

Keeleandmete olulisus keeletehnoloogia lahenduste arendamises

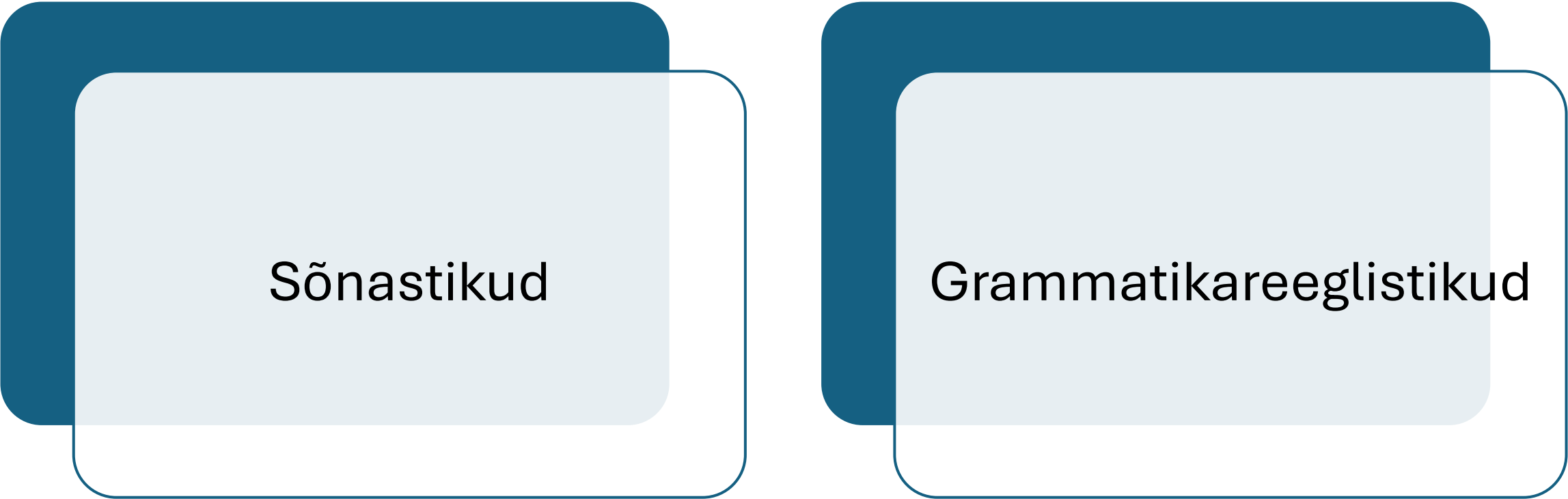
Importance of language data for the development of Language Technology solutions

Euroopa keeleandmeruumi töötuba, 11.06.2025

Kairit Sirts

TÜ keeletehnoloogia kaasprofessor

Reeglipõhised andmestikud



The diagram consists of two identical visual elements side-by-side. Each element is a light blue rounded rectangle with a dark blue border, containing a dark blue rounded rectangle with a light blue border. The text is centered within the dark blue inner rectangle.

Sõnastikud

Grammatikareeglistikud

Märgendatud andmestikud

Näiteks: pärisnimede tuvastamine

märgendid	-	Nimi	-	-	-
sisendlause	Kui	Arno	isaga	koolimajja	jõudis

Näide: grammatikavead

Grammatiliselt vigane
sisendlause

Tahaks veel rääkida Eesti **suve pealinnast** –
Pärnu**s**. See on tõesti **kuurort linn**.

Parandtud lause

Tahaks veel rääkida Eesti **suvepealinnast** –
Pärnu**st**. See on tõesti **kuurortlinn**.

Märgendatud andmestikud

Sisukokkuvõtted

Küsimustele vastused

Tõlgitud laused

Tekstide kategooriad: teemad, võtmesõnad, meelsus jne

Biologilised alused

Kaebamise neuroloogilised alused peituvad aju emotsionaalsetes piirkondades, eriti mandeltuuma aktiivsuses ja prefrontaalse ajukoore kontrollifunktsioonides. Neurootilised inimesed, kellel on madalam serotoniini ja ebastabiilne dopamiini tase, kalduvad rohkem kaebama. Perfektsionism ja katastroofimõtlemine aktiveerivad aju stressisüsteemi ülemäära, luues kroonilise kaebamise tsükli.

võimendaj	See ei ole lihtsalt kommunikatsiooniprobleem – see on süsteemne kriis, kus oleme kaotanud tasakaalu. Patsiendikoolide loomine võiks aidata inimestel paremini mõista tervishoiusüsteemi toimimist ja realistlikke ootusi seada.
-----------	---

Kaebajatega konstruktiivne toimetulek nõuab tasakaalukat lähenemist, mis arvestab nende bioloogiliste ja psühholoogiliste eripäradega. Oluline on valideerida tundeid, suunata lahenduste otsimisele ja seada piire pidevale negatiivsusele. Krooniliste kaebajate puhul võib olla vajalik soovitada psühholoogilist abi, eriti kui kaebamine seostub depressiooniga või neurootilisusega. Enda vaimse tervise kaitsmine on oluline – pidev negatiivne energia kuulamine võib mõjutada ka kuulaja meeleolu ja muuta neid ennast kaebekamaks.

