CLARIN-LT: Lithuanian Language Resources for the Age of AI

Andrius Utka

Vytautas Magnus University

Institute of Digital Resources and Intedisciplinary Research (SITTI)

VYTAUTO
DIDŽIOJO
UNIVERSITETAS
MCMXXII

SITTI
SKAITMENINIŲ IŠTEKLIŲ IR
TARPDISCIPLININIŲ TYRIMŲ INSTITUTAS

CLARIN-LT

Plan of the presentation

1. CLARIN-LT is 11 years old
2. Context. The Age of AI
3. Lithuanian RRF language projects
4. The concept of “large corpus”
5. Project: Lithuanian language corpus and Lithuanian language models
6. A final note

CLARIN-LT is 11 years old

ALKSN

R. Petrauskaitė

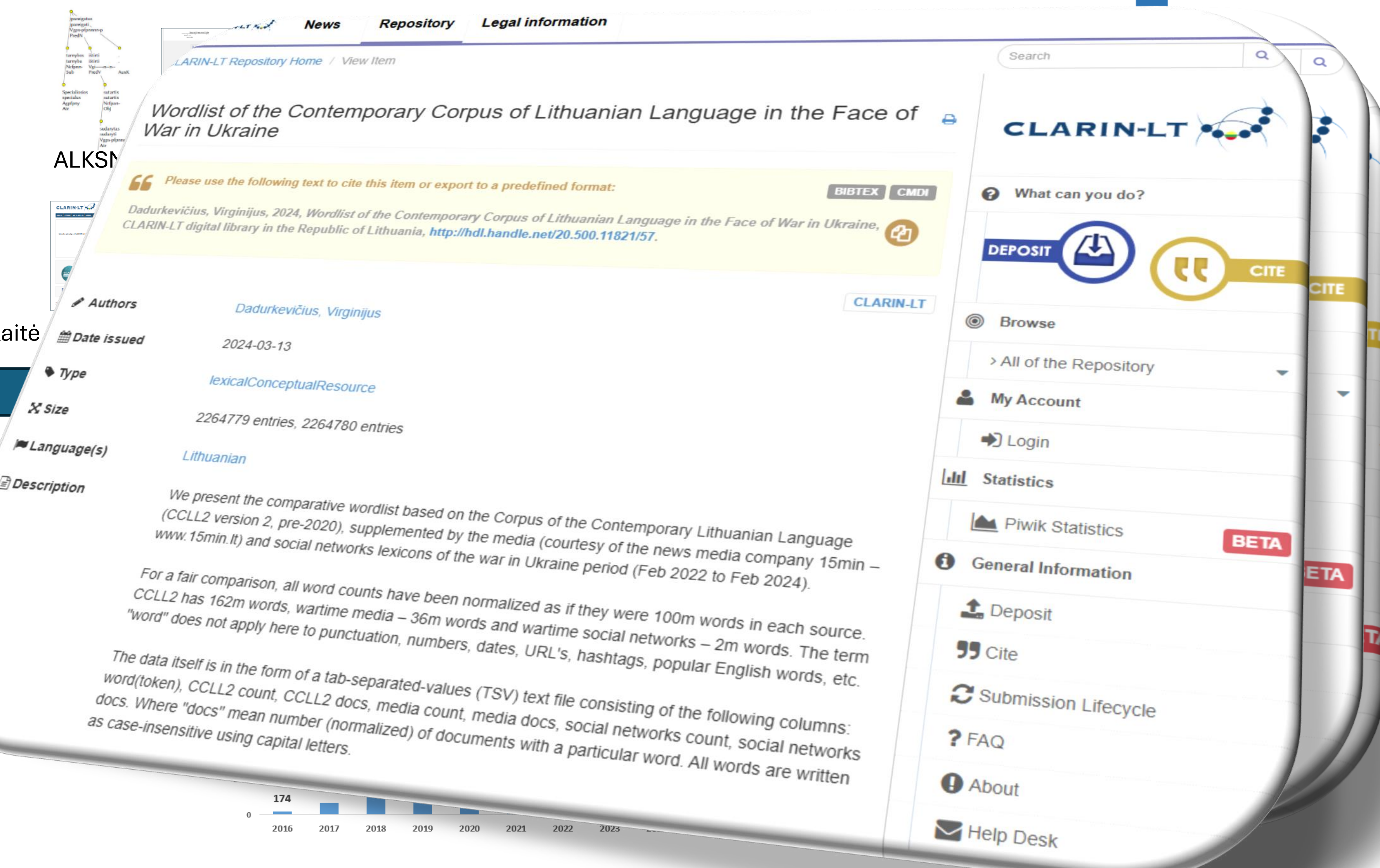
201

Oc

Li

ha

CL



Wordlist of the Contemporary Corpus of Lithuanian Language in the Face of War in Ukraine

Please use the following text to cite this item or export to a predefined format: BIBTEX CMDI

Dadurkevičius, Virginijus, 2024, Wordlist of the Contemporary Corpus of Lithuanian Language in the Face of War in Ukraine, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/57.

Authors	Dadurkevičius, Virginijus
Date issued	2024-03-13
Type	lexicalConceptualResource
Size	2264779 entries, 2264780 entries
Language(s)	Lithuanian

CLARIN-LT

Description

We present the comparative wordlist based on the Corpus of the Contemporary Lithuanian Language (CCLL2 version 2, pre-2020), supplemented by the media (courtesy of the news media company 15min – www.15min.lt) and social networks lexicons of the war in Ukraine period (Feb 2022 to Feb 2024).

For a fair comparison, all word counts have been normalized as if they were 100m words in each source. CCLL2 has 162m words, wartime media – 36m words and wartime social networks – 2m words. The term "word" does not apply here to punctuation, numbers, dates, URL's, hashtags, popular English words, etc.

The data itself is in the form of a tab-separated-values (TSV) text file consisting of the following columns: word(token), CCLL2 count, CCLL2 docs, media count, media docs, social networks count, social networks docs. Where "docs" mean number (normalized) of documents with a particular word. All words are written as case-insensitive using capital letters.

Immediate CLARIN-LT plans

- Update the DSpace database (from version 5.4 to 7);

✓Expand the disk capacity of the CLARIN repository to receive results of RRF projects;

- Besides some bigger challenges
  - o Certification of the CLARIN-LT data repository (i.e., obtaining CoreTrustSeal) and becoming a B Type center.
  - o Finding a twin center, i.e. a place where we can store backup copies of CLARIN-LT resources.
  - o Defeating data-collecting robots.

