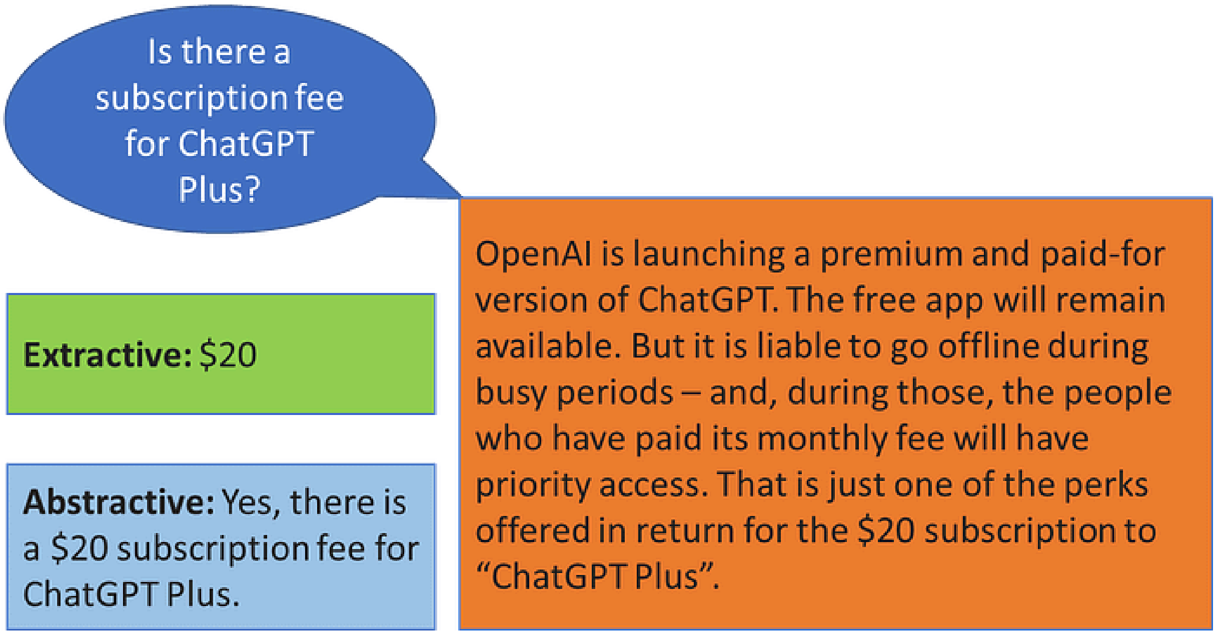
< Go to the original



Is there a subscription fee for ChatGPT Plus?

Extractive: $20

Abstractive: Yes, there is a $20 subscription fee for ChatGPT Plus.

OpenAI is launching a premium and paid-for version of ChatGPT. The free app will remain available. But it is liable to go offline during busy periods – and, during those, the people who have paid its monthly fee will have priority access. That is just one of the perks offered in return for the $20 subscription to "ChatGPT Plus".

Extractive vs Generative Q&A — Which is better for your business?

The arrival of ChatGPT hints at a new era of search engines, this tutorial dives into the 2 basic types of AI based question answering

Skanda Vivek

Follow

tds Towards Data Science a11y-light ~6 min read · February 6, 2023 (Updated: April 13, 2023) · Free: No

Transformer models introduced in 2017 have led to a breakthrough in solving hard language related tasks. Variations of the original transformer architecture in models like BERT, GPT, etc. trained on large amounts of text data have produced state of the art results on language related tasks.

