

Lietuvių kalbos ištekliai: žvelgiant iš CLARIN-LT perspektyvos

Andrius Utkas

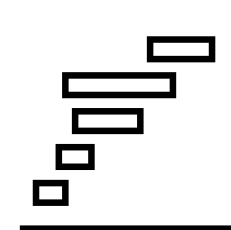
Vytauto Didžiojo universitetas

Skaitmeninių išteklių ir tarpdisciplininių tyrimų institutas



Pranešimo planas

1. Kontekstas
2. Naujos kartos Lietuva projektai
 1. Projektų trukmė
 2. Projektai ir kuriami ištekliai
3. „Naujos kartos Lietuva“ projektai ir CLARIN-LT
4. Projektas: Bendrojo lietuvių kalbos tekstyno ir vektorizuotų lietuvių kalbos modelių sukūrimas



Kontekstas

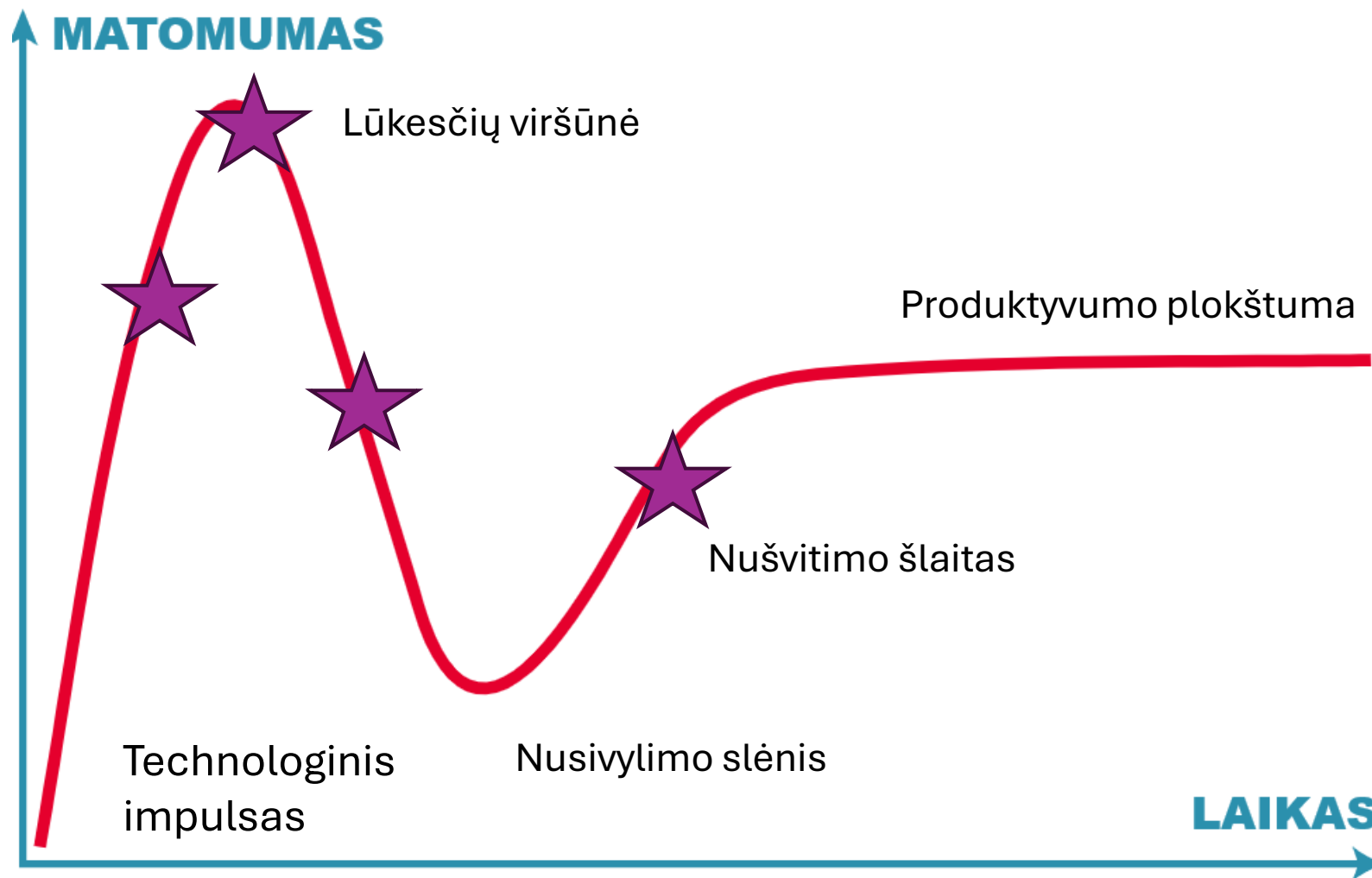
Technologinis impulsas

ChatGPT



- ChatGPT atvertas **2022 m. lapkričio 30 d.**
- Pirmas pokalbių robotas, kurio kalbos modelis buvo apmokintas su 300 mlrd. žodžių tekstynu ir su 175 mlrd. parametru.
