

Tildes pieredze ar HPC



Ceļš līdz 30B LVM



Mārcis Pinnis

Mākslīgā intelekta
attīstības vadītājs
marcis.pinnis@tilde.lv

Par ko šodien jums stāstīšu?



Neliels ieskats tajā,
ko mēs darām un
kā mēs nonācām
pie vajadzības pēc
HPC jaudām.



Sadarbība ar RTU
HPC jomā.

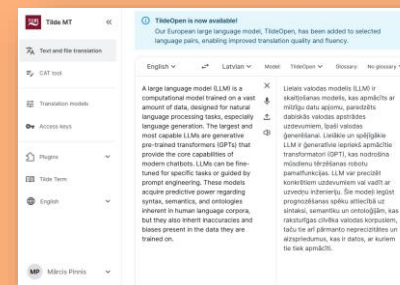


TildeOpen LLM – Uz LUMI
superdatora no nulles
apmācīts liels valodas
modelis.

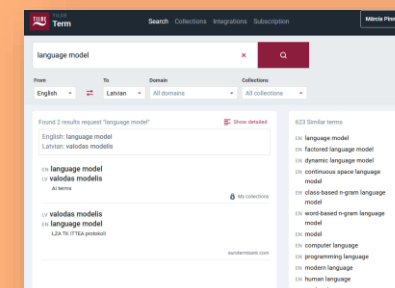
Ko mēs darām?

- **Apmācām valodas modeļus:**
 - tulkošanai;
 - runas atpazīšanai un sintēzei;
 - atbildēšanai uz jautājumiem;
 - kopsavilkumu ģenerēšanai;
 - teksta rediģēšanai;
 - utt.
- **Izstrādājam praktiskus valodas tehnoloģiju lietojumus:**
 - vispārīgai lietošanai.
 - pielāgotus konkrētas organizācijas vajadzībām.
- **Sniedzam datu gatavošanas, tulkošanas un lokalizācijas pakalpojumus.**

