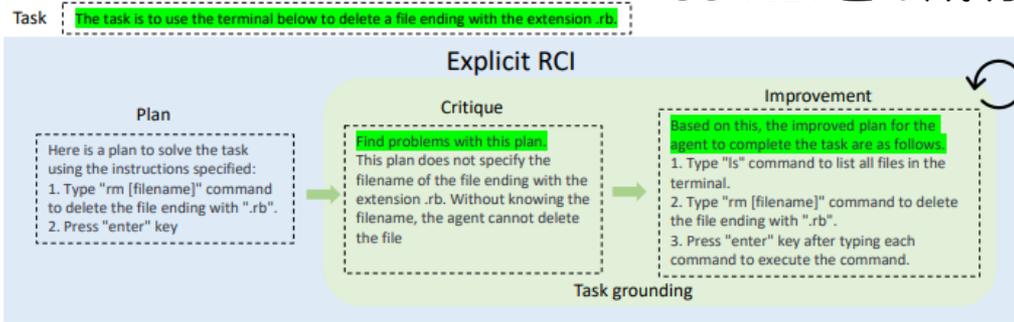
- LLM自身に生成結果を見直させるプロンプトで、**計画を自己改善**する Recursively Criticizes and Improves (RCI) を提案。
- MiniWoB++の複数タスクにおいてfew-shotでCC-Netと互角以上の成功率。



Use the terminal below to delete a file ending with the extension .rb

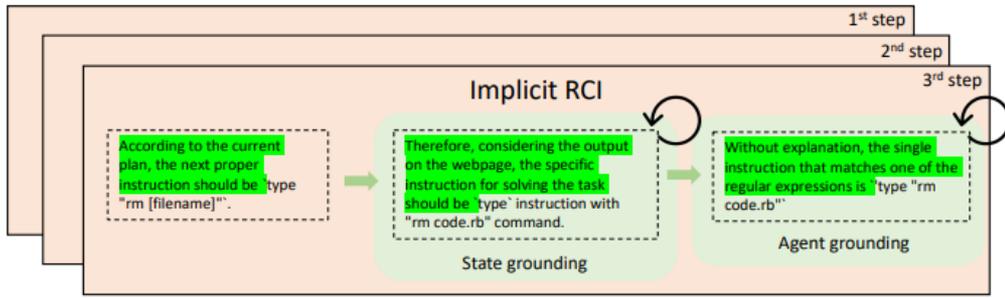
```
terminal
Welcome! Type help for a list
of available commands.
Last login: Mon Jan 30 2023
user$
```

Initial state



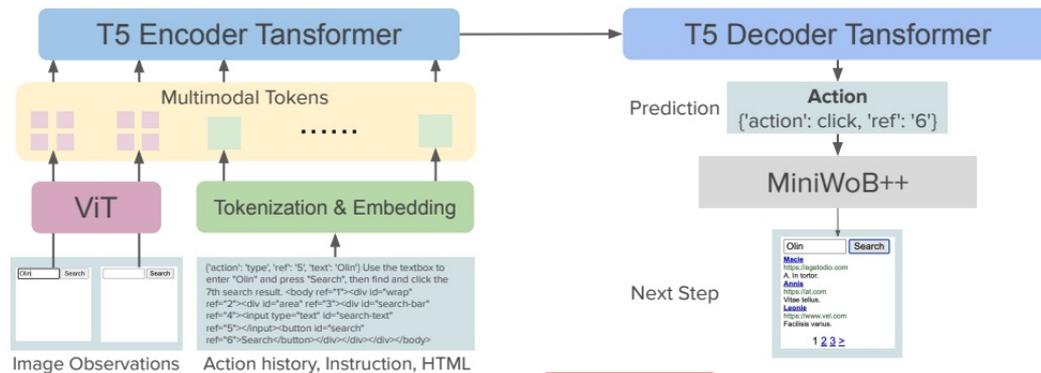
```
terminal
Welcome! Type help for a list
of available commands.
Last login: Mon Jan 30 2023
user$ ls
alloy.jpg code.rb file.gpg
navigator.txt skynet.html
thunderbird.html
user$
```

