



ChatGPT and Conversational AI with Reinforcement Learning from Human Feedback (RLHF)

Time: 2023/02/17 (Fri) 17:50-19:30

Place: R5117, Soochow University, Taiwan

Host: Prof. Shih-Wei Liao, NTU and Inclusion Technology, NSTC, Taiwan



Min-Yuh Day, Ph.D.
Associate Professor

Institute of Information Management, National Taipei University

<https://web.ntpu.edu.tw/~myday>





Min-Yuh Day, Ph.D.

aws academy

Accredited Educator

aws certified

Solutions Architect

Associate

aws certified

Cloud Practitioner

**Associate Professor, Information Management, NTPU
Visiting Scholar, IIS, Academia Sinica
Ph.D., Information Management, NTU**

Director, Intelligent Financial Innovation Technology, IFIT Lab, IM, NTPU

Associate Director, Fintech and Green Finance Center, NTPU

Publications Co-Chairs, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013-)

Program Co-Chair, IEEE International Workshop on Empirical Methods for Recognizing Inference in TExt (IEEE EM-RITE 2012-)

Publications Chair, The IEEE International Conference on Information Reuse and Integration for Data Science (IEEE IRI 2007-)



國立臺北大學
National Taipei University



Outline

- ChatGPT and Conversational AI
 - Generative Pre-trained Transformer (GPT)
 - Large Language Model (LLM)
- Reinforcement Learning from Human Feedback (RLHF)
- Generative AI
- AI for Social Good (AI4SG)

OpenAI ChatGPT

[API](#)[RESEARCH](#)[BLOG](#)[ABOUT](#)

ChatGPT: Optimizing Language Models for Dialogue

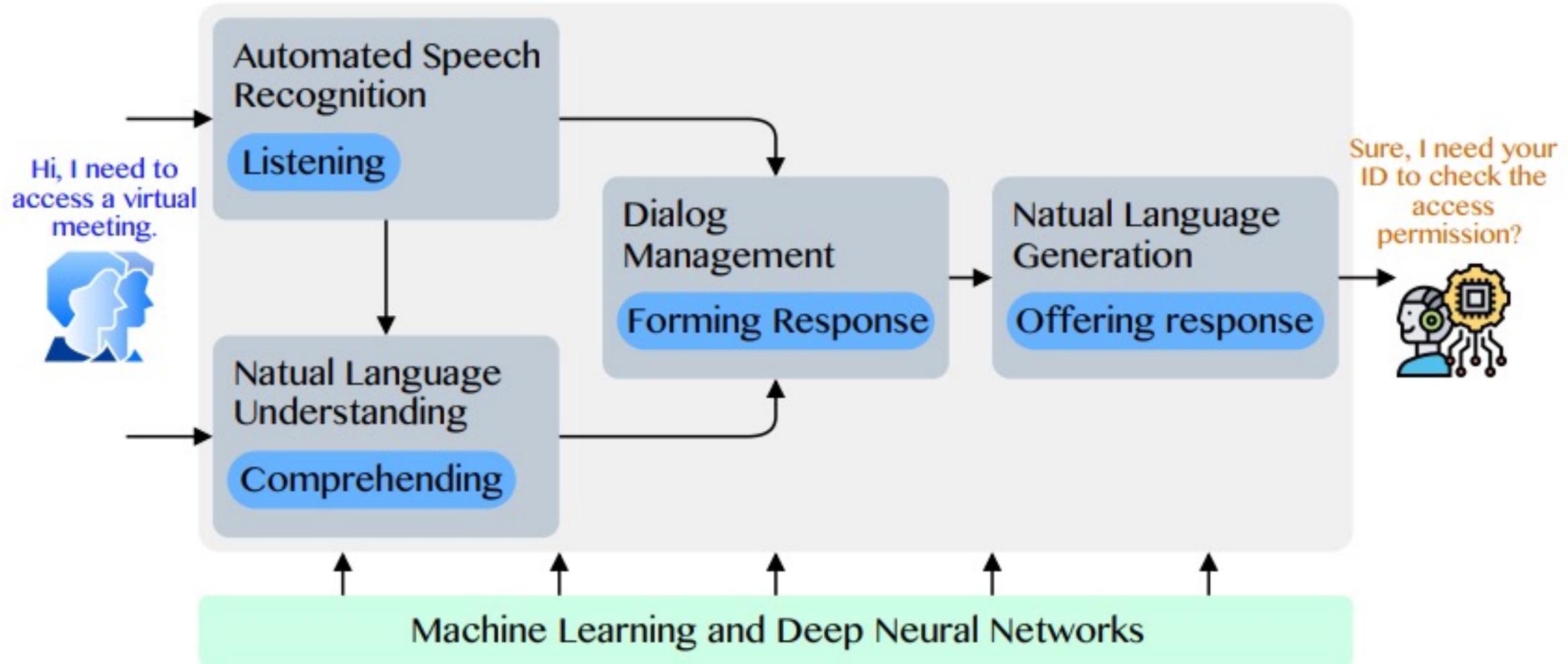
We've trained a model called ChatGPT which interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer followup questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests. ChatGPT is a sibling model to InstructGPT, which is trained to follow an instruction in a prompt and provide a detailed response.



Source: <https://openai.com/blog/chatgpt/>

Conversational AI

to deliver contextual and personal experience to users



Source: Huynh-The, Thien, Quoc-Viet Pham, Xuan-Qui Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim (2022).

"Artificial Intelligence for the Metaverse: A Survey." arXiv preprint arXiv:2202.10336.

OpenAI ChatGPT

ChatGPT

This is a free research preview.



Our goal is to get external feedback in order to improve our systems and make them safer.



While we have safeguards in place, the system may occasionally generate incorrect or misleading information and produce offensive or biased content. It is not intended to give advice.

OpenAI ChatGPT

ChatGPT

How we collect data



Conversations may be reviewed by our AI trainers to improve our systems.



Please don't share any sensitive information in your conversations.

OpenAI ChatGPT

ChatGPT



Examples

"Explain quantum computing in simple terms" →



Capabilities

Remembers what user said earlier in the conversation



Limitations

May occasionally generate incorrect information

"Got any creative ideas for a 10 year old's birthday?" →

Allows user to provide follow-up corrections

May occasionally produce harmful instructions or biased content

"How do I make an HTTP request in Javascript?" →

Trained to decline inappropriate requests

Limited knowledge of world and events after 2021



Free Research Preview: ChatGPT is optimized for dialogue. Our goal is to make AI systems more natural to interact with, and your feedback will help us improve our systems and make them safer.

Source: <https://chat.openai.com/chat>

OpenAI ChatGPT

IM

What is the future of AI?

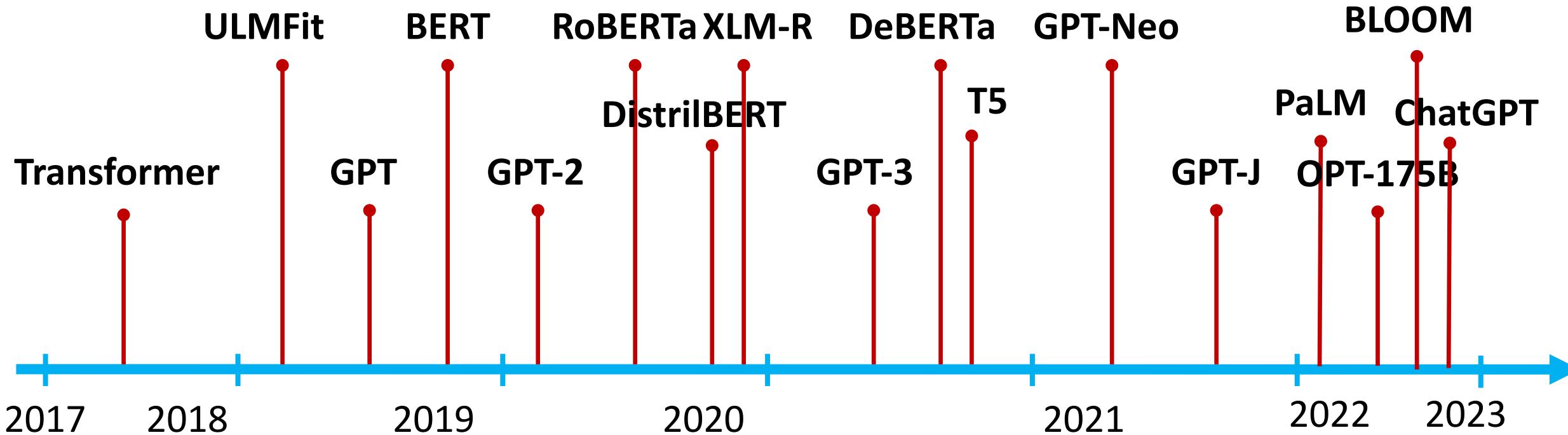


It is difficult to predict the exact future of AI, but many experts believe that it will continue to play an increasingly important role in our lives. Some potential future developments in the field of AI include:

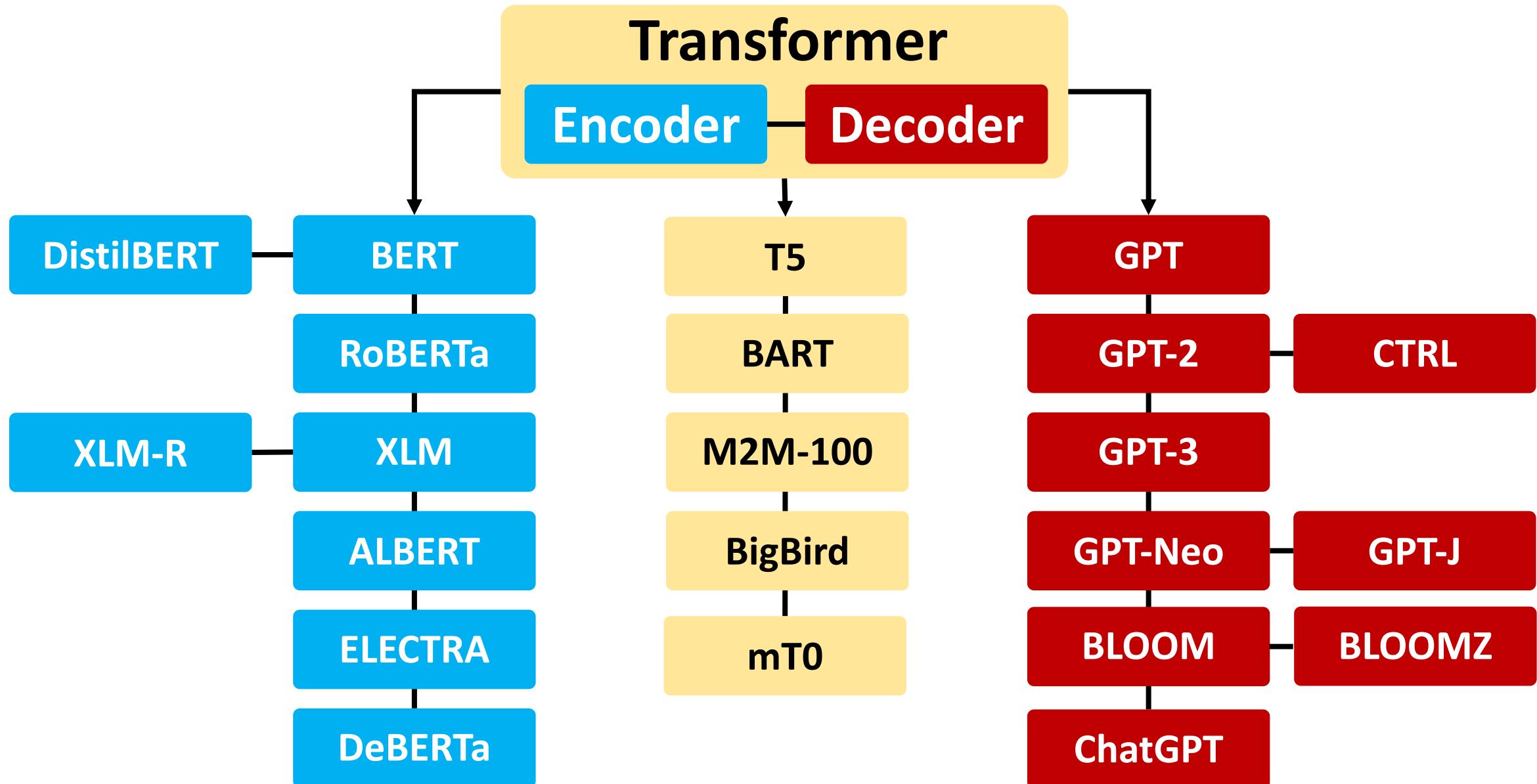
- Improved natural language processing, allowing AI systems to better understand and respond to human speech and writing.
- Enhanced ability to learn and adapt, allowing AI systems to become more autonomous and capable of handling a wider range of tasks.
- Greater integration with the internet of things (IoT), allowing AI systems to control and interact with a wide range of connected devices in the home and workplace.
- Continued advances in machine learning, allowing AI systems to make more accurate predictions and decisions based on data.
- Increased use of AI in healthcare, transportation, and other industries, leading to more efficient and effective processes and services.
- Continued development of ethical and regulatory frameworks to ensure that AI is developed and used responsibly.

Ultimately, the future of AI will depend on how it is developed and used by individuals, organizations, and society as a whole.

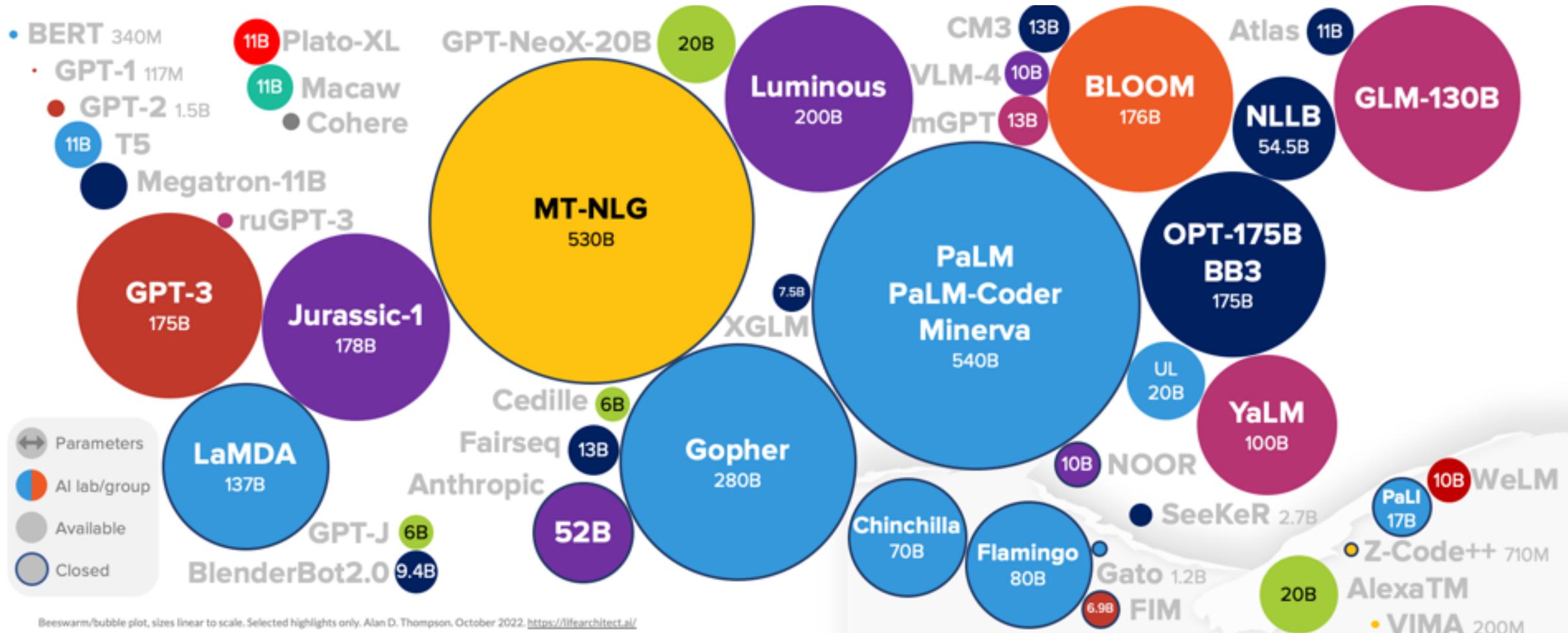
The Transformers Timeline



Transformer Models



Language Models Sizes (GPT-3, PaLM, BLOOM)



ChatGPT: Optimizing Language Models for Dialogue

Step 1

Collect demonstration data and train a supervised policy.

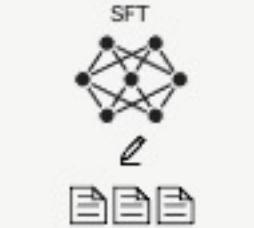
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



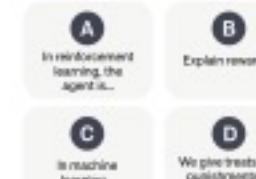
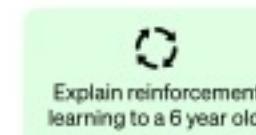
This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

Collect comparison data and train a reward model.

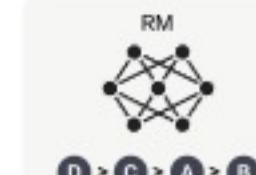
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.



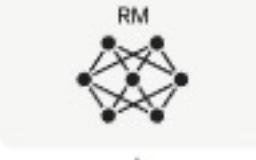
The PPO model is initialized from the supervised policy.



The policy generates an output.

Once upon a time...

The reward model calculates a reward for the output.



