Technology and Principles Behind ChatGPT and Similar Models

Mladen Fernežir, Lead Data Scientist & Co-Founder

VELEBIT AI

About Velebit AI

- AI custom R&D
- AI consultancy
- Fast prototyping
- Images, text, tabular data
- Team with 8 years of experience
- Data engineering
- Deployment and monitoring

photomath INFINUM

Trikoder njuškalo markt.de Der Marktplatz für Deutschland

WILLHABEN.AT® bolha.com TURBULENT DYNAMICS

uniri Jožef Stefan Institute Ljubljana, Slovenija

Creative Fabrica infostud GRUPA computer systems

VELEBIT AI

Co-founders

Davor

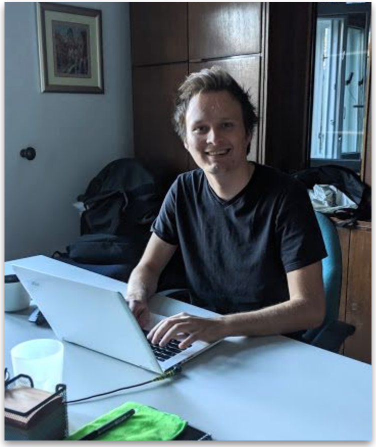
CEO

Ivan

Machine Learning Engineer

Mladen

Data Scientist



Tomislav

Data Engineer

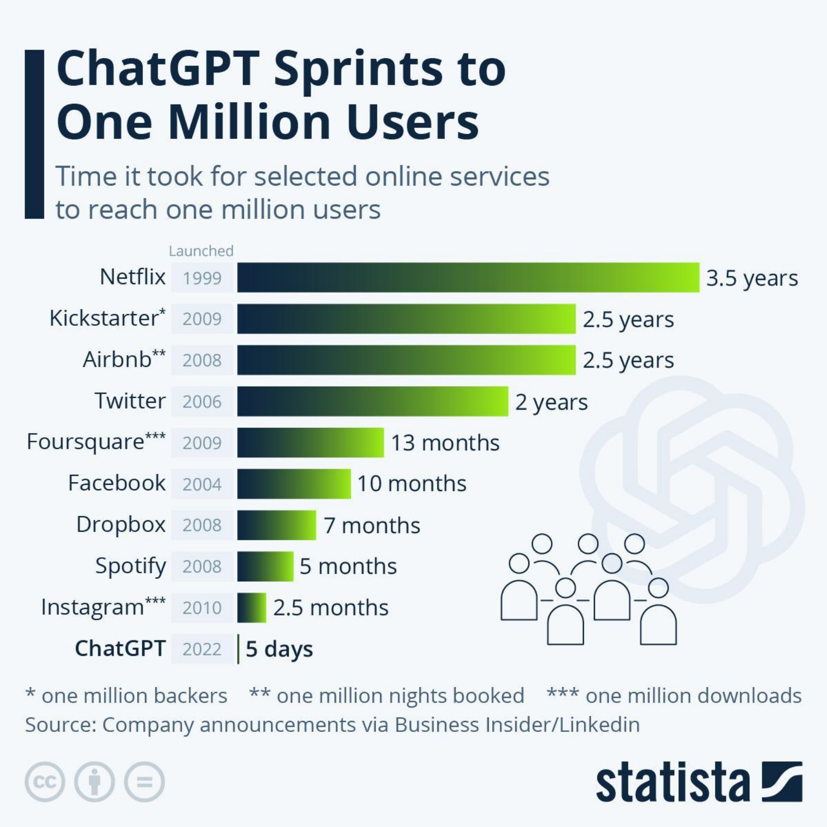
Outline

- Introduction
- ChatGPT Basics
- Alignment Research
- Challenges & Concerns
- Language Models in Velebit AI
- Current Outlook
- Educational Resources

Introduction

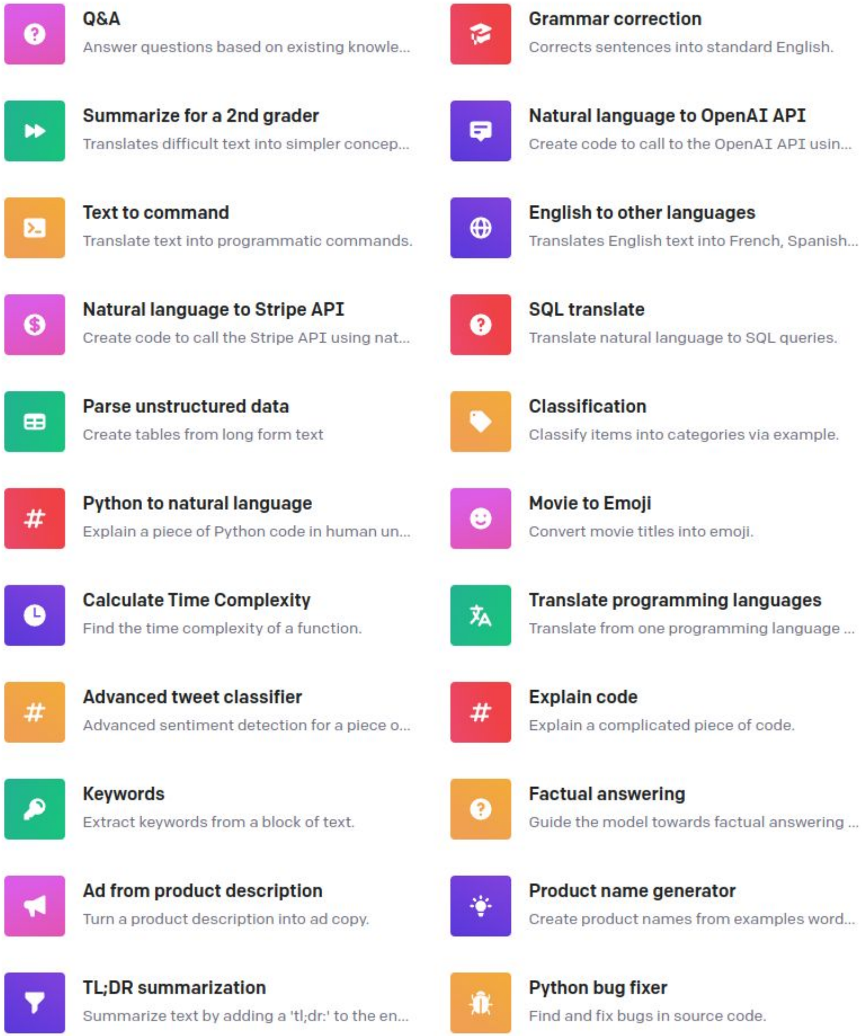
Massive ChatGPT adoption

- ChatGPT took the Internet by storm
- Just 5 days to 1 million users
- How Disruptive is ChatGPT and Why?
- How innovative is it as a technological breakthrough?
- How does OpenAI compare to others?



* one million backers ** one million nights booked *** one million downloads

Source: Company announcements via Business Insider/Linkedin



Q&A
Answer questions based on existing knowle...

Grammar correction
Corrects sentences into standard English.

Summarize for a 2nd grader
Translates difficult text into simpler concep...

Natural language to OpenAI API
Create code to call to the OpenAI API usin...

Text to command
Translate text into programmatic commands.

English to other languages
Translates English text into French, Spanish...

Natural language to Stripe API
Create code to call the Stripe API using nat...

SQL translate
Translate natural language to SQL queries.

Parse unstructured data
Create tables from long form text

Classification
Classify items into categories via example.

Python to natural language
Explain a piece of Python code in human un...

Movie to Emoji
Convert movie titles into emoji.

Calculate Time Complexity
Find the time complexity of a function.

Translate programming languages
Translate from one programming language ...

Advanced tweet classifier
Advanced sentiment detection for a piece o...

Explain code
Explain a complicated piece of code.

Keywords
Extract keywords from a block of text.

