AI for subject indexing and descriptive cataloguing – a civilized approach?

Osma Suominen

KI – konkurrent eller kompanjong?
Oslo, 24 April 2026

NATIONAL LIBRARY OF FINLAND

NATIONAL LIBRARY OF FINLAND

About me

Fediverse/Mastodon:
@osma@sigmoid.social

LinkedIn:
osmasuominen

GitHub:
@osma

Osma Suominen
Information Systems Specialist, National Library of Finland

Doctoral thesis "*Methods for Building Semantic Portals*"
Semantic Computing Research Group, Aalto University, 2013
Supervisor Professor Eero Hyvönen

Joined the National Library in 2013
to set up the Finto.fi thesaurus and ontology service

Working on opening up bibiliographic metadata as Linked Data (Fennica-LD) and automated subject indexing (Annif)

Open source software projects e.g.:

- Skosify – Validation and QA tool for SKOS vocabularies
- Skosmos – SKOS vocabulary publishing tool
- Annif – Tool for automated subject indexing and classification

Languages spoken: Finnish, Swedish (~Norwegian), English, Estonian

OUR VISION: **Bildung** at the heart of society

Swedish: **Bildning** (dannelse) i kärnan av samhället

Photo: NatLibFi / Frida Lönnroos

vs.

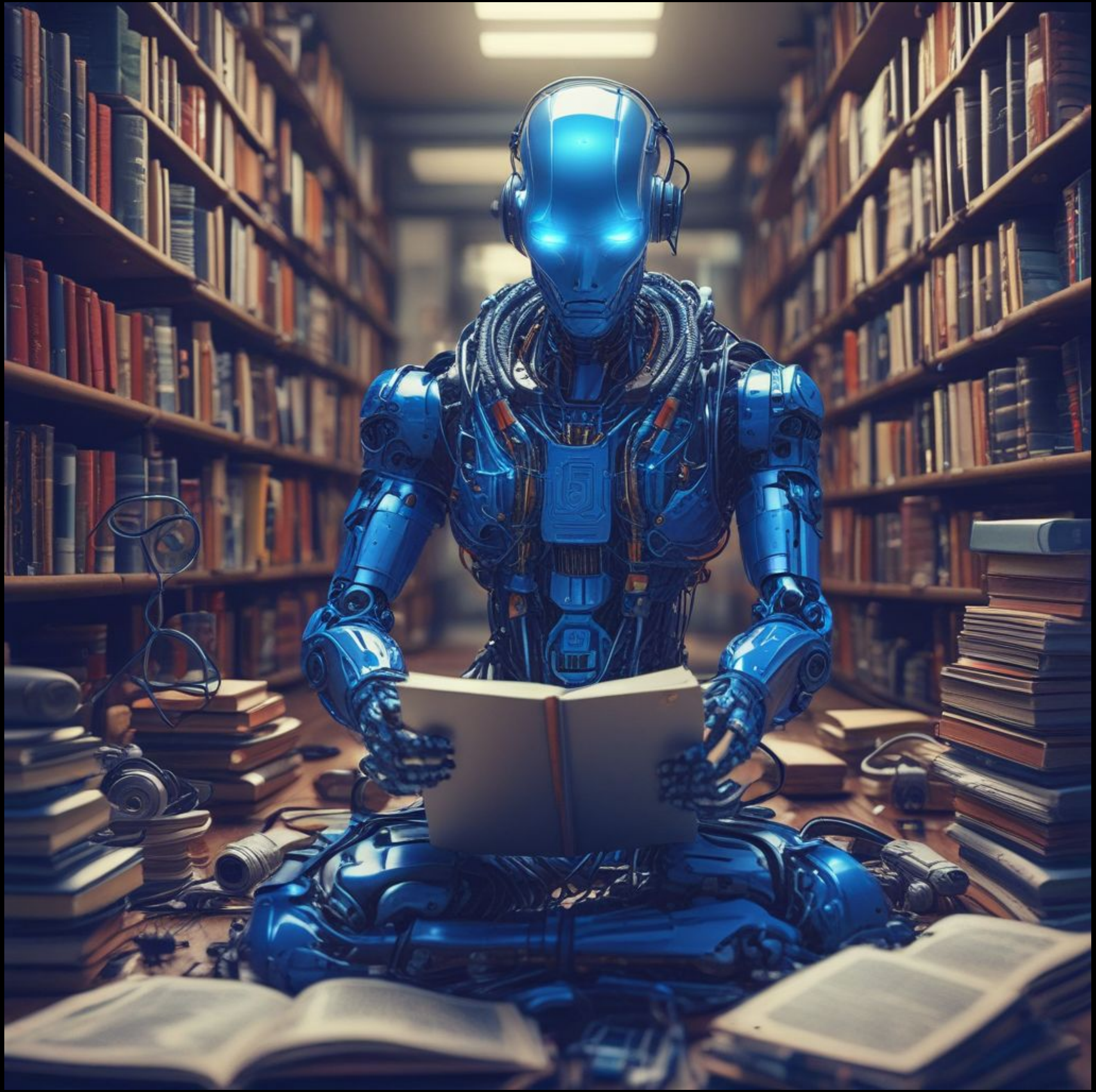


Image: Stable Diffusion XL
Prompt: An AI robot looking at books, electric blue, mechanic, shiny, lots of cables

There are many problems with AI…

As a 🧑‍💻🧑‍🔧, I am ✨🤩 about AI

As a 🧑‍🏫🤔, I am 😱🥶 about AI

How can we develop practical AI solutions
without amplifying the worst problems of AI?

Automated subject indexing: Annif and Finto AI

How can I annotate my documents or data with concepts from Finto?

annif

developed since 2017

General purpose open source **tool** for automated subject indexing and classification

