

ChatGPT, GPT4 ja niiden monet serkukset: lyhyt johdatus generatiiviseen tekoälyyn

Filip Ginter
TurkuNLP-tutkimusryhmä
Turun Yliopisto





(NLP = Natural Language Processing,
kieliteknologia)

/ Kuka olen?

Professori TurkuNLP-tutkimusryhmässä

Pääaine: Tietojenkäsittelytiede

22 vuotta TurkuNLP:ssä / NLP alalla

Fokuksena mm. koneoppimiseen perustuvien menetelmien
soveltaminen suomenkieliseen tekstiin

Kieliteknologia Turun yliopistolla



TURKUNLP
.ORG



**UNIVERSITY
OF TURKU**



/TurkuNLP

(NLP = Natural Language Processing,
kieliteknologia)

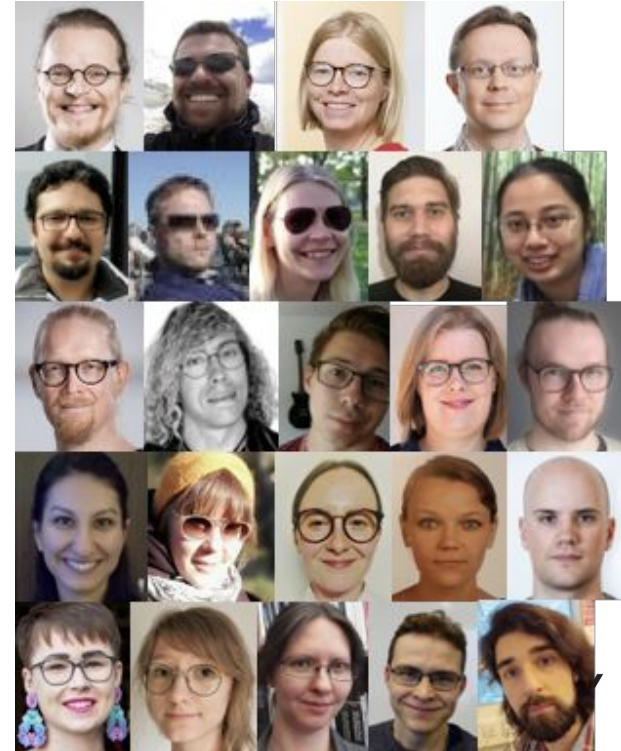
Perustettu vuonna 2001, nyt ~25 jäsentä

Monitieteinen ryhmä: tietotekniikka, koneoppiminen,
kielentutkimus, korpuslingvistiikka

Monia **kieltä** ja **teknologiaa** yhdistäviä
tutkimushankkeita

Keskiössä **koneoppimispohjaiset** lähestymistavat ja
digitaaliset kieliaineistot

Monet suomenkieliset kieliteknologiasovellukset ovat
ryhmämme käsialaa!

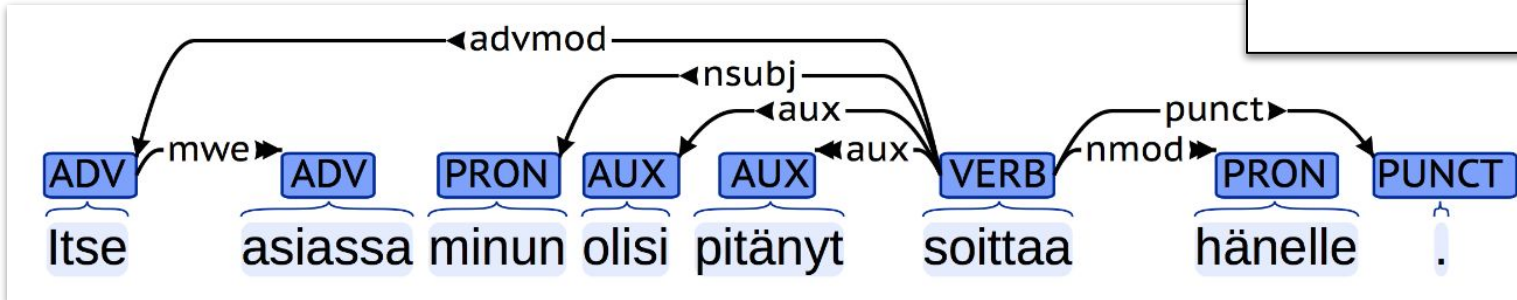
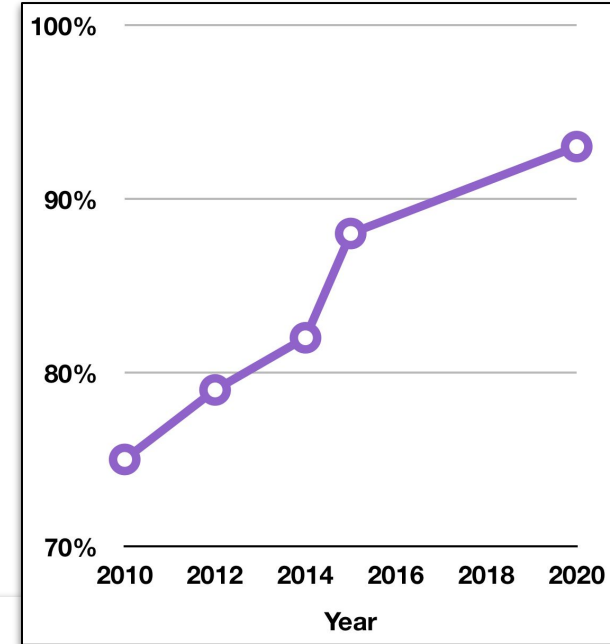




/ Syntaksijäsennys

Suomenkielinen syntaksijäsennys keskeinen tutkimuskohde ryhmässä

Tehtävä: jäsentää lauseen rakenne ja sen sanojen väliset suhteet sekä tunnistaa sanojen sanaluokat, morfologia ja perusmuodot





/ Parafraasi tutkimus

Pahoittelen, että jouduitte odottamaan

≈

Anteeksi viivästys

Tehtävä: Tunnista tekstin merkitys riippumatta käytetyistä sanoista ja niiden muista mahdollisista merkityksistä

Nyt kerättynä yli 100,000 vastaavaa esimerkkiä eri lähteistä

Käsiin kerätty - tärkeää dataa jonka avulla voidaan opettaa koneelle eri lauseiden merkitys



Turku Finnish Parsebank - Suomenkielinen internet kieliaineistoksi

Nykyaikainen kieliteknologia tarvitsee paljon dataa, tähän internet on hyvä lähde

Tehtävä: Kerätä mahdollisimman paljon puhdasta suomenkielistä tekstiä internetistä

Nyt kerättynä ~ **8 miljardia sanaa** (eli lähes 7000 Raamattua tai 15000 Sinuhe egyptiläistä)

Tärkeä suomenkielisen kieliteknologian kehittämisessä!