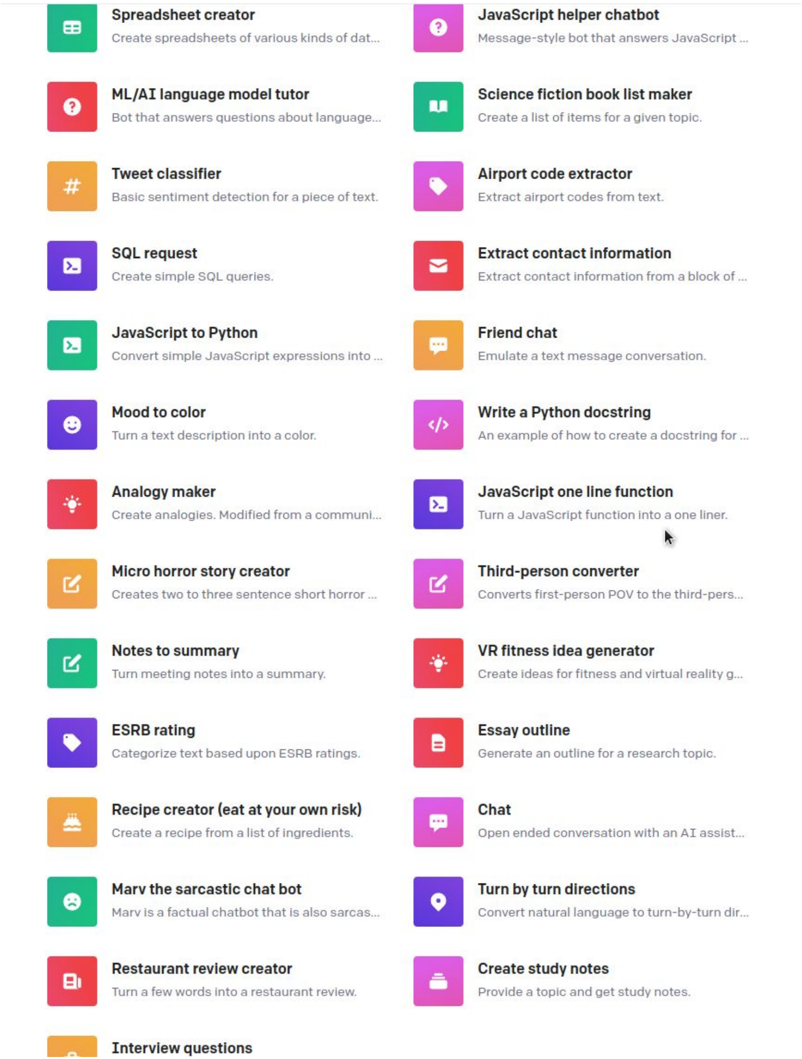
Factual answering
Guide the model towards factual answering ...

Ad from product description
Turn a product description into ad copy.

Product name generator
Create product names from examples word...

TL;DR summarization
Summarize text by adding a 'tl;dr:' to the en...

Python bug fixer
Find and fix bugs in source code.



Spreadsheet creator
Create spreadsheets of various kinds of dat...

JavaScript helper chatbot
Message-style bot that answers JavaScript ...

ML/AI language model tutor
Bot that answers questions about language...

Science fiction book list maker
Create a list of items for a given topic.

Tweet classifier
Basic sentiment detection for a piece of text.

Airport code extractor
Extract airport codes from text.

SQL request
Create simple SQL queries.

Extract contact information
Extract contact information from a block of ...

JavaScript to Python
Convert simple JavaScript expressions into ...

Friend chat
Emulate a text message conversation.

Mood to color
Turn a text description into a color.

Write a Python docstring
An example of how to create a docstring for ...

Analogy maker
Create analogies. Modified from a communi...

JavaScript one line function
Turn a JavaScript function into a one liner.

Micro horror story creator
Creates two to three sentence short horror ...

Third-person converter
Converts first-person POV to the third-pers...

Notes to summary
Turn meeting notes into a summary.

VR fitness idea generator
Create ideas for fitness and virtual reality g...

ESRB rating
Categorize text based upon ESRB ratings.

Essay outline
Generate an outline for a research topic.

Recipe creator (eat at your own risk)
Create a recipe from a list of ingredients.

Chat
Open ended conversation with an AI assist...

Marv the sarcastic chat bot
Marv is a factual chatbot that is also sarcas...

Turn by turn directions
Convert natural language to turn-by-turn dir...

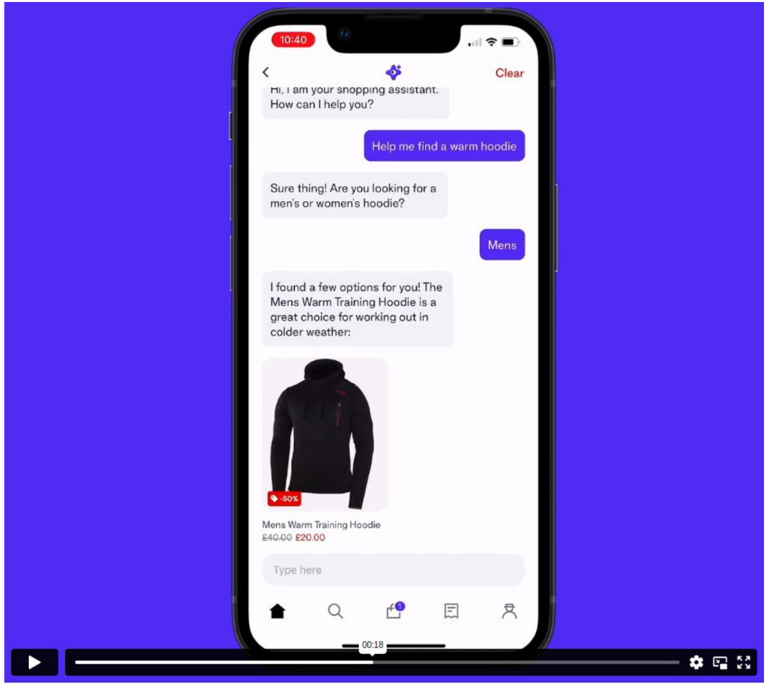
Restaurant review creator
Turn a few words into a restaurant review.

Create study notes
Provide a topic and get study notes.

Interview questions

ChatGPT Plugins

- Plugins add extra functionality
- Possible to call external APIs for different tasks
- Wolfram Alpha, Internet Browsing, Python Interpreter, Knowledge Retrieval, Shopping, etc.



Hi, I am your shopping assistant. How can I help you?

Help me find a warm hoodie

Sure thing! Are you looking for a men's or women's hoodie?

Mens

I found a few options for you! The Mens Warm Training Hoodie is a great choice for working out in colder weather:

Mens Warm Training Hoodie
~~£40.00~~ £20.00

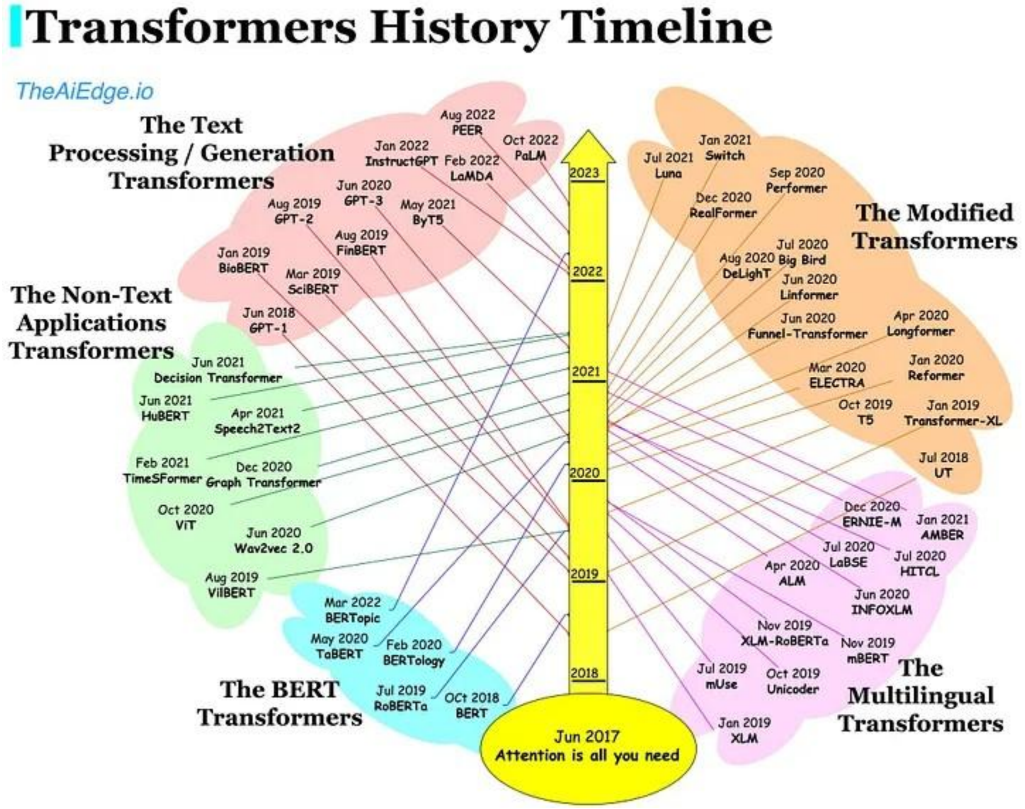
Shopify's Shop app

https://openai.com/blog/introducing-chatgpt-and-whisper-apis (March 1, 2023)
https://openai.com/blog/chatgpt-plugins (March 23, 2023)