- Iki 2023 m. sausio mėn. tuo metu ji tapo sparčiausiai populiarėjančia programine įranga istorijoje, sulaukė daugiau nei 100 mln. naudotojų ir prisidėjo prie to, kad "OpenAI" vertė išaugo iki 29 mlrd. JAV dolerių.
- Iš karto po to *Google*, *Baidu* ir *Meta* drastiškai padidino investicijas į analogiškus produktus BARD, Ernie Bot, LLaMA...

Gartner išpūstų lūkesčių ciklas (*Gartner hype cycle*)

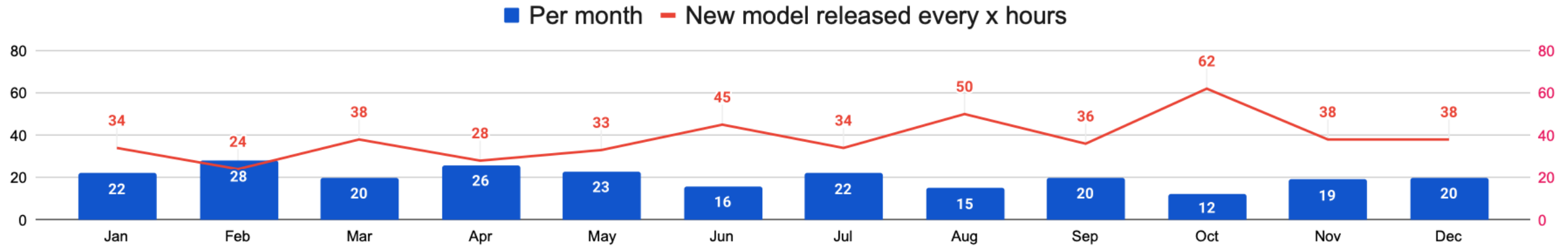


AI Hype
„*DI burbulas*“

LLMs RELEASED PER MONTH (2024)

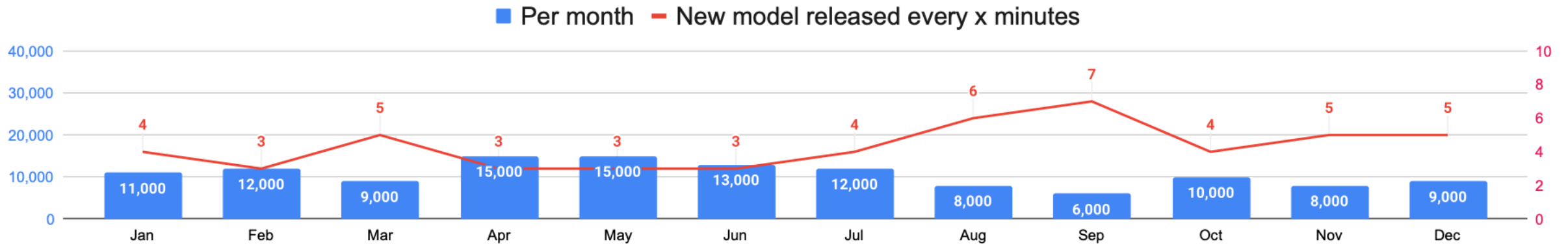
New major models released per month/x hours

LifeArchitect.ai/models (data from LifeArchitect.ai/models-table)