The reward is used to update the policy using PPO.

r_k

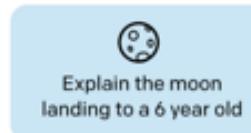
Training language models to follow instructions with human feedback

InstructGPT and GPT 3.5

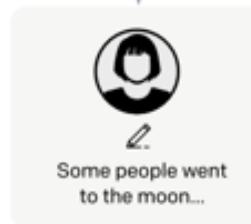
Step 1

Collect demonstration data, and train a supervised policy.

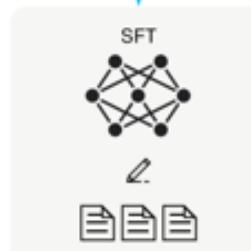
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



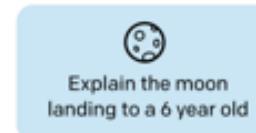
This data is used to fine-tune GPT-3 with supervised learning.



Step 2

Collect comparison data, and train a reward model.

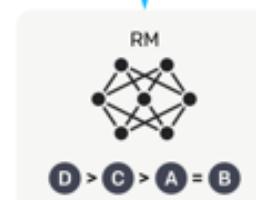
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.



This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using reinforcement learning.

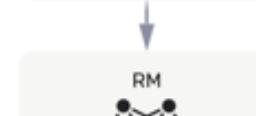
A new prompt is sampled from the dataset.



The policy generates an output.



Once upon a time...



The reward model calculates a reward for the output.



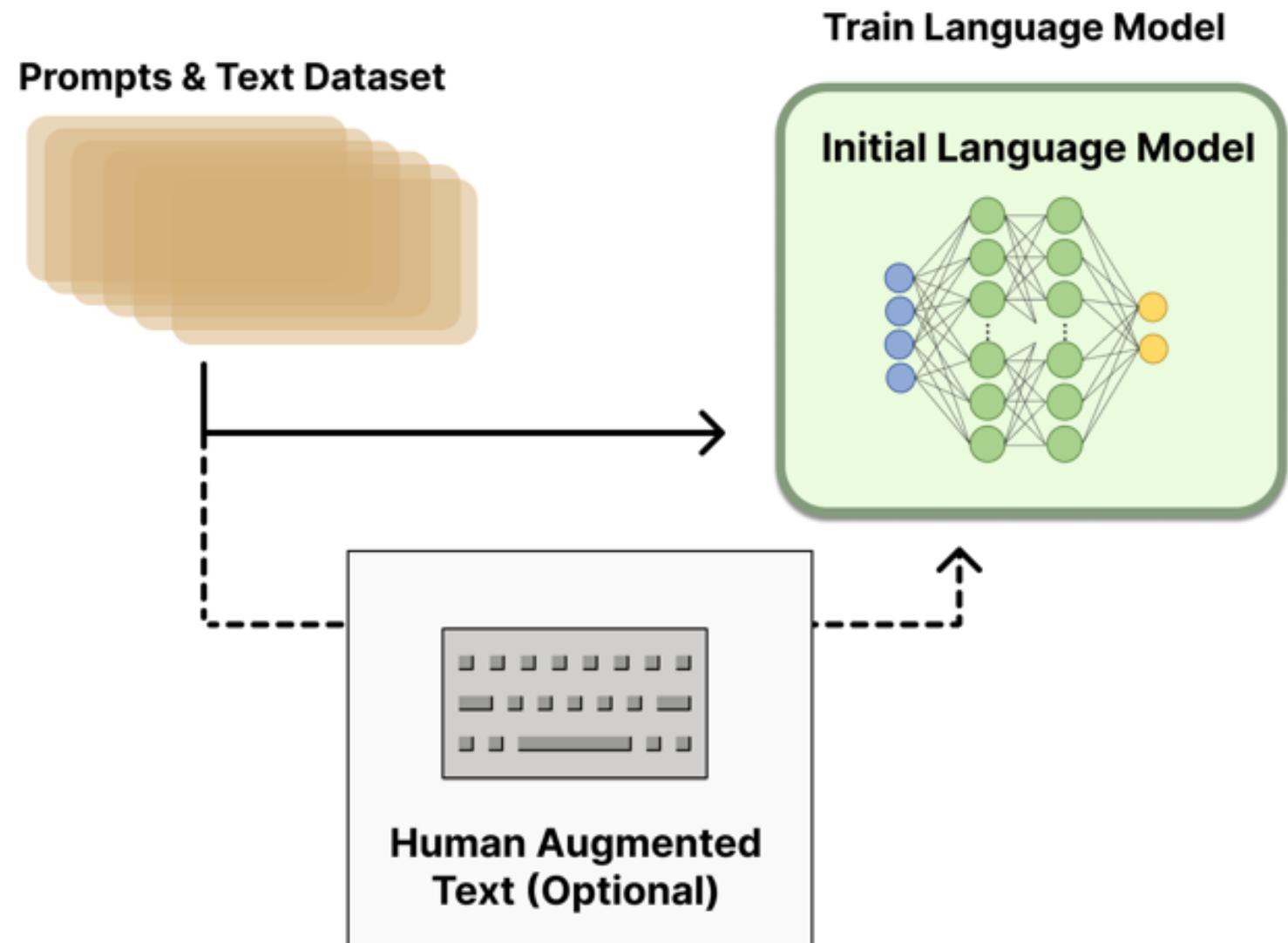
The reward is used to update the policy using PPO.

Reinforcement Learning from Human Feedback (RLHF)

- 1. Pretraining a Language Model (LM)**
- 2. Gathering Data and Training a Reward Model**
- 3. Fine-tuning the LM with Reinforcement Learning**

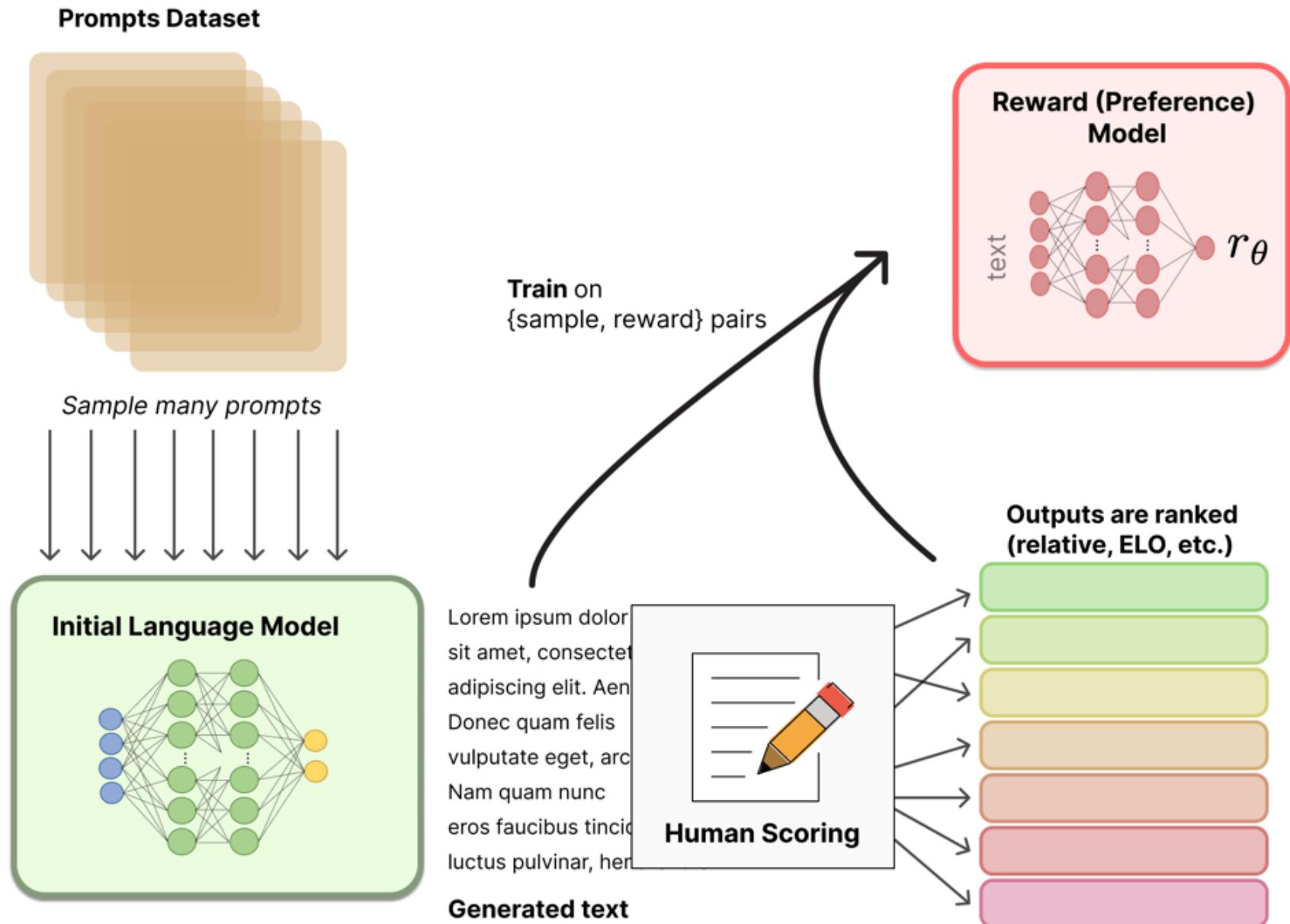
Reinforcement Learning from Human Feedback (RLHF)

**Step 1. Pretraining
a Language Model
(LM)**



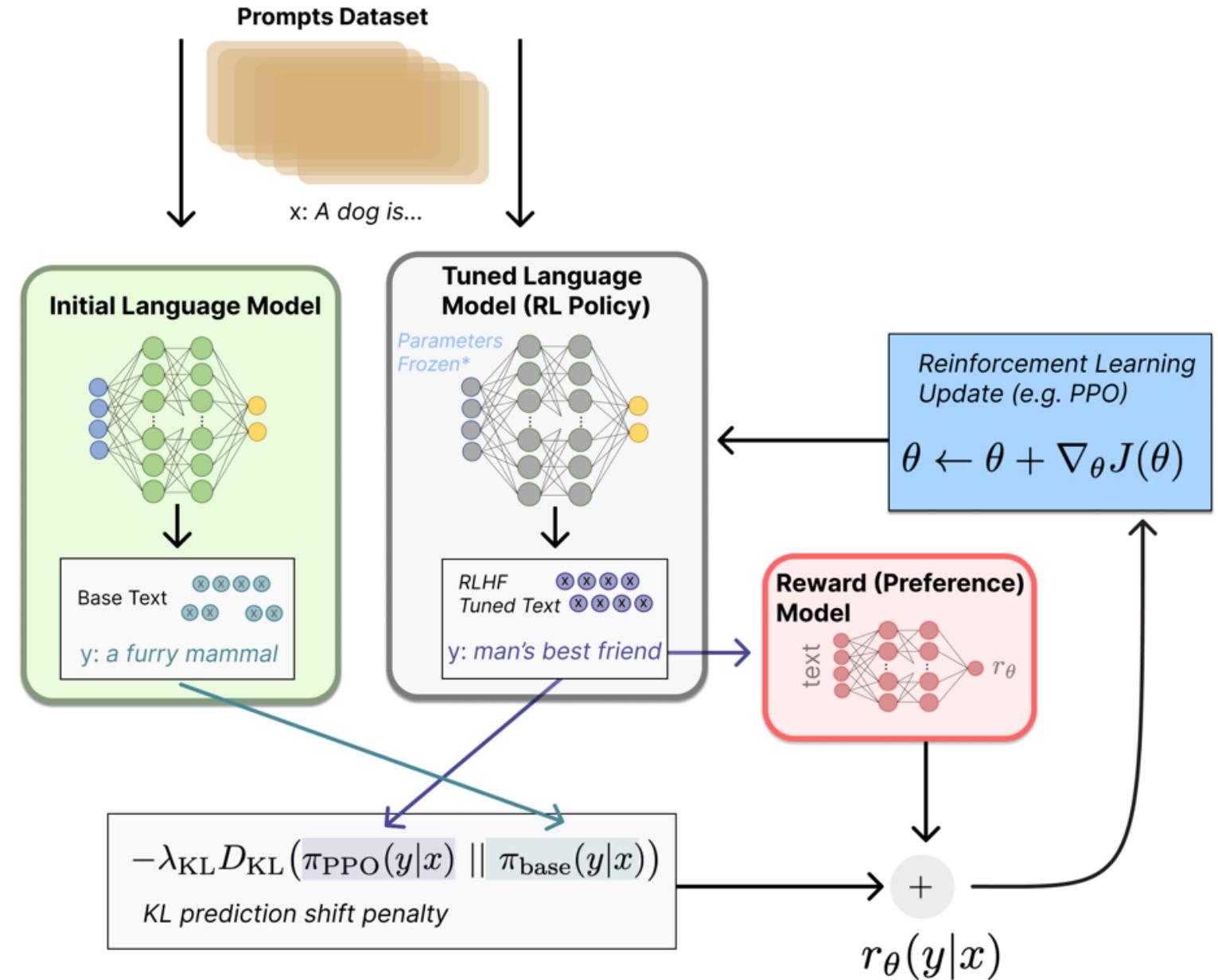
Reinforcement Learning from Human Feedback (RLHF)

Step 2. Gathering Data and Training a Reward Model



Reinforcement Learning from Human Feedback (RLHF)

Step 3. Fine-tuning the LM with Reinforcement Learning



4 Approaches of AI

<p>2. Thinking Humanly: The Cognitive Modeling Approach</p>	<p>3. Thinking Rationally: The “Laws of Thought” Approach</p>
<p>1. Acting Humanly: The Turing Test Approach (1950)</p>	<p>4. Acting Rationally: The Rational Agent Approach</p>

AI Acting Humanly: The Turing Test Approach

(Alan Turing, 1950)

- Knowledge Representation
- Automated Reasoning
- Machine Learning (ML)
 - Deep Learning (DL)
- Computer Vision (Image, Video)
- Natural Language Processing (NLP)
- Robotics

Can machines think?

- Alan Turing rejected the question “Can machines think?” and replaced it with a behavioral test.
 - Alan Turing anticipated many objections to the possibility of thinking machines.
 - Concentrate on their systems’ performance on practical tasks
 - rather than the ability to imitate humans.
 - Consciousness remains a mystery.