ChatGPT Basics

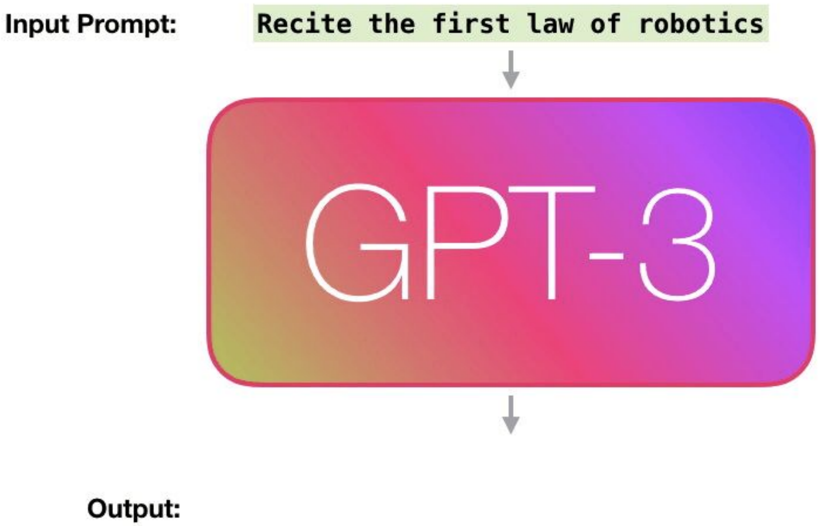
Large Language Models

- ChatGPT is a LLM
- A type of a Transformer neural network
- GPT family: predicting the next probable word



Transformer self-supervised learning

- GPT takes into account all previous words to predict the next probable word
- We can add prompts as inputs for guidance



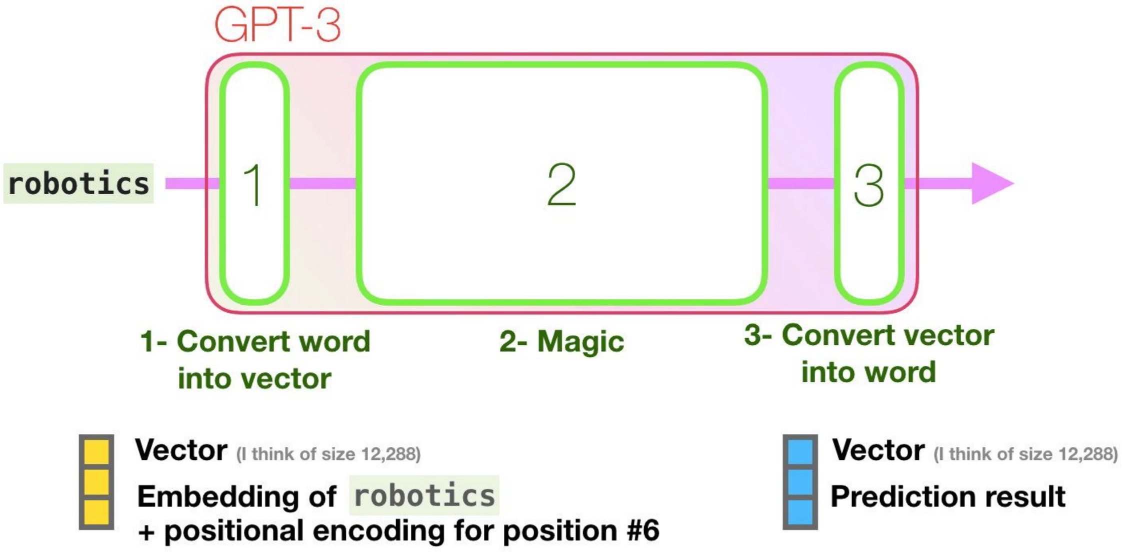
Input Prompt: `Recite the first law of robotics`

Output:

https://jalammar.github.io/how-gpt3-works-visualizations-animations/

GPT elements

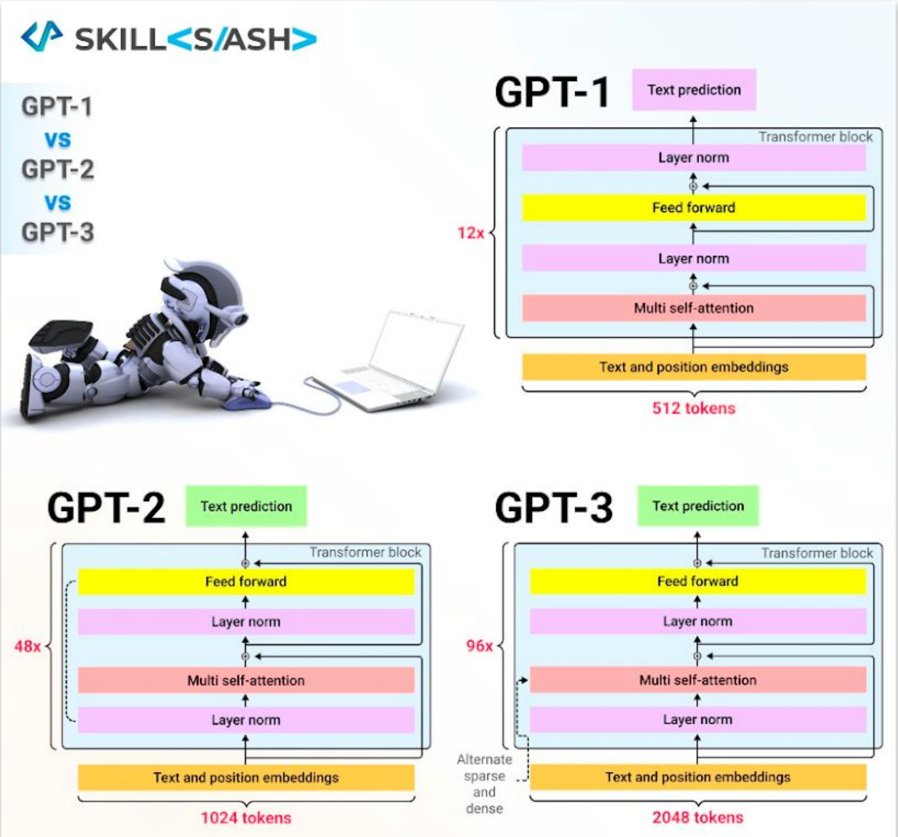
- Input words are converted to vectors (embeddings)
- We also add embeddings for each word position


