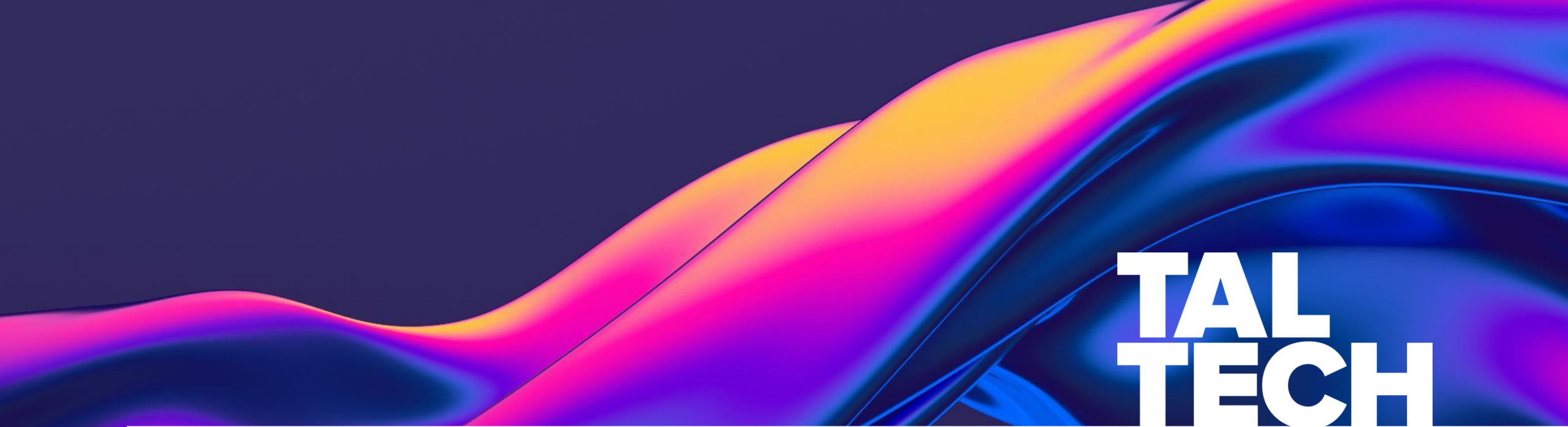




**TAL
TECH**

TEHISARU JA KEELEANDMED **VÄIKE SISSEJUHATUS**

Tanel Alumäe
Tarkvarateaduse instituut

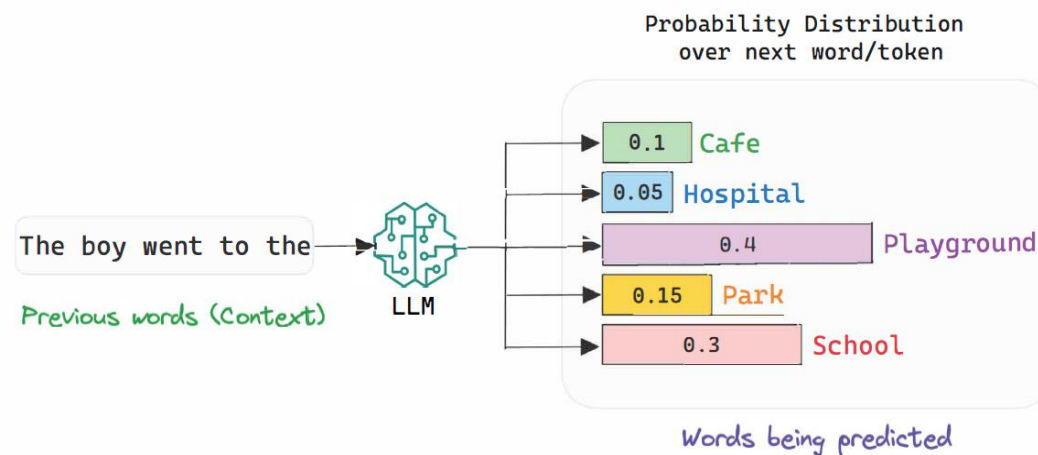
**TALLINN UNIVERSITY
OF TECHNOLOGY**

KUIDAS: EELTREENIMINE

- Keelemudel: *järgmise sõna ennustaja*
- Keelemudelid on ülisuured närvivõrgud, mis on eeltreenitud suure tekstikogumi peal (pmts. terve internet)
- Keelemudeli treenimisloogika:
 - Antud on 1000 (näiteks) eelmist sõna internetileheküljelt. Mis on järgmine sõna?
- Igast tekstist saame väga palju treeningnäiteid (*prompt* → sõna)
 - Gustav → Adolfi
 - Gustav Adolfi → Gümnaasium
 - Gustav Adolfi Gümnaasium → (
 - Gustav Adolfi Gümnaasium (→ lühend
 - Gustav Adolfi Gümnaasium (lühend → GAG

Gustav Adolfi Gümnaasium (lühend **GAG**) on üldhariduskool Tallinnas. See on Eesti vanim katkematult tegutsenud kool, mille asutas Rootsi kuningas Gustav II Adolf 1631. aastal.^[2]

Kooli põhiruumid paiknesid kuni 1990. aastani endise Tallinna Püha Miikaeli kloostri läänetiivas. 1990. aastal sai kool oma kättesse ka kloostri idatiiva. Võimlahoone teisel pool Suur-Kloostri tänavat valmis 1912. aastal. Tänapäeval õpivad peamajas (Suur-Kloostri 16) progümnaasiumi ja gümnaasiumi õpilased, Kotzebue õppehoones (Vana-Kalamaja 9) õpivad põhikooli I ja II kooliastme õpilased.^[2]



JÄRGMISE SÕNA ENNUSTAMINE == INTELEKT?

- Mida on vaja, et järgmist sõna hästi ennustada?
 - Teadmised maailma asjadest
 - Teadmised grammatikast või keelest üldisemalt
 - Teadmised mitmest keelest, oskus tõlkida