Generative AI

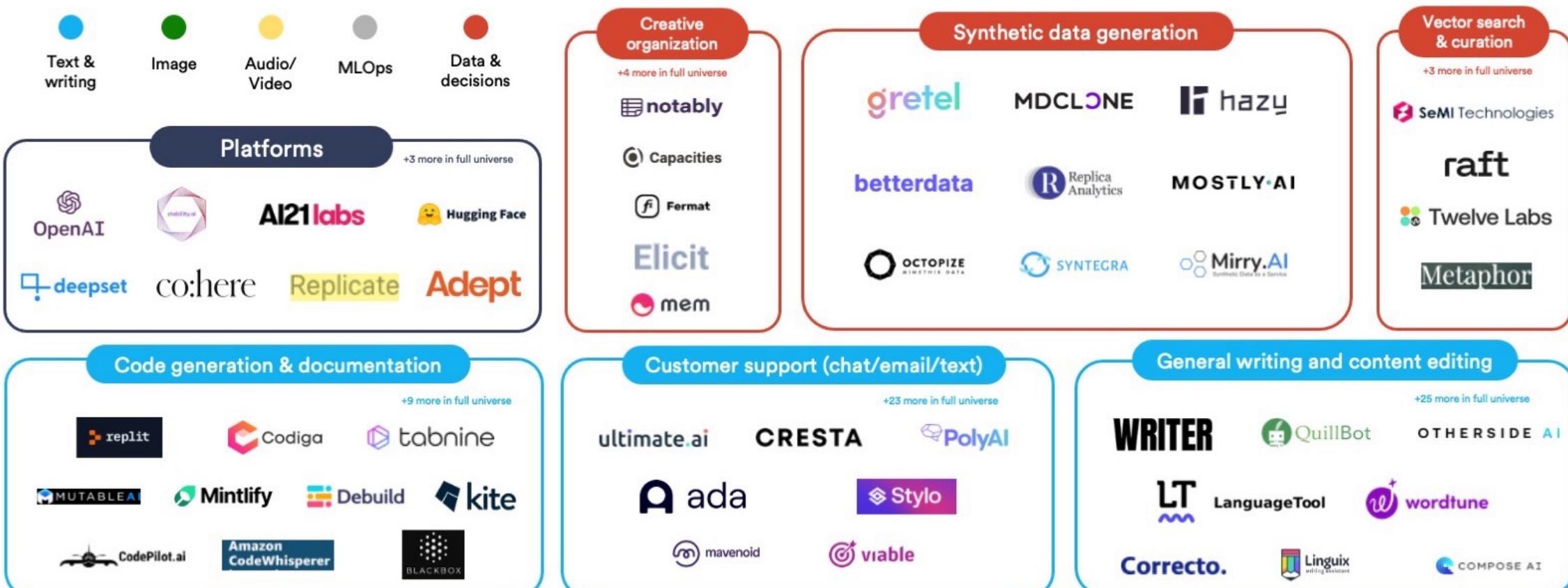
Gen AI

Generative AI

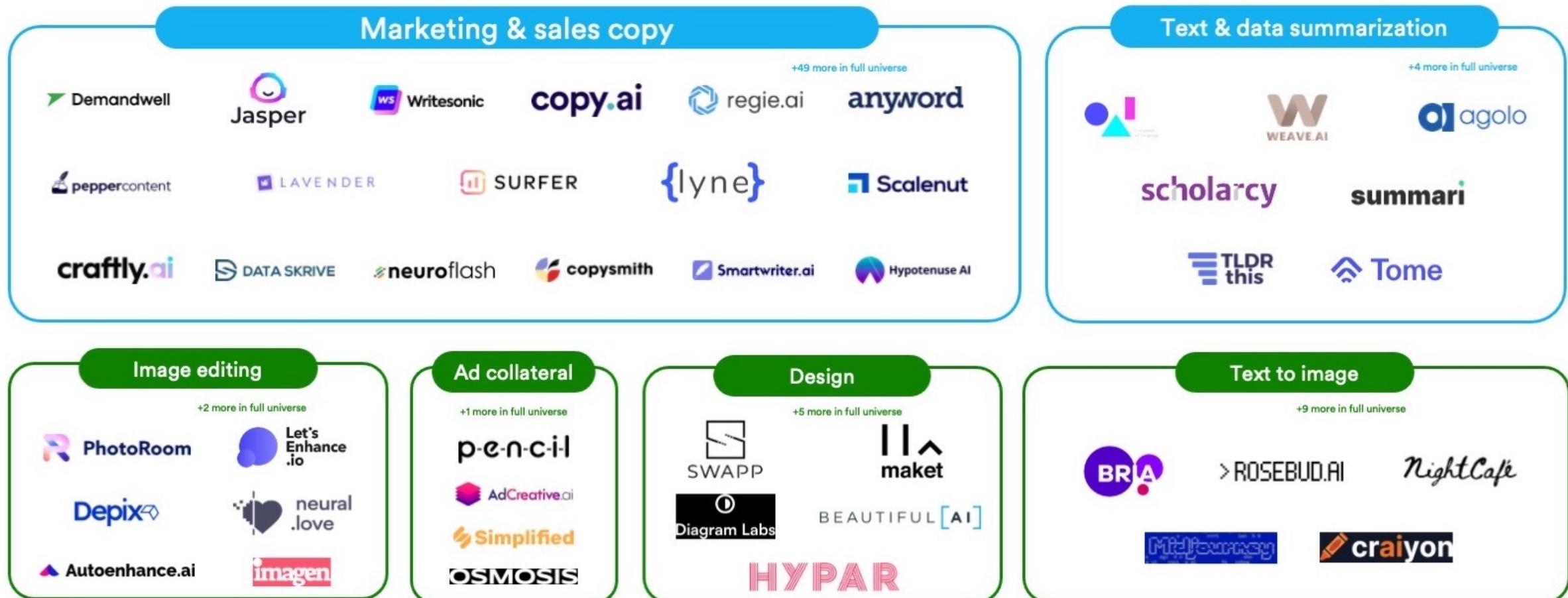
BASE10 TREND MAP: GENERATIVE AI

Base10

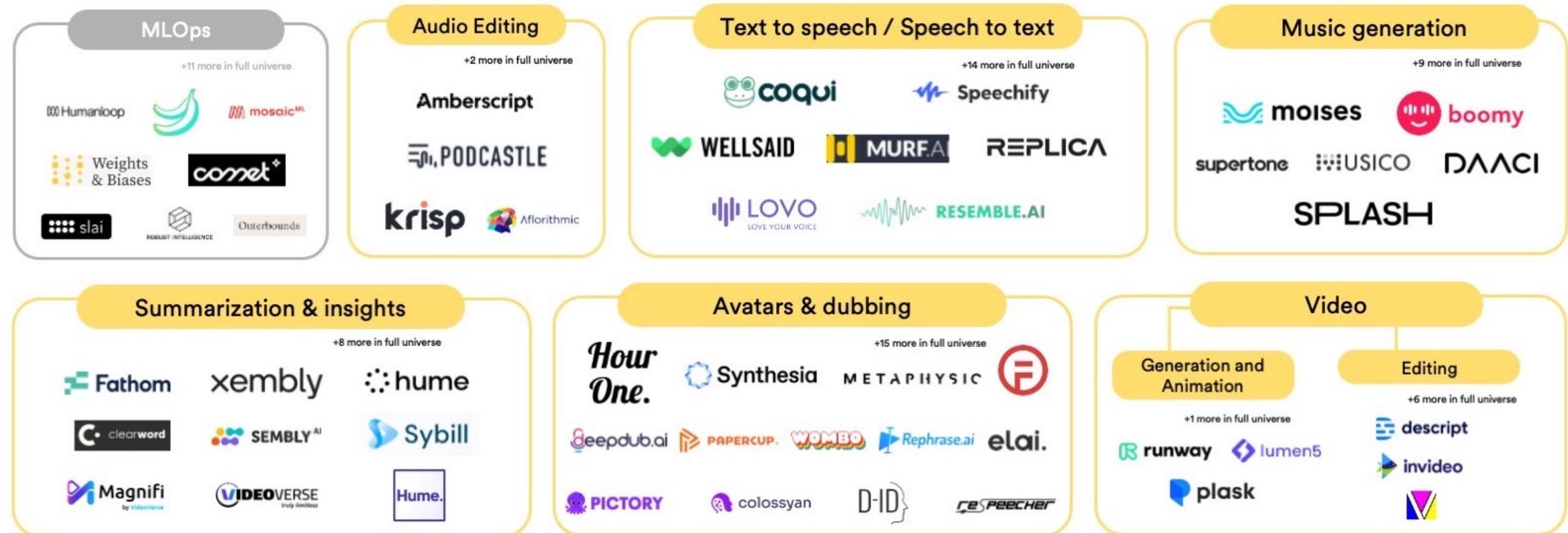
Companies are grouped based on medium produced and segmented by use case within each medium. Companies that offer products across segments are grouped in the segment of the core product offering.



Generative AI



Generative AI



DALL·E 2

Create original, realistic images and art from a text description.
It can combine concepts, attributes, and styles.

TEXT DESCRIPTION

An astronaut Teddy bears A
bowl of soup

riding a horse lounging in a
tropical resort in space playing
basketball with cats in space

in a photorealistic style in the
style of Andy Warhol as a pencil
drawing

DALL·E 2



Stable Diffusion

Hugging Face Search models, datasets, users...

Models Datasets Spaces Docs Solutions Pricing

Spaces: stabilityai/stable-diffusion like 1.89k Running

App Files Community 241 Linked Models

Stable Diffusion Demo

Stable Diffusion is a state of the art text-to-image model that generates images from text.

For faster generation and forthcoming API access you can try [DreamStudio Beta](#)

an insect robot preparing a delicious meal

Generate image



<https://huggingface.co/spaces/stabilityai/stable-diffusion>

Stable Diffusion Colab

wocitezuma / stable-diffusion-colab Public

Notifications Fork 7 Star 31

<> Code Issues Pull requests Actions Projects Wiki Security Insights

main ▾ 1 branch 0 tags Go to file Code ▾

woctezuma README: add a reference for sampler schedules	37bc02d 24 days ago	18 commits
LICENSE	Initial commit	27 days ago
README.md	README: add a reference for sampler schedules	24 days ago
stable_diffusion.ipynb	Allow to choose the scheduler	25 days ago

About
Colab notebook to run Stable Diffusion.
github.com/CompVis/stable-diffusion

deep-learning colab image-generation
text-to-image diffusion text2image
colaboratory google-colab
colab-notebook google-colaboratory
google-colab-notebook
text-to-image-synthesis huggingface
diffusion-models
text-to-image-generation latent-diffusion
stable-diffusion huggingface-diffusers
diffusers stable-diffusion-diffusers

☰ README.md

Stable-Diffusion-Colab

The goal of this repository is to provide a Colab notebook to run the text-to-image "Stable Diffusion" model [1].

Usage

- Run [stable_diffusion.ipynb](#) 

Readme MIT license 31 stars 2 watching

<https://github.com/wocitezuma/stable-diffusion-colab>

Lexica Art: Search Stable Diffusion images and prompts

Lexica

Search images

Search

Lexica

The Stable Diffusion search engine

Join the Discord

Search for an image

Columns: 10

https://lexica.art/

AnyFace: Free-style Text-to-Face Synthesis and Manipulation

AnyFace: Free-style Text-to-Face Synthesis and Manipulation

Jianxin Sun^{1,2*}, Qiyao Deng^{1,2*}, Qi Li^{1,2 †}, Muyi Sun¹, Min Ren^{1,2}, Zhenan Sun^{1,2}

¹ Center for Research on Intelligent Perception and Computing, NLPR, CASIA

² School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS)

{jianxin.sun, dengqiyao, muyi.sun, min.ren}@cripac.ia.ac.cn, {qli, znsun}@nlpr.ia.ac.cn



- (1) This is a young man with a melon seed face.
- (2) He has wheat skin, big eyes and slightly bushy eyebrows.
- (3) He has medium-length black hair.
- (4) The man is smiling with his mouth slightly open.
- (5) He wears black-rimmed glasses and no beard

Source



(1)



(1) - (2)



(1) - (3)



(1) - (4)



(1) - (5)

(a) One caption vs Multi-caption



She graduated with a PhD.



He looks very knowledgeable.

(b) Open-world text descriptions



He has black hair and beard.



Source

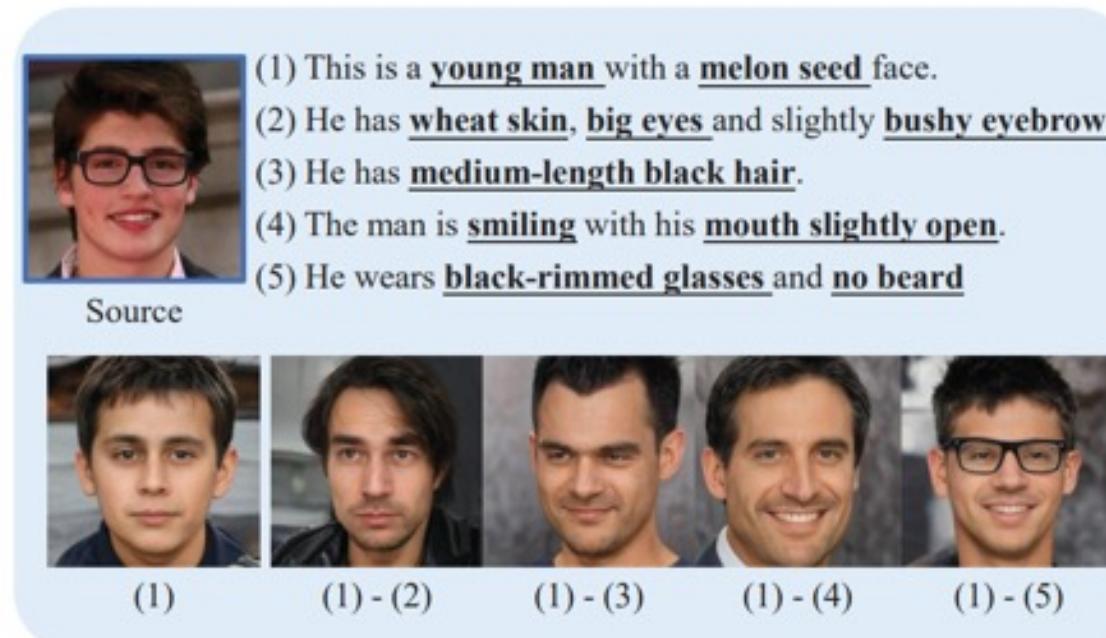
(c) Face manipulation

Figure 1. Our AnyFace framework can be used for real-life applications. (a) Face image synthesis with optical captions. The top left is the source face. (b) Open-world face synthesis with out-of-dataset descriptions. (c) Text-guided face manipulation with continuous control. Given source images, AnyFace can manipulate faces with continuous changes. The arrow indicates the increasing relevance to the text.

Source: Sun, Jianxin, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. (2022)

"AnyFace: Free-style Text-to-Face Synthesis and Manipulation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18687-18696.

AnyFace: Free-style Text-to-Face Synthesis and Manipulation



(a) One caption vs Multi-caption

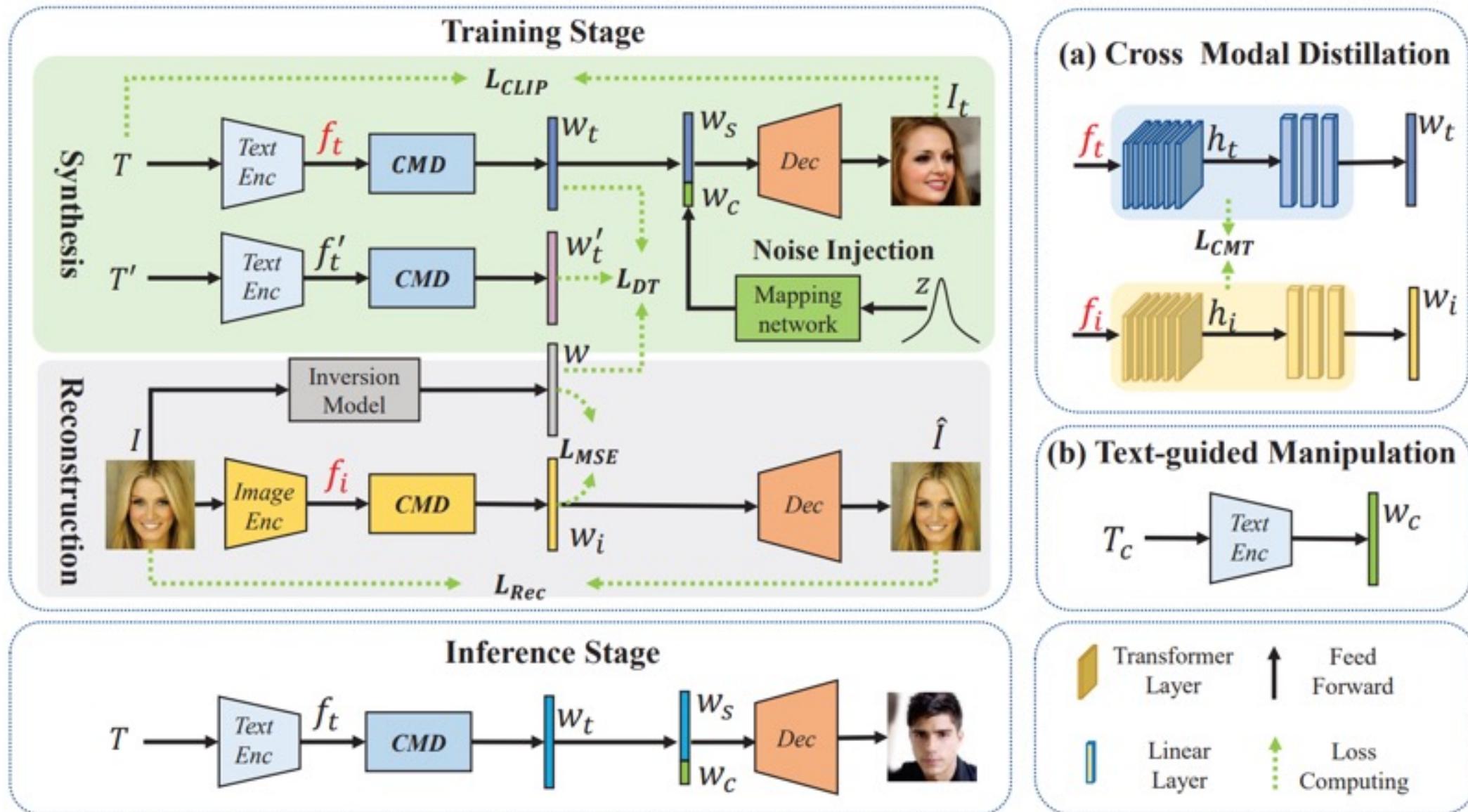


Methods	AttnGAN [31]	DFGAN [25]	RiFeGAN [1]	SEA-T2F [24]	CIGAN [28]	TediGAN-B [30]	AnyFace
Single Model	✓	✓	✓	✓	✓	-	✓
One Generator	-	✓	-	-	✓	✓	✓
Multi-caption	-	-	✓	✓	-	-	✓
High Resolution	-	-	-	-	✓	✓	✓
Manipulation	-	-	-	-	✓	✓	✓
Open-world	-	-	-	-	-	✓	✓

Source: Sun, Jianxin, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. (2022)

"AnyFace: Free-style Text-to-Face Synthesis and Manipulation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18687-18696.

AnyFace: Free-style Text-to-Face Synthesis and Manipulation



Source: Sun, Jianxin, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. (2022)

"AnyFace: Free-style Text-to-Face Synthesis and Manipulation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18687-18696.

AnyFace: Free-style Text-to-Face Synthesis and Manipulation

The person wears lipstick.
She has blond hair, and
pale skin. She is attractive.



The woman has wavy hair,
black hair, and arched eyebrows. She is young. She is wearing heavy makeup.



She is wearing lipstick. She has high cheekbones, wavy hair, bushy eyebrows, and oval face. She is attractive.



He has mouth slightly open,
wavy hair, bushy eyebrows, and oval face. He is attractive, and young. He has no beard.