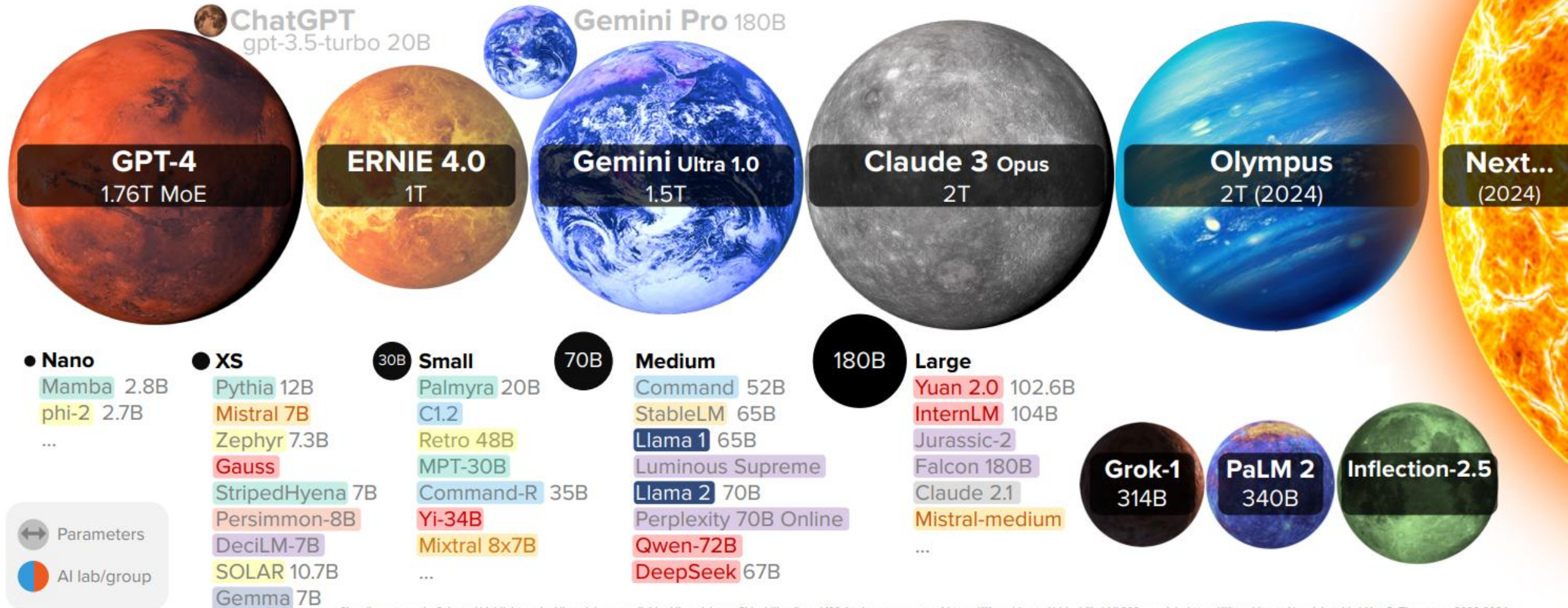


New derivative models released per month/x minutes

LifeArchitect.ai/models (data from Hugging Face)



LARGE LANGUAGE MODEL HIGHLIGHTS (MAR/2024)



Sizes linear to scale. Selected highlights only. All models are available. All models are Chinchilla-aligned (20:1 tokens:parameters) <https://lilearchitect.ai/chinchilla/> All 300+ models: <https://lilearchitect.ai/models-table/> Alan D. Thompson. 2023-2024.

Naujos kartos Lietuva projektai



Finansuoja
Europos Sąjunga
NextGenerationEU



NAUJOS KARTOS
LIETUVA

„Naujos kartos Lietuva“ projektai

- Finansuojami 16 projektų, investuojama apie 26 mln. eurų.
- Projektai finansuojami Ekonomikos gaivinimo ir atsparumo didinimo plano „Naujos kartos Lietuva“ lėšomis (RRF).
- 11 projektų vykdo Valstybės skaitmeninių sprendimų agentūra (VSSA).
- 5 vykdo mokslo institucijos.