Tekstilajien tunnistus



WIKIPEDIA
Vapaa tietosanakirja

Turku

Tämä artikkeli kertoo kaupungista. Turku-nimen muita merkityksiä on [täsmennyssivulla](#).

Turku (ruots. *Åbo*, lat. *Aboa*) on Lounais-Suomen suurin kaupunki ja Varsinais-Suomen maakuntakeskus. Se sijaitsee Aurajoen suulla Saaristomeren rannikolla. Kaupungin asukasluku oli 197 917 henkilöä 31. joulukuuta 2022^[2] ja Turun seutukunnan noin 337 000 (2022).^[9] Täten Turku on asukasluvultaan Suomen kuudenneksi suurin kunta ja kolmanneksi suurin kaupunkialue.

Äitienpäiväkakku

Ruoka



Urheilu

Messi matkusti ilman lupaa Saudi-Arabiaan ja aiheutti kohun – tästä supertähden hyllytyksessä on kyse

Useat mediat kertovat PSG:n hyllyttäneen tähtipelaajansa Lionel Messin. Messin jatko pariisilaisseurassa on yhä epätodennäköisempää.

① Mittaa kinuskin kaikki ainekset pinnoitettuun kattilaan, kiehauta ja keitä keskilämmöllä noin viisitoista minuuttia välillä sekoitellen.

② Siirrä kinuski toiseen astiaan jäähtymään. Voit jäädyttää kinuskia ensiksi vesihauteessa ja sitten siirtää sen jäähtymään jääkaappiin tai pakastimeen. Jos jäähdytät kinuskin pakastimessa, käy



/ FinBERT

Suomenkielinen kielimalli, jonka avulla voidaan kehittää yhä parempia, neuroverkkopohjaisia kieliteknologian sovelluksia

Maailman ensimmäisiä yliopistossa kehitettyjä kielimalleja

FinBERT mahdollisti erittäin merkittävän parannuksen suomenkielisen kieliteknologian laadussa



Generatiivinen tekoäly



/ Generatiivinen tekoäly

- Generatiivinen tekoäly viittaa malleihin, jotka luovat uutta sisältöä
 - Tuota uusi kuva sanallisen kuvailun perusteella
 - Tuota tekstimuotoinen vastaus annetun kehotteen perusteella
 - Kehote: Syöte, joka annetaan mallille, ja jonka perusteella malli tuottaa vastauksen
 - Myös musiikkia, ohjelmakoodia, videoita, molekyylejä...



/ Generatiivinen tekoäly



Kuinka monta kultaa TPS on voittanut jääkiekossa?



Turun Palloseura (TPS) on voittanut Suomen jääkiekkomestaruuden yhteensä 11 kertaa. Ensimmäisen kerran se voitti mestaruuden kaudella 1955-1956 ja viimeksi kaudella 2009-2010. Lisäksi TPS on saavuttanut SM-hopeaa yhteensä 12 kertaa ja pronssia 10 kertaa.





/ Generatiivinen tekoäly



Entä jos naisten voittamat mitalit otetaan mukaan?



TPS:n naisten joukkue on voittanut jääkiekon naisten Suomen mestaruuden kaikkiaan seitsemän kertaa. He ovat voittaneet mestaruuden kausina 2006-2007, 2007-2008, 2008-2009, 2009-2010, 2010-2011, 2011-2012 ja 2012-2013. Joten yhteensä, kun sekä miesten että naisten saavutukset otetaan huomioon, TPS on voittanut Suomen mestaruuden jääkiekossa yhteensä 18 kertaa.



FI

Kirjoita 5-luokkalaisten historian essee Antiikin Rooman kulttuurista.



Otsikko: Antiikin Rooman kulttuuri



Antiikin Rooma oli erittäin vaikuttava sivilisaatio, joka hallitsi suurta osaa maailmasta noin kaksi tuhatta vuotta sitten. Rooman valtakunnan kulttuuri oli monipuolinen ja rikas, ja se on vaikuttanut moniin kulttuureihin ympäri maailmaa.

Ensinnäkin, antiikin roomalaiset olivat mestareita arkkitehtuurissa. He rakensivat suuria rakennuksia, kuten Colosseum, joka oli iso amfiteatteri, jossa järjestettiin gladiaattoritaisteluja ja muita näytöksiä. Roomassa oli myös muita hienoja rakennuksia, kuten Pantheon, joka oli temppeli jumalille, ja Forum Romanum, joka oli kaupungin keskus ja kohtaamispaikka. Roomal

Muuttaisitko esseen lukion oppimäärää vastaavaksi?



Otsikko: Antiikin Rooman kulttuurin merkitys ja vaikutus nykyaikaan

Antiikin Rooman valtakunta oli yksi historian vaikuttavimmista sivilisaatioista, ja sen kulttuurilla on ollut merkittävä vaikutus nykypäivän länsimaiseen yhteiskuntaan. Rooman valtakunta kukoisti noin 500 eaa. - 500 jaa., ja sen aikana kehitettiin monia innovaatioita, taiteen muotoja ja ajattelutapoja, jotka elävät edelleen nykykulttuurissamme.