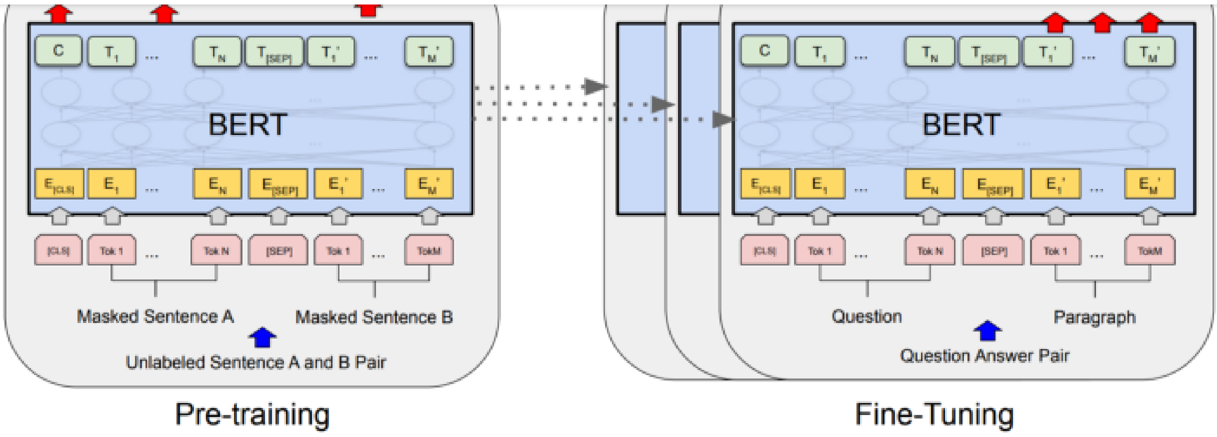
faster and at a fraction of the cost. I believe this will revolutionize industries in the coming decade.

A typical task is extracting information from text. Question Answering is a powerful information extraction tool, whereby models can be trained to extract specific bits of information through complex queries. Think about the potential time and money saved by AI models in answering hard questions from legal documents, instead of asking an experienced lawyer or hiring an intern to pour over the document for hours. Let's take a dive into the 2 basic types of AI based QA: Extractive vs Abstractive.

Extractive QA

The BERT transformer model was released in 2019 by the Google Language team. BERT was trained on unlabeled text data by masking words and training the model to predict masked words based on context. This masked word prediction is a common test, administered to gauge language proficiency.

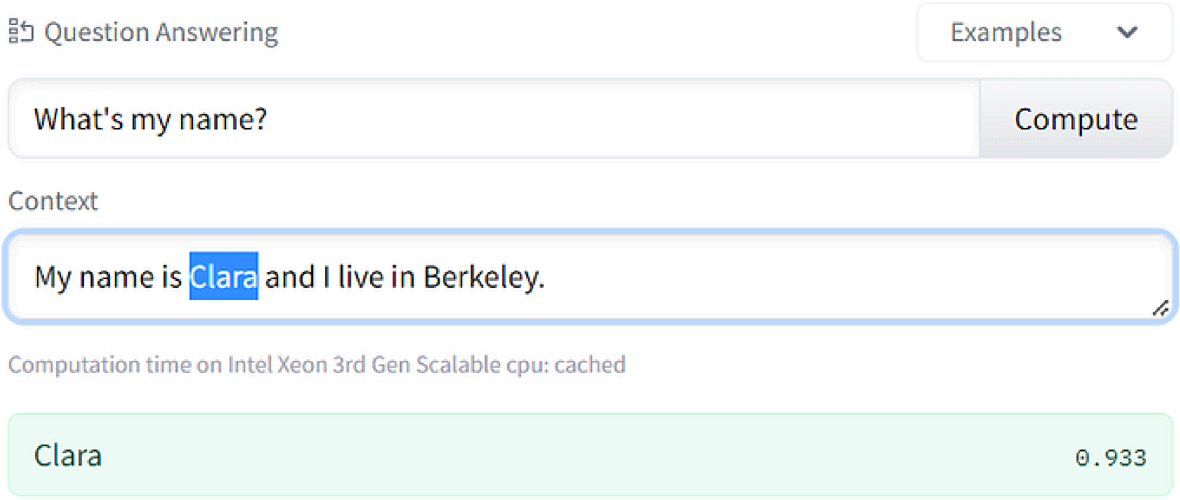
After training the model, BERT was later fine-tuned on multiple tasks. In particular, BERT was fine-tuned on hundreds of thousands of question answer pairs from the SQUAD dataset, consisting of questions posed on Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding passage.