GPT-3
robotics
1
2
3
1- Convert word into vector
2- Magic
3- Convert vector into word
Vector (I think of size 12,288)
Embedding of robotics
+ positional encoding for position #6
Vector (I think of size 12,288)
Prediction result

https://jalammar.github.io/how-gpt3-works-visualizations-animations/

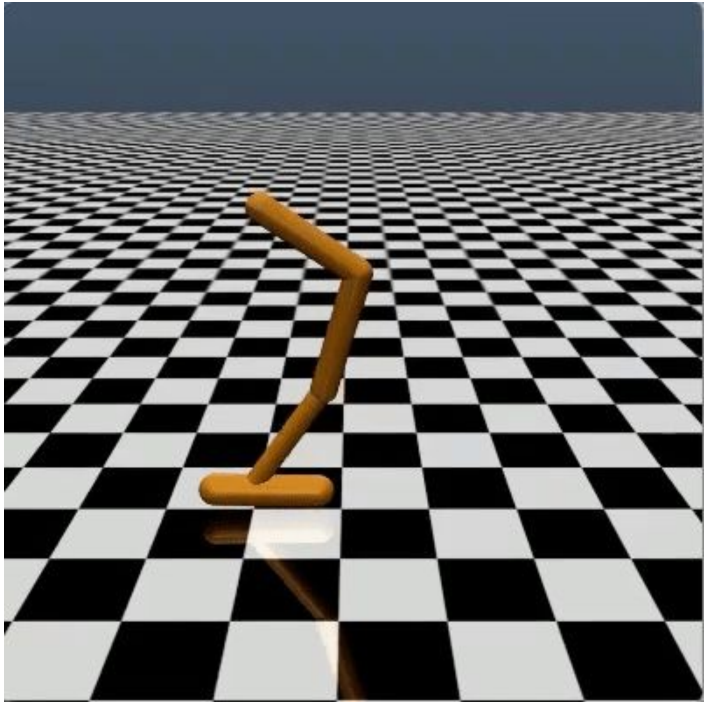
GPT 1, 2, 3

- GPT 1, 2, 3 progression
- Larger models
- Larger datasets
- More tokens
- All decoder only
- Fundamentally the same



Reinforcement Learning addition

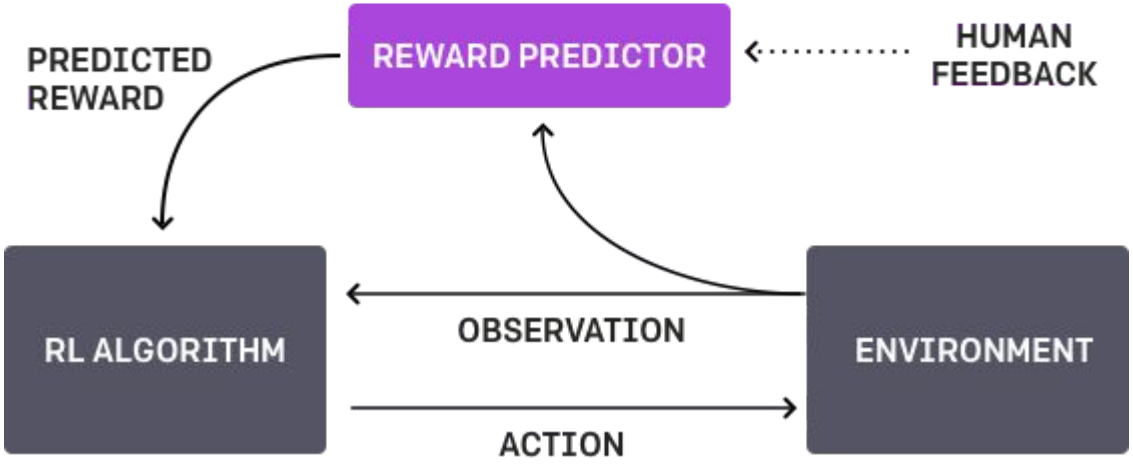
- How to teach an agent to do the backflip?
- Ask human raters whether a flip A was better than a flip B



https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/

Reinforcement Learning addition

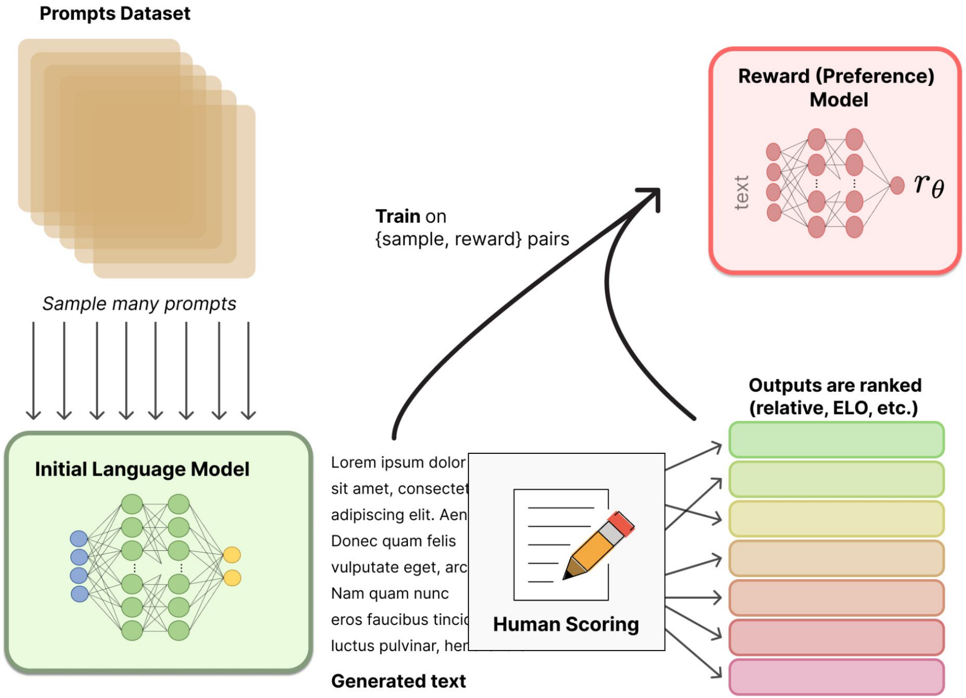
- New Idea: improve LLMs such as GPT by adding well-known techniques from Reinforcement Learning



https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/

New reward model

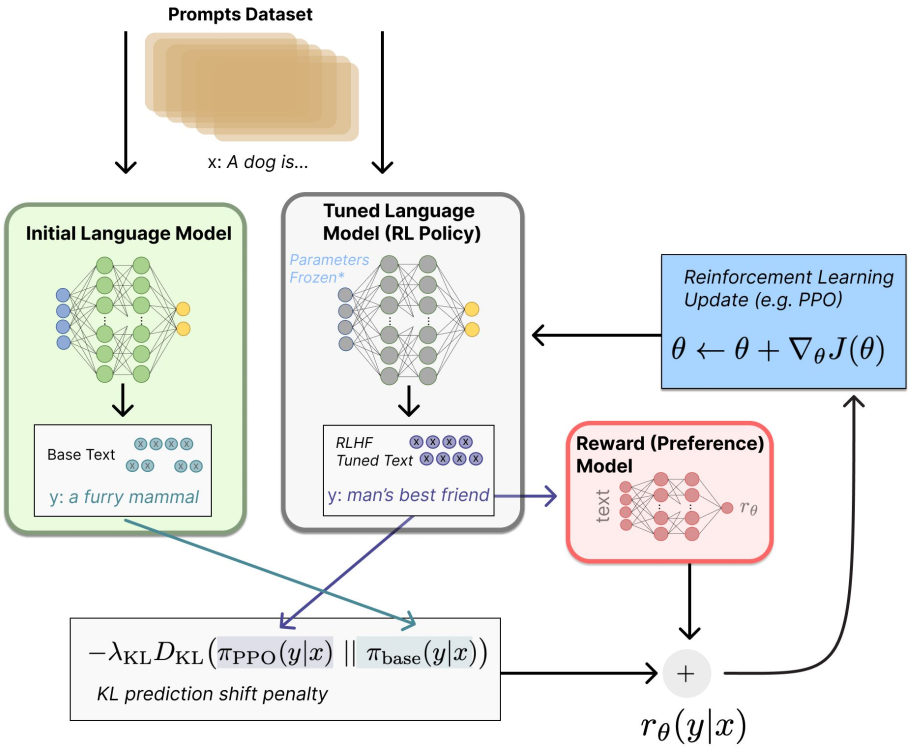
- Basic GPT 3 can be toxic, biased, and not in-line with user intent (prompt)
- We can use human raters to judge different GPT outputs