AttnGAN

SEA-T2F

TediGAN-B

Ours w/o L_{DT}

Ours w/o L_{CMT}

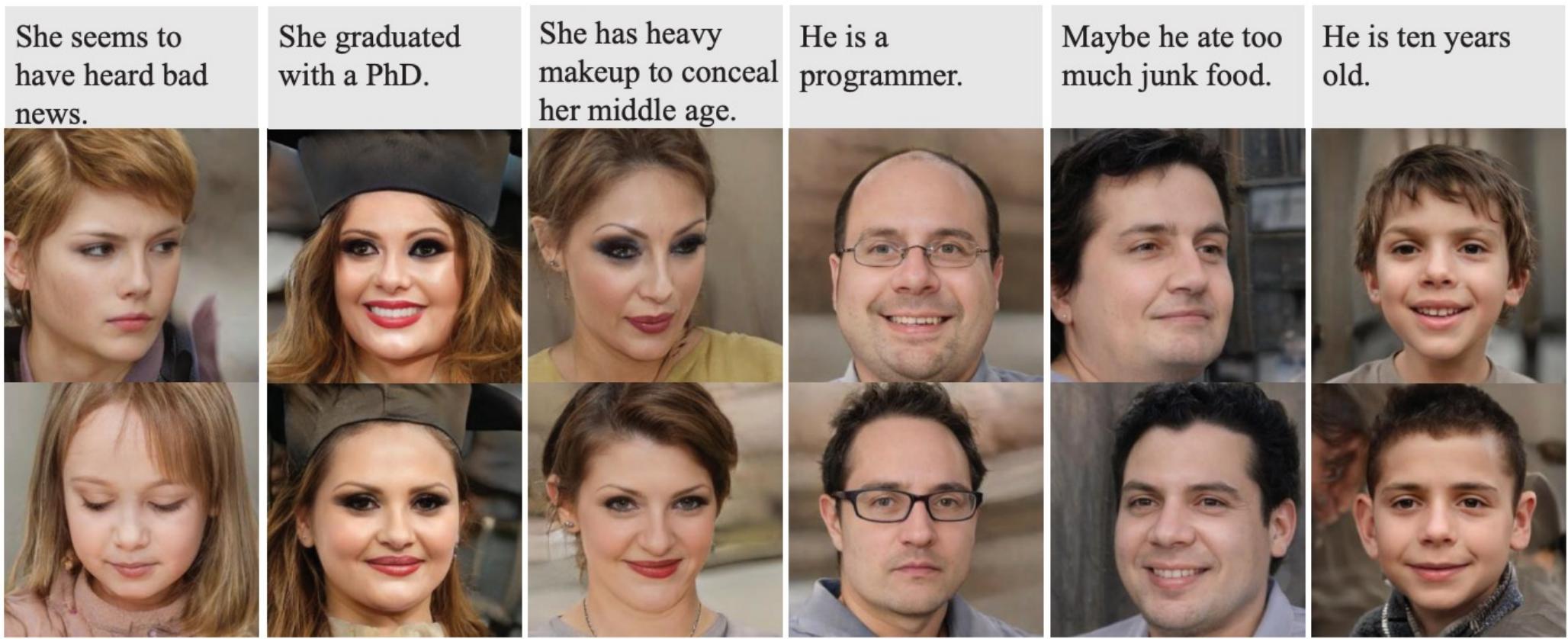
Ours

Source: Sun, Jianxin, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. (2022)

"AnyFace: Free-style Text-to-Face Synthesis and Manipulation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18687-18696.

AnyFace: Free-style Text-to-Face Synthesis and Manipulation

AnyFace



TediGAN-B



Source: Sun, Jianxin, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. (2022)

"AnyFace: Free-style Text-to-Face Synthesis and Manipulation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18687-18696.

AnyFace: Free-style Text-to-Face Synthesis and Manipulation

Text-guided Face Manipulation



Source

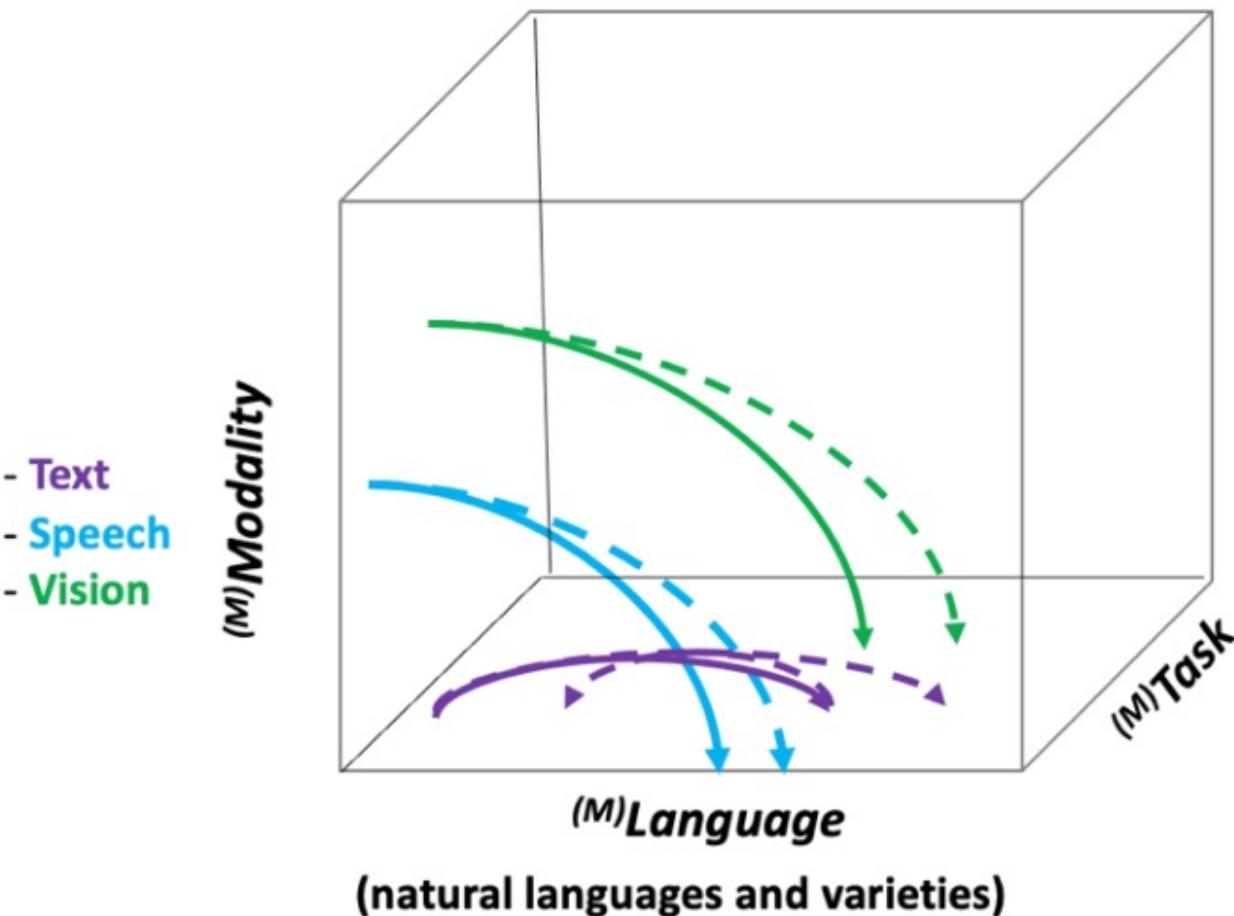


Source: Sun, Jianxin, Qiyao Deng, Qi Li, Muyi Sun, Min Ren, and Zhenan Sun. (2022)

"AnyFace: Free-style Text-to-Face Synthesis and Manipulation." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18687-18696.

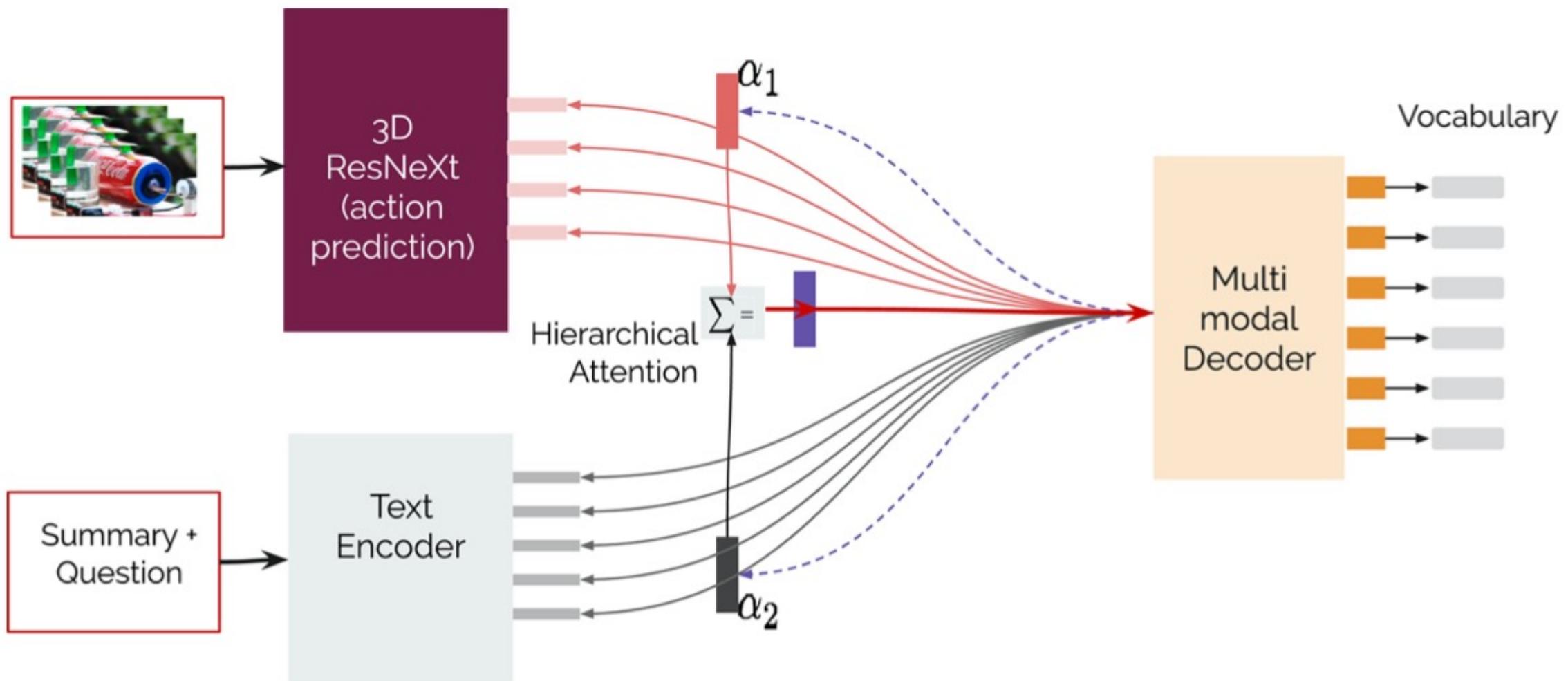
NLG from a Multilingual, Multimodal and Multi-task perspective

Multi³(Natural Language) Generation



- Recognize and transcribe speech (ASR)
- Translate from one language to another (MT)
- Describe, ask or answer questions or converse about visual objects (Captioning, VQA, Visual Dialogue, ...)

Text-and-Video Dialog Generation Models with Hierarchical Attention



Source: Erdem, Erkut, Menekse Kuyu, Semih Yagcioglu, Anette Frank, Letitia Parcalabescu, Barbara Plank, Andrii Babii et al.

"Neural Natural Language Generation: A Survey on Multilinguality, Multimodality, Controllability and Learning." Journal of Artificial Intelligence Research 73 (2022): 1131-1207.

Multimodal Few-Shot Learning with Frozen Language Models

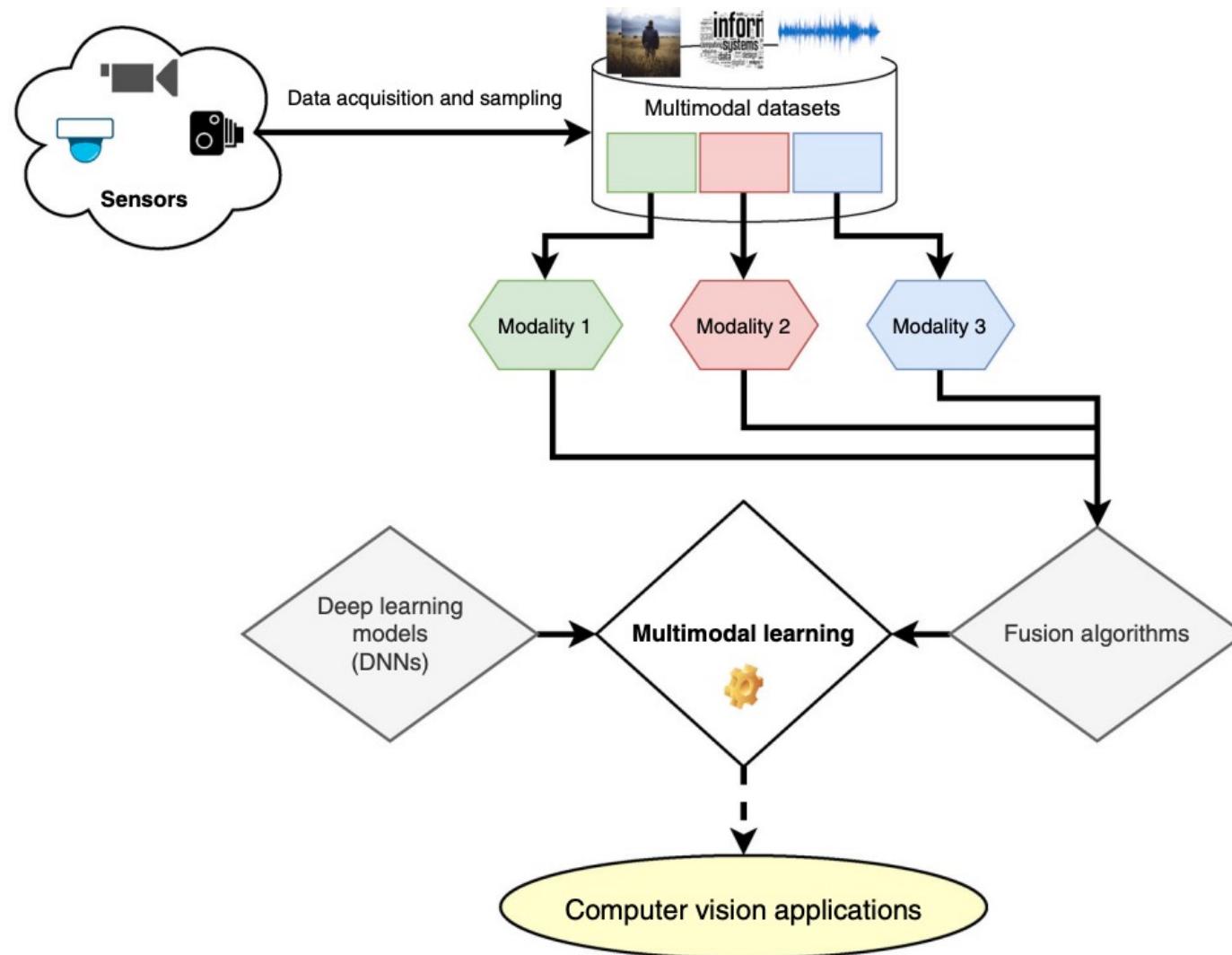
	This person is like 😊.		This person is like 😞.		This person is like	Model Completion 选拨. <EOS>
	This was invented by Zacharias Janssen.		This was invented by Thomas Edison.		This was invented by	Model Completion the Wright brothers. <EOS>
	With one of these I can drive around a track, overtaking other cars and taking corners at speed		With one of these I can take off from a city and fly across the sky to somewhere on the other side of the world		With one of these I can	Model Completion break into a secure building, unlock the door and walk right in <EOS>

Curated samples with about five seeds required to get past well-known language model failure modes of either repeating text for the prompt or emitting text that does not pertain to the image.

These samples demonstrate the ability to generate open-ended outputs that adapt to both images and text, and to make use of facts that it has learned during language-only pre-training.

Multimodal Pipeline

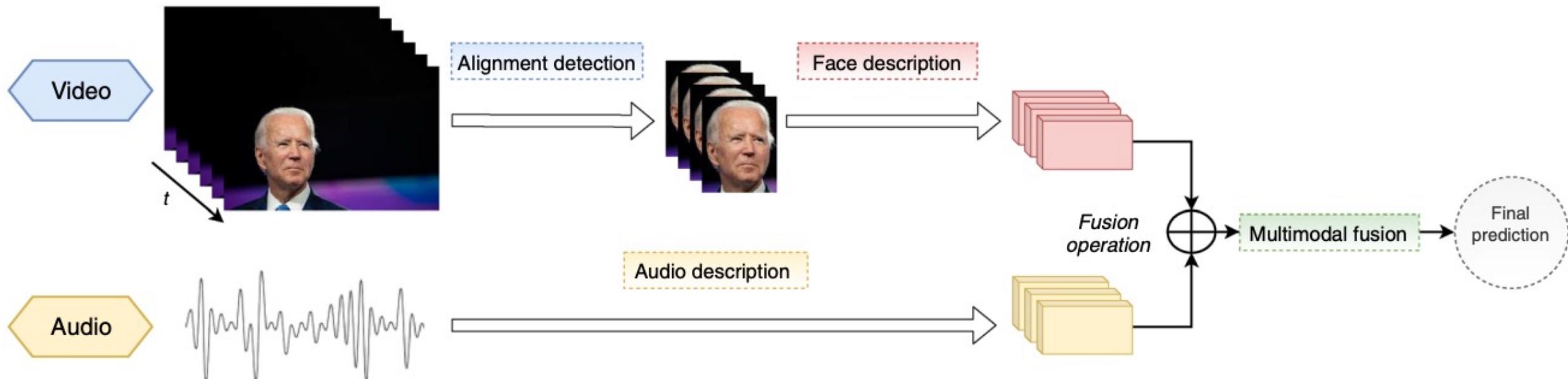
that includes three different modalities (Image, Text, Audio)



Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).

"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

Video and Audio Multimodal Fusion



Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).

