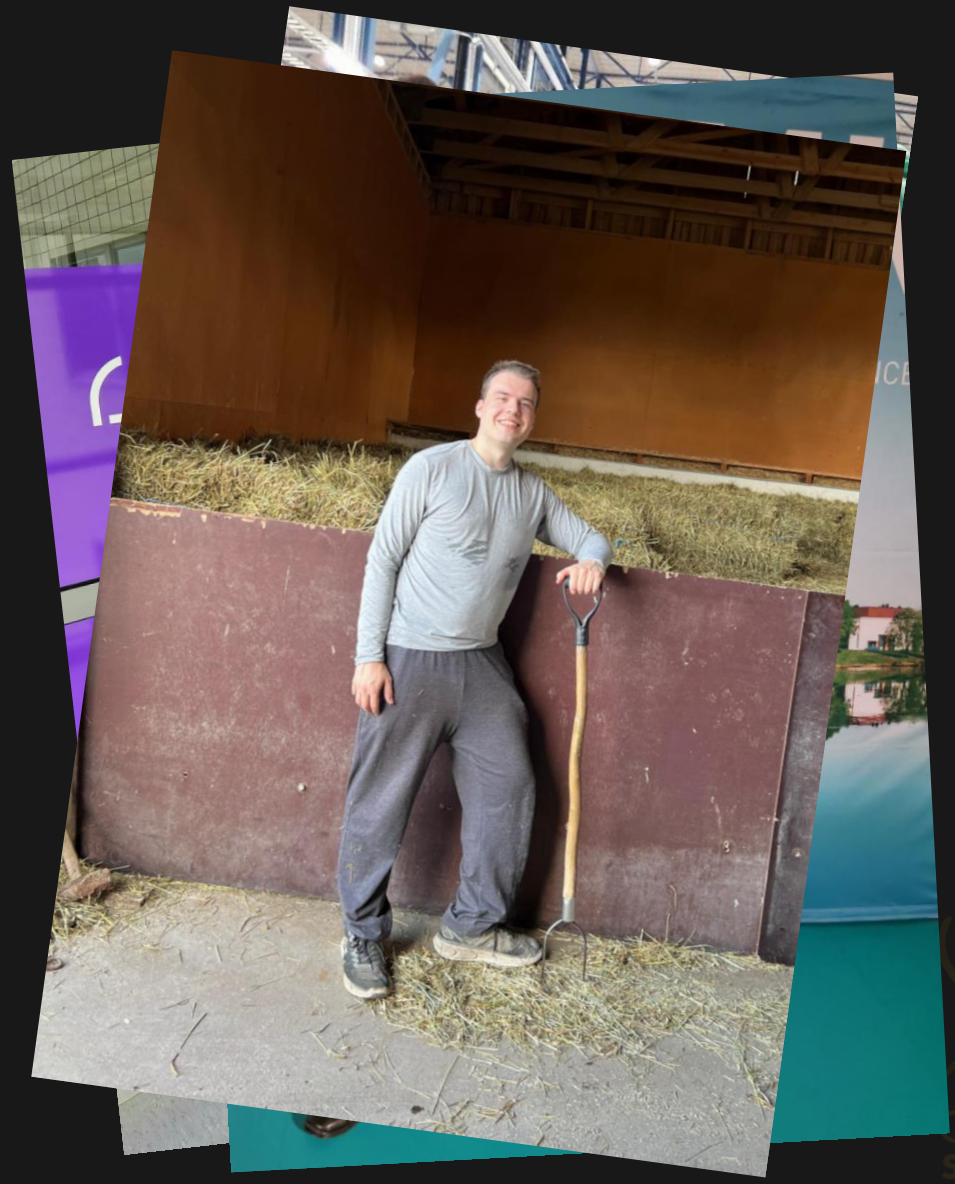
Paikalliset kielimallit tekevät MITÄ!?!

GPT Lab Seinäjoki

42 views • right now

Ketä? Juha Ala-Rantala

- **Tutkimusavustaja**
  GPT Lab Seinäjoki (2024->)
- **Tietotekniikan DI maisteritutkinto**
  Tampereen Yliopisto (2023->)
- **Ohjelmistosuunnittelija / -kehittäjä**
  Prima Power / Finn Power (2018-2022)
- **Tietotekniikan insinööri**
  SeAMK (2015-2019)
- **Moon Härmästä**
  “lähdin maatilaa pakoon it alalle
  – mutta mitä kauemmin on it alalla
  – sitä enemmän alkaa maatila kutsua”



GPT Lab Seinäjoki

TIETOPAKETTI:

PAIKALLISET

KOODAAVAT

KIELIMALLIT

GPT Lab
Seinäjoki

Inspiraatio-osuus

Konkreettinen esimerkki
mitä tullaan rakentamaan

(avaa vscode nyt)

SISÄLTÖ

GPT Lab
Seinäjoki

Sisältö

- Tavoitteet
- Taustaa
- Paikalliset kielimallit
- Työkalun rakentaminen
- Työpaja
- Johtopäätökset

GPT Lab
Seinäjoki

TAVOITTEET

GPT Lab
Seinäjoki

Tavoitteet

Jakaa yleissivistävää tietoa (paikallisista) kielimalleista:

Parametrit, kvantisointi, konteksti ja niiden merkitys

GPT Lab Seinäjoki

Tavoitteet

Vertailla keskeisiä tekoälyratkaisuja ohjelmoinnissa: ChatGPT, GitHub Copilot vs paikalliset kielimallit

GPT Lab Seinäjoki

Tavoitteet

Esitellä käytännön esimerkkejä tekoälyn käytöstä ohjelmointitehtävissä:

“älykkäämpi" autocomplete, promptaus, testaaminen

GPT Lab Seinäjoki

Tavoitteet

Opastaa asennuksessa ja konfiguroinnissa:

paikallisen kielimallin käyttöönotto ohjelmointia varten

GPT Lab Seinäjoki

Tavoitteet

Innostaa keskustelua tekoälyn mahdollisuuksista, kehityksestä ja käytöstä:

rohkaista kokeilemaan uusia työkaluja

GPT Lab Seinäjoki

TAUSTAA

GPT Lab
Seinäjoki

Taustaa

Mitä tutkittu, miten tutkittu, miksi tutkittu?

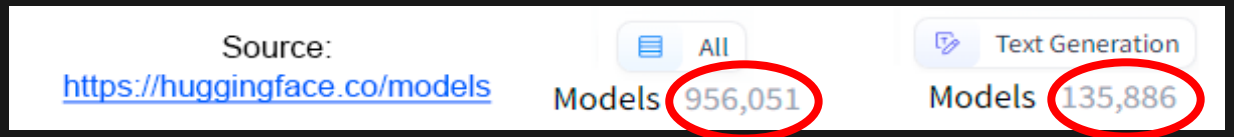
1. Voiko paikalliset kielimallit korvata isot ja maksulliset?
2. Mitä paikallisia kielimalleja on olemassa ohjelmointia varten?
3. Mitä tarkoittavat termit kuten parametrimäärä, kvantisointi, konteksti?
4. Miten niitä voi ajaa ja käyttää omalla koneella?
5. Mitä vaatii koneelta pyörittää mitäkin mallia? Laskentateho, muisti?
6. Miten eri mallit on lisensoitu? Mitkä ovat yleisimmät?
7. Mitä VSCode plugineita on olemassa? Miten ne eroavat toisistaan?

GPT Lab Seinäjoki

Taustaa

Mikä on paikallisten (koodaavien) kielimallien tilanne?

- Kehittyvät nopeasti, jatkuvasti tulee uusia ja parempia
- Isoin alusta kaikille malleille on Huggingface
  Vuosi sitten oli 300.000 mallia, kuukausi sitten, nyt: Models 1,034,534



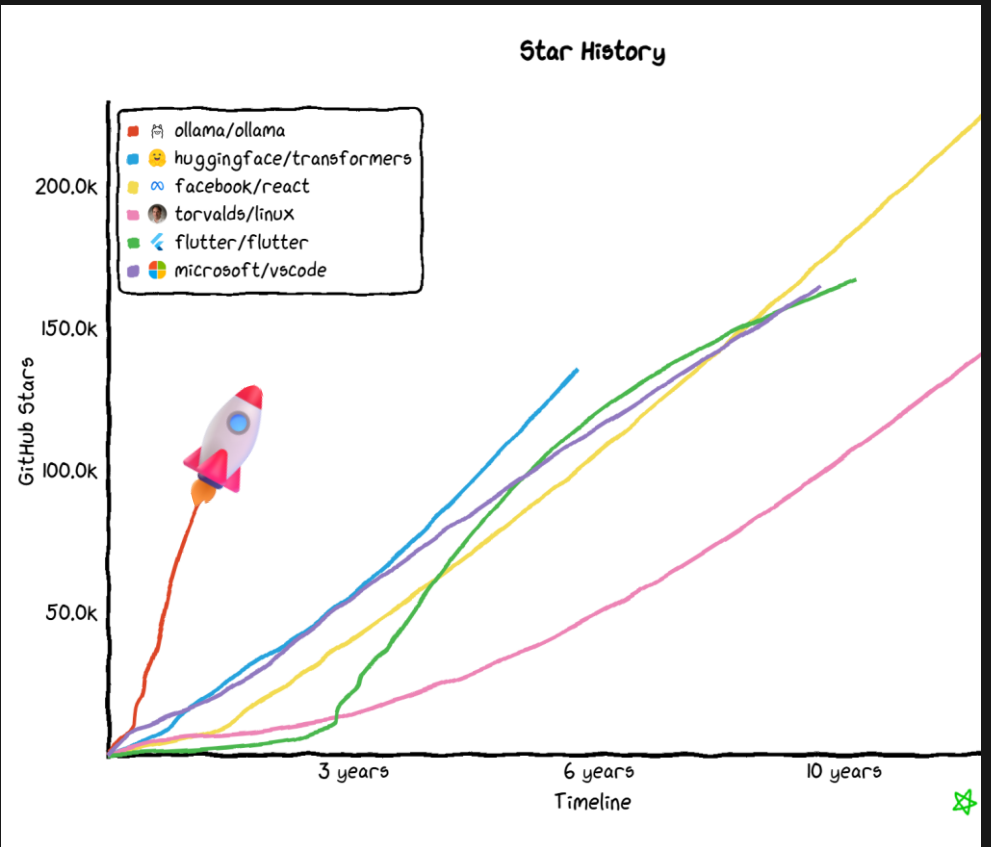
- Ohjelmia (provider) mallien ajamiseen on paljon:
- VSCode plugineita paljon:

Taustaa

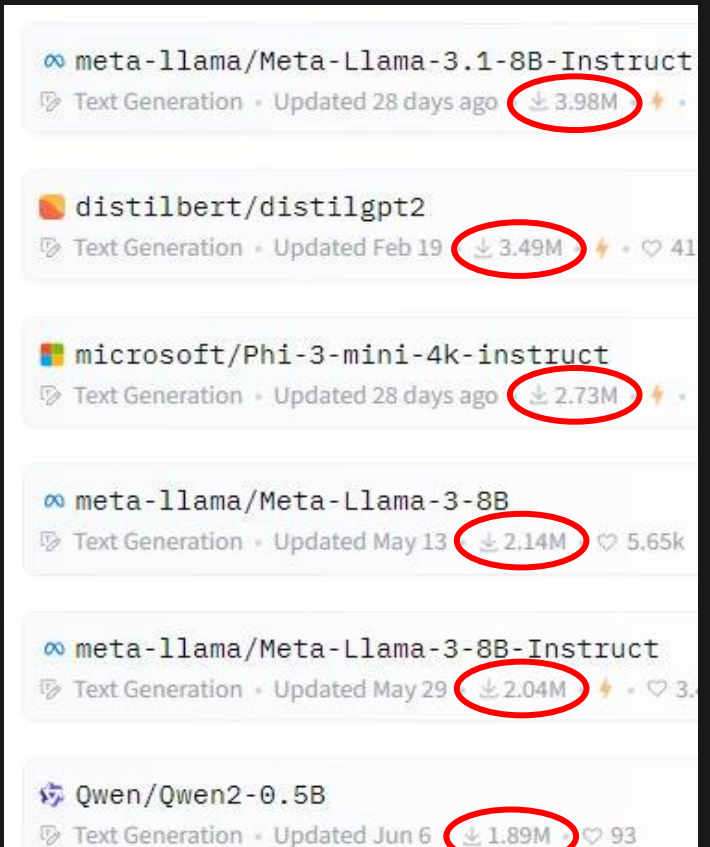
Ollama yksi nopeimmin kasvavista GitHub repoista ikinä

Suosituimpien mallien latausmäärät miljoonissa

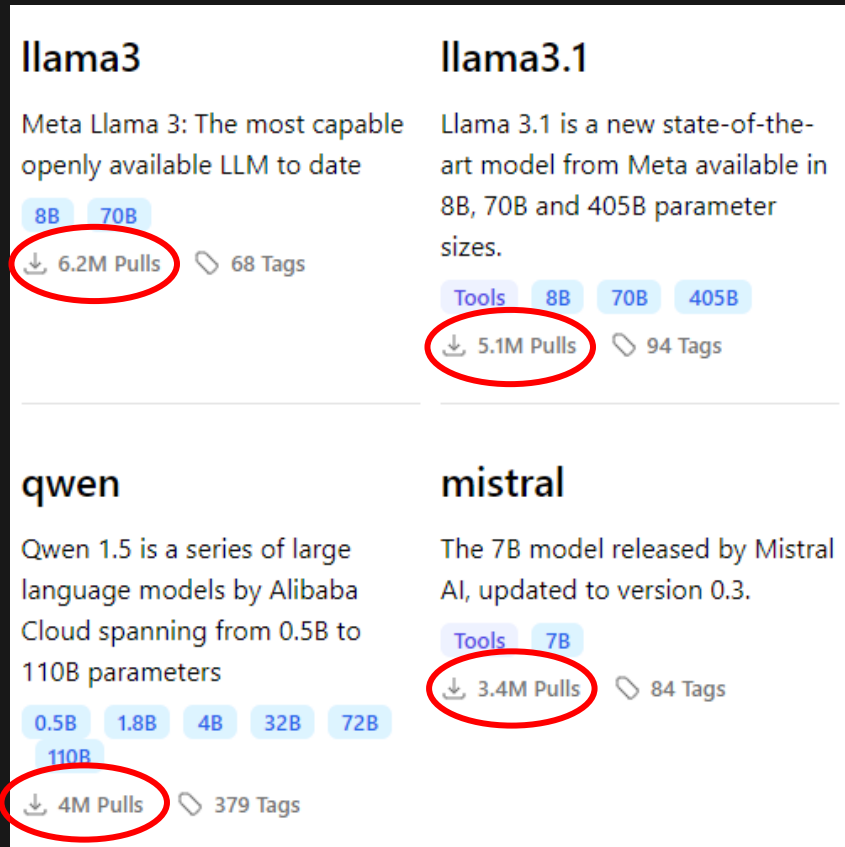
github.com



huggingface.co/models



ollama.com/library



Taustaa

Mitä tulee paikallisista malleista huomioida?

- Laskentatehojen vaativuus
  - Kielimallien ajaminen vaatii laskentatehoa ja muistia
- Tarkkuus ja konteksti
  - Pienemmillä malleilla on rajallisempi kyky ymmärtää kyselyitä ja tuottaa vastauksia
- Kielituki
  - Pienet mallit on pääasiassa koulutettu englanninkielisellä materiaalilla
- Lisenssikysymykset
  - Useilla malleilla on omat lisenssit jotka rajoittavat niiden käyttöä

GPT Lab Seinäjoki

Taustaa

ChatGPT / GitHub Copilot / Claude VS Paikalliset Kielimallit

+ Ei vaadi laskentatehoa omalta koneelta

+ Laaja osaaminen

+ Helppo käyttöönotto

- Maksaa

- Yksityisyys / data menee palvelimelle

- Muuttuu jatkuvasti

- Internetyhteys

+ “Ilmainen”

+ Yksityisyys / kaikki pysyy omalla koneella

+ Omassa hallinnassa

+ Ei internetyhteyttä

- Vaatii laskentatehoa

- Suppea osaaminen

- Käyttöönoton monimutkaisuus

Taustaa

Paikallinen malli hakkaa ChatGPT ohjelmoinnissa?

	Qwen2.5 Coder 32B Instruct	DeepSeek Coder V2 Instruct	DeepSeek Coder 33B Instruct	CodeStral 22B	GPT-4o 2024-08.06	Claude 3.5 Sonnet 2024-10.22
HumanEval	**92.7**	88.4	79.3	78.1	92.1	92.1
MBPP	**90.2**	89.2	81.2	73.3	86.8	91.0
EvalPlus Average	**86.3**	83.8	74.9	73.5	84.4	85.9
MultiPL-E	79.4	**79.9**	69.2	70.2	79.1	83.8
McEval	**65.9**	62.9	54.3	50.5	65.8	66.5
LiveCodeBench 2024 07 - 2024 11	**31.4**	27.9	21.3	22.6	34.6	31.6
CRUXEval-O CoT	**83.4**	75.1	50.6	63.5	89.2	87.2
BigCodeBench Instruct Average	**38.3**	36.3	29.8	29.4	37.6	34.5
Aider Pass@2	**73.7**	72.9	59.4	51.1	71.4	86.5
Spider	**85.1**	81.3	73.8	76.6	79.8	74.6
BIRD-SQL	**58.4**	51.9	45.6	46.2	54.2	49.5
CodeArena v.s. GPT-4 Turbo 0409	**68.9**	57.4	16.8	21.7	69.1	78.1

GPT Lab Seinäjoki

PAIKALLISET KIELIMALLIT

GPT Lab Seinäjoki

Paikalliset kielimallit

Jos avaat https://huggingface.co/models?pipeline_tag=text-generation niin mitä näet?

Paljon sekavia nimiä malleilla, mitä ihmettä nämä kaikki oikein tarkoittavat?

Paikalliset kielimallit

Mitä pitävät sisällään?

- Kielimalli koostuu parametreista, joita säädetään painoilla, jotka määrittelevät mallin toiminnan eli miten malli tulkitsee ja käsittelee kieltä
- Parametrimäärä: kyvykkyys
- Kvantisointi (quantization): painojen tarkkuus
- Konteksti (context): mallin käsittelykapasiteetti
- Koulutusmateriaali: opettaminen
- Hienosäätö (fine-tuning): erikoistuminen
- Template: syötteen mukauttaminen
- Nämä kaikki vaikuttavan mallin ajamisen vaatimuksiin, tarkkuuteen, kykyihin, ulosannin laatuun, tiedon käsittelyyn
- Lisenssi

GPT Lab Seinäjoki

Kielimalli

Miten kielimalli toimii käytännössä? Miten se tuottaa tekstiä?

- Kielimalli on kuin tilastoihin perustuva matemaattinen ennustaja, joka laskee, mikä sana on todennäköisin seuraavaksi tekstissä

This visualization illustrates how likely different word successions are according to a large language model.