Current state



WebGUM [Fruta(UTokyo)+, 2023]

- MiniWoB++の学習データを既存モデルやスクリプトを使って自動で用意し、Flan-T5言語モデルをファインチューニング。
- より少ないデータ量で教師あり学習によるCC-Netを超える性能。
- MiniWoB++で学習後、WebShopにファインチューニングしベースラインを上回る。汎用的モデル構造と言語モデルの指示遂行力が効果的。



Methods	Training	Modality	Pre-trained Models	Dataset	Success Rate
CC-Net (Humphreys et al., 2022)	SL	DOM+Image	ResNet	2.4M	32.0%
WebN-T5 (Gur et al., 2022)	SL	HTML	T5-XL	12K	48.4%
WGE (Liu et al., 2018)	SL+RL	DOM	-	12K+	64.6%
CC-Net (Humphreys et al., 2022)	SL+RL	DOM+Image	ResNet	2.4M+	96.4%
WebGUM (Ours)	SL	HTML	Flan-T5-XL	347K	61.5%
WebGUM (Ours)	SL	HTML+Image	Flan-T5-XL, ViT-B16	347K	66.1%

MiniWoB++

Methods	Training	Models	Score	Success Rate
Rule	-	-	45.6	9.6%
IL	SL	BART, BERT	59.9	29.1%
IL+RL	SL+RL	BART, BERT	62.4	28.7%
Act	In-context	PaLM-540B	62.3	30.1%
ReAct	In-context	PaLM-540B	66.6	40.0%
WebGUM	SL	Flan-T5-XL	67.5	45.0%

