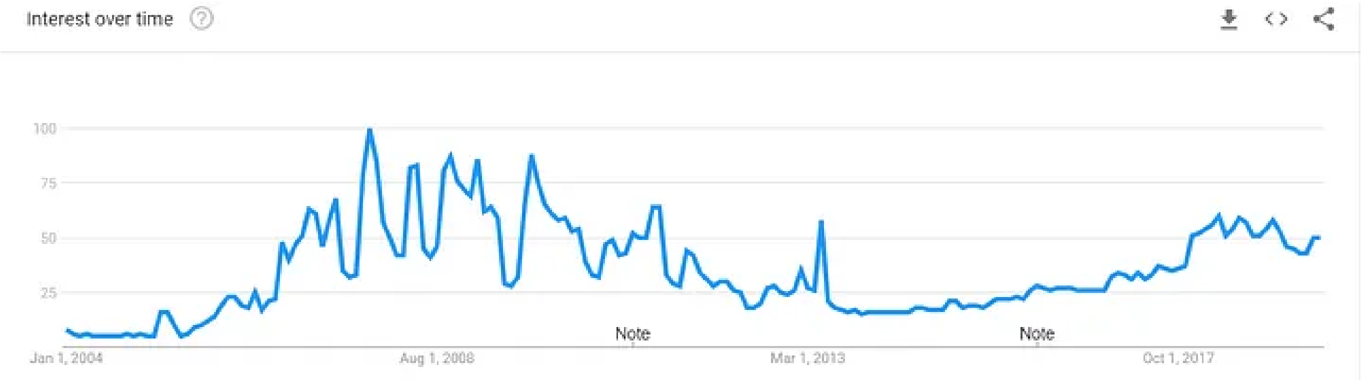
Can Michael Scott be recreated with AI?

Eugene Olkhov · Follow

Published in CompassRed Data Blog · 5 min read · Jul 30, 2019

The Office has been a popular show for over a decade. The show had high popularity during its 9 seasons on air, and a new wave of popularity hit after it stopped airing thanks to Netflix.

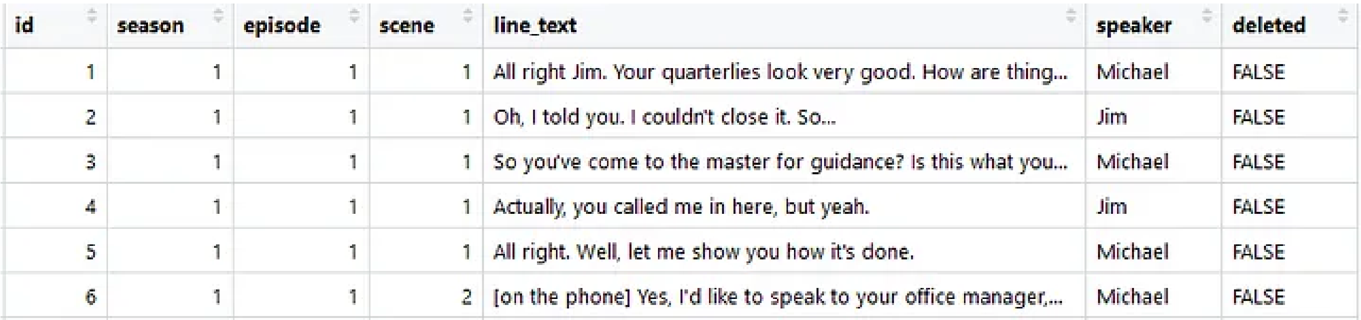


Michael Scott has been a unique and loved TV character in many people's hearts. Michael Scott was different — and when he

left towards the end of the series' run, people were devastated that one of their favorite characters was no longer on the show and led to an emotional reaction from many fans. It is because of how iconic Michael Scott is, that I figured he would be an excellent candidate for being emulated by AI. More specifically, I want to know if an AI model can generate Michael Scott-esque written speech.