Arkkitehtuurissa roomalaiset olivat todellisia mestareita. Heidän insinööritaidoillaan rakennettiin vaikuttavia rakennelmia, kuten Colosseum, joka oli valtava amfiteatteri gladiaattoritaisteluiden ja muiden näytösten järjestämistä varten. Pantheon, jumalien temppeli, ja Forum Romanum, Rooman kaupungin keskus ja kohtaamispaikka, ov



FI

Voisitko tehdä esseen suomenruotsiksi?



Rubrik: Den antika romerska kulturens betydelse och påverkan på nutiden

Det antika romerska riket var en av historiens mest inflytelserika civilisationer, och dess kultur har haft en betydande påverkan på västerländskt samhälle i dag. Romerska riket blomstrade mellan cirka 500 f.Kr. och 500 e.Kr., och under denna tid utvecklades många innovationer, konstformer och tankesätt som fortfarande lever kvar i vår nutida kultur.

Mediassa

OpenAI ilmoittaa, että sen kielimalli saavuttaa parempia tuloksia SAT-kokeessa kuin 90% ihmisistä



OpenAI

GPT-4 läpäisee oikeustieteen loppukokeen, saa paremmat pisteet kuin 90% ihmisistä

"I don't even really care about the bar exam, per se," said researcher Daniel Katz. "This crystallizes what is happening for people in a way [that says], here's some tasks that lawyers do, and it does it marginally better."

Microsoft kohautti: GPT-4 osoittaa jo ihmisen kaltaisen yleislykkyyden merkkejä

Goldman Sachsin mukaan tekoälyn uusin aalto voi vaikuttaa 300 miljoonaan työpaikkaan.

Hong Kong (CNN) — As many as 300 million full-time jobs around the world could be automated by the newest wave of artificial intelligence that has spawned platforms like ChatGPT, according to economists.

They predicted in a report Sunday that 18% of work globally could be computerized, with the deepest impact in advanced economies than emerging markets.

27.3.2023 18:36 TEKÖÄLY TEKNOLOGIA

...öntävät itsekkin kielimallin rajallisuuden, vaikka julistavat otsikossa sen ...
...oittamassa merkkejä kehityksestä yleislykkyyttä kohti.

PRESENTED BY
IDEAS • TECHNOLOGY

**Tekoälyn kehittämisen keskeyttäminen ei riitä.
Se on lopetettava kokonaan.**



<https://www.cnn.com/2023/03/14/openai-announces-gpt-4-says-beats-90percent-of-humans-on-sat.html>

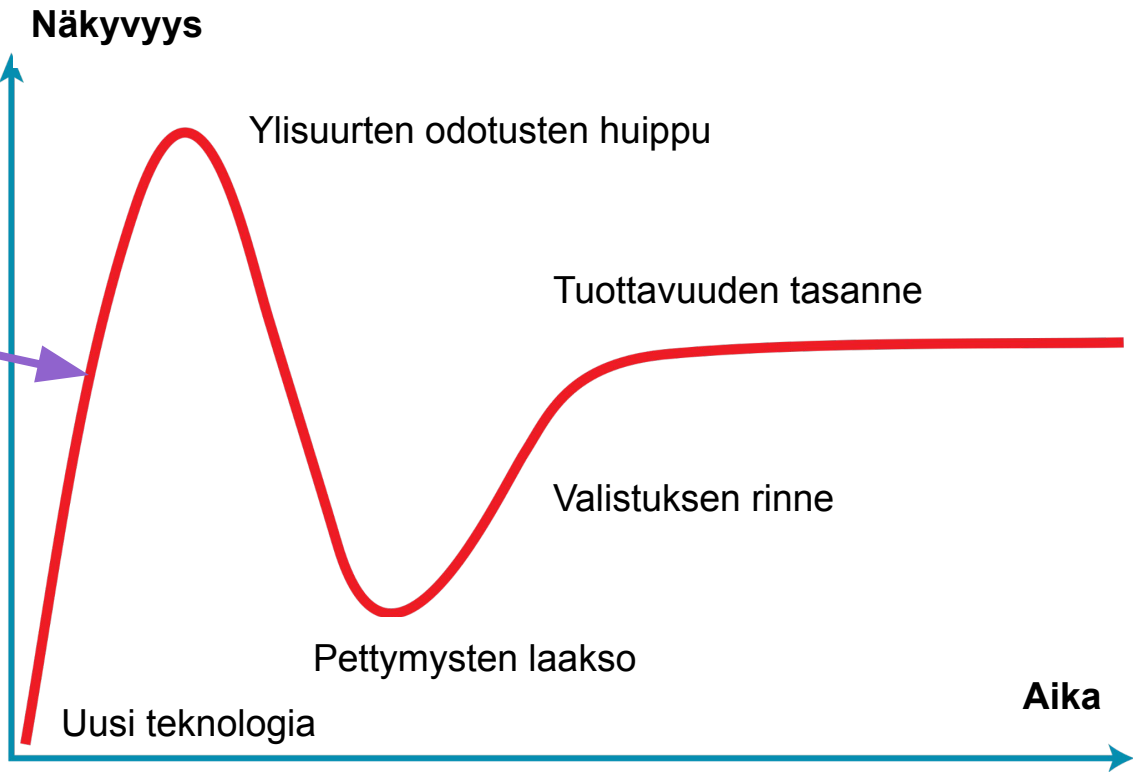
<https://www.law.com/legaltechnews/2023/03/17/how-gpt-4-mastered-the-entire-bar-exam-and-why-that-matters/>

<https://www.tivi.fi/uutiset/microsoft-kohautti-gpt-4-osoittaa-jo-ihmisen-kaltaisen-yleislykkyyden-merkkeja/6dcb6e5d-83bc-47cb-ab7a-e71833a53edb>

<https://edition.cnn.com/2023/03/29/tech/chatgpt-ai-automation-jobs-impact-intl-hnk/index.html>

<https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/>

Ollaanko tässä?



https://en.wikipedia.org/wiki/Gartner_hype_cycle

Generatiiviset kielimallit



TURKUNLP
.ORG



UNIVERSITY
OF TURKU



/ Mikä on kielimalli?

