LandingAI

Agents, Inference and Hardware

Andrew Ng

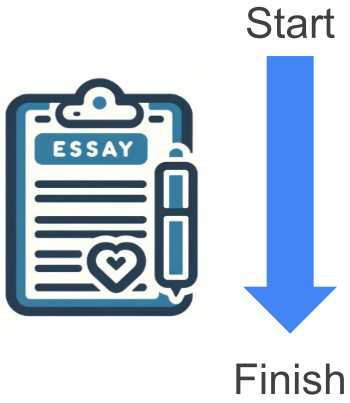
Technical Trends In AI

- **On-device AI.** Instead of running an LLM in the cloud, run it on your own laptop, phone or industrial PC.
- **Image/Video analysis.** LLMs brought us the text processing revolution. The visual processing revolution is coming – not just generation, but analysis. This will affect, manufacturing, life sciences, self-driving, retail, etc.
- **AI Agentic Workflows.** Given an instruction (“research topic X for me”) software that can carry out a sequence of steps to generate a result.

LLM-Based Agents

Non-agentic workflow (zero-shot)

Please type out an essay on topic X from start to finish in one go, without using backspace.



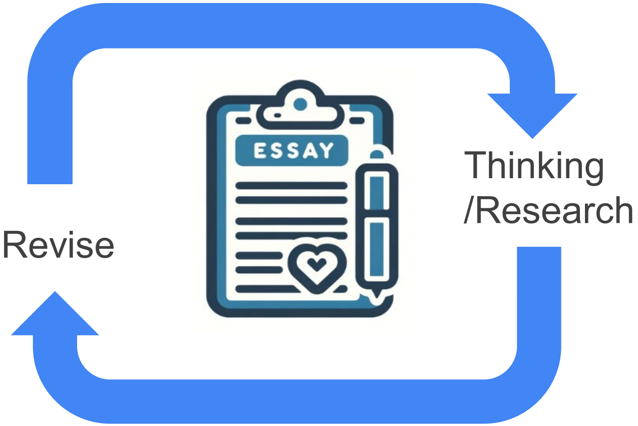
Start

Finish

Agentic workflow

- Write an essay outline on topic X
- Do you need any web research?
- Write a first draft.
- Consider what parts need revision or more research.
- Revise your draft

….