Before jumping right in to the AI work, let's do some exploration of how Michael speaks and how his language has changed over the course of the seven seasons of The Office he was in.

I was able to obtain data with every spoken line from The Office on Reddit, provided by u/misunderstoodpoetry. I imported the data into R for the exploration. The first few rows of the data are as follows:



id	season	episode	scene	line_text	speaker	deleted
1	1	1	1	All right Jim. Your quarterlies look very good. How are thing...	Michael	FALSE
2	1	1	1	Oh, I told you. I couldn't close it. So...	Jim	FALSE
3	1	1	1	So you've come to the master for guidance? Is this what you...	Michael	FALSE
4	1	1	1	Actually, you called me in here, but yeah.	Jim	FALSE
5	1	1	1	All right. Well, let me show you how it's done.	Michael	FALSE
6	1	1	2	[on the phone] Yes, I'd like to speak to your office manager,...	Michael	FALSE

Cleaning and shaping

First step of every data science project is to clean the data. In this case, minimal cleaning was needed. I removed a few of the unnecessary columns `id`, `deleted`, and removed text in brackets from the `line_text` column. An example of this is seen in row 6.

```
office_cleaned <-
  office_df %>%
  select(-c(id, deleted)) %>%
  mutate(line_text = str_replace_all(line_text, "\\[.+?\\]", ""))
```

Next, I wanted to collapse all of the lines spoken so that it would be possible to tokenize (shape to one word per line) them later. In case I wanted to look at multiple characters across seasons later, I grouped the data by speaker and season. But, since we're only interested in Michael Scott at the moment, we can filter to lines spoken by only him:

```
office_grouped <-
office_cleaned %>%
  group_by(season, speaker) %>%
  summarise(text = paste(line_text, collapse = " "))

#Filter to just Michael

office_michael <-
  office_grouped %>%
  filter(speaker == "Michael")
```