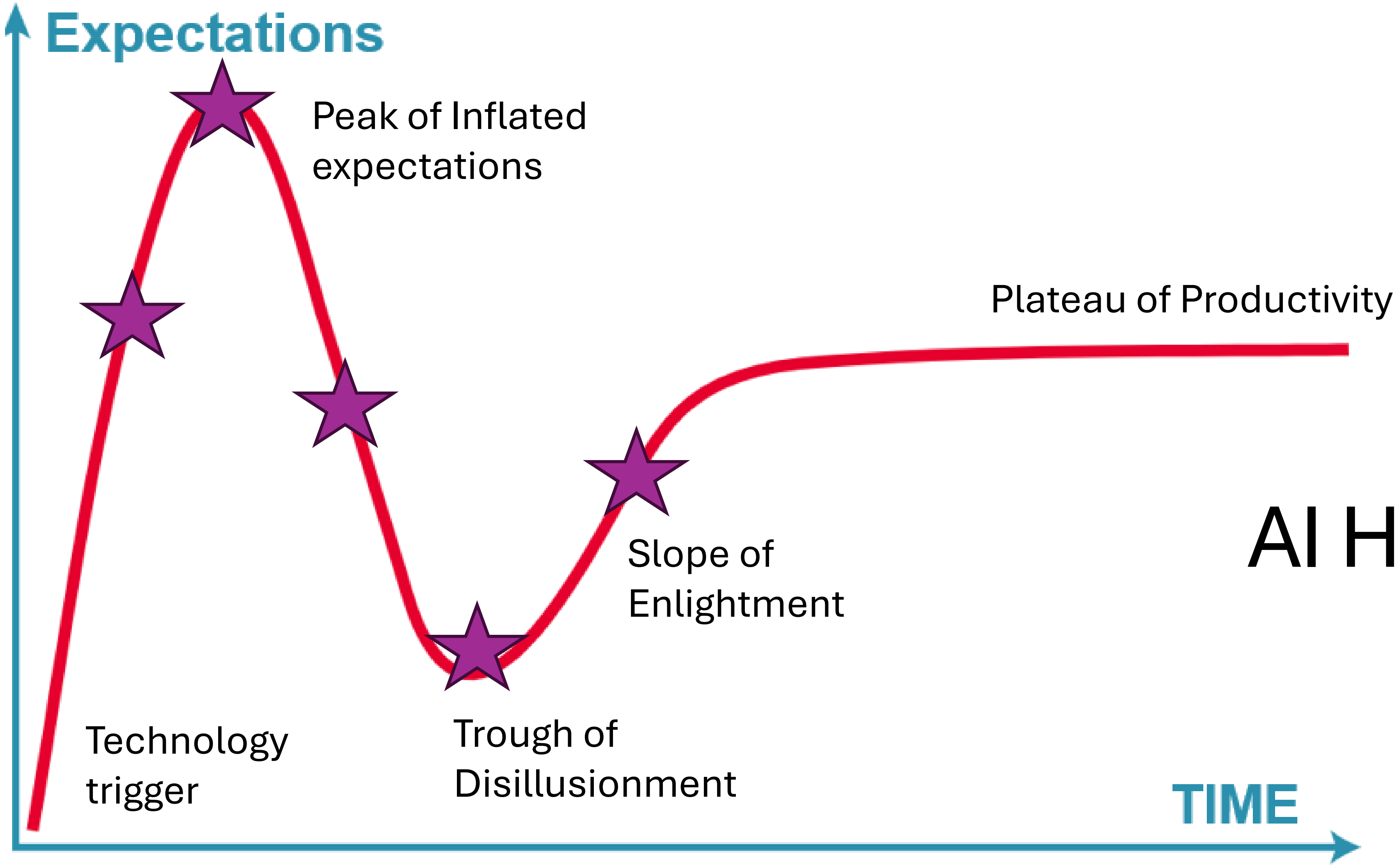
Context

Do you know what important event happened for Language and AI approximately 3 years and 4 months ago?

ChatGPT

- ChatGPT launched on **November 30, 2022**.
- It was the first chatbot based on such a large language model (175B) that was opened to the public .
- By January 2023, it had become the fastest-growing software in history, attracting more than 100 million users and contributing to OpenAI‘s value rising to $29 billion.
- Immediately afterwards, *Google*, *Baidu*, and *Meta* dramatically increased their investments in similar products such as BARD, Ernie Bot, LLaMA, and others.

Gartner hype cycle



Expectations

Peak of Inflated expectations

Plateau of Productivity

Slope of Enlightment

Technology trigger

Trough of Disillusionment

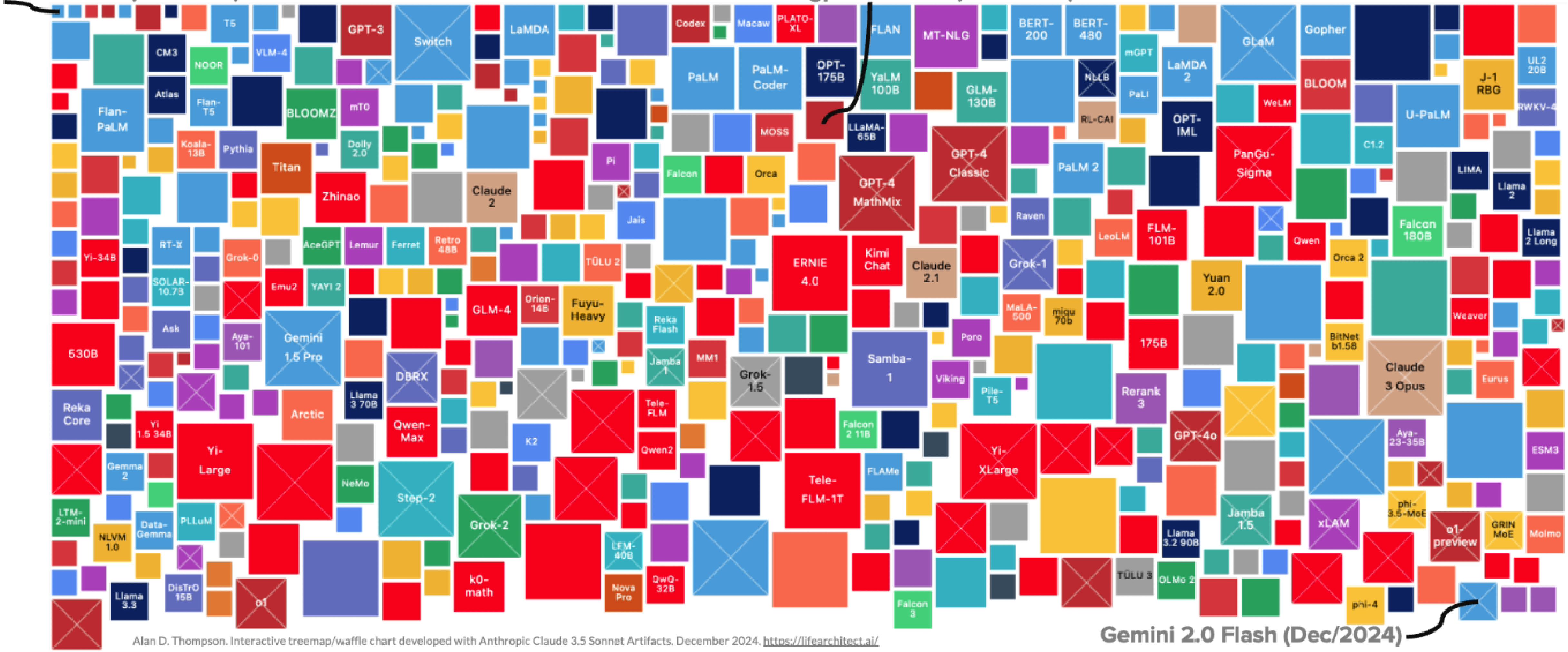
TIME

AI Hype

LARGE LANGUAGE MODEL HIGHLIGHTS 2017–2024

Transformer (Jun/2017)

ChatGPT gpt-3.5-turbo (Nov/2022)



Alan D. Thompson. Interactive treemap/waffle chart developed with Anthropic Claude 3.5 Sonnet Artifacts. December 2024. https://lifearchitect.ai/