For example, if I count from ten to one; ten, nine, eight, seven, etc., you notice how the probability increases as the count progresses. However the fac stead of continuing counting down was very surprising from the lan of view.

prob	0.005%
logprob	-10.0017
␣six	92.759%
␣four	1.507%
␣and	1.007%
␣6	0.734%
␣five	0.674%
␣seven	0.617%
␣two	0.512%
␣sixth	0.354%
␣eight	0.339%
␣go	0.116%

```
const http = require('http');

const server = http.createServer((req, res) => {
  // Routing
  if (req.url === '/') {
    res.writeHead(200, { 'Content-Type': 'text/plain' });
    res.end('Hello, world!');
  } else if (req.url === '/about') {
    res.writeHead(200, { 'Content-Type': 'text/plain' });
    res.end('About page');
  } else {
    res.writeHead(404, { 'Content-Type': 'text/plain' });
    res.end('Page not found');
  }
});

const port = 3000;
server.listen(port, () => {
  console.log(`Server listening on port ${port}`);
});
```

GPT Lab Seinäjoki

Parametrit

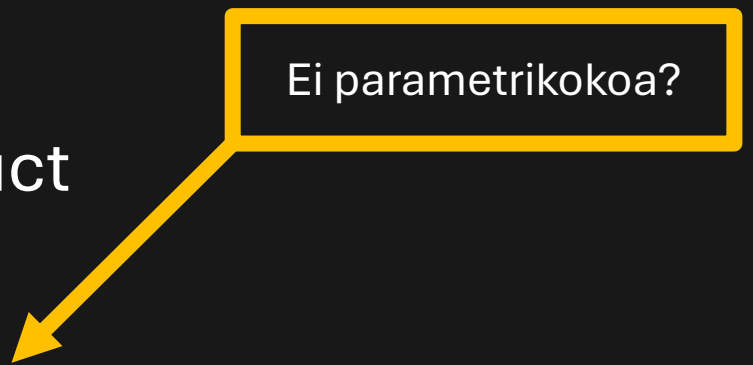
Mallin parametrimäärä löytyy usein sen nimestä

- Qwen2.5-0.5b-instruct
- Yi-Coder-1.5b-Chat-GGUF
- Zamba2-2.7b-instruct
- Polyglot-ko-3.8b
- Deepseek-coder-6.7b-instruct
- Llama-3.1-8b-instruct
- Phi-3.5-medium-4k-instruct
- Deepseek-moe-16b-base
- Llama-2-70b-chat-hf

Parametrit

Mallin parametrimäärä löytyy usein sen nimestä

- Qwen2.5-0.5b-instruct
- Yi-Coder-1.5b-Chat-GGUF
- Zamba2-2.7b-instruct
- Polyglot-ko-3.8b
- Deepseek-coder-6.7b-instruct
- Llama-3.1-8b-instruct
- Phi-3.5-medium-4k-instruct
- Deepseek-moe-16b-base
- Llama-2-70b-chat-hf

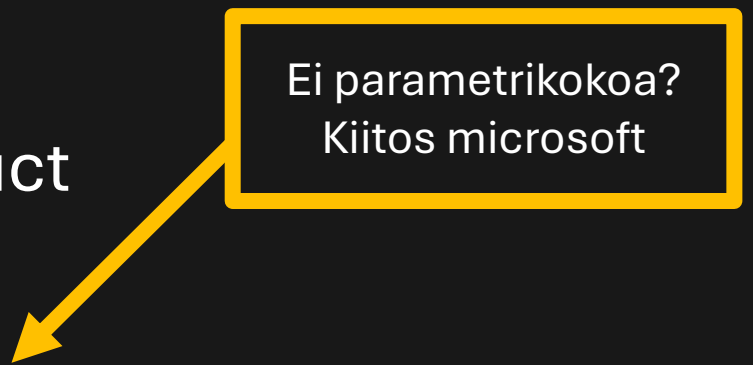


GPT Lab Seinäjoki

Parametrit

Mallin parametrimäärä löytyy usein sen nimestä

- Qwen2.5-0.5b-instruct
- Yi-Coder-1.5b-Chat-GGUF
- Zamba2-2.7b-instruct
- Polyglot-ko-3.8b
- Deepseek-coder-6.7b-instruct
- Llama-3.1-8b-instruct
- Phi-3.5-medium-4k-instruct
- Deepseek-moe-16b-base
- Llama-2-70b-chat-hf



GPT Lab Seinäjoki

Parametrit

Mallin koko, kompleksisuus ja kyvykkyys eli mihin malli pystyy

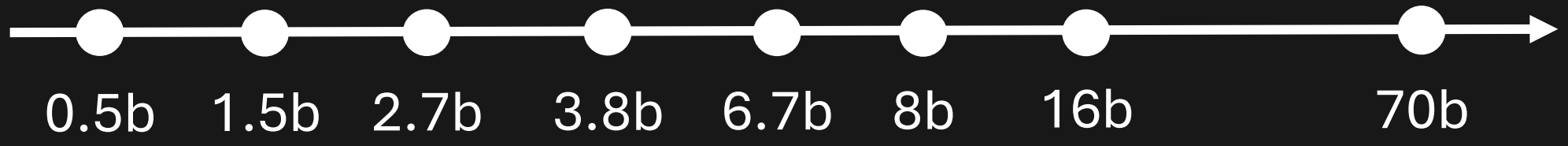
- 0.5b
- 1.5b
- 2.7b
- 3.8b
- 6.7b
- 8b
- 16b
- 70b

Mutta mitä ihmettä nämä numerot tarkoittavat?

GPT Lab Seinäjoki

Parametrit

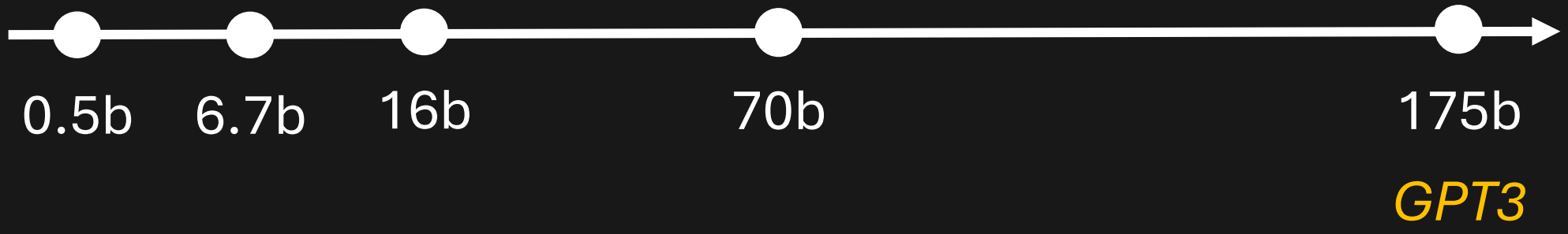
Mallin koko, miljardeja parametreja



GPT Lab Seinäjoki

Parametrit

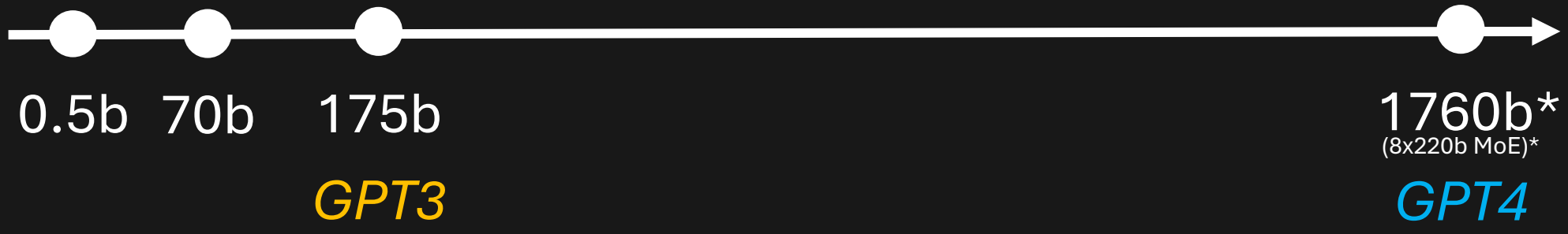
Vertailukohteeksi



GPT Lab Seinäjoki

Parametrit

There’s always a bigger fish



* Speculation https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/

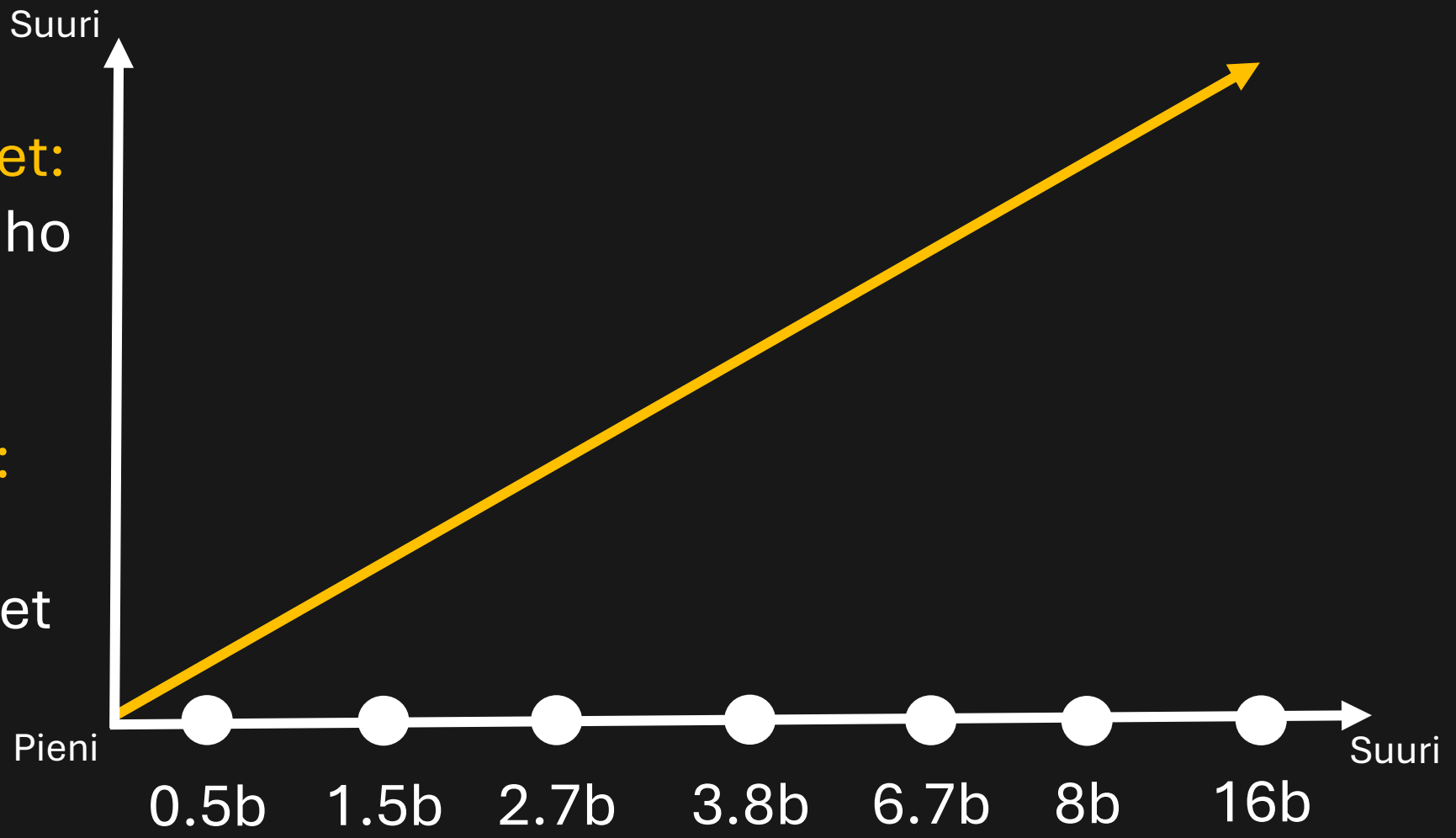
GPT Lab Seinäjoki

Parametrit

Parametrien vaikutus mallin vaatimuksiin ja ulosantiin (nyrkkisääntö)

Vaatimukset:
Laskentateho
Muisti

Ulosanti:
Laatu
Kyvykkyydet



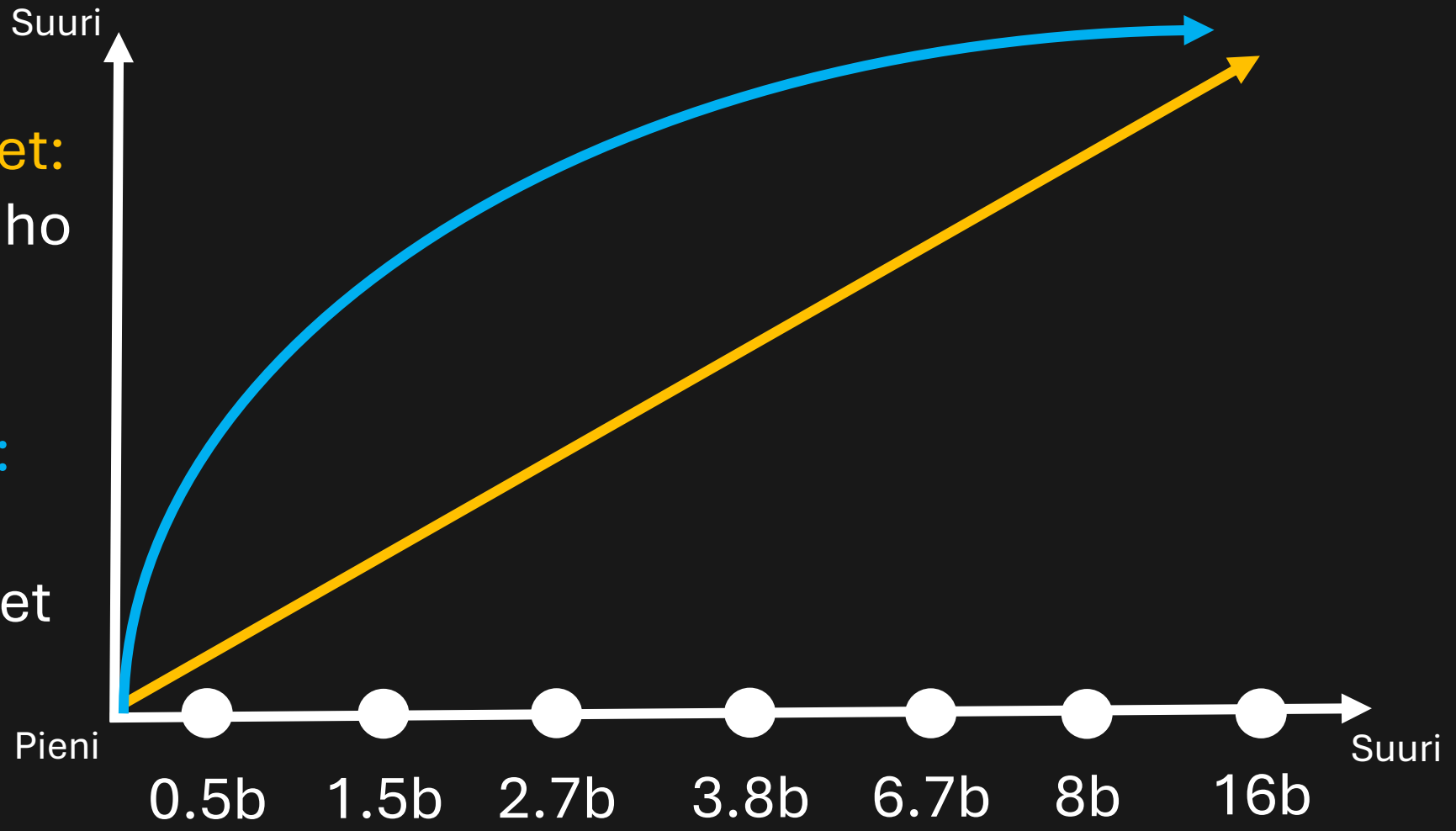
GPT Lab Seinäjoki

Parametrit

Parametrien vaikutus mallin vaatimuksiin ja ulosantiin (lähempänä totuutta)

Vaatimukset:
Laskentateho
Muisti

Ulosanti:
Laatu
Kyvykkyydet



GPT Lab Seinäjoki

Kvantisointi

Miksi malleja ajetaan kvantisointuna?

Kvantisoimalla voidaan merkittävästi vähentää laskentatehoa ja muistinkäyttöä, ilman että ulosannin tarkkuus kärsii huomattavasti

GPT Lab Seinäjoki

Kvantisointi

Paikallisten kielimallien monet eri kvantisointitekniikat ja tiedostomuodot

Tekniikka	CPU	NVIDIA GPU	AMD GPU	Apple Metal	Tiedosto-muoto	Info
GGML	✓	✓	✗	✓	.ggml	Ensimmäinen helppojakoinen kuluttajakoneilla ajettava muoto
GGUF	✓	✓	?	✓	.gguf	GGML uudempi paranneltu versio lisäominaisuuksilla
GPTQ	✗	✓	✓	✗	.gptq	Optimisoitu ajamaan malleja näytönohjaimella
AWQ	✗	✓	✓	✗	Useampi	Uudempi, parempi ja nopeampi kuin GPTQ, CPU offloading
EXL2	✗	✓	✓	✗	.exl2	Ideana ajaa isoja malleja nopeasti pienellä kvantisoinnilla
AQLM	✓	✓	✗	✗	Useampi	Myös ideana saada isoja malleja ajettavaksi pienellä kvantisoinnilla

Kvantisointi

Tässä tapauksessa meitä eniten kiinnostaa

Tekniikka	CPU	NVIDIA GPU	AMD GPU	Apple Metal	Tiedosto-muoto	Info
GGML	✓	✓	✗	✓	.ggml	Ensimmäinen helppojakoinen kuluttajakoneilla ajettava muoto
GGUF	✓	✓	?	✓	.gguf	GGML uudempi paranneltu versio lisäominaisuuksilla
GPTQ	✗	✓	✓	✗	.gptq	Optimisoitu ajamaan malleja näytönohjaimella
AWQ	✗	✓	✓	✗	Useampi	Uudempi, parempi ja nopeampi kuin GPTQ, CPU offloading
EXL2	✗	✓	✓	✗	.exl2	Ideana ajaa isoja malleja nopeasti pienellä kvantisoinnilla
AQLM	✓	✓	✗	✗	Useampi	Myös ideana saada isoja malleja ajettavaksi pienellä kvantisoinnilla

Kvantisointi

Yhden mallin kaikki kvantisoinnit (GGUF)

- Qwen2.5-1.5b-instruct-FP16
- Qwen2.5-1.5b-instruct-Q8_0
- Qwen2.5-1.5b-instruct-Q6_K
- Qwen2.5-1.5b-instruct-Q5_K_M
- Qwen2.5-1.5b-instruct-Q5_K_S
- Qwen2.5-1.5b-instruct-Q5_1
- Qwen2.5-1.5b-instruct-Q5_0
- Qwen2.5-1.5b-instruct-Q4_K_M
- Qwen2.5-1.5b-instruct-Q4_K_S
- Qwen2.5-1.5b-instruct-Q4_1
- Qwen2.5-1.5b-instruct-Q4_0
- Qwen2.5-1.5b-instruct-Q3_K_L
- Qwen2.5-1.5b-instruct-Q3_K_M
- Qwen2.5-1.5b-instruct-Q3_K_S
- Qwen2.5-1.5b-instruct-Q2_K