Kaebamise kalduvus tuleneb seega neuroloogiliste, psühholoogiliste ja kultuuriliste tegurite keerukas koosmõjust. Kontrolli allikas mängib kesksel rollil, kultuurilised eripärad kujundavad kollektiivseid mustrid, ja digitaalne ajastus lisandub uus dimensioon, eriti tervishoius. Kaebamine ei ole pelgalt „halva iseloomu“ küsimus, vaid õpitud käitumismuster, mida saab muuta teadlikkuse ja oskuste arendamise kaudu.

depressiooni ja ärevusega seotud probleemide eest kui mõni poliitik ütleb „arstid peavad rohkem patsienti kuulama“ või „tervishoiu probleem hoopis kommunikatsioonis“, tunnen, kuidas mu kolleegid „vajuvad.“

Kliiniku juhina näen iga päev, kuidas head arstid võitlevad mitte ainult haigustega, vaid ka kasvava kaebuste ja kriitikatulvaga. See mõjutab nende motivatsiooni ja professionaalset eneseusku. Samas mõistan ka patsientide frustratsiooni – nad tulevad hirmul ja ootustega, mis ei pruugi alati täituda. Kuid me ei saa lasta sellel spiraalil edasi keerelda.

Me ei saa kaotada generatsiooni häid arste, kes lahkuvad meditsiinist mitte seetõttu, et ravida on raske, vaid seetõttu, et igapäevane süüdistamine muudab selle talumatuks. Lahendusi peitub vastastikusel mõistmises ja kommunikatsiooni parandamises. Lõppkokkuvõttes on meditsiini tuum alati olnud ja jääb inimese, kes aitab teist inimest. See põhimõte peab jääma kesksel ka kõigi muutuste keskel.

Treenimine märgendamata andmetel

Suured keelemudelid (ChatGPT ja muud tekstirobotid) on treenitud märgendamata andmetel.

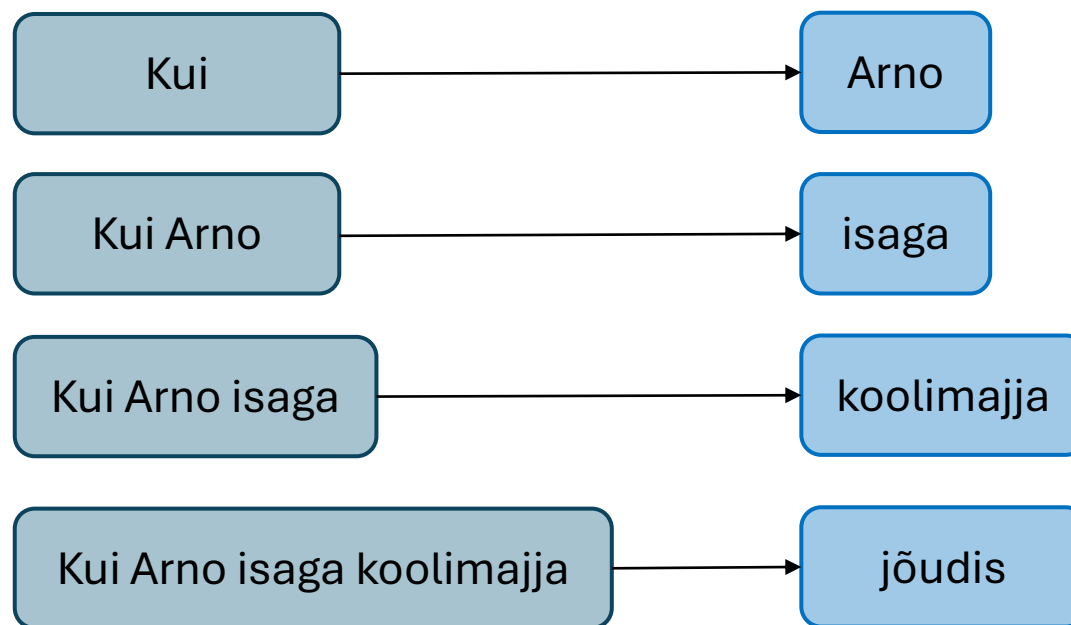
Mudel ennustab igal sammul järgmist sõna eelneva konteksti baasil.

