Turning AI into applications

Creating Value through Customisation, Tailoring, and Quality Assessment

Jussi Karlgren, SiloGen

Building a European AI flagship

SILOGEN

Production-grade AI for the enterprise

6 COUNTRIES

10 OFFICES

300+ AI EXPERTS

150+ PHDS

200+ PRODUCTION-LEVEL AI

LLMs will become an integral component in software products

but

SOFTWARE PRODUCTS NEED TRUSTWORTHY AND ROBUS HIGH QUALITY LMS FOR SPECIFIC USE CASES…

Enterprise software
ERP & CRM & Enterprise search & AI Assistant

Media software
Media monitoring, summarisation, and analysis
Automatic media production for specific channels

Healthcare software
Medical co-pilot
Personal health advisor

Legal software
Legal co-pilot
Compliance & regulatory monitoring

Finance & insurance software
Market analysis & research
AI Assistant

… AND WHILE TODAY'S LLMS PROVIDE RAW MATERIAL THEY ARE NOT THE FINISHED PRODUCT …


High amount of content errors


Lack of domain specificity


Not effective in all languages


Computationally inefficient


Undetermined adherence to regulation & data security demands

… WHICH HAS BECOME APPARENT IN MANY CASE STUDIES.

"ChatGPT Is Losing Users. Is The Artificial Intelligence Craze Over?," Forbes, Jul '23

"The AI bot has picked an answer for you. Here's how often it's bad," Washington Post, Apr '23

"Trying Microsoft's new AI chatbot search engine, some answers are uh-oh," Washington Post, Feb '23

"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

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

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

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

"We tested a new ChatGPT-detector for teachers. It flagged an innocent student," Washington Post, Apr '23

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

We need specialised models!

Value creation comes from application and integration

Specialisation is the path to relevance

vertical

bespoke

Turning base models into specialised models for products

Data initiative

annotation, synthetic data generation with specialised LLMs, proprietary task- and application specific data

Fine tuning

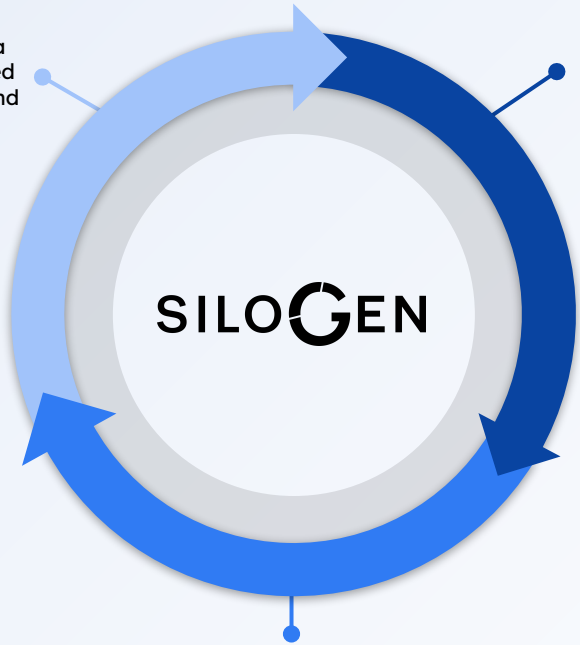
human feedback, instruction fine tuning, and reinforcement learning with AI feedback

Base models

Multilingual high quality base models

many represented in this room:

Poro, GPT-SW3...



Specialised LLM API

Quality evaluation

standard benchmarks, domain specific tests, application specific quality assessment

Three steps to enterprise-grade LLMs

Base

High-quality & compliant base models
Independent and transparent open source models that cover multiple languages