BERT Transformer Architecture from https://arxiv.org/abs/1810.04805

The architecture of BERT and BERT-like models compose one-half of the original transformer architecture proposed in the 2017 paper, known as the encoder. In this model, ***E*** denotes the token embeddings wherein the original sentence of length *M* is converted to a length *M'* (BERT used the WordPiece embeddings). The final hidden vector *T* can be used to predict which part of the text represents the start of the answer and the end of the answer using a softmax.

RoBERTa is a variation of BERT that modified key hyperparameters during training and improved overall performance. Let's look at the output of a fine-tuned RoBERTa model on huggingface released by deepset. As you can see below, in extractive QA the answering you are limited to text contained within the original context:



What's my name?

Context

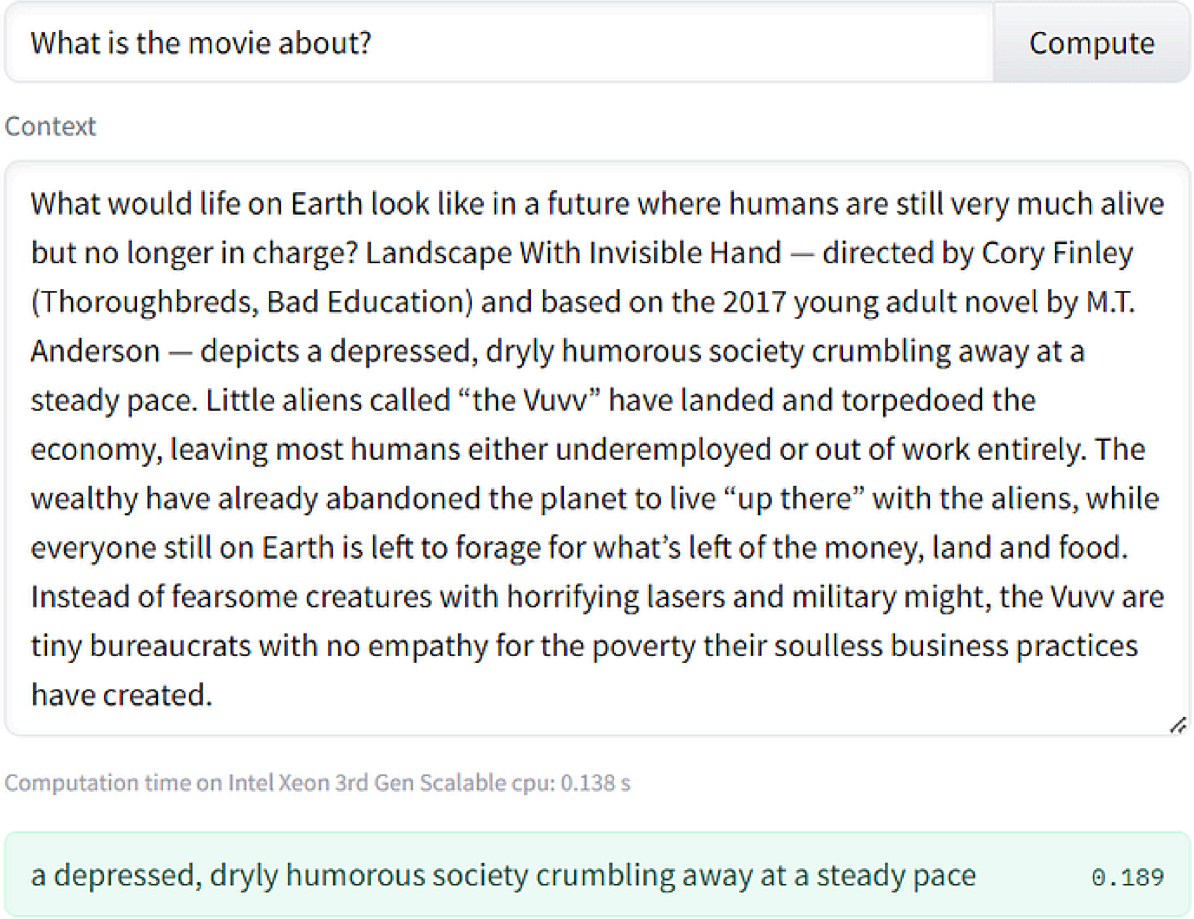
My name is Clara and I live in Berkeley.