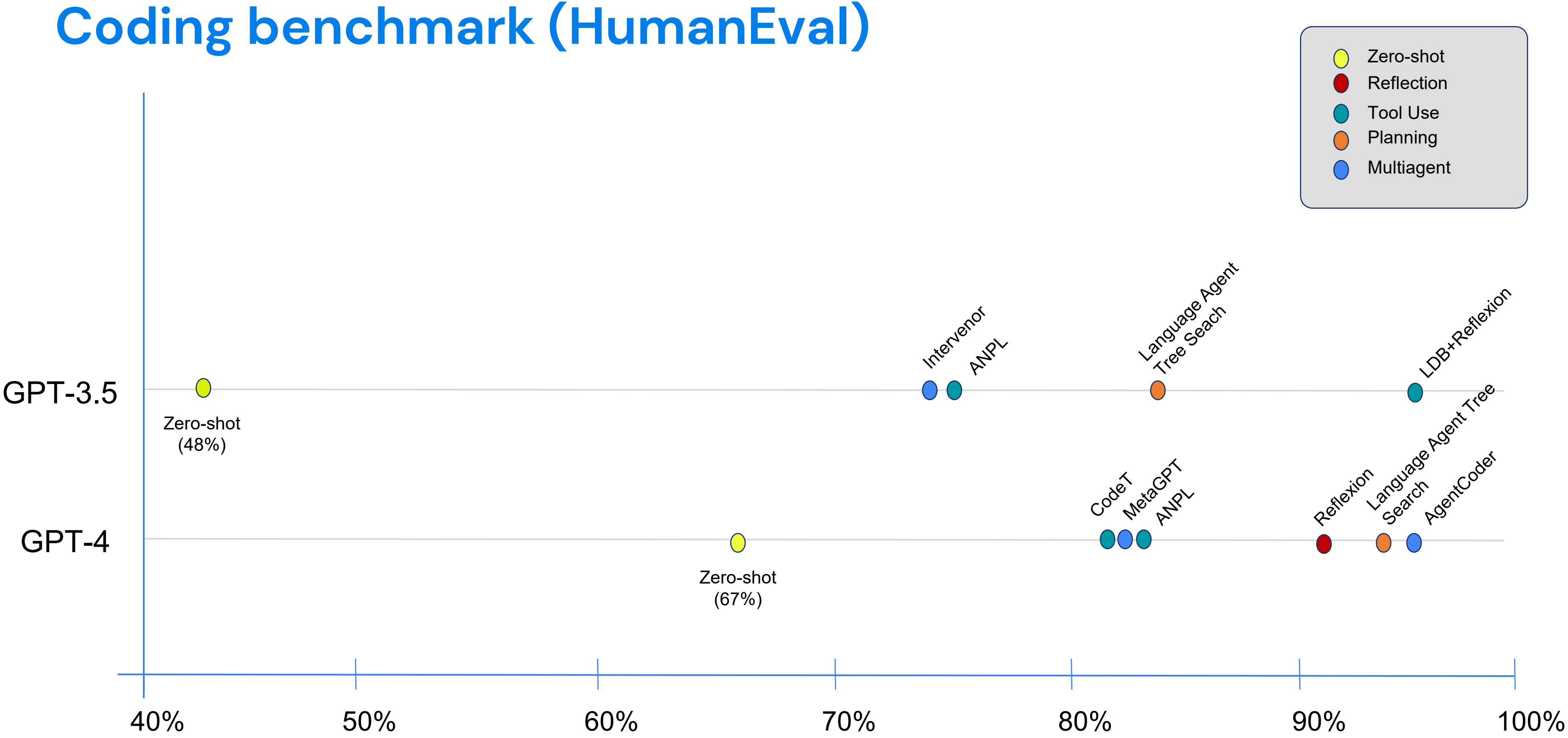


Thinking /Research

Revise

LandingAI

Coding benchmark (HumanEval)



[Thanks to Joaquin Dominguez and John Santerre (DeepLearning.AI) for help with analysis.]

LandingAI

Reflection with LLMs

Here’s code intended for {task}:

```
def do_task (x):
     ...
```

Check the code carefully for correctness, style and efficiency, and give constructive criticism for how to improve it.

Please write code for {task}

```
def do_task(x): ...
```

```
def do_task_v2(x):
```

```
def do_task_v3(x):
```

There’s a bug on line 5. Fix it by …

It failed Unit Test 3. Try changing …

Coder
Agent (LLM)

Related work:

- Self-Refine: Iterative Refinement with Self-Feedback, Madaan et al. (2023)
- Reflexion: Language Agents with Verbal Reinforcement Learning, Shinn et al., (2023)

LandingAI

Reflection with LLMs

Please write code for {task}

```
def do_task(x): ...
```

```
def do_task_v2(x):
```

```
def do_task_v3(x):
```

Coder Agent (LLM)

Here's code intended for {task}:

```
def do_task (x):
    ...
```

Check the code carefully for correctness, style and efficiency, and give constructive criticism for how to improve it.

There's a bug on line 5. Fix it by …

It failed Unit Test 3. Try changing …

Critic Agent (LLM)

Recommended reading:

- Self-Refine: Iterative Refinement with Self-Feedback, Madaan et al. (2023)
- Reflexion: Language Agents with Verbal Reinforcement Learning, Shinn et al., (2023)

Agentic Reasoning Design Patterns

1. Reflection

2. Multi-agent collaboration

3. Tool use (API calls)

4. Planning (decide on steps for task)

The importance of inference

1. Agentic workloads use many more tokens than zero-shot prompting.

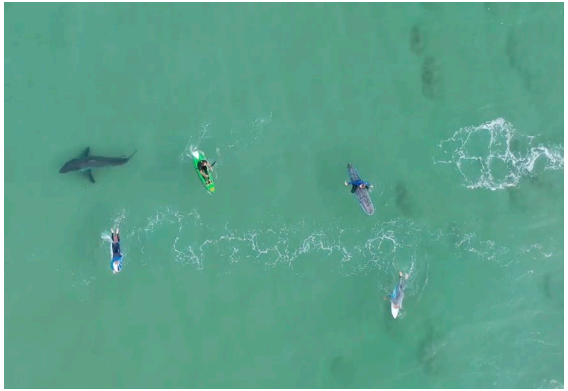
2. Fast, low cost token generation will be a huge performance driver. Open weight models (like Llama 3.1) also make it easier for providers to compete directly on inference price and speed.