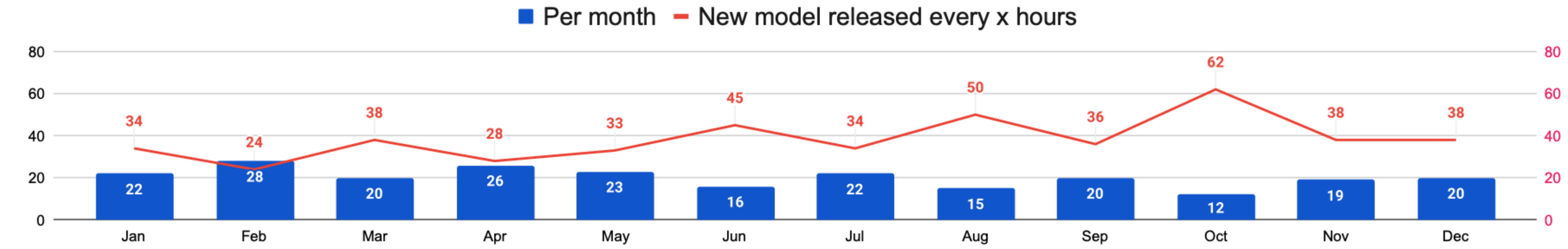
Gemini 2.0 Flash (Dec/2024)

LifeArchitect.ai/models-table

LLMs RELEASED PER MONTH (2024)

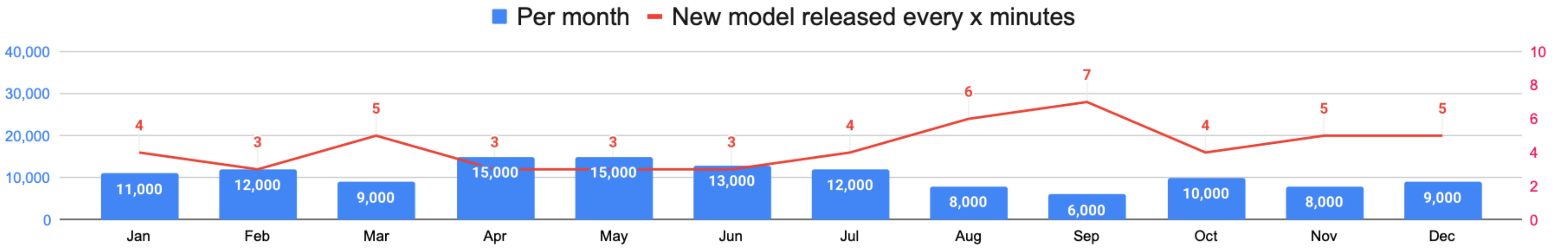
New major models released per month/x hours

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



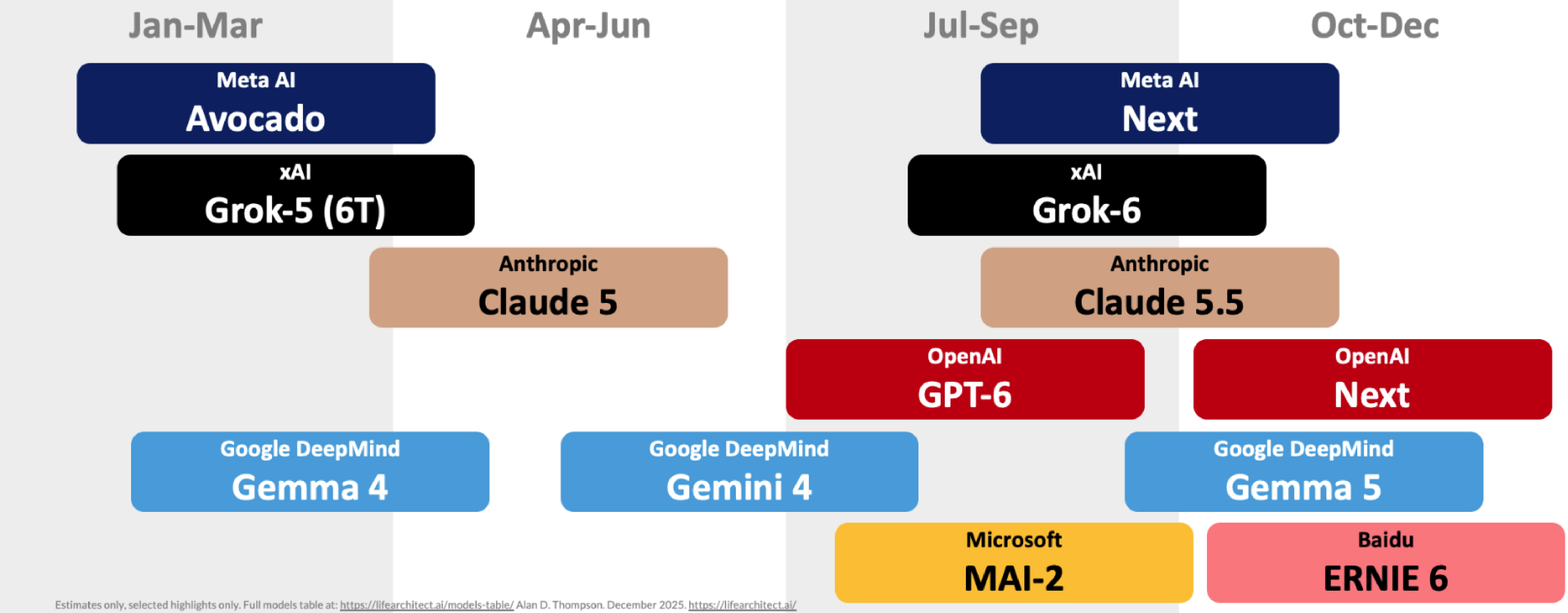
New derivative models released per month/x minutes

LifeArchitect.ai/models (data from Hugging Face)



LLMs: NEXT PUBLIC RELEASES IN 2026

ESTIMATES AS OF DECEMBER 2025



Jan-Mar

Apr-Jun

Jul-Sep

Oct-Dec

Meta AI
Avocado

xAI
Grok-5 (6T)

