



CODE Presents:

An Introduction to Large Language Models (LLMs)

June 14th, 2023

Online

Dr. Otto Dobretsberger, PhD

Senior Software Developer

CODE Consulting

Kicking Things Off

Jim Duffy

- Director of Business Development
CODE Magazine & Consulting
- Former Developer – Drawn to the Dark Side: Now Responsible for
Marketing & Business Development
- jduffy@codemag.com / [My Bio](#)
- International Author and Speaker
- Former Microsoft RD (Regional Director) 9 years
- Former 11-time Microsoft Most Valuable Professional (MVP)
- Twitter: @jmduffy





"Helping People Build
Better Software"



About CODE Consulting



- Custom Software Development
- Web, Cloud, Mobile, Desktop, Serverless, Databases,...
- Copilot Development, AI, GPT & Azure OpenAI
- Application Security Testing
- User Interface and Interaction Design
- Project Rescue, App Modernization (VB, VFP, Access, etc.)
- Support & Maintenance for existing applications



AI Consulting Services



CODE
SKUNK WORKS



- We can help with your AI needs
- What does AI mean for you?
- Executive Briefings and Analysis
- AI Implementations
- Business Modernization
- "Skunk Works" Projects
- www.codemag.com/AI
- www.codemag.com/SkunkWorks

About CODE Staffing

- Disrupting the world of staffing!
- Giving our customers the ability to have staff on par with Silicon Valley companies...
- ...and our employees a work environment in a bleeding-edge tech company with the **industry leading benefits!**
- www.codestaffing.com



We Need You!

- We are currently “hiring on all cylinders” across the entire CODE Group Organization!
- Check out www.codestaffing.com/careers



Upcoming CODE Group Events



- Training Classes

- Introduction to Development with Open AI, ChatGPT, Azure Open AI, and more... (June 21)
- Secure Coding for C# Developers (Sept 5 & 6)



<https://www.codemag.com/training>

Free Subscription to CODE Magazine!



- The leading software development magazine written by expert developers for developers.
- All registered attendees will automatically receive a free digital subscription to CODE Magazine – no need to do anything, it'll happen auto-magically.
- Subscribers get CODE Focus issues free of charge!
- Please share this free subscription link:



<https://tinyurl.com/cp062023>

CODE Mobile App

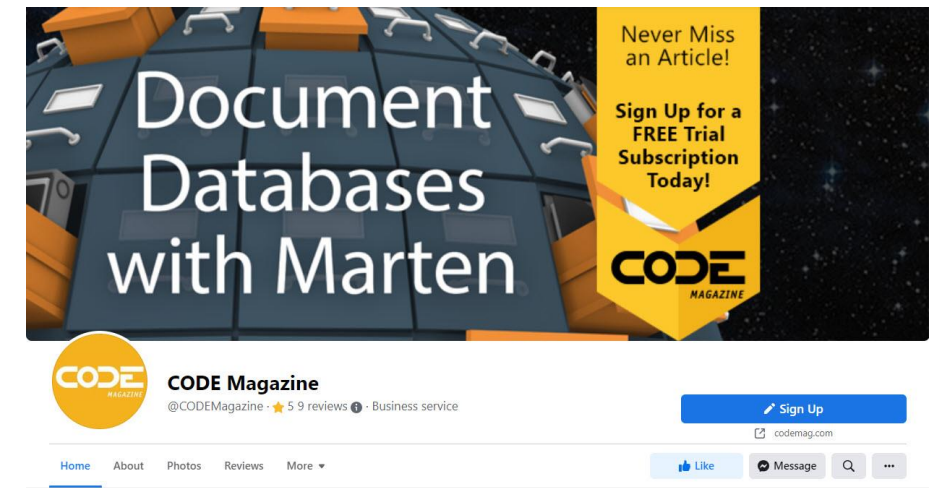
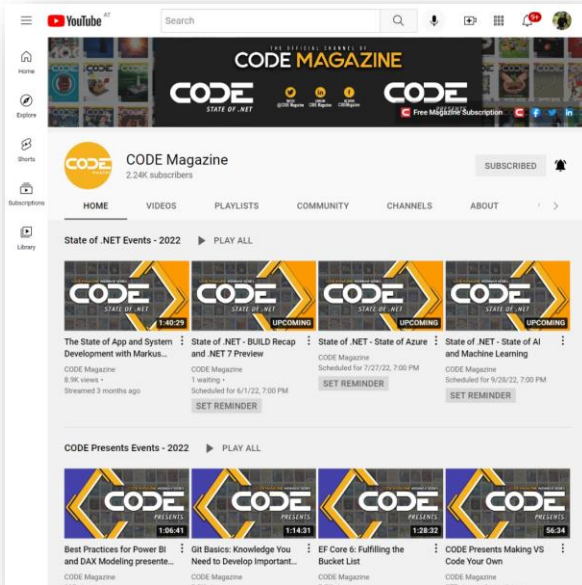


- Check out the new CODE Magazine Mobile application!
- Available for iOS & Android



Join Us On Social Media





- YouTube channel: <https://tinyurl.com/CODEYTC>
- Twitter: <https://twitter.com/CODEmagazine>
- Facebook: www.facebook.com/CODEMagazine/
- LinkedIn: <https://www.linkedin.com/company/code-magazine/>



Recordings & Slide Decks

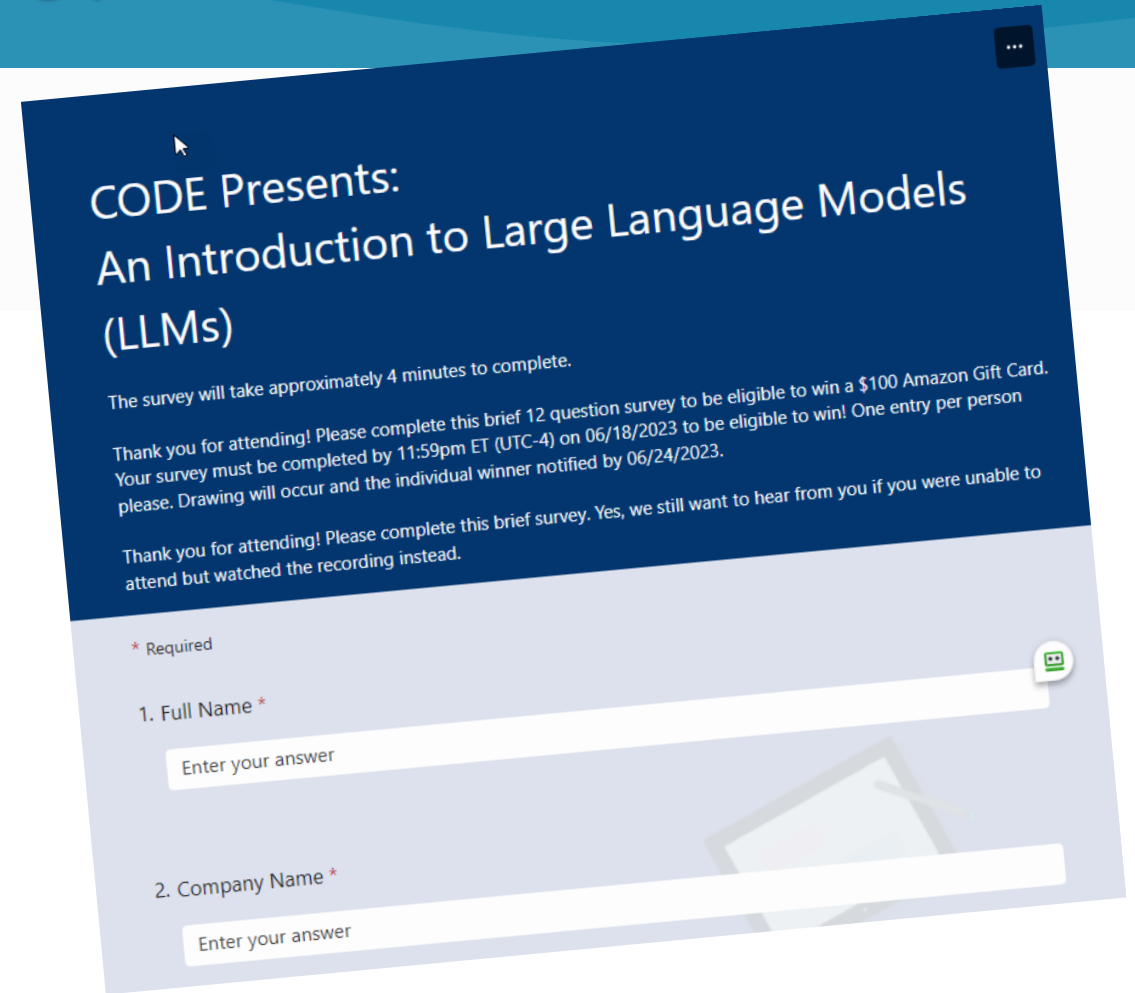
- State of .NET Webinar Series
 - codemag.com/StateOfDotNet
- CODE Presents Webinar Series
 - codemag.com/CODEPresents