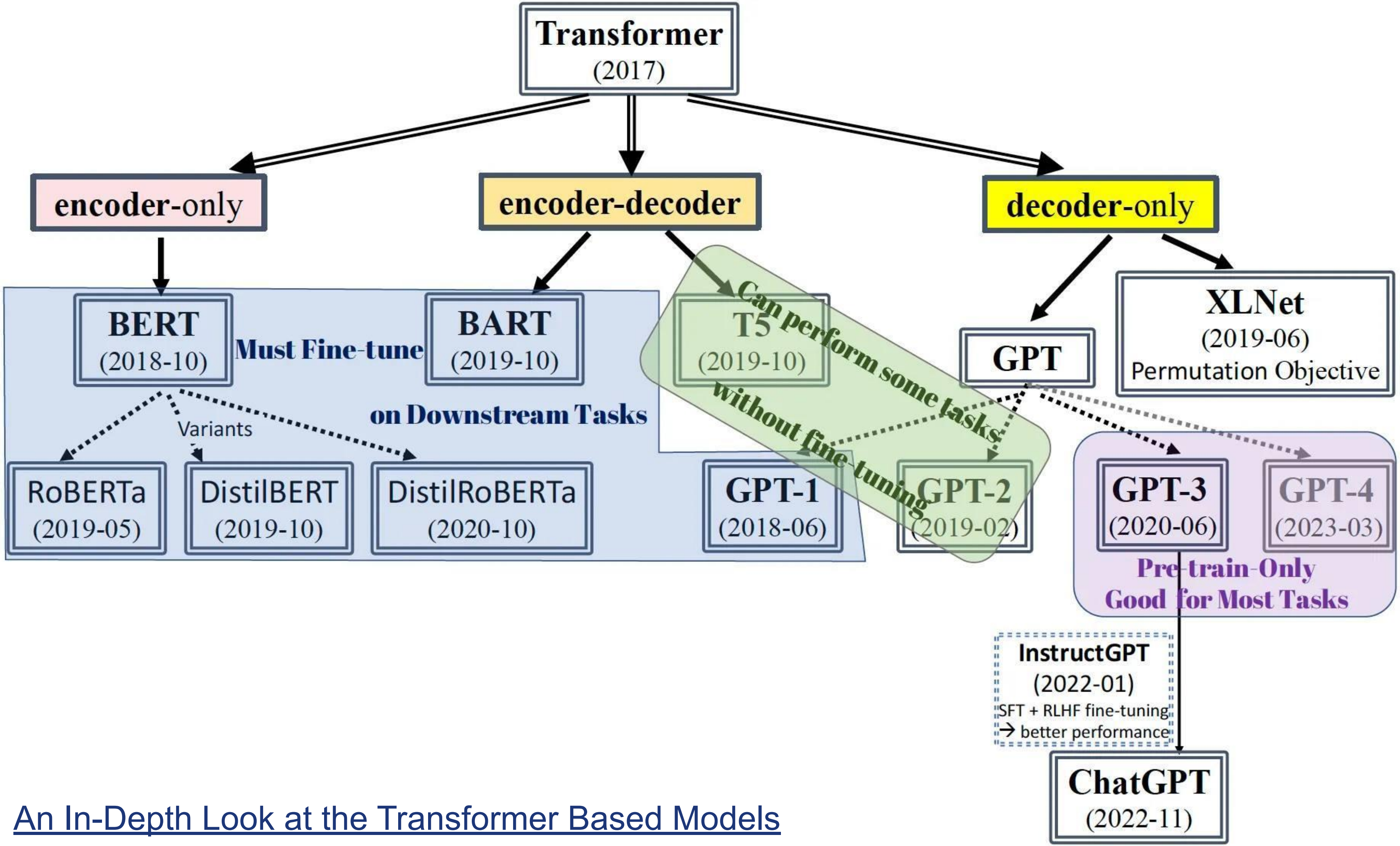
https://huggingface.co/blog/rlhf

RLHF for GPT 3

- We use the new reward model from human feedback as a basis to update the policy by which the new language model creates words


Prompts Dataset
x: A dog is...
Initial Language Model
Base Text
y: a furry mammal
Tuned Language Model (RL Policy)
Parameters Frozen*
RLHF Tuned Text
y: man's best friend
Reinforcement Learning Update (e.g. PPO)
$\theta \leftarrow \theta + \nabla_\theta J(\theta)$
Reward (Preference) Model
text
r_θ
$-\lambda_{\text{KL}} D_{\text{KL}}(\pi_{\text{PPO}}(y|x) \,||\, \pi_{\text{base}}(y|x))$
KL prediction shift penalty
+
$r_\theta(y|x)$

https://huggingface.co/blog/rlhf



An In-Depth Look at the Transformer Based Models

Alignment Research

InstructGPT

- OpenAI already had a very similar model replacing GPT 3 in their API: InstructGPT
- The same approach as ChatGPT, but without the large public attention
- They call it their first Alignment Research product
- text-davinci-003 in the API is Instruct GPT, 3.5 series like ChatGPT

https://platform.openai.com/docs/model-index-for-researchers

What's Alignment Research?

- Users prefer InstructGPT / ChatGPT to basic GPT
- It is the RLHF part that aligns human intent and some predefined human values to model outputs
- OpenAI (and others) want safe, unbiased, useful AI, aligned with human interests
- Alignment Research is a broad research area: for now it is (mostly) about human language, but it will be any AI action in the future

https://openai.com/blog/our-approach-to-alignment-research/

Challenges and Concerns

ChatGPT limitations and challenges

- Multiple known limitations to ChatGPT
- Issues of factual correctness, bias, and toxicity
- Questions of values
- It is still just a model predicting statistically likely words, but to please the human raters
- Hallucinations instead of facts
- Confident, but making things up
- Legal issues & copyright
- Ethical and educational challenges

Do androids dream of electric sheep?

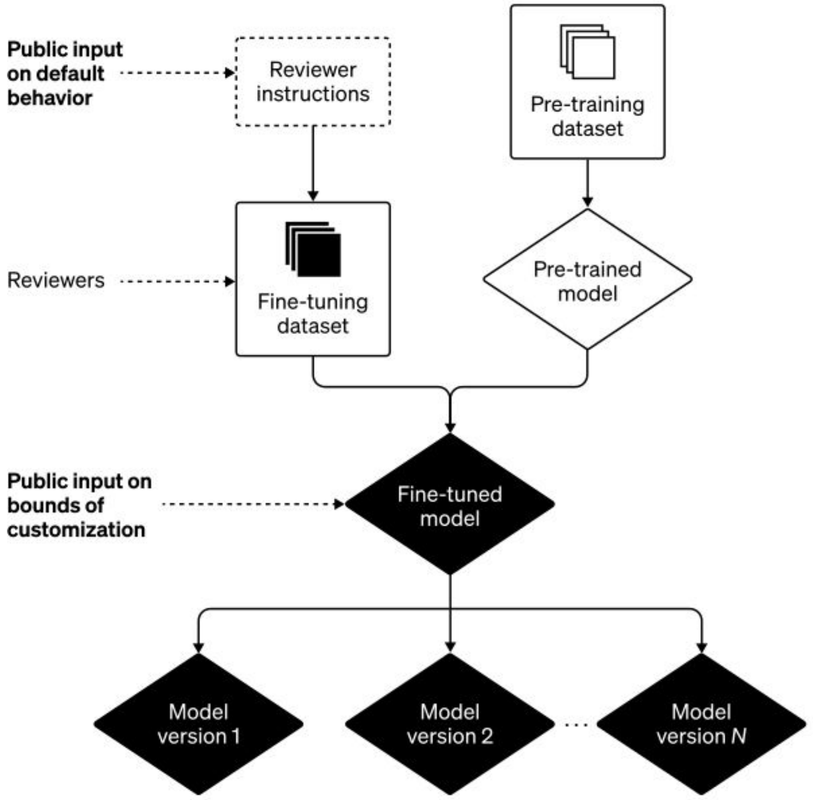
- Let's check if the ChatGPT can pass the Voight-Kampff Test!
- The question is which values and behavior do we want to mimic
- Can Bing's "ChatBPD"?