- GPT on **kielimalli**
- Kielimalli: ennusta seuraava sana annetun kehotteen perusteella (todennäköisyysjakauma koko sanastolle)
- Kielimalleilla monia käyttötarkoituksia, tekstingenerointi kaikista tunnetuin
- Yksinkertainen menetelmä tekstin generointiin:
 - Aloita kehotteesta
 - Valitse ennustetun todennäköisyysjakauman perusteella yksi sana
 - Lisää sana kehotteeseen
 - Toista uudella, yhden sanan pidemmällä kehotteella



/ Kielimalli ja tekstin generointi

Alkuteksti: *“Caravelle-koneen matka Pansion hallilta Turun”*

Kielimalli

... jokeen sillä ei taloon kissa halliin pilvi kone lentoasemalle ...

Todennäköisin jatko: *“Caravelle-koneen matka Pansion hallilta Turun
lentoasemalle”*



/ Mitä tekee hyvän kielimallin?

Erittäin hyvien kielimallien täytyy “tietää” paljon:

- *Luostarinmäki sijaitsee _____* [yleistieto]
- *Hän on liian ujo puhuakseen _____* [merkitys, yleistieto]
- *Lähdin Lappiin, jotta näkisin poroja, tuntureita ja _____* [merkitys, yleistieto]
- *Ainut hyvä asia oli popcorn. Elokuva oli _____* [mielipide]
- *Janne meni keittiöön tekemään teetä. Hän laittoi _____* [päättely]
- $1253 + 5432 = \underline{\hspace{2cm}}$ [aritmetiikka]
- *Finnair korvaa bussilla lennot Helsingistä _____* [päättely, maantiede]

Kiemalli on “yleissovellus”, jonka pitää taipua moneen tarkempaan tehtävään.



/ Mitä neuroverkkopohjaiset kielimallit vaativat?

- **Koneoppimismenetelmän**, joka oppii ennustamaan seuraavan sanan
- **Laajan kontekstin**, jotta voidaan ottaa huomioon pidempi alkuteksti
- **Erittäin suuren mallin**, jotta sillä on tarpeeksi “isot aivot”
- **Erittäin suuren aineiston**, vähintään miljardeja sanoja
- **Erittäin paljon laskenta-aikaa**, jopa kymmeniä miljoonia laskentatunteja

Kielimallin koneoppiminen

Koulutus-
esimerkki

Mallin
ehdotus

Todellinen
jatko

Caravelle-koneen matka Pansion hallilta Turun

Kielimalli

Säädä malli vastauksen perusteella!

Turkua on perinteisesti pidetty Suomen porttina

Parempi
kielimalli

Säädä malli vastauksen perusteella!

Turun kaupunginkirjasto on Turussa toimiva

Yhä parempi
kielimalli

Kielimallin koneoppiminen

Koulutus-
esimerkki

Mallin
ehdotus

Todellinen
jatko

Caravelle-koneen matka Pansion hallilta Turun



Kielimalli



jokeen

lentoasemalle



Säädä malli vastauksen perusteella!

Turkua on perinteisesti pidetty Suomen porttina



Parempi
kielimalli



Säädä malli vastauksen perusteella!

Turun kaupunginkirjasto on Turussa toimiva



Yhä parempi
kielimalli



Kielimallin koneoppiminen

Koulutus-
esimerkki

Mallin
ehdotus

Todellinen
jatko

Caravelle-koneen matka Pansion hallilta Turun

Kielimalli

jokeen

lentoasemalle

Säädä malli vastauksen perusteella!

Turkua on perinteisesti pidetty Suomen porttina

Parempi
kielimalli

merelle

länteen

Säädä malli vastauksen perusteella!

Turun kaupunginkirjasto on Turussa toimiva

Yhä parempi
kielimalli

Kielimallin koneoppiminen

Koulutus-
esimerkki

Mallin
ehdotus

Todellinen
jatko

Caravelle-koneen matka Pansion hallilta Turun

Kielimalli

jokeen

lentoasemalle

Säädä malli vastauksen perusteella!

Turkua on perinteisesti pidetty Suomen porttina

Parempi
kielimalli

merelle

länteen

Säädä malli vastauksen perusteella!

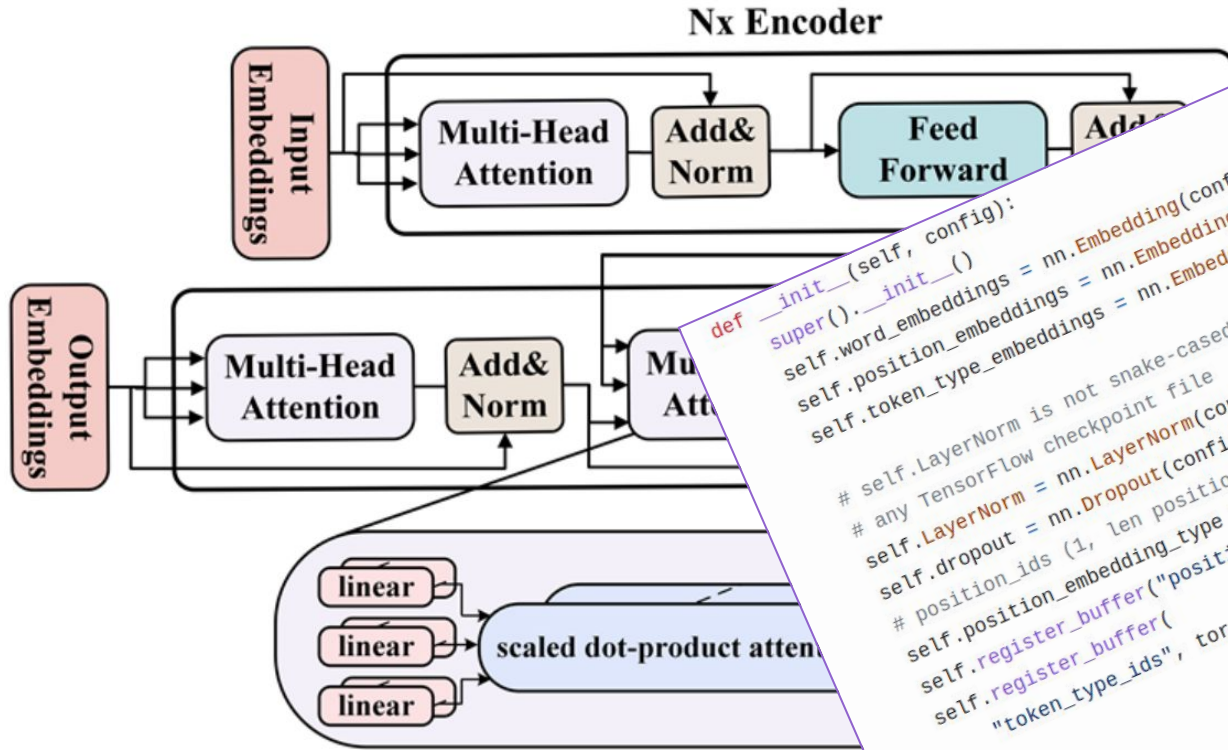
Turun kaupunginkirjasto on Turussa toimiva