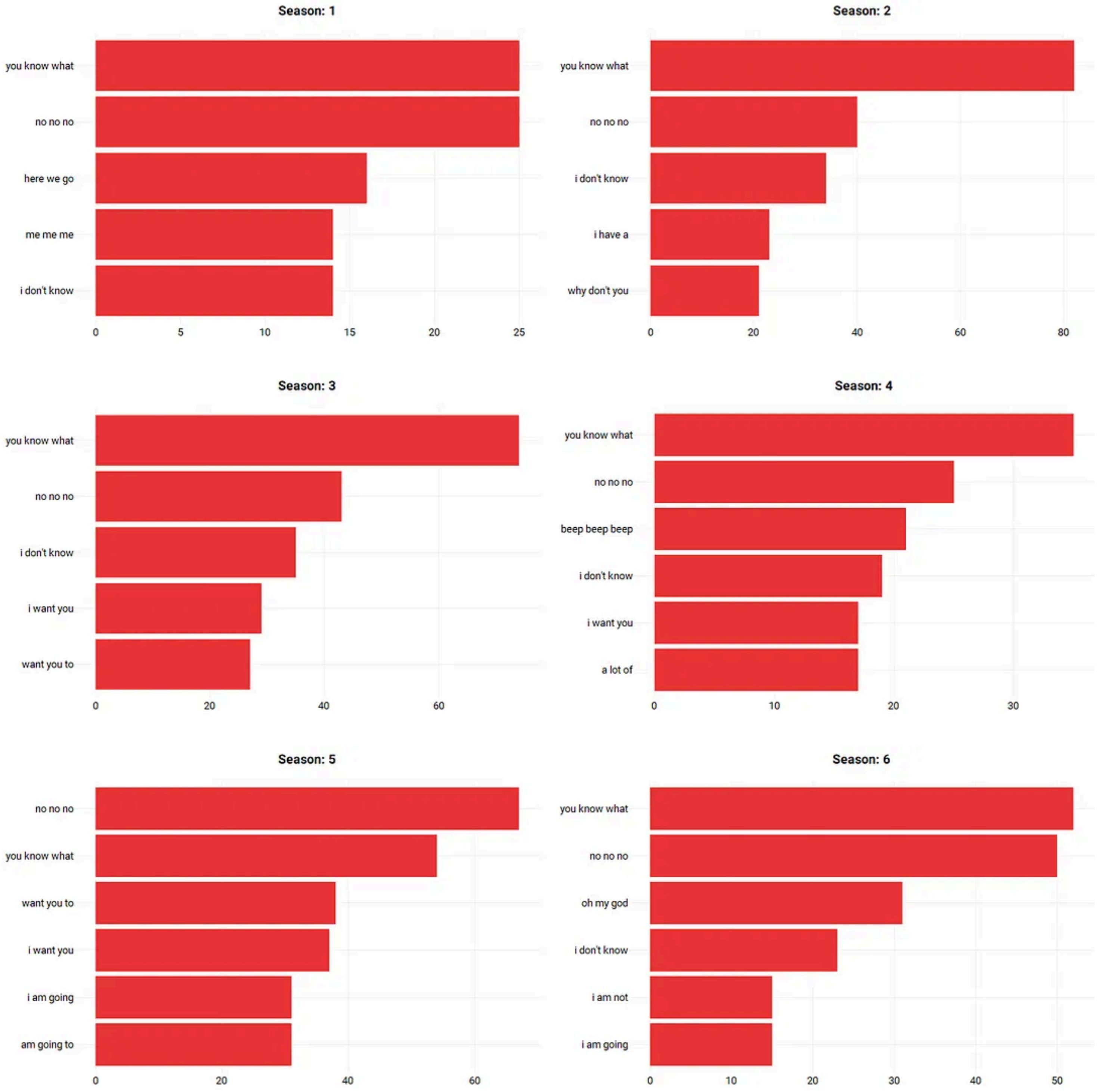
Finally, we can use the `tidytext` package to easily clean the data further, tokenize, and remove stop-words. I also added a few custom stop-words after a seeing that the stop-words list in `tidytext` missed a few meaningless words that Michael Scott says often (e.g., yeah, hey, um). In this case, I decided to use 3-grams instead of looking at single words.