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

Clara 0.933

RoBERTa fine-tuned QA model output

However, the answer is not always the best. As you can see below, for a movie review the answer I would have chosen would have been "*What would life on Earth look like in a future where humans are still very much alive but no longer in charge*"



What is the movie about?

Context

What would life on Earth look like in a future where humans are still very much alive but no longer in charge? Landscape With Invisible Hand — directed by Cory Finley (Thoroughbreds, Bad Education) and based on the 2017 young adult novel by M.T. Anderson — depicts a depressed, dryly humorous society crumbling away at a steady pace. Little aliens called "the Vuvv" have landed and torpedoed the economy, leaving most humans either underemployed or out of work entirely. The wealthy have already abandoned the planet to live "up there" with the aliens, while everyone still on Earth is left to forage for what's left of the money, land and food. Instead of fearsome creatures with horrifying lasers and military might, the Vuvv are tiny bureaucrats with no empathy for the poverty their soulless business practices have created.

Computation time on Intel Xeon 3rd Gen Scalable cpu: 0.138 s

a depressed, dryly humorous society crumbling away at a steady pace 0.189

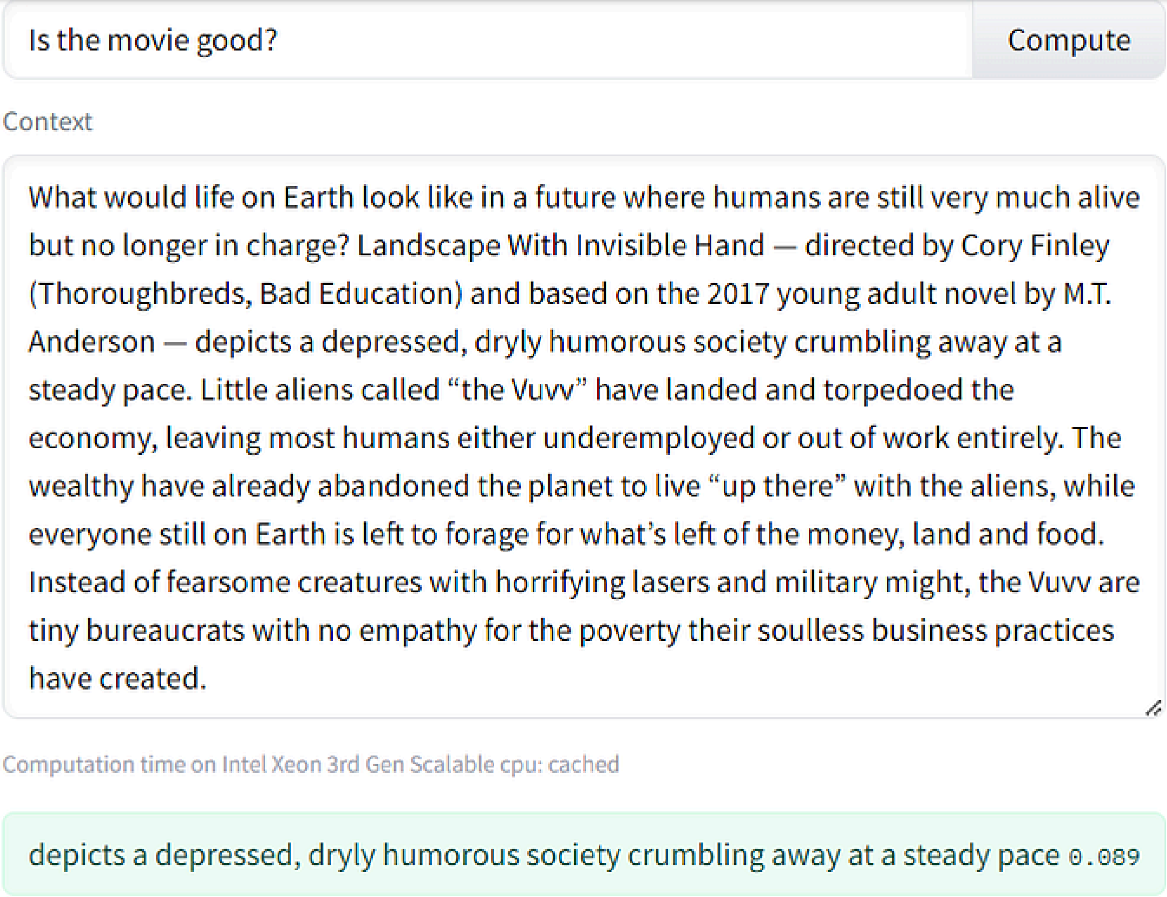
The solution to getting more relevant results is fine-tuning. In the article below, I have discussed how to fine-tune extractive question answering models on the HuggingFace hub using custom data. Fine-tuning based on just a few thousand examples can vastly improve performance, sometimes by **more than 50%.**