Kvantisointi

Kvantisointi bittitarkkuudet ja tekniikka

- Qwen2.5-1.5b-instruct-FP16
- Qwen2.5-1.5b-instruct-Q8_0
- Qwen2.5-1.5b-instruct-Q6_K
- Qwen2.5-1.5b-instruct-Q5_K_M
- Qwen2.5-1.5b-instruct-Q5_K_S
- Qwen2.5-1.5b-instruct-Q5_1
- Qwen2.5-1.5b-instruct-Q5_0
- Qwen2.5-1.5b-instruct-Q4_K_M
- Qwen2.5-1.5b-instruct-Q4_K_S
- Qwen2.5-1.5b-instruct-Q4_1
- Qwen2.5-1.5b-instruct-Q4_0
- Qwen2.5-1.5b-instruct-Q3_K_L
- Qwen2.5-1.5b-instruct-Q3_K_M
- Qwen2.5-1.5b-instruct-Q3_K_S
- Qwen2.5-1.5b-instruct-Q2_K

Kvantisointi

Kvantisointi bittitarkkuudet

- Qwen2.5-1.5b-instruct-FP16
- Qwen2.5-1.5b-instruct-Q8_0
- Qwen2.5-1.5b-instruct-Q6_K
- Qwen2.5-1.5b-instruct-Q5_K_M
- Qwen2.5-1.5b-instruct-Q5_K_S
- Qwen2.5-1.5b-instruct-Q5_1
- Qwen2.5-1.5b-instruct-Q5_0
- Qwen2.5-1.5b-instruct-Q4_K_M
- Qwen2.5-1.5b-instruct-Q4_K_S
- Qwen2.5-1.5b-instruct-Q4_1
- Qwen2.5-1.5b-instruct-Q4_0
- Qwen2.5-1.5b-instruct-Q3_K_L
- Qwen2.5-1.5b-instruct-Q3_K_M
- Qwen2.5-1.5b-instruct-Q3_K_S
- Qwen2.5-1.5b-instruct-Q2_K

Kvantisointi

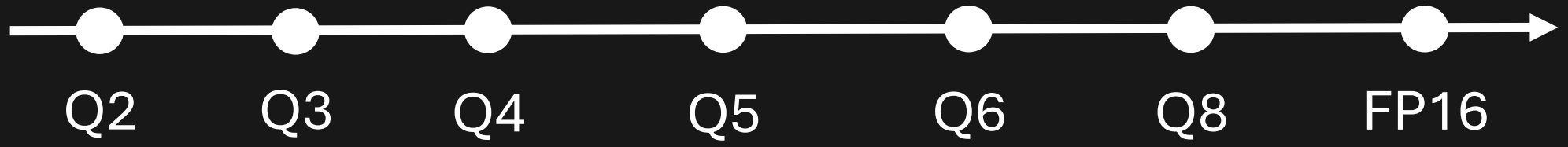
Mallin koon pienentäminen vähentämällä bittitarkkuutta

- FP16
- Q8
- Q6
- Q5
- Q4
- Q3
- Q2

GPT Lab Seinäjoki

Kvantisointi

Hahmottamisen helpottamiseksi

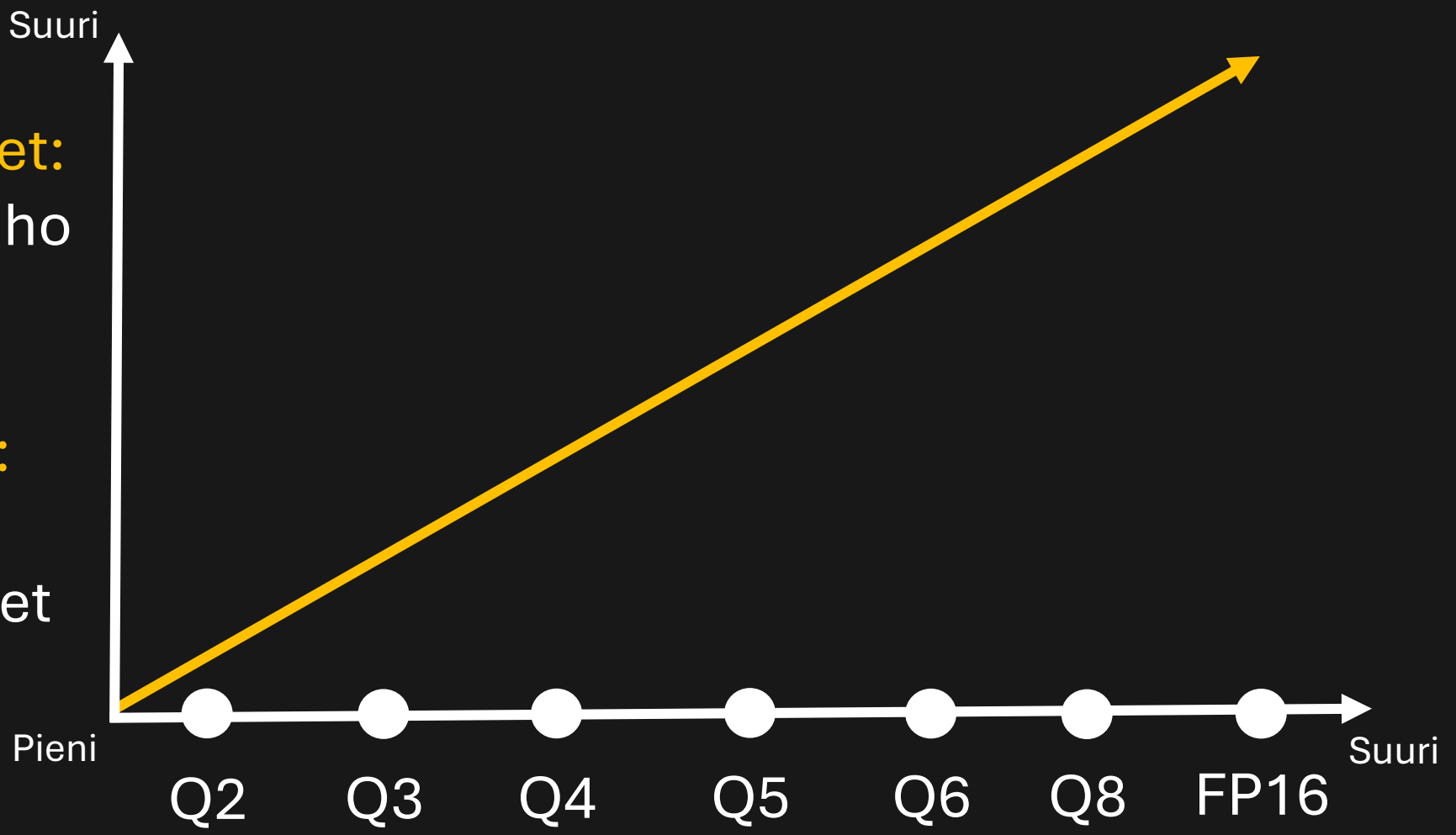


Kvantisointi

Kvantisoinnin vaikutus mallin vaatimuksiin ja ulosantiin (nyrkkisääntö)

Vaatimukset:
Laskentateho
Muisti

Ulosanti:
Laatu
Kyvykkyydet



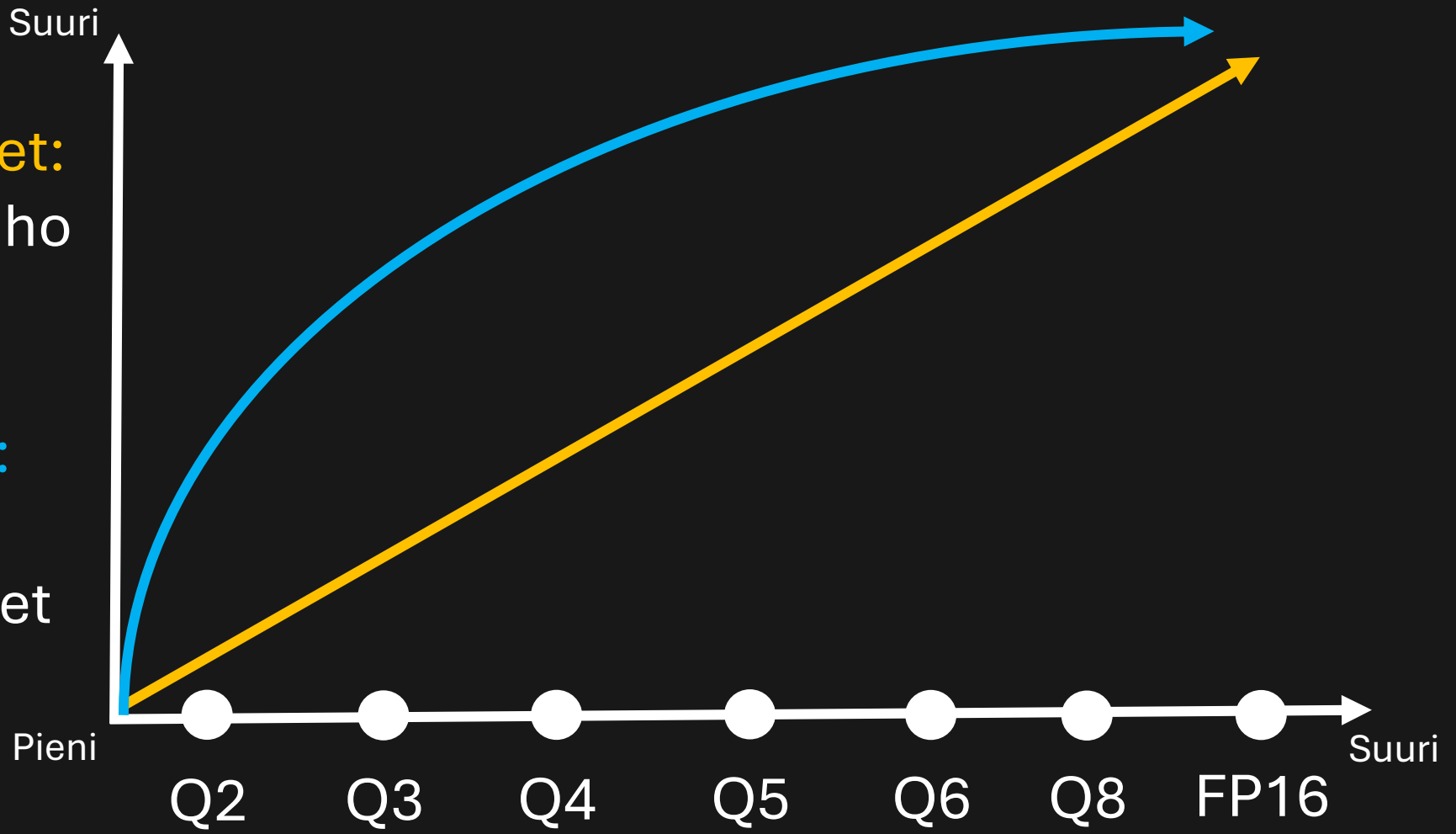
GPT Lab Seinäjoki

Kvantisointi

Kvantisoinnin vaikutus mallin vaatimuksiin ja ulosantiin (lähempänä totuutta)

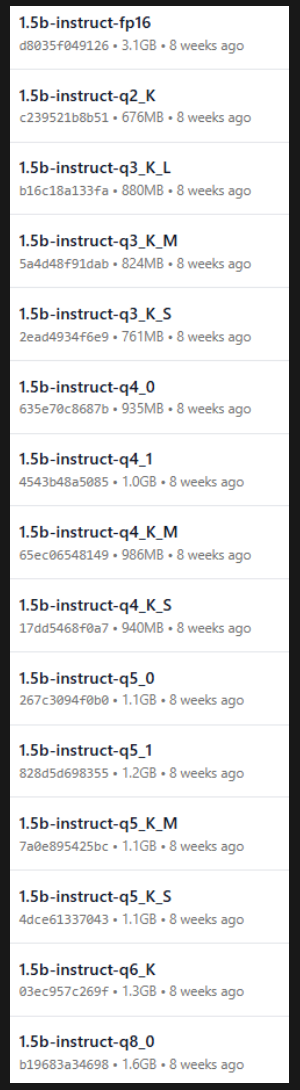
Vaatimukset:
Laskentateho
Muisti

Ulosanti:
Laatu
Kyvykkyydet



GPT Lab Seinäjoki

Kvantisointi



Versiot, k-kvantti

- Q[]_0, Q[]_1
  - Kaikki bitit samalla tarkkuudella, vanha ja uudempi versio
  - Vanhoja kvantisointi tekniikoita tässä kohtaa (legacy)
- Q[]_K
  - Uudempi kvantisointi (k-quant), eri kokoisia bittejä painoissa
  - Sama koko ja nopeus, mutta tarkempi
- Q[]_K_(S/M/L)
  - Eri kokoisia k-kvantteja (small / medium / large)
  - Käytännössä vaikuttaa bittien määrään painoa kohden
- IQ[]_(XXS/S/M)
  - Tärkeysmatriisi (importance matrix) kehitetty parantamaan pieniä kvantisointeja (<=Q3) jotka ovat muuten lähes käyttökelvottomia

Konteksti

Mitä tarkoittaa konteksti?

Kuinka paljon tietoa kielimalli voi käsitellä kerrallaan:

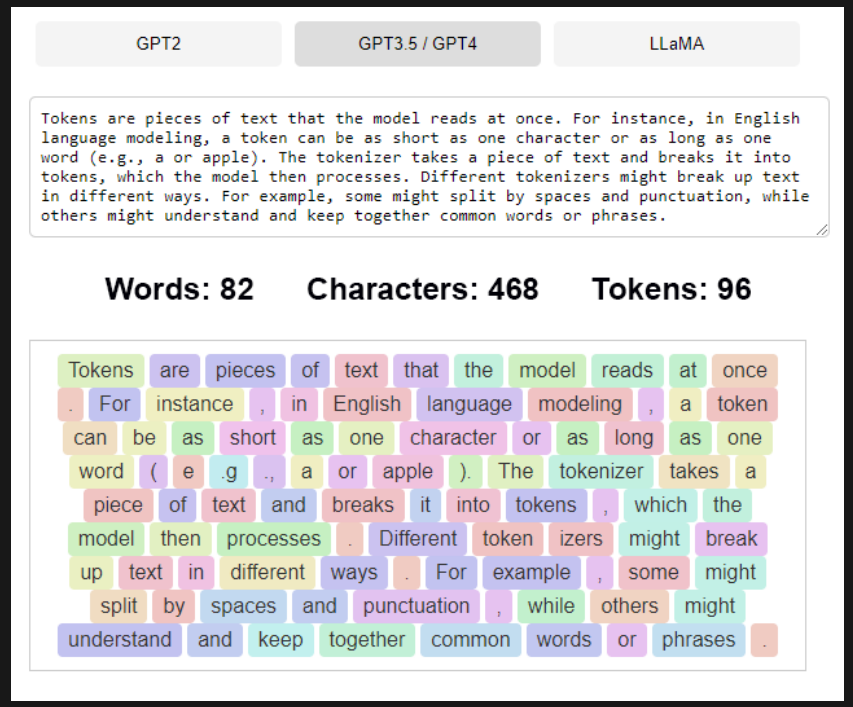
kuinka monta tokenia kielimallin muistiin mahtuu, joiden perusteella kielimalli tuottaa uutta tekstiä

GPT Lab Seinäjoki

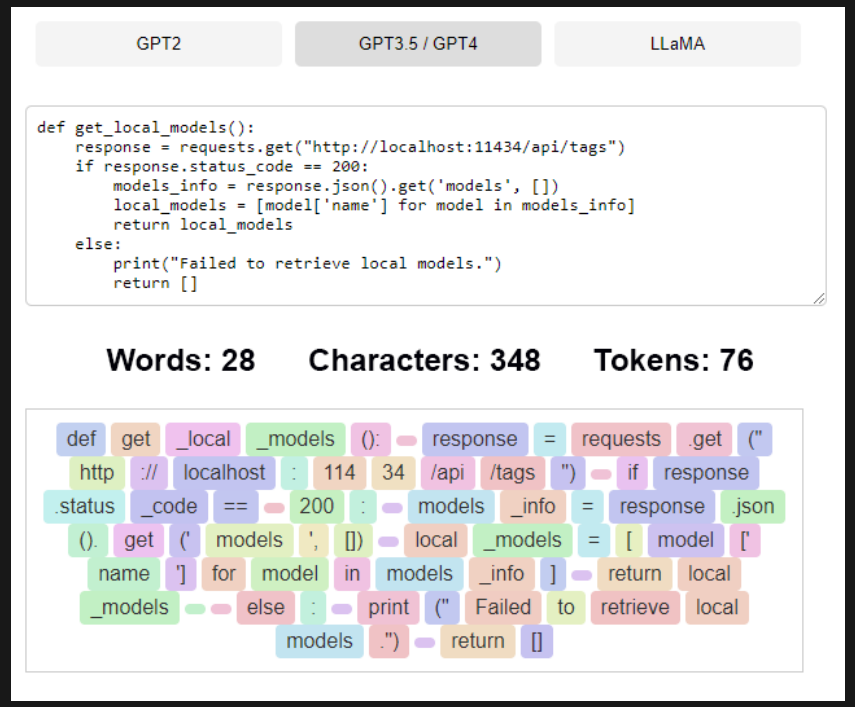
Konteksti

Mitä ovat “tokenit”?

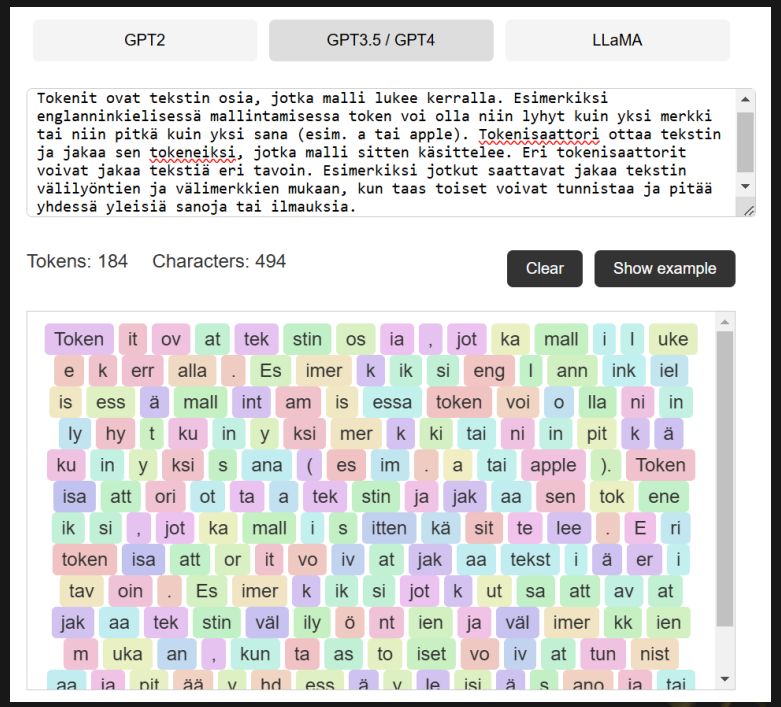
Englanti



Koodi



Suomi

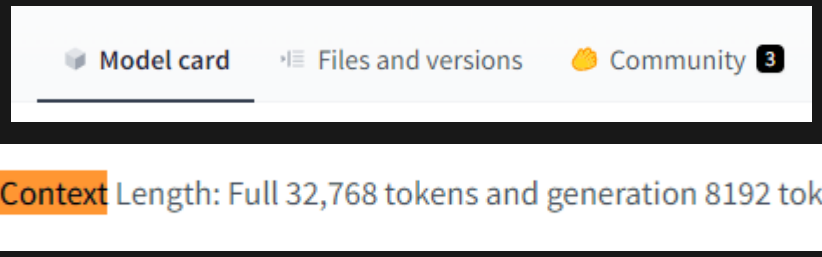


Konteksti

Miten tiedän kuinka paljon mallin konteksti on?