Blade Runner city, AI imagination

Alignment Research is controversial

- 1. Improve default behavior
- 2. Define your AI’s values, within broad bounds
- 3. Public input on defaults and hard bounds

How should AI systems behave, and who should decide?
Open AI, Feb 16 2023



Language Models in Velebit AI

Language Model Development

- Collaboration with UNIRI on the InfoCov project
- Base language model for Croatian:
  - CroSloEngual BERT,
  - BERTić* [bert-ich] /bɜrtitʃ/ - A transformer language model for Bosnian, Croatian, Montenegrin and Serbian
- Self-supervised tuning to COVID specific Croatian data
- Supervised COVID sentiment classification
- Supervised retweet prediction

BERTić model self-supervised tuning

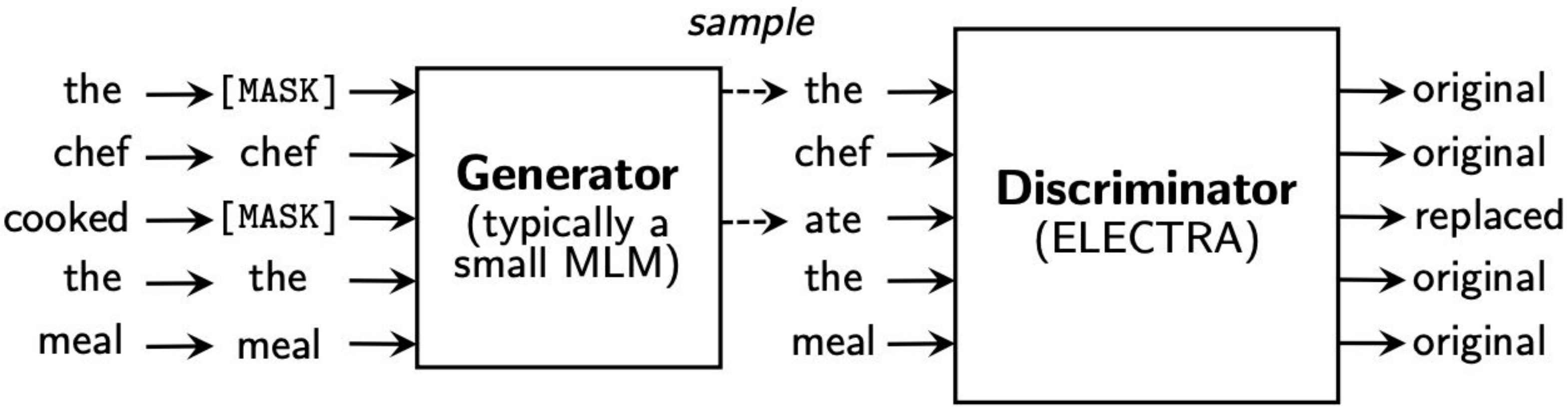


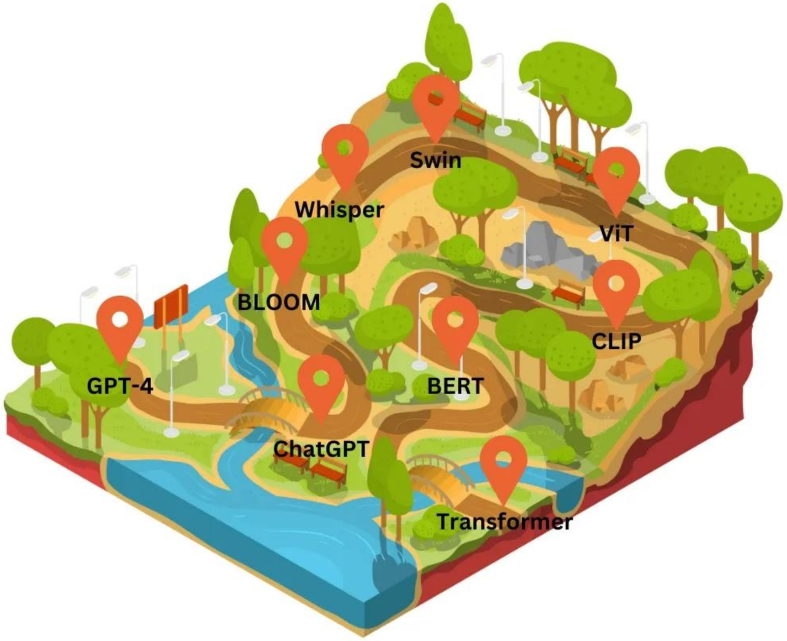
Figure 2: An overview of replaced token detection. The generator can be any model that produces an output distribution over tokens, but we usually use a small masked language model that is trained jointly with the discriminator. Although the models are structured like in a GAN, we train the generator with maximum likelihood rather than adversarially due to the difficulty of applying GANs to text. After pre-training, we throw out the generator and only fine-tune the discriminator (the ELECTRA model) on downstream tasks.

Retweet Prediction

- Content features extracted from a transformer language model
- Tabular features representing Twitter users and their interactions (categorical and numerical)
- Different types of classification algorithms: MLP, Random Forest, LightGBM, NODE, TabNet, Category Embedding Model
- https://github.com/InfoCoV/Multi-Cro-CoV-cseBERT

Other Projects & Transformers

- Automatic Text and Image Categorization
- Image and Text Similarity
- 2D and 3D Object Detection & Segmentation
- Item Tagging and Attribute Prediction



The Map Of Transformers

Current Outlook

We will align to behaviors and actions

- ChatGPT is just the beginning
- We've entered the time of Alignment Research
- Research and products already underway for better factual understanding, and integration with search
- Many companies have the same technology and understanding, besides OpenAI
- Google, Meta, Microsoft, DeepMind, Anthropic, ...
- Some other tools to try: you.com, perplexity.ai

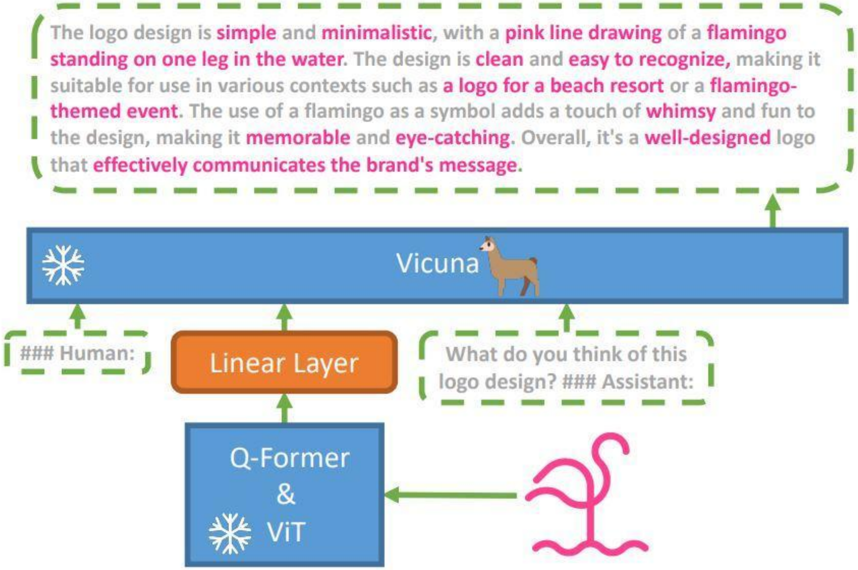
Open Source Explosion of Models

- LLaMA
- Alpaca
- GPT4ALL
- Vicuna
- Dolly
- StableLM
- Open Assistant Models
- …

"A Stochastic Parrot, flat design, vector art" — Stable Diffusion XL

List of Open Sourced Fine-Tuned Large Language Models (LLM)

Open Source MiniGPT 4



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.