Recent Event Recordings

State of .NET - The State of .NET App and System Development 2/23/2022 	State of .NET - .NET Conf Recap 11/17/2021 
State of .NET - .NET 6 Preview 10/27/2021 	State of .NET - Azure Architecture 9/29/2021 

Event Survey – Win \$100!

- Complete this very short 12 question survey for a chance at a \$100 Amazon Gift Card!
- Survey must be completed by 11:59pm ET on Friday 6/16/2023 to be eligible!
- THIS SLIDE WILL BE REPEATED AT THE END AND SURVEY LINK REPEATED IN THE CHAT WINDOW!



CODE Presents:
An Introduction to Large Language Models (LLMs)

The survey will take approximately 4 minutes to complete.

Thank you for attending! Please complete this brief 12 question survey to be eligible to win a \$100 Amazon Gift Card. Your survey must be completed by 11:59pm ET (UTC-4) on 06/18/2023 to be eligible to win! One entry per person please. Drawing will occur and the individual winner notified by 06/24/2023.

Thank you for attending! Please complete this brief survey. Yes, we still want to hear from you if you were unable to attend but watched the recording instead.

* Required

1. Full Name *

Enter your answer

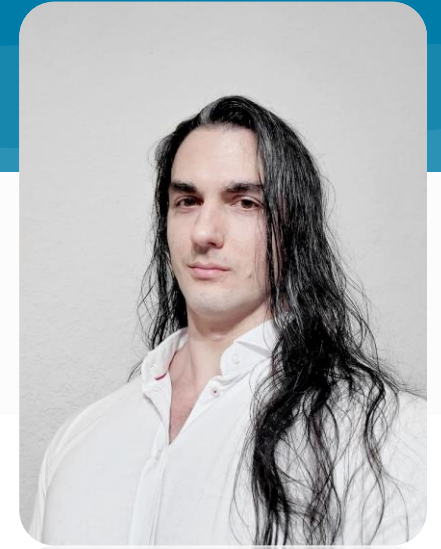
2. Company Name *

Enter your answer

<https://tinyurl.com/cp062023survey>

About the Presenter

- **Dr. Otto Dobretsberger**
- Senior Software Developer
- Founding Photino Team Member
- Professor at University of Houston
- PhD in Computer Science
- [Email: ottod@codemag.com](mailto:ottod@codemag.com)
- <https://www.codemag.com/People/Bio/Otto.Dobretsberger>
- ...and able to deadlift 700lbs+
(Texas State record holder)