Fine-Tune Transformer Models For Question Answering On Custom Data

A tutorial on fine-tuning the Hugging Face RoBERTa QA Model on custom data and obtaining significant performance boosts

towardsdatascience.com

However, extractive QA does not do so well in cases where the answer is not explicitly in the context like below.



Is the movie good?

What would life on Earth look like in a future where humans are still very much alive but no longer in charge? Landscape With Invisible Hand — directed by Cory Finley (Thoroughbreds, Bad Education) and based on the 2017 young adult novel by M.T. Anderson — depicts a depressed, dryly humorous society crumbling away at a steady pace. Little aliens called “the Vuvv” have landed and torpedoed the economy, leaving most humans either underemployed or out of work entirely. The wealthy have already abandoned the planet to live “up there” with the aliens, while everyone still on Earth is left to forage for what’s left of the money, land and food. Instead of fearsome creatures with horrifying lasers and military might, the Vuvv are tiny bureaucrats with no empathy for the poverty their soulless business practices have created.

Computation time on Intel Xeon 3rd Gen Scalable cpu: cached

depicts a depressed, dryly humorous society crumbling away at a steady pace 0.089

Model yielding useless results when the answer is not explicitly present

This issue can be circumvented by appending "ANSWERNOTFOUND" and fine-tuning on these cases so that the model does not yield an answer when it is unsure.

