

This is a conversation between a human and an advanced A.I. robot! The robot can give answers about the great mysteries of life.

Human: Hello!

Robot: Hello Human!

Human: What is the meaning of life?

Robot:

There is no one answer to this question. To some people, life might mean spending time with family, others might believe that the meaning of life is to help others. There are many different interpretations of what the meaning of life is, and it ultimately comes down to what you believe.

family, others might believe that the meaning of life is to help others. There are many different interpretations of what the meaning of life is, and it ultimately comes down to what you believe.

CHATBOT

TI XPEIAZETAI ENA CHATBOT;

Μία διεπαφή επικοινωνίας

Χρειαζόμαστε έναν τρόπο για να επικοινωνούμε με το chatbot μας. Πχ. Μία ιστοσελίδα, μία mobile εφαρμογή κτλ.



Έναν τρόπο να είναι

πάντα διαθέσιμο

Χρειαζόμαστε έναν server για να φιλοξενήσει την εφαρμογή μας

Κατανόηση κειμένου

Το chatbot μας θα πρέπει να «καταλαβαίνει» τις ερωτήσεις του χρήστη προκειμένου να δίνει χρήσιμες απαντήσεις

Ικανότητα Απάντησης

• Το chatbot θα πρέπει να μπορεί να συνθέτει κείμενο και να το στέλνει πίσω στον χρήστη

ΤΟ ΔΙΚΟ ΜΑΣ CHATBOT

Μία διεπαφή επικοινωνίας





Έναν τρόπο να είναι πάντα διαθέσιμο



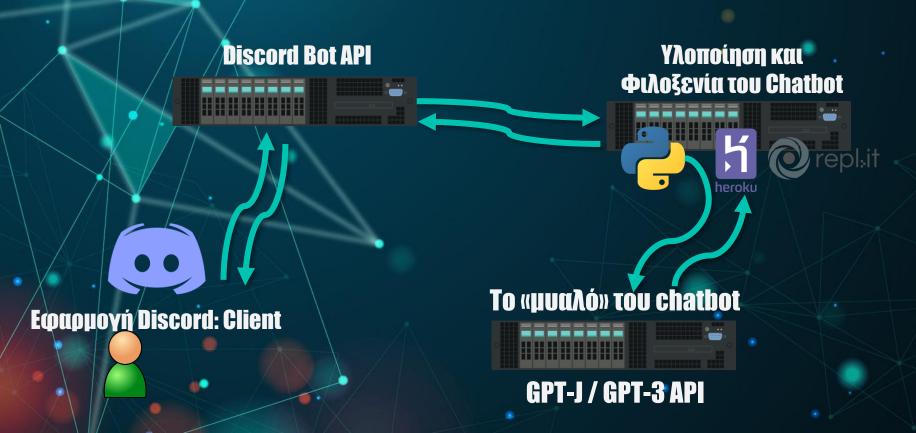
Κατανόηση κειμένου

GPT-J / GPT-3

Ικανότητα Απάντησης

GPT-J/GPT-3

H APXITEKTONIKH TOY CHATBOT ΜΑΣ



APIS

API = Application programming interface Διεπαφή προγραμματισμού εφαρμογών

Πως μπορούμε να κάνουμε χρήση μίας κεντρικής υπηρεσίας; **API** Web εφαρμογή Mobile Εφαρμογή

ΠΑΡΑΔΕΙΓΜΑ: Πως φτιάχνω μία εφαρμογή που απεικονίζει την κίνηση στους δρόμους σε πραγματικό χρόνο;

ΕΠΙΛΟΓΗ 1

- Αγοράζω μερικούςδορυφόρους
- Τους στέλνω σε τροχιά
- Αγοράζω Servers για την συλλογή δεδομένων και την εξυπηρέτηση των χρηστών
- Φτιάχνω μία εφαρμογή

ΕΠΙΛΟΓΗ 2

- Σρησιμοποιώ το API από τους χάρτες της Google
- Φτιάχνω μία εφαρμογή

ΠΩΣ ΛΕΙΤΟΥΡΓΕΙ ΕΝΑ ΑΡΙ

Στέλνουμε μια ερώτηση σε έναν server

www.foveri.upiresia.kairou
.gr/oKairosZwntana?poli=
Athina&xwra=Ellada

```
Παίρνουμε μία απάντηση
{
"Temperature": 20,
"Humidity": 60
}
```

ΓΛΩΣΙΚΑ ΜΟΝΤΕΛΑ GPT

GPT;

Generative Pre-trained Transformer

Παραγωγικά

Σκοπός είναι να παράγουμε νέα δεδομένα, στην προκειμένη περίπτωση νέες λέξεις. Προ-Εκπαιδευμένα [†]

Το μοντέλο μας έχει εκπαιδευτεί σε τεράστια σετ δεδομένων.