Kui Arno isaga koolimajja jõudis, olid tunnid juba  alanud

Konteks **Järgmine sõna**

Treenimine märgendamata andmetel

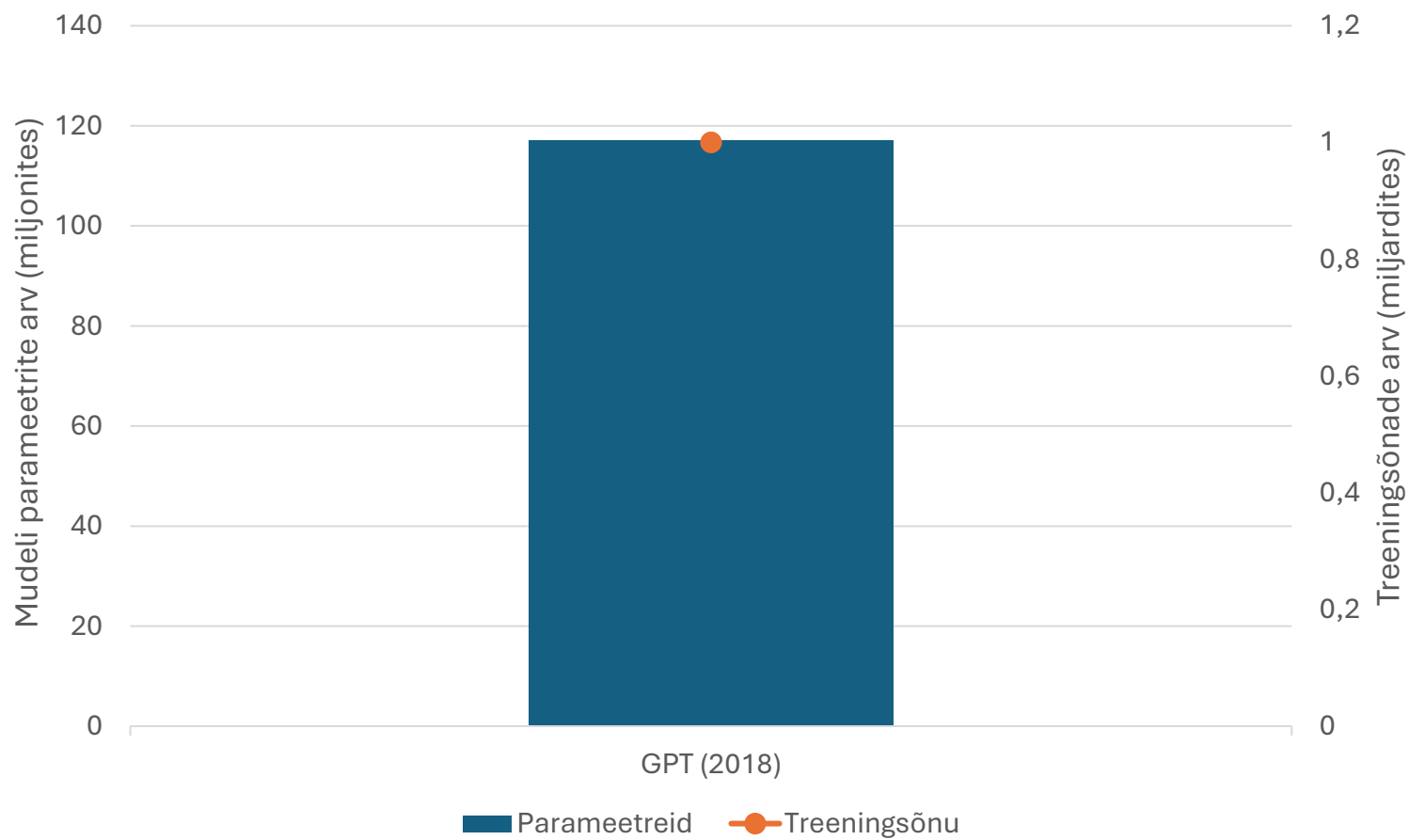
Kui Arno isaga koolimajja jõudis