- Mallin nimessä
  - phi-3-mini-4k-instruct
  - phi-3-small-128k-instruct
  - phi-3-medium-8k-instruct
- Useimmiten ei ole nimessä
  - eli miten voi ottaa selvää?

https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct



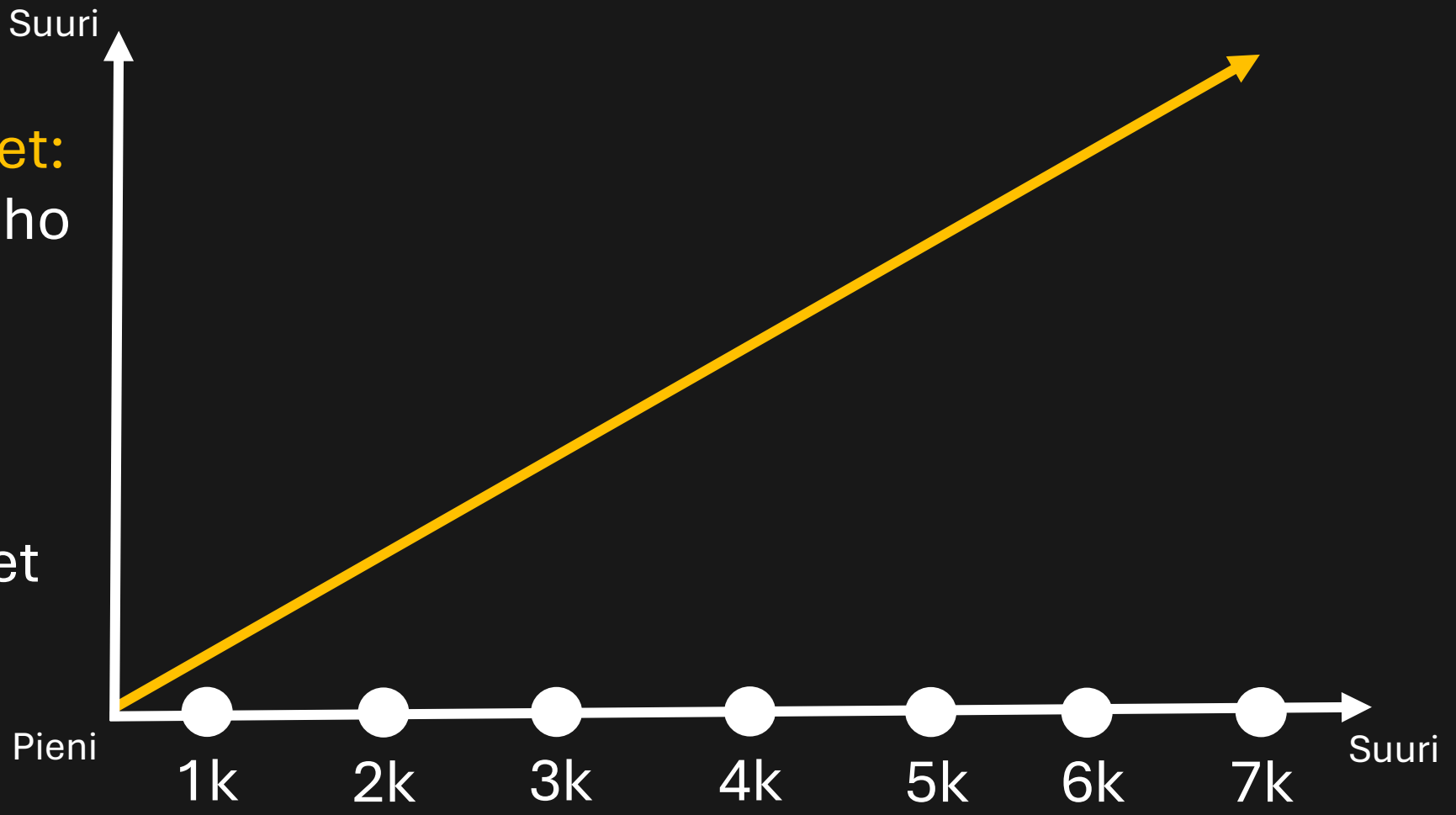
- Konteksti vie paljon muistia
- Mallia ei tarvitse ajaa täydellä kontekstilla
- Ollama ajaa malleja oletuksena 2048 tokenin kontekstilla
- Mallien kontekstit vaihtelevat 1024-128000
- Vaikka mallille voi antaa 128000 tokenia niin enemmän ei ole aina parempi

Konteksti

Konteksin vaikutus mallin vaatimuksiin ja ulosantiin (nyrkkisääntö)

Vaatimukset:
Laskentateho
Muisti

Ulosanti:
Laatu
Kyvykkyydet



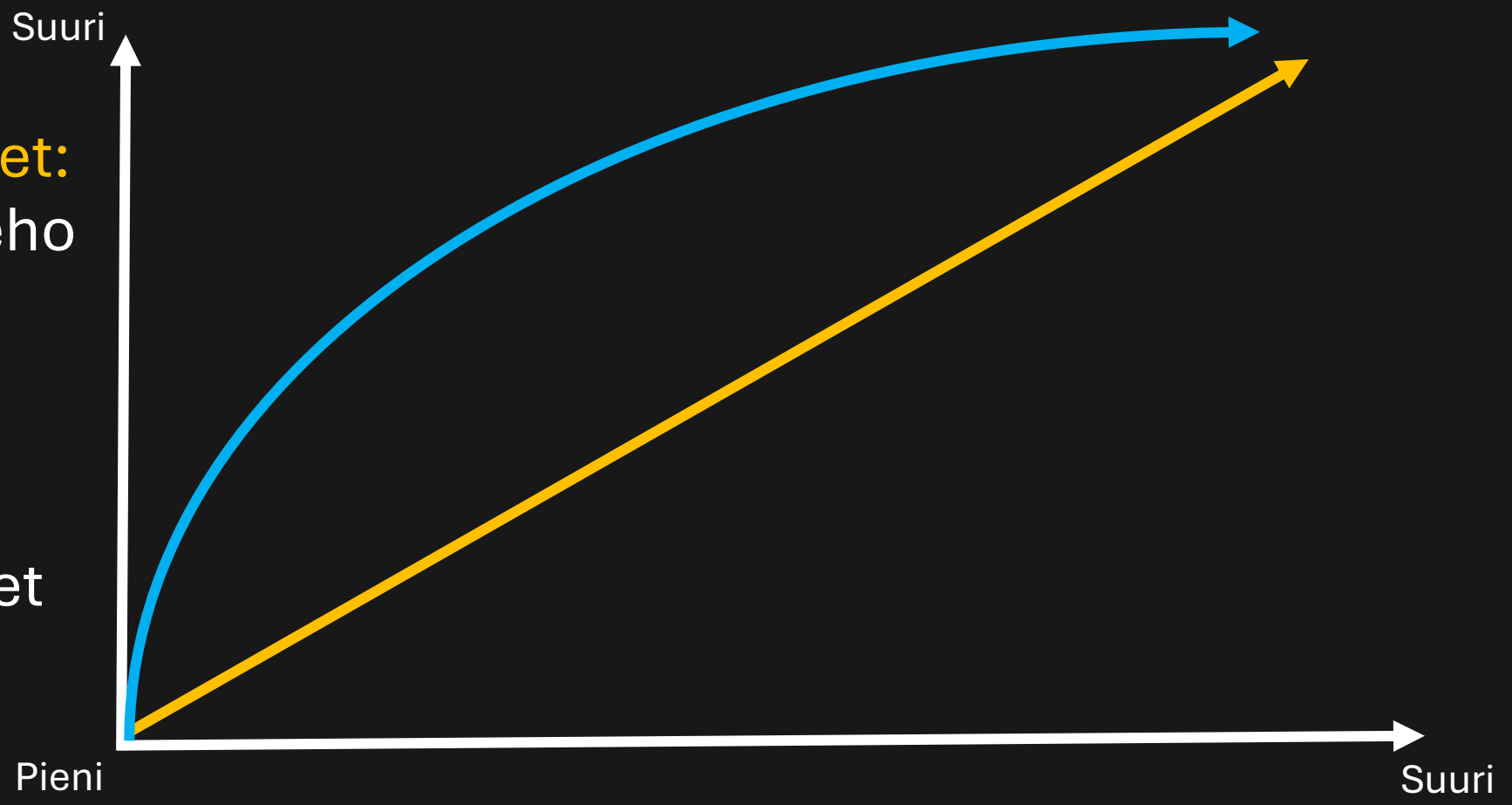
GPT Lab Seinäjoki

Konteksti

Konteksin vaikutus mallin vaatimuksiin ja ulosantiin (ehkä lähempänä totuutta)

Vaatimukset:
Laskentateho
Muisti

Ulosanti:
Laatu
Kyvykkyydet



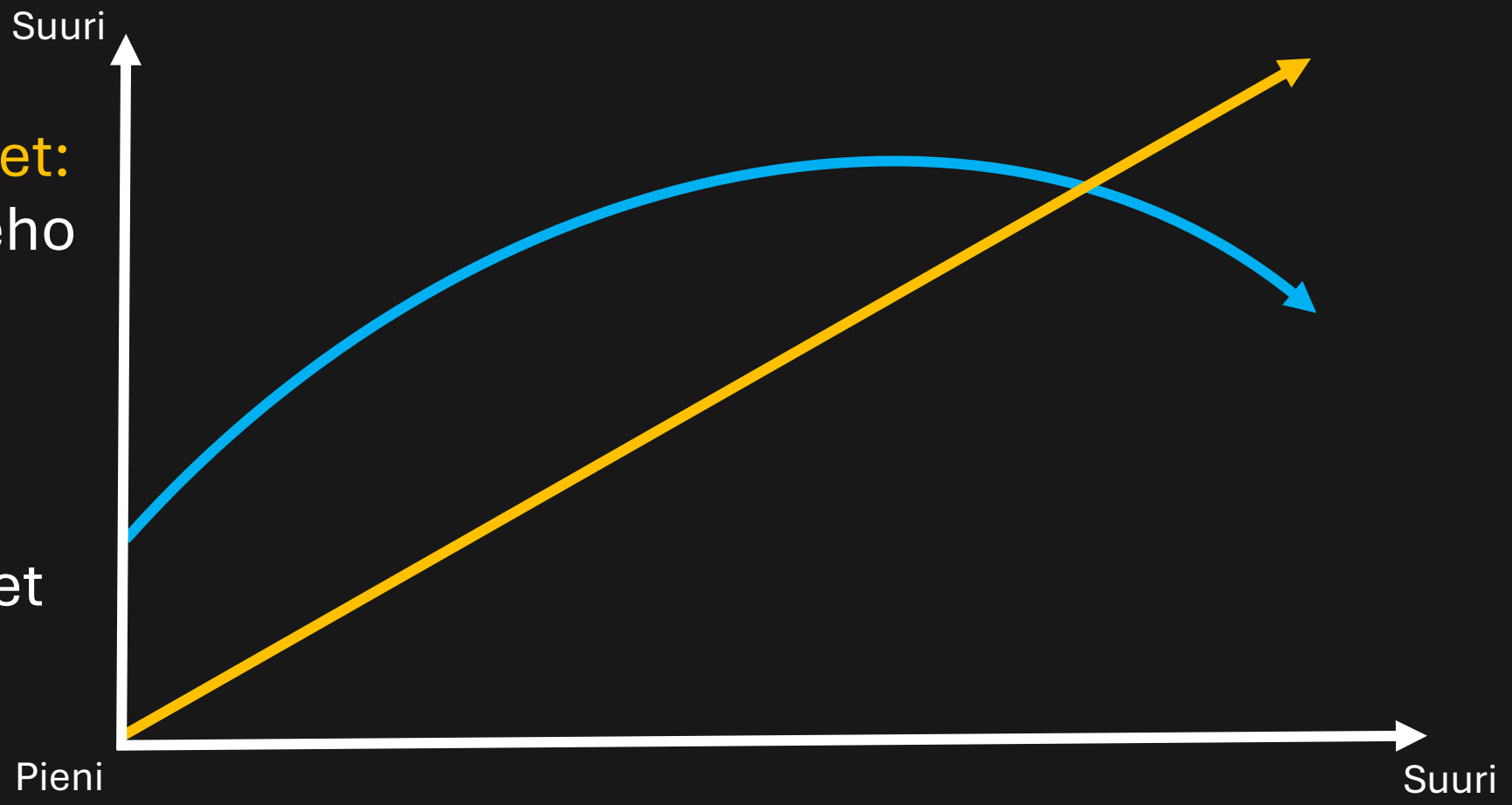
GPT Lab Seinäjoki

Konteksti

Konteksin vaikutus mallin vaatimuksiin ja ulosantiin (ehkä vielä lähempänä totuutta)

Vaatimukset:
Laskentateho
Muisti

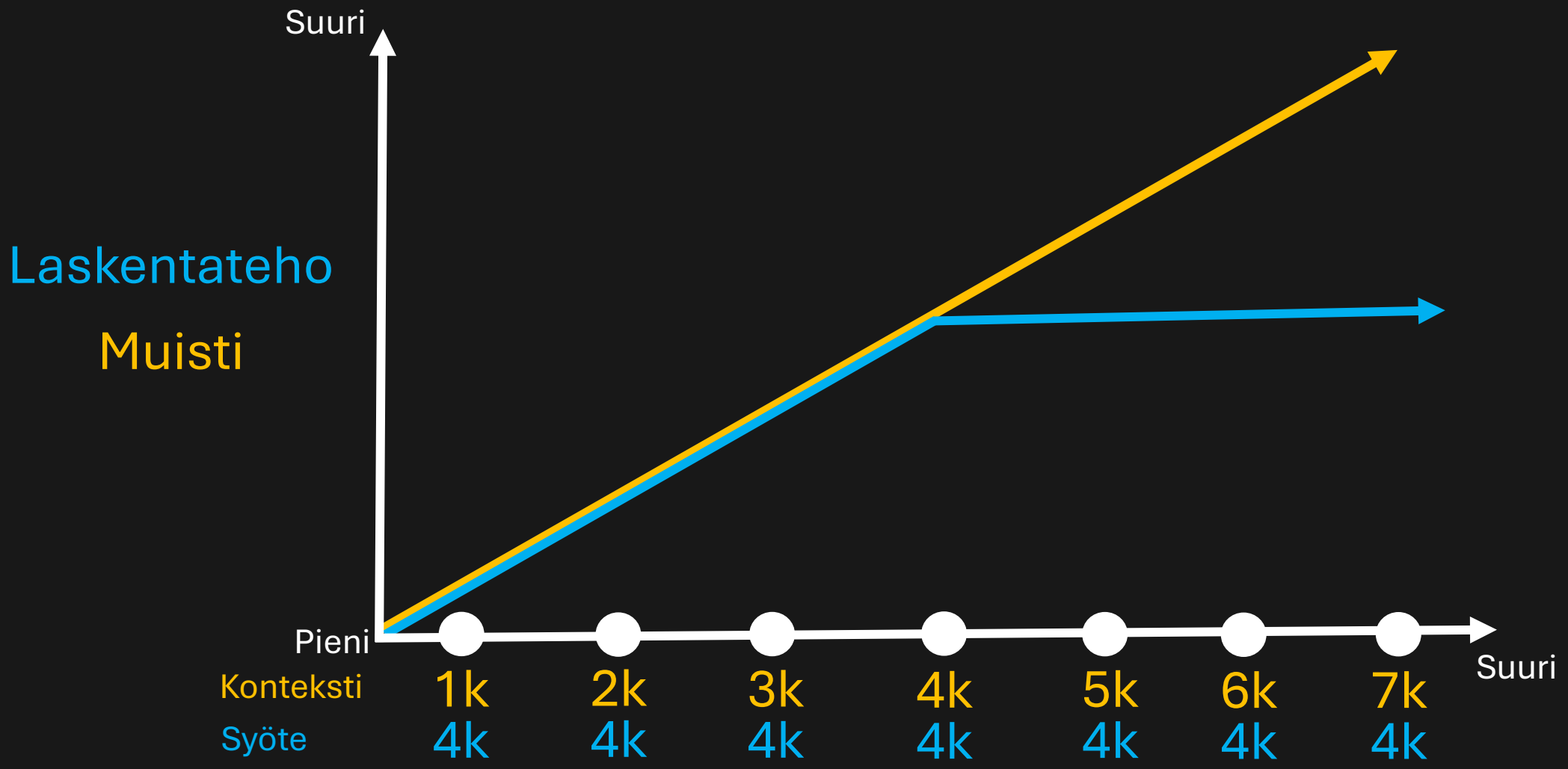
Ulosanti:
Laatu
Kyvykkyydet



GPT Lab Seinäjoki

Konteksti

Konteksin vaikutus mallin vaatimuksiin: laskentateho ja muisti (konteksti vs syöte)



Parametrit+Kvantisointi

Isompi malli? Pienempi kvantisaatio?

- Ulosannin kannalta: paras on ajaa mahdollisimman isoa mallia
- Esimerkkinä 1.5b-Q8 vs 3b-Q4 niin 3b-Q4 antaa parempaa ulosantia
- Pienissä malleissa (0.5b-3b) kannattaa pysyä Q6-Q8
- Isoimmissa malleissa (7b+) kannattaa pysyä Q4-Q6
- Q2 ja Q3 hallusinoivat liikaa - Q6, Q8 ja FP16 välillä ei paljoa eroa
- Uudemmat pienemmät mallit voivat tarjota parempaa ulosantia kuin vanhat isommat mallit
- Joissain tutkimuksissa Q4 on optimaalisin kvantisointi kaikille

GPT Lab Seinäjoki

Koulutusmateriaali

Mitä se vaikuttaa?

- Koulutusmateriaalilla on ratkaiseva vaikutus miten kielimalli toimii ja minkälaisia tehtäviä se pystyy suorittamaan
- Jokainen malli on rajoitettu siihen tietoon, mitä sille on syötetty
- Kielimalli toistaa koulutusdatan sisältöä
- Toistaa myös koulutusdatassa olevia virheellisiä tietoja
- Koulutusmateriaali ei aina sisällä uusinta tietoa

GPT Lab Seinäjoki

Hienosäätö

Mitä hienosäätö (fine-tuning) tarkoittaa?

- Koulutettu kielimalli mukautetaan erityistehtävään kouluttamalla sitä uudelleen pienellä tehtäväkohtaisella aineistolla
- Mallin tarkkuus ja suorituskyky paranevat tietyissä käyttötapauksissa

GPT Lab Seinäjoki

Hienosäätö

Mitä niitä on ja mitä ne tarkoittavat? (ympäripyöreästi)

Meta-Llama-3.1-8B

- Base
  - Koulutettu malli joka käytännössä tuottaa vain tekstiä

8b-instruct-q5_K_S 5.6GB

- Instruct
  - Tehtävien suorittaminen ohjeiden mukaan

v1.5-chat 4.2GB

- Chat
  - Vuoropuhekeskustelu

7b-code 4.2GB

- Code
  - Koodin tuottaminen ja täydentäminen

8b-text-q4_0 4.7GB

- Text
  - Tekstin käsittely ja tuottaminen

GPT Lab Seinäjoki

Hienosäätö

Mitä se on käytännössä?

