Europe's AI flagship

SILOGEN

Specialized large language models (LLM)

October 2023

Aarne Talman
Head of Technology, SiloGen

- Natural language processing (NLP) researcher and leader with 20 years of industry experience.
- Worked in software development, product management (search engines), management consulting, academic research and technology leadership positions.
- Visiting Researcher in Language Technology at University of Helsinki.
- Academic research focuses on natural language understanding.
- Currently leading Silo AI's generative AI and LLM technology development at SiloGen.

SiloGen is positioned to be a leader in GenAI…

6	COUNTRIES
10	OFFICES
300+	AI EXPERTS
150+	PHDS
200+	PRODUCTION-LEVEL AI

SILO AI LAYS THE FOUNDATION …

Europe's largest private AI lab

Unique team of 300+ PhDs and AI scientists/engineers, incl. NLP & LLMs

Trusted AI partner to industry leading companies across the world

Data-centric AI development platform

… FOR EUROPE'S GenAI FLAGSHIP

Team with experience to build end to end LLMs, product and scale operations

Developing leading open multilingual language models on LUMI supercomputer

Family of specialized LLMs

LLM platform to finetune & deploy LLMs

SILOGEN

“ChatGPT is the 2 minute trailer for the 50 year movie, most of the plot is not shown yet”

“AI will have impact comparable to aircraft and computers” – US DoD

SILOGEN

GenAI narrative: This time is different?

GOAL	FSD	AGI	
TRIGGER	TESLA	OpenAI	Microsoft
FOLLOWERS	cruise WAYMO	⇩	How is this time different?
FSD / AGI	***"Tesla recalls 362,758 vehicles, says Full Self-Driving Beta software may cause crashes,"*** CNBC, Feb '23 ***"Tesla's self-driving efforts rank last in study of autonomous driving firms,"*** Business Insider, Apr '23	⇩	
Production-grade AI	***"Waymo is using AI to simulate autonomous vehicle camera data,"*** VentureBeat, May '20 ***"Tesla's Autopilot Depends on a Deluge of Data,"*** IEEE Spectrum, Aug '22	?	

Large language models (LLM) create value as part of (software) products

Enterprise
- ERP/CRM search
- AI Assistant

Legal
- Legal co-pilot
- Compliance & regulatory monitoring

Media
- Media search
- Create & monitor content

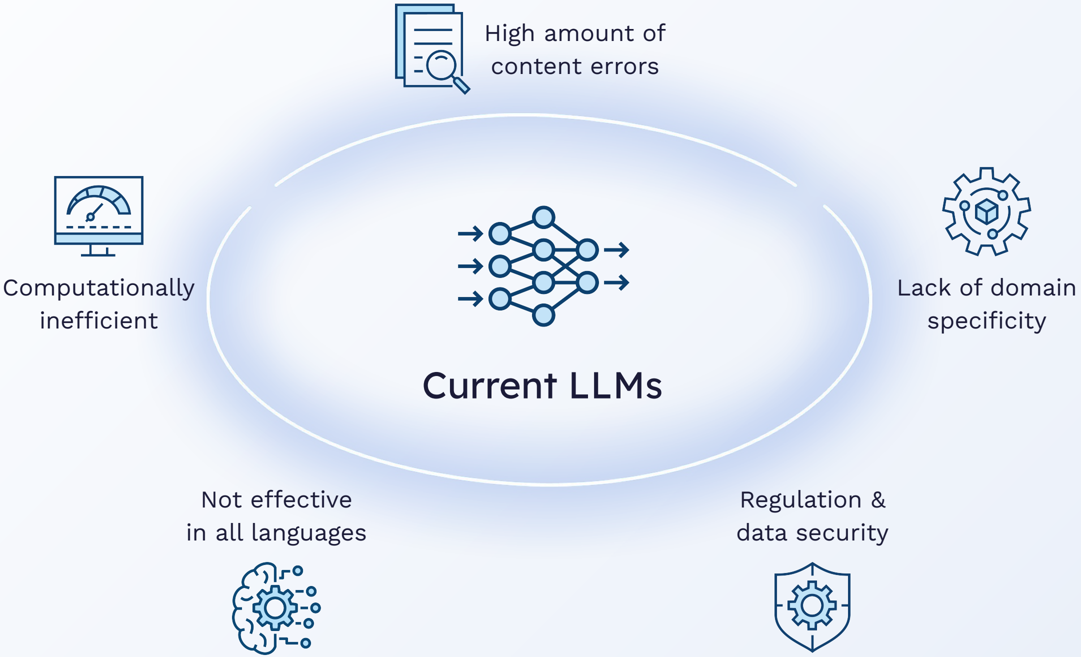
Finance & insurance
- Market analysis & research
- Service for investment advice