Finansuoja
Europos Sąjunga
NextGenerationEU



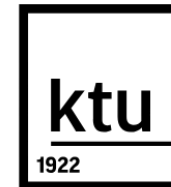
NAUJOS KARTOS
LIETUVA



VALSTYBĖS SKAITMENINIŲ SPRENDIMŲ AGENTŪRA



Vilniaus
universitetas



kauno
technologijos
universitetas



VYTAUTO
DIDŽIOJO
UNIVERSITETAS
MCMXXII

„Naujos kartos Lietuva“ projektai (2)

- Projektai skirti mašininio mokymosi ir dirbtinio intelekto sistemoms tobulinti.
- Visi projektuose sukurti ištekliai turi būti atviri, nemokami ir prieinami tiek mokslui, tiek verslui, tiek visai visuomenei.
- Iš esmės. ištekliai turi atitikti FAIR principus, t. y. būti randami (**F**indable), atvirai prieinami (**A**ccessible), interoperabilūs (**I**nteroperable) ir perpanaudojami (**R**eusable).
- KUR?

Metaduomenys



Duomenys

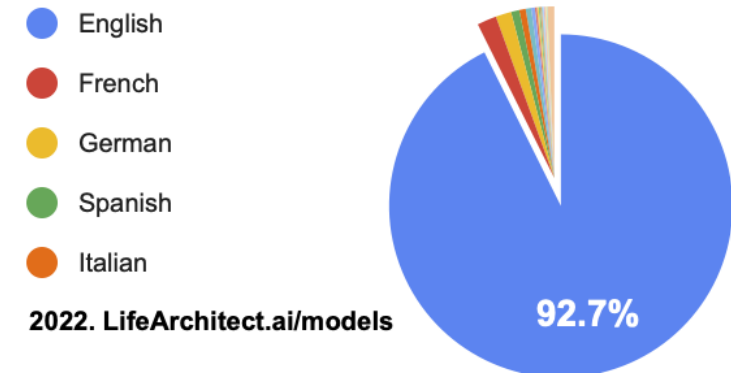


AR KITUR...

Kam to reikia

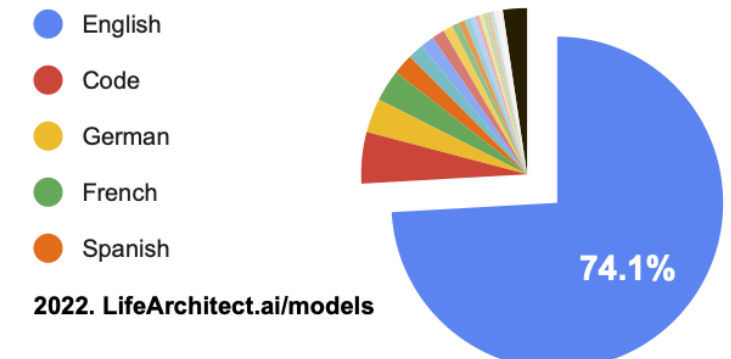
- Investavimas į išteklius remiasi šiomis svarbiausiomis prielaidomis:
 - Didieji komerciniai modeliai, visų pirma orientuojasi į anglų kalbą, todėl jei norime gerinti lietuvių kalbos palaikymą turime patys imtis iniciatyvos. Be to, jų naudojimas kartais būna brangus ir dažnai negalime valdyti jų kokybės;
 - Dabartinėms generatyvinio dirbtinio intelekto technologijoms reikalingas milžiniškas kiekis duomenų, todėl norint pagerinti lietuvių kalbos palaikymą turime juos kaupti;
 - Technologijos ir sistemos greitai sensta, jų palaikymas brangiai kainuoja, tuo tarpu duomenų senėjimas yra lėtesnis ir duomenys gali būti perpanaudoti daugelio vystytojų.