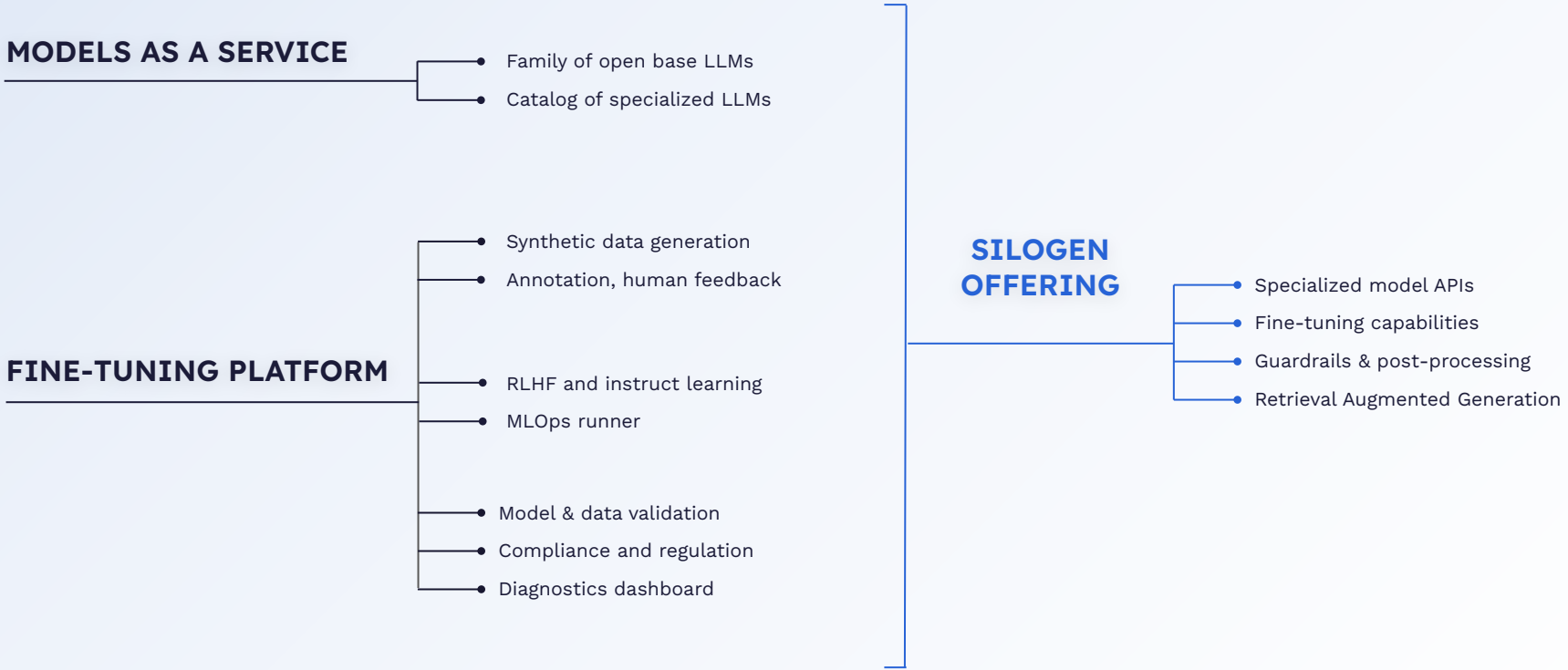
Platform

Data-centric LLM development
Platform to build models for specific domains & industries through fine-tuning and data generation

Studio

Specialised model APIs with controls
Specialised models with fine-tuning, guardrails, post-processing & retrieval augmented generation

SiloGen's domain-specific LLM offering



Example use cases for specialized models

BUSINESS CRITICAL B2B CO-PILOT
Technical sales and customer operations in high-risk and high reward B2B environments

HIGH-SENSITIVITY B2C CO-PILOT
Virtual sales agents co-pilots for high sensitivity and high regulation environments (e.g. retail banking)

HEALTHCARE CO-PILOT
Healthcare-specialized model tailored for healthcare data structures and terminology

Specialised models from SiloGen improve performance

compared to base models

with respect to

Domain knowledge

Truthfulness within domain

Robustness

- SiloGen contains both quantitative benchmarking of models against generic evaluation benchmarks as well as industry-specific benchmarks developed for each vertical offering.
- Example intrinsic evaluation (healthcare): Base model: **51.53** perplexity, Fine-tuned specialized model: **27.58** perplexity (lower better)

CUSTOMER	QUALITATIVE EVALUATION
tietoevry	Tietoevry's medical expert team did qualitative assessment and found that the output of specialized model significantly outperformed generic models.
Allianz	More performant specialized LLMs for code generation on standard code generation benchmarks
yle	YLE's journalists evaluated our specialized model and found that it performs significantly better in the given use case of generating news articles with specific style requirements.
RightHub®	RightHub's IP legal expert evaluated the specialized model output and verified its high performance in the IP and patent search use case.