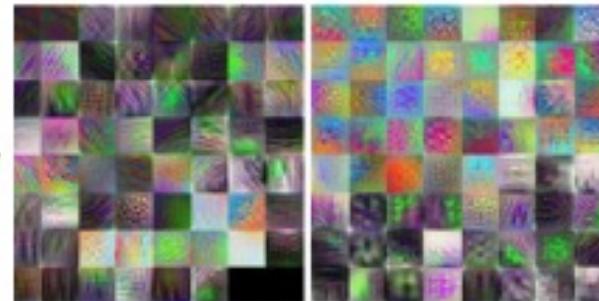
"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

Visual and Textual Representation

Image



Visual representations (Dense)



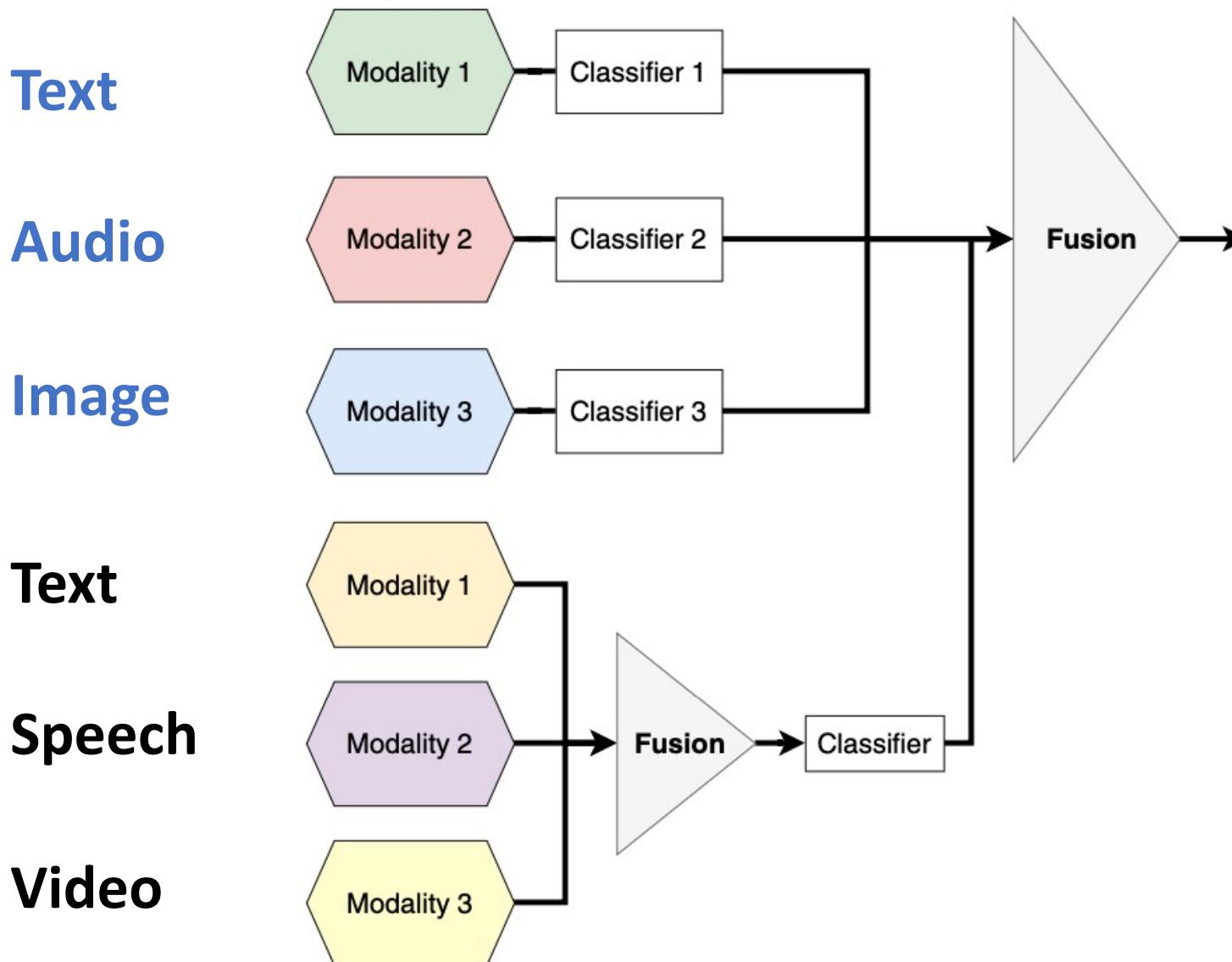
Text

This is the oldest and most important defensive work to have been built along the North African coastline by the Arab conquerors in the early days of Islam. Founded in 796, this building underwent several modifications during the medieval period. Initially, it formed a quadrilateral and then was composed of four buildings giving onto two inner courtyards.

Textual representations (Sparse)



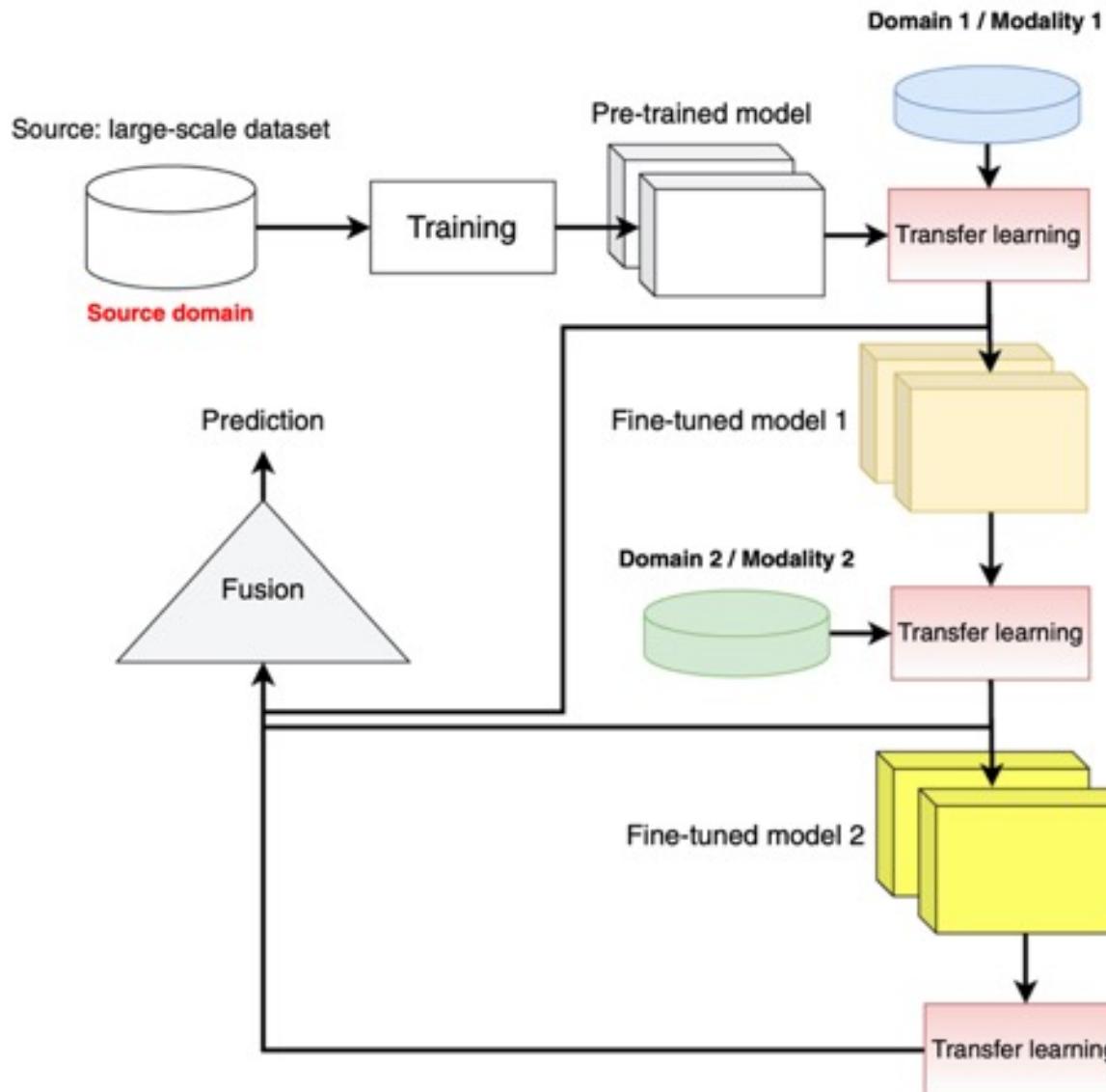
Hybrid Multimodal Data Fusion



Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).

"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

Multimodal Transfer Learning

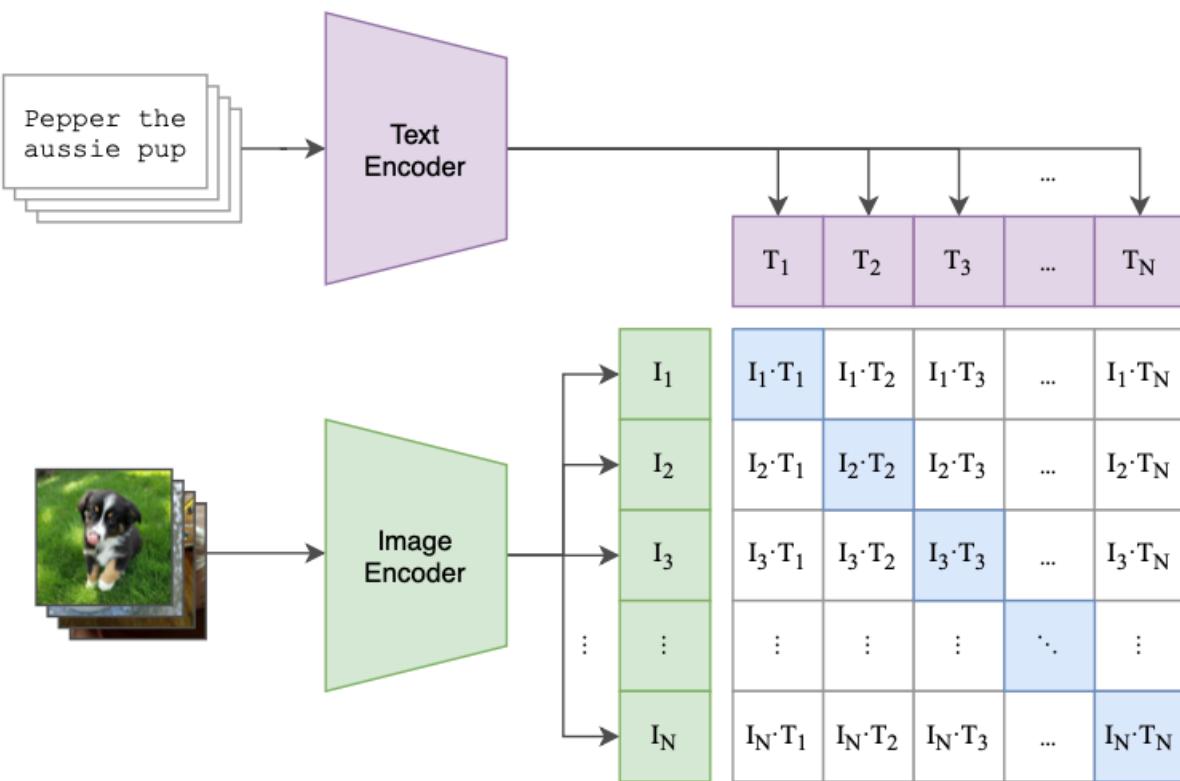


Source: Bayoudh, Khaled, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa (2022).

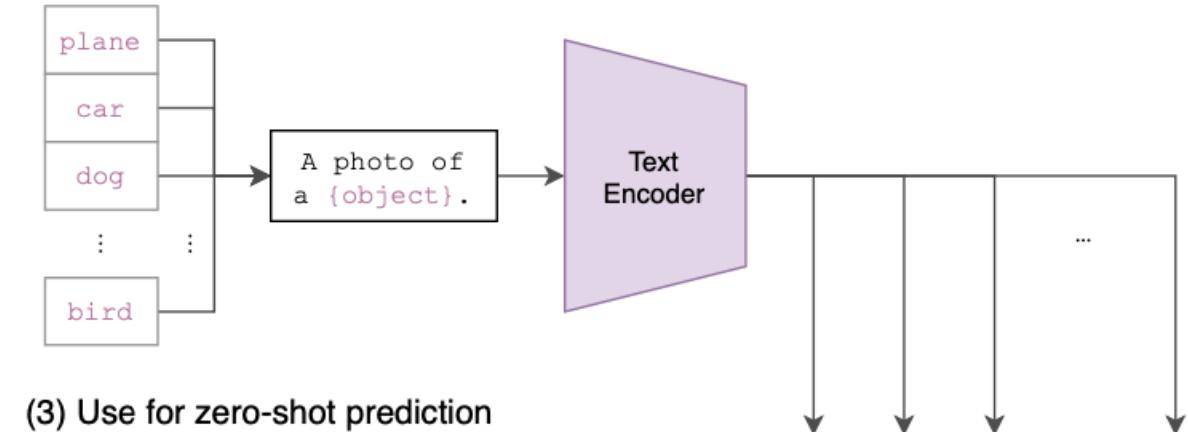
"A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets." The Visual Computer 38, no. 8: 2939-2970.

CLIP: Learning Transferable Visual Models From Natural Language Supervision

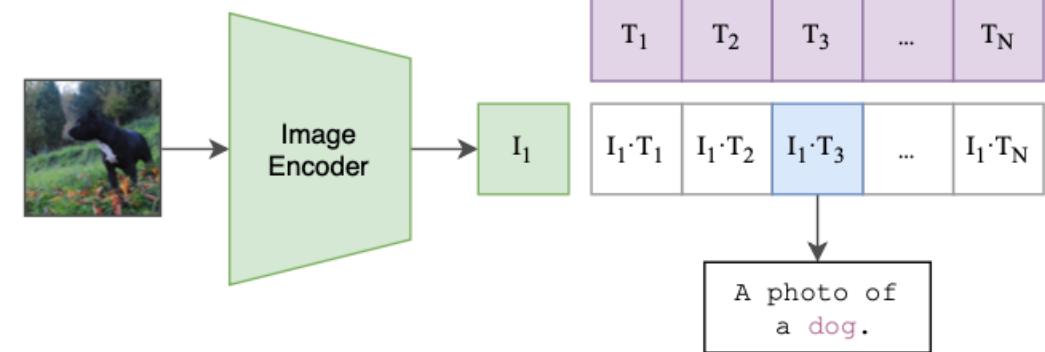
(1) Contrastive pre-training



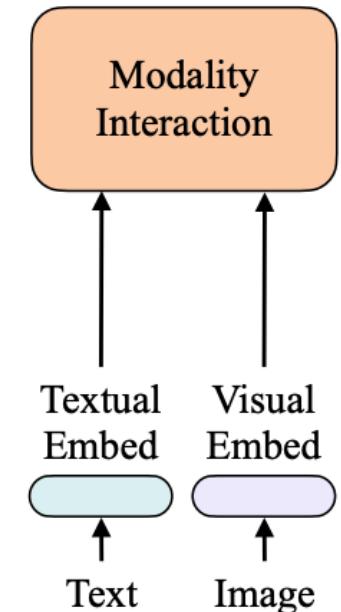
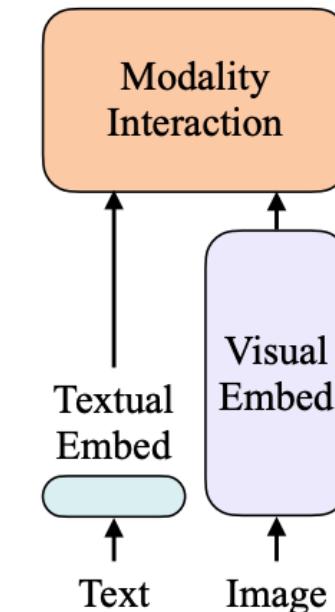
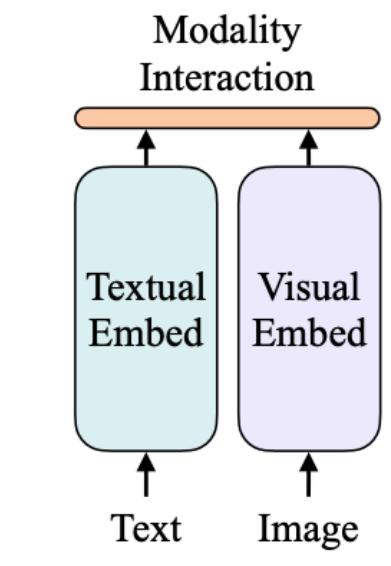
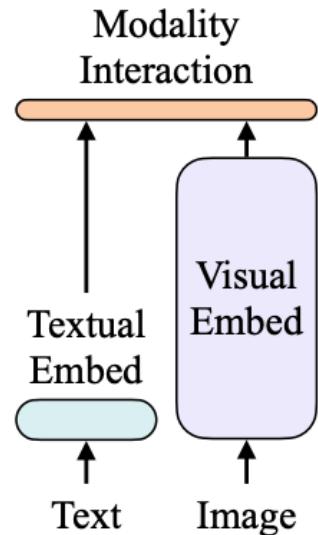
(2) Create dataset classifier from label text