GPT-3 - 90 languages



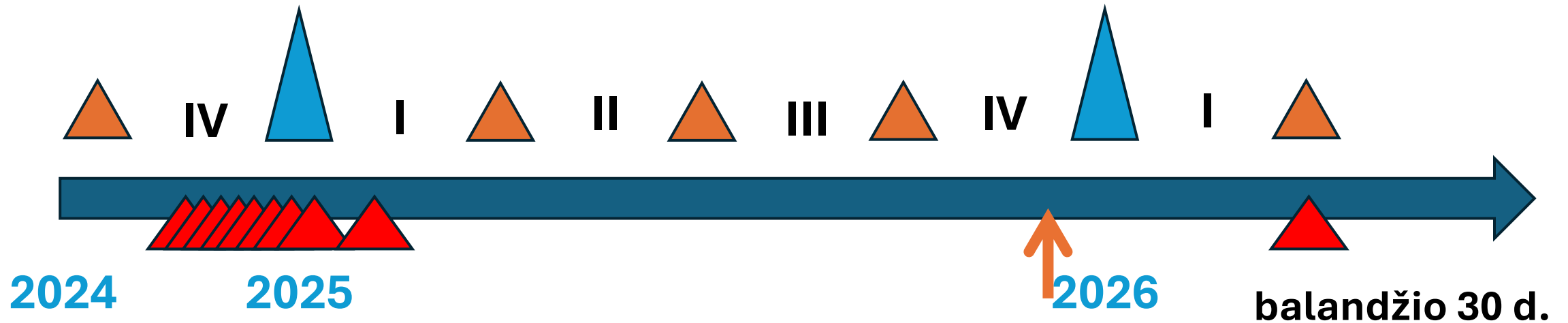
2022. LifeArchitect.ai/models

PaLM - 122 languages



2022. LifeArchitect.ai/models

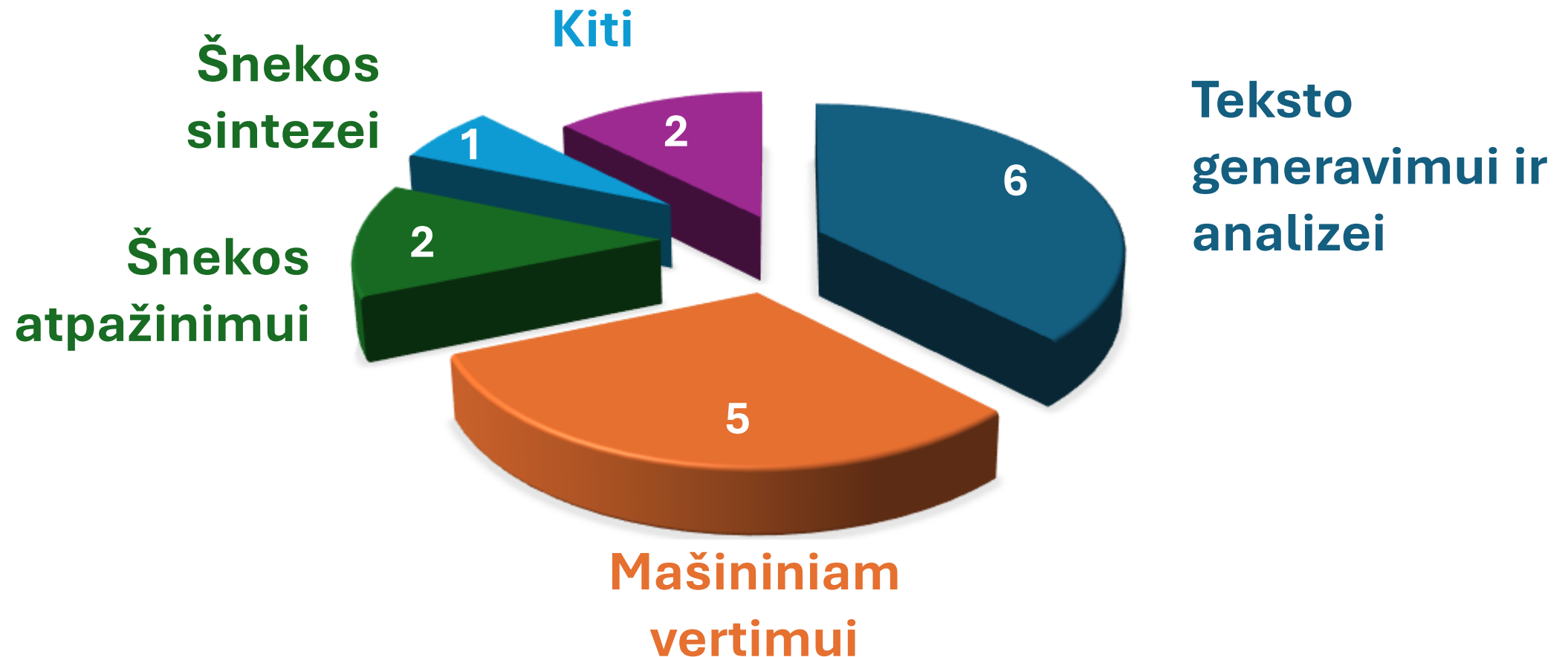
„Naujos kartos Lietuva“ projektų trukmė



- Projektų trukmė 15-18 mėnesių
- Beveik 2/3 projektų vykdymo trukmės jau praėjo

„Naujos kartos Lietuva“ projektai

- 16 vykdomų projektų skirti skirtingiems tikslams įgyvendinti:



Teksto ištekliai skirti teksto generavimui, analizei ir kitiems uždaviniams

1. Bendrojo lietuvių kalbos tekstyno ir vektorizuotų modelių sukūrimas	VSSA	Tekstynas (3,5 mlrd.), 2 kalbos modeliai
2. Mišrių lietuvių kalbos tekstų santraukų/abstraktų tekstynų sukūrimas	VDU, VU	Spec. tekstynai (4 tekstynai)
3. Klausimų-atsakymų porų tekstyno sukūrimas	VSSA	Spec. tekstynas (15 mln.)
4. Nuasmeninimo tekstyno sukūrimas (NUS)	KTU	Spec. tekstynas
5. Morfologiškai ir sintaksiškai anotėtų tekstynų modeliai apmokymui (auksiniai standartai)	VDU	Spec. tekstynai (10 mln. žodžių)
6. Melagingos informacijos automatinio identifikavimo tekstyno sukūrimas	VSSA	Spec. tekstynas (5 tūkst.), k. modelis

Teksto ištekliai skirti mašininiam vertimui

1. Daugiakalbių ir vienkablių tekstynų sukūrimas (ukrainiečių, norvegų, švedų, danų, ispanų)	VSSA	Lygiagretūs ir vienkalbiai tekstynai
2. Vienkalbių ir daugiakalbių tekstynų atnaujinimas (EN, DE, FR, PL, LT)	VSSA	Lygiagretūs ir vienkalbiai tekstynai