Multilingual, supports many vocabularies

Code on GitHub, website with test form and API

Global development and user community; user forum **annif-users** on Google Groups

annif.org

fintoai

launched in 2020

Automated subject indexing **service** for production use, based on Annif

Supports indexing with the General Finnish Ontology YSO in Finnish, Swedish and English language

Web user interface and API service

Intended to support subject cataloguers in Finland regardless of institution (GLAMs, public administration); sister project to the Finto vocabulary service

ai.finto.fi

Demonstration

Automated subject indexing using Finto AI

Based on the Annif toolkit that we develop

Languages: Finnish, Swedish or English

No LLM needed! *

* In fact, LLMs are pretty bad at subject indexing when prompted in a typical naïve way.

We took part in the two LLMs4Subjects challenges in 2025.
Our hybrid system (Annif + data pre-processing with LLMs) performed better than pure LLM systems.

Metadata extraction and descriptive cataloguing: Meteor + LLM and BIBRA

How can I get basic metadata out of electronic grey literature publications for making a catalogue record?

NATIONAL LIBRARY OF FINLAND

Grey literature?

reports

working papers

government documents

white papers

preprints

theses

...

Semi-formal non-commercial PDFs published on the web – lots of them! Often with lacking metadata.

Image made with DreamStudio (based on Stable Diffusion)
Prompt: "A big pile of papers, reports, documents, PDF files, word documents, powerpoint slides, posters, articles and books forming a wave in the style of Hokusai"

Why extract metadata from grey literature PDFs?