Vicuna

Human:

Linear Layer

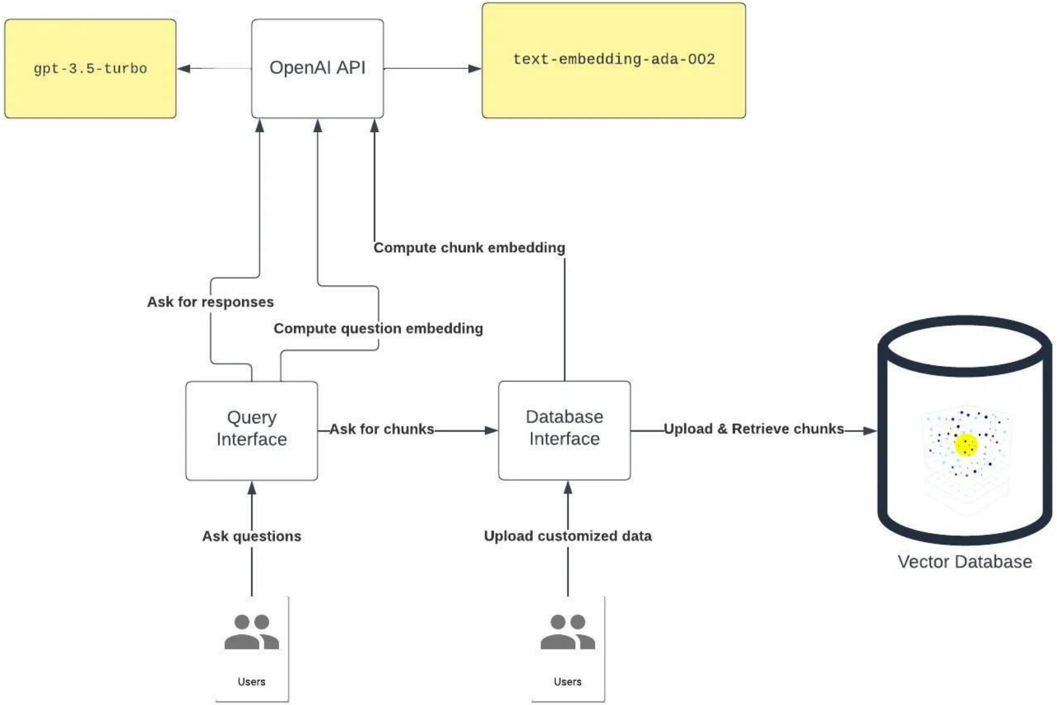
What do you think of this logo design? ### Assistant:

Q-Former & ViT

Figure 1: **The architecture of MiniGPT-4.** It consists of a vision encoder with a pretrained ViT and Q-Former, a single linear projection layer, and an advanced Vicuna large language model. MiniGPT-4 only requires training the linear projection layer to align the visual features with the Vicuna.

MiniGPT-4:Enhancing Vision-language Understanding with Advanced Large Language Models

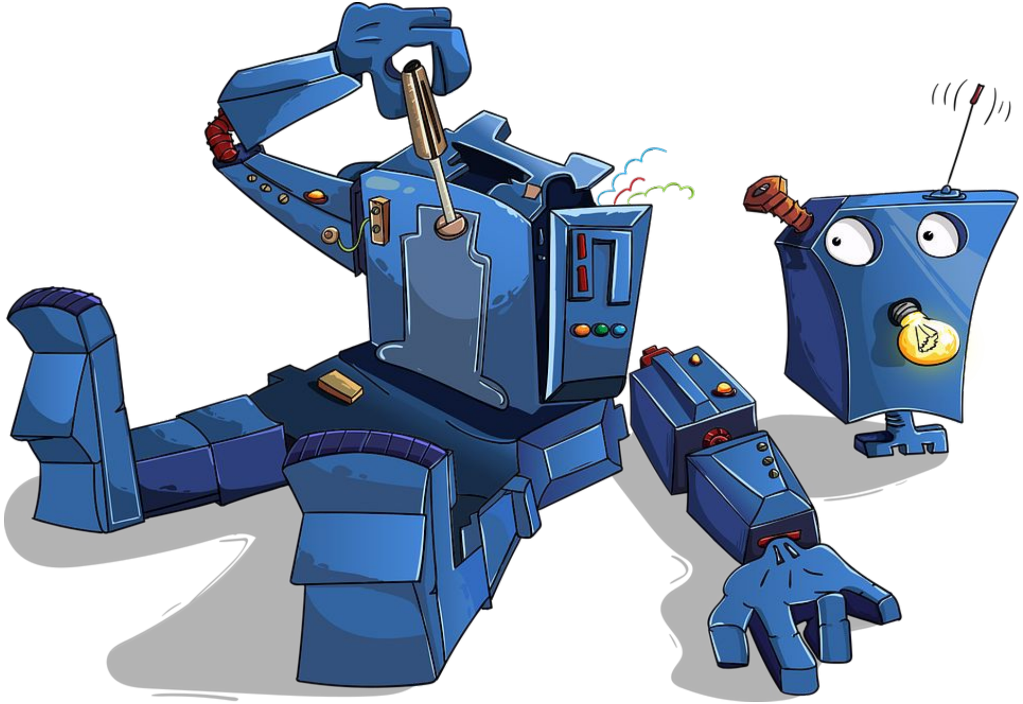
Add External Memory



[Enhancing ChatGPT With Infinite External Memory Using Vector Database and ChatGPT Retrieval Plugin]

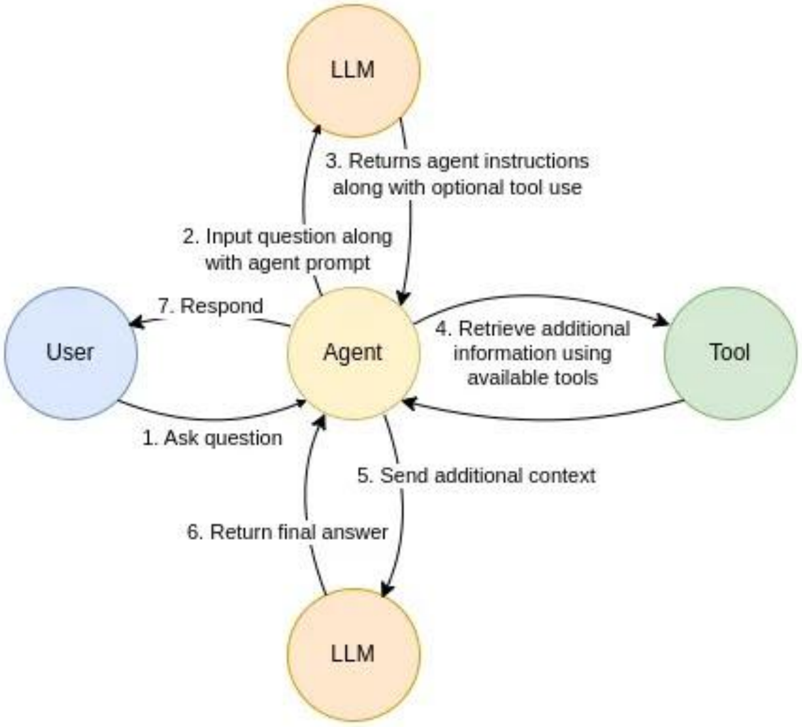
Multiple external tools

- Toolformer by Meta
- LLM that learns to use external tools
- calculator, Q&A system, search engines, translation, calendar
- Feb 9, 2023



https://syncedreview.com/2023/02/16/meta-ai-upfs-toolformer-enabling-language-models-to-teach-themselves-to-use-external-tools/

Agent Development

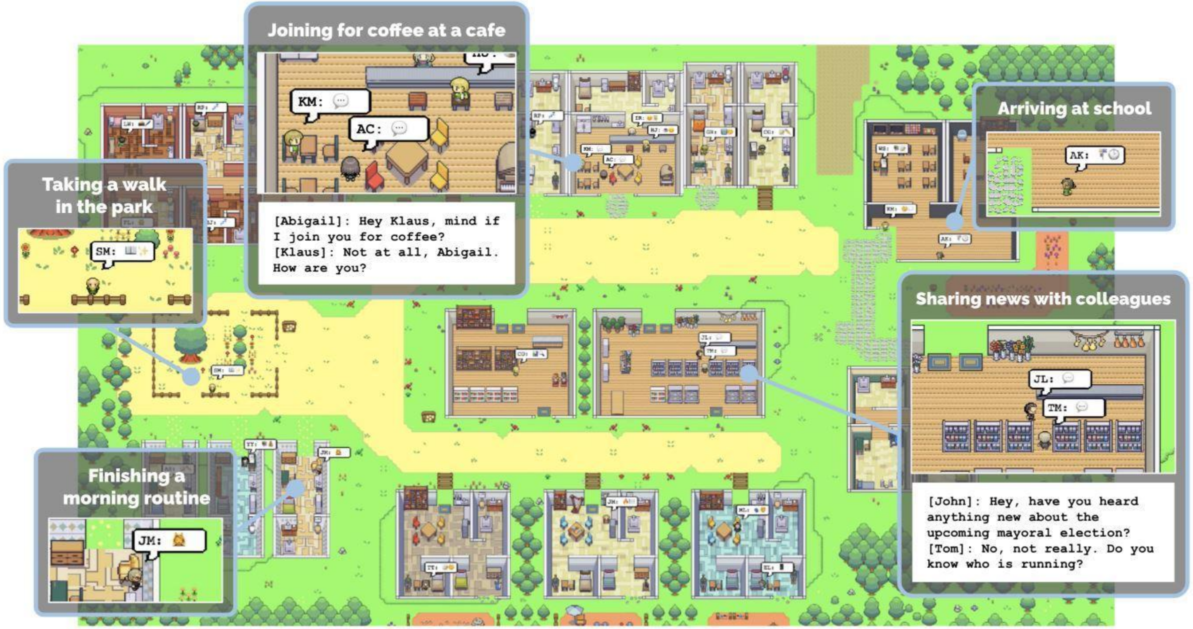

LLM
3. Returns agent instructions along with optional tool use
2. Input question along with agent prompt
7. Respond
User
Agent
4. Retrieve additional information using available tools
Tool
1. Ask question
5. Send additional context
6. Return final answer
LLM