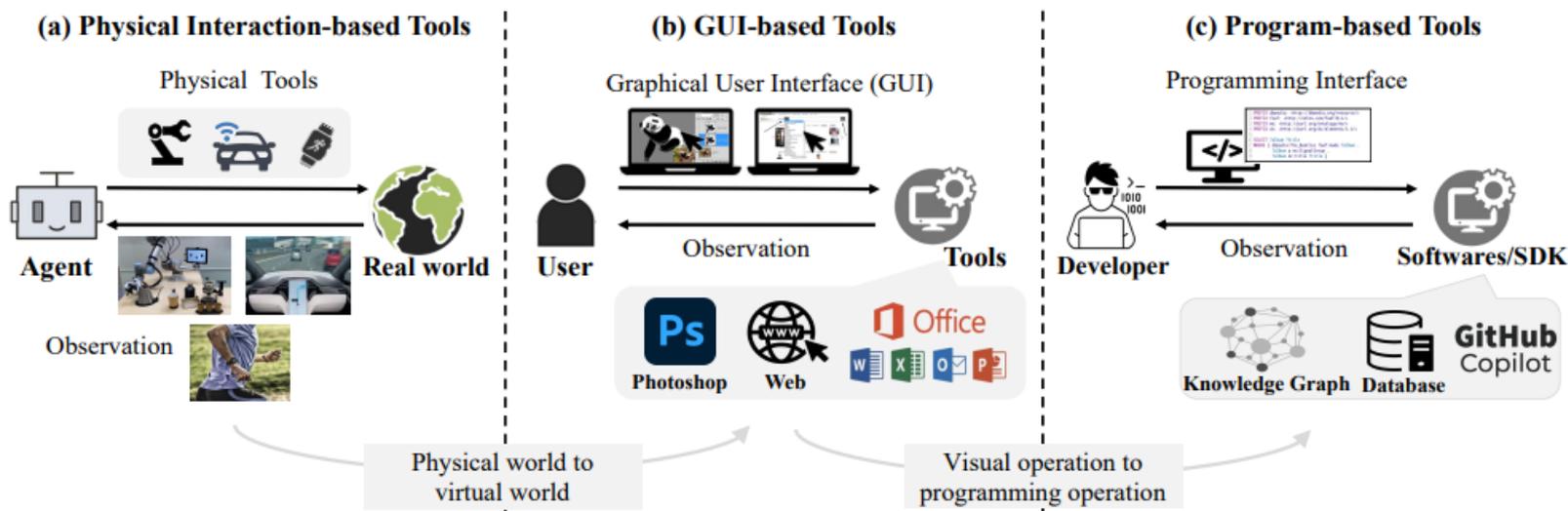
WebShop



今後の方向性

LLM・基盤モデル + ツールのパラダイム

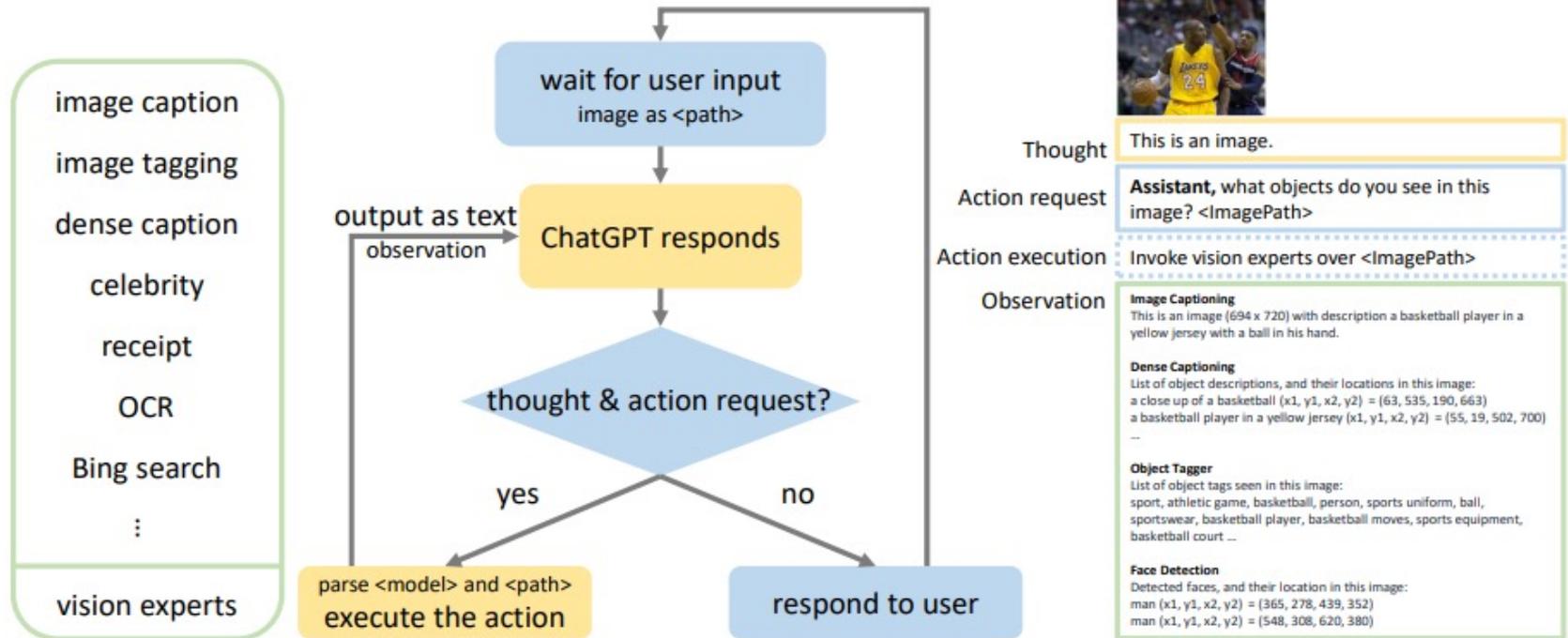
- **Tool learning** [Qin(Tsinghua University)+, 2023]
言語における強力な汎化能力によって、LLMは言語を媒介して**ツール** (言語I/F + 物理的な道具、GUI、プログラム) を使えるようになった。
- **TaskMatrix.AI** [Liang(Microsoft)+, 2023]
既存の基盤モデルが**他のAIモデル、システムAPI**を利用して、デジタルと物理の両方の領域で多様なタスクを遂行するエコシステムの構想