3. Sintetinių lygiagrečių tekstynų sukūrimas (LT-EN, LT-FR, LT-DE)	VSSA	Lygiagretūs tekstynai
4. Medicinos vienkablio ir lygiagrečiųjų tekstynų sukūrimas	VSSA	Medicinos vienkabliai ir lygiagretūs tekstynai
5. Gynybos ir saugumo vienkablio ir lygiagrečiųjų tekstynų sukūrimas (EN-LT)	VSSA	Gynybos vienkabliai (10 mln.) ir lygiagretus tekstynas

Šnekos ištekliai - šnekos atpažinimas

1. Lietuvių kalbos garsyno sukūrimas (šnekos atpažinimo tikslams)	VU, VDU, LKI	Garsynas (10 tūkst. val.)
2. Lietuvių kalbos medicinos garsyno sukūrimas	VSSA	Medicinos garsynas (400 val.)

Šnekos ištekliai - šnekos sintezė

1. Lietuvių kalbos garsyno (šnekos sintezės tikslams neuroniniams balsams generuoti) sukūrimas	VSSA	Garsynas (400 val.)
--	------	---------------------

Kiti ištekliai

1. Kalbos paveldo transformacija ir lietuvių kalbos erdvinių duomenų išteklių sukūrimas	LKI	Skaitmenin., įv. ištekliai ir geoduomenys
2. Žmogaus fenotipo ontologijos lietuvių kalba sukūrimas	VSSA	Ontologija (13 tūkst. sąv.)

„Naujos kartos Lietuva“ projektai
ir CLARIN-LT

„Naujos kartos Lietuva“ ir CLARIN-LT

- ✓ Dėka „Naujos kartos Lietuva“ projektų tam tikra dalis išteklių pateks į CLARIN-LT saugyklą.
- ✓ (Tiesa, nėra tiksliai žinoma, kurie ištekliai ir kiek pateks į CLARIN saugyklą).
- ✓ Bet ruoštis reikia ...