Anthropic
Claude 5

Google DeepMind
Gemma 4

Google DeepMind
Gemini 4

Meta AI
Next

xAI
Grok-6

Anthropic
Claude 5.5

OpenAI
GPT-6

OpenAI
Next

Google DeepMind
Gemma 5

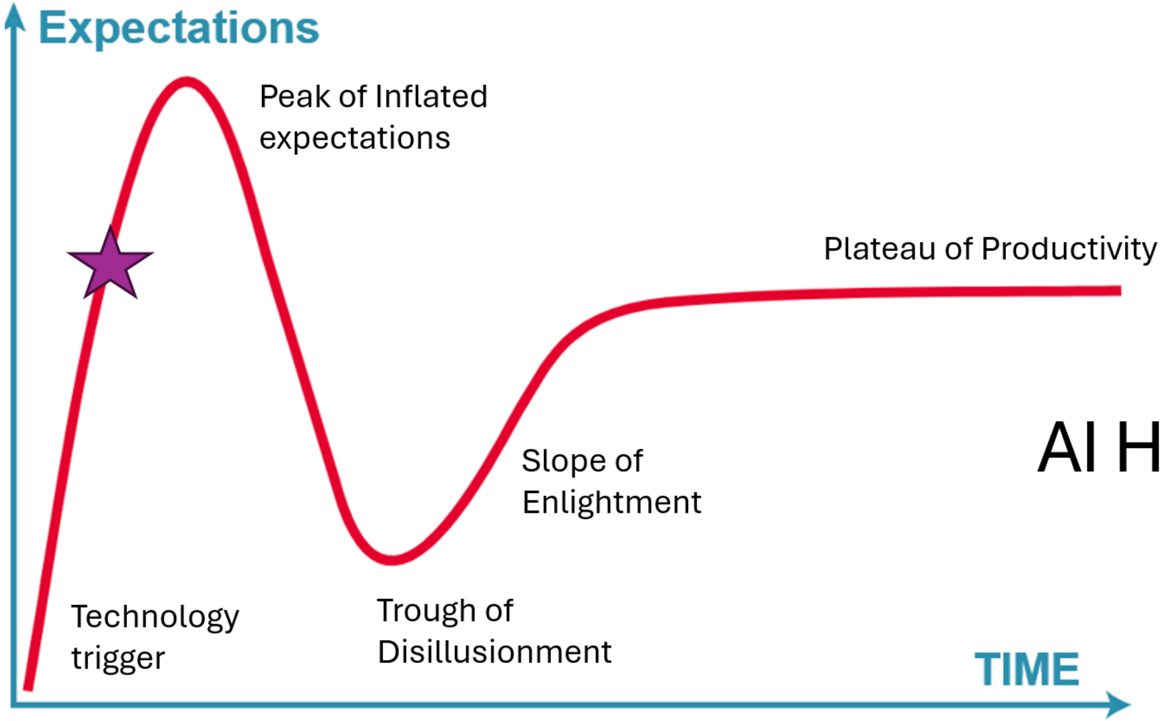
Microsoft
MAI-2

Baidu
ERNIE 6

Estimates only, selected highlights only. Full models table at: https://lifearchitect.ai/models-table/ Alan D. Thompson. December 2025. https://lifearchitect.ai/

We are still racing …

- We have‘t reached the peak of inflated expectations.
- Now we have started not only about Artificial General Intelligence (AGI), but also about Artificial Super Intelligence (ASI).



AI Hype

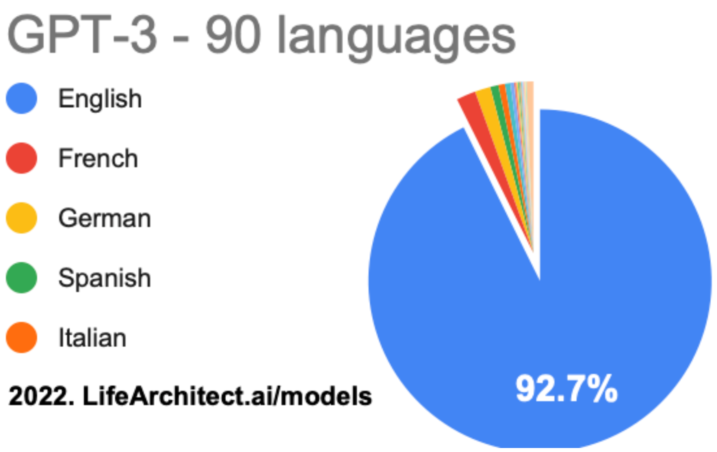
Lithuanian RRF projects

Funded by
the European Union
NextGenerationEU

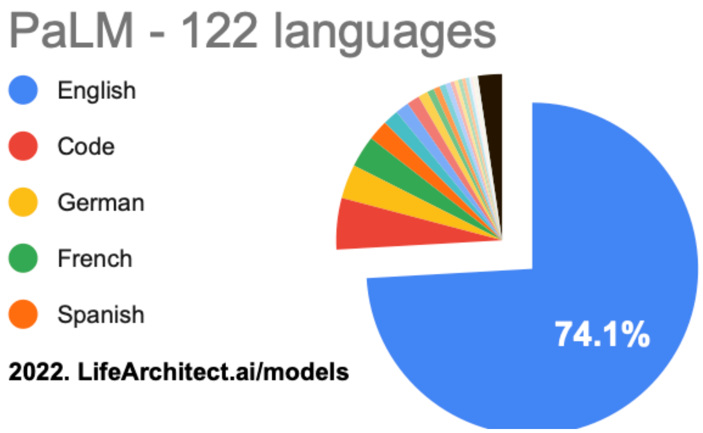
NEW GENERATION
LITHUANIA

Why do we need to invest into our own resources?

- Large commercial models are primarily focused on English, so if we want to improve support of our languages, we must take the initiative ourselves.
- Current generative AI technologies require enormous amounts of quality data, so we need to help developers (even the largest ones) to collect our national data, in order to improve support for our languages.
- Data ages more slowly than technologies or systems.


GPT-3 - 90 languages

2022. LifeArchitect.ai/models


PaLM - 122 languages

2022. LifeArchitect.ai/models

RRF „New Generation Lithuania“ projects

- 16 projects are being funded, with total investments of around €26 million.
- The projects are being funded by the Economic Recovery and Resilience Facility fund "New Generation Lithuania" (RRF).
- 11 projects are being implemented by the State Digital Solutions Agency (VSSA).
- 5 are being implemented by scientific institutions.

Funded by the European Union
NextGenerationEU

NEW GENERATION LITHUANIA

STATE DIGITAL SOLUTIONS AGENCY

LIETUVIŲ KALBOS INSTITUTAS

VILNIAUS UNIVERSITETAS · 1579 · UNIVERSITAS VILNENSIS
Vilniaus universitetas

ktu 1922
kauno technologijos universitetas

VYTAUTI MAGNI UNIVERSITAS CAUNAE
VYTAUTO DIDŽIOJO UNIVERSITETAS
MCMXXII

RRF „New Generation Lithuania“ projects

- The projects are aimed at collecting language resources for improving machine learning and AI systems.
- All compiled resources in the projects must be open, free of charge, and accessible to science, business, and the general public.
- Essentially, resources must comply with the FAIR principles, i.e., be Findable, Accessible, Interoperable, and Reusable.
- WHERE?

Metadata

LIETUVOS ATVIRŲ DUOMENŲ PORTALAS

https://data.gov.lt/

Data

CLARIN-LT

Hugging Face

MIDAS mokslo duomenų archyvas

GitHub

OR ELSEWHERE...

Types of projects

Text (6)

Speech Recognition (2)

Machine translation (5)

Speech Synthesis (1)

Other (2)

TEXT

1. Lithuanian corpus, BERT and GPT models	VSSA	3.5B words, 2 LM
2. Summarisation corpora	VDU/VU	4 corpora
3. Question-answer corpus	VSSA	15M pairs
4. Anonymisation corpus	KTU	Spec. corpus
5. Morpho-syntactic annotated corpus	VDU	10M corpora (2)
6. Fake news corpus	VSSA	5K examples

Machine translation

1. Multi- and monolingual corpora (UKR, NOR, SWE, DAN, ESP)	VSSA
2. Multi- and monolingual corpora (EN, DE, FR, PL, LT)	VSSA
3. Synthetic parallel corpora (LT-EN, LT-FR, LT-DE)	VSSA
4. Medical parallel and monolingual corpora	VSSA
5. Defense and security parallel and monolingual corpora (EN-LT)	VSSA