Large Language Models

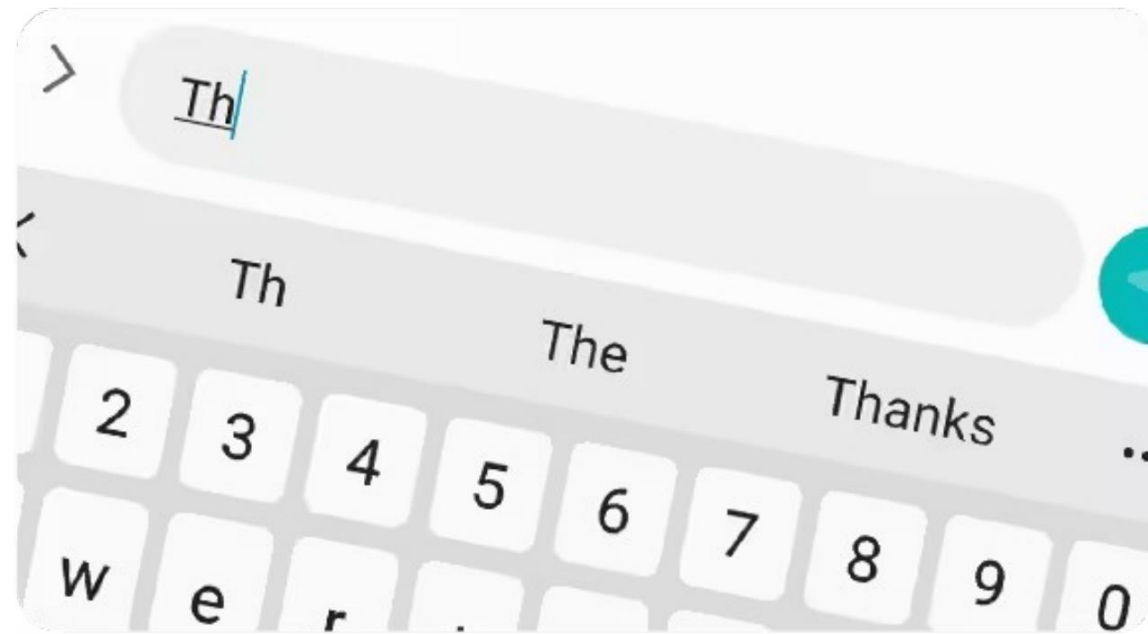
Otto Dobretsberger, PhD

Sr. Software Dev

CODE Magazine, CODE Consulting, CODE Staffing

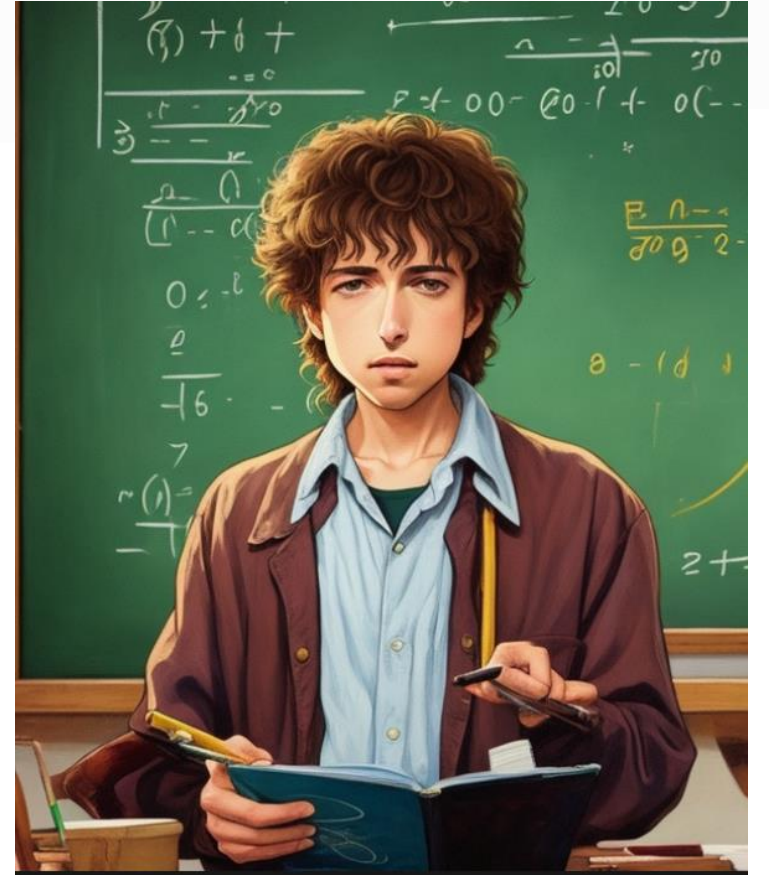
Large Language Models

LLMs based on frequencies

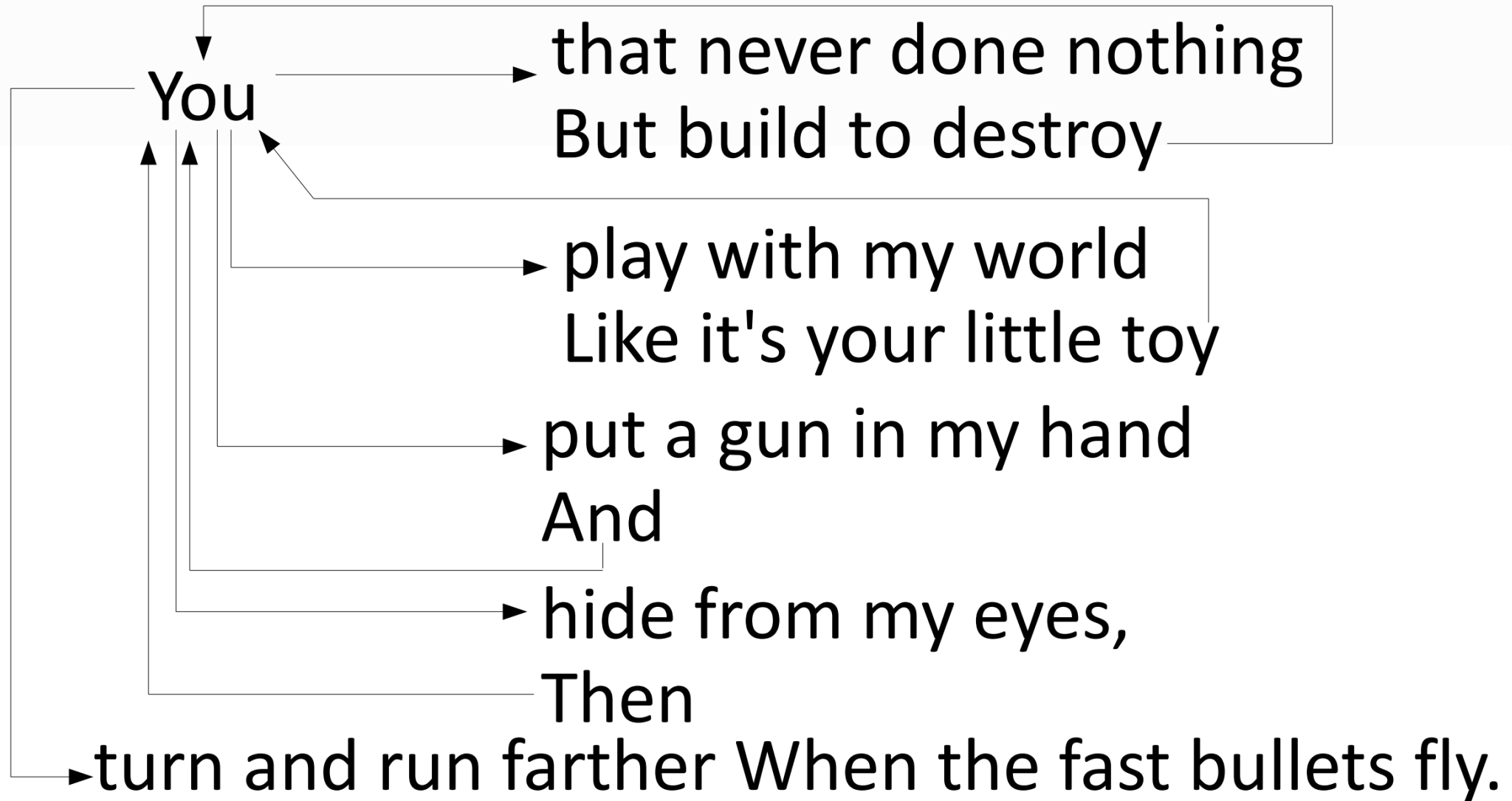


Large Language Models with Bob Dylan

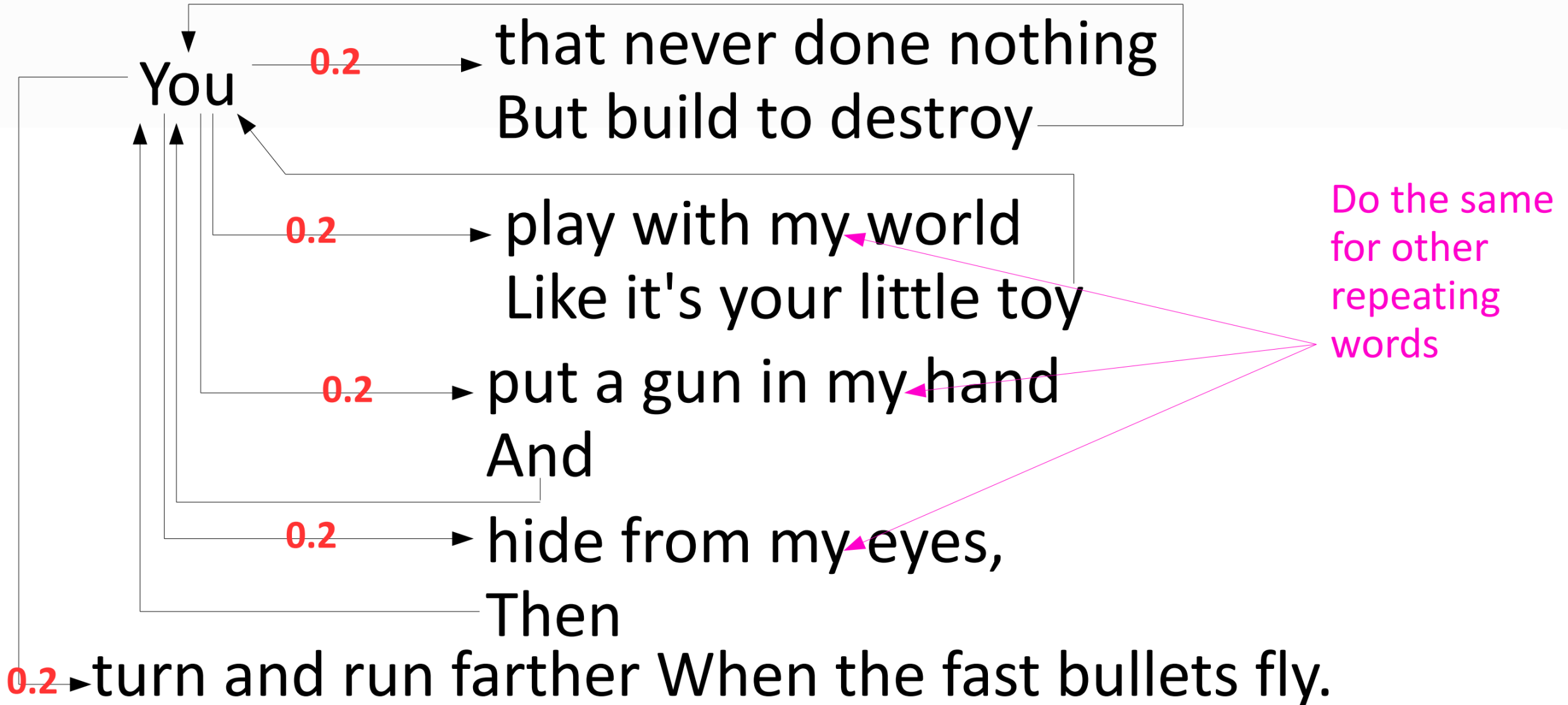
You that never done nothing
But build to destroy,
You play with my world
Like it's your little toy.
You put a gun in my hand
And **you** hide from my eyes,
Then **you** turn and run farther
When the fast bullets fly.