Finansuoja
Europos Sąjunga
NextGenerationEU



NAUJOS KARTOS
LIETUVA



VSSA ir CLARIN bendradarbiavimo sutartis

VDU pasirašė bendradarbiavimo sutartį su Valstybės skaitmeninių sprendimų agentūra

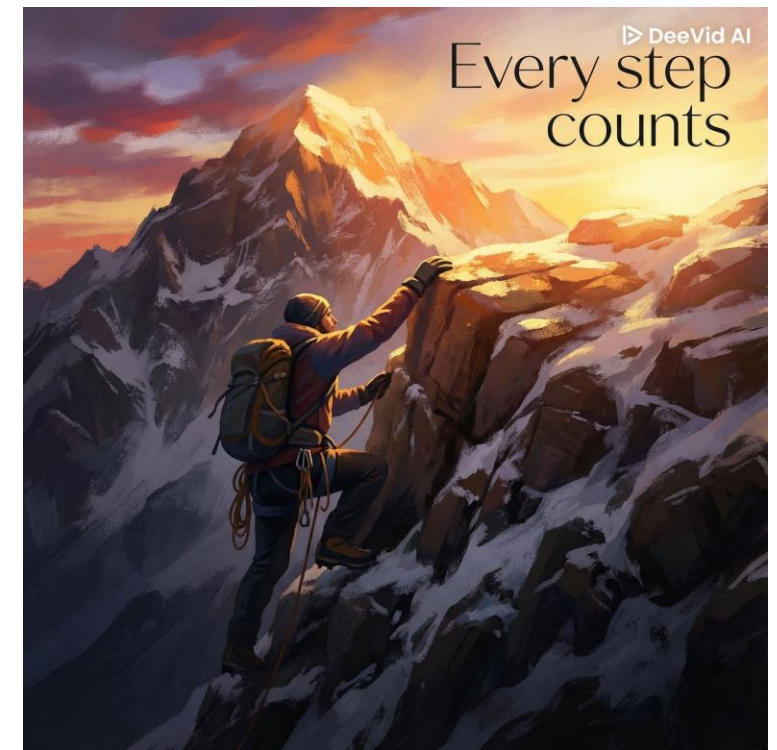
2025 Balandžio 1



VSSA įsipareigojo teikti kokybiškus lietuvių kalbos išteklius, skirtus dirbtinio intelekto (DI) poreikiams, taip pat naudotinus moksliniuose tyrimuose. VDU įsipareigojo, kad šie lietuvių kalbos ištekliai bus atvirai prieinami visuomenei VDU kuruojamoje CLARIN-LT atvirųjų kalbos duomenų saugykloje.

Artimiausi CLARIN-LT planai

- ✓ Atnaujinti DSpace duomenų bazę (iš 5.4 versijos į 7);
- ✓ Išplėsti CLARIN saugyklos diskinę talpą;
- ✓ Be to, laukia ir didesni iššūkiai:
 - ✓ CLARIN-LT duomenų saugyklos sertifikavimas (t.y. gauti CoreTrustSeal) ir tapimas B tipo centru.
 - ✓ Įsigijimas centrų-dvynio (twin centre), t. y. vietą kur galėsime saugoti išteklių atsargines kopijas.
 - ✓ Duomenis renkančių robotų nugalėjimas.



Projektas:
Bendrojo lietuvių kalbos tekstyno ir
vektorizuotų lietuvių kalbos modelių
sukūrimas

Bendrojo lietuvių kalbos tekstyno ir vektorizuotų lietuvių kalbos modelių sukūrimas

- **Projekto trukmė:**
- 2024 m. gruodis – 2026 m. balandis (17 mėnesių).
- **Projekto vykdytojas:** [Valstybės skaitmeninių sprendimų agentūra \(VSSA\)](#)
- **Projekto vadovas:** Arminas Rakauskas
 - **Pirkimo vertė:** 5 100 000 eurų
 - **Tiekėjų grupė:** VDU, UAB Neurotechnology, UAB Tilde Lietuva, MB Krilas
 - **Tiekėjų grupės vadovas:** Andrius Utkas



NAUJOS KARTOS
LIETUVA



Finansuoja
Europos Sąjunga
NextGenerationEU



MB KRILAS

Projekto uždaviniai ir rezultatai

- Sukaupti Bendrąjį lietuvių kalbos tekstyną (BLKT) – 3,5 mlrd.
- Sukurti du lietuvių kalbos vektorizuotus modelius:
 - Mažąjį lietuvių kalbos vektorizuotas modelis – BERT architektūra
 - <https://data.gov.lt/datasets/3923/>
 - <https://huggingface.co/VSSA-SDSA>
 - Didįjį lietuvių kalbos vektorizuotą modelį – GPT architektūra
- Sukurti programinės įrangos sprendimą, kuris leistų generuoti lietuvišką tekstą.