Healthcare
- Medical co-pilot
- Personal health advisor

Industrial
- Field services co-pilot
- Customer service chat-bots

THE PROBLEM

Current generic LLMs aren't trustworthy enough to be deployed into products



Highlighting the need for trustworthy and reliable specialized models

We are already seeing a lot of public discussion and news coverage about LLM failures

"Snapchat tried to make a safe AI. It chats with me about booze and sex." Washington Post, Mar '23

"National Eating Disorders Association takes its AI chatbot offline after complaints of 'harmful' advice," CNN, Jun '23

"Plagued with errors: A news outlet's decision to write stories with AI backfires," CNN, Jan '23

"Lawyer used ChatGPT in court - and cited fake cases. A judge is considering sanctions," Forbes, Jun '23

"Mozilla pauses error-prone AI Explain feature in MDN," The Register, Jul '23

"Google shares lose $100 billion after company's AI chatbot makes an error during demo," CNN, Feb '23

Current LLMs are not compliant with European regulation

Grading Foundation Model Providers' Compliance with the Draft EU AI Act

Source: Stanford Research on Foundation Models (CRFM), Institute for Human-Centered Artificial Intelligence (HAI)

	OpenAI	cohere	stability.ai	ANTHROPIC	Google	BigScience	Meta	AI21labs	ALEPH ALPHA	EleutherAI	
Draft AI Act Requirements	**GPT-4**	**Cohere Command**	**Stable Diffusion v2**	**Claude**	**PaLM 2**	**BLOOM**	**LLaMA**	**Jurassic-2**	**Luminous**	**GPT-NeoX**	**Totals**
Data sources	●○○○	●●●○	●●●●	○○○○	●●○○	●●●●	●●●●	○○○○	○○○○	●●●●	22
Data governance	●●○○	●●●○	●●○○	○○○○	●●●○	●●●●	●●○○	○○○○	○○○○	●●●○	19
Copyrighted data	○○○○	○○○○	○○○○	○○○○	○○○○	●●●○	○○○○	○○○○	○○○○	●●●●	7
Compute	○○○○	○○○○	●●●●	○○○○	○○○○	●●●●	●●●●	○○○○	●○○○	●●●●	17
Energy	○○○○	●○○○	●●●○	○○○○	○○○○	●●●●	●●●●	○○○○	○○○○	●●●●	16
Capabilities & limitations	●●●●	●●●○	●●●●	●○○○	●●●●	●●●○	●●○○	●●○○	●○○○	●●●○	27
Risks & mitigations	●●●○	●●○○	●○○○	●○○○	●●●○	●●○○	●○○○	●●○○	○○○○	●○○○	16
Evaluations	●●●●	●●○○	○○○○	○○○○	●●○○	●●●○	●●○○	○○○○	●○○○	●○○○	15
Testing	●●●○	●●○○	○○○○	○○○○	●●○○	●●○○	○○○○	●○○○	○○○○	○○○○	10
Machine-generated content	●●●○	●●●○	○○○○	●●●○	●●●○	●●●○	○○○○	●●●○	●○○○	●●○○	21
Member states	●●○○	○○○○	○○○○	●●○○	●●●●	○○○○	○○○○	○○○○	●○○○	○○○○	9
Downstream documentation	●●●○	●●●●	●●●●	○○○○	●●●●	●●●●	●●○○	○○○○	○○○○	●●●○	24
Totals	25 / 48	23 / 48	22 / 48	7 / 48	27 / 48	36 / 48	21 / 48	8 / 48	5 / 48	29 / 48	

Source: Do Foundation Model Providers Comply with the Draft EU AI Act?, Stanford University, https://crfm.stanford.edu/2023/06/15/eu-ai-act.html

Current LLM developers are not Transparent

Foundation Model Transparency Index Scores by Major Dimensions of Transparency, 2023

Source: 2023 Foundation Model Transparency Index

Major Dimensions of Transparency	Meta	BigScience	OpenAI	stability.ai	Google	ANTHROPIC	cohere	AI21labs	Inflection	amazon	
	Llama 2	BLOOMZ	GPT-4	Stable Diffusion 2	PaLM 2	Claude 2	Command	Jurassic-2	Inflection-1	Titan Text	**Average**
Data	40%	60%	20%	40%	20%	0%	20%	0%	0%	0%	**20%**
Labor	29%	86%	14%	14%	0%	29%	0%	0%	0%	0%	**17%**
Compute	57%	14%	14%	57%	14%	0%	14%	0%	0%	0%	**17%**
Methods	75%	100%	50%	100%	75%	75%	0%	0%	0%	0%	**48%**
Model Basics	100%	100%	50%	83%	67%	67%	50%	33%	50%	33%	**63%**
Model Access	100%	100%	67%	100%	33%	33%	67%	33%	0%	33%	**57%**
Capabilities	60%	80%	100%	40%	80%	80%	60%	60%	40%	20%	**62%**
Risks	57%	0%	57%	14%	29%	29%	29%	29%	0%	0%	**24%**
Mitigations	60%	0%	60%	0%	40%	40%	20%	0%	20%	20%	**26%**
Distribution	71%	71%	57%	71%	71%	57%	57%	43%	43%	43%	**59%**
Usage Policy	40%	20%	80%	40%	60%	60%	40%	20%	60%	20%	**44%**
Feedback	33%	33%	33%	33%	33%	33%	33%	33%	33%	0%	**30%**
Impact	14%	14%	14%	14%	14%	0%	14%	14%	14%	0%	**11%**
Average	**57%**	**52%**	**47%**	**47%**	**41%**	**39%**	**31%**	**20%**	**20%**	**13%**	