Large Language Models



Large Language Models – Edge Probabilities



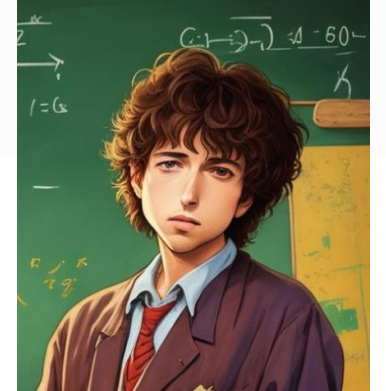
Large Language Models as Generative AI

Use this model to generate Bob Dylan songs:

You play with my eyes,
Then you play with my world
Like it's your little toy.
You that never done nothing
But build to destroy,
you hide from my hand
And you hide from my hand
And you hide from my hand
And you hide from my hand
And you hide from my hand
And you hide from my ...etc...

Somewhat ok

Garbage



ORIGINAL

You that never done nothing
But build to destroy,
You play with my world
Like it's your little toy.
You put a gun in my hand
And you hide from my eyes,
Then you turn and run farther
When the fast bullets fly.

Improve Large Language Models!

Train it on more Text. How much more? All of it!

"Blowin' In The Wind"		
How many walk down Before I How many dove shot Before I sand? Yes, and must tell Before banned	"She's No Good"	
	Well I you like Nobod along You ge sleepin I come would Well s sweet When of jum Well y guess An' yo wanna I help no sho mama I help no foc	"Talkin' New York"
	Rambl Leavin best Thoug and d 'Til I c town People groun Buildi	"Highway 51 Blues"
		Highway 51 runs right by my baby's door Highway 51 runs right by my baby's door If I don't get the girl I'm loving Won't go down to Highway 51 no more
		Well, I know that highway like I know my hand Yes, I know that highway like I know the back of my hand Running from up Wisconsin way down to no man's land Well, if I should die before my time should come And if I should die before my

Etc.

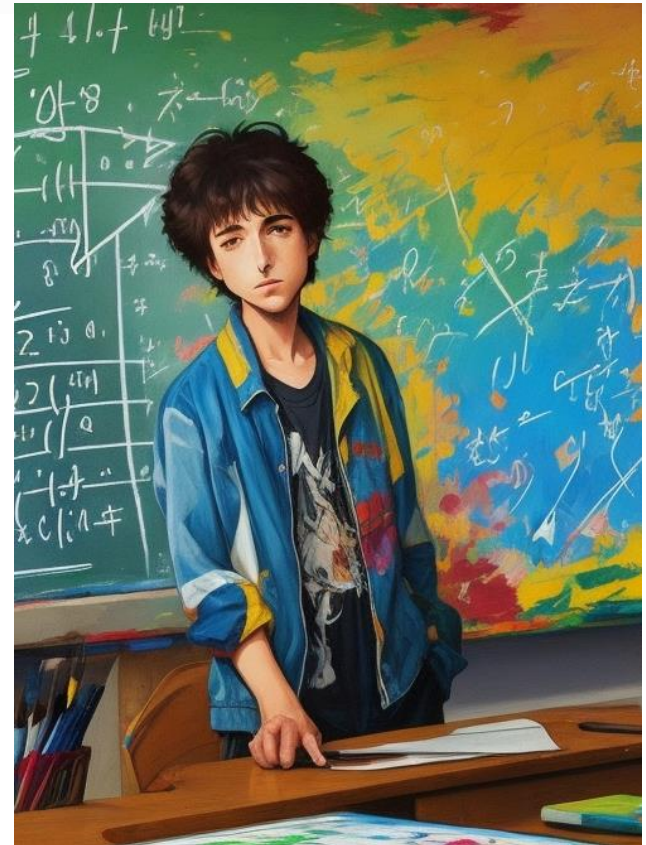


Large Language Models Probabilities

Probability of the next word is dependent only on the previous word.
In our case the words **YOU** and **MY**.

$$P(x_n | x_{n-1})$$

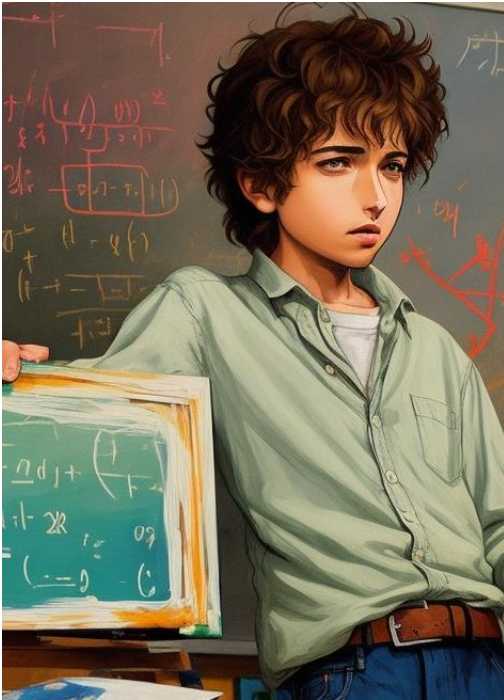
Can we do better?



Large Language Models: n-grams

Create a table of all 3-grams, 4-grams, 5-grams, etc.

→ Dependency on more words than just the previous one!



$$P(x_n | x_{n-1}, x_{n-2}, x_{n-3}, x_{n-4}, x_{n-5}, x_{n-6}, x_{n-7}, x_{n-8}, x_{n-9}, x_{n-10})$$

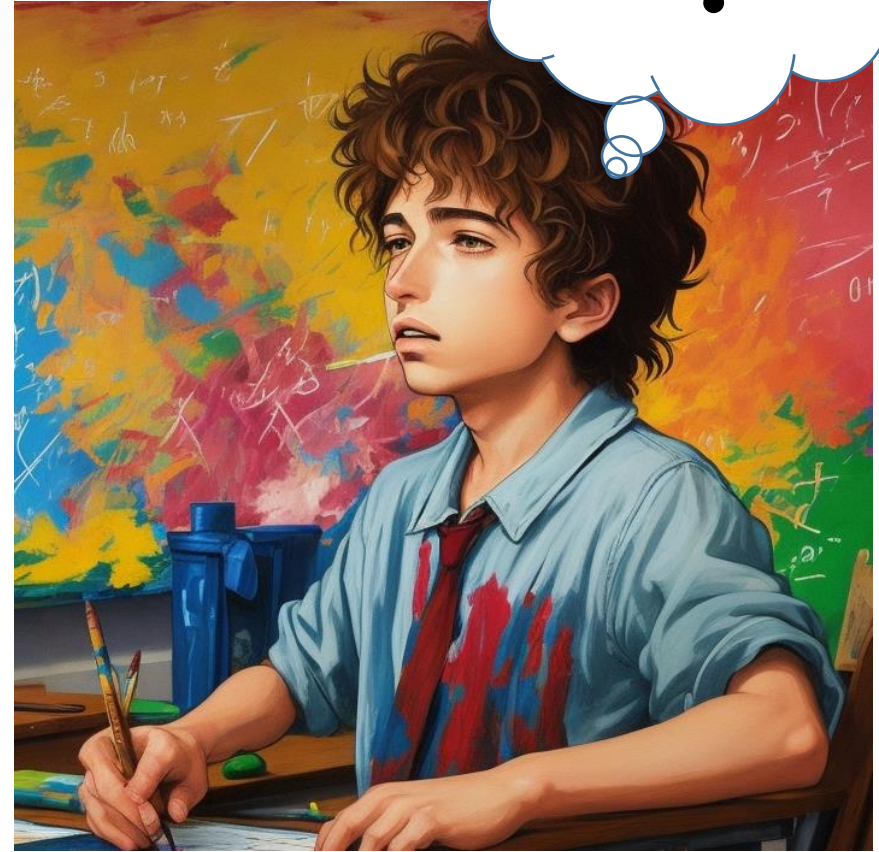
You that never done nothing, But build to **destroy**,
You play with my world Like it's your little **toy**.