**Kontekst ehk
Sisend**

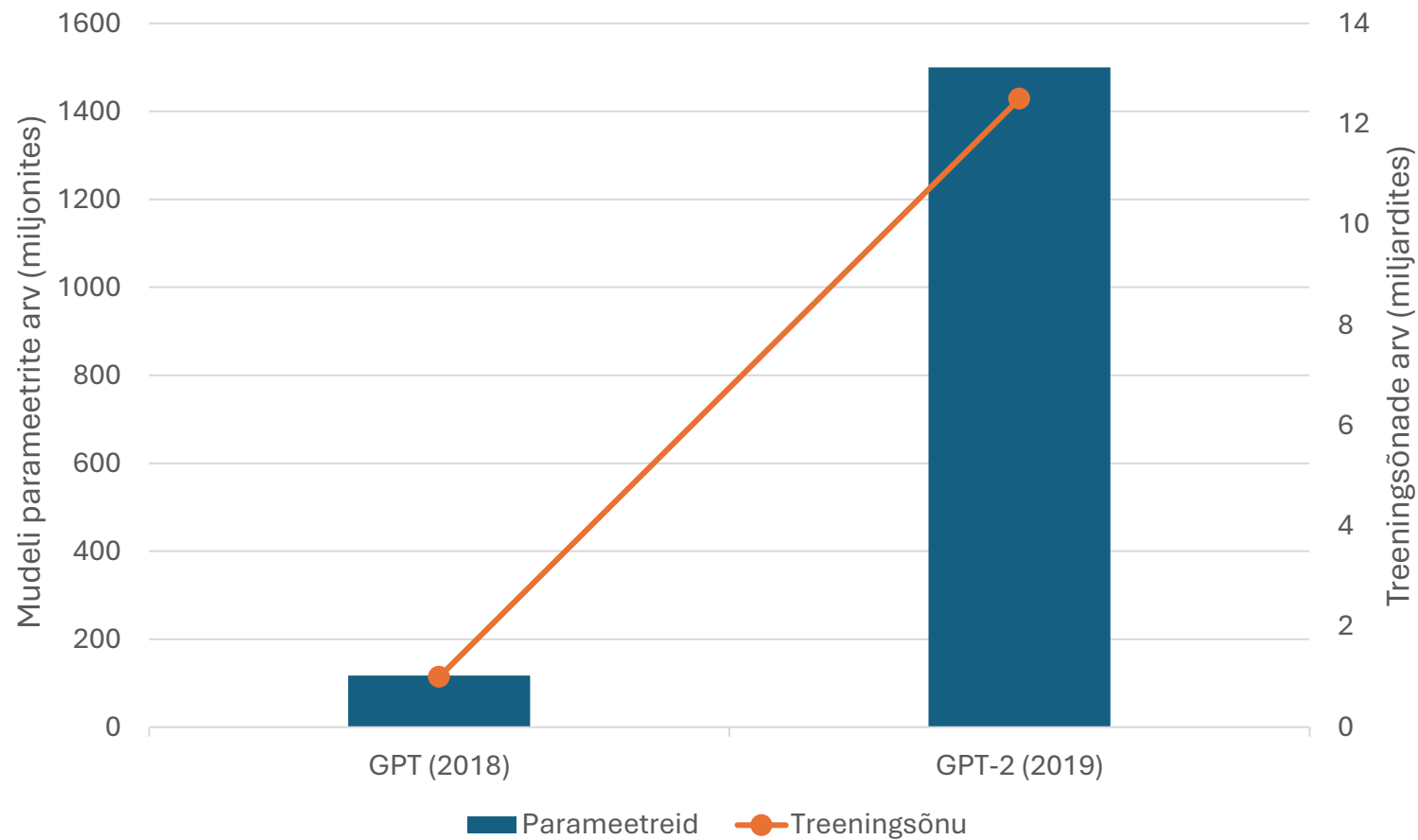
**Järgmine sõna ehk
Märgend**

OpenAI GPT mudelid



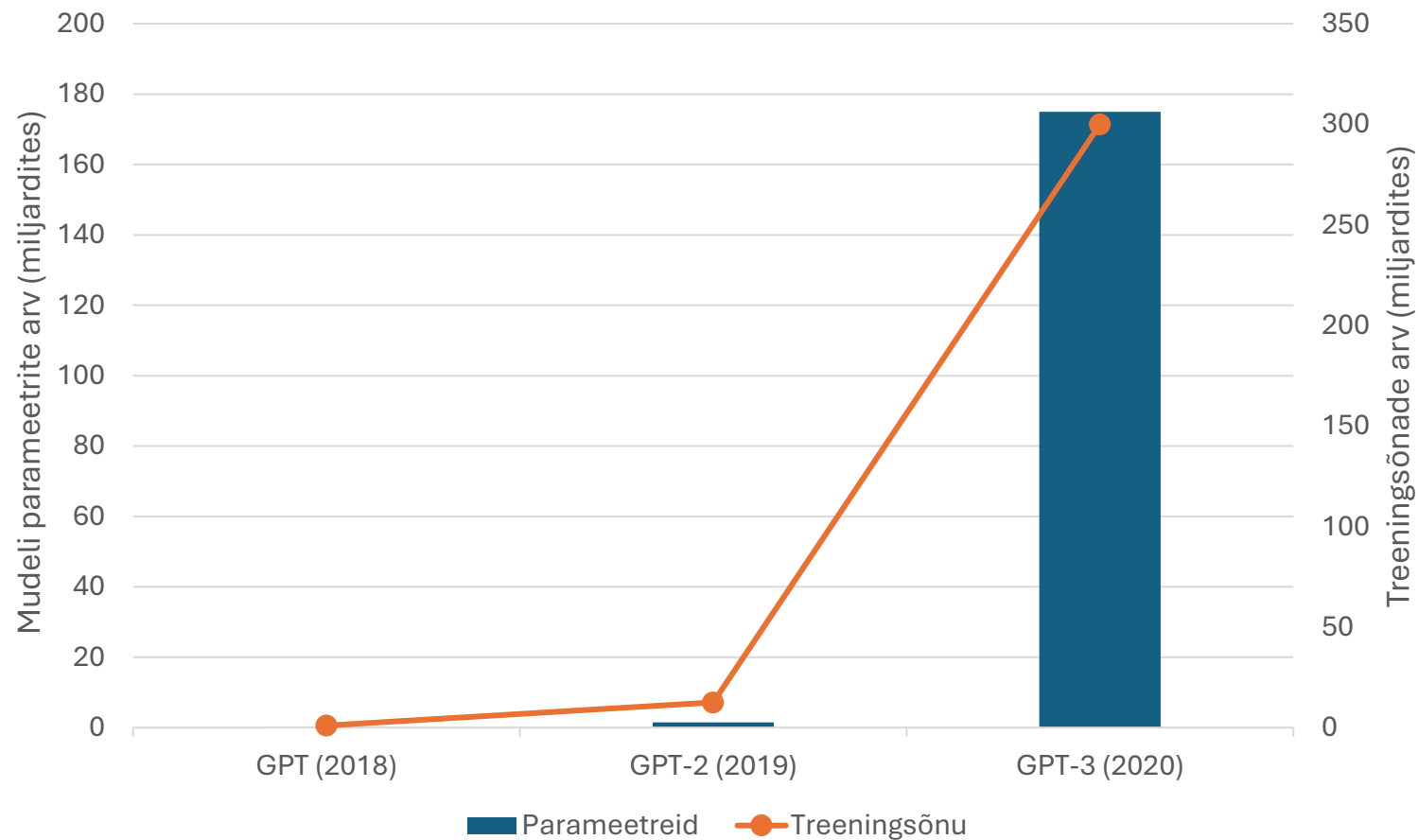
Märkus: treeningsõnade arv on ligikaudne