 - Matemaatika, mustrite ja teadmiste seostamine
 - Ehk teisisõnu, inimintellekt?

Gustav Adolphi Gümnaasium (lühend **GAG**) on üldhariduskool Tallinnas. See on Eesti vanim katkematult tegutsenud kool, mille asutas Rootsi kuningas Gustav II Adolf 1631. aastal.^[2]

Kooli põhiruumid paiknesid kuni 1990. aastani endise Tallinna Püha Miikaeli kloostri läänetiivas. 1990. aastal sai kool oma käsutusse ka kloostri idatiiva. Võimlahoone teisel pool Suur-Kloostri tänavat valmis 1912. aastal. Tänapäeval õpivad peamajas (Suur-Kloostri 16) progümnaasiumi ja gümnaasiumi õpilased, Kotzebue õppehoones (Vana-Kalamaja 9) õpivad põhikooli I ja II kooliastme õpilased.^[2]

TEADMISED KEELEMUDELIS

- Keelemudelil pole teadmusbaasi
- Kõik tema teadmised on peidetud hägusalt tema parameetritesse, mis kõik töötavad ühe eesmärgi nimel
- **Eesmärk: ennustada täpsemini järgmist sõna**
- Teadmised pole eriti usaldusväärsed ja mudel ei suuda neid alati "abstrahheerida"
- Siiski, kogu mudeli oskus tuleneb treeningtekstide "lugemisest"
- Tekstide-põhisele eeltreenimisele järgneb keelemudeli peenhäälestamine vestlusrobotiks, aga sellele kulub alla 1% arvutusressussi

⚡ Default (GPT-3.5)



Millal toimus Jüriöö ülestõus?



Jüriöö ülestõus toimus 1343. aastal Eestis. See oli ülestõus, mida juhtisid eestlased ning mille eesmärk oli võidelda Taani, Saksa ja Rootsi võimu vastu, mis oli tollal Eestis kehtestatud. Ülestõus sai alguse 1343. aasta 23. aprillil (Jüriöö) ning kestis mitme kuu jooksul.

⚡ Default (GPT-3.5)

Kas sa oskad nimetada mõne sündmuse, mis toimus 1343. aasta 23. aprillil?