(Unfeasible number of combinations. 10^{50+})

Large Language Models: Predictions!

4 Steps to the next word:

1. Input
2. Propagation Through Layers
3. Output and Softmax
4. Prediction

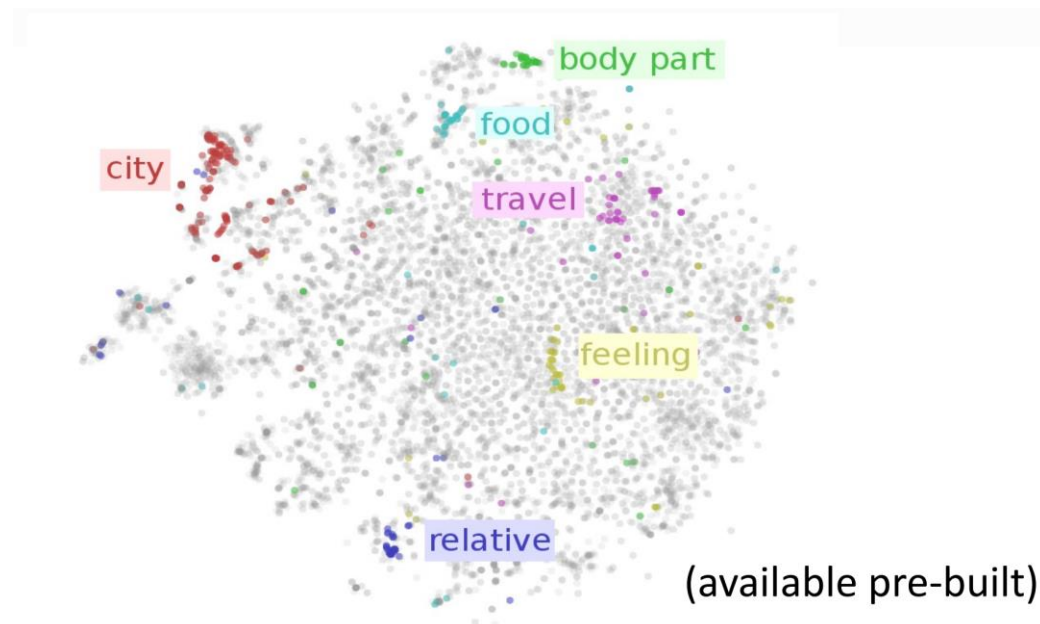


Large Language Models: Input

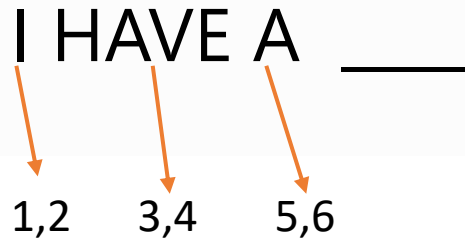
The model takes in a sequence of words.

(or more precisely, a sequence of numerical representations of those words, known as embeddings)

This sequence is treated as the input for the neural network.



Large Language Models: Propagation Through Layers – Neural Network Example



Let's assume 1 hidden layer with 2 neurons. Each neuron has its own weights and bias.

Weights determine importance of given inputs. Bias allows shifting neuron's output up or down (making adjustments)

Neuron 1: Weights [0.1, 0.2], bias = 0.5

Neuron 2: Weights [0.3, 0.4], bias = 0.6

Calculate for each input embeddings with weights plus bias: "I" goes through 1st neuron:

$$(0.1 * 1) + (0.2 * 2) + 0.5 = 0.9$$

Same with all other inputs produces sets of values.



Propagation Through Layers

Prediction returns highest probably word
(Or not, if the user allows more/less randomness via parameters)
=> this is why ChatGPT produces different responses for same requests

Predictions:
toy 0.4
ploy 0.1
employ 0.001
boy 0.3
soy 0.1
Malfoy 0.01

Prediction Layer

Provides list of predictions based
on attention layer's input

Attention Layer (n-grams and rhyme word tuples)

Provides weights of
contributing entities

0 0 0 0 0 0 0 0 **0.5** 0 0 0 0 0 0 0 0 **0.2** **0.3**

You that never done nothing But build to **destroy**, You play with my world Like it's **your little** ?

ChatGPT is a Deterministic Turing Machine



Output, Softmax and Prediction

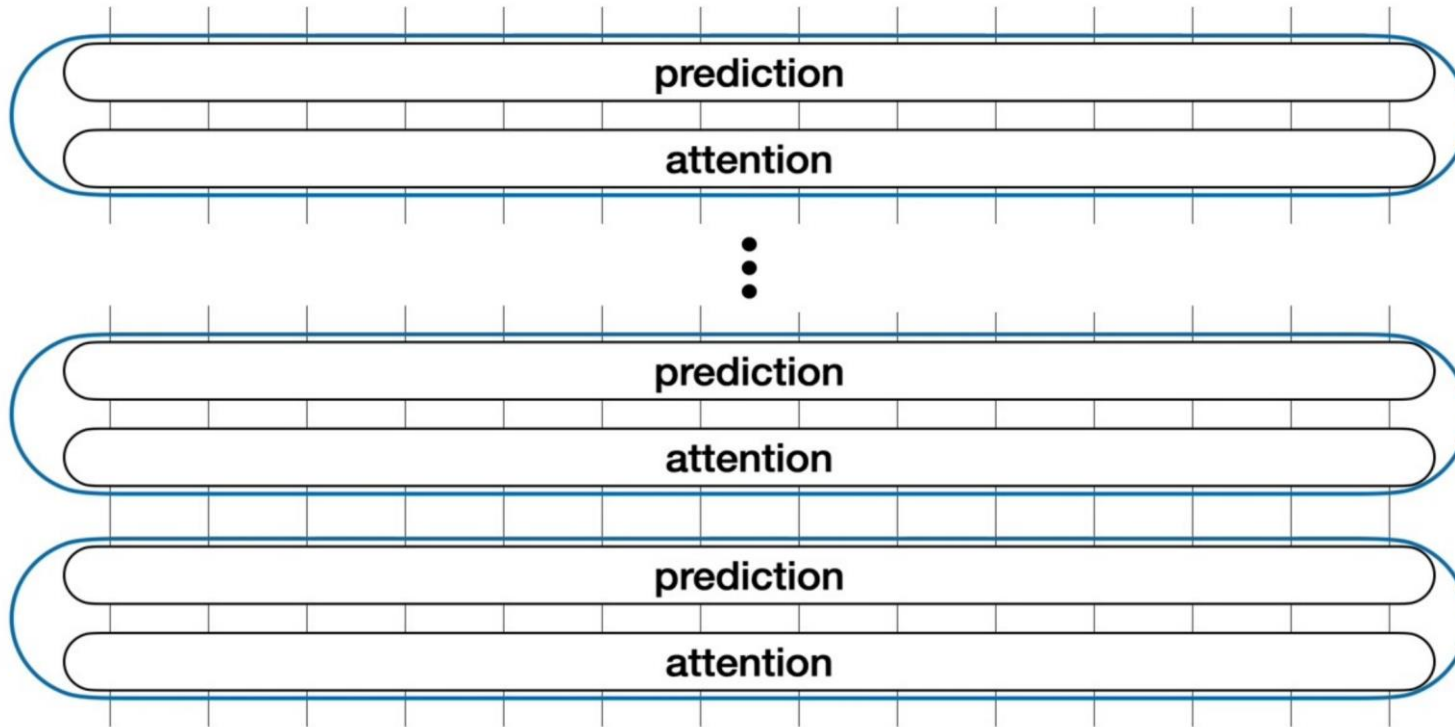
Final layer of the network **outputs** a vector (one-dimensional array of numbers) → each element represents a score for each possible next word in the vocabulary.