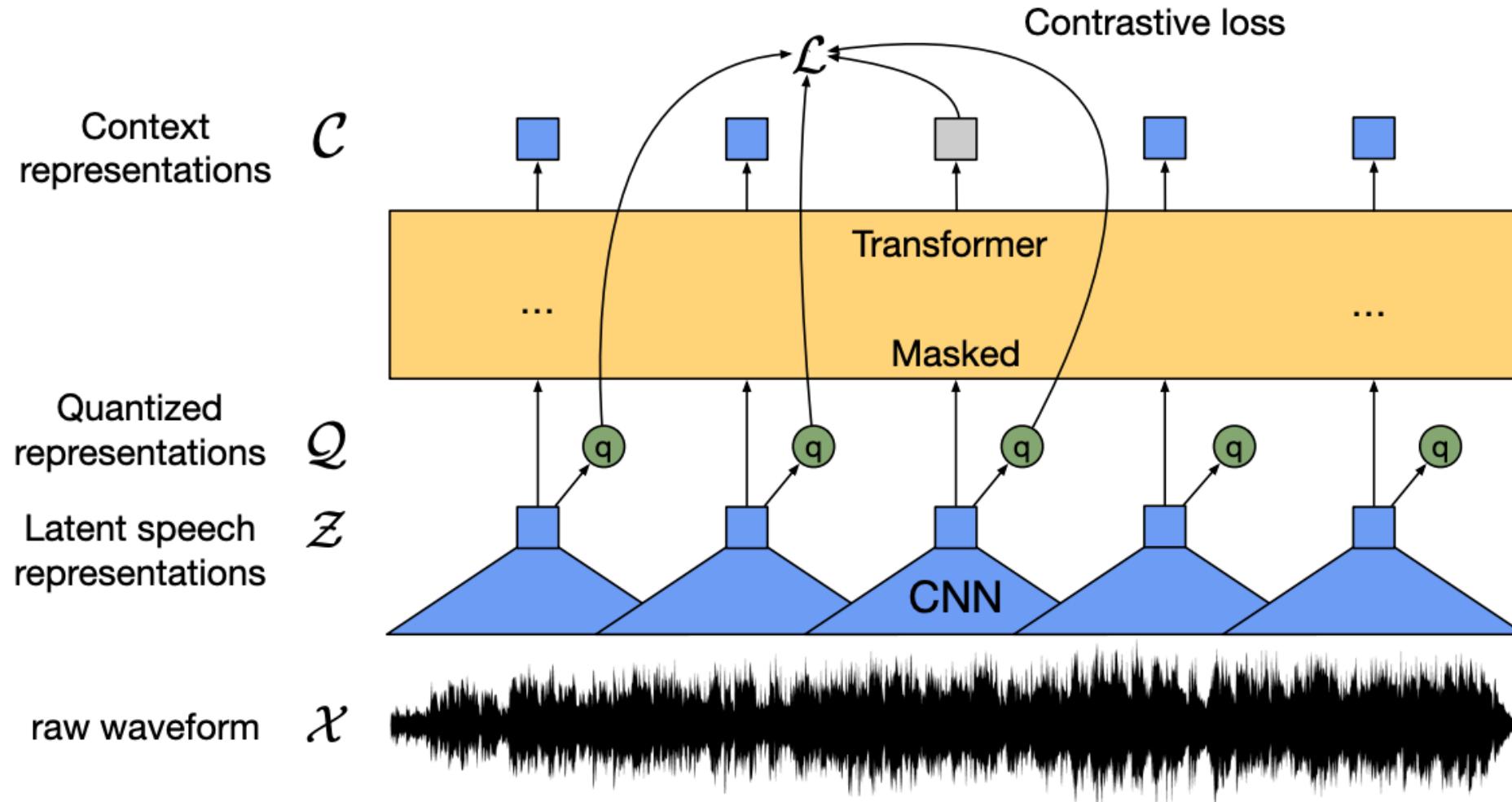
(3) Use for zero-shot prediction



ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision



wav2vec 2.0: A framework for self-supervised learning of speech representations

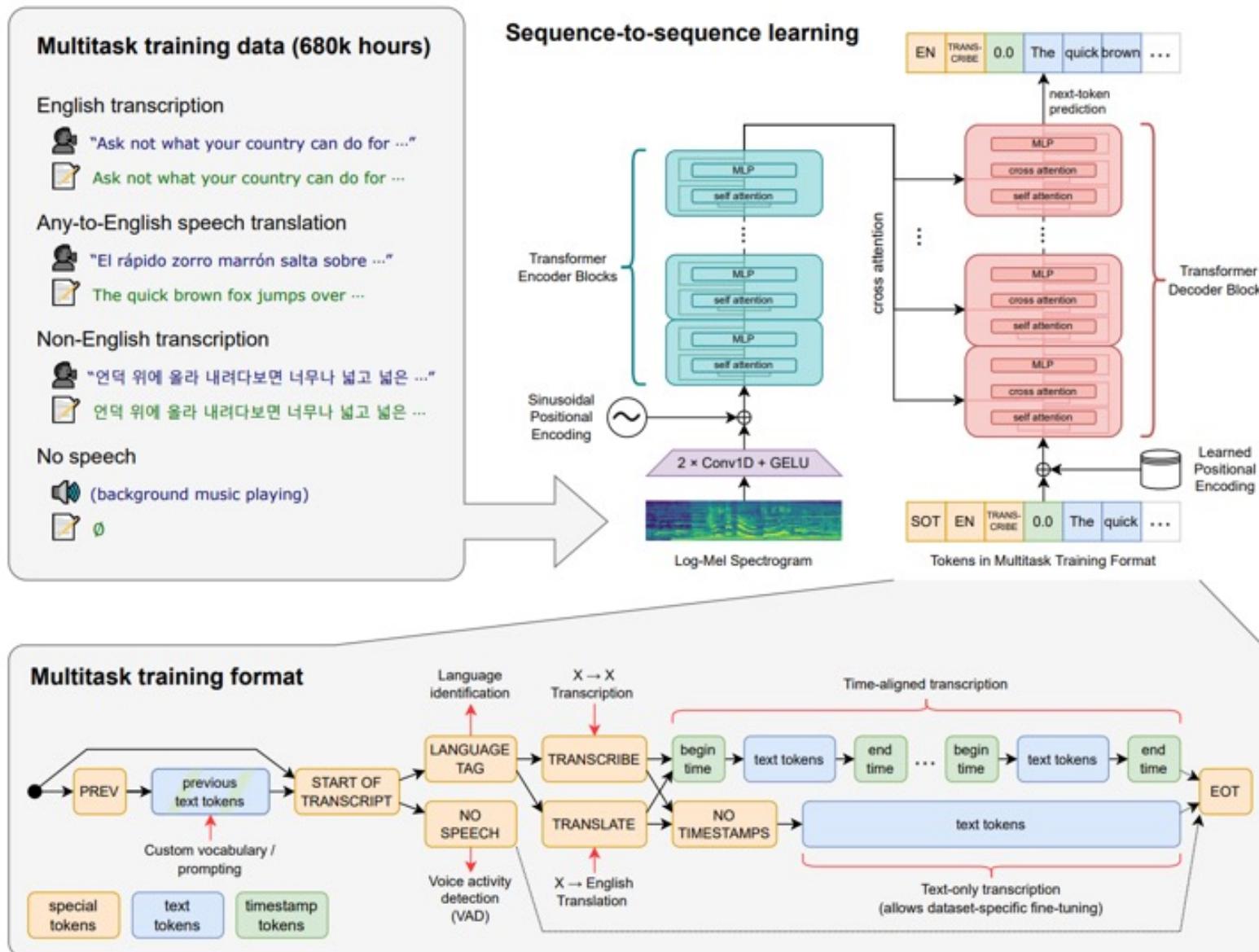


Source: Baevski, Alexei, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli.

"wav2vec 2.0: A framework for self-supervised learning of speech representations." Advances in Neural Information Processing Systems 33 (2020): 12449-12460.

Whisper:

Robust Speech Recognition via Large-Scale Weak Supervision



Microsoft Azure

Text to Speech (TTS)

Text SSML

You can replace this text with any text you wish. You can either write in this text box or paste your own text here.

Try different languages and voices. Change the speed and the pitch of the voice. You can even tweak the SSML (Speech Synthesis Markup Language) to control how the different sections of the text sound. Click on SSML above to give it a try!

Enjoy using Text to Speech!

Language

English (United States)

Voice

Jenny (Neural)

Speaking style

General

Speaking speed: 1.00

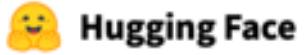


Pitch: 0.00



Play

Hugging Face



Search models, datasets

Models

Datasets

Spaces

Docs

Solutions

Pricing



Log In

Sign Up



The AI community building the future.

Build, train and deploy state of the art models powered by
the reference open source in machine learning.

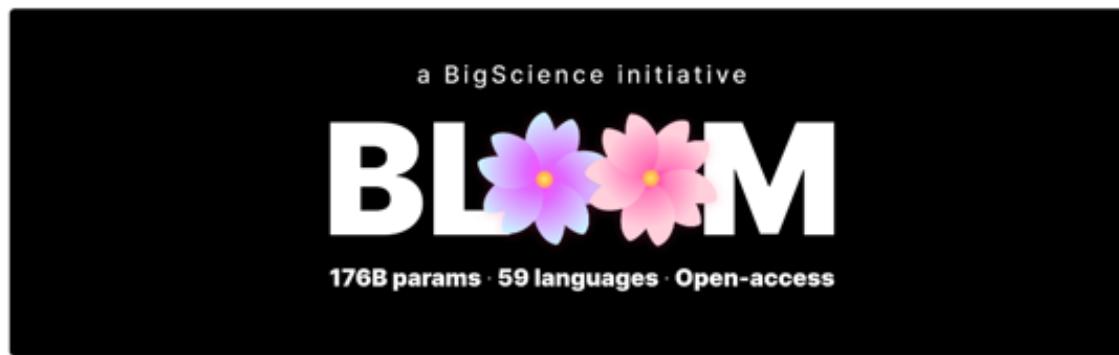
Star

58,696

<https://huggingface.co/>

BLOOM

BigScience Large Open-science Open-access Multilingual Language Model



BigScience Large Open-science Open-access Multilingual Language Model

Version 1.3 / 6 July 2022

Current Checkpoint: **Training Iteration 95000**

Total seen tokens: **366B**

Downloads last month
12,875



⚡ Hosted inference API ⓘ

🖨️ Text Generation

Groups

Examples

I love bloom. Super simple, but so effective! I went through a similar process a couple of years ago when!

sampling greedy

ⓘ [BLOOM prompting tips](#)

Switch to "greedy" for more accurate completion e.g. math/history/translations (but which may be repetitive/less inventive)

Compute

⌘+Enter

1.3

OpenAI Whisper



Search models

Models

Datasets

Spaces

Docs

Solutions

Pricing



Spaces: openai/whisper □ like 422 Running

App

Files

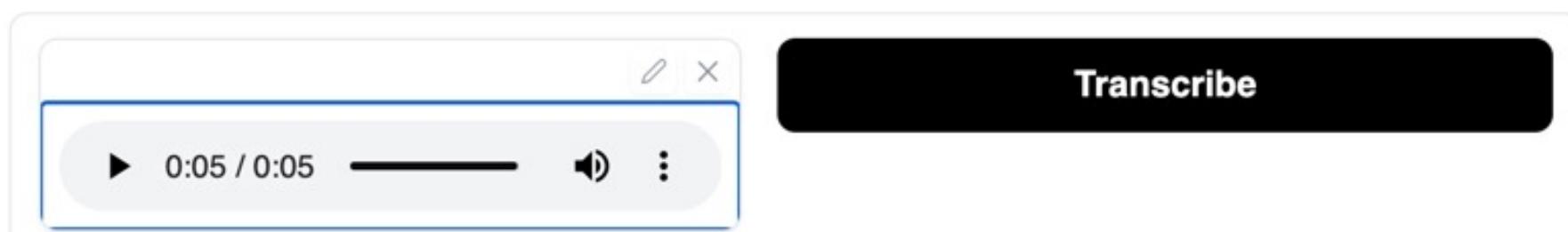
Community 49

Whisper

Whisper is a general-purpose speech recognition model. It is trained on a large dataset of diverse audio and is also a multi-task model that can perform multilingual speech recognition as well as speech translation and language identification. This demo cuts audio after around 30 secs.

You can skip the queue by using google colab for the space:

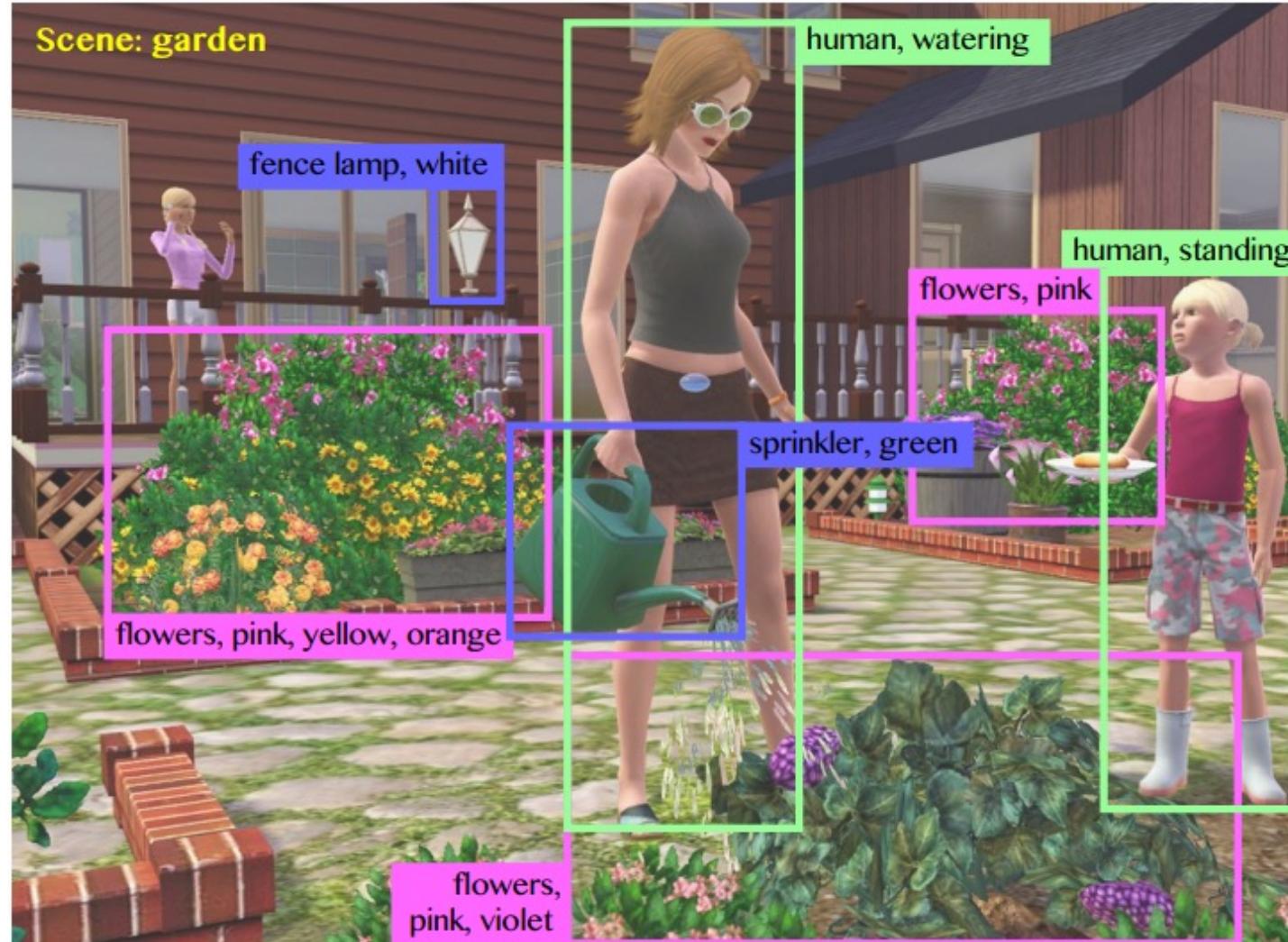
Open in Colab



Source: <https://huggingface.co/spaces/openai/whisper>

Computer vision in the metaverse

with scene understanding, object detection, and human action/activity recognition



Source: Huynh-The, Thien, Quoc-Viet Pham, Xuan-Qui Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim (2022).

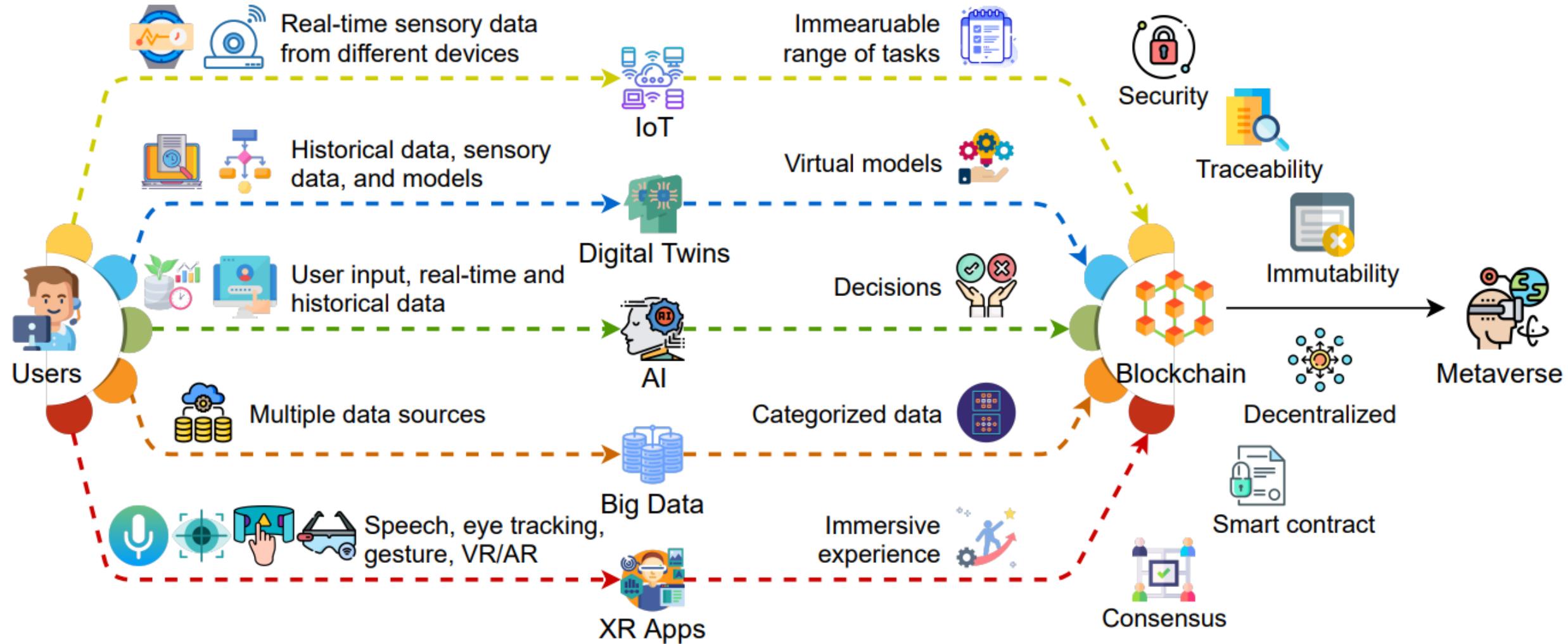
"Artificial Intelligence for the Metaverse: A Survey." arXiv preprint arXiv:2202.10336.