> Πχ GPT-3: Πάνω από 500GB κειμένου



Δίκτυα Μετασχηματιστών

Αυτό αναφέρεται στην δομή των νευρωνικών δικτύων του μοντέλου.

Οι μετασχηματιστές είναι δίκτυα που βασίζονται στον μηχανισμό της «Προσοχής» και είναι ιδανικά για διεργασίες τύπου seq2seq. (Μετατροπή σειράς σε σειρά)

ΓΛΩΣΣΙΚΑ MONTEΛΑ GPT

- Συναντήσαμε τον όρο GPT για πρώτη φορά το **2018** με την δημοσίευση του πρώτου **GPT** γλωσσικού μοντέλου από την OpenAl.
- Στην συνέχεια το **2019** είχαμε την επόμενη μεγαλύτερη έκδοση το **GPT-2**.
- Και το 2020 είχαμε το ακόμα μεγαλύτερο GPT-3.
- Ο στόχος ενός τέτοιου γλωσσικού μοντέλου είναι αρκετά απλός:
 - Σκοπός είναι η πρόβλεψη της επόμενης πιο κατάλληλης λέξης σε μία πρόταση.

Για παράδειγμα:

- Η γάτα έσπρωξε το ποτήρι από το τραπέζι και το ποτήρι έπεσε και... έσπασε



TI MNOPEI NA KANEI

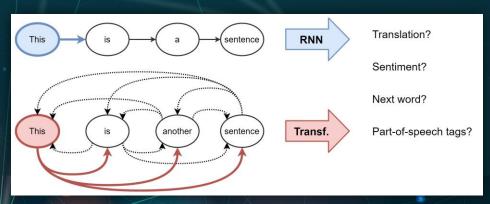
- . Παραγωγή κειμένου
- . Ανάλυση συναισθημάτων
- . Κατηγοριοποίηση εννοιών
- . Απάντηση σε ερωτήσεις
- . Διόρθωση κειμένου (Ορθογραφία/Γραμματική)
- . Μετάφραση
- . Περίληψη

ΔΙΚΤΥΑ ΜΕΤΑΣΧΗΜΑΤΙΣΤΩΝ

Μέχρι πρόσφατα ο πιο διαδεδομένος τρόπος για να αντιμετωπίσουμε σειρές-ακολουθίες από λέξεις ήταν τα γνωστά RNNs – Recurrent Neural Networks.

Το πρόβλημα με τα RNN είναι ότι η επεξεργασία των πληροφοριών γίνεται σειριακά

Με τους μετασχηματιστές εισάγεται η έννοια της «προσοχής» και μπορούμε να επεξεργαζόμαστε λέξεις παράλληλα.

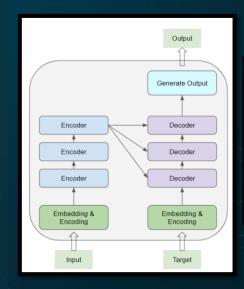


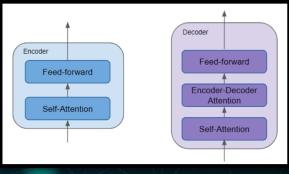
ΔΙΚΤΎΑ ΜΕΤΑΣΧΗΜΑΤΙΣΤΩΝ

Ο μηχανισμός της «Προσοχής» μας επιτρέπει να επικεντρωνόμαστε στις λέξεις που μπορεί να είναι πιο σημαντικές στην πρόταση μας.