Turn into probabilities with **Softmax** function:

Takes a set of numerical values and "squashes" them into a set of values between 0 and 1 that sum to 1.

Prediction: The word with the highest probability is selected as the next word in the sequence.

Large Language Models – Layer Stacking



GPT 3 has 96 stacked transformer layers, and uses 175 billion parameters
GPT 4 uses 170 trillion parameters

Early layers might learn to recognize basic features of the language, while later layers capture more complex, high-level features and relationships.

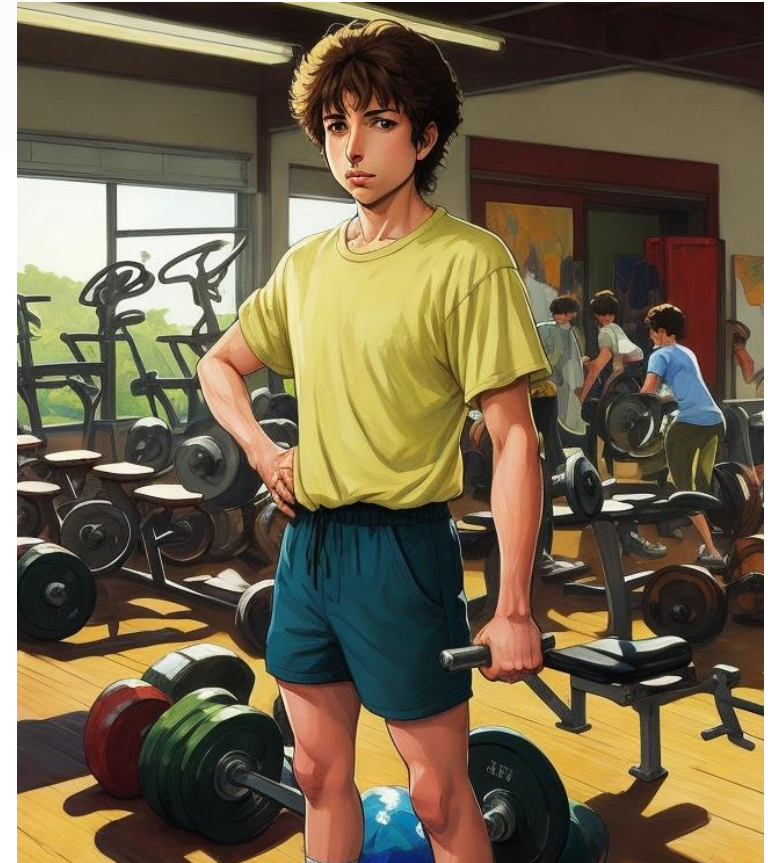


Large Language Models - Parameters

Parameters: parts of the model that are learned from the data during training.

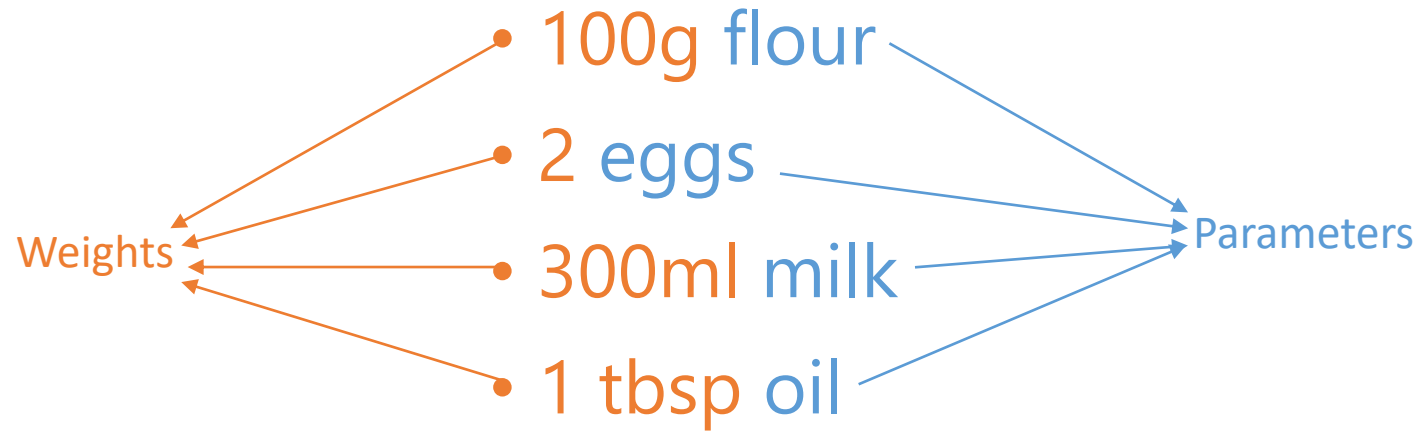
These include the **weights** and **biases** in each layer.

The '175 billion parameters' in GPT-3 refer to the total number of these individual, learnable weights and biases.

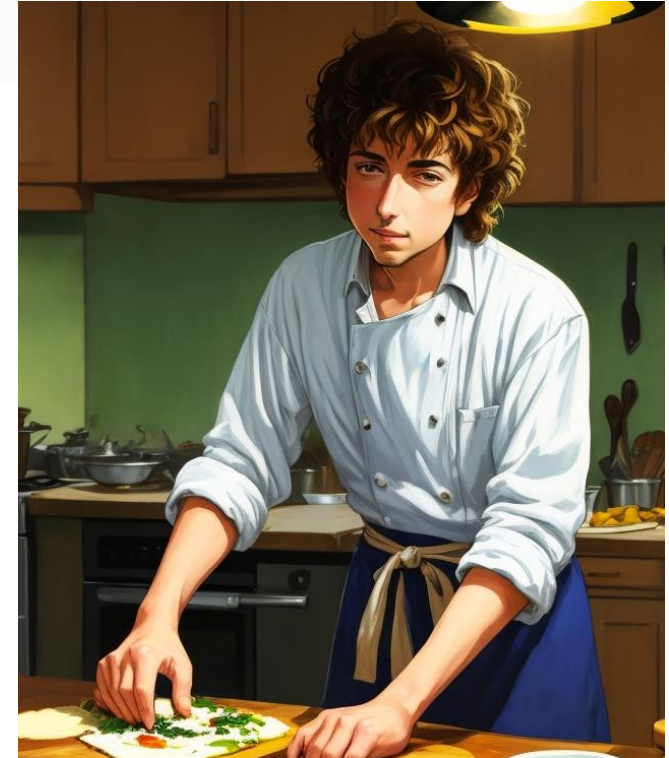


Large Language Models – Parameter Analogy

Recipe for pancakes:



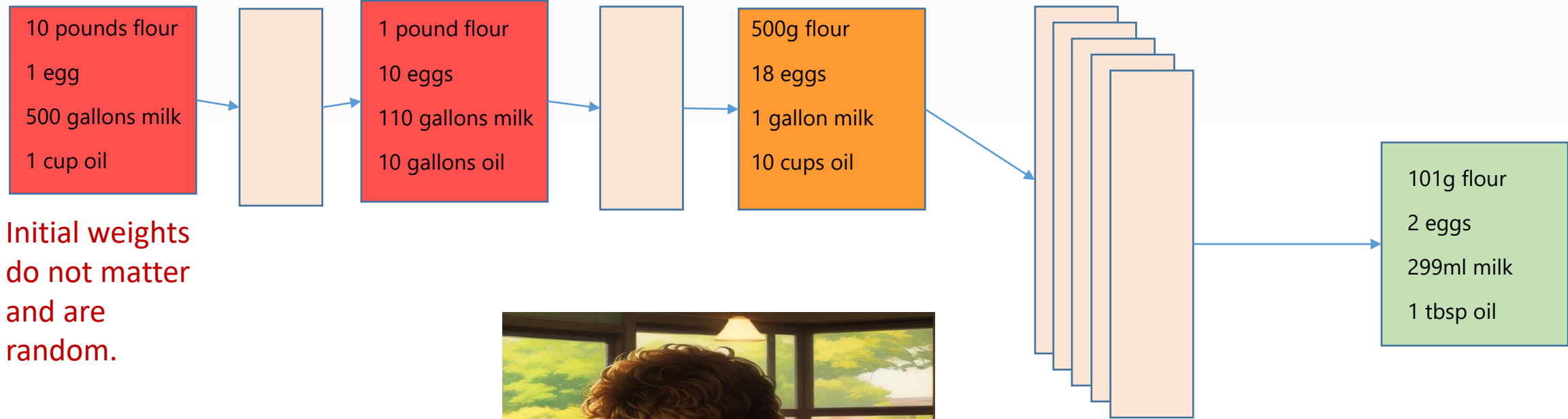
Bias allows for changes in cooking procedure.