AI for Social Good (AI4SG)

Source: Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle Belgrave et al. (2020)

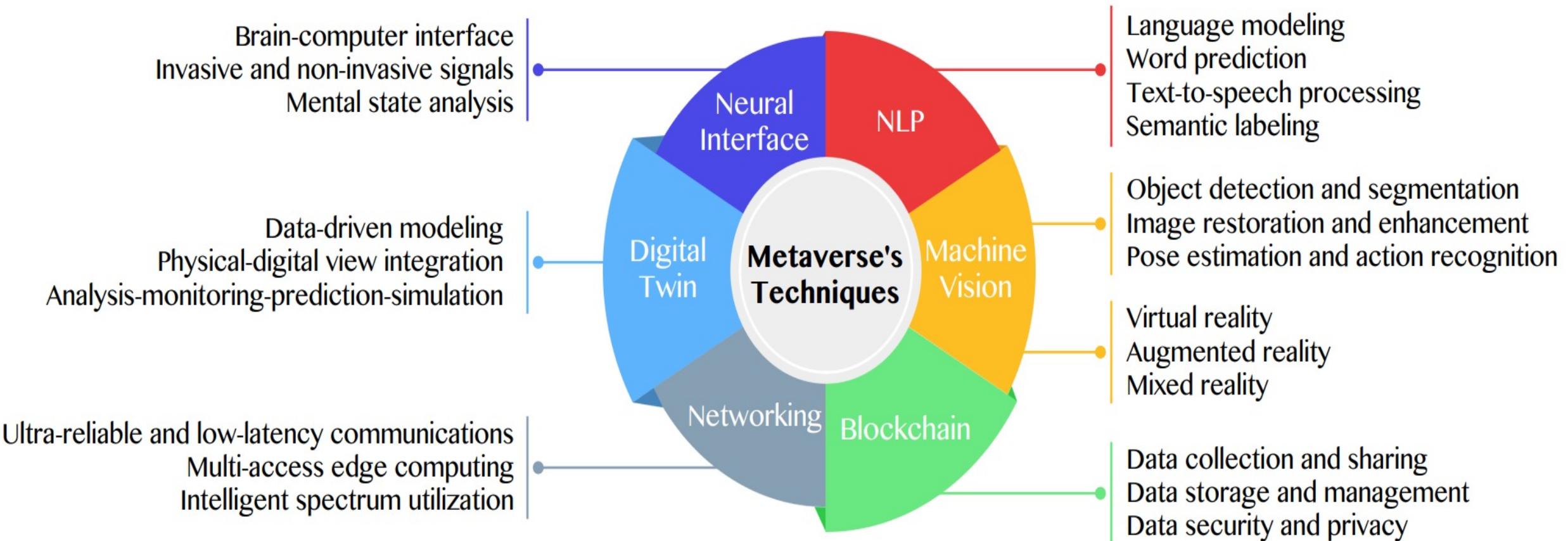
"AI for social good: unlocking the opportunity for positive impact." Nature Communications 11, no. 1: 1-6.

Key Enabling Technologies of the Metaverse



Primary Technical Aspects in the Metaverse

AI with ML algorithms and DL architectures
is advancing the user experience in the virtual world

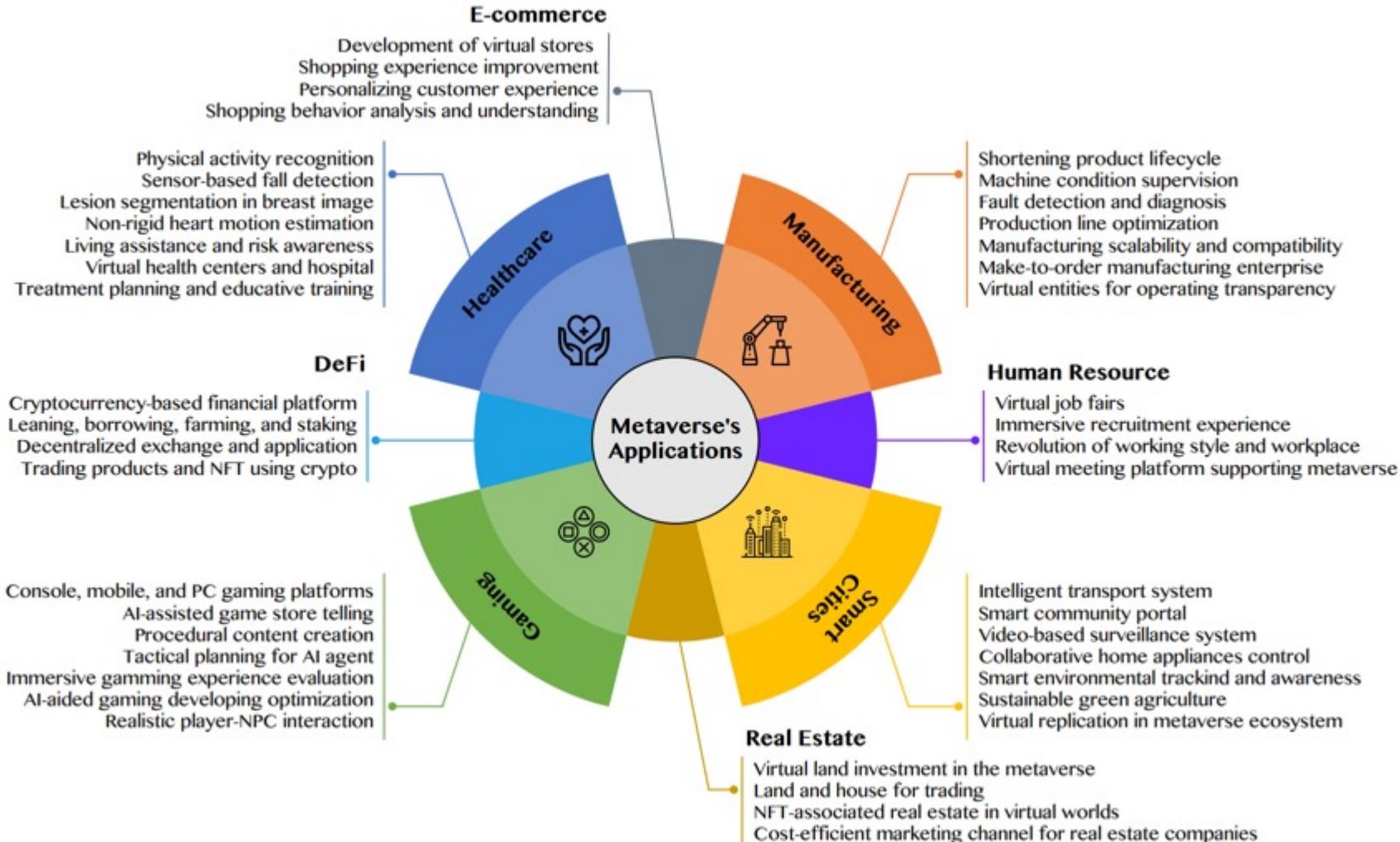


Source: Huynh-The, Thien, Quoc-Viet Pham, Xuan-Qui Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim (2022).

"Artificial Intelligence for the Metaverse: A Survey." arXiv preprint arXiv:2202.10336.

AI for the Metaverse in the Application Aspects

healthcare, manufacturing, smart cities, gaming
E-commerce, human resources, real estate, and DeFi



Source: Huynh-The, Thien, Quoc-Viet Pham, Xuan-Qui Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim (2022).

"Artificial Intelligence for the Metaverse: A Survey." arXiv preprint arXiv:2202.10336.

Sustainable Development Goals (SDGs)

1 NO
POVERTY



2 ZERO
HUNGER



3 GOOD HEALTH
AND WELL-BEING



4 QUALITY
EDUCATION



5 GENDER
EQUALITY



6 CLEAN WATER
AND SANITATION



7 AFFORDABLE AND
CLEAN ENERGY



8 DECENT WORK AND
ECONOMIC GROWTH



9 INDUSTRY, INNOVATION
AND INFRASTRUCTURE



10 REDUCED
INEQUALITIES



11 SUSTAINABLE CITIES
AND COMMUNITIES



12 RESPONSIBLE
CONSUMPTION
AND PRODUCTION



13 CLIMATE
ACTION



14 LIFE
BELOW WATER



15 LIFE
ON LAND



16 PEACE, JUSTICE
AND STRONG
INSTITUTIONS



17 PARTNERSHIPS
FOR THE GOALS




**SUSTAINABLE
DEVELOPMENT
GOALS**

Sustainable Development Goals (SDGs) and 5P

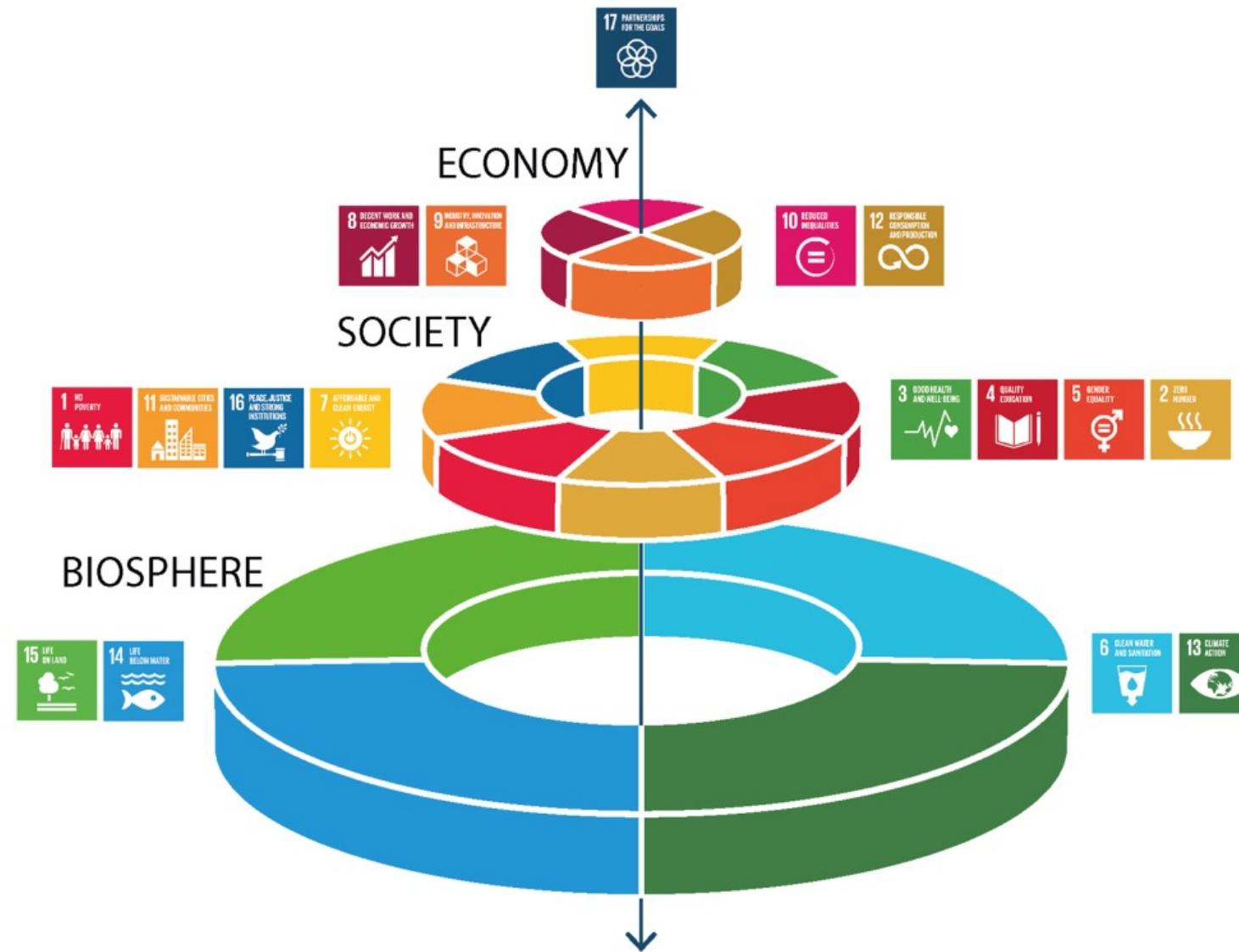
Partnership

Peace

Prosperity

People

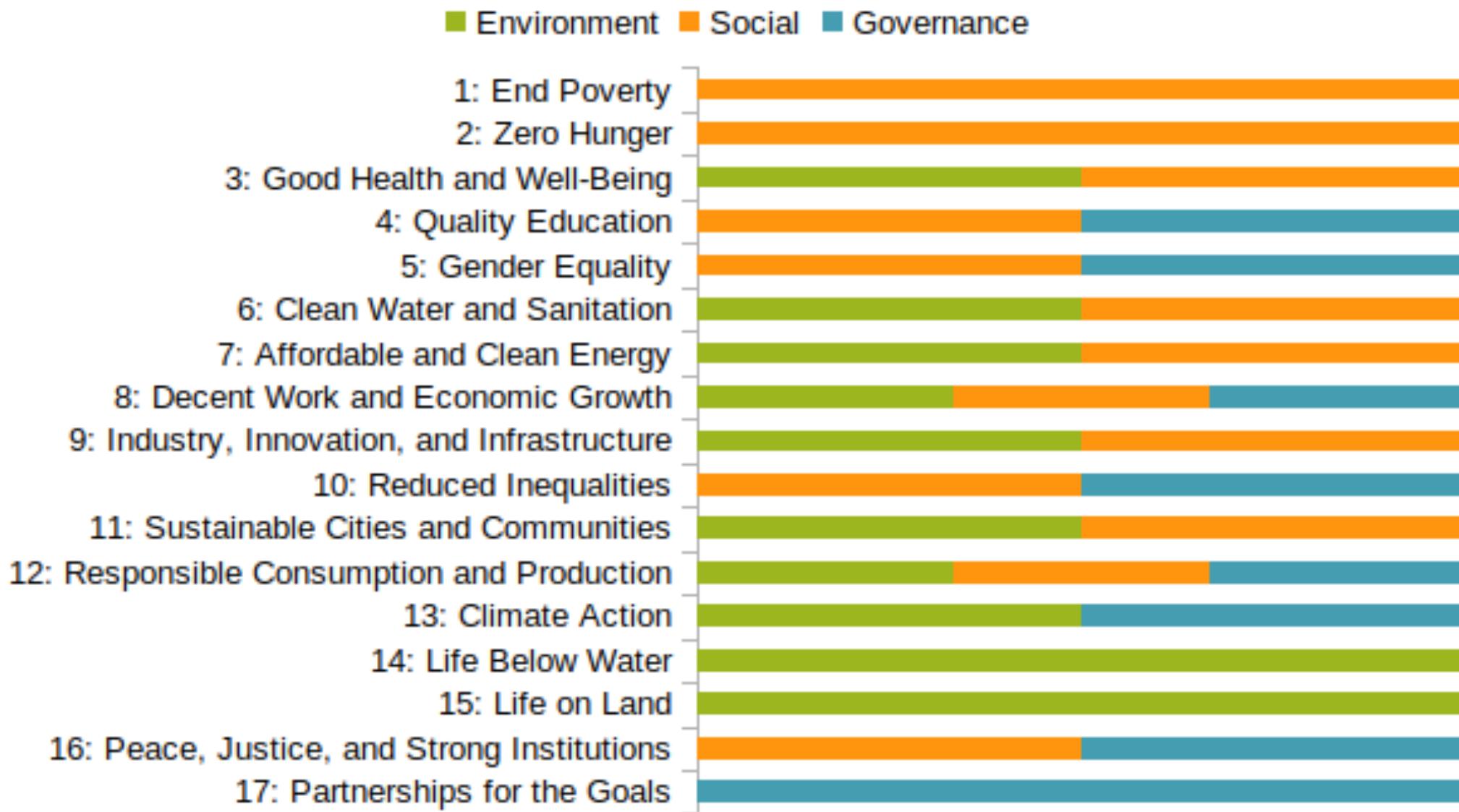
Planet



ESG to 17 SDGs



ESG to 17 SDGs



AI for Social Good (AI4SG)

AI for Sustainable Development

AI4SG 10 Guidelines

- **AI Technology (G1, G2, G3)**
- **Applications (G4, G5, G6, G7, G8)**
- **Data Handling (G9, G10)**

AI4SG 10 Guidelines

AI Technology (G1, G2, G3)

- **G1: Expectations of what is possible with AI need to be well-grounded.**
- **G2: There is value in simple solutions.**
- **G3: Applications of AI need to be inclusive and accessible, and reviewed at every stage for ethics and human rights compliance.**

AI4SG 10 Guidelines

Applications (G4, G5, G6, G7, G8)