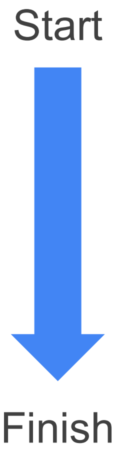
3. Training remains important, but I hope our community will also invest significantly in inference. (E.g., SambaNova, Cerebras, Groq). Also see benchmarks at artificialanalysis.ai

LMM-Based Agents

Non-agentic workflow (zero-shot)

Watch this video and tell me if any surfer is within 10m of a shark

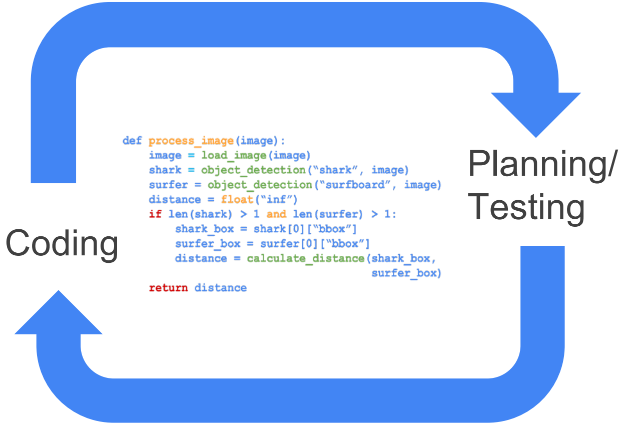




Video credit: Crazy Nature [youtu.be/2_HfdtljBKk?si=3rKt7StDATeuFYjS]

Agentic workflow

1. Detect the sharks (bounding boxes)
2. Detect the surfers (bounding boxes)
3. Compute distances between sharks & surfers
4. Determine if any are <10m
5. Iterate through Steps 1-4 for all frames in video



```
def process_image(image):
    image = load_image(image)
    shark = object_detection("shark", image)
    surfer = object_detection("surfboard", image)
    distance = float("inf")
    if len(shark) > 1 and len(surfer) > 1:
        shark_box = shark[0]["bbox"]
        surfer_box = surfer[0]["bbox"]
        distance = calculate_distance(shark_box,
                                      surfer_box)
    return distance
```

LandingAI

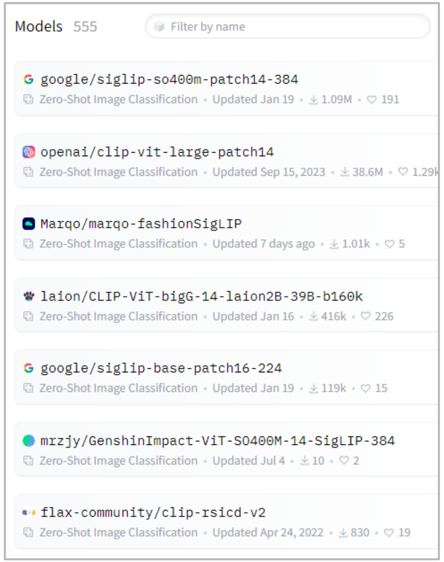
Demo:

Mercedes Logo

Tool use: Computer Vision Models

+222.9% Y/Y
144 → 465

Zero-Shot Image Classification



Models 555 | Filter by name

google/siglip-so400m-patch14-384
Zero-Shot Image Classification • Updated Jan 19 • ↓ 1.09M • ♡ 191

openai/clip-vit-large-patch14
Zero-Shot Image Classification • Updated Sep 15, 2023 • ↓ 38.6M • ♡ 1.29k

Marqo/marqo-fashionSigLIP
Zero-Shot Image Classification • Updated 7 days ago • ↓ 1.01k • ♡ 5

laion/CLIP-ViT-bigG-14-laion2B-39B-b160k
Zero-Shot Image Classification • Updated Jan 16 • ↓ 416k • ♡ 226

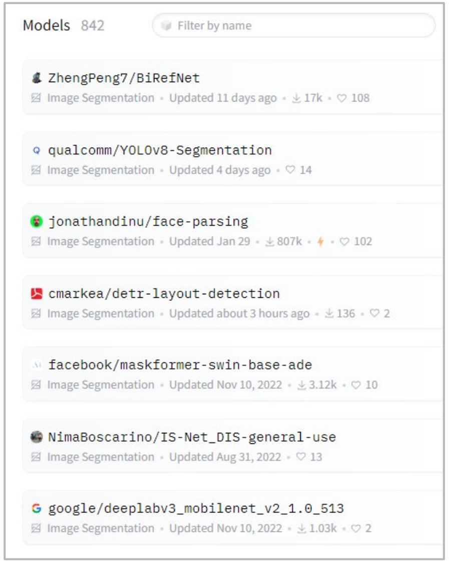
google/siglip-base-patch16-224
Zero-Shot Image Classification • Updated Jan 19 • ↓ 119k • ♡ 15

mrzjy/GenshinImpact-ViT-SO400M-14-SigLIP-384
Zero-Shot Image Classification • Updated Jul 4 • ↓ 10 • ♡ 2

flax-community/clip-rsicd-v2
Zero-Shot Image Classification • Updated Apr 24, 2022 • ↓ 830 • ♡ 19