Instruct

System	Question	Answer
Vastaa selkeästi ja täsmällisesti kysymyksiin.	Kuinka kauan kananmunia pitää keittää?	Keitä kananmunia 6–8 minuuttia saadaksesi pehmeästi kypsät munat ja 10–12 minuuttia saadaksesi täysin kypsät munat.
Selitä annettu prosessi askel askeleelta ja käytä yksinkertaista kieltä.	Miten luon uuden tiedoston Linuxissa?	Voit luoda uuden tiedoston Linuxissa käyttämällä `touch`-komentoa. Esimerkiksi, kirjoita `touch tiedosto.txt`, jolloin luot tiedoston nimeltä 'tiedosto.txt'.

Chat

System	User	Assistant
Olet avulias ja ystävällinen avustaja. Vastaa lyhyesti ja selkeästi.	Miten voin tehdä kahvia ilman kahvinkeitintä?	Käytä suodatinpaperia ja kaada kuumaa vettä jauhetun kahvin päälle hitaasti mukiin.
Olet tekninen asiantuntija, joka selittää asiat yksityiskohtaisesti mutta ymmärrettävästi.	Mikä on ohjelmointikielen Python etu verrattuna muihin kieliin?	Python on helppokäyttöinen, sillä sen syntaksi on yksinkertainen ja muistuttaa luonnollista kieltä. Lisäksi sillä on laaja kirjasto tuki eri sovelluksiin, kuten datan analysointiin ja koneoppimiseen.

Template

Mikä se on?

qwen2.5-coder:1.5b-base / template 96f5a2272876 · 117B

```
{{- if .Suffix }}<|fim_prefix|>{{ .Prompt }}<|fim_suffix|>{{ .Suffix }}
<|fim_middle|>{{ else }}{{ .Prompt }}{{ end }}
```

- Templaten tärkein tehtävä on varmistaa että kielimalli saa syötteen koulutusmateriaalia vastaavassa muodossa
- Tämä parantaa mallin kykyä tuottaa parempia vastauksia
- Usein erotellaan viestien roolit, kuten user, assistant, system

Template

Esimerkkinä base vs instruct

qwen2.5-coder:1.5b-base / template 96f5a2272876 · 117B

```
{{- if .Suffix }}<|fim_prefix|>{{ .Prompt }}<|fim_suffix|>{{ .Suffix }}
<|fim_middle|>{{ else }}{{ .Prompt }}{{ end }}
```

- Base-mallit tuottavat vain tekstiä, ennustavat seuraavaa sanaa, niiden template on yksinkertainen tekstin yhdistelyyn keskittyvä

- Instruct-mallit on hienosäädetty noudattamaan ohjeita ja suorittamaan tehtäviä, niiden template on monimutkainen, sisältäen paljon ohjeistusta, ja useita rooleja

qwen2.5-coder:1.5b-instruct / template e94a8ecb9327 · 1.6kB

```
{{- if .Suffix }}<|fim_prefix|>{{ .Prompt }}<|fim_suffix|>{{ .Suffix }}
<|fim_middle|>
{{- else if .Messages }}
{{- if or .System .Tools }}<|im_start|>system
{{- if .System }}
{{ .System }}
{{- end }}
{{- if .Tools }}

# Tools

You may call one or more functions to assist with the user query.

You are provided with function signatures within <tools></tools> XML tags:
<tools>
{{- range .Tools }}
{"type": "function", "function": {{ .Function }}}
{{- end }}
</tools>

For each function call, return a json object with function name and arguments
within <tool_call></tool_call> XML tags:
<tool_call>
{"name": <function-name>, "arguments": <args-json-object>}
</tool_call>
{{- end }}<|im_end|>
{{ end }}
{{- range $i, $_ := .Messages }}
{{- $last := eq (len (slice $.Messages $i)) 1 -}}
{{- if eq .Role "user" }}<|im_start|>user
{{ .Content }}<|im_end|>
{{ else if eq .Role "assistant" }}<|im_start|>assistant
{{ if .Content }}{{ .Content }}
{{- else if .ToolCalls }}<tool_call>
{{ range .ToolCalls }}{"name": "{{ .Function.Name }}", "arguments": {{
.Function.Arguments }}}
{{ end }}</tool_call>
{{- end }}{{ if not $last }}<|im_end|>
{{ end }}
{{- else if eq .Role "tool" }}<|im_start|>user
<tool_response>
{{ .Content }}
</tool_response><|im_end|>
{{ end }}
{{- if and (ne .Role "assistant") $last }}<|im_start|>assistant
{{ end }}
{{- end }}
{{- else }}
{{- if .System }}<|im_start|>system
{{ .System }}<|im_end|>
{{ end }}{{ if .Prompt }}<|im_start|>user
{{ .Prompt }}<|im_end|>
{{ end }}<|im_start|>assistant
{{ end }}{{ .Response }}{{ if .Response }}<|im_end|>{{ end }}
```

Lisenssit

Mikä on johtopäätös?

- Kielimalleilla on usein oma lisenssi erilaisilla ehdoilla
  - Joissain lisensseissä annetaan enemmän oikeuksia käyttää mallia kuinka tahtoo ja toisissa rajoitetaan käyttöä enemmän
- Toinen yleinen lisenssi on Apache 2.0
  - Saa käyttää kaupallisesti
- Jotkut mallit on myös lisensoitu MIT
  - Saa tehdä mitä haluaa

GPT Lab Seinäjoki

TYÖKALUN RAKENTAMINEN

GPT Lab Seinäjoki

Työkalun rakentaminen

Mitä tarvitaan?

1. Kielimalli
2. Provider
3. VSCode
4. Plugin

GPT Lab Seinäjoki

Kielimalli

Miten valitaan oikea malli?

1. Mallin toimii riittävän nopeasti

Kuinka paljon tehoa ja muistia tarvitaan käytännössä?

2. Malli tuottaa käyttökelpoista ulosantia

Miten sitä voidaan mitata ja verrata?

GPT Lab Seinäjoki

Kielimalli

Miten sitä voidaan mitata ja verrata?

Kehitetty paljon eri menetelmiä:
Ihmiset arvioivat, toinen tekoäly arvioi, ohjelmallisesti arvioidaan

Tuloksista on tehty paljon listoja

Paljon, paljon listoja

GPT Lab Seinäjoki

Kielimalli

EvalPlus AI Coders Leaderboard

HumanEval + MBPP (Mostly Basic Python Problems)
https://evalplus.github.io/leaderboard.html

#	Model	pass@1
1	GPT-4-Turbo (April 2024)✨	86.6
2	DeepSeek-Coder-V2-Instruct✨	82.3
3	GPT-4-Turbo (Nov 2023)✨	81.7
4	GPT-4 (May 2023)✨	79.3
5	CodeQwen1.5-7B-Chat✨	78.7
6	claude-3-opus (Mar 2024)✨	77.4
7	DeepSeek-Coder-33B-instruct✨	75
8	OpenCodeInterpreter-DS-33B✨💙	73.8
9	WizardCoder-33B-V1.1✨	73.2
10	Artigenz-Coder-DS-6.7B✨	72.6

Kielimalli

LLM Capability as a Coding Assistant

How often do LLMs generate an acceptable answer?

https://prollm.toqan.ai/leaderboard/coding-assistant

#	Model	Provider	Size	Acceptance ? ↑↓
1	**GPT-4 Turbo** gpt-4-turbo-2024-04-09	OpenAI	—	0.88
2	**Claude-v3.5 Sonnet** claude-3-5-sonnet-20240620	Anthropic	—	0.871
3	**GPT-4o** gpt-4o-2024-05-13	OpenAI	—	0.848
4	**GPT-4o** gpt-4o-2024-08-06	OpenAI	—	0.83
5	**Mistral Large 2** mistral-large-2407	Mistral	123 B	0.819
6	**WizardLM-2 8×22B** alpindale/WizardLM-2-8×22B	Microsoft	141 B	0.802

GPT Lab Seinäjoki

Kielimalli

Aider Code Editing Leaderboard

Can LLMs follow a prompt to edit existing code successfully?
https://aider.chat/docs/leaderboards/

Model	Percent completed correctly	Percent using correct edit format	Command	Edit format
o1-preview (whole)	79.7%	100.0%	`aider --model o1-preview`	whole
claude-3.5-sonnet	77.4%	99.2%	`aider --sonnet`	diff
o1-preview (diff)	75.2%	84.2%	`aider --model o1-preview`	diff
DeepSeek Coder V2 0724 (deprecated)	72.9%	97.7%	`aider --model deepseek/deepseek-coder`	diff
gpt-4o-2024-05-13	72.9%	96.2%	`aider`	diff
DeepSeek V2.5	72.2%	96.2%	`aider --deepseek`	diff

Kielimalli

Code Completion Leaderboard

Can LLMs solve realistic, practical and challenging tasks?
https://huggingface.co/spaces/bigcode/bigcodebench-leaderboard
https://bigcode-bench.github.io/

Model	Complete	Instruct	Average	Elo Rating
GPT-4-Turbo-2024-04-09	35.1	29.1	32.1	1220
GPT-4o-2024-08-06	36.5	25	30.8	1213
Gemini-1.5-Pro-Exp-0827	31.8	27	29.4	1190
DeepSeek-Coder-V2-Instruct (2024-07-24)	33.1	25.7	29.4	1206
Claude-3.5-Sonnet-20240620	33.1	25.7	29.4	1189
o1-Preview-2024-09-12 (temperature=1)	34.5	23	28.8	1169
DeepSeek-V2-Chat (2024-06-28)	32.4	25	28.7	1192
Reflection-Llama-3.1-70B	33.1	23	28.1	1173
Gemini-1.5-Pro-Exp-0801	29.7	25	27.4	1162
o1-Mini-2024-09-12 (temperature=1)	27	27.7	27.4	1150
GPT-4o-2024-05-13	29.1	25	27.1	1157
DeepSeek-Coder-V2-Instruct	29.7	24.3	27	1171
Llama-3.1-405B-Instruct	30.4	22.3	26.4	1153

Kielimalli

Dubesor LLM Benchmark Table

Small-scale manual performance comparison using weighted rating system

https://dubesor.de/benchtable

Model (85)	TOTAL	Pass	Refine	Fail	Refusal	$ mTok	Reason	STEM	Utility	Code	Censor
#1 GPT-4 Turbo	83.8%	64	9	10	0	$26.00	80.6%	84.5%	78.9%	91.0%	87.8%
#2 gpt2-chatbot †	81.1%	62	6	13	0	n/a	86.8%	73.3%	65.8%	77.9%	100.0%
#3 GPT-4o	71.8%	57	7	19	0	$13.00	68.3%	58.0%	83.1%	85.0%	81.9%
#4 Chatgpt-4o-latest (2024-10)	71.6%	57	7	19	0	$13.00	70.3%	60.0%	88.9%	72.0%	80.5%
#5 ChatGPT o1-preview	70.0%	55	8	16	4	$208.44	81.7%	52.5%	73.7%	78.2%	54.7%
#6 Llama 3.1 405B Instruct bf16	69.2%	55	8	20	0	$4.00	68.0%	60.2%	82.1%	67.6%	80.4%
#7 Grok-2	65.0%	48	15	20	0	$6.36	69.8%	59.8%	58.2%	53.0%	85.9%

Kielimalli

EQ-Bench

Emotional Intelligence Benchmark for LLMs

https://eqbench.com/

Model	Params	EQ-Bench*	MAGI-Hard†	Combined
Meta-Llama-3.1-405B-Instruct	405	83.0	83.81	83.41
claude-3-5-sonnet-20240620		86.36	78.8	82.58
gpt-4o		83.51	80.86	82.19
gpt-4-turbo-2024-04-09		86.35	77.74	82.04
NEW RYS-XLarge-Base	78	85.05	78.3	81.67
gpt-4-0613		84.79	77.85	81.32
gpt-4-0314		85.73	75.67	80.70
NEW RYS-XLarge	78	84.55	76.83	80.69
gpt-4-1106-preview		86.05	74.96	80.50
gpt-4-0125-preview		83.87	76.83	80.35
NEW Hermes-3-Llama-3.1-405B	405	82.79	76.23	79.51
claude-3-opus-20240229		82.19	76.55	79.37
mistral-large-2407	123	85.05	72.37	78.71
Qwen2-72B-Instruct	72	81.35	75.74	78.54

Kielimalli

GPT Lab Seinäjoki tutkimustyötä

Name	Tune/Base	Parameters	Quantization	Released	License	GitHub	HuggingFace	Ollama
DeepSeek-Coder-V2-Lite	Instruct	16b	Q4_0	June -24	Proprietary	2k	36k	226k
*	*	*	Q5_K_M	*	*	*	*	*
*	*							

Models, Tokens per second, RAM

Model	Parameters	Fine-tune	Quantization	Context	CPU Win	lowGPU Win	Macbook M
Different models, parameters, tune							
Yi-Coder 1.5b							
yi-coder	1.5b	chat	q2_K	2048	30,74	37,27	49,81
yi-coder	1.5b	chat	q3_K_S	2048	27,6	29,9	43,15
yi-coder	1.5b	chat	q3_K_M	2048	25,73	28,35	42,77
yi-coder	1.5b	chat	q3_K_L	2048	24,54	22,72	41,14
yi-coder	1.5b	chat	q4_0	2048	24,44	15,54	49,46
yi-coder	1.5b	chat	q4_1	2048	22,34	15,27	46,83
yi-coder	1.5b	chat	q4_K_S	2048	22,73	18,42	46,03
yi-coder	1.5b	chat	q4_K_M	2048	21,08	27,96	45,64
yi-coder	1.5b	chat	q5_0	2048	19,7	19,26	44,5
yi-coder	1.5b	chat	q5_1	2048	17,76	12,94	42,1
yi-coder	1.5b	chat	q5_K_S	2048	18,58	14,91	40,98
yi-coder	1.5b	chat	q5_K_M	2048	18,05	23,43	40,63
yi-coder	1.5b	chat	q6_K	2048	16,36	18,63	38,62
yi-coder	1.5b	chat	q8_0	2048	13,8	13,26	32,25
yi-coder	1.5b	chat	fp16	2048	8,04	0,97	19,08
Yi-Coder 9b							
yi-coder	9b	chat	q2_K	2048	5,69	2,68	12,33
yi-coder	9b	chat	q3_K_S	2048	4,76	2,2	9,78
yi-coder	9b	chat	q3_K_M	2048	4,61	2,41	9,92
yi-coder	9b	chat	q3_K_L	2048	4,29	2,42	9,27
yi-coder	9b	chat	q4_0	2048	4,22	1,98	10,93
yi-coder	9b	chat	q4_1	2048	3,77	2,7	10,02
yi-coder	9b	chat	q4_K_S	2048	4,19	3,48	10,72
yi-coder	9b	chat	q4_K_M	2048	3,93	3,01	10,36

Kielimalli

Tiedoksi

Seuraavissa taulukoissa ja kuvaajissa olevat numerot on ajettu pienellä syötteellä (~100 tokenia) ja Ollama oletuskontekstilla (2048 tokenia)

GPT Lab Seinäjoki

Kielimalli

Laskentatehot

Qwen2.5 – Laptop CPU (i5-1335U)

Tokens Per Second (TPS)

Kvantisointi Parametrit	Q2 _K	Q3 _K_S	Q3 _K_M	Q3 _K_L	Q4 _0	Q4 _1	Q4 _K_S	Q4 _K_M	Q5 _0	Q5 _1	Q5 _K_S	Q5 _K_M	Q6 _K	Q8 _0	FP16
0.5b	44.1	45.9	43.3	42.6	44.6	41.3	41.3	40.1	41.1	39.2	38.9	37.7	31.6	29.8	17.4
1.5b	24.2	19.6	19.2	17.7	18.7	16.6	17.7	17.1	16.0	14.9	16.1	15.8	14.4	12.3	7.0
3b	15.1	13.4	12.6	11.6	11.2	9.8	10.6	10.3	9.1	8.1	9.0	8.8	7.8	6.5	3.5
7b	6.8	5.8	5.5	5.2	5.0	4.5	5.0	4.7	4.2	3.9	4.3	4.2	3.7	2.9	0

Kielimalli

Laskentatehot

Kvantisoinnin vaikutus muistinkäyttöön

Qwen2.5 – Laptop CPU (i5-1335U)

Tokens Per Second (TPS)

Kvantisointi Parametrit	Q2 _K	Q3 _K_S	Q3 _K_M	Q3 _K_L	Q4 _0	Q4 _1	Q4 _K_S	Q4 _K_M	Q5 _0	Q5 _1	Q5 _K_S	Q5 _K_M	Q6 _K	Q8 _0	FP16
0.5b	44.1	45.9	43.3	42.6	44.6	41.3	41.3	40.1	41.1	39.2	38.9	37.7	31.6	29.8	17.4
1.5b	24.2	19.6	19.2	17.7	18.7	16.6	17.7	17.1	16.0	14.9	16.1	15.8	14.4	12.3	7.0
3b	15.1	13.4	12.6	11.6	11.2	9.8	10.6	10.3	9.1	8.1	9.0	8.8	7.8	6.5	3.5
7b	6.8	5.8	5.5	5.2	5.0	4.5	5.0	4.7	4.2	3.9	4.3	4.2	3.7	2.9	0

Kielimalli

Laskentatehot

"Q6, Q8 ja FP16 antavat käytännössä samaa ulosantia"

Qwen2.5 – Laptop CPU (i5-1335U)

Tokens Per Second (TPS)

Kvantisointi Parametrit	Q2 _K	Q3 _K_S	Q3 _K_M	Q3 _K_L	Q4 _0	Q4 _1	Q4 _K_S	Q4 _K_M	Q5 _0	Q5 _1	Q5 _K_S	Q5 _K_M	Q6 _K	Q8 _0	FP16
0.5b	44.1	45.9	43.3	42.6	44.6	41.3	41.3	40.1	41.1	39.2	38.9	37.7	31.6	29.8	17.4
1.5b	24.2	19.6	19.2	17.7	18.7	16.6	17.7	17.1	16.0	14.9	16.1	15.8	14.4	[illegible]	7.0
3b	15.1	13.4	12.6	11.6	11.2	9.8	10.6	10.3	9.1	8.1	9.0	8.8	7.8	6.5	3.5
7b	6.8	5.8	5.5	5.2	5.0	4.5	5.0	4.7	4.2	3.9	4.3	4.2	3.7	2.9	0

GPT Lab Seinäjoki

Kielimalli

Laskentatehot

“Mieluummin ajaa isompaa pienemmällä kvantisoinnilla kuin pienempää isommalla”

Qwen2.5 – Laptop CPU (i5-1335U)

Tokens Per Second (TPS)

Kvantisointi Parametrit	Q2 _K	Q3 _K_S	Q3 _K_M	Q3 _K_L	Q4 _0	Q4 _1	Q4 _K_S	Q4 _K_M	Q5 _0	Q5 _1	Q5 _K_S	Q5 _K_M	Q6 _K	Q8 _0	FP16
0.5b	44.1	45.9	43.3	42.6	44.6	41.3	41.3	40.1	41.1	39.2	38.9	37.7	31.6	29.8	17.4
1.5b	24.2	19.6	19.2	17.7	18.7	16.6	17.7	17.1	16.0	14.9	16.1	15.8	14.4	[illegible]	7.0
3b	15.1	13.4	12.6	11.6	11.2	9.8	10.6	10.3	9.1	8.1	9.0	8.8	7.8	6.5	3.5
7b	6.8	5.8	5.5	5.2	5.0	4.5	5.0	4.7	4.2	3.9	4.3	4.2	3.7	2.9	0

GPT Lab Seinäjoki

Kielimalli

Laskentatehot

"Q4 on tutkimusten mukaan optimaalisin"