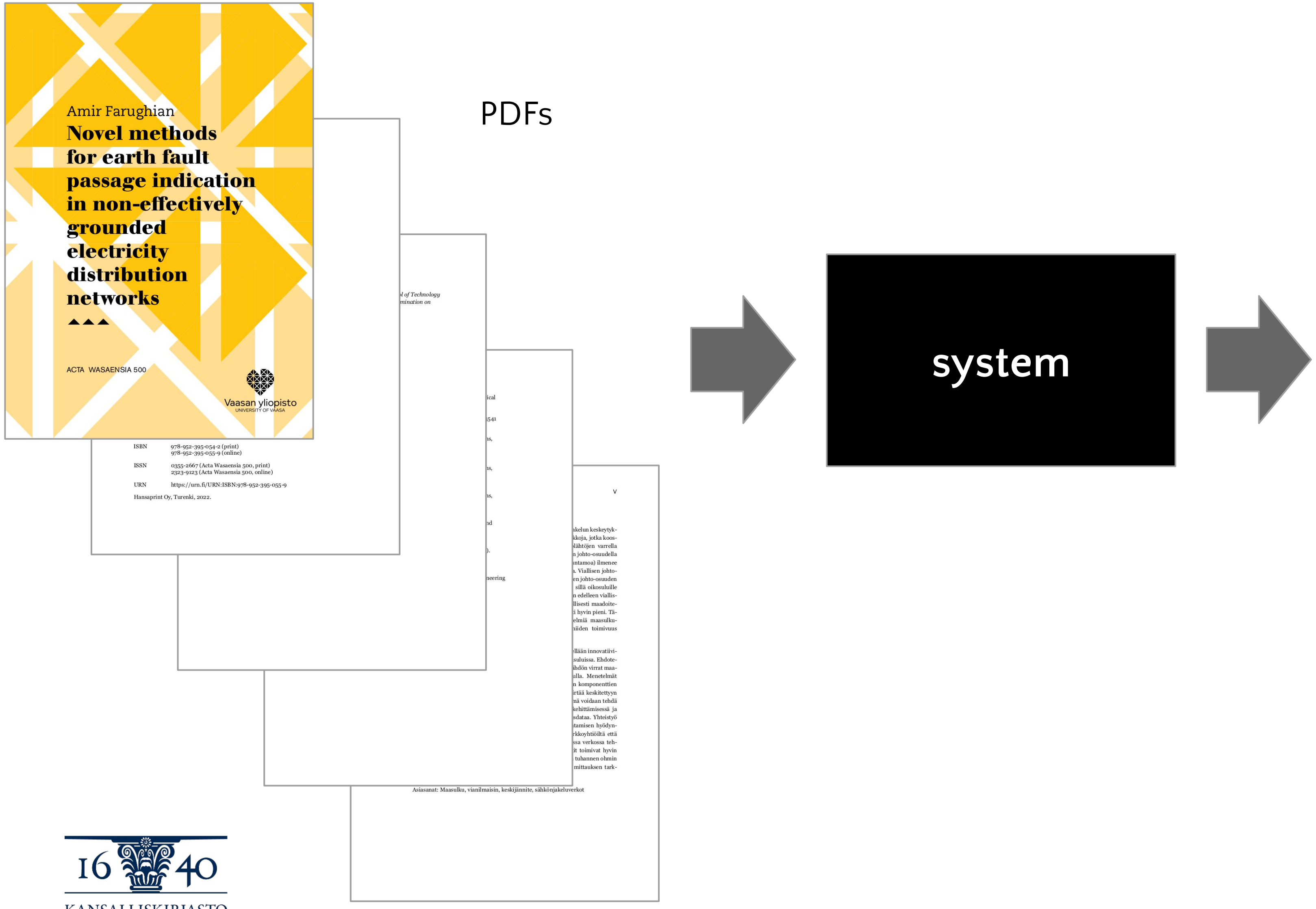
1. help users of **digital repositories** who need to enter metadata when uploading PDFs

2. in **web archiving**, get more information about harvested PDF files

3. ease **cataloguing** of e.g. government reports and doctoral theses

16 40 NATIONAL LIBRARY OF FINLAND

Extracting bibliographic metadata



title: Novel methods for earth fault passage indication in non-effectively grounded electricity distribution networks

creator: Farughian, Amir

publisher: University of Vaasa

faculty: School of Technology and Innovations

date: 2022

e-issn: 2323-9123

p-issn: 0355-2667

e-isbn: 978-952-395-055-9

p-isbn: 978-952-395-054-2

ispartofseries: Acta Wasaensia

numberinseries: 500

Our solution

1. FinGreyLit data set for training and evaluation
   currently 1,600 PDFs with curated metadata, 4 languages, on GitHub

2. GreyLitLM series of fine-tuned small language models
   fine-tuned using 75% of the FinGreyLit documents

3. Meteor tool, extended to use an external LLM service