UIの観点から見たツールのカテゴリー [Qin(Tsinghua University)+, 2023]

例: MM-REACT [Yang(Microsoft)+, 2023]

- キャプション生成やOCRの結果をテキスト化してLLMに伝える **vision experts** によって、LLMをマルチモーダルに拡張。
- LLM: ChatGPT / gpt-3.5-turbo / GPT-4
- Vision expert: Azure Cognitive Services API



例: MM-REACT [Yang(Microsoft)+, 2023]

- Image captioning, dense Captioning, object tagger, OCR を駆使して入力画像をテキスト化する様子

①ユーザーが画像を入力



Convert image to `<ImagePath>` and send to ChatGPT

Thought 1 ChatGPT: This is an image.

Action 1 Assistant, what objects do you see in this image? `<ImagePath>`
Invoke vision experts, including image captioning, dense captioning, object tagger.

Observation 1 Image Captioning
Assistant: This is an image (747 x 720) with description a chalkboard sign with a wooden frame.

Dense Captioning

List of object descriptions, and their locations in this image:
a chalkboard sign with writing on it (x1,y1,x2,y2) = (4, 6, 720, 702)
a close up of a blackboard (x1,y1,x2,y2) = (118, 323, 579, 497)
a chalk drawing of a slice of pizza (x1,y1,x2,y2) = (590, 47, 734, 231)
a white letter on a black board (x1,y1,x2,y2) = (101, 531, 619, 653)
a letter on a black surface (x1,y1,x2,y2) = (475, 433, 525, 496)
a white letter on a grey background (x1,y1,x2,y2) = (142, 620, 360, 689)
a heart drawn in chalk (x1,y1,x2,y2) = (55, 543, 113, 611)

Object Tagger

List of object tags seen in this image:
text, blackboard, handwriting, chalk, slate, menu

②LLMが使うexpertを指定して画像をテキスト化

<https://arxiv.org/pdf/2303.11381.pdf>

Thought 2 ChatGPT: This image is likely a menu or a chalkboard sign. Let me ask for more information about the text.

Action 2 Assistant, what are the OCR texts in this image? `<ImagePath>`
Invoke OCR

Observation 2 OCR
This image contains handwritten text

List of texts (words) seen in this image:
Today's
Specials!
7AM-
create your
Own Pizza
\$9.99
Schicken marsala
asparagus, soup of
the day mashed notat
each

ChatGPT: This image is a menu with Today's Specials and create your own pizza for \$9.99. The soup of the day is mashed notat and chicken marsala with asparagus. The text is in English and German.

③質問

I am getting just two custom pizza for me and my friend. How much should I pay in total?

ChatGPT: Based on the information you provided, you are ordering two custom pizzas. The price of each custom pizza is \$9.99. So, the total cost of two custom pizzas would be \$9.99 x 2 = \$19.98.

④回答

- ソフトウェアロボットによる自動化の研究は「LLM・基盤モデル + ツールのパラダイム」の下で加速していくと私たちは考える。
(Web Navigationは多様なサービスへのアクセス手段として重要)
- 大きく2つの方向性がある。
 - A. ツール (モデルとPC環境の仲介) の改良
 - B. 頭脳 (LLM・基盤モデル) の改良
- また、一定の性能を有した自動化モデルを作れるようになってきたため、今後は協働に適したインターフェースへの関心が高まると考える。

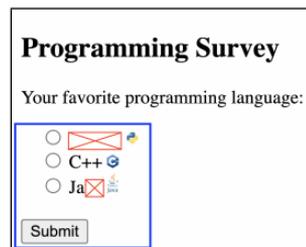
A. ツール (モデルとPC環境の仲介) の改良の方向性

GUI Expert

- GUIとLLM・基盤モデルの仲介をする GUI Expert
 - 全てのソフトウェアがLLMとの直接対話に対応するとは限らない。
 - 真の汎用性の実現にはウェブブラウザ (DOM) など特定のソフトウェアに依存しないキー、マウス操作のような手段を確保したい。
 - PC上の身体の実現？
- 既存の資源を使ってGUI Expertは構築できるか。
 - UI・画像理解のデータ資源を活用
 - GUIを文書ととらえることで文書画像読解の技術が役立つ可能性