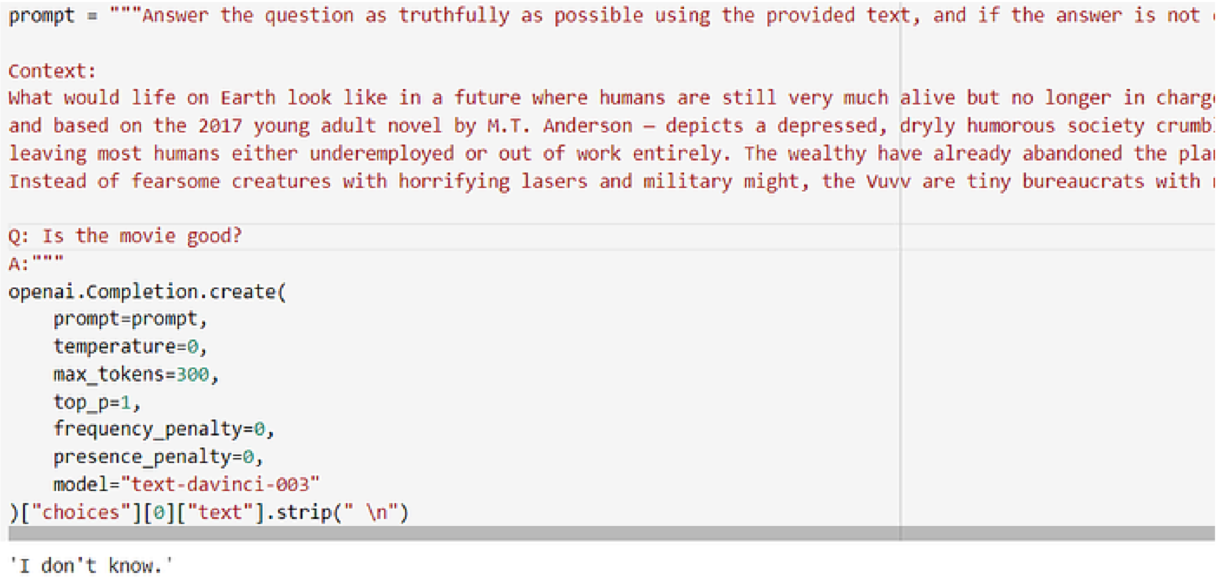
Abstractive QA

While ChatGPT has taken the whole world by storm recently, the original GPT model was released before BERT. GPT models use the decoder layer of the original 2017 Transformer. GPT models are trained to predict the next word in a sequence in an unsupervised manner. Next, they are fine-tuned in a supervised fashion. For QA, GPT models are presented during fine-tuning with multiple answer choices across numerous examples, and they are trained to pick the right choice. One important difference at inference is that GPT

Currently, OpenAI has 4 major language models that they offer API access to:

1. Ada ($0.0004 / 1K tokens — Fastest)
2. Babbage ($0.0005 / 1K tokens)
3. Curie ($0.0020 / 1K tokens)
4. Davinci ($0.0200 / 1K tokens — Most powerful)

For reference, 1K tokens is basically 750 words that you send in to the API to process. So let's see how this model does for similar questions:



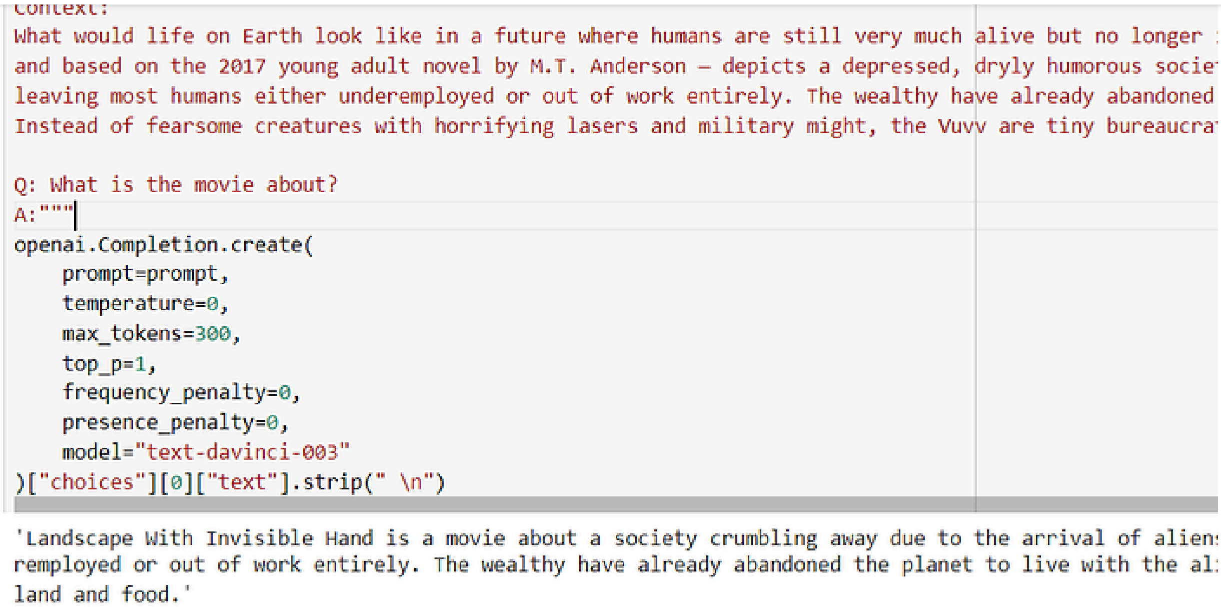
```
prompt = """Answer the question as truthfully as possible using the provided text, and if the answer is not

Context:
What would life on Earth look like in a future where humans are still very much alive but no longer in charg
and based on the 2017 young adult novel by M.T. Anderson – depicts a depressed, dryly humorous society crumb
leaving most humans either underemployed or out of work entirely. The wealthy have already abandoned the pla
Instead of fearsome creatures with horrifying lasers and military might, the Vuvv are tiny bureaucrats with

Q: Is the movie good?
A:"""
openai.Completion.create(
    prompt=prompt,
    temperature=0,
    max_tokens=300,
    top_p=1,
    frequency_penalty=0,
    presence_penalty=0,
    model="text-davinci-003"
)["choices"][0]["text"].strip(" \n")
```

```
'I don't know.'
```