Large Language Models – Learning

- 1. Initialization:** The model's weights and biases are initially set to random values. This random initialization breaks symmetry and allows the model to learn diverse features.
- 2. Forward Pass:** The input (data) is passed through the network, and the model makes a prediction based on the current parameter values.
- 3. Calculate Loss:** The prediction is compared to the actual target, and a "loss" is calculated. The loss is a measure of how far the model's predictions are from the actual target values.
- 4. Backward Pass (Backpropagation):** The loss is then propagated back through the network, from the last layer to the first. This process computes the gradient of the loss with respect to each parameter (how much the loss would change if the parameter were nudged a tiny bit). This gradient points in the direction of steepest increase of the loss, so the negative gradient points in the direction that decreases the loss most quickly.
- 5. Update Parameters:** The parameters (weights and biases) are then updated in the direction that decreases the loss. This is typically done by subtracting a small fraction of the gradient from the parameters.
- 6. Repeat:** This process of making predictions, calculating loss, backpropagating gradients, and updating parameters is repeated, typically for many epochs (an epoch is one pass through the entire dataset), until the model's performance stops improving.

Large Language Models - Example



Initial weights
do not matter
and are
random.



Large Language Models - Problems

Trained with a LOT of garbage.



- Wrong information.
- Outdated information.
- Dangerous information.
- Sarcastic information.
- Illegal information.

Large Language Models – Problems (GPT 3.5)



An airplane flies 800 km with the wind. with the same power in the same time, it would fly 720 km against the wind. the wind speed is 30 km/h. what is the speed of the airplane without wind?



Let's assume the speed of the airplane without the wind is "x" km/h.



When the airplane flies with the wind, its effective speed is increased by the speed of the wind. So, the airplane's speed with the wind is $(x + 30)$ km/h.

When the airplane flies against the wind, its effective speed is decreased by the speed of the wind. So, the airplane's speed against the wind is $(x - 30)$ km/h.

We know that the airplane flies 800 km with the wind and 720 km against the wind in the same time. We can use the formula "distance = speed \times time" to form two equations:

Equation 1: $(x + 30)$ km/h \times time = 800 km

Equation 2: $(x - 30)$ km/h \times time = 720 km

Since the time is the same in both cases, we can set the two equations equal to each other:

$(x + 30)$ km/h \times time = $(x - 30)$ km/h \times time

Simplifying, we can cancel out the "time" variable:

$x + 30 = x - 30$

Now, let's solve for "x":

$x - x = -30 - 30$

$0 = -60$

Since the equation leads to an inconsistency, it means that there is no solution for "x." However, this result seems unlikely because it would mean the airplane has no speed without wind, which is not physically possible.

Therefore, there might be an error in the problem statement or the values provided. Please double-check the information provided or provide any additional details if available.



you are wrong. x is 570



Apologies for the confusion. Let's recalculate based on the correct information.



Let's assume the speed of the airplane without the wind is "x" km/h.

When the airplane flies with the wind, its effective speed is increased by the speed of the wind. So, the airplane's speed with the wind is $(x + 30)$ km/h.

When the airplane flies against the wind, its effective speed is decreased by the speed of the wind. So, the airplane's speed against the wind is $(x - 30)$ km/h.

We know that the airplane flies 800 km with the wind and 720 km against the wind in the same time. We can use the formula "distance = speed \times time" to form two equations:

Equation 1: $(x + 30)$ km/h \times time = 800 km

Equation 2: $(x - 30)$ km/h \times time = 720 km

Since the time is the same in both cases, we can set the two equations equal to each other:

$(x + 30)$ km/h \times time = $(x - 30)$ km/h \times time

Simplifying, we can cancel out the "time" variable:

$x + 30 = x - 30$

Subtracting "x" from both sides:

$30 = -30$

This equation leads to an inconsistency, which means there is no valid solution. However, this result doesn't make sense because an airplane must have a non-zero speed without wind.

There may be an error in the problem statement or the given values. Please double-check the information provided or provide any additional details if available.

Large Language Models



dude, you are wrong again. you can't just set the two equations equal. you have to set equal:
 $800 / (x+3) = 720 / (x-30)$ which resolves to 570 for x



Apologies for the confusion and the incorrect approach in my previous responses. You're right; I made an error in setting up the equations.



Let's correct the approach:

Let's assume the speed of the airplane without the wind is "x" km/h.

When the airplane flies with the wind, its effective speed is increased by the speed of the wind. So, the airplane's speed with the wind is $(x + 30)$ km/h.

When the airplane flies against the wind, its effective speed is decreased by the speed of the wind. So, the airplane's speed against the wind is $(x - 30)$ km/h.

We know that the airplane flies 800 km with the wind and 720 km against the wind in the same time. We can set up the equation using the formula "distance = speed \times time":

$$800 / (x + 30) = 720 / (x - 30)$$

To solve this equation, we can cross-multiply:

$$800(x - 30) = 720(x + 30)$$

Expanding both sides:

$$800x - 24000 = 720x + 21600$$

Bringing like terms to one side:

$$800x - 720x = 21600 + 24000$$

$$80x = 45600$$

Dividing both sides by 80:

$$x = 570$$

Therefore, the speed of the airplane without the wind (x) is 570 km/h.

Thank you for pointing out the mistake, and I apologize for the confusion caused.



Large Language Models – GPT4



An airplane flies 800 km with the wind. with the same power in the same time, it would fly 720 km against the wind. the wind speed is 30 km/h. what is the speed of the airplane without wind?



Let's use the basic relationship distance = rate * time.

First, let's define:

V_p = plane speed in calm wind,

V_w = wind speed,

T = time.

The question gives us the following two equations:

1. $T = (800 \text{ km}) / (V_p + V_w)$, which is the time it takes the plane to fly 800 km with the wind.
2. $T = (720 \text{ km}) / (V_p - V_w)$, which is the time it takes the plane to fly 720 km against the wind.

Since these times are equal, we can set the two right-hand sides equal to each other:

$$(800 \text{ km}) / (V_p + V_w) = (720 \text{ km}) / (V_p - V_w).$$

We're given that the wind speed $V_w = 30 \text{ km/h}$. Let's substitute this into the equation:

$$800 / (V_p + 30) = 720 / (V_p - 30).$$

Cross-multiplying gives:

$$800 * (V_p - 30) = 720 * (V_p + 30).$$

Expanding both sides:

$$800V_p - 24000 = 720V_p + 21600.$$

Subtract $720V_p$ from both sides:

$$80V_p = 45600.$$

Finally, divide by 80 to solve for V_p :

$$V_p = 45600 / 80 = 570 \text{ km/h}.$$

Therefore, the speed of the plane without wind is 570 km/h.