Agent steps:

1. User asks question
2. Question is send to an LLM along with the Agent prompt
3. LLM responds with further instructions either to immediately answer the user or use tools for additional information
4. Retrieve additional information
5 & 6. LLM constructs a final answer based on additional context

[Integrating Neo4j into the LangChain ecosystem]()

Agents Simulating Human Behavior



Joining for coffee at a cafe
KM:
AC:
Taking a walk in the park
SM:
Arriving at school
AK:
Sharing news with colleagues
JL:
TM:
Finishing a morning routine
JM:

[Abigail]: Hey Klaus, mind if I join you for coffee?
[Klaus]: Not at all, Abigail. How are you?

[John]: Hey, have you heard anything new about the upcoming mayoral election?
[Tom]: No, not really. Do you know who is running?

Figure 1: Generative agents create believable simulacra of human behavior for interactive applications. In this work, we demonstrate generative agents by populating a sandbox environment, reminiscent of The Sims, with twenty-five agents. Users can observe and intervene as agents they plan their days, share news, form relationships, and coordinate group activities.

Generative Agents: Interactive Simulacra of Human Behavior

Let's Call Other AI Models

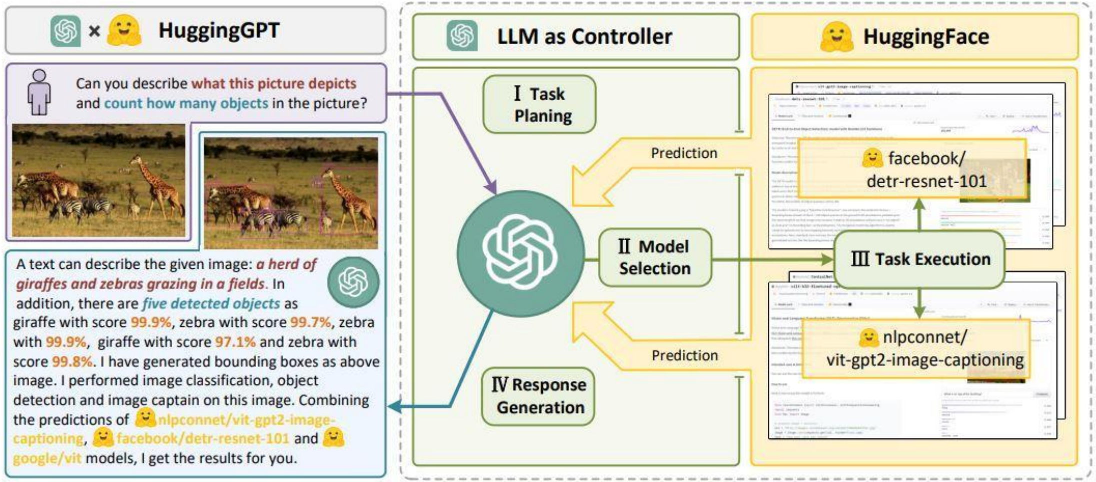
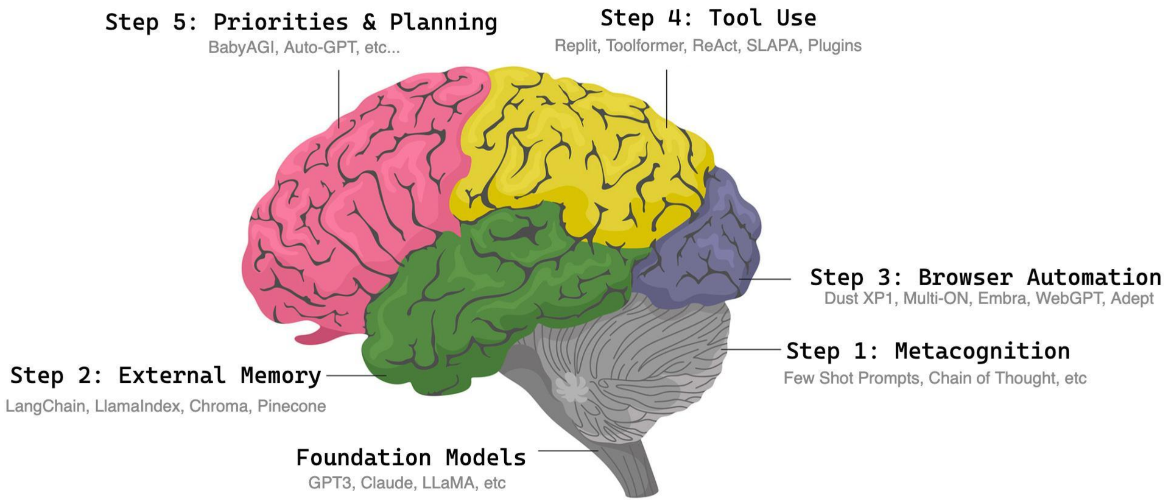

× HuggingGPT
Can you describe what this picture depicts and count how many objects in the picture?
A text can describe the given image: a herd of giraffes and zebras grazing in a fields. In addition, there are five detected objects as giraffe with score 99.9%, zebra with score 99.7%, zebra with 99.9%, giraffe with score 97.1% and zebra with score 99.8%. I have generated bounding boxes as above image. I performed image classification, object detection and image captain on this image. Combining the predictions of nlpconnet/vit-gpt2-image-captioning, facebook/detr-resnet-101 and google/vit models, I get the results for you.
LLM as Controller
HuggingFace
I Task Planing
Prediction
facebook/ detr-resnet-101
II Model Selection
III Task Execution
Prediction
nlpconnet/ vit-gpt2-image-captioning
IV Response Generation

Figure 1: *Language serves as an interface for LLMs (e.g., ChatGPT) to connect numerous AI models (e.g., those in Hugging Face) for solving complicated AI tasks.* In this concept, an LLM acts as a controller, managing and organizing the cooperation of expert models. The LLM first plans a list of tasks based on the user request and then assigns expert models to each task. After the experts execute the tasks, the LLM collects the results and responds to the user.

[HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace]

Roadmap to Autonomy

The Anatomy of Autonomy


Step 5: Priorities & Planning
BabyAGI, Auto-GPT, etc...
Step 4: Tool Use
Replit, Toolformer, ReAct, SLAPA, Plugins
Step 3: Browser Automation
Dust XP1, Multi-ON, Embra, WebGPT, Adept
Step 1: Metacognition
Few Shot Prompts, Chain of Thought, etc
Step 2: External Memory
LangChain, LlamaIndex, Chroma, Pinecone
Foundation Models
GPT3, Claude, LLaMA, etc

The Anatomy of Autonomy: Why Agents are the next AI Killer App after ChatGPT

Educational Resources

Some educational starting points

- Understanding Large Language Models, https://substack.com/inbox/post/115060492 by Sebastian Raschka
- A minimal PyTorch GPT implementation, https://github.com/karpathy/minGPT by Andrej Karpathy
- Annotated PyTorch Paper Implementations, https://nn.labml.ai/index.html by labml.ai

Thank you for your interest!

Mladen Ferležir

Lead Data Scientist | Co-founder

mladen.fernezir@velebit.ai | velebit.ai

Velebit AI LLC