Qwen2.5 – Laptop CPU (i5-1335U)

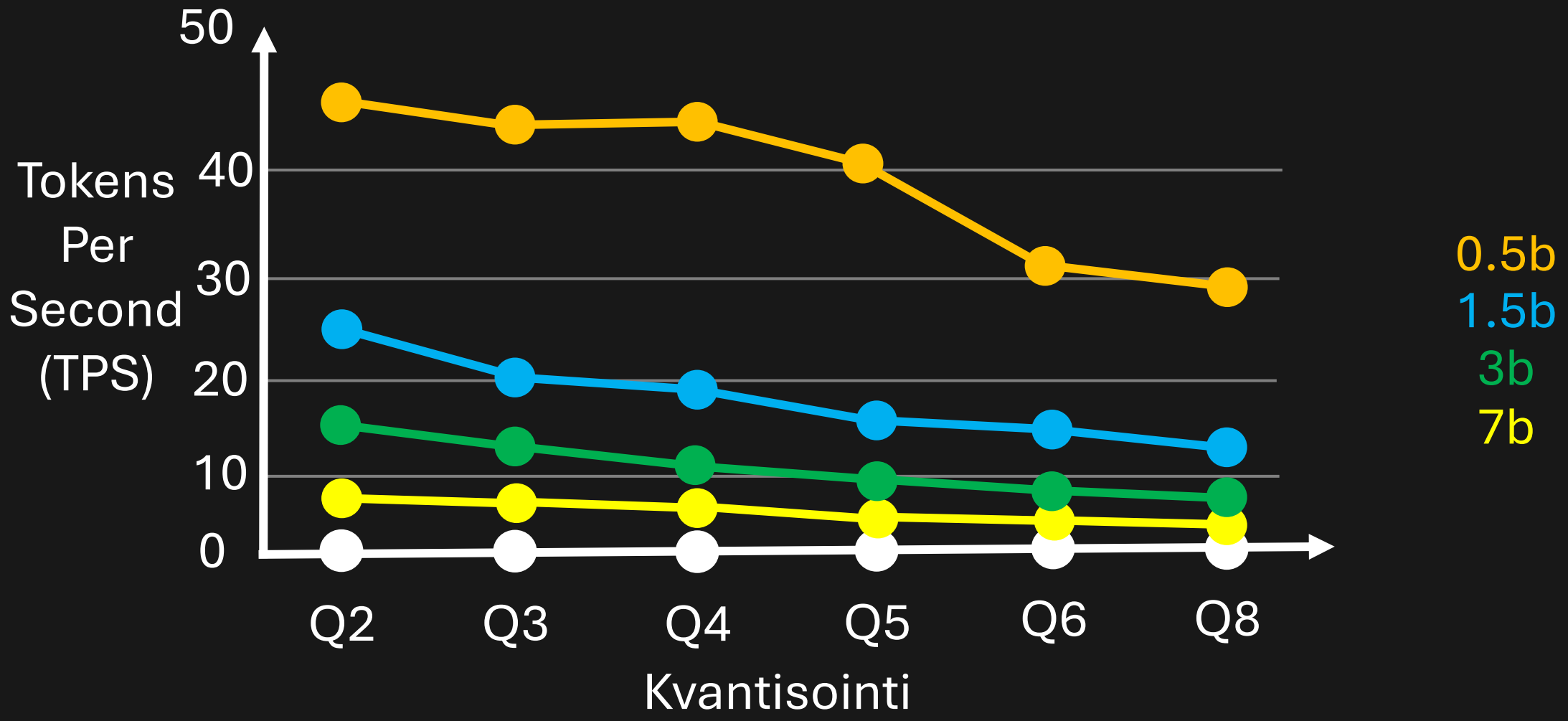
Tokens Per Second (TPS)

Kvantisointi Parametrit	Q2 _K	Q3 _K_S	Q3 _K_M	Q3 _K_L	Q4 _0	Q4 _1	Q4 _K_S	Q4 _K_M	Q5 _0	Q5 _1	Q5 _K_S	Q5 _K_M	Q6 _K	Q8 _0	FP16
0.5b	44.1	45.9	43.3	42.6	44.6	41.3	41.3	40.1	41.1	39.2	38.9	37.7	31.6	29.8	17.4
1.5b	24.2	19.6	19.2	17.7	18.7	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	7.0
3b	15.1	13.4	12.6	11.6	11.2	[illegible]	10.6	10.3	9.1	8.1	9.0	8.8	7.8	6.5	3.5
7b	6.8	5.8	5.5	5.2	5.0	4.5	5.0	4.7	4.2	3.9	4.3	4.2	3.7	2.9	0

Kielimalli

Laskentatehot (yhdellä laitteella eri parametrimäärät)

Qwen2.5 – Laptop CPU (i5-1335U)

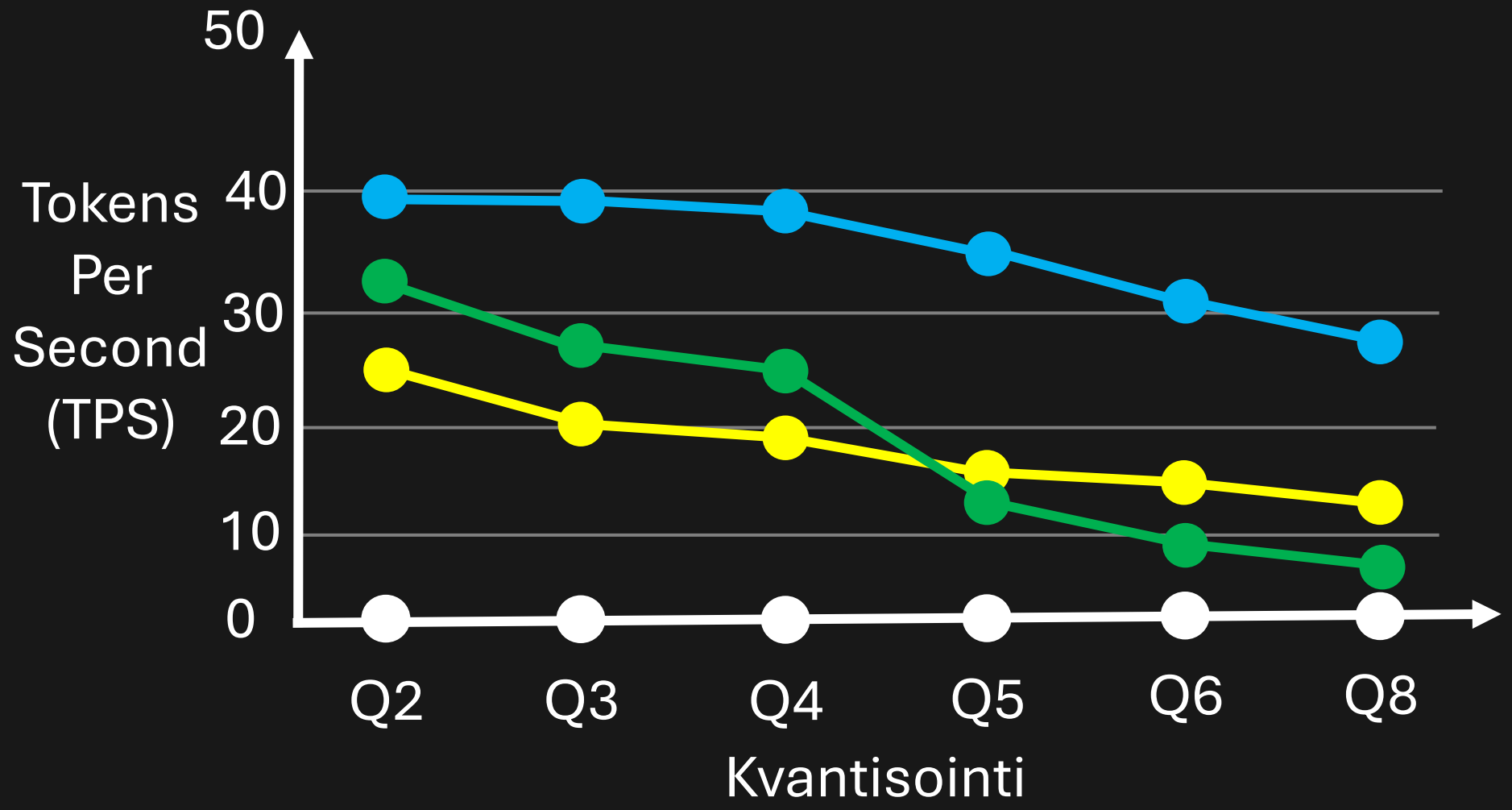


GPT Lab Seinäjoki

Kielimalli

Laskentatehot (sama parametrimäärä eri laitteilla)

Qwen2.5-1.5b-instruct



Macbook M1

HX450

(old low end laptop gpu)

I5-1335U

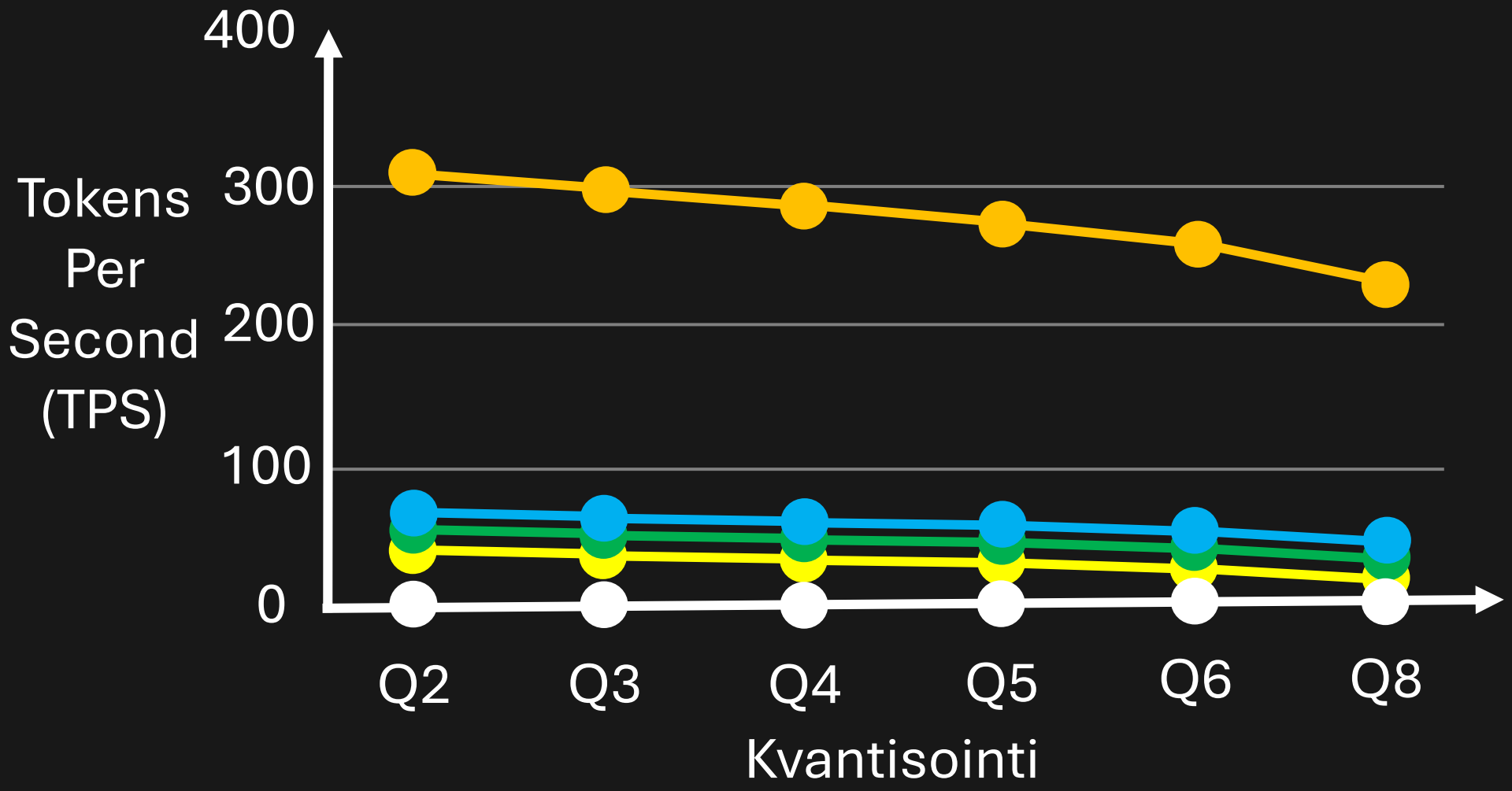
(intel laptop cpu)

GPT Lab Seinäjoki

Kielimalli

Laskentatehot (sama parametrimäärä eri laitteilla)

Qwen2.5-1.5b-instruct



RTX4090
(high end desktop gpu)

Macbook M1

HX450
(old low end laptop gpu)

I5-1335U
(intel laptop cpu)

GPT Lab Seinäjoki

Kielimalli

Muistinkäyttö

Qwen2.5

RAM (Gb)

Kvantisointi Parametrit	Q2 _K	Q3 _K_S	Q3 _K_M	Q3 _K_L	Q4 _0	Q4 _1	Q4 _K_S	Q4 _K_M	Q5 _0	Q5 _1	Q5 _K_S	Q5 _K_M	Q6 _K	Q8 _0	FP16
0.5b	0.72	0.72	0.74	0.75	0.73	0.75	0.77	0.78	0.78	0.8	0.79	0.8	0.89	0.91	1.34
1.5b	1.16	1.24	1.3	1.36	1.41	1.49	1.42	1.46	1.57	1.65	1.57	1.59	1.73	2.09	2.09
3b	1.79	1.96	2.1	2.2	2.32	2.48	2.33	2.42	2.65	2.81	2.65	2.7	3	3.71	6.47
7b	3.64	4.04	4.35	4.62	4.88	5.27	4.91	5.13	5.66	6.06	5.66	5.79	6.5	8.15	14.45

Kielimalli

Muistinkäyttö

Kvantisoinnin vaikutus muistinkäyttöön

Qwen2.5

RAM (Gb)

Kvantisointi Parametrit	Q2 _K	Q3 _K_S	Q3 _K_M	Q3 _K_L	Q4 _0	Q4 _1	Q4 _K_S	Q4 _K_M	Q5 _0	Q5 _1	Q5 _K_S	Q5 _K_M	Q6 _K	Q8 _0	FP16
0.5b	0.72	0.72	0.74	0.75	0.73	0.75	0.77	0.78	0.78	0.8	0.79	0.8	0.89	0.91	1.34
1.5b	1.16	1.24	1.3	1.36	1.41	1.49	1.42	1.46	1.57	1.65	1.57	1.59	1.73	2.09	2.09
3b	1.79	1.96	2.1	2.2	2.32	2.48	2.33	2.42	2.65	2.81	2.65	2.7	3	3.71	6.47
7b	3.64	4.04	4.35	4.62	4.88	5.27	4.91	5.13	5.66	6.06	5.66	5.79	6.5	8.15	14.45

Kielimalli

Muistinkäyttö

Kvantisoinnin vaikutus muistinkäyttöön, isoimmissa malleissa näkyy enemmän

Qwen2.5

RAM (Gb)

Kvantisointi Parametrit	Q2 _K	Q3 _K_S	Q3 _K_M	Q3 _K_L	Q4 _0	Q4 _1	Q4 _K_S	Q4 _K_M	Q5 _0	Q5 _1	Q5 _K_S	Q5 _K_M	Q6 _K	Q8 _0	FP16
0.5b	0.72	0.72	0.74	0.75	0.73	0.75	0.77	0.78	0.78	0.8	0.79	0.8	0.89	0.91	1.34
1.5b	1.16	1.24	1.3	1.36	1.41	1.49	1.42	1.46	1.57	1.65	1.57	1.59	1.73	[illegible]	2.09
3b	1.79	1.96	2.1	2.2	2.32	2.48	2.33	2.42	2.65	2.81	2.65	2.7	3	[illegible]	6.47
7b	3.64	4.04	4.35	4.62	4.88	5.27	4.91	5.13	5.66	6.06	5.66	5.79	6.5	[illegible]	14.45

Kielimalli

Muistinkäyttö

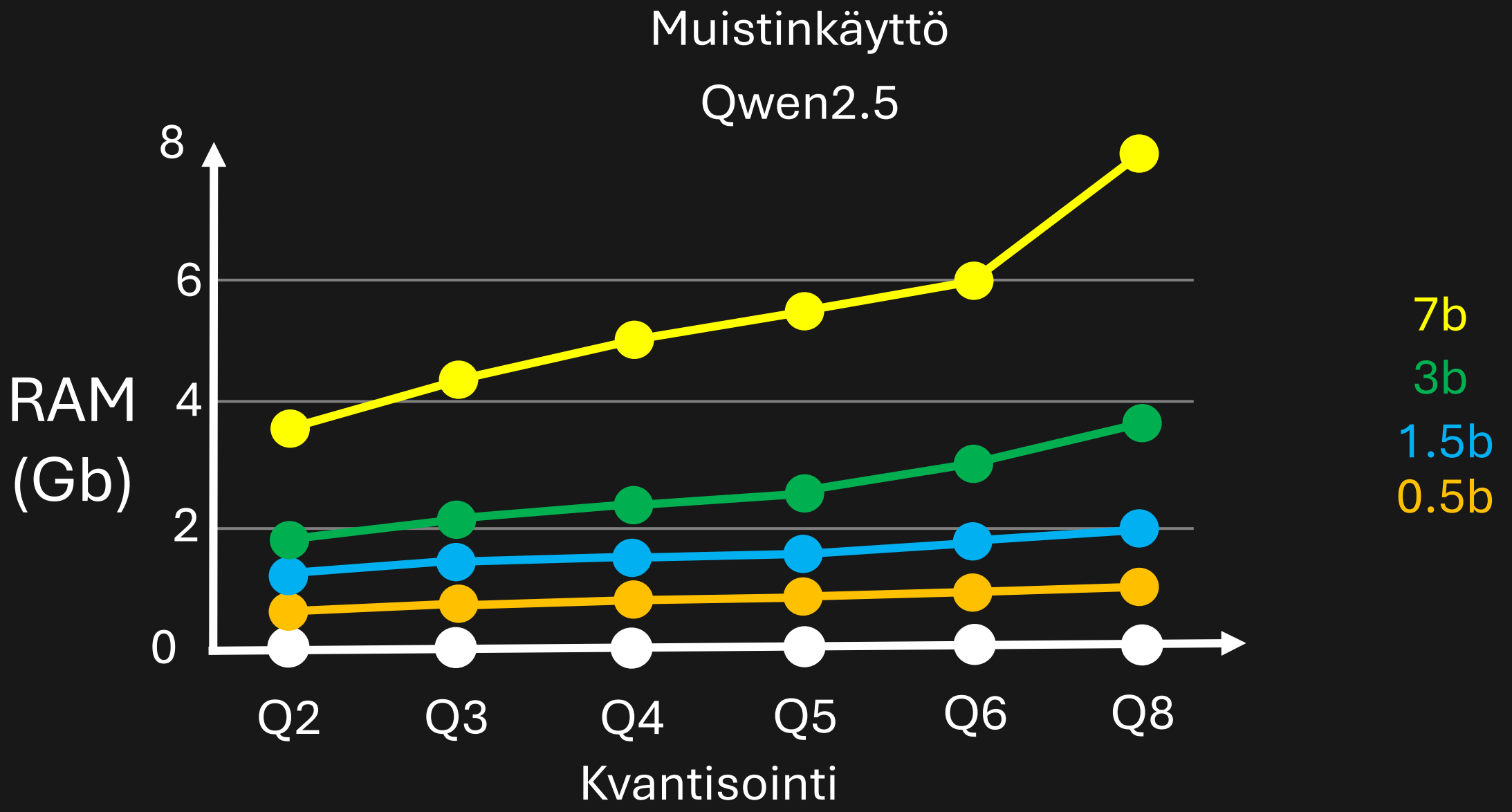
Kvantisoinnin vaikutus muistinkäyttöön, isoimmissa malleissa näkyy enemmän