:



→

```
<<<Python>
<img_src=py_logo.png img_alt=Python>>
<<C++>
<img_src=cpp_logo.png img_alt=C++>>
<<Java>
<img_src=java_logo.png img_alt=Java>>
<Submit>>
```

文書画像を見たまま読むPix2Struct [Lee(Google)+, 2022]

<https://arxiv.org/pdf/2210.03347.pdf>

協働に適したインターフェースの追求

- 既存研究は与えられたひとつの指示文を遂行する能力に注力して来た (大規模言語モデル前はそれで十分難しかった)。
- 今後は環境とのインタラクション、ユーザーとのインタラクションの双方を考慮した研究開発が本格化する。



AgentGPTのwebUI

目標を入れたあと、人とどうか関わるべきか (Auto-GPTやgodmodeなどはフィードバックを入れられる)。

シンプルなチャットUIに比べるとかなり情報量が多い。どのように提示するか

セクションのまとめ

- ソフトウェアロボットによる自動化は、PC環境に閉じているが、理解・行動にまたがる技術の融合が必要であり、実世界の縮図のような対象。
- 研究タスクとして Web navigation、UI・画面理解のタスクを紹介。
- Web navigation モデルの動向
 - 効率的な探索で改良を目指した研究
 - 大規模なデモンストレーションデータを用いた性能向上
 - LLM・基盤モデルを用いた効率的なモデル構築着実な性能向上もまだ人とはギャップ有り。
- 今後「LLM・基盤モデル + ツールのパラダイム」のもとで研究が加速。
 - ツール: GUI Expert・PC上の身体
 - 頭脳: 仮想世界・実世界をつなぐ基盤モデル
 - 協働に適したインターフェースの追求

- GPT-4は人レベルの汎用的な自然言語処理能力を身につけた
 - 視覚との融合理解については、おそらくまだ人のレベルには達していないが、かなり近いレベルまで到達したように見える
 - LLMがマルチモーダルなハブとして視覚以外の理解・生成にも同様のアプローチで広がっていくだろう
- これからPC上で人を対話的にアシスタントするAIが普及していく
 - 多くのソフトウェアが、LLMから操作して貰いやすくするためにAPIを備え、“co-pilot化”するであろう
 - あらゆるソフトウェアに対応できる真の汎用性に向けて、PC上の身体性（マウスやキーボードの操作、OSという世界のモデル化）の獲得が次の大きな未解決課題と考えている
- その先にはロボットとLLMが繋がっていく
 - Webに変わる事前学習用のデータ・環境を用意できるか？
 - LLMが持つプランニング能力（世界モデル）は実世界でも通用するのか？
 - あらゆる環境で人と自然に協調可能なAIロボットを作れるか？

参考文献 (Part 1)