Speech recognition

1. Speech corpus	VU, VDU, LKI	10K hours
2. Medical speech corpus	VSSA	400 hours

Speech synthesis

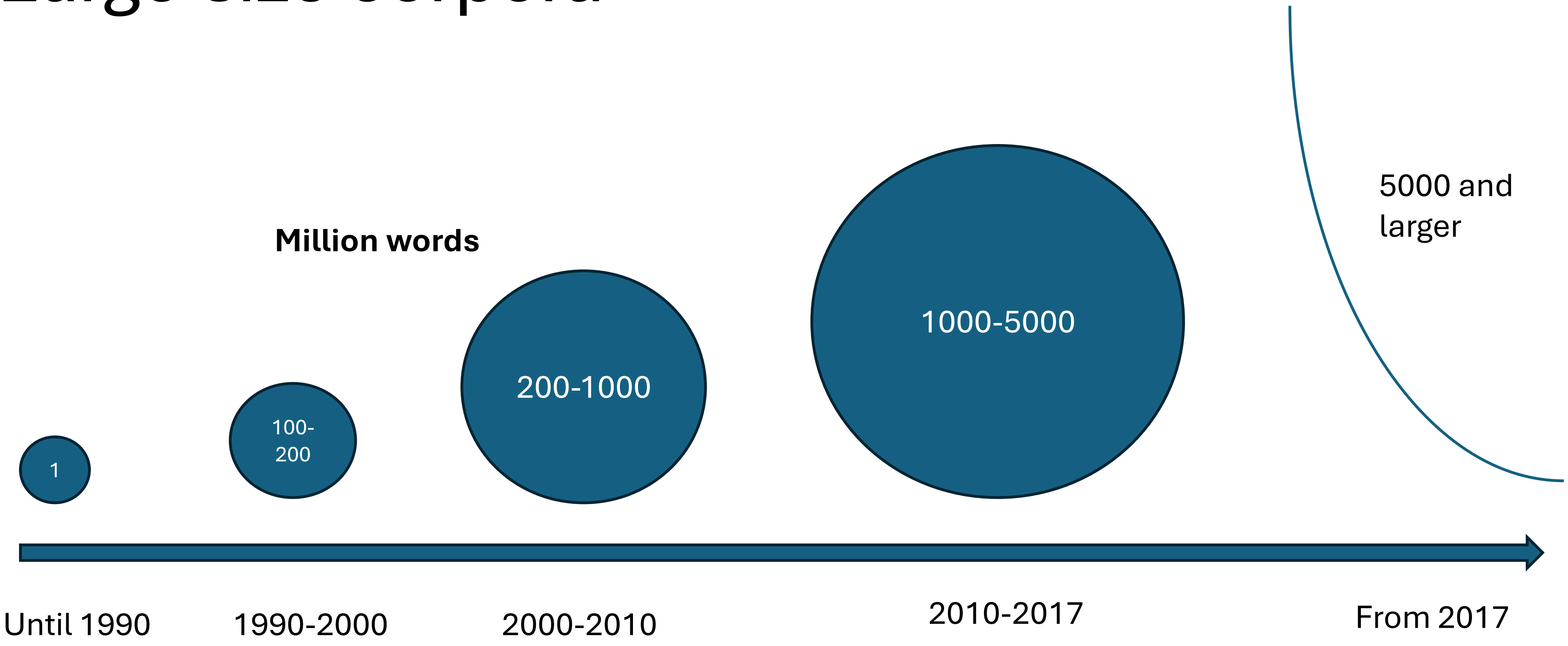
1. Speech synthesis corpus	VSSA	400 hours

Other

1. Language heritage transformation and Creation and language GIS resources	LKI	Various resource and geo data
2. Human Phenotype Ontology (HPO)	VSSA	Ontology (13K concepts)

The concept of “large corpus”

Large size corpora



How much is 1 billion words?

1B words

□ ~15,000 books
□ ~1,400,000 internet pages
□ ~ 2,000,000 messages

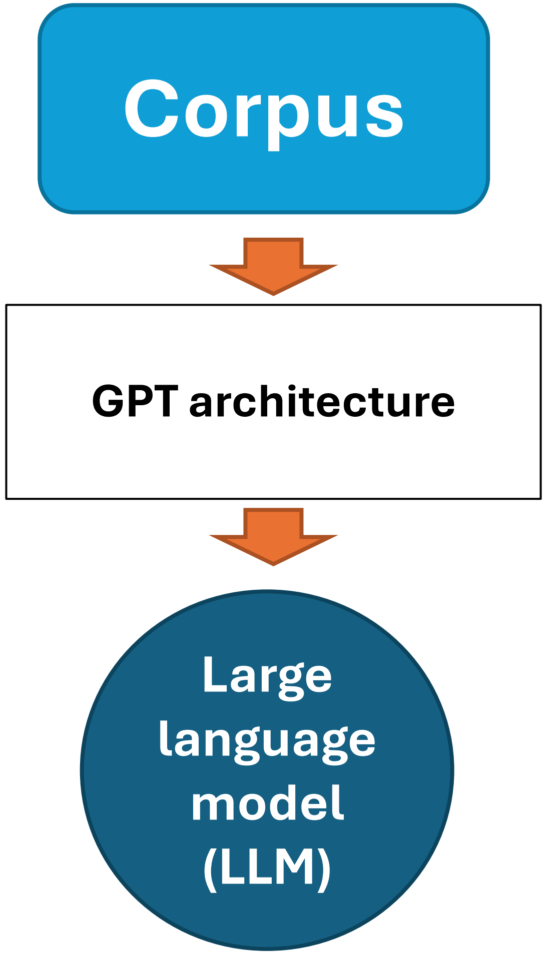
What amount of text is reachable by Google (2022)?

□ **LAT -**	~53B words
□ **LIT –**	~56B words
□ **EST –**	**~81B words**

Generated by ChatGPT 4 with my edit

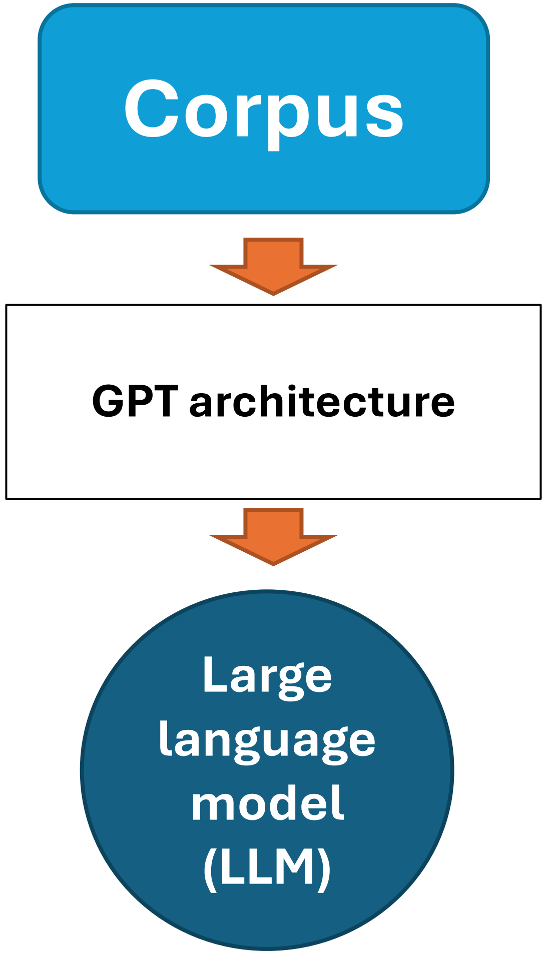
V. Dadurkevičius ir A. Utka (2022)

Why do we need large size corpora?



- If you want to train a large language model using GPT (*generative pretrained transfromer*) architecture,
- you need an enormous amount of language data.

Why do we need large size corpora?



Corpus

GPT architecture

Large language model (LLM)

300B words

175B parameters

- The model is "smarter“, when it has been trained on larger number of parameters.
- The amount of data should be significantly greater than the number of parameters in order to avoid overfitting.