Qwen2.5

RAM (Gb)

Kvantisointi Parametrit	Q2 _K	Q3 _K_S	Q3 _K_M	Q3 _K_L	Q4 _0	Q4 _1	Q4 _K_S	Q4 _K_M	Q5 _0	Q5 _1	Q5 _K_S	Q5 _K_M	Q6 _K	Q8 _0	FP16
0.5b	0.72	0.72	0.74	0.75	0.73	0.75	0.77	0.78	0.78	0.8	0.79	0.8	0.89	0.91	1.34
1.5b	1.16	1.24	1.3	1.36	1.41	1.49	1.52	1.56	1.65	1.74	1.77	1.89	2.08	2.2	2.09
3b	1.79	1.96	2.1	2.2	2.32	2.49	2.69	2.72	2.85	2.94	2.95	3.1	3.3	3.51	6.47
7b	3.64	4.04	4.35	4.62	4.88	5.27	5.31	5.43	5.68	6.06	6.09	6.29	6.5	6.95	14.45

Kielimalli

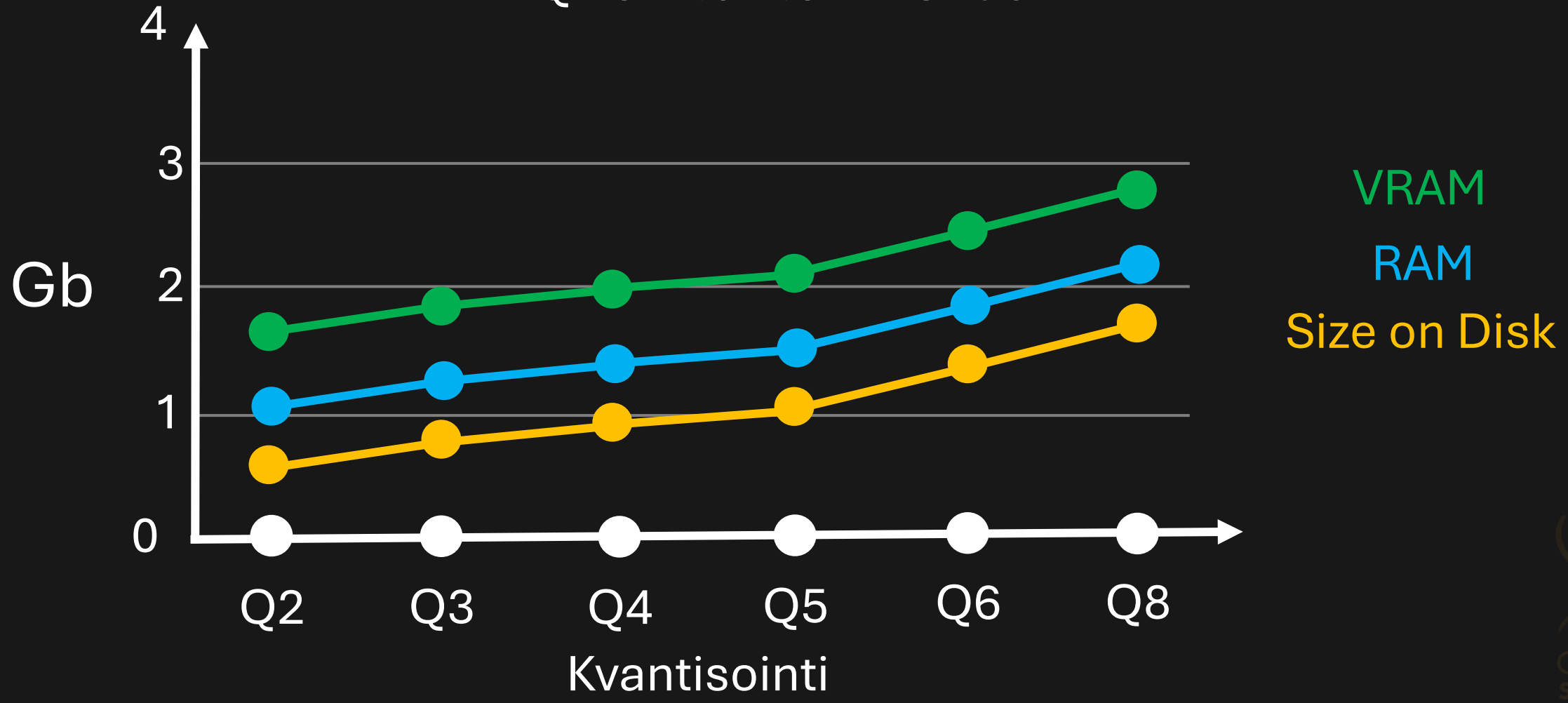


GPT Lab Seinäjoki

Kielimalli

Muistinkäyttö

Qwen2.5-1.5B-Instruct



GPT Lab Seinäjoki

Kielimalli

Kontekstin vaikutus laskentatehoon ja muistinkäyttöön (RTX4090)

Phi3-3.8b-instruct-Q4 (Tokens Per Second)

Konteksti Syöte	1024	2048	3072	4096	5120	6144	7168	8192	9216
Syöte = 1024	191	160	163	163	166	162	164	163	161
Syöte = Konteksi	187	193	179	175	164	153	151	142	138

Phi3-3.8b-instruct-Q4 (VRAM (Gb))

Konteksti Syöte	1024	2048	3072	4096	5120	6144	7168	8192	9216
Syöte = 1024	4.28	6.07	7.87	9.67	11.46	13.26	15.06	16.86	18.65
Syöte = Konteksi	4.28	6.07	7.87	9.67	11.46	13.26	15.06	16.86	18.65

Kielimalli

Nelinkertaisen kontekstin vaikutus laskentatehoihin

Phi3-3.8b-instruct-Q4 (Tokens Per Second)

Konteksti Syöte	1024	2048	3072	4096	5120	6144	7168	8192	9216
Syöte = 1024	191	160	163	163	160	162	164	163	161
Syöte = Konteksi	187	193	179	175	164	153	151	142	138

Phi3-3.8b-instruct-Q4 (VRAM (Gb))

Konteksti Syöte	1024	2048	3072	4096	5120	6144	7168	8192	9216
Syöte = 1024	4.28	6.07	7.87	9.67	11.46	13.26	15.06	16.86	18.65
Syöte = Konteksi	4.28	6.07	7.87	9.67	11.46	13.26	15.06	16.86	18.65

Kielimalli

Nelinkertaisen kontekstin vaikutus muistinkäyttöön

Phi3-3.8b-instruct-Q4 (Tokens Per Second)

Konteksti Syöte	1024	2048	3072	4096	5120	6144	7168	8192	9216
Syöte = 1024	191	160	163	163	166	162	164	163	161
Syöte = Konteksi	187	193	179	175	164	153	151	142	138

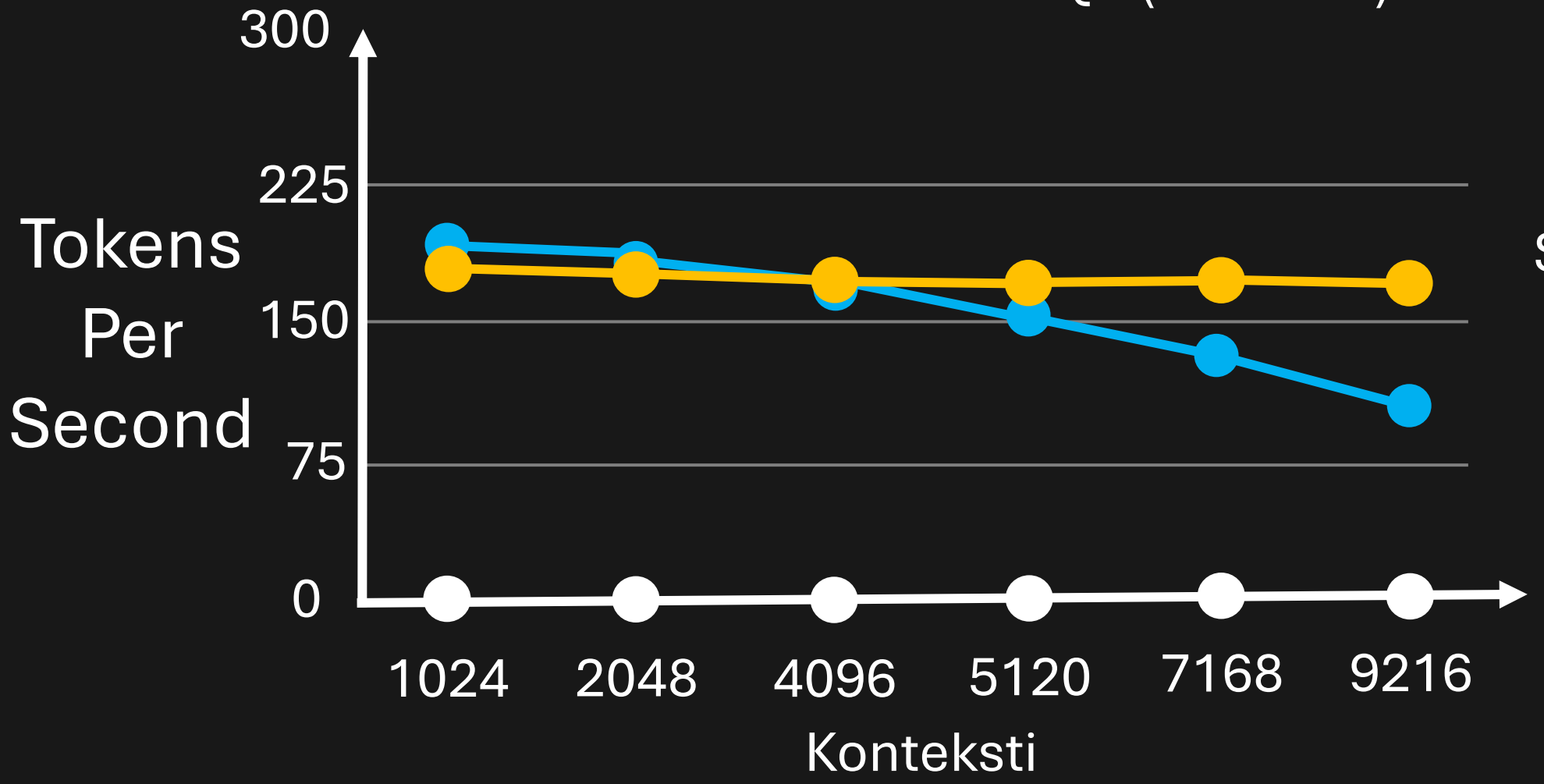
Phi3-3.8b-instruct-Q4 (VRAM (Gb))

Konteksti Syöte	1024	2048	3072	4096	5120	6144	7168	8192	9216
Syöte = 1024	4.28	6.07	7.87	9.67	11.46	13.26	15.06	16.86	18.65
Syöte = Konteksi	4.28	6.07	7.87	9.67	11.46	13.26	15.06	16.86	18.65

Kielimalli

Konteksin vaikutus laskentatehoihin

Phi3-3.8b-Instruct-Q4 (RTX4090)



Syöte (tokenia):

Konteksti

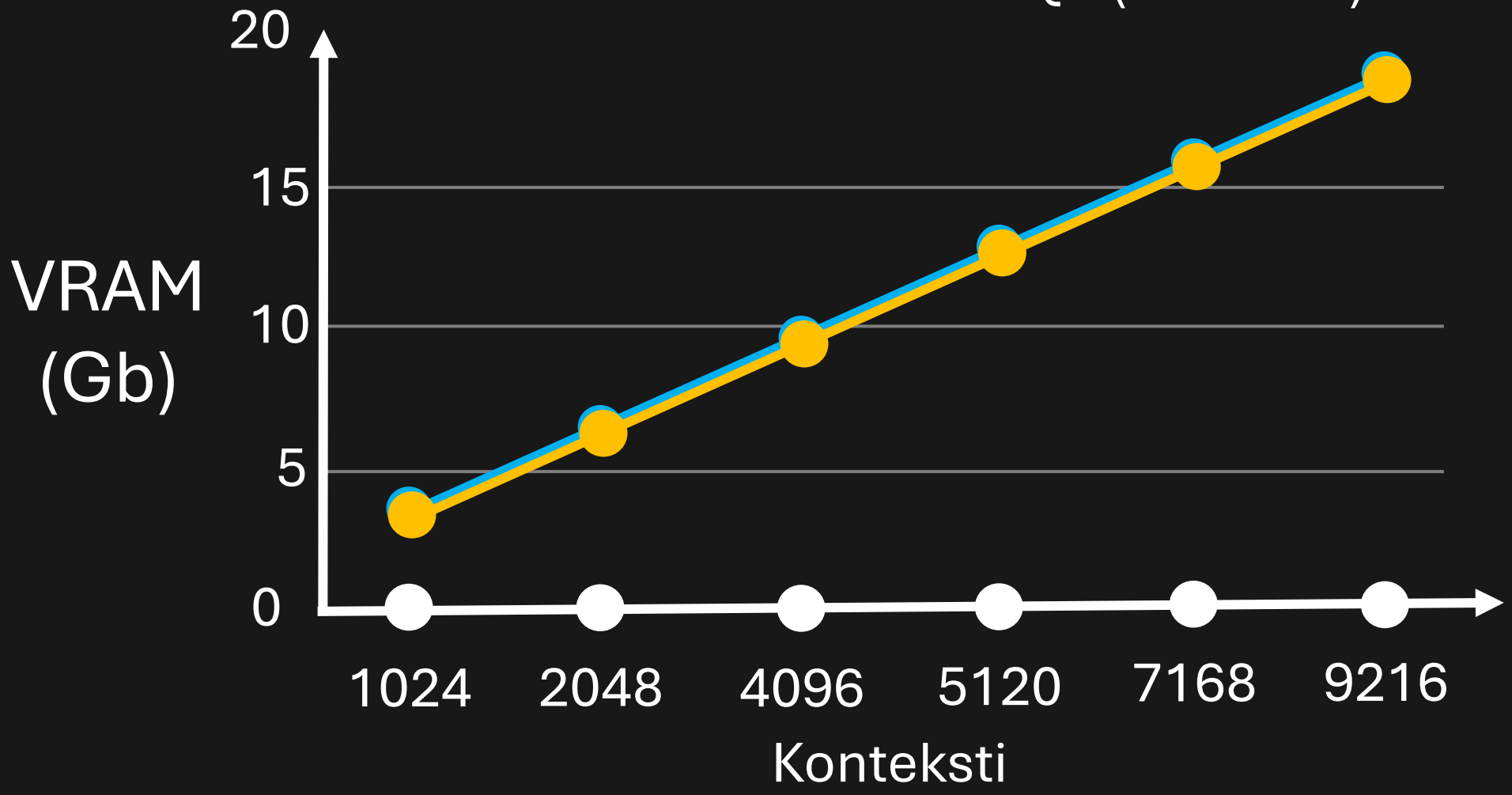
1024

GPT Lab Seinäjoki

Kielimalli

Konteksin vaikutus muistinkäyttöön

Phi3-3.8b-Instruct-Q4 (RTX4090)



Syöte (tokenia):

Konteksti

1024

GPT Lab Seinäjoki

Kielimalli

Miten valitaan oikea malli?

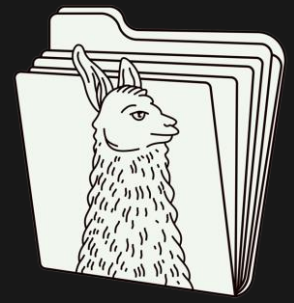
Valitaan malli joka on

1. Isoin mahdollinen
2. Sopivalla kvantisoinnilla
3. Jota koneen tehot riittävät ajamaan
4. Koulutettu ja hienosäädetty vastaamaan tarpeitamme
5. Arvioitu antavan parasta ulosantia

Provider

Mitä vaihtoehtoja on?

- Llama.cpp
- Ollama
- koboldcpp
- LM Studio
- vLLM
- TabbyML
- Llamafile
- LiteLLM
- Text Generation WebUI
- Open WebUI



Provider

Kaikkien rakastama taulukko

Provider	Lisenssi	GitHub Stars	Avoin Lähdekoodi	Alustat	Llama.cpp	Info
Ollama	MIT	92k	✓		✓	Suosituin, helppokäyttöisin, oma kirjasto
Llamafile	Apache	20k	✓		✓	Yksi ajettava tiedosto, mukana tuleva llama.cpp versio voi olla vanha, Windowsissa .exe rajoitukset
LLaMA C++	MIT	66k	✓		✓	Suurin osa muista pohjautuu tähän, lataa omat mallit, säädä itse parametrit
vLLM	Apache	28k	✓		✗	Vain Linux, tarkoitettu palvelimille ja tehokkaille näytönohjaimille
LM Studio	Proprietary	-	✗		✓	Helppokäyttöinen, lataa mallit huggingface, suosittu
TabbyML	Apache	21k	✓		✗	VSCode plugin joka voi ajaa myös itse malleja
koboldcpp	GNU	5k	✓		✓	Käytännössä Llama.cpp lisäominaisuuksilla

Plugin

Mitä vaihtoehtoja on?

Aktiiviset

- Continue
- Twinny
- TabbyML
- CodeGPT
- qodo

Epäaktiiviset

- LLM-VSCode
- Rubberduck
- Privy
- FireCoder

Pariohjelmoija

- Aider

qodo

Continue

CODE GPT

Plugin

Vielä yksi taulukko

Plugin	Lisenssi	GitHub Stars	VSCode Marketplace	Avoin Lähdekoodi	Paikalliset Mallit	Info
Continue	Apache	15k	260k	✓	✓	Monipuolinen, eniten ominaisuuksia, aktiivisin
Twinny	MIT	3k	20k	✓	✓	Pienempi, yksinkertaisempi
TabbyML	Apache	21k	33k	✓	✓	Ajaa itse mallit, hyvä autocomplete, vaatii docker tai enemmän setup
CODE GPT	Proprietary	-	1.4m	✗	✗	Osittain ilmainen, osittain maksullinen
qodo	Proprietary	-	450k	✗	✗	Entinen CodeiumAI, osittain ilmainen, osittain maksullinen

TYÖPAJA

GPT Lab
Seinäjoki

Työpaja

Mitä valittiin ja miksi?