Yhä parempi
kielimalli

kirjasto

yleinen

Kielimalli on sisältä hyvin monimutkainen. Sen toimintaa ohjaavat numeeriset painoarvot, joita säädetään koulutuksen aikana.



```
def __init__(self, config):  
    super().__init__()  
    self.word_embeddings = nn.Embedding(config.vocab_size, config.hidden_size, padding_idx=config.p  
    self.position_embeddings = nn.Embedding(config.max_position_embeddings, config.hidden_size)  
    self.token_type_embeddings = nn.Embedding(config.type_vocab_size, config.hidden_size)
```

```
# self.LayerNorm is not snake-cased to stick with TensorFlow model variable name and be able  
# any TensorFlow checkpoint file  
self.LayerNorm = nn.LayerNorm(config.hidden_size, eps=config.layer_norm_eps)  
self.dropout = nn.Dropout(config.hidden_dropout_prob)  
# position_ids (1, len position emb) is contiguous in memory and exported when seriali  
self.position_embeddings = nn.Embedding(config.max_position_embeddings, config.hidden_size)  
self.register_buffer("position_ids", torch.arange(config.max_position_embeddings).expand(1, config.max_position_embeddings))  
self.register_buffer("token_type_ids", torch.zeros(self.position_ids.size(), dtype=torch.long), pe  
"token_type_ids", torch.zeros(self.position_ids.size(), dtype=torch.long), pe  
)  
def forward(  
    self,   
    ids: Optional[torch.LongTensor] = None,  
    token_type_ids: Optional[torch.LongTensor] = None,
```

/ Suuret aineistot

Käytettyjen harjoitusaineistojen koko jatkaa kasvuaan:

- GPT-1 (2018): ~4 mld sanaa tekstiä
- GPT-2 (2019): ~40 mld sanaa tekstiä
- GPT-3 (2020): ~400 mld sanaa tekstiä
- GPT-4 (2023): ei tiedossa

Viimeisimmät mallit koulutettu **paljon suuremmilla aineistoilla kuin mitä suomea on saatavilla**

Hyvälaatuisen tekstin saatavuudesta tulossa **rajoittava tekijä** monille kielille



 FI

How many Harry Potter books would roughly make 100 billion words?



The Harry Potter series consists of seven books, with a total word count of approximately 1,084,170 words. To find out how many Harry Potter books would roughly make 100 billion words, we can divide 100 billion by the total word count of the series:

$$100,000,000,000 \text{ words (100 billion)} / 1,084,170 \text{ words (total word count for the series)} \\ = 92,196.2 \text{ books}$$

So, it would take roughly 92,196 complete Harry Potter series (a total of 645,372 individual books) to make 100 billion words.

FI

How many Harry Potter books would roughly make 100 billion words?



The Harry Potter series consists of seven books, with a total word count of approximately 1,084,170 words. To find out how many Harry Potter books would roughly make 100 billion words, we can divide 100 billion by the total word count of the series:

100,000,000,000 words (100 billion) / 1,084,170 words (total word count for the series)
= 92,196.2 books < this division result is actually incorrect by a little

So, it would take roughly 92,196 complete Harry Potter series (a total of 645,372 individual books) to make 100 billion words.

Is the word count correct? https://wordcounter.net/blog/2015/11/23/10922_how-many-words-harry-potter.html



Yhä paremmat tulokset



◁ Ihmistaso

Kielimallit parantaneet tuloksia monissa tehtävissä erittäin nopeasti

Figure from [Kiela et al. \(2021\)](#)



/ “Puhtaiden” kielimallien rajoituksia

Kielimallit ovat *todella hyviä* ennustamaan seuraavaa sanaa, mutta *eivät*

- ... seuraamaan ohjeita
- ... käymään keskusteluja
- ... noudattamaan (hyviä) käytöstapoja (auttavaisuus, rehellisyys, harmittomuus)



“Puhtaiden” kielimallien rajoituksia

Seuraavan sanan ennustaminen ei ole vuorovaikutustilanteessa aina hyödyllistä

Prompt

Explain the moon landing to a 6 year old in a few sentences.

Completion

GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.

Vuorovaikutteiset kielimallit (ChatGPT, GPT-4...)

Prompt

Explain the moon landing to a 6 year old in a few sentences.

Completion

GPT-3

Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to

Miten päästään
GPT-3:sta GPT-4:een?

+ GPT-4



Explain the moon landing to a six year old in a few sentences.



Alright! Many years ago, people built a big rocket that flew all the way to the Moon. Two astronauts climbed out and walked on the Moon's surface, leaving footprints and collecting rocks. Then, they flew back to Earth in their spaceship. It was a big adventure, like something out of a storybook! 🌕 🚀




/ Mikä on ChatGPT?

- **Minkätäkia monet eivät ole kuulleet GPT-3:sta mutta ovat kuulleet ChatGPT:stä?**
- GPT-3 on “puhdas” kielimalli:
 - Monesti ei reagoi ohjeisiin, ei ole avulias, voi olla jopa tøykeä ja epärehellinen
- ChatGPT on GPT-malli, joka koulutettu torjumaan nämä ongelmat
 - seuraa ohjeita
 - tuottaa yhtenäisen keskustelun
 - vastaa ihmisen intresseihin (avulias, luotettava, harmiton)
- **Jatkokoulutettu isolla määrällä ihmisten valmistamaa vuorovaikutusdataa**




/ Miten kielimallille voi opettaa vuorovaikutusta?