SiloGen base models being trained now

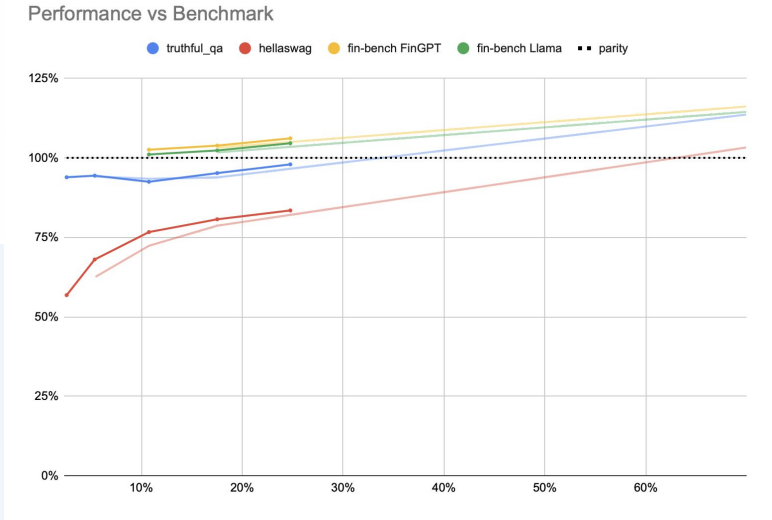
- Developed together with TurkuNLP
- Trained on LUMI
- Will support all official European languages
- Focus currently on EN, FI, code
- Data collected in the EU funded HPLT project
- Mostly monolingual data
- Cross-lingual benefits already apparent
- Models will be released openly

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)

Base models at state of the art performance

Checkpoint release in Nov 2023!

PORO



SILOGEN

How can we tell if we are doing the right thing?

monitoring

tracking standard tests and research test suites

benchmarking

comparing systems and system components wrt requirements and quality concerns

validation

end-to-end system quality assessment

base › vertical › bespoke

Top-level quality criteria

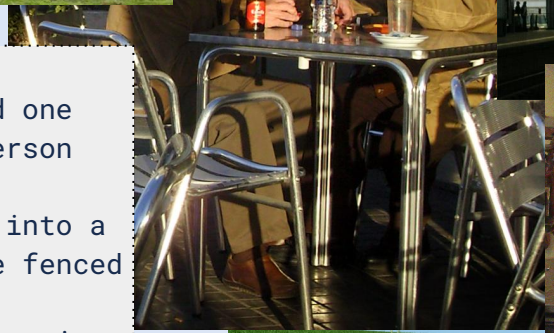
Language	*Are system utterances correct and well-formed language?*
Discourse	*Is the conversation fluent over the several turns of a session?*
Social awareness	*Is the output of a system appropriate for the conversation and the parties engaged in it?*
Consistency	*Does the system produce robust output to varied input?*
Veracity	*Is the output of a system truthful?*
Topical competence	*Does the system know the topic enough for its output to be trusted?*
Compliance	*Is the content of the model and the output compliant with regulatory constraints?*
Common sense	*Is the system capable to reason using language?*
Effectiveness	*Does the system hold to budgets with respect to time, computational effort, and hardware?*
Creativity	*Is interaction with the system interesting, delightful, and fun?*

Quality assessment is the key to trustworthy systems

Quality tests and culture

Hellaswag is an excellent test, but:

- "Two people are seen sitting before a wave pool and one leads another out onto the water on a board. The person ..."
- "A cowboy rides a horse out of a corral and enters into a fenced off area. The man rides his horse out of the fenced are and throws a rope ..."
- "A small group of people are seen sitting around a casino table speaking to one another and playing a game of poker ..."



Cross-lingual benefits are real ...

... but tests need to be translated and customised!

Cultural aspects are part of translation ...

... and this extends to instruction tuning

ELOQUENT

CLEF shared task lab for evaluation of generative language model quality

"Is your LLM really clever? Can it mark its own homework?"

Task 1 - Topical Competence: *Does your generative language model know what it is talking about?*
Task 2 - Hallucinogen: *Is this text true? Or make-believe?*
Task 3 - Robustness: *Will your generative language model respond with the same content to all of us?*
Task 4 - Voight-Kampff: *Has a machine or a human author put together these words?*

The first ELOQUENT edition will happen in 2024

- Fall 2023: discussion and task formulation
- January 2024: tasks open
- April 2024: registration for participation closes
- May 2024: submission deadline of experimental runs from participants
- September 2024: workshop in Grenoble

SiloGen: Jussi Karlgren and Aarne Talman
RISE ICT: Liane Guillou, Luise Dürlich, Evangelia Gogoulou, Joakim Nivre
AI Sweden: Magnus Sahlgren

Open high quality base models
Use case-directed customisation and specialisation
Task-directed tailoring for bespoke solutions
Continuous quality assessment
Ongoing research collaborations

SiloGen for trustworthy generative AI