- **G4: Goals and use cases should be clear and well-defined.**
- **G5: Deep, long-term partnerships are required to solve large problems successfully.**
- **G6: Planning needs to align incentives, and factor in the limitations of both communities.**
- **G7: Establishing and maintaining trust is key to overcoming organisational barriers.**
- **G8: Options for reducing the development cost of AI solutions should be explored.**

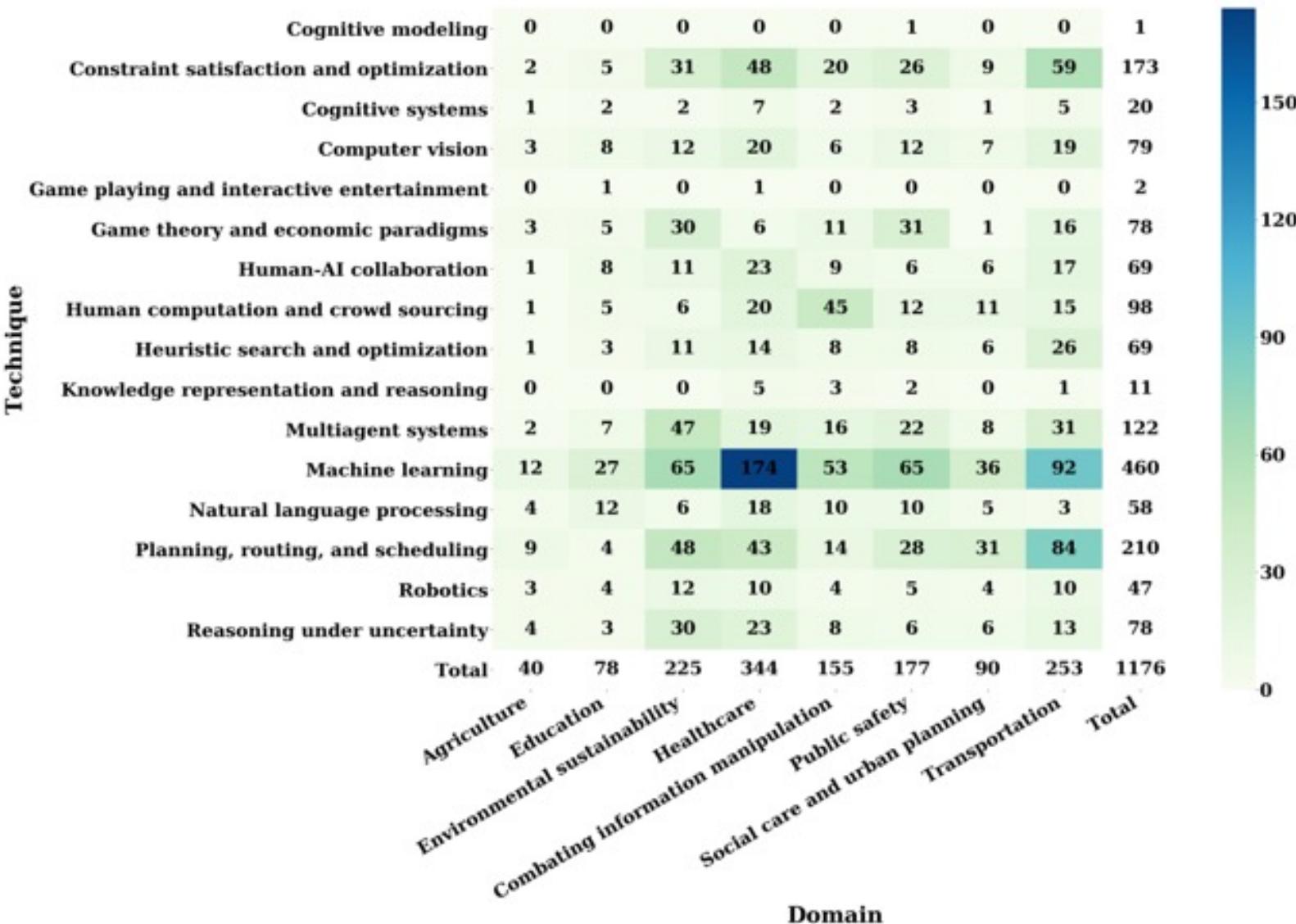
AI4SG 10 Guidelines

Data Handling (G9, G10)

- **G9: Improving data readiness is key.**
- **G10: Data must be processed securely, with utmost respect for human rights and privacy.**

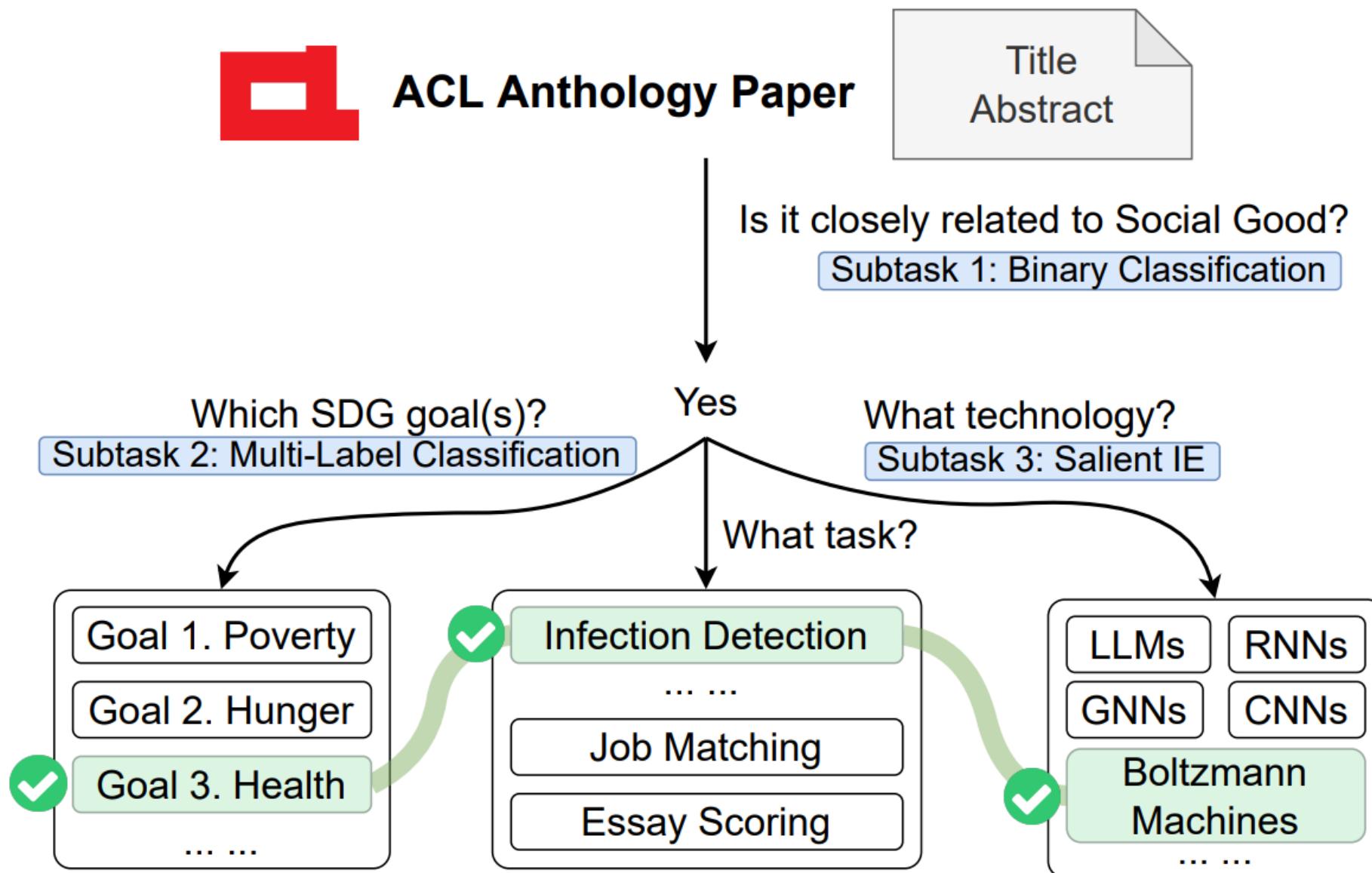
AI for Social Good (AI4SG)

Domains and Techniques

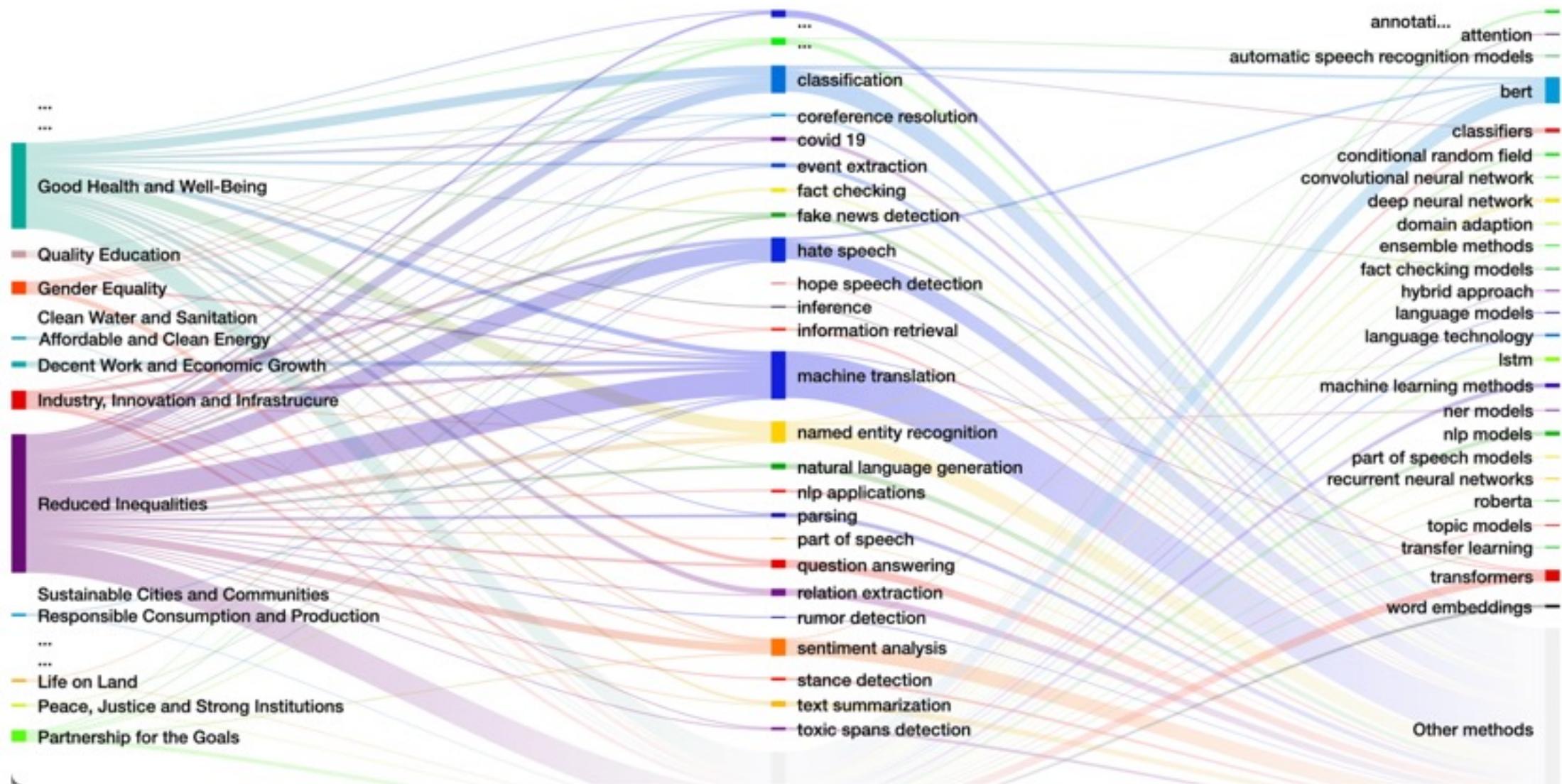


Source: Zheyuan Shi, Ryan, Claire Wang, and Fei Fang (2020). "Artificial intelligence for social good: A survey." arXiv preprint arXiv:2001.01818.

NLP for Social Good (NLP4SG)



NLP for Social Good (NLP4SG) Visualization



Summary

- ChatGPT and Conversational AI
 - Generative Pre-trained Transformer (GPT)
 - Large Language Model (LLM)
- Reinforcement Learning from Human Feedback (RLHF)
- Generative AI
- AI for Social Good (AI4SG)

Acknowledgments: Research Projects

1. Applying AI technology to construct knowledge graphs of cryptocurrency anti-money laundering: a few-shot learning model
 - MOST, 110-2410-H-305-013-MY2, 2021/08/01~2023/07/31
2. Fintech Green Finance for Carbon Market Index, Corporate Finance, and Environmental Policies. Carbon Emission Sentiment Index with AI Text Analytics
 - NTPU, 112-NTPU_ORDA-F-003 , 2023/01/01~2024/12/31
3. Research on speech processing, synthesis, recognition, and sentence construction of people with language disabilities. Multimodal Cross-lingual Task-Oriented Dialogue System
 - NTPU, 112-NTPU_ORDA-F-004, 2023/01/01~2025/12/31
4. Use deep learning to identify commercially dental implant systems - observational study
 - USTP-NTPU-TMU, USTP-NTPU-TMU-112-01, 2023/01/01~2023/12/31
5. Metaverse Avatar Automatic Metadata Generation Module
 - FormosaVerse x NTPU, NTPU-111A413E01, 2022/12/01~2023/11/30
6. Establishment and Implementation of Smart Assistive Technology for Dementia Care and Its Socio-Economic Impacts. Intelligent, individualized and precise care with smart AT and system integration
 - MOST, 111-2627-M-038-001-, 2022/08/01~2023/07/31



Q & A

ChatGPT and Conversational AI with Reinforcement Learning from Human Feedback (RLHF)

Time: 2023/02/17 (Fri) 17:50-19:30

Place: R5117, Soochow University, Taiwan

Host: Prof. Shih-Wei Liao, NTU and Inclusion Technology, NSTC, Taiwan



Min-Yuh Day, Ph.D.
Associate Professor

[Institute of Information Management, National Taipei University](#)

<https://web.ntpu.edu.tw/~myday>



References

- Longbing Cao (2022). "Decentralized ai: Edge intelligence and smart blockchain, metaverse, web3, and desc." IEEE Intelligent Systems 37, no. 3: 6-19.
- Qinglin Yang, Yetong Zhao, Huawei Huang, Zehui Xiong, Jiawen Kang, and Zibin Zheng (2022). "Fusing blockchain and AI with metaverse: A survey." IEEE Open Journal of the Computer Society 3 : 122-136.
- Russell Belk, Mariam Humayun, and Myriam Brouard (2022). "Money, possessions, and ownership in the Metaverse: NFTs, cryptocurrencies, Web3 and Wild Markets." Journal of Business Research 153: 198-205.
- Thien Huynh-The, Quoc-Viet Pham, Xuan-Qui Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim (2022). "Artificial Intelligence for the Metaverse: A Survey." arXiv preprint arXiv:2202.10336.
- Thippa Reddy Gadekallu, Thien Huynh-The, Weizheng Wang, Gokul Yenduri, Pasika Ranaweera, Quoc-Viet Pham, Daniel Benevides da Costa, and Madhusanka Liyanage (2022). "Blockchain for the Metaverse: A Review." arXiv preprint arXiv:2203.09738.
- Dan Sheridan, James Harris, Frank Wear, Jerry Cowell Jr, Easton Wong, and Abbas Yazdinejad (2022). "Web3 Challenges and Opportunities for the Market." arXiv preprint arXiv:2209.02446.
- Yves Hilpisch (2020), Artificial Intelligence in Finance: A Python-Based Guide, O'Reilly Media.
- Aurélien Géron (2019), Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Edition, O'Reilly Media.
- Yves Hilpisch (2018), Python for Finance: Mastering Data-Driven Finance, 2nd Edition, O'Reilly Media.
- Paolo Sironi (2016), FinTech Innovation: From Robo-Advisors to Goal Based Investing and Gamification, Wiley.
- Yuxing Yan (2017), Python for Finance: Apply powerful finance models and quantitative analysis with Python, Second Edition, Packt Publishing
- Campbell R. Harvey, Ashwin Ramachandran, Joey Santoro, Fred Ehrsam (2021), DeFi and the Future of Finance, Wiley
- Matt Fortnow and QuHarrison Terry (2021), The NFT Handbook - How to Create, Sell and Buy Non-Fungible Tokens, Wiley
- Parma Bains, Mohamed Diaby, Dimitris Drakopoulos, Julia Faltermeier, Federico Grinberg, Evan Papageorgiou, Dmitri Petrov, Patrick Schneider, and Nobu Sugimoto (2021), The Crypto Ecosystem and Financial Stability Challenges, International Monetary Fund, October 2021
- Henrik Skaug Sætra (2021) "A Framework for Evaluating and Disclosing the ESG Related Impacts of AI with the SDGs." Sustainability 13, no. 15: 8503
- Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle Belgrave et al. (2020) "AI for social good: unlocking the opportunity for positive impact." Nature Communications 11, no. 1: 1-6.
- Zheyuan Shi, Ryan, Claire Wang, and Fei Fang (2020). "Artificial intelligence for social good: A survey." arXiv preprint arXiv:2001.01818.
- Fernando Gonzalez, Zhijing Jin, Jad Beydoun, Bernhard Schölkopf, Tom Hope, Rada Mihalcea, and Mrinmaya Sachan (2022). "How Is NLP Addressing the 17 UN Sustainability Goals? A Challenge Set of Social Good Paper Classification and Information Extraction."
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.