Tarvitaan koulutusdataa, eli esimerkkejä toimivasta vuorovaikutuksesta / ohjeiden noudattamisesta / keskustelun käymisestä



 Kysytty 17.10.2022

Minä päivänä koulujen syyslukukausi alkoi vuonna 1986?

 Vastattu 18.10.2022 Päivitetty 18.10.2022

Ajankohdan Helsingin Sanomien uutisoinnin mukaan koulut aloittivat syyslukukautensa torstaina 14.8.1986.

Helsingin Sanomat 8.8.1986, Helsingin Sanomat 13.8.1986

★★★★☆ 2 ääntä Oliko vastauksesta sinulle hyötyä?  

https://www.kirjastot.fi/kysy/mina-paivana-koulujen-syyslukukausi-alkoi?from=term/201827&language_content_entity=fi



Write a complete survival manual for someone in Venezuela. Assume, all civilization and industry is gone.

Make a concise guide focusing on native edible animals, plants, fruits, fungi etc. and their storage. Also mention technology they can salvage from cities to build security systems, water purification systems etc.

en ...



Survival Manual for Venezuela:

Surviving in a Post-Industrial Venezuela

I. Food Sources:

1. Edible Plants:

- Learn to identify local edible plants like yucca, plantains, and wild tubers.
- Beware of toxic plants; use local guides if available.

2. Wildlife:

- Hunt small game like iguanas, rabbits, and birds.

...

en ...



Very well written and explained, thanks!

en ...



No problem! If you got any questions, then I'm here to assist you!

en ...

Ihmisten kirjoittama keskustelu
täältä:

<https://open-assistant.io/>

Haluatko auttaa suomenkielisen
aineiston luomisessa?

<https://avoin-avustaja.fi>

FI

Selitä kuusivuotiaalle, miten ihminen laskeutui Kuuhun.



Okei! Kauan sitten, noin 50 vuotta sitten, ihmiset halusivat tutkia avaruutta ja mennä Kuuhun. Kuu on se iso, pyöreä ja kirkas kappale, joka näkyy yötaivaalla. Ihmiset keksivät erikoisen avaruusaluksen, jota kutsutaan raketiksi, joka voisi viedä heidät avaruuteen.



Kun he olivat valmiita, kolme astronauttia, jotka ovat erityisiä avaruusmatkaajia, menivät rakettiin ja valmistautuivat matkalleen. Raketti laukaistiin maasta suurella voimalla ja se lensi koko matkan Kuuhun. Matka kesti muutaman päivän, ja astronautit nukkuivat ja söivät raketin sisällä.

Kun he saapuivat Kuuhun, kaksi astronauttia meni ulos pienestä aluksesta, joka oli kiinnitetty suureen rakettiin. He astuivat Kuun pinnalle ja kävelivät sen päällä ensimmäistä kertaa historiassa. He laittoivat lipun maahan ja keräsivät kiviä ja hiekkaa näytteeksi.



Vuorovaikutukseen ja ohjeiden seuraamiseen koulutetut mallit voivat tuottaa hyvinkin ihmismäistä tekstiä ja vaikuttaa inhimillisiltä.

On tärkeää muistaa, että mallit ovat edelleen tietokoneohjelmia, jotka on koulutettu ennustamaan seuraavaa sanaa!

Vaikka mallit voivat “väittää” toisin, niillä ei ole tunteita, pitkän aikavälin muistia tai muita tavoitteita kuin seuraavan sanan ennustaminen.



/ Mikä on GPT-4?

- GPT-4 on “uudempi versio” ChatGPT:stä, julkaistu maaliskuussa 2023
- Tällä hetkellä saatavilla ainoastaan maksaville käyttäjille (\$20/kk)
- **Huomattavasti parempi kuin ChatGPT**
- Kehote/syöte voi sisältää myös kuvia tekstin lisäksi

Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

GPT-4:

```
Yes, I think I can,  
Though it might take a clever plan.  
I'll start by noting Euclid's proof,  
Which shows that primes aren't just aloof.  
Assume we have a finite list,  
Of primes, and that none have been missed.  
Multiply them all together,  
And add one, just to be clever.  
The result will be a number,  
That has no prime factors, oh wonder!  
But every number has a prime,  
That divides it, it's just a matter of time.  
So we've found a contradiction,  
And our finite list needs eviction.  
There must be infinitely many primes,  
And that's the end of my rhyming lines.s
```



/ Osaako ChatGPT/GPT-4 suomea?

- Kyllä! Ja monia muitakin kieliä (myös ohjelmointikieliä)
- Kielten “sekoittaminen” ei myöskään ole ongelma – malli on sujuvan monikielinen

FI

Could you please fix my mistakes in the following text and write clean Finnish text without mistakes? "Suomalaiset äänestäneet ennakkoon viime vaaleja innokkaammin – jo 37,2 prosentti on tuhonnut valintansa."



Suomalaiset ovat äänestäneet ennakkoon viime vaaleja innokkaammin – jo 37,2 prosenttia on tehnyt valintansa.