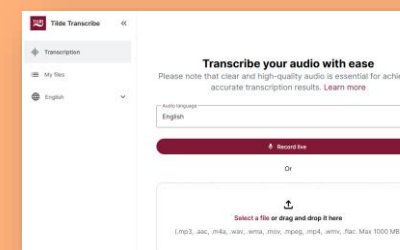
Tilde MT



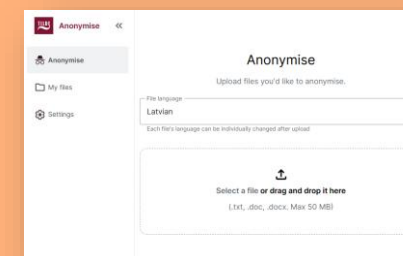
Tilde Terminology



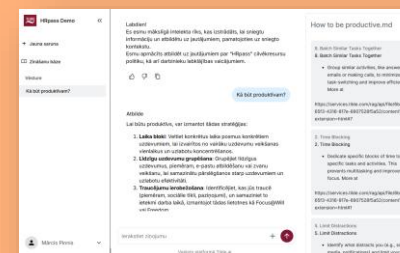
Tilde Transcribe



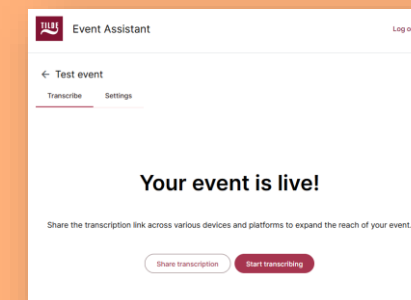
Tilde Anonymise



Tilde Enterprise Assistant



Tilde Event Assistant



**Kā mēs nonācām
pie vajadzības
pēc HPC jaudām?**

Valodas modeļu apmācība ir skaitļošanas ziņā intensīvs process

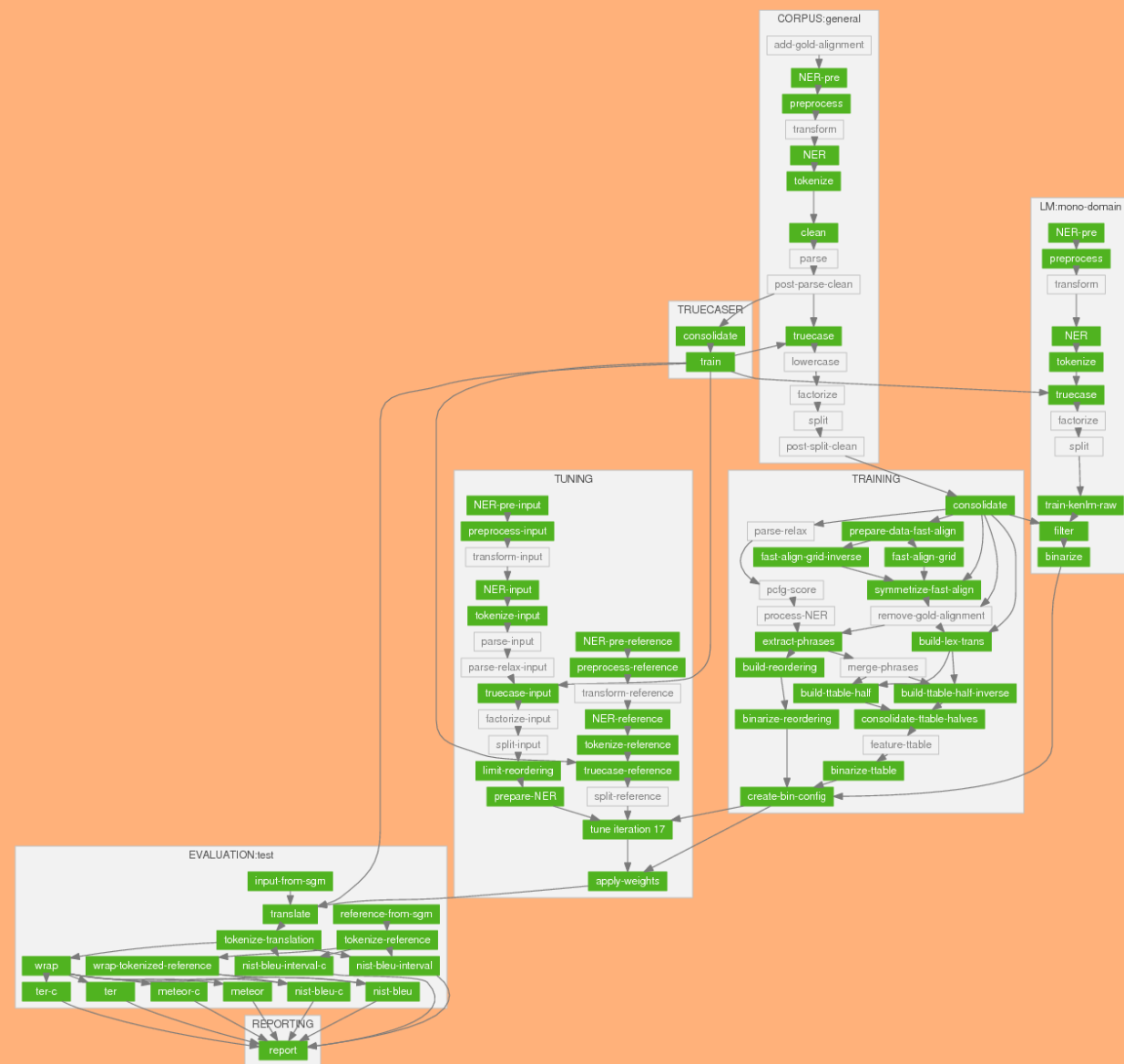
Kam ir nepieciešami lielākie
skaitļošanas resursi?

Kā piemēru apskatīsimies
Tildes mašintulkošanas
tehnoloģiju attīstību.



Tildei liela mēroga valodas modeļu apmācība aizsākās 2010. gadā

- 2010. gadā Tilde radīja mašīntulkošanas sistēmu apmācības platformu [Lets MT](#)
- Platforma (joprojām) izvietota **Amazon EC2**
- **Automātiska mezglu mērogošana**
- **Izkliedēta mašīntulkošanas sistēmu apmācība**
 - Datu apstrādei - 200-500 paralēli procesi
 - Tulkošanas modeļu apmācībai - līdz 200 paralēli procesi
 - Statistiskās mašīntulkošanas (SMT) modeļi apmācās ~5 dienās



Tehnoloģiju paaudžu maiņa prasa lielāku skaitļošanas resursu

2015	2016	2017	2018
• Pirmā NMT sistēma latviešu valodai • GRU tīkls • 2GB VRAM GPU • 5 dienas / modelis (1GPU)	• Pirmās publiskās NMT sistēmas Baltijas valodām • GRU tīkls • 12GB VRAM GPU • 7 dienas / modelis (1GPU)	• Pasaulē labākās NMT sistēmas latviešu valodai* • MLSTM tīkls • 12GB VRAM GPU • 14 dienas / modelis (2GPU)	• Pasaulē labākās NMT sistēmas igauņu valodai* • Transformer tīkls • 12GB VRAM GPU • 14-21 diena / modelis (2GPU)

* 2016.-2020. gadā mēs aktīvi piedalījāmies starptautiskās Mašīntulkošanas konferences sacensībās, kurās tika veikti neatkarīgi sistēmu vērtējumi/salīdzinājumi.