ChatGPT (GPT 3.5)
800GB

Development of the General Lithuanian Language Corpus and Vectorized Lithuanian Language Models

Funded by
the European Union
NextGenerationEU

NEW GENERATION
LITHUANIA

Project goals and tasks

- Compile the Lithuanian General corpus – 3.5B words
- Create two language models:
  - BERT architecture (based on 50% of corpus)

    https://data.gov.lt/datasets/3923/

    https://huggingface.co/VSSA-SDSA
  - GPT model (is under construction)

    The release date is planned in April, 2026

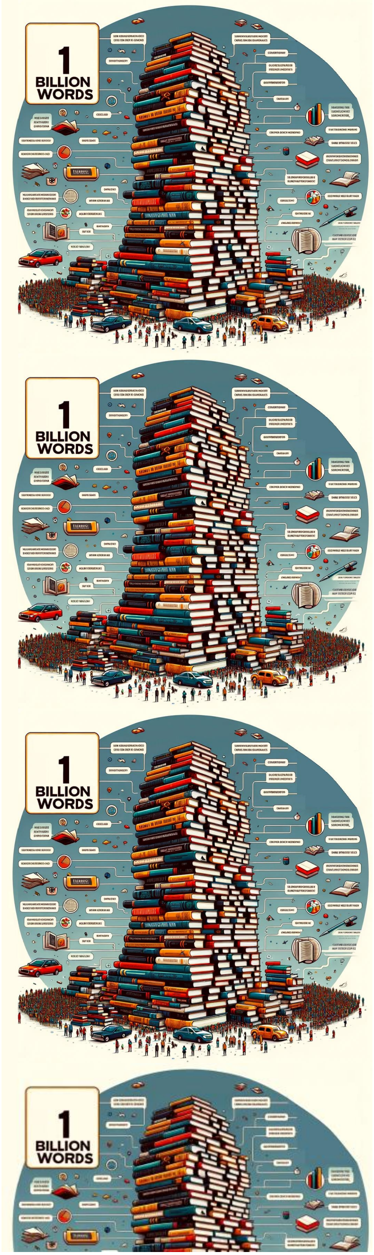
VSSA VALSTYBĖS SKAITMENINIŲ SPRENDIMŲ AGENTŪRA

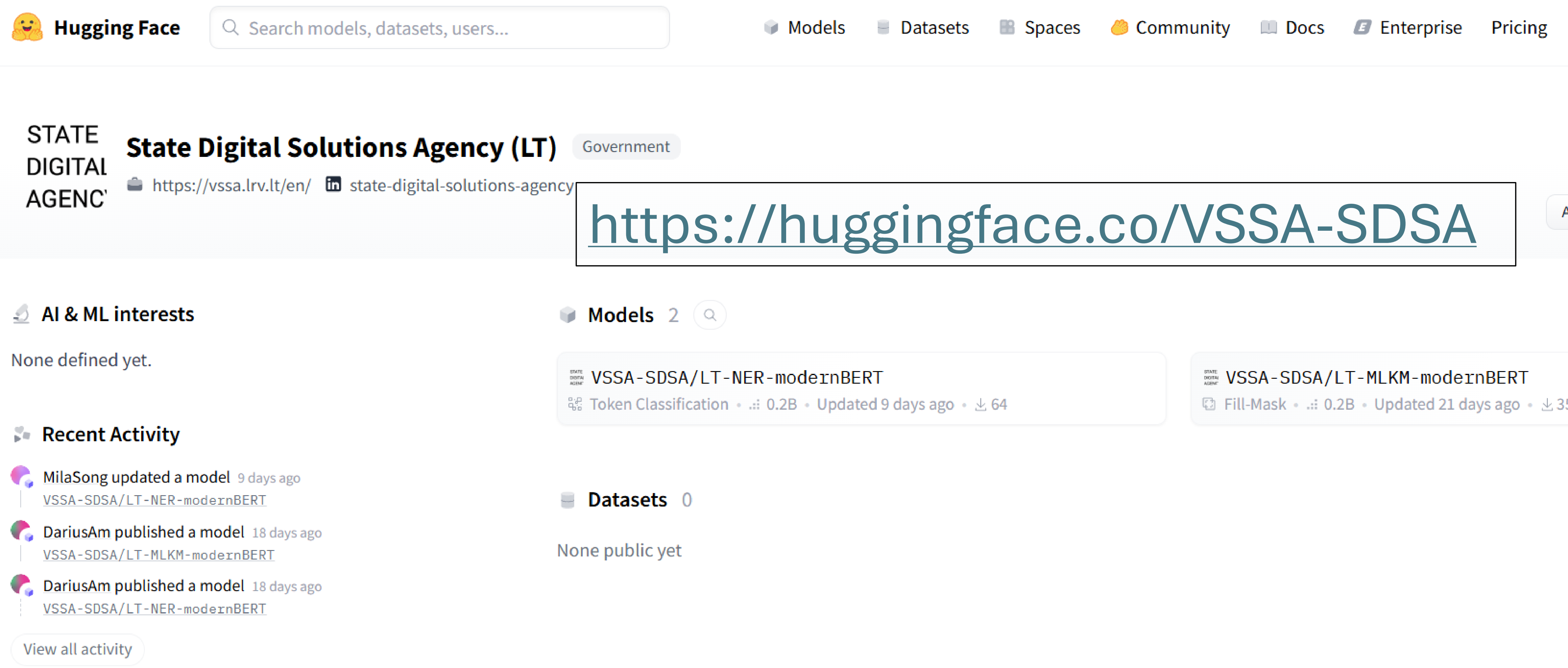
UNIVERSITAS VYTAUTI MAGNI CAUNAE

NEUROtechnology

TILDE

MB KRILAS





State Digital Solutions Agency (LT) Government

https://vssa.lrv.lt/en/ 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

The textual data wasn’t easy …

- When collecting data, we had to ensure that it is collected ethically.
- Data cannot not include private information (GDPR requirements).
- We cannot simply scrape data that is publicly available on the web – we need to get permission from data providers.
- Therefore, a large part of the data in the corpus is licensed.
- Due to copyright restrictions, the part of books and fiction texts in the corpus is very small.