STATE
DIGITAL
AGENCY

State Digital Solutions Agency (LT)

Government

 <https://vssa.lrv.lt/en/>  [state-digital-solutions-agency](https://www.linkedin.com/company/state-digital-solutions-agency/)

<https://huggingface.co/VSSA-SDSA>

AI & ML interests

None defined yet.

Recent Activity

-  MilaSong updated a model 9 days ago
[VSSA-SDSA/LT-NER-modernBERT](#)
-  DariusAm published a model 18 days ago
[VSSA-SDSA/LT-MLKM-modernBERT](#)
-  DariusAm published a model 18 days ago
[VSSA-SDSA/LT-NER-modernBERT](#)

[View all activity](#)

Models 2



VSSA-SDSA/LT-NER-modernBERT



Token Classification • 0.2B • Updated 9 days ago • 64



VSSA-SDSA/LT-MLKM-modernBERT



Fill-Mask • 0.2B • Updated 21 days ago • 35

Datasets 0

None public yet

Duomenys tapo vertybe

- Kaupdami duomenis turime užtikrinti etišką duomenų rinkimą.
- Negalime tiesiog surinkti internete viešai pateikiamus duomenis – turime susiderinti su duomenų valdytojais.
- Todėl didelė dalis tekstyno duomenų yra licencijuoti.
- Deja, dėl autorinių teisių ribojimo knygų ir grožinės literatūros tekstų dalis tekstyne bus maža.

Duomenų tiekėjai



BLKT (54,6%) - šaltiniai

Nr.	Šaltiniai	Tipas	Sritis	Žodžių skaičius	%
1	delfi.lt	zin	naujienų portalai	702 957 035	37,445%
2	lrytas.lt	zin	naujienų portalai	406 814 522	21,670%
3	lrt.lt	zin	naujienų portalai	234 786 835	12,506%
4	ve.lt	zin	naujienų portalai	157 808 789	8,406%
5	eur-lex.europa.eu	dok	ES dokumentai	105 893 846	5,641%
6	GOV_LT	neg	Interneto tekstai	89 489 638	4,767%
7	lrs.lt.stenogramos	sak	stenogramos	79 213 991	4,220%
8	lrs.lt	neg	kiti	39 671 302	2,113%
9	Vikipedija	neg	moksiniai tekstai	34 742 507	1,851%
10	VDU CRIS	neg	moksiniai tekstai	7 399 476	0,394%
11	urm.lt	neg	Interneto tekstai	5 972 814	0,318%
12	Švenčionių kraštas	zin	naujienų portalai	5 035 426	0,268%
13	lrkt.lt	dok	LT dokumentai	4 163 254	0,222%
14	e-tar.lt	dok	LT dokumentai	2 291 061	0,122%
15	Lituanistika_DB	neg	moksiniai tekstai	929 106	0,049%
16	Gargždapis	neg	Interneto tekstai	98 983	0,005%
17	Humanitarų meka	neg	Interneto tekstai	50 288	0,003%
				1 877 318 873	100,000%

BLKT (54,6%) – teksto tipai

Teksto tipai	Žodžiai	%
gro	0	0%
dok	112,348,161	6.0%
neg	178,354,114	9.5%
sak	79,213,991	4.2%
zin	1,507,402,607	80.3%
Viso:	1,877,318,873	100.00%

Baigiamosios žinutės

- Jeigu viskas pavyks, tai 2026 m. metais su
 - atnaujinta duomenų baze,
 - „Naujos kartos Lietuva“ ištekliais,
 - kokybės sertifikatu bei B centro statusu
- ... CLARIN-LT turėtų įgyti naują kvėpavimą.
- Bet ar tai atsitiks priklauso nuo mūsų darbo ir sėkmės.



Ačīū už dėmesį!