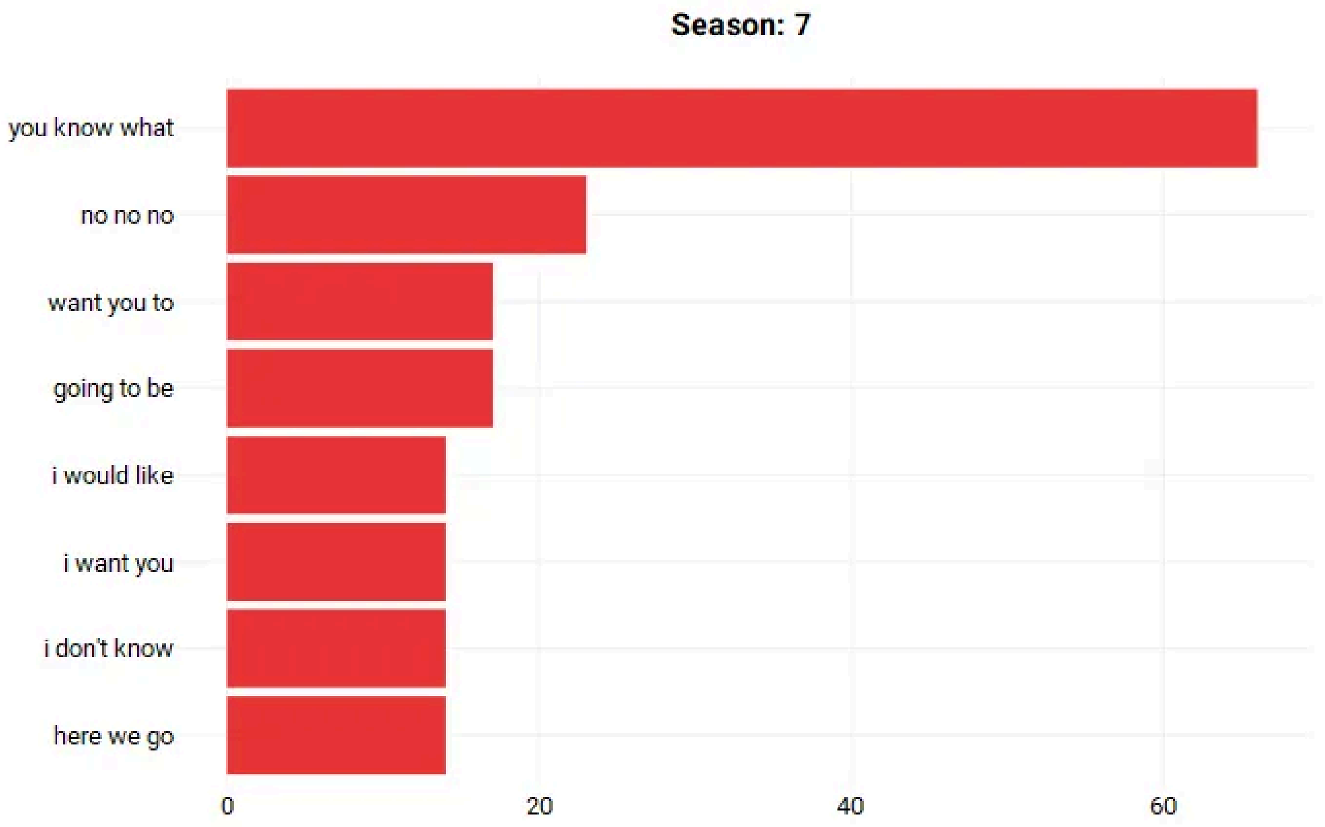
```
michael_tokenized <-
  office_michael %>%
  ungroup() %>%
  unnest_tokens(word, text, token = "ngrams", n = 3) %>%
  filter(!word %in% stop_words_custom$word,
         str_detect(word, "[a-z]"))
```

Michael Scott's evolution across the seasons

So, what does Michael's speech look like across the seasons, and does it change at all?

Short story — not really. Let’s take a look at the top tri-grams spoken by Michael per season:





As we can see, Michael's speech is very consistent throughout his 7 seasons. His top tri-grams consist of talking about himself, asking others to do something (which makes sense given his managerial position). From this, we can see that Michael Scott is quite self-centered.

Text Generation

Now that we have obtained some insights about how Michael Scott speaks, we can take all of his text and feed it into a neural network, and see if the network can output some text that sounds like something our dear Michael Scott would say.

There are three major types of neural networks: Multi-Layer Perceptions (MLP), Convolutional Neural Networks (CNN), and Recurrent Neural

Networks (RNN). For this task, we will be using an RNN because RNNs are able to learn sequences, such as sequences in speech patterns.

Unfortunately, text generation models take a long time to train and requires a system with a GPU to speed up the process. Luckily, I was able to find a shortcut. Google Collab is a great resource that provides a notebook environment with no setup requirements. It also has GPU capabilities which make running these types of models faster. I also found a helpful Python library `textgenrnn` that made training a model and generating text easier.

Thanks to Max Woolf (author of `textgenrnn`) for providing the package as well as some existing code to make my first pass at generating text easier.

Results

So, how did the first pass at training the model and generating text go? The output of the model spit out roughly 2,000 words of what it thought Michael Scott sounds like. Most of it was incoherent (although maybe that is indicative of Michael Scott after all), but there were some interesting parts.