1. OpenAI, Introducing ChatGPT. Available at <https://openai.com/blog/chatgpt>, November 30, 2022
2. OpenAI: GPT-4 Technical Report. CoRR abs/2303.08774 (2023)
3. Tom B. Brown et al.: Language Models are Few-Shot Learners. NeurIPS 2020 / CoRR abs/2005.14165 (2020)
4. Long Ouyang et al.: Training language models to follow instructions with human feedback. CoRR abs/2203.02155 (2022)
5. Sébastien Bubeck et al.: Sparks of Artificial General Intelligence: Early experiments with GPT-4. CoRR abs/2303.12712 (2023)
6. Daniel Martin Katz+, GPT-4 Passes the Bar Exam. <http://dx.doi.org/10.2139/ssrn.4389233>, March 15, 2023
7. Alexander Pan et al.: Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the MACHIAVELLI Benchmark. CoRR abs/2304.03279 (2023)
8. Baolin Peng et al.: Instruction Tuning with GPT-4. CoRR abs/2304.03277 (2023)
9. Yizhong Wang et al.: Self-Instruct: Aligning Language Model with Self Generated Instructions. CoRR abs/2212.10560 (2022)
10. Hugo Touvron et al.: LLaMA: Open and Efficient Foundation Language Models. CoRR abs/2302.13971 (2023)
11. OpenAI: GPT-4. <https://openai.com/research/gpt-4>, March 14, 2023.
12. Geewook Kim et al.: OCR-Free Document Understanding Transformer. ECCV (28) 2022: 498-517
13. Shaohan Huang et al.: Language Is Not All You Need: Aligning Perception with Language Models. CoRR abs/2302.14045 (2023)
14. OpenAI: ChatGPT plugins. <https://openai.com/blog/chatgpt-plugins>, March 23, 2023.
15. Microsoft: Introducing Microsoft 365 Copilot – your copilot for work. <https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/>, March 16, 2023.
16. Yongliang Shen et al.: HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace. CoRR abs/2303.17580 (2023)
17. Auto-GPT. i ment, <https://github.com/Significant-Gravitas/Auto-GPT>
18. Llama.cpp. <https://github.com/ggerganov/llama.cpp>
19. Deyao Zhu et al.: MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. CoRR abs/2304.10592 (2023)
20. Junnan Li et al.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. CoRR abs/2301.12597 (2023)
21. UC Berkeley, CMU, Stanford, MBZUAI, and UC San Diego: Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, <https://vicuna.lmsys.org/>, 2023/03/19

1. Shi, Tianlin, et al. "World of bits: An open-domain platform for web-based agents." International Conference on Machine Learning. PMLR, 2017.
2. Liu, Evan Zheran, et al. "Reinforcement Learning on Web Interfaces using Workflow-Guided Exploration." International Conference on Learning Representations.(2018).
3. Yao, Shunyu, et al. "WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents." Advances in Neural Information Processing Systems. (2022)
4. Deka, Biplab, et al. "Rico: A mobile app dataset for building data-driven design applications." Proceedings of the 30th annual ACM symposium on user interface software and technology. 2017.
5. Li, Yang, et al. "Widget Captioning: Generating Natural Language Description for Mobile User Interface Elements." Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020.
6. Wang, Bryan, et al. "Screen2words: Automatic mobile UI summarization with multimodal learning." The 34th Annual ACM Symposium on User Interface Software and Technology. 2021.
7. Wu, Jason, et al. "WebUI: A Dataset for Enhancing Visual UI Understanding with Web Semantics." Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023.
8. Gur, Izzeddin, et al. "Learning to navigate the web." arXiv preprint arXiv:1812.09195 (2018).
9. Humphreys, Peter C., et al. "A data-driven approach for learning to control computers." International Conference on Machine Learning. PMLR, 2022.
10. Yao, Shunyu, et al. "ReAct: Synergizing Reasoning and Acting in Language Models." NeurIPS 2022 Foundation Models for Decision Making Workshop.
11. Kim, Geunwoo, Pierre Baldi, and Stephen McAleer. "Language models can solve computer tasks." arXiv preprint arXiv:2303.17491 (2023).
12. Furuta, Hiroki, et al. "Instruction-Finetuned Foundation Models for Multimodal Web Navigation." Workshop on Reincarnating Reinforcement Learning at ICLR 2023.
13. Qin, Yujia, et al. "Tool learning with foundation models." arXiv preprint arXiv:2304.08354 (2023).
14. Liang, Yaobo, et al. "Taskmatrix. ai: Completing tasks by connecting foundation models with millions of apis." arXiv preprint arXiv:2303.16434 (2023).
15. Yang, Zhengyuan, et al. "Mm-react: Prompting chatgpt for multimodal reasoning and action." arXiv preprint arXiv:2303.11381 (2023).
16. Lee, Kenton, et al. "Pix2Struct: Screenshot parsing as pretraining for visual language understanding." arXiv preprint arXiv:2210.03347 (2022).
17. Driess, Danny, et al. "Palm-e: An embodied multimodal language model." arXiv preprint arXiv:2303.03378 (2023).