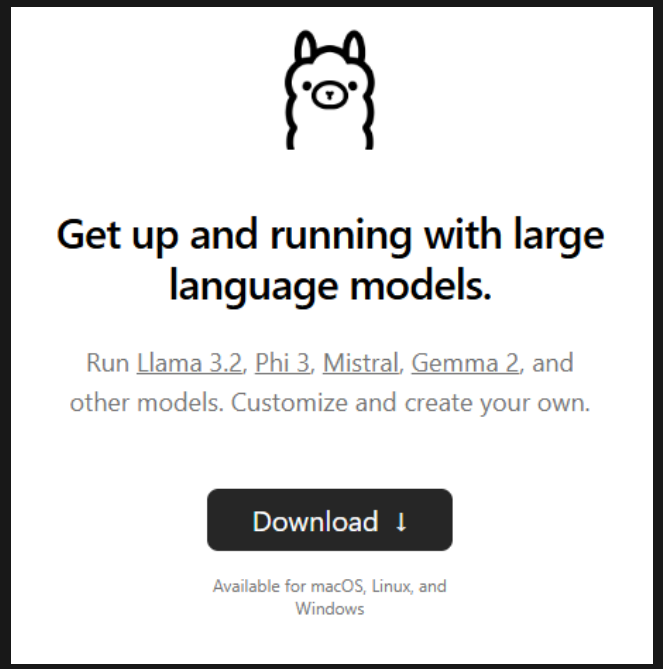
- Provider: Ollama
  - Suosituin ja helpoin käyttää
  - Ajovalmiit mallit, oma kirjasto
  - Toimii jokaisella alustalla
  - Ei vaadi admin oikeuksia, installer purkaa tiedostot
- Plugin: Continue
  - Suosituin ja eniten ominaisuuksia
- Kielimalli: Qwen2.5-Coder-0.5b/1.5b (Q4_K_M)
  - Autocomplete: Base, Chat: Instruct
  - Tällä hetkellä paras malli ohjelmointiin
  - Pyörii jokaisen koneella ja toimii hyvin pieneksi malliksi

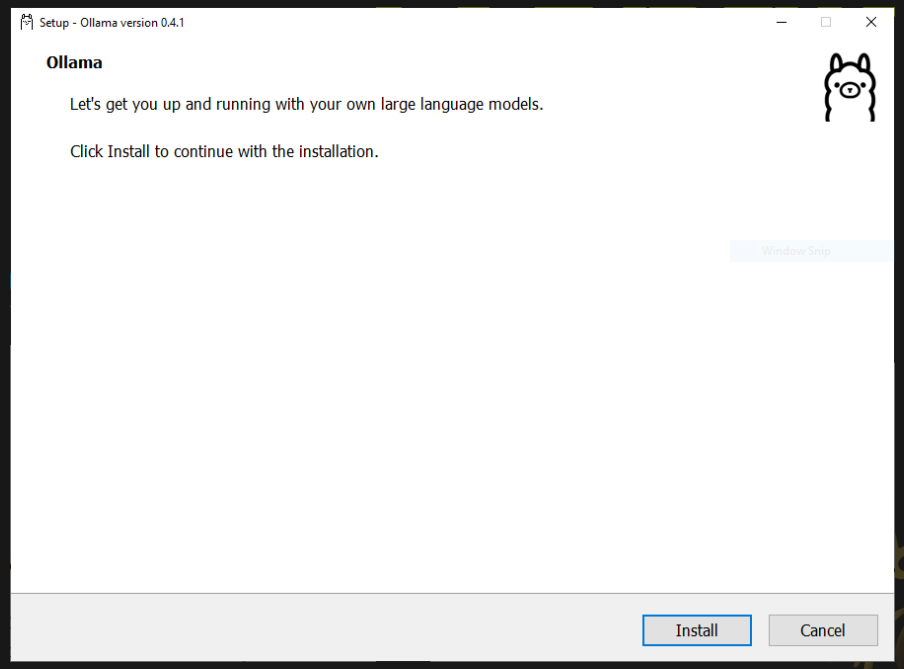
GPT Lab Seinäjoki

Työpaja

Ollama asennus (itse ohjelma)

https://ollama.com/

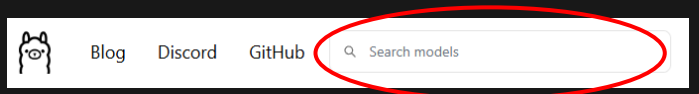


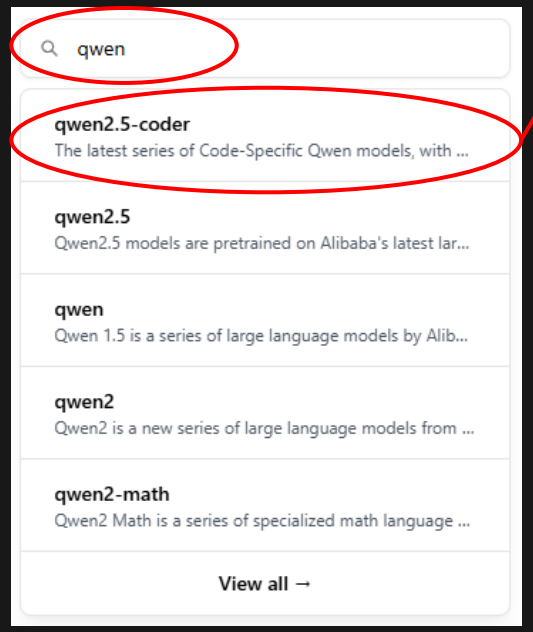


Työpaja

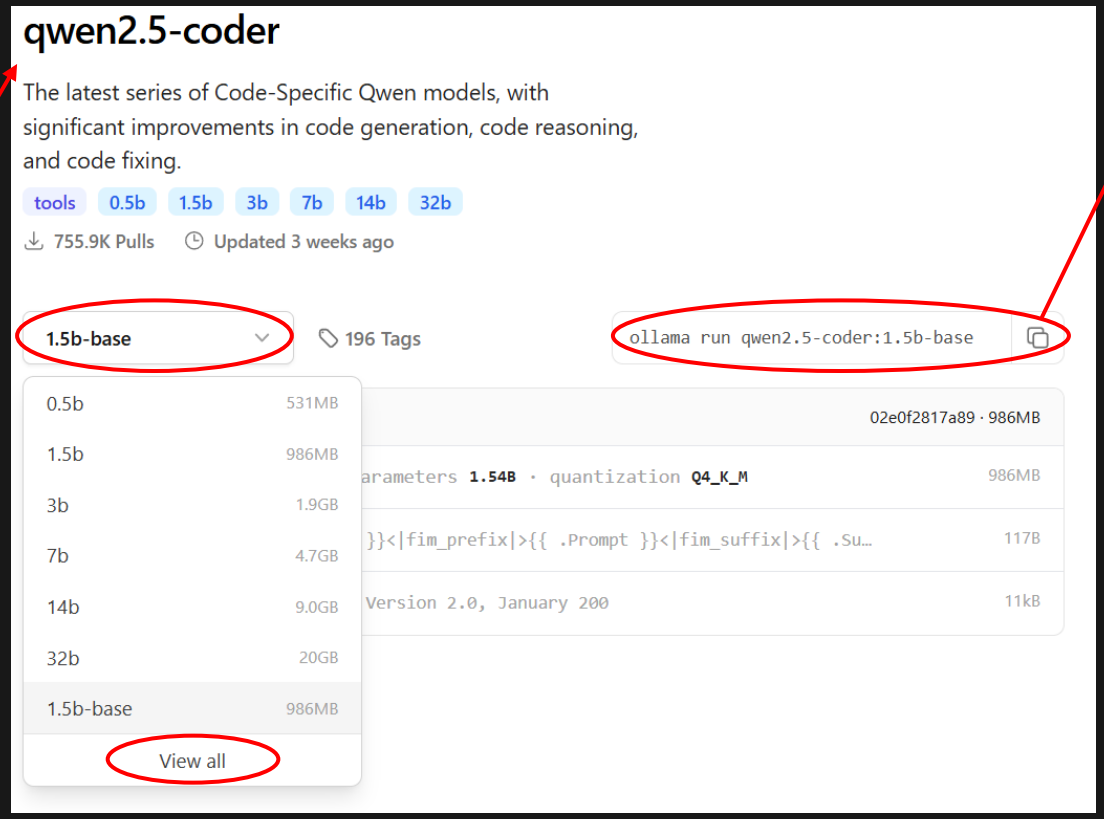
Ollama asennus (mallit)

1. Avaa https://ollama.com/





2. Lataa (qwen2.5-coder) 0.5b-base, 0.5b-instruct, 1.5b-base ja 1.5b-instruct



(ollama pull qwen2.5-coder:0.5b-base)
(ollama pull qwen2.5-coder:0.5b-instruct)
(ollama pull qwen2.5-coder:1.5b-base)
(ollama pull qwen2.5-coder:1.5b-instruct)

```
C:\Users\pmjual>ollama pull qwen2.5-coder:0.5b-base
pulling manifest
pulling 313e7db38447...  19%
```

ollama list näyttää mitä malleja on ladattu

```
C:\Users\pmjual>ollama list
NAME
qwen2.5-coder:0.5b-base
qwen2.5-coder:0.5b-instruct
qwen2.5-coder:1.5b-instruct
qwen2.5-coder:1.5b-base
```

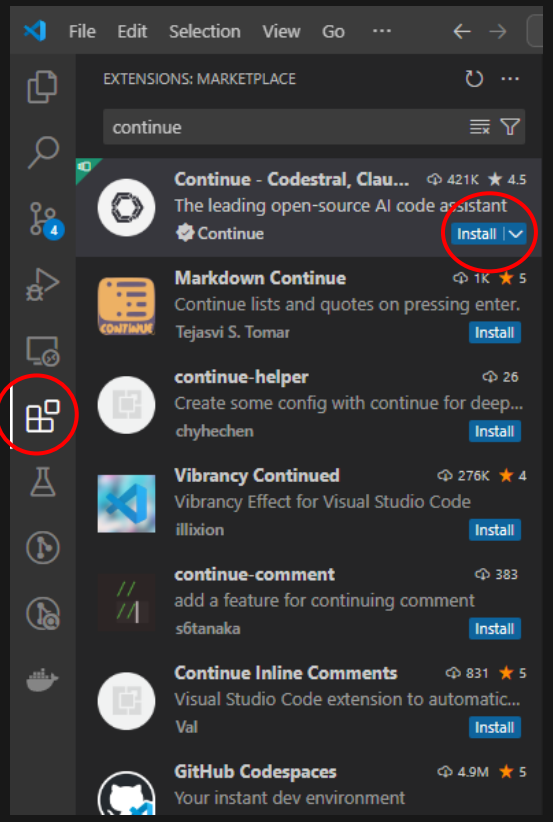
ollama ps näyttää mitä malleja on ajossa

```
C:\Users\pmjual>ollama ps
NAME
qwen2.5-coder:0.5b-base
```

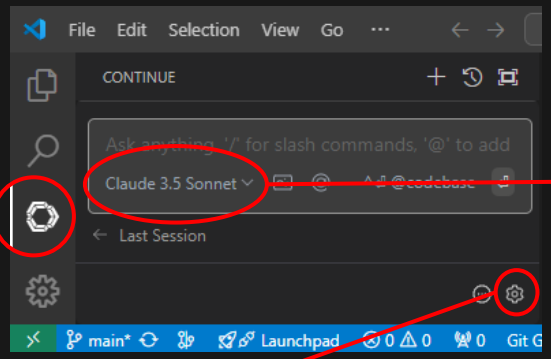
Työpaja

Continue asennus https://code.visualstudio.com/

1. Asenna continue



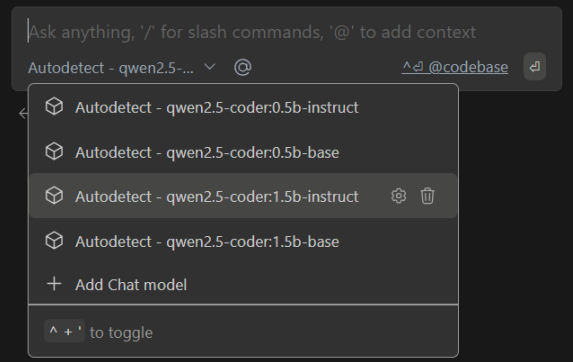
2. Rattaasta settings



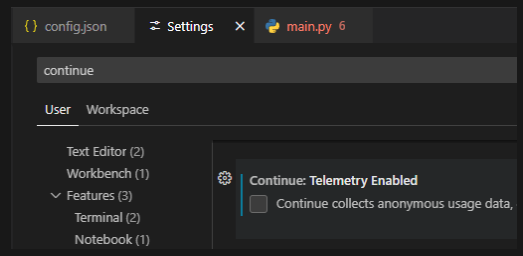
3. config.json (tallenna)

```
"models": [
  {
    "title": "Autodetect",
    "provider": "ollama",
    "model": "AUTODETECT"
  }
],
"tabAutocompleteModel": {
  "title": "NotAutodetect",
  "provider": "ollama",
  "model": "qwen2.5-coder:1.5b-base
},
```

4. Dropdown valikosta instruct-malli



5.Disable telemetry



Voit kokeilla sekä 0.5b (nopeampi, epätarkempi) ja 1.5b (hitaampi, tarkempi) Tai vaikka ladata itse isompia malleja (3b)

Base mallit autocomplete ehdotuksia varten

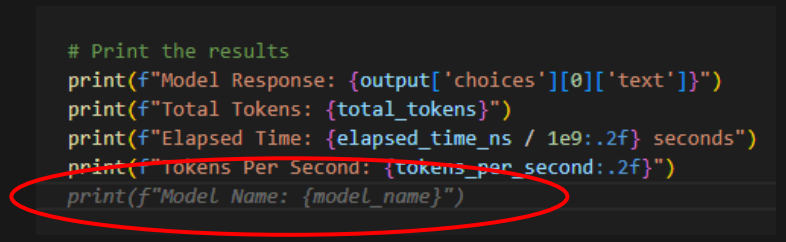
Instruct mallit kaikkeen muuhun (chat, ...)

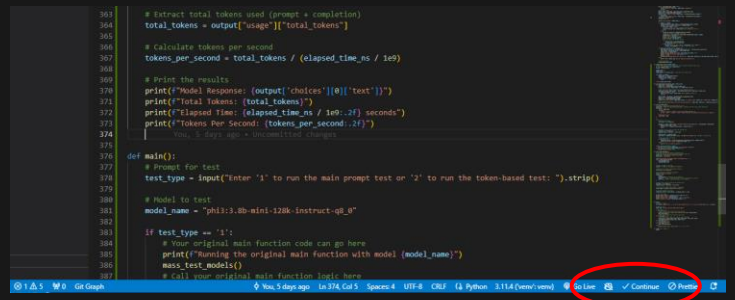
GPT Lab Seinäjoki

Työpaja

Continue Käyttöohjeet

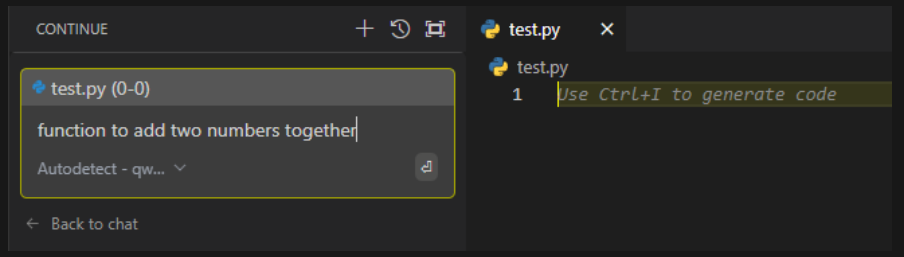
Autocomplete (TAB)



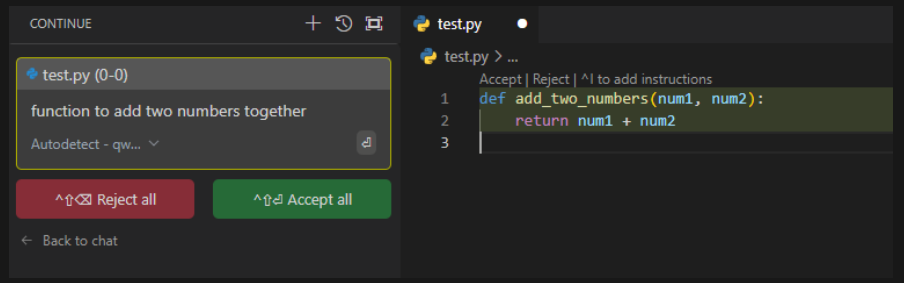


Disable Copilot

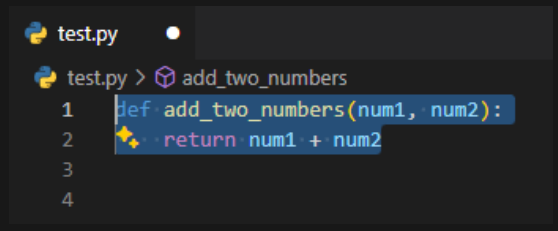
Prompting (CTRL+I)



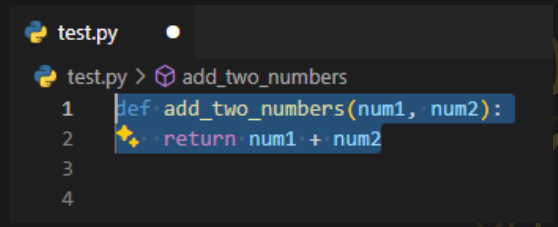
Accept (CTRL+SHIFT+ENTER)
Reject (CTRL+SHIFT+RETURN)



Editing (CTRL+I)



Chatting (CTRL+L)



Työpaja

Käytännön esimerkkejä käytöstä ja käyttötapauksista

(avaa vscode nyt)

GPT Lab Seinäjoki

Työpaja

Cheat Sheet

ollama help
ollama list
ollama ps
ollama pull <model>
ollama run <model>
ollama stop <model>
ollama rm <model>

Continue

Autocomplete
(TAB)

Prompting / Editing
(CTRL+I)

Chatting
(CTRL+L)

Accept
(CTRL+SHIFT+ENTER)

Reject
(CTRL+SHIFT+RETURN)
(CTRL+Z)

GPT Lab
Seinäjoki

JOHTOPÄÄTÖKSET

GPT Lab
Seinäjoki

Johtopäätökset

Mitä mieltä ollaan?

- Hyvin pienet mallit tuottavat yllättävän hyvää tekstiä, mutta eivät kuitenkaan aivan vielä pärjää isoille malleille eli päivittäiseen käyttöön vain poikkeustilanteissa, tosin vuoden päästä voi olla jo hyvin eri tilanne
- Keskikokoiset mallit ovat jo erittäin kilpailukykyisiä isojen kanssa, mutta vaativat jo paljon laskentatehoa eli mikäli löytyy tehokas näytönohjain, niin näitä voi jo suositella
- Paikalliset kielimallit ovat tulleet todella pitkän matkan jo vuodessa, uusia parempia tulee jatkuvasti ja tekniikat kehittyvät
- Visual Studio Code pluginit ovat hyvin kehittyvässä vaiheessa

GPT Lab Seinäjoki

Juha Ala-Rantala

THE END

Toivottavasti edes joku nautti?

GPT Lab Seinäjoki

ETELÄ-POHJANMAAN liitto

UCS SEINÄJOEN YLIOPISTOKESKUS

GPT Lab Seinäjoki

Raskun Jussin leikkuupöytä

GPT Lab Seinäjoki

Mixture of Experts (MoE)

Yksi menetelmistä ulosannin parantamiseksi

- MoE? Mixture of Experts?
  - Malli koostuu useasta erikoistuneesta osasta
  - Syötteestä riippuen vastausta tuottaessa vain tietyt niistä aktivoituvat
  - Paljon parametreja, suuri muistikäyttö
  - Suosio kasvaa, GPT4o on MoE, Grok on MoE, Llama4 tulee olemaan MoE?
  - Paikallisten mallien nimissä: MoE tai NxNB