Davinci OpenAI Model based on GPT3 for QA



```
context:
What would life on Earth look like in a future where humans are still very much alive but no longer
and based on the 2017 young adult novel by M.T. Anderson – depicts a depressed, dryly humorous socie
leaving most humans either underemployed or out of work entirely. The wealthy have already abandoned
Instead of fearsome creatures with horrifying lasers and military might, the Vuvv are tiny bureaucra

Q: What is the movie about?
A:"""
openai.Completion.create(
    prompt=prompt,
    temperature=0,
    max_tokens=300,
    top_p=1,
    frequency_penalty=0,
    presence_penalty=0,
    model="text-davinci-003"
)["choices"][0]["text"].strip(" \n")
```

```
'Landscape With Invisible Hand is a movie about a society crumbling away due to the arrival of alien
remployed or out of work entirely. The wealthy have already abandoned the planet to live with the al
land and food.'
```

Davinci OpenAI Model based on GPT3 for QA

As you can see, the Davinci model does pretty well in summarizing movie plots as well as saying "I don't know" when the answer is not clearly in the context.

Which Model is Better — Abstractive or Extractive??

You might be tempted to say that OpenAI's abstractive QA is clearly superior to extractive QA models. However, that is where the business case matters. I'll break it down below:

Cost

The Davinci model is clearly more expensive, at a large enough scale. It amounts to 0.02$ per 1K tokens which might as well be 0.02$ for 1–10 queries. Whereas hosting a model from Hugging Face on AWS might amount to a fraction of the cost, at 0.5 cents to 1$ per hour running thousands or more queries every hour.

Output

might not be satisfied by dry extractive answers that paraphrase the text. However, if you are doing post processing on the answers obtained — say storing numbers in a database, abstractive QA might be a hindrance as you need to use additional logic to strip out extra words.

Customizability

OpenAI API usage requires reliance on OpenAI servers. While they do make it possible to fine-tune their models on custom data, it is not possible to host these models on separate infrastructures like AWS. But you can take open-source models on Hugging Face and create APIs on AWS, and not have to rely any more on Hugging Face for model serving. This is powerful in that it allows companies to keep all the infrastructure in-house and rely only on cloud providers like AWS.

One thing I would like to point out is that Hugging Face does also support abstractive QA models. In fact, they released a text2text generation model Flan T5 on the model hub recently. But I have noticed that this model does not perform as well on QA tasks as the Davinci GPT-3 model. Very soon, I expect Hugging Face to also host open-source fine-tuned models like the Davinci GPT-3 model.

I hope this article was a useful walkthrough in using AI for question answering. In conjunction with existing methods for information retrieval and searching through large amounts of data, AI based information extraction can help extract needles from haystacks, and greatly improve efficiency in extracting essential details from large amounts of data, previously possible only through human comprehension.

custom text is now live! Answer domain specific questions 3 easy steps!

1. ***Upload a URL or paste a text and hit the search button***
2. ***Ask a question specific to the context and hit query***
3. ***Get your answer!***

Feel free to use and let me know your feedback!

If you are not yet a Medium member and want to support writers like me, feel free to sign-up through my referral link: https://skanda-vivek.medium.com/membership

For weekly data-based perspectives subscribe here!

#data-science #machine-learning #chatgpt #business #tutorial