> *Hey Dwight. I'm going to change to see you. I am going to talk to you all today. So it's not going to be a baby. I want to have a secret. I saw it I want you to take a fant for states of the party is the cards and I spent the office. It's my friend and I think I guess this is good. That is good Thank God. Hey hey she is in the way to be done. I don't understand that he is a reality of the car and that you don't have to do this. I need to get to the guys. I think I should have been a warehouse. The Michael Scott Pam is going to be supposed better. All right Does it What the hell isn't the heart with any sense.*

While still mostly incoherent, there definitely are some phrases in there that sound close to something Michael would say.

So, maybe right now AI cannot duplicate Michael Scott just yet. However, given this is just the first pass with a fairly quick model training method, it appears that a more thorough method could provide results closer to what we were looking for.

Machine Learning AI The Office NLP Michael Scott

Published in CompassRed Data Blog

Follow

249 Followers · Last published Dec 8, 2021

We live for data and analytics.

Written by Eugene Olkhov

Follow

28 Followers · 8 Following

Responses (1)

What are your thoughts?

Respond

mrunal bokil
Jul 30, 2019

•••

Great AI model Eugene! It’s actually good start for text mining where you already know what the responses could be.

Reply

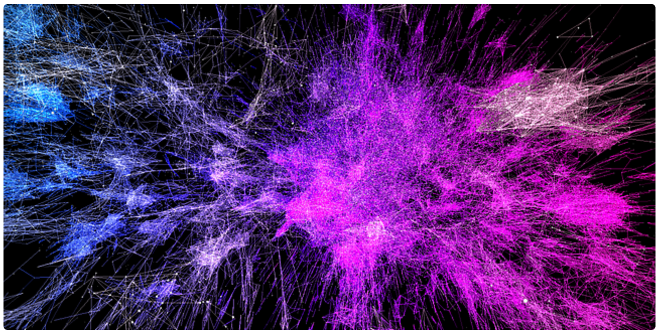
More from Eugene Olkhov and CompassRed Data Blog

In CompassRed Data Blog by Eugene Olkhov

How to Host and Share RMarkdown Files

When I first learned how to use RMarkdown, I was in absolute awe. Moving from...

Jul 22, 2020 👏 74



In CompassRed Data Blog by Aayush Dua

Introduction to Word Embeddings and its Applications

Intro to Word2Vec, Glove, ELMo and fastText

Aug 5, 2020 👏 111 💬 1

In CompassRed Data B... by CompassRed Data L...

How to Create a Click Heat Map for Your Website using Google...

Have you ever seen one of those fancy heat maps that tools like Hotjar and CrazyEgg ca...

Aug 24, 2018 👏 633 💬 19

In CompassRed Data Blog by Eugene Olkhov

Current and Future State of Recommender Systems

In this day and age, we have so many choices when it comes to making online purchases,...

Sep 10, 2019 👏 11

See all from Eugene Olkhov

See all from CompassRed Data Blog

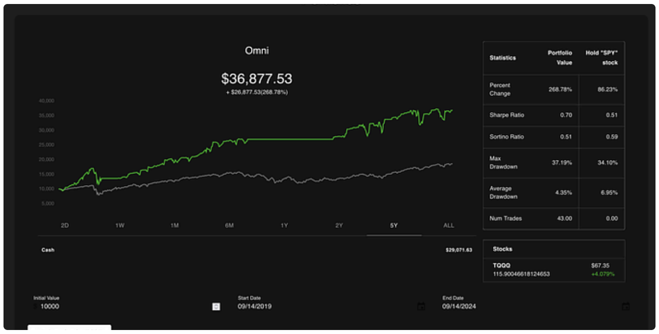
Recommended from Medium

Jessica Stillman

Jeff Bezos Says the 1-Hour Rule Makes Him Smarter. New...

Jeff Bezos's morning routine has long included the one-hour rule. New...

Oct 30, 2024 23K 640



Statistics	Portfolio Value	Hold "SPY" stock
Percent Change	268.78%	86.23%
Sharpe Ratio	0.70	0.51
Sortino Ratio	0.51	0.59
Max Drawdown	37.19%	34.10%
Average Drawdown	4.35%	6.95%
Num Trades	43.00	0.00

In DataDrivenInvestor by Austin Starks

I used OpenAI's o1 model to develop a trading strategy. It is...