OpenAI GPT mudelid



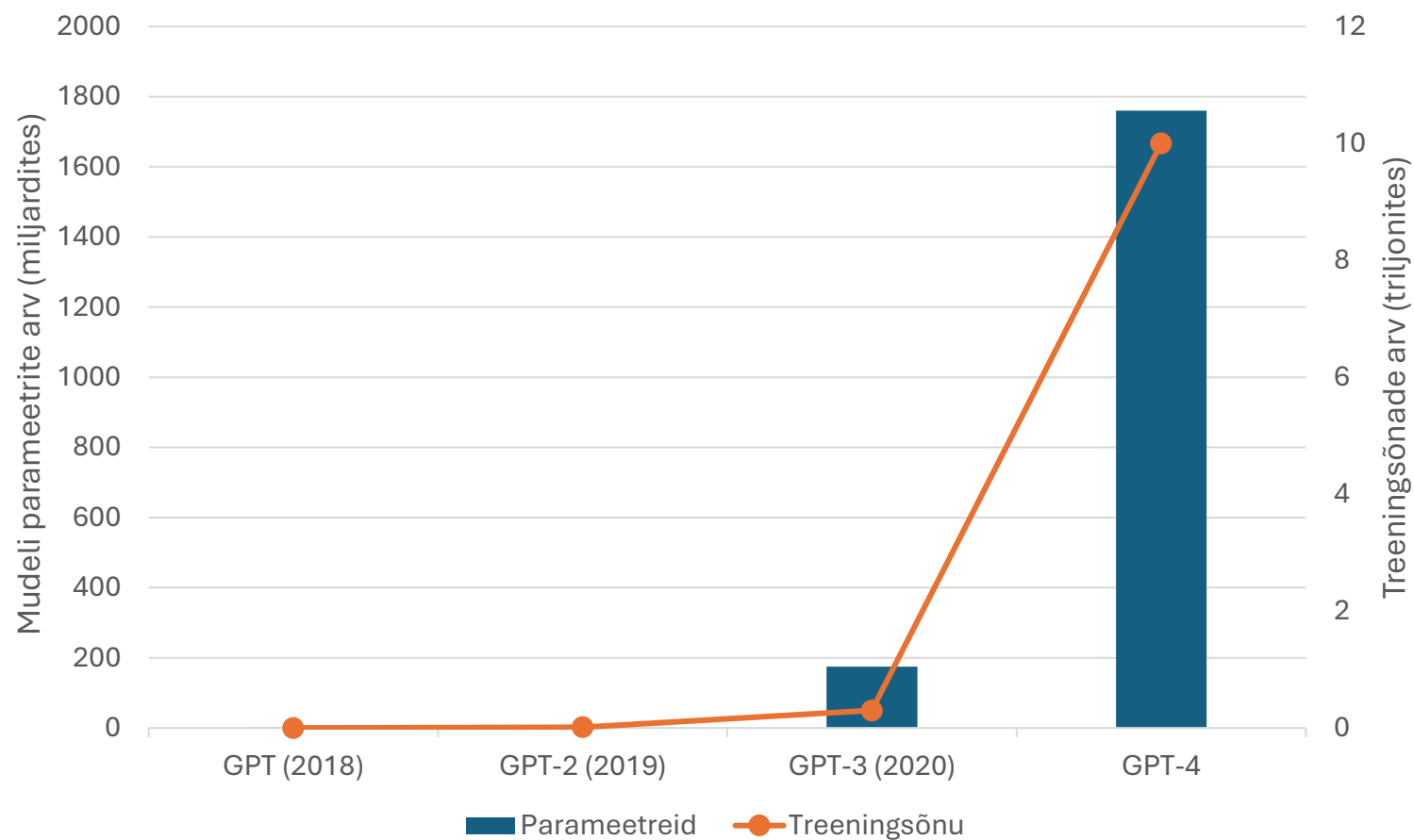
Märkus: treeningsõnade arv on ligikaudne