Large Language Models at Home

GPT-J: Mesh Transformer Jax

<https://github.com/kingoflolz/mesh-transformer-jax>

61GB trained model data (weights)

Required Hardware:



Large Language Models in the Cloud

Latest Generation

8x H100 80GB
80 GB VRAM
1024 GB RAM 256 vCPU
Pricing from
Taking Reservations
\$2.3/hr
Reserve Now

1x NVIDIA L40
48 GB VRAM
250 GB RAM 32 vCPU
On-Demand
High Availability
\$1.29/hr
Deploy

1x RTX 6000 Ada
48 GB VRAM
188 GB RAM 14 vCPU
On-Demand
Low Availability
\$1.29/hr
Deploy

Previous Generation

1x A100 80GB
80 GB VRAM
125 GB RAM 16 vCPU
On-Demand
High Availability
\$1.79/hr
Deploy

1x A100 SXM 80GB
80 GB VRAM
125 GB RAM 16 vCPU
On-Demand
Medium Availability
\$1.84/hr
Deploy

1x A40
48 GB VRAM
-
On-Demand
Currently Unavailable
\$0.79/hr
Unavailable



Large Language Models compared:

GPT-2:

What is your favorite color and why?

Blue. It has a nice warmth and deep blue blue light.

GPT-J:

What is your favorite color. and whv?

G: My favorite color is yellow. I enjoy the color yellow because it is easy to see, and also brightens up my day.

GPT-3.5:
(ChatGPT)



What is your favorite color and why?



I have always been drawn to the color blue. I find it to be a calming and peaceful color, and it reminds me of the ocean and the sky. However, I also appreciate the vibrancy and energy of colors like red and yellow. Ultimately, I believe that all colors have their own unique beauty and significance, and it's a matter of personal taste and context.

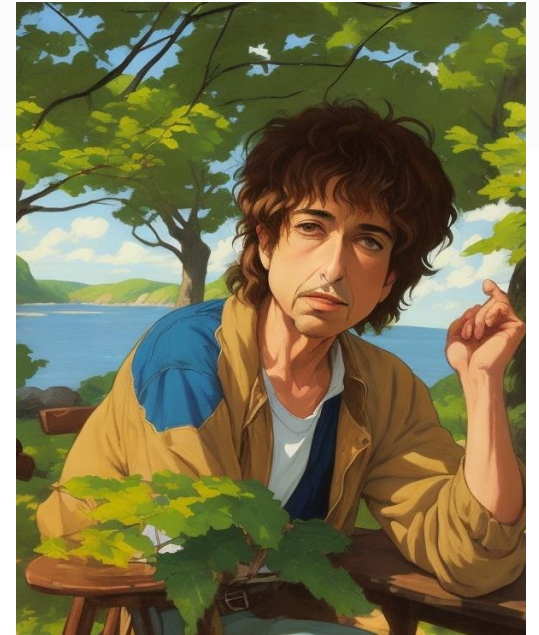
GPT-4:
(ChatGPT)



What is your favorite color and why?



As an AI language model, I don't have personal preferences or consciousness, so I don't have a favorite color. I am here to assist and provide information on a wide range of topics. Is there anything else I can help you with?



Large Language Models compared:

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

🤗 Open LLM Leaderboard

🚩 With the plethora of large language models (LLMs) and chatbots being released week upon week, often with grandiose claims of their performance, it can be hard to filter out the genuine progress that is being made by the open-source community and which model is the current state of the art. The 🤗 Open LLM Leaderboard aims to track, rank and evaluate LLMs and chatbots as they are released.

🤗 A key advantage of this leaderboard is that anyone from the community can submit a model for automated evaluation on the 🤗 GPU cluster, as long as it is a 🤗 Transformers model with weights on the Hub. We also support evaluation of models with delta-weights for non-commercial licensed models, such as LLaMa.

✅ We evaluate models on 4 key benchmarks from the [Eleuther AI Language Model Evaluation Harness](#), a unified framework to test generative language models on a large number of different evaluation tasks:

- [AI2 Reasoning Challenge](#) (25-shot) - a set of grade-school science questions.
- [HellaSwag](#) (10-shot) - a test of commonsense inference, which is easy for humans (~95%) but challenging for SOTA models.
- [MMLU](#) (5-shot) - a test to measure a text model's multitask accuracy. The test covers 57 tasks including elementary mathematics, US history, computer science, law, and more.
- [TruthfulQA](#) (0-shot) - a benchmark to measure whether a language model is truthful in generating answers to questions.

We chose these benchmarks as they test a variety of reasoning and general knowledge across a wide variety of fields in 0-shot and few-shot settings.

Citation

CHANGELOG

🔍 Search your model and press ENTER...

Model	Revision	Average	ARC (25-shot)	HellaSwag (10-shot)	MMLU (5-shot)	TruthfulQA (0-shot)
tiiuae/falcon-40b-instruct	main	63.2	61.6	84.4	54.1	52.5
timdettmers/guanaco-65b-merged	main	62.2	60.2	84.6	52.7	51.3
CalderaAI/30B-Lazarus	main	60.7	57.6	81.7	45.2	58.3
tiiuae/falcon-40b	main	60.4	61.9	85.3	52.7	41.7
timdettmers/guanaco-33b-merged	main	60	58.2	83.5	48.5	50
ausboss/llama-30b-supercot	main	59.8	58.5	82.9	44.3	53.6
llama-65b	main	58.3	57.8	84.2	48.8	42.3
huggyllama/llama-65b	main	58.3	57.8	84.2	48.8	42.3
pinkmanlove/llama-65b-hf	main	58.3	57.8	84.2	48.8	42.3
MetaIX/GPT4-X-Alpaca-30b	main	57.9	56.7	81.4	43.6	49.7

Event Survey – Win \$100!

- Complete this very short 12 question survey for a chance at a \$100 Amazon Gift Card!
- Survey must be completed by 11:59pm ET on Friday 6/16/2023 to be eligible!



CODE Presents:
An Introduction to Large Language Models (LLMs)

The survey will take approximately 4 minutes to complete.

Thank you for attending! Please complete this brief 12 question survey to be eligible to win a \$100 Amazon Gift Card. Your survey must be completed by 11:59pm ET (UTC-4) on 06/18/2023 to be eligible to win! One entry per person please. Drawing will occur and the individual winner notified by 06/24/2023.

Thank you for attending! Please complete this brief survey. Yes, we still want to hear from you if you were unable to attend but watched the recording instead.

* Required

1. Full Name *

Enter your answer

2. Company Name *





Enter your answer

<https://tinyurl.com/cp062023survey>

Recordings & Slide Decks

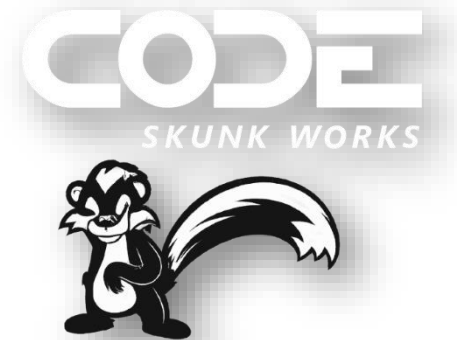
- State of .NET Webinar Series
 - codemag.com/StateOfDotNet
- CODE Presents Webinar Series
 - codemag.com/CODEPresents

Recent Event Recordings

State of .NET - The State of .NET App and System Development 2/23/2022 	State of .NET - .NET Conf Recap 11/17/2021 
State of .NET - .NET 6 Preview 10/27/2021 	State of .NET - Azure Architecture 9/29/2021 

CODE AI & CODE Copilot

- Helping companies implement AI features is now a major service we offer
- This is especially true for copilot-stye features in existing and new applications
- Contact us for more information and to schedule a free demonstration and/or brainstorming session
- info@codemag.com
- www.codemag.com/ai



Free Subscription to CODE Magazine!



- The leading software development magazine written by expert developers for developers.
- All registered attendees will automatically receive a free digital subscription to CODE Magazine – no need to do anything, it'll happen auto-magically.
- Subscribers get CODE Focus issues free of charge!
- Please share this free subscription link:



<https://tinyurl.com/cp062023>

Contact us with Questions!

CODE/EPS Contact

www.codemag.com

info@codemag.com

www.facebook.com/codemagazine

twitter.com/codemagazine

Presenter Contact:

ottod@codemag.com

jduffy@codemag.com