Source: The Foundation Model Transparency Index, Stanford University, https://crfm.stanford.edu/fmti/

The Solution

Our approach to trustworthy, reliable and private LLMs for industrial use cases

Base Models

High quality and compliant base models that cover multiple languages

Specialization

Specialization of base models to specific domains and industries **through fine tuning** with domain specific data

Controls

Complement specialized LLMs with **retrieval augmented generation (RAG), controls and guardrails**

Evolution of Large Language Models

	BASE	CO-PILOT	CONVERSE	AGENTS
Model description	Models that generate and summarize information	Co-pilot AI built on instruction-tuned dialogue models	Conversational AI built on instruction-tuned dialogue models	Agent AI built on RL to perform autonomous tasks
Domains & use cases	Generic sentence completion, generation and summarization	Augments experts through generation and summarization capabilities	Conversational agents with trustworthy end-to-end capability	Autonomous agents that drive external systems and take actions

Current focus (BASE, CO-PILOT)

In progress (CONVERSE, AGENTS)

SiloGen base models - Multilingual LLMs pre-trained from scratch (coming early next year)

- Developed in collaboration with TurkuNLP
- Trained on LUMI supercomputer
- Support for all the European languages
- Data collected in the EU funded HPLT project
- Trained with large amount of data (over 2.4T tokens)

Model size (parameters)	Tokens	Comparison
7B	2.4T	2T tokens (7B LLaMa 2)
15B	2.4T	2T tokens (13B LLaMa 2)
30B	2.4T	2T tokens (34B LLaMa 2)
60B	2.4T	2T tokens (70B LLaMa 2)
120B	4.8T	366B tokens (174B BLOOM)

Specialization improves model accuracy while being more cost efficient and trustworthy

Accuracy

- Specialization on a particular domain makes an LLM **produce better responses**
  - 20 %-point increase on medical QA [Singhal et al., 2023]
  - Up to 40% improvement on coding tasks [Rozière et al., 2023]

Cost & Efficiency

- in-win-win: **smaller, better, cheaper** models
  - 16B code-specialised model improves by 14% over PaLM 540B on coding tasks [Li et al., 2023]
  - From 22% to 89% accuracy on SQL Generation (vs 79% accuracy for GPT4) [source]

Alignment

- Specialization steers models to make them more **helpful**, **honest**, and **harmless (HHH)**
  - Responses from a 1.3B model fine-tuned with HF is preferred to a 175B model [Ouyang et al., 2022]
  - Aligning models with HHH does *not* come at the cost of accuracy [Bai et al., 2022]
  - Fine-tuning improves significantly on dialogue quality + safety + groundedness [Thoppilan et al., 2022]

Example: Technical documentation co-pilot

How to use QUrlQuery removeQueryItem() method?

Based on the provided sources, the `removeQueryItem()` method of `QUrlQuery` can be used to remove a query item from a URL. The method takes a `QByteArray` key as an argument, which is the name of the query item to be removed.

Here's an example of how to use the `removeQueryItem()` method:

c Copy code

```c
QUrlQuery query;
query.addQueryItem("key1", "value1");
query.addQueryItem("key2", "value2");

// Remove the query item with key "key1"
query.removeQueryItem("key1");

// Print the updated query string
qDebug() << query.query();
```

In this example, the `removeQueryItem()` method is called with the key "key1", which removes the corresponding query item from the URL. The updated query string is then printed to the debug output.

Example customer cases

Customer services BPO Customer Services LLM	**Healthcare software vendor** Healthcare LLM	**Data analytics company** Custom LLM for code generation	**Global asset management company** Custom LLM for finance use cases
IP & Patent database startup LLM for patent search	**Media company** LLM for media search	**Research company** LLM for content creation	**NextGen intranet startup** LLM for generating intranet site design

Aarne Talman

aarne.talman@silo.ai
+358 40 871 8712
https://silo.ai