OpenAI GPT mudelid



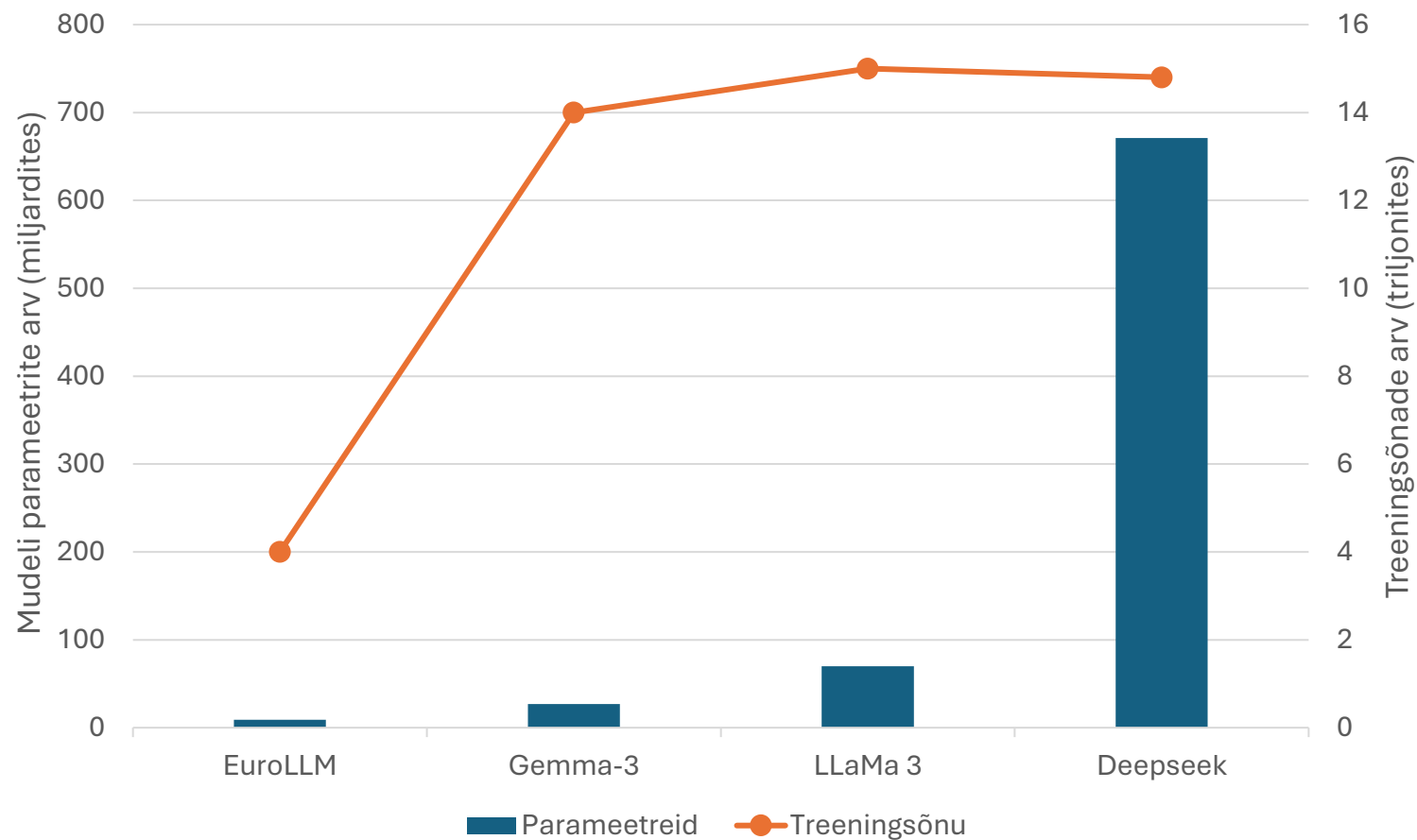
Märkus: treeningsõnade arv on ligikaudne

OpenAI GPT mudelid



Märkus: treeningsõnade arv on ligikaudne

Avatud mudelid



Märkus: treeningsõnade arv on ligikaudne

Kui palju teksti on 15 triljonit sõna?