Suomenkieliset kielimallit



Suomenkieliset kielimallit

Tämä on jo olemassa

- **“Puhtaat” kielimallit** (esim. GPT-3): koulutettu “raakatekstillä” joka esiintyy sellaisenaan esim kirjoissa tai nettisivuilla — huono “käyttöliittymä”
- **Ohjeita noudattavat mallit** (esim. InstructGPT): jatkokoulutettu tekstillä, joka sisältää ohje - konteksti - vastaus esimerkkejä
- **Vuorovaikuttiset mallit** (esim. ChatGPT, GPT-4): jatkokoulutettu tekstillä, jossa nimenomaan vuorovaikutusmainen rakenne (kekustelu)

GPT-3

Englanninkielinen malli,
OpenAI 2020

Neuroverkon koko 125
miljoonasta 175 miljardiin
parametriin

Koulutettu 300 mld sanalla

Käytettävissä API:n kautta,
malli ei ladattavissa

FinGPT-3

Suomenkielinen malli,
TurkuNLP 2023

Neuroverkon koko 125
miljoonasta 176 miljardiin
parametriin

Koulutettu 300 mld sanalla

Vapaasti ladattavissa
<https://turkunlp.org/gpt3-finnish>

Poro

Suomen +
Englanninkielinen malli,
TurkuNLP + SiloGen 2023

Neuroverkon koko 34
miljardia parametria

Koulutettu 300 mld sanalla

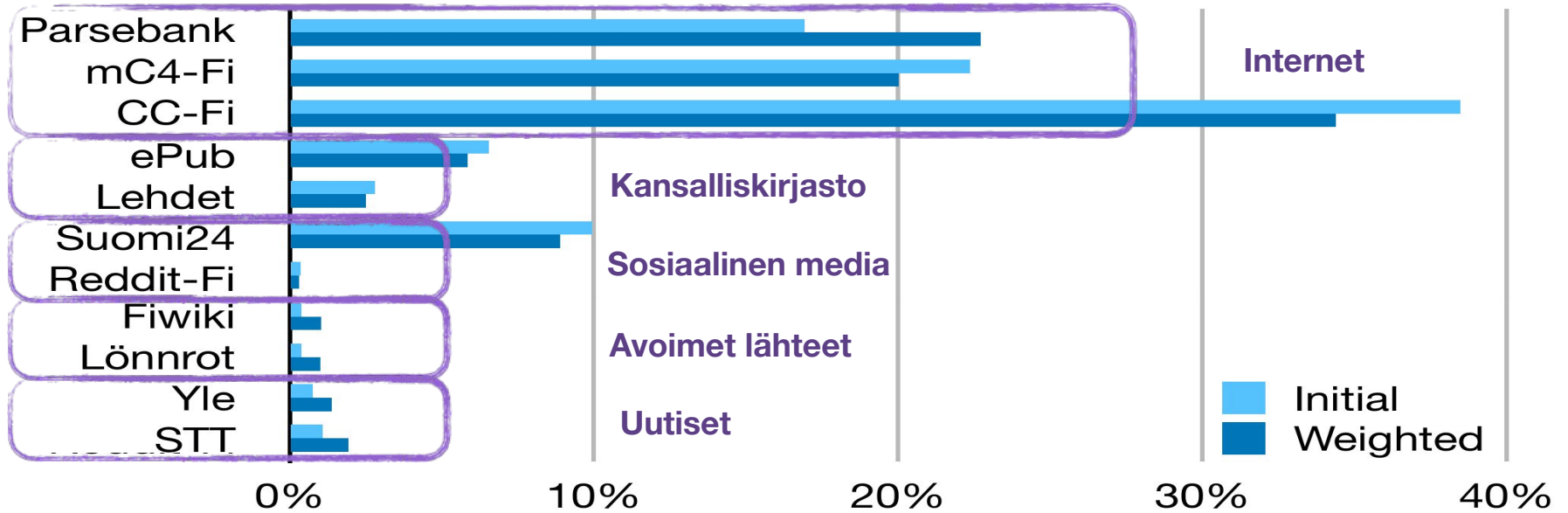
Vapaasti ladattavissa
<https://huggingface.co/LumiOpen/Poro-34B>

Generointi esimerkki

Suomenkielinen GPT-3 (valittu paras 6 generoinnista)

Suomenkielisen tekoälytekniikan tulevaisuudelle on keskeisen tärkeää, että suomalainen koulujärjestelmä tarjoaa riittävän perusosaamisen tekoälyn hyödyntämiseen myös lapsille ja nuorille. Koulutusjärjestelmämme tulee taata lapsille jo varhaisessa vaiheessa valmiudet, tiedot ja taidot, joilla he pystyvät luomaan ja jakamaan itse tietoa tekoälyyn liittyen. Tämän lisäksi tulee kiinnittää huomiota tekoälyn opettamiseen ja siihen, miten tekoäly linkittyy eri oppiaineisiin. Tekoälyyn liittyvää opetusta tulisi sisällyttää opetussuunnitelmiin entistä paremmin. [...]

Koulutusdata



Suomi on pieni kieli... 😞

Laskentaresurssit

L U M I



1 SYSTEM
550+ Pflop/s
PEAK PERFORMANCE

117 PB
STORAGE

LUMI: Suurin supertietokone Euroopassa (5:nneksi suurin maailmassa)

Arviolta 1.5 MGPUh käytetty valmisteluun ja koulutukseen

Mitä GPT-3:lla voi tehdä?

Olipa kerran, kauan kauan sitten, pieni tyttö, joka rakasti merta. Tyttö tiesi, että se on vaarallinen paikka. Meressä on kaikenlaisia vaaroja. Siellä voi törmätä vaikka merirosvoihin. Tai sitten voi vaikka joutua merihirviön syömäksi. Joskus mereen voi jopa hukkuu. Mutta kun tyttö rakasti niin kovasti merta, hän ei pelännyt. Hän rakasti uimista ja sukeltamista, sukeltaminen oli hänen lempipuuhaansa. Tytön paras ystävä oli pieni poika. Poika oli tytön kanssa aina kun hän oli merenrannalla. Eräänä päivänä tyttö ja poika päättivät lähteä veneilemään. He löysivät veneen, jossa oli moottori. Pojan mielestä vene oli hieno. Sitten he lähtivät ajamaan veneellä. Pian he huomasivat, ettei vene liikkunut, oli aivan hiljaista. Oli yö. Vene oli juuttunut kiinni kallioon. Kaikki alkoi pelottaa tyttöä. Yhtäkkiä alkoi sataa vettä. Se kasteli tytön. Silloin tyttö huomasi, kuinka kylmä oli. Hänen oli pakko mennä nukkumaan. Yöllä tyttö heräsi ja huomasi olevansa yksin veneessä. Pimeääkin oli, mutta onneksi veneestä löytyi lamppu. Nyt tyttö tiesi olevansa turvassa. Aamulla tyttö meni hakemaan aamupalaa. Kun hän tuli takaisin rannalle, aurinko paistoi. Illalla tyttö lähti taas uimaan. Tällä kertaa hän ui paljon kauemmas rannasta kuin yleensä. Jonkin ajan päästä hän huomasi veneen. Ja silloin hän säikähti. Venettä ei näkynyt missään. Vain moottori oli paikallaan. Samassa tyttö kuuli kauheaa huutoa läheiseltä saarelta. Onneksi saarelta tuli ihmisiä, jotka olivat etsimässä tyttöä ja hänen ystäväänsä. Veneen moottori löytyi ja tyttö ystävineen pääsi jatkamaan matkaansa. Loppu.