Kahjuks ei ole minu teadmiste andmebaasist võimalik leida konkreetset sündmust, mis toimus 1343. aasta 23. aprillil. See oli väga ammu ja paljude minevikusündmuste kohta on säilinud vaid piiratud teave. Kui soovite rohkem teavet sellel kuupäeval toimunud sündmuste kohta, võiksite pöörduda ajalooliste arhiivide, ajaloolaste või spetsialiseeritud ajalooallikate poole.

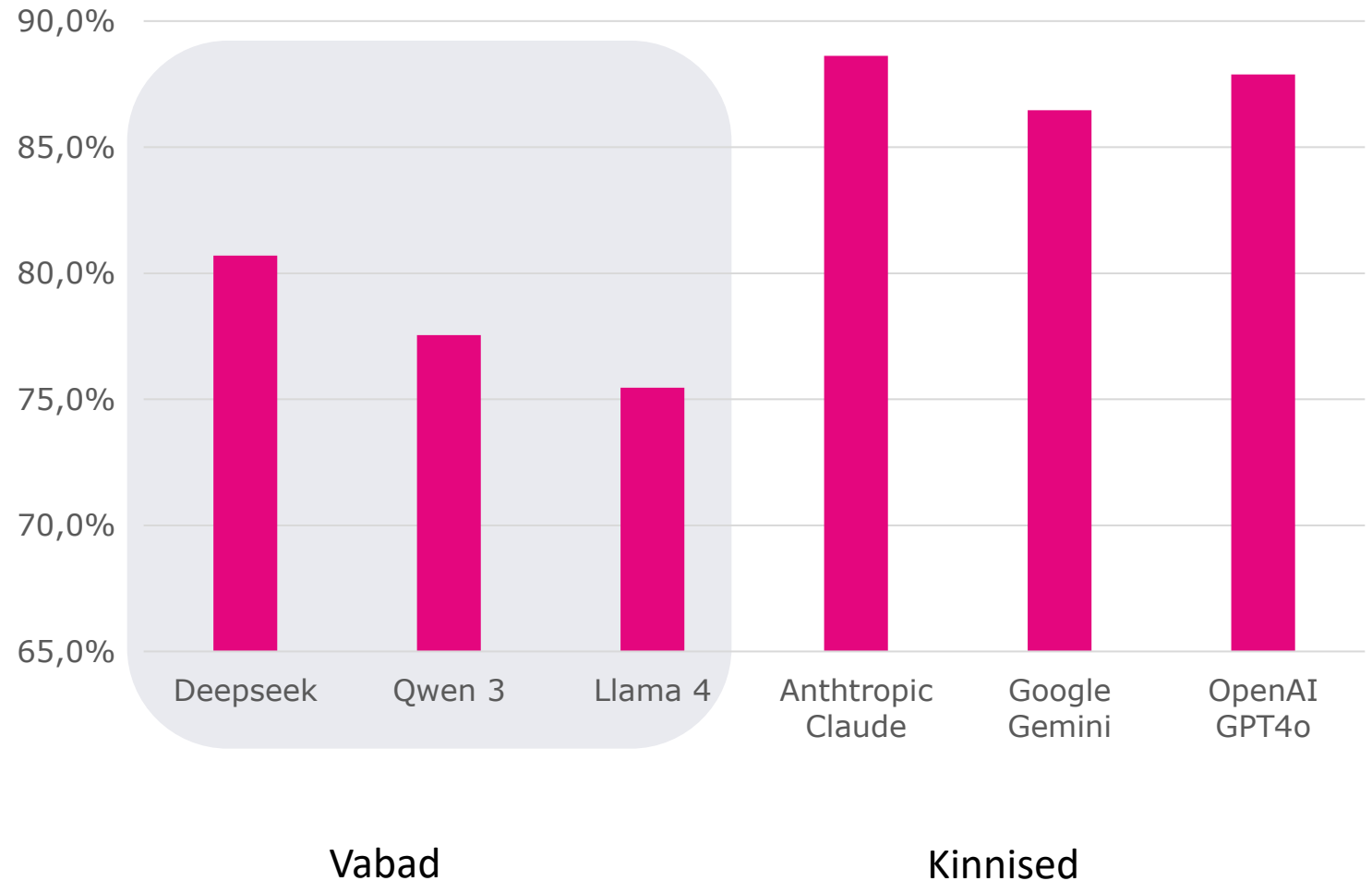
KEELEMUDELID: VABAD JA KINNISED

- Tehisaruks kasutatavad keelemudelid saab jaotada vabadeks ja kinnisteks
- **Kinnised:** keelemudel jookseb teenusepakkuja serveris, seda alla laadida ei saa
 - *OpenAI ChatGPT, Google Gemini, Anthropic Claude*
- **Vabad:**
 - Mudelit saab **alla laadida**
 - Mudelit saab jooksutada **oma serveris** või oma arvutis (väiksemate mudelite puhul)
 - Mudelit saab **peenhäälestada** (s.t. treenida edasi oma andmetega)
 - Näiteks: Meta AI **Llama**, **Deepseek**, **Qwen**
 - Mudelil on litsents, mis võib piirata teatud tingimustel kasutamist
 - Miks vabad? Keelemudeli treenimine maksab miljoneid eurosid!
 - Keerulised strateegilised valikud

VABADE JA KINNISTE KEELEMUDELITE VÕRDlus

- Keelemudeleid saab võrrelda, kasutades spetsiaalseid testandmestikke
- Meil on koostatud Eesti põhikooli- ja gümnaasiumi eksamite põhjal keelemudelite test
- Näiteks küsimus ajaloost:
 - *Mis linna lähistel toimusid Landeswehri sõja otsustavad lahingud, mille võitmist tähistatakse Eestis võidupühana?*
 - Variandid: Rapla, Rakvere, **Võnnu**, Narva

Eksamiküsimustele vastamise täpsus



BAROMEETER.AI

- Hiljuti avalikustatud beenirakendus *baromeeter.ai*
- Vestle kahe juhuslikult valitud mudeliga ja vali, kumb on parem
- Esimesed 8 on kinnised mudelid
- Parim avatud mudel *Deepseek*

Mudel A

Kes on kõige kiirem mereimetaja?

Kõige kiirem mereimetaja on delfiin. Täpsemalt, see tiitel kuulub tavaliselt lühinokast delfiinile (Stenella clymene), kes suudab saavutada kiirusi kuni 60 km/h (umbes 37 mph). Delfiinid on tuntud oma erakordse kiiruse ja agility poolest, mis aitab neil kiskjatest põgeneda ja saaki püüda.

A on parem

Mudel B

Kes on kõige kiirem mereimetaja?

Kõige kiirem mereimetaja on mõõkvaal, kes suudab ujuda kiirusega kuni 56 km/h.

B on parem

Mudel	▲	Skoor	▲
Gemini-2.5-Pro-Exp-03-25		1590	
GPT-4o-2024-05-13		1570	
Grok-3...(beta)		1553	
Claude-3.5-Sonnet (20241022)		1534	
Claude-3.7-Sonnet (20250219)		1524	
GPT-4-Turbo-2024-04-09		1514	
Claude-3-Opus...(20240229)		1500	
Gemini-2.0-Flash-001		1500	
<u>DeepSeek-V3...(0324)</u>		1495	
Gemini-2.0-Flash-Lite-001		1477	
Gemini-1.5-Pro-002		1452	
Grok-3-mini...(beta)		1455	
Claude-3.5-Haiku (20241022)		1396	
Claude-3-Haiku...(20240307)		1394	
<u>Gemma-3-27B-it</u>		1392	
Gemini-1.5-Flash-002		1389	

KINNISED MUDELID: EESTI KEEL VS INGLISE KEEL

- **MMLU-PRO**: vastusevariantidega küsimused inglise keeles
- Tõlgitud automaatselt eesti keelde
- Näiteks:
 - *Küsimus: Kuidas nimetatakse vahet varade ja kohustuste vahel?*
 - *Variandid:*

*A: Bilanss B: Võlg C: Kulud D: Netoväärtus E: Kasum F: **Omakapital***
G: Investeering H: Rahavoog I: Tulu

Võrdlus:

- Kinnised mudelid
 - Inglise keeles: **OpenAI ChatGPT: 78%**
 - Eesti keeles: **OpenAI ChatGPT: 50%**
- Vabad mudelid
 - Inglise keeles: Llama 3.1 (70B): 56%
 - Eesti keeles: Llama 3.1 (70B): 35%

KEELEMUDELI KVALITEET VS TREENINGANDMED

- Mitmed uuringud on näidanud, et keelemudeli vastuste kvaliteet mingis keeles esitatud küsimustele on tugevas korrelatsioonis vastava keele treeningandmete mahuga
 - Mida rohkem eestikeelseid andmeid, seda paremini ka tehisaru eesti keeles töötab
- Samas, oluline on ka **kvaliteet** ja **mitmekesisus**
 - Kvaliteetsed andmed annavad parema tulemuse kui samas mahus mitte-kvaliteetsed andmed
- Mis on kvaliteetsed andmed?
 - Inimese kirjutatud haridusliku väärtusega tekstid
 - Sotsiokultuuriliselt sisukad tekstid
- Mittekvaliteetsed andmed: masintõlgitud lehed, sisuturundus jms

EESTIKEELSESED TEKSTIANDMED

- EKI ühendkorpus: **4.3 miljardit** sõna
- Teised rahvusvaheliselt kogutud andmestikud – HPLT, CulturaX, jms
- Kokku - 25 miljardit sõna
- Aga: palju duplitseeritud andmeid, peale filtreerimist kokku **7.5 miljardit**
- Need andmed on kõik internetist ja vabalt kasutatavad ka kommertsmodelite treenimiseks (EL autoriõiguse masinõppe erand)
- Postimehe grupist ca 160 miljonit sõna on juba EKI ühendkorpuses, neilt saaks juurde 320 miljonit sõna (kokku on ca 480 miljonit sõna). Leping EKTБ104 projektiga on läbirääkimistes.
- **Rahvusraamatukoguga** on EKTБ104 projektile leping:
 - **3.5M toimikut**, kuid mitte kõik ei ole digitekst
 - Neil on ka veebiarhiiv, võimalik et kattub juba avalike andmetega
 - Juristid on nõus, et kinniste andmetega (nät. RR) treenitud mudeli võib avalikustada
- ERR on pigem vähem koostöövalmis (küsub raha)

KEELEANDMED

- Strateegiliselt on hea, et võimalikult palju eestikeelseid andmeid oleks vabalt kasutatavad, ka kinniste mudelite jaoks
- On arvata, et kommertsmodelid on alati sammukese ees, paljude ülesannete jaoks on nende kasutamine mõistlik (vähemalt lühikeses perspektiivis)
- Palju häid eestikeelseid mudeleid == valikuvabadus, konkurents
- Eesti riik ei suuda ilmselt tavakasutaja jaoks konkurentsivõimelist kasutajakogemust pakkuda
- Siiski, paljud institutsionaalsed ja tundlikud kasutusjuhud eeldavad majasiseste mudelite kasutamist, mis eeldab hea vaba keelemudeli olemasolu

MIDA TEHA?

- Vaba keelemudel, mis oskab eesti keelt juba algusest peale, on parem kui eesti keelele peenhäälestatud mudel
- Uusi vabasid mudeleid tuleb pidevalt
 - Ideaalis nad ei peaks vajama peenhäälestamist
 - Palju on teenusepakkujaid, mis vabasid mudeleid üliodavalt (efektiivselt) jooksutavad
 - Aga nad ei viitsiks hakata eesti keele spetsiifilist mudelit serveerima
 - S.t., lootus kasutada meie endi treenitud "Eesti keelemudelit" ei ole jätkusuutlik
- Selleks, et uued mudelid eesti keeles head oleks, tuleb head eesti treeningandmed võimalikult lihtsalt kättesaadavaks teha
- Ja andmete kureerimisega pidevalt edasi tegeleda!

PIKK PERSPEKTIIV

- Tehisaru tulek on paratamatu
- See võtab üle terved tööstusharud ja tegevusalad
 - Agendid
- Oht: järjest suuremad rahavood liiguvad Eestist välja
- Hea eesti keele mudeli vajadus:
 - Et tehisaru oleks eesti keeles võrreldaval tasemel kui inglise keeles
 - Muidu me ei räägi temaga eesti keeles
 - Et saaksime tehisaru jooksutada Eestis või majasiseselt
 - Julgeolek
 - Privaatsus