Ühes romaanis keskmiselt ligikaudu 90,000 sõna

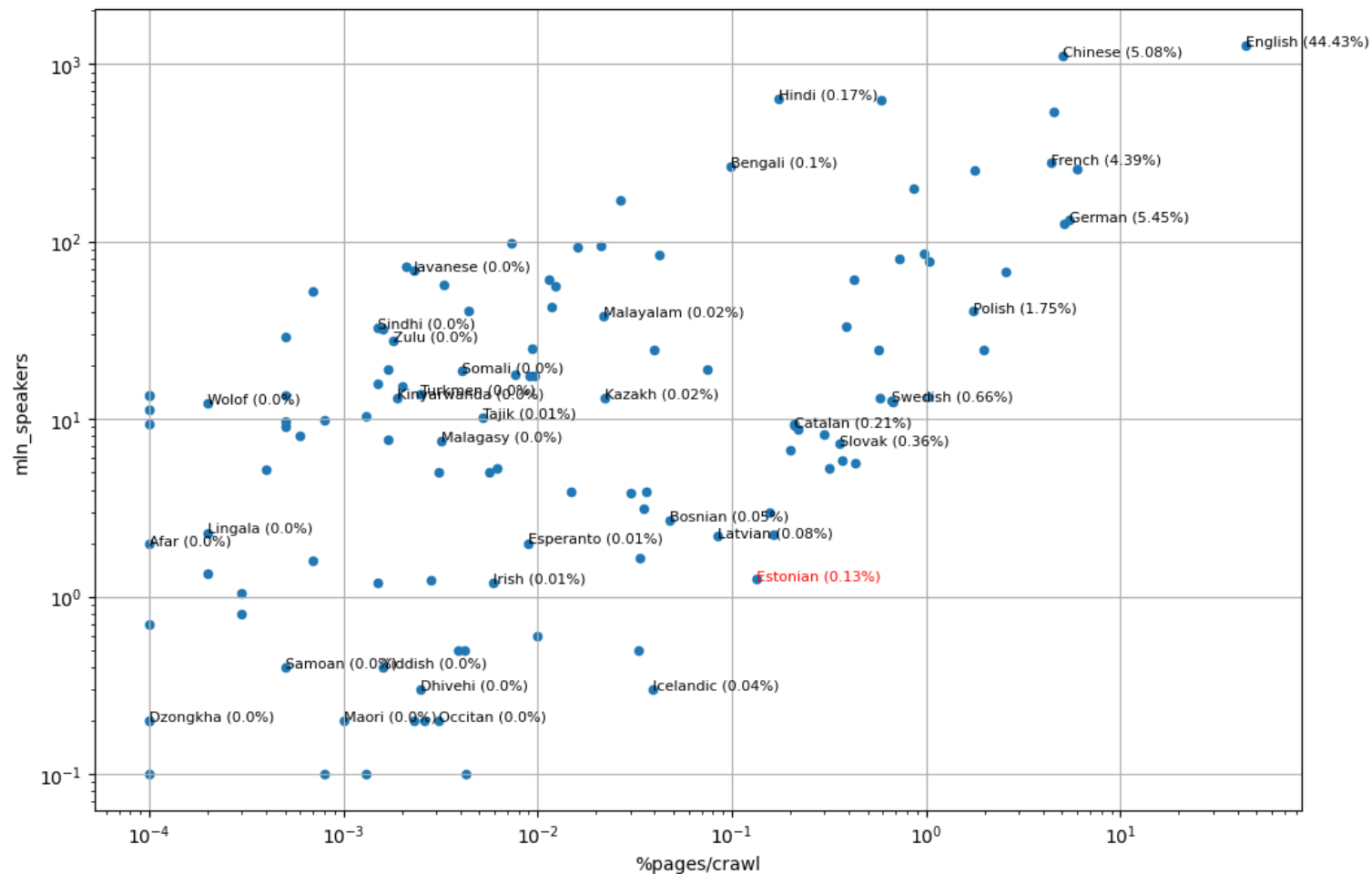
$15,000,000,000,000 / 90,000 \approx$ **170 miljonit romaani**

Vahemikus 1990-2024 eestikeelseid trükiseid ilmunud ligi **100,000** (allikas: Statistikaameti andmebaas)

Keskmiselt **2850 trükist aastas**

170 miljoni eestikeelse trükise kogunemiseks selles tempos kuluks ligikaudu **60,000 aastat**

Eestikeelse teksti osakaal CommonCrawl'is



Olemasolevad eestikeelsed tekstid

Eesti Ühendkorpus: ligikaudu **4.3 miljardit** sõna

Erinevatest veebikorpustest: ligikaudu **20.8 miljardit** sõna

- Puhastades ja duplikaate eemaldades jääb järele ligikaudu **5.8 miljardit** sõna

Ehk kokku on meil praegu ligikaudu **10 miljardit** sõna

Milleks neid andmeid kasutada saab?

Eeltöödeldud
korpused
baaskeelemudelite
arendajatele
kasutamiseks

Meie enda
pingutused avatud
mudelite
edasiõppimiseks

Kust saaks andmeid juurde?

Hoiatus! Hinnangud on väga umbkaudsed ja võivad osutuda ebatäpseteks.

- 1) Meediaväljaanded: hinnanguliselt 1 miljard sõna
 - Vaja kokkuleppeid meediamajadega
- 2) Eestikeelsed raamatud ja teosed: kuni 1 triljon sõna
 - Vaja ühiskondlikku kokkulepet autoriõiguse all olevate teoste kasutamiseks
- 3) Veebikorpustest: tuleb pidevalt juurde mingis mahus
 - Üha rohkem on veebis masintõlgitud ning tehisaruga genereeritud tekste



Kas märgendatud andmeid ka veel vaja läheb?