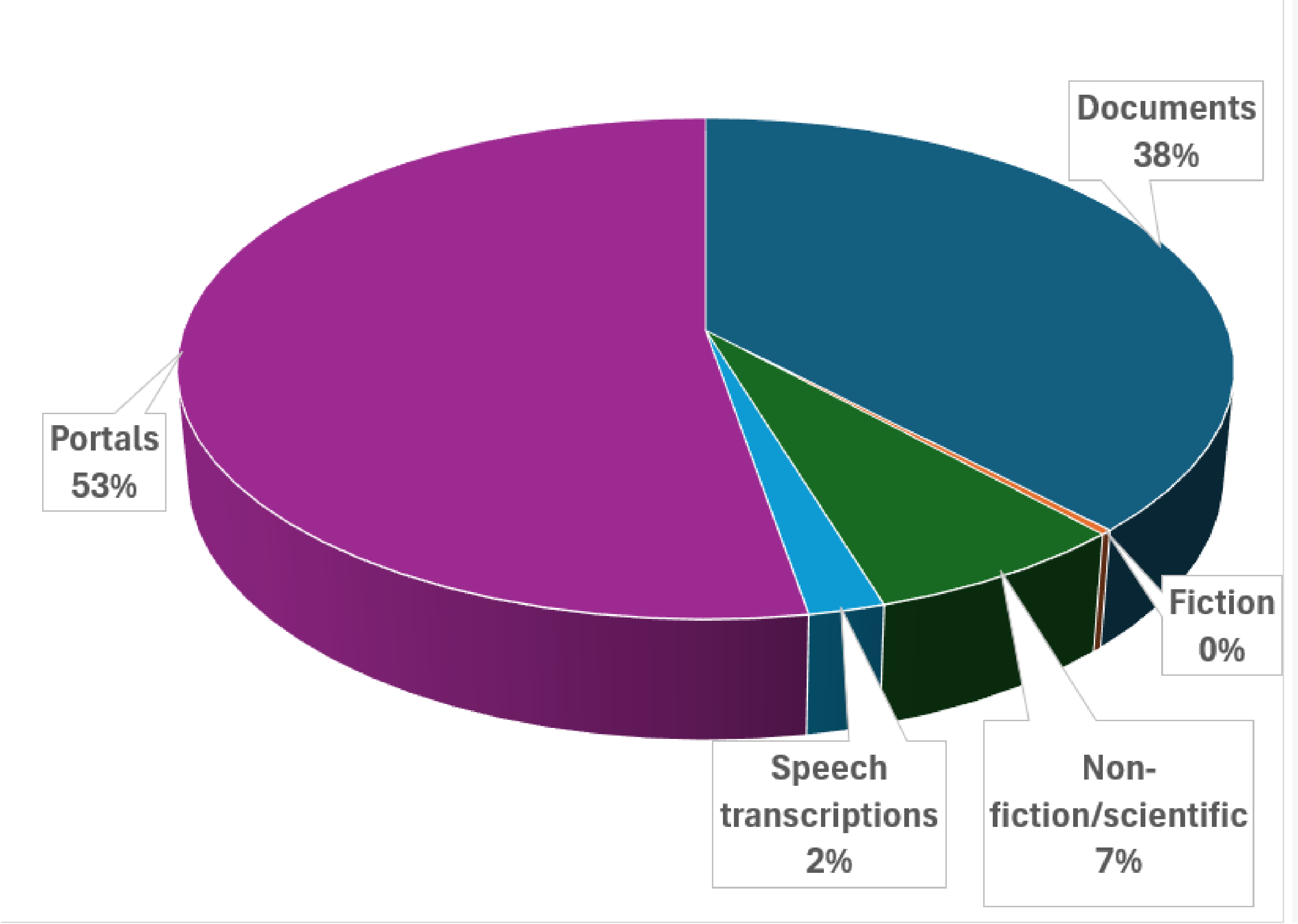
Major data providers licensed data

Delfi
LRT | .lt
Vakarų ekspresas VE
Lrytas
ŠVENČIONIŲ kraštas
Gargždapilis
sena
AUTO GIDAS
15
bernardinai.lt
Humanitarų meka
ELTA
KAS VYKSTA Kaune

Corpus structure

Type of texts	Alpha words
Documents	1509524702
Fiction	11849476
Non-fiction/scientific	280706171
Speech transcriptions	80483739
Portals	2095911732
Total	**3978475820**

Periods	Part
1922-1940	0,28%
1941-1989	0,26%
1990-2004	5,78%
2005-2026	93,68%



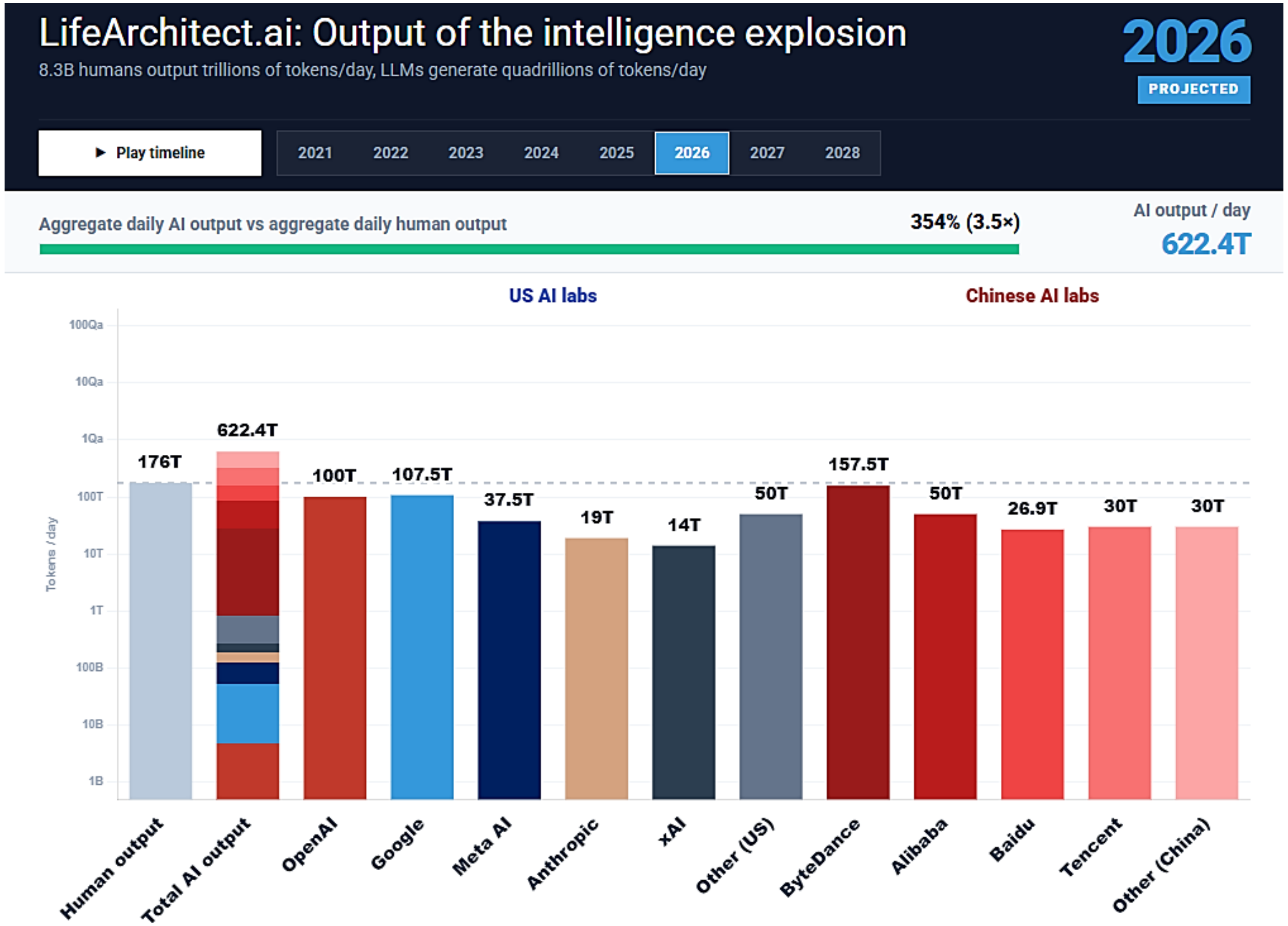
An finally

AI explosion

Alan D. Thompson:

In 2026 major AI systems ***per day*** *will generate more text than the entire humanity.*

(based on the assumption that a person on average produces 16K words per day (Mehl et al. 2007, Hoffman et al. 2022))



LifeArchitect.ai: Output of the intelligence explosion

8.3B humans output trillions of tokens/day, LLMs generate quadrillions of tokens/day

2026

PROJECTED

Aggregate daily AI output vs aggregate daily human output 354% (3.5×)