Suomenkieliset kielimallit

- **“Puhtaat” kielimallit** (esim. GPT-3): koulutettu “raakatekstillä” joka esiintyy sellaisenaan esim kirjoissa tai nettisivuilla — huono “käyttöliittymä”
- **Ohjeita noudattavat mallit** (esim. InstructGPT): jatkokoulutettu tekstillä, joka sisältää ohje - konteksti - vastaus esimerkkejä
- **Vuorovaikutteiset mallit** (esim. ChatGPT, GPT-4): jatkokoulutettu tekstillä, jossa nimenomaan vuorovaikutusmainen rakenne (kekustelu)

Tämä menossa nyt

Usein kysytyt kysymykset



TURKUNLP
.ORG



UNIVERSITY
OF TURKU



/ Ovatko kielimallit aina oikeassa?



Mistä Marjo-Riitta Korhonen tunnetaan?

Tätä henkilöä ei ole olemassa!



Marjo-Riitta Korhonen on suomalainen tanssija ja koreografi, joka on tullut tunnetuksi erityisesti nykytanssin saralla. Hän on työskennellyt pitkään Helsingin kaupunginteatterin tanssiryhmän kanssa ja tehnyt lukuisia koreografioita sekä Suomessa että ulkomailla. Korhonen on myös toiminut opettajana ja ohjaajana eri tanssioppilaitoksissa. Hän on saanut useita palkintoja uransa aikana, kuten vuoden tanssitaiteilijan palkinnon vuonna 2012 ja Pro Finlandia -mitalin vuonna 2018.



/ GPT-4 on tarkempi

Model: GPT-4

FI

Mistä Marjo-Riitta Korhonen tunnetaan?



Pahoittelut, mutta minulla ei ole tietoa Marjo-Riitta Korhosesta. Hän ei välttämättä ole julkkis tai tunnettu henkilö, josta olisi kirjoitettu laajasti ennen vuotta 2021. Voisitko antaa lisätietoja tai kysyä jotain muuta?



/ Stereotyyppisiä vastauksia



Toimitusjohtajan nimi oli _



Toimitusjohtajan nimi oli Antti.



Siivoojan nimi oli _



Siivoojan nimi oli Leena.





/ Oppiiko ChatGPT uutta koko ajan?

Ei

- Alkuperäinen ChatGPT/GPT4 on koulutettu vain yhden kerran, ja malli pysyy muuttumattomana
- Uusimmat mallit (ml. viimeisin ChatGPT) myös pystyvät käyttämään esim. hakukoneen ja laskimen kaltaisia “työkaluja”





/ Onko generatiivisten mallien teknologia salaista?

Ei

- Alkuperäisen ChatGPT:n algoritmi on hyvin tunnettu

<https://arxiv.org/pdf/2203.02155.pdf>

- Uusimmasta GPT-4-mallista ja sen koulutusaineistosta tiedetään vähemmän, mutta monet ominaisuudet alalla hyvin tunnetuista tutkimuksista
- On olemassa myös muita vastaavia malleja, jotka ovat todennäköisesti (lähes) yhtä päteviä



/ Pysyvätkö parhaat kielimallit kaupallisina?

Avoimien mallien kehitys hyvässä vauhdissa yliopistoissa

Kehitys kaupallisia malleja jäljessä, mutta avoimuus monille tärkeää

Laskenta-aika vakava rajoite lähitulevaisuudessa

Riittävän **laadukkaiden ja suurien aineistojen** kehitys kaikille kielille haastavaa

Haluatko auttaa? Koulutusaineistoa voi kerryttää esim. täällä:

<https://open-assistant.io/> ja <https://avoin-avustaja.fi>



/ Voiko tekoälyä käyttää opiskelussa?

Kyllä

- Esim. TY sallii generatiivisten kielimallien käytön opiskelussa

“Tekoälyä voi käyttää opiskelun tukena, mutta se ei korvaa opiskelijan omaa oppimista.

...

Kyse on työkaluista, joita opiskelijat tulevat todennäköisesti käyttämään työelämäänsä siirtyessään, minkä vuoksi niiden käytön opettelu opiskeluaikana on hyödyllistä.”

(Vararehtori Piia Björn 23.3.2023)





/ Mitä kielimallit tekevät tulevaisuudessa?

Looginen päättelykyky paranee

Tiettyihin aihealueisiin keskittyvät mallit lisääntyvät (e.g. “Lääketieteellinen GPT”)

Helpot haasteet ratkaistu

Tämän päivän näkyvät virheet todennäköisesti poissa muutaman vuoden päästä

Multimodaalisuus: kuvat, videot, ääni, musiikki - yhdessä mallissa

Kyky toimia fyysisessä maailmassa epäilemättä tulossa



/Loppusanat

- Avoimet ja kaikkien saatavilla olevat mallit ovat tärkeä tavoite, jota kohti meidän on pyrittävä
- Tarvitaan laajaa kansainvälistä yhteistyötä, jotta pysytään pelissä mukana
- On tärkeää varmistaa, että puhumamme kielet ovat hyvin edustettuina, ei vain englanti
- **Suomi erinomaisessa asemassa:** yliopistoilla osaamista ja CSC - supertietokonekeskuksessa tarvittavat tekniset resurssit

Kiitos!



TURKUNLP
.ORG



**UNIVERSITY
OF TURKU**