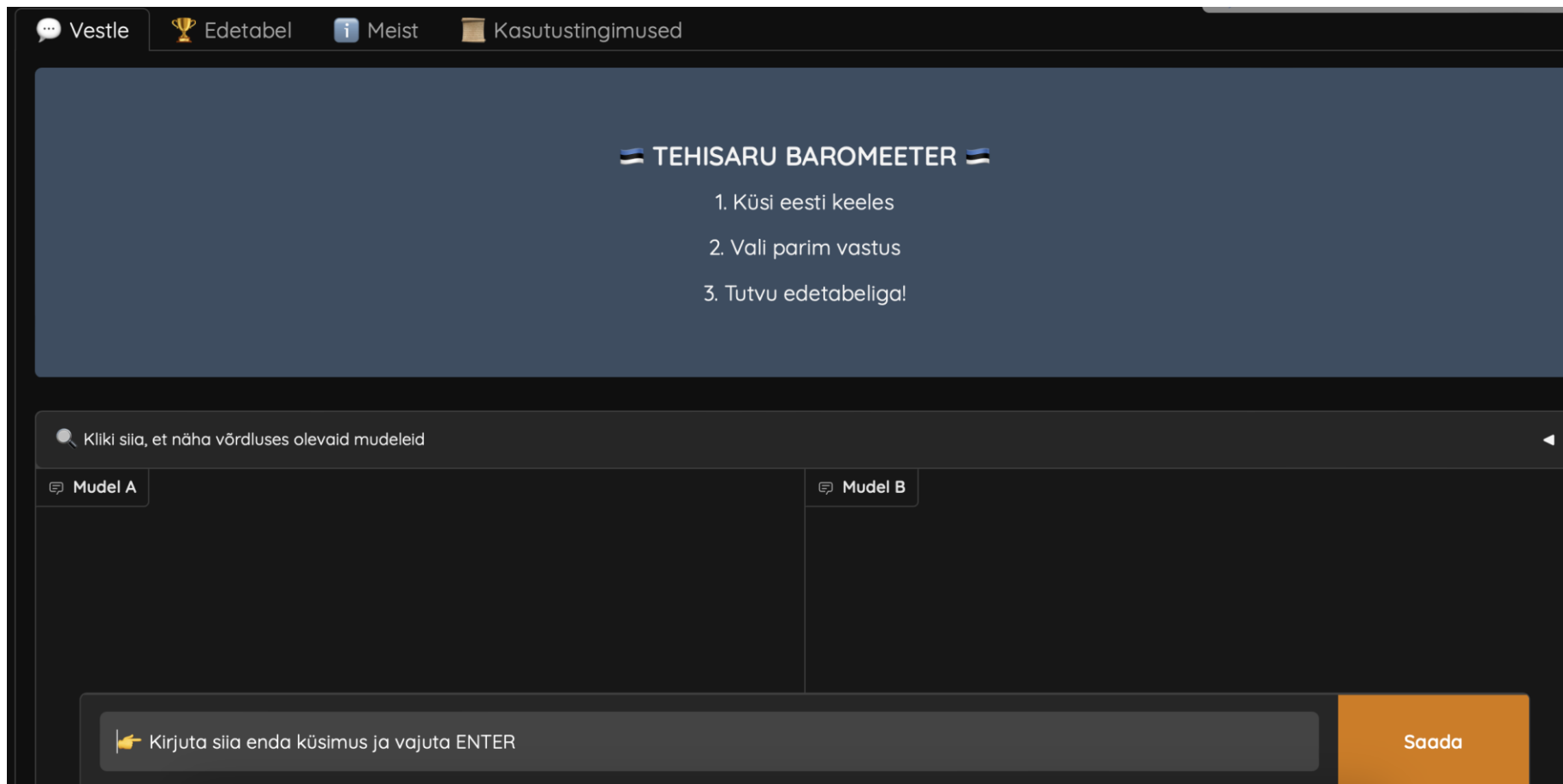
Jah, kindlasti!

- 1) Spetsiifilised ülesanded, mis ei vaja üldoskustega suuri keelemudeleid, näiteks:
 - Nimeüksuste tuvastamine dokumentide anonüümimiseks
 - Tekstide kategoriseerimine etteantud klassidesse
- 2) Suurte keelemudelite teadmiste ja oskuste hindamine
 - Erinevaid oskusi hindavad testandmestikud
- 3) Suurte keelemudelite joondamine
 - Mudelite peenhäälestamine instruktsioon-vastus paaridel

How do you heal gum after tooth extraction?

Gum healing after a tooth extraction is an important part of the overall healing process for tissue damage caused by the tooth extraction. In general, it is recommended to focus on good oral hygiene and diet to promote healthy gum tissue and avoid complications. Specifically, it is recommended to avoid eating hard or crunchy foods in the first few days, to brush the remaining teeth and gums gently but thoroughly, to rinse the mouth frequently with warm saltwater solution (1/4 tsp of table salt in an 8oz glass of lukewarm water), and to avoid tobacco products and acidic foods and beverages. As the gum tissue heals, it is important to continue good oral hygiene and dietary habits to avoid inflammation, infection, and accumulation of dental plaque and tartar. If you have any specific questions or concerns about your own recovery process, you may want to speak with a dental professional to discuss specific treatment recommendations tailored to your particular circumstances.

baromeeter.ai



Kas reeglipõhist keeletehnoloogiat ka veel vaja läheb?

Jah, kindlasti!

- 1) Keeleteaduses ja korpuslingvistikas
 - Näiteks: sõnaliikide ja morfoloogiliste atribuutide tuvastamiseks
- 2) Väga spetsiifiliste ülesannete lahendamiseks
 - Näiteks suurest hulgast tekstidest teatud kriteeriumitele vastavate tekstide filtreerimine
- 3) Suurtele keelemudelitele eksplitsiitselt sõnatähenduste ja keelereeglite õpetamine?
 - Sõnastike ja grammatikaõpikute abil

Kokkuvõtteks

Keeletehnoloogia areng sõltub andmetest

- Viimastel aastatel sõltub see üha enam **märgendamata** tekstiandmetest, mida on vaja väga suures mahus
- Samas on jätkuvalt vaja ka inimeste poolt erineval viisil märgendatud andmeid
- Oma spetsiifiline koht on jätkuvalt ka keele seaduspärasusi reeglite abil kirjeldavatel andmetel

Aitäh! Küsimused?

Kontakt: kairit.sirts@ut.ee