Η επιτυχία των μετασχηματιστών κρύβεται στον μηχανισμό της «Αυτοπροσοχής»

Πρακτικά για κάθε σειρά από λέξεις που βάζουμε σαν είσοδο, ο μηχανισμός μας υπολογίζει τον συσχετισμό της κάθε λέξης με όλες τις υπόλοιπες.





ΔΙΚΤΥΑ ΜΕΤΑΣΧΗΜΑΤΙΣΤΩΝ

Το γατί έριξε το ποτήρι και αυτό έσπασε





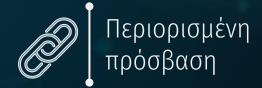
GPT-3







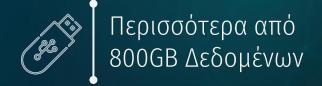




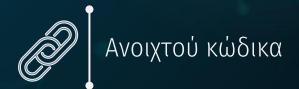
GPT-j











ΣΥΓΚΡΙΣΗ ΜΟΝΤΕΛΩΝ

	GPT-1	GPT-2	GPT-3	GPT-j
Παράμετροι	117M	1.5B	175B	6B
Dataset	BooksCorpus 11.038 Βιβλία ~1GB	WebText 40GB 8M έγγραφα	Common Crawl, WebText2, Books1, Books2 and Wikipedia 570 GB	The Pile 825 GB

The Pile Dataset

Component	Raw Size	Weight	Epochs	Effective Size	Mean Document Size
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
/ PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
Books3 [†]	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
ArXiv	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB
Github	95.16 GiB	7.59%	1.0	95.16 GiB	5.25 KiB
FreeLaw	51.15 GiB	6.12%	1.5	76.73 GiB	15.06 KiB
Stack Exchange	32.20 GiB	5.13%	2.0	64.39 GiB	2.16 KiB
USPTO Backgrounds	22.90 GiB	3.65%	2.0	45.81 GiB	4.08 KiB
PubMed Abstracts	19.26 GiB	3.07%	2.0	38.53 GiB	1.30 KiB
Gutenberg (PG-19) [†]	10.88 GiB	2.17%	2.5	27.19 GiB	398.73 KiB
OpenSubtitles [†]	12.98 GiB	1.55%	1.5	19.47 GiB	30.48 KiB
Wikipedia (en) [†]	6.38 GiB	1.53%	3.0	19.13 GiB	1.11 KiB
DM Mathematics [†]	7.75 GiB	1.24%	2.0	15.49 GiB	8.00 KiB
Ubuntu IRC	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB
BookCorpus2	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB
EuroParl [†]	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB
HackerNews	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB
YoutubeSubtitles	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB
PhilPapers	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB
NIH ExPorter	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB
Enron Emails†	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB
The Pile	825.18 GiB			1254.20 GiB	5.91 KiB

GITHUB

Βρίσκουμε τον κώδικα μας στο GitHub

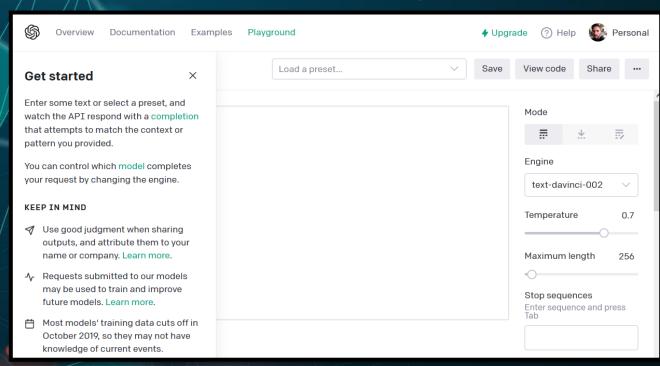
https://github.com/fzantalis/gpt-discord-bot