It literally took one try. I was shocked.

Sep 15, 2024 8.9K 230

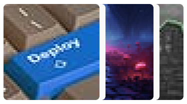
Lists

Natural Language Processing

1943 stories · 1597 saves

The New Chatbots: ChatGPT, Bard, and Beyond

12 stories · 550 saves



Predictive Modeling w/ Python

20 stories · 1833 saves

Practical Guides to Machine Learning

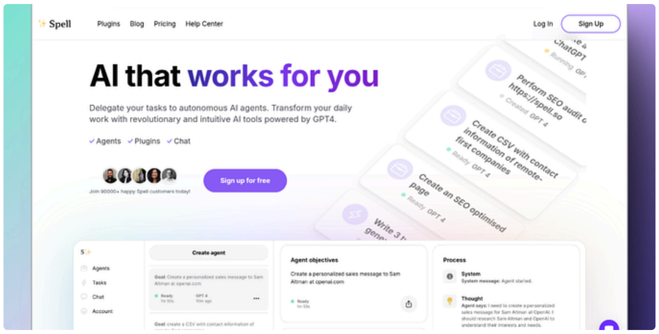
10 stories · 2208 saves

In AI Advances by Manpreet Singh

Goodbye RAG? Gemini 2.0 Flash Have Just Killed It!

Alright!!!

Feb 10 2.1K 79


Spell Plugins Blog Pricing Help Center Log In Sign Up

AI that works for you

Delegate your tasks to autonomous AI agents. Transform your daily work with revolutionary and intuitive AI tools powered by GPT4.

✓ Agents ✓ Plugins ✓ Chat

Join 90000+ happy Spell customers today!

Sign up for free

ChatGPT
Running GPT

Perform SEO audit o https://spell.so
Created GPT 4

Create CSV with contact information of remote-first companies
Ready GPT 4

Create an SEO optimised page
Ready GPT 4

Write 3 t gene

Agents
Tasks
Chat
Account

Create agent

Goal: Create a personalized sales message to Sam Altman at openai.com
Ready 1m 50s GPT 4 10m ago

Goal: create a CSV with contact information of

Agent objectives

Create a personalized sales message to Sam Altman at openai.com

Ready 1m 50s

Process

System
System message: Agent started.

Thought
Agent says: I need to create a personalized sales message for Sam Altman at OpenAI. I should research Sam Altman and OpenAI to understand their interests and needs.

Mohit Vaswani

6 AI Agents That Are So Good, They Feel Illegal

AI agents are the future because they can replace all the manual work with automation...

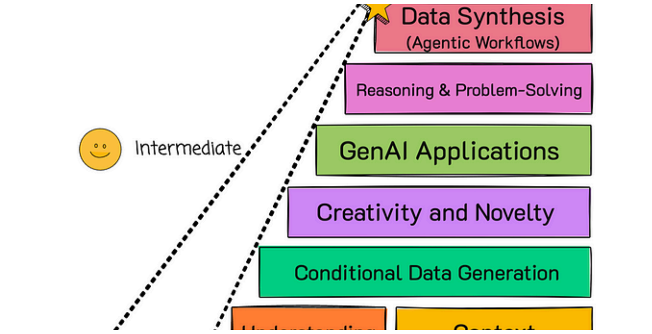
Jan 11 3.9K 129

In Level Up Coding by Jayden Levitt

Warren Buffett Just Sold $133 Billion in Stock: Does He Know...

It's a stock market signal that's becoming hard to ignore.

Feb 5 4.3K 104



Cobus Greyling

Why The Focus Has Shifted from AI Agents to Agentic Workflows

We find ourselves on a stairway from where Large Language Models were introduced to...

Feb 5 877 21

See more recommendations