+186.4% Y/Y
<200 → 842

Image Segmentation



Models 842 | Filter by name

ZhengPeng7/BiRefNet
Image Segmentation • Updated 11 days ago • ↓ 17k • ♡ 108

qualcomm/YOLOv8-Segmentation
Image Segmentation • Updated 4 days ago • ♡ 14

jonathandinu/face-parsing
Image Segmentation • Updated Jan 29 • ↓ 807k • ⚡ • ♡ 102

cmarkea/detr-layout-detection
Image Segmentation • Updated about 3 hours ago • ↓ 136 • ♡ 2

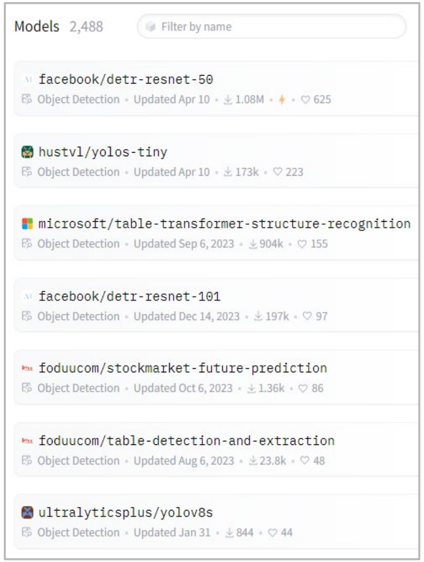
facebook/maskformer-swin-base-ade
Image Segmentation • Updated Nov 10, 2022 • ↓ 3.12k • ♡ 10

NimaBoscarino/IS-Net_DIS-general-use
Image Segmentation • Updated Aug 31, 2022 • ♡ 13

google/deeplabv3_mobilenet_v2_1.0_513
Image Segmentation • Updated Nov 10, 2022 • ↓ 1.03k • ♡ 2

+384.2% Y/Y
514 → 2.5k

Object Detection



Models 2,488 | Filter by name

facebook/detr-resnet-50
Object Detection • Updated Apr 10 • ↓ 1.08M • ⚡ • ♡ 625

hustvl/yolos-tiny
Object Detection • Updated Apr 10 • ↓ 173k • ♡ 223

microsoft/table-transformer-structure-recognition
Object Detection • Updated Sep 6, 2023 • ↓ 904k • ♡ 155

facebook/detr-resnet-101
Object Detection • Updated Dec 14, 2023 • ↓ 197k • ♡ 97

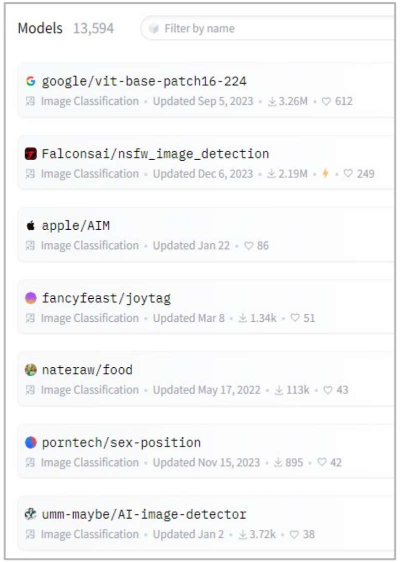
foduucom/stockmarket-future-prediction
Object Detection • Updated Oct 6, 2023 • ↓ 1.36k • ♡ 86

foduucom/table-detection-and-extraction
Object Detection • Updated Aug 6, 2023 • ↓ 23.8k • ♡ 48

ultralyticsplus/yolov8s
Object Detection • Updated Jan 31 • ↓ 844 • ♡ 44

+214.3% Y/Y
4.3k → 13.6k

Image Classification



Models 13,594 | Filter by name

google/vit-base-patch16-224
Image Classification • Updated Sep 5, 2023 • ↓ 3.26M • ♡ 612

Falconsai/nsfw_image_detection
Image Classification • Updated Dec 6, 2023 • ↓ 2.19M • ⚡ • ♡ 249

apple/AIM
Image Classification • Updated Jan 22 • ♡ 86

fancyfeast/joytag
Image Classification • Updated Mar 8 • ↓ 1.34k • ♡ 51

nateraw/food
Image Classification • Updated May 17, 2022 • ↓ 113k • ♡ 43

porntech/sex-position
Image Classification • Updated Nov 15, 2023 • ↓ 895 • ♡ 42

umm-maybe/AI-image-detector
Image Classification • Updated Jan 2 • ↓ 3.72k • ♡ 38

Source: Huggingface.co

LandingAI

Resources

- Try VisionAgent for free at va.landing.ai
- Open source code on github

LandingAI team here today

Steve Ackley

Ian Rysdale

LandingAI

Fireside Chat

Andrew Ng
Executive Chairman, LandingAI

Baskar Sridharan
VP of AI/ML services & infrastructure at AWS at