AI output / day 622.4T

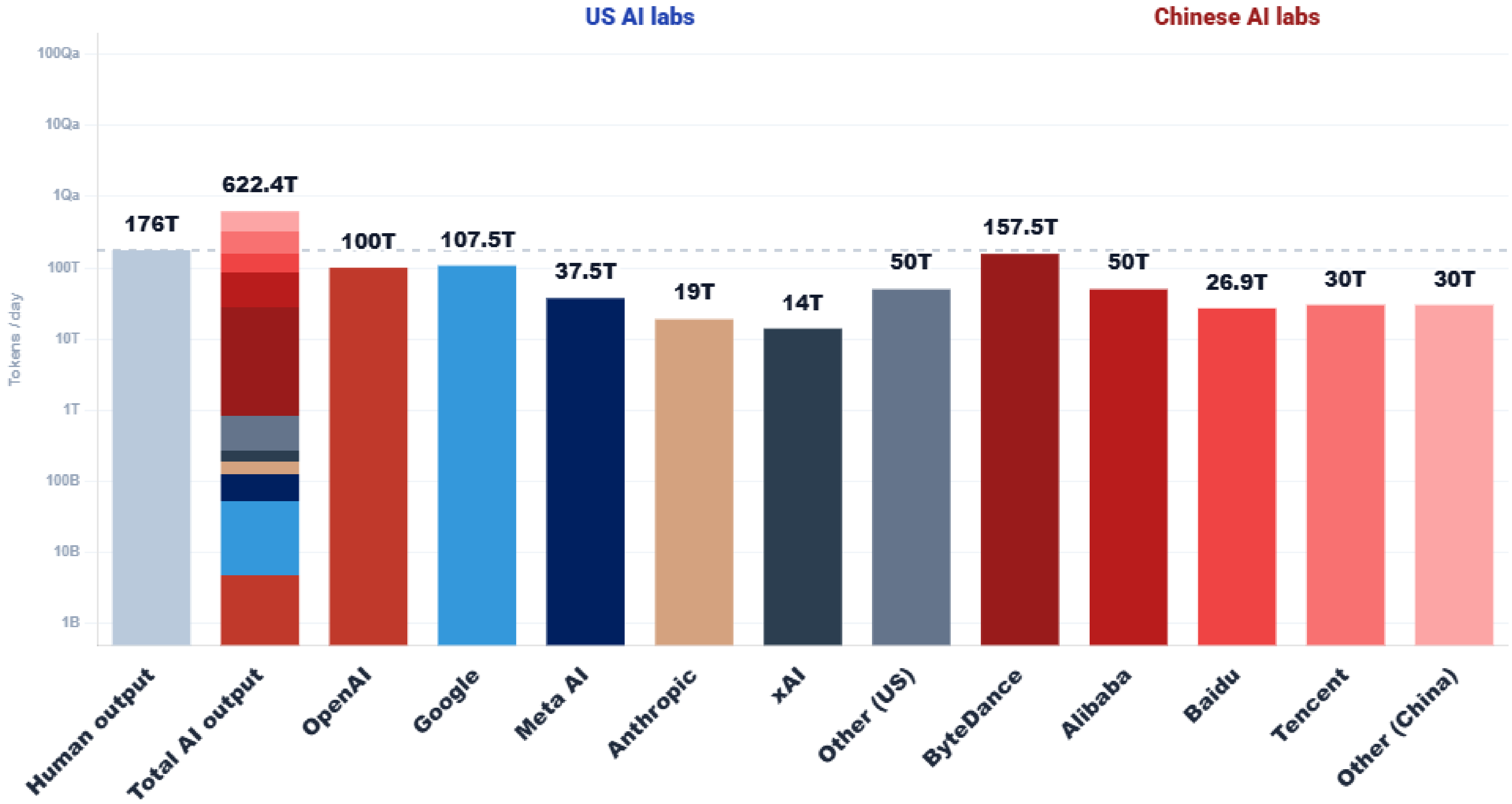
Source: https://lifearchitect.ai/intelligence-explosion.html

Aggregate daily AI output vs aggregate daily human output

354% (3.5×)

AI output / day

622.4T



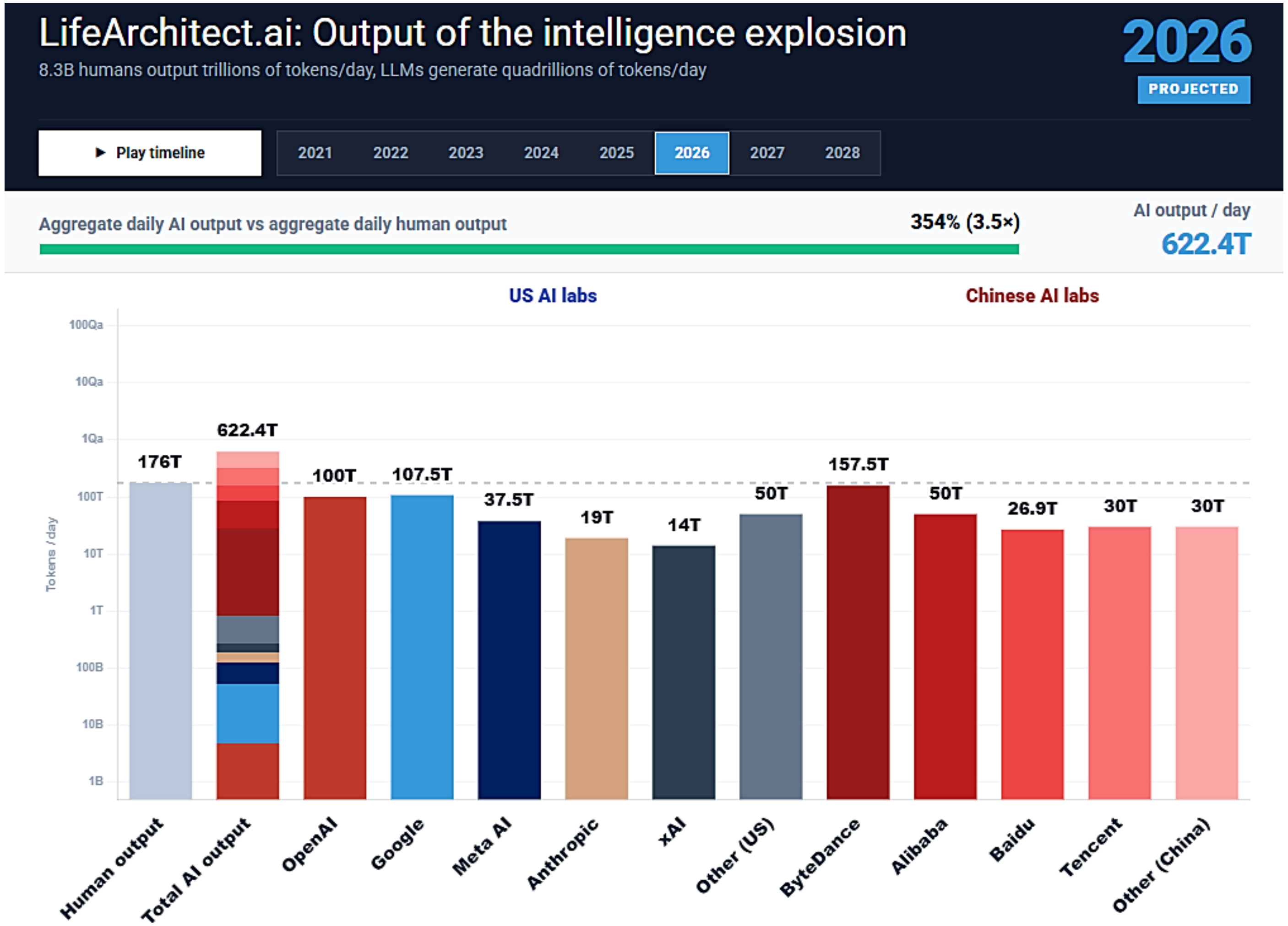
AI explosion

Alan D. Thompson:

*In 2026 major AI systems **per year** will generate more text than the entire humanity.*

(based on the assumption that a person on average produces 16K tokens per day (Mehl et al. 2007, Hoffman et al. 2022))

What does it mean for humanity and for all of us?



LifeArchitect.ai: Output of the intelligence explosion

8.3B humans output trillions of tokens/day, LLMs generate quadrillions of tokens/day

2026 PROJECTED

Aggregate daily AI output vs aggregate daily human output 354% (3.5×)

AI output / day 622.4T

Source: https://lifearchitect.ai/intelligence-explosion.html

Thank you!