Tildes infrastruktūra neironu tīklu apmācībai

- Datus mašīntulkošanas sistēmu apmācībai apstrādājam Tilde MT platformā
 - Amazon EC2
- Modeļus apmācām lokāli
 - **Izmaksu ziņā izdevīgākais variants**
 - Esam apmācījuši tūkstošiem NMT modeļus kopš 2015. gada.
- Tildes lokālā GPU infrastruktūra šobrīd:
 - 38x NVIDIA GPU
 - Dažādas paaudzes ar 11-143GB VRAM
 - Jaunākie:
 - 4x H200 (141GB)
 - 6x L40 (48GB)

Sadarbība ar RTU HPC centru

Mašīntulkošanas konferences (WMT) sacensības

- No 2016.-2021. gadam piedalījāties starptautiskās WMT sacensībās.
- 1* gadā uz 3 mēnešiem īsā laikā bija jāapmāca apmēram 100 dažādi neironu tīklu valodas modeļi.
- 2016.-2018. un 2021. gadā mums pietika ar Amazon EC2 un lokālajiem resursiem.
- 2019. un 2020. gadā mums lokālās infrastruktūras attīstība netika līdzī – **RTU izpalīdzēja ar jaudām.**



Sadarbība ar RTU HPC centru

WMT 2019

- Uz RTU HPC apmācīti ~10 NMT modeļi (EN<->LT)
- V100 GPU (2GPU vienlaicīgi)
- Transformer Base arhitektūra (ar variācijām)
 - Sākot ar 57 miljoniem parametru

WMT 2020

- Uz RTU HPC apmācīti ~8 NMT modeļi (EN<->PL)
- V100 GPU (2GPU vienlaicīgi)
- Transformer Big arhitektūra
 - Ap 203 miljoniem parametru

Ieguvumi no sadarbības ar RTU HPC centru

- Iepazinām aktuālu akadēmisku HPC klasteri darbībā:
 - klastera programmatūru un spējas;
 - uzdevumu veidošanu un darbināšanu;
 - uzdevumu izpildes ierobežojumus.
- **Šī pieredze mums ļāva neuzkāpt uz grābekļiem LUMI.**

TildeOpen LLM

30B

TildeOpen – Latvijā radīts Eiropas lielais valodas modelis

TildeOpen iniciatīvai **tika piešķirti 2 miljoni GPU stundu** LUMI superdatorā, Somijā, ar Eiropas Komisijas “Large AI Grand Challenge” finansējumu.

Pamatmodelis

Ar fokusu uz Eiropas valodām

> 30 miljardi parametru

Vienmērīga valodu reprezentācija

Atvērts tālākai komerciālai un nekomerciālai
pielāgošanai

<https://huggingface.co/TildeAI/TildeOpen-30b>



AI-BOOST

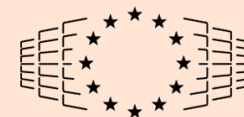


EuroHPC
Joint Undertaking

L U M I



Apmācīts, izmantojot ~2M GPUh LUMI superdatorā



EuroHPC
Joint Undertaking

LUMI jauda:

- 11 712 AMD MI250x GPU
- #1 Eiropā (apmācības laikā)
- #5 globāli (apmācības laikā)

Modeļa konfigurācija

Llama 3 līdzīga arhitektūra

- 60 slāņi, kartējuma vektora garums 6 144, FFN vektora garums 21 504, GQA uzmanības mehānisms ar 48 galvām un 8 grupām (KV galvām)
- SwiGLU aktivācijas funkcija
- RMSNorm normalizācijas metode
- YaRN metode konteksta loga paplašināšanai