GPT-3 PLAYGROUND

https://beta.openai.com/playground



TextSynth GPT-j Playground

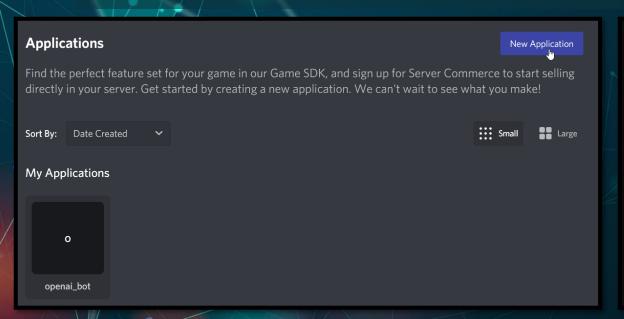
https://textsynth.com/playground.html

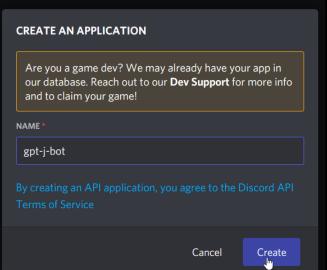
TextSynth	Home	Playgro	ound	Docun	nentatio	n Pri	cing Login	Sign L	lp					
									English mode the French la					
model. Mor		-						1-5 101	the Henchia	nguage.	GFT-NEOX	200 13 (ne large	31
Type a text	and let th	ie neura	l netwo	ork co	mplete it	t. Each	try returns a	differer	nt randomly ch	osen co	mpletion.			
Model: GP	T-J 6B	~	top-k:	40	top-p:	0.9	temperature	1.0	max tokens:	200	stop at:			
Select an e		It will be	e auton	natica	lly comp	leted	with a (hopefu	ılly) rea	llistic text.					
Complete	Toyt													

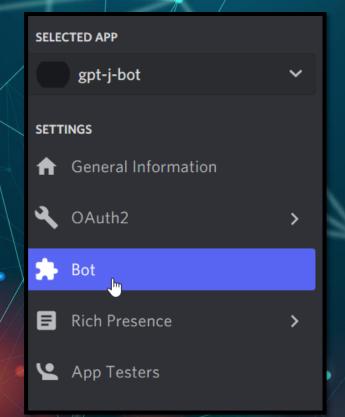
DISCORD BOT

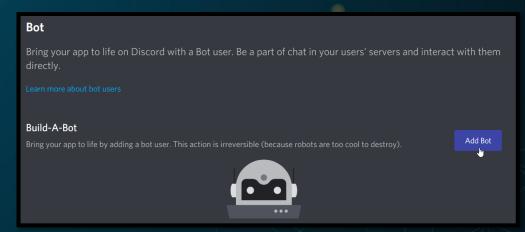
https://discord.com/developers/applications

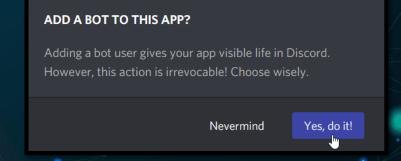
https://discordpy.readthedocs.io/en/stable/discord.html







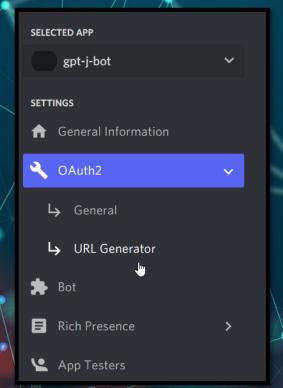




Αντιγράφουμε το API key και το κρατάμε κάπου για αργότερα

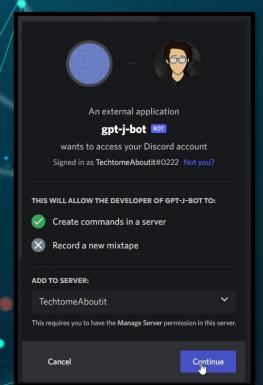


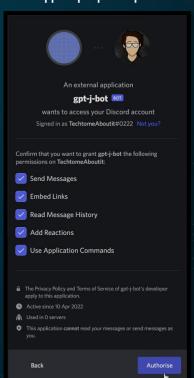
Αντιγράφουμε το generated URL στο τέλος



		✓ applications.commands
	✓ bot	
OT PERMISSIONS		
GENERAL PERMISSIONS	TEXT PERMISSIONS	VOICE PERMISSIONS
	✓ Send Messages	
	✓ Embed Links	
	✓ Read Message History	
	✓ Add Reactions	
	✓ Use Slash Commands	
SENERATED URL		

Βάζουμε στον browser το URL που αντιγράψαμε πριν

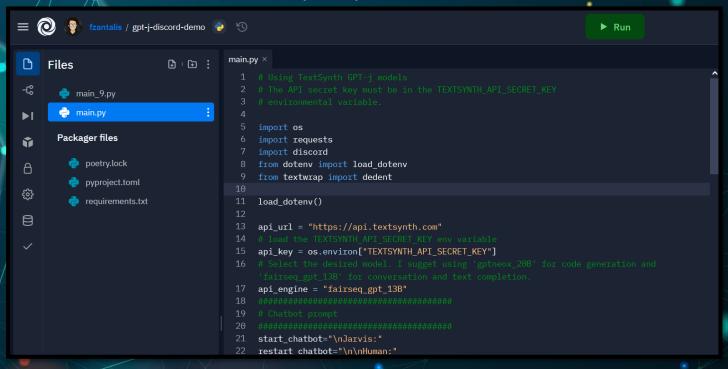




REPLIT WEB IDE

Δοκιμάζουμε την Εφαρμογή μας στο Replit

https://replit.com



Τα διαθέσιμα μοντέλα του GPT-j

- gptj_6B: Το βασικό μοντέλο του GPT-j των 6 δις παραμέτρων. Η βασική του γλώσσα είναι τα αγγλικά αλλά μπορεί να μιλήσει και σε άλλες γλώσσες. Ξέρει επίσης αρκετές γλώσσες προγραμματισμού.
- boris_6B: Είναι μία τροποποιημένη έκδοση του GPT-j εκπαιδευμένο στα Γαλλικά
- fairseq_gpt_13B: Είναι ένα μοντέλο 13 δις παραμέτρων. Δεν έχει εκπαιδευτεί στα ίδια δεδομένα με το GPT-j, συνεπώς δεν είναι τόσο καλό στις γλώσσες προγραμματισμού, αλλά αποδίδει καλύτερα στην απλή ομιλία στα Αγγλικά.
- gptneox 20B: Είναι το μεγαλύτερο ανοιχτό αγγλικό μοντέλο που έχουμε αυτή την στιγμή με 20 δις παραμέτρους. Έχει εκπαιδευτεί στα ίδια δεδομένα με το GPT-j οπότε είναι πολύ καλό στον κώδικα

Οι ρυθμίσεις του GPT-j

- max_tokens: optional int (default = 100)
 Κάθε token είναι 4 χαρακτήρες. Ορίζει το μέγιστο μέγεθος της απάντησης.
- temperature: optional double (default = 1)
 Υψηλότερες τιμές θα δώσουν πιθανότατα πιο πρωτότυπες απαντήσεις, αλλά ίσως να έχουν και λιγότερη συνοχή
- ο **top_k** optional integer (range: 1 to 1000, default = 40) Ορίζει το πλήθος των πιθανών απαντήσεων από τις οποίες βγαίνει το τελικό αποτέλεσμα. Ισχύει ότι ισχύει και στο temperature για την συνοχή.
- top_p: optional double (range: 0 to 1, default = 0.9)
 Η πιθανότητα επιλογής της κάθε υποψήφιας απάντησης θα πρέπει να είναι μεγαλύτερη ή ίση με το top_p
- stop: optional string or array of string (default = null)
 Η παραγωγή κειμένου θα σταματήσει όταν συναντήσει τον χαρακτήρα stop.
- presence_penalty: optional number (range: -2 to 2, default = 0)
 Θετικές τιμές του presence_penalty κάνουν τις λέξεις που έχουν χρησιμοποιηθεί ήδη λιγότερο πιθανές να επαναχρησιμοποιηθούν.
- frequency_penalty: optional number (range: -2 to 2, default = 0)
 Παρόμοιο με το presence_penalty αλλά παίζει ρόλο και η συχνότητα που έχει εμφανιστεί η κάθε λέξη

HEROKU

https://dashboard.heroku.com/new-app

Create New App	
App name	
gpt-j-bot-ttmai	•
gpt-j-bot-ttmai is available	
Choose a region	
Europe	\$
Add to pipeline	
Create app	

Στο μενού «Deployment method» συνδέουμε τον λογαριασμό μας στο GitHub για να παίρνει τον κώδικα του project μας.

Deployment method







Container Registry
Use Heroku CLI

Connect to GitHub

Connect this app to GitHub to enable code diffs and deploys.

View your code diffs on GitHub

Connect your app to a GitHub repository to see commit diffs in the activity log.

Deploy changes with GitHub

Connecting to a repository will allow you to deploy a branch to your app.

Automatic deploys from GitHub

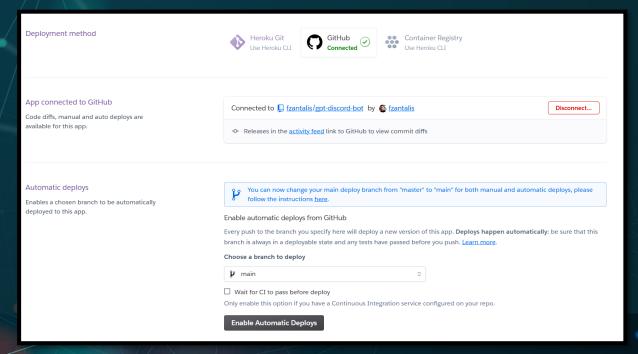
Select a branch to deploy automatically whenever it is pushed to.

Create review apps in pipelines

Pipelines connected to GitHub can enable review apps, and create apps for new pull requests. Learn more.

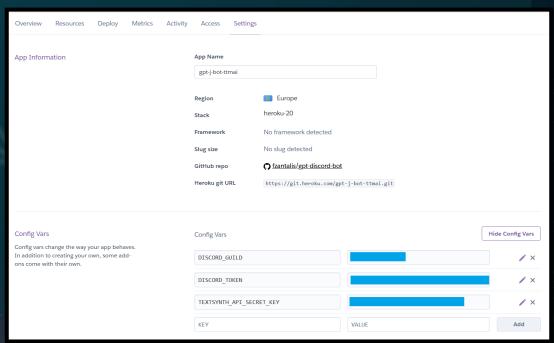
Connect to GitHub

Επιλέγουμε να ενεργοποιήσουμε τα αυτόματα Deploys έτσι ώστε κάθε φορά που κάνουμε αλλαγή στον κώδικα μας να ανανεώνεται η εφαρμογή

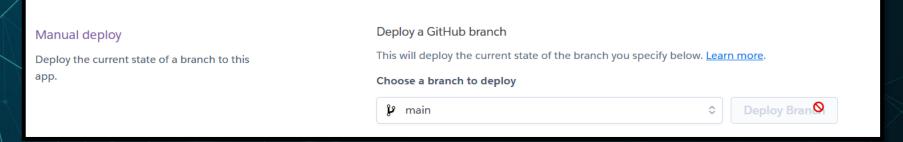


Στις ρυθμίσεις, πάμε στην κατηγορία Config Vars και προσθέτουμε τις βασικές μεταβλητές μας με τα API keys του Discord, του TextSynth και το GUILD του discord.

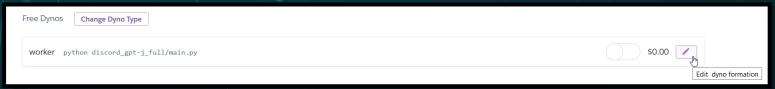
DISCORD_GUILD
DISCORD_TOKEN
TEXTSYNTH_API_SECRET_KEY

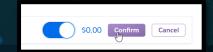


Κάνουμε το πρώτο Deploy της εφαρμογής



Στην καρτέλα "resources" ενεργοποιούμε τον worker της εφαρμογής μας





Πηγές

- Sequence Transduction with Recurrent Neural Networks: https://arxiv.org/abs/1211.3711
- Language Models are Unsupervised Multitask Learners https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf
- Language Models are Few-Shot Learners https://arxiv.org/pdf/2005.14165.pdf
- Attention Is All You Need https://arxiv.org/pdf/1706.03762.pdf
- https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2
- https://6b.eleuther.ai/
- The Pile: An 800GB Dataset of Diverse Text for Language Modeling https://arxiv.org/pdf/2101.00027.pdf
- https://textsynth.com/documentation.html
- https://discordpy.readthedocs.io/en/stable/discord.html

Ευχαριστώ!

Tech to me About it









CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.