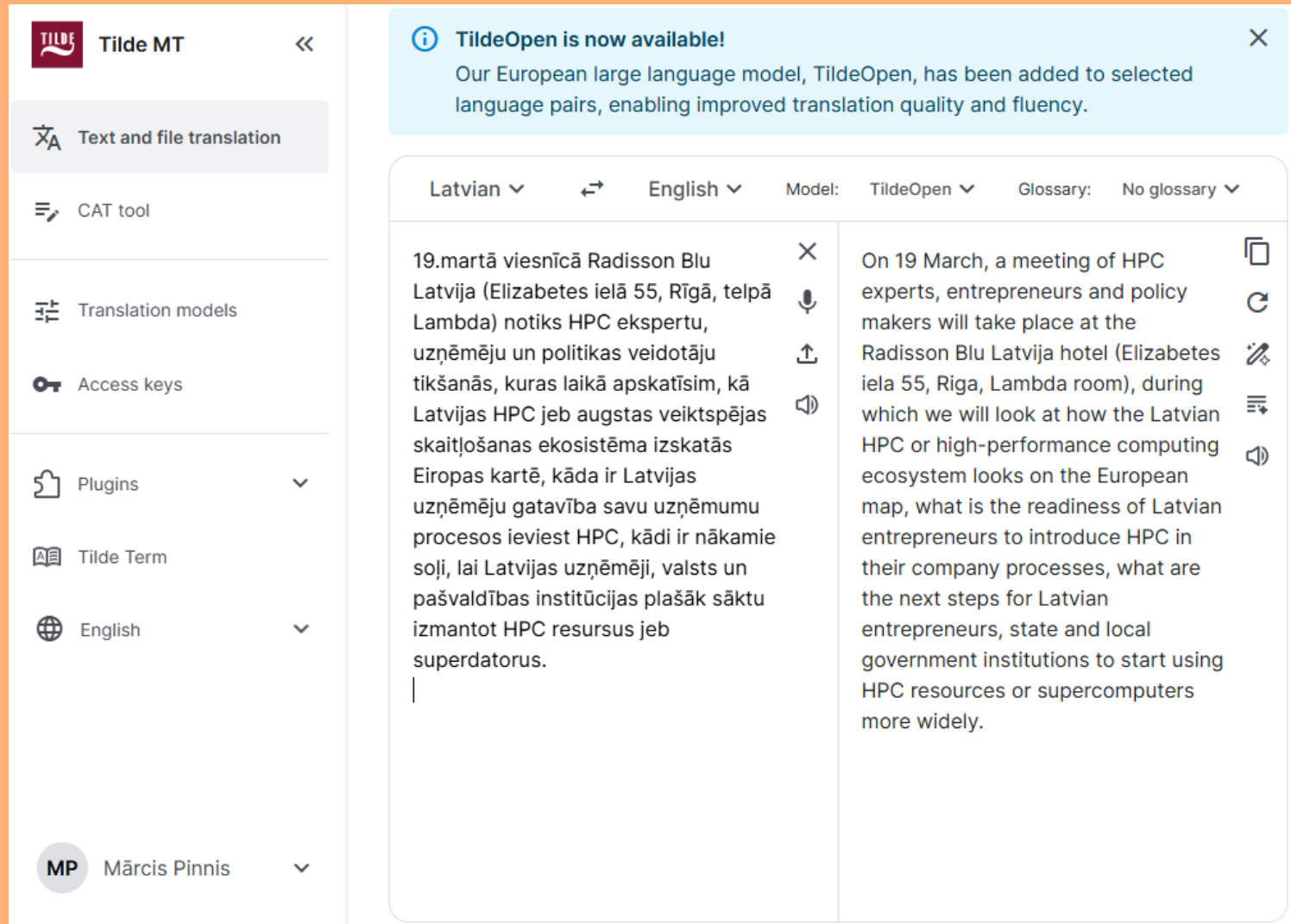
Skaitļošanas resursi

768 GPU ar 276k tekstvienību/s caurlaidspēju un 83% mērogošanas efektivitāti

- 1.55M GPUh / 2T tekstvienības pamatmodeļa apmācībai
- 0.35M GPUh / 0.18T tekstvienības konteksta loga paplašināšanai līdz 64K tekstvienībām
- 0.1M GPUh modeļa instruēšanas eksperimentiem

TildeOpen LLM instruēšana

- TildeOpen LLM 30B ir pamatmodelis.
- Pamatmodelis ir jāinstruē.
- Kas nepieciešams instruēšanai?
 - Pieeja HPC klasterim.
 - 1K-10K GPUh (atkarībā no uzdevuma un datu apjoma).
- Kad modelis ir instruēts, to vēl var uzlabot, apmācot adapterus.
 - Labāks atbalsts jomas datiem.
 - Vajadzīgs 1x H200 GPU vai 2x L40 GPU.
 - Nevajag pieeju HPC klasterim.
- Pirmais lietojums, kuram publiski esam instruējuši TildeOpen ir tulkošana
 - Instruēto modeli var lietot Tilde MT platformā: <https://translate.tilde.ai>



Paldies!

Laiks jautājumiem!



Mārcis Pinnis

Mākslīgā intelekta
attīstības vadītājs
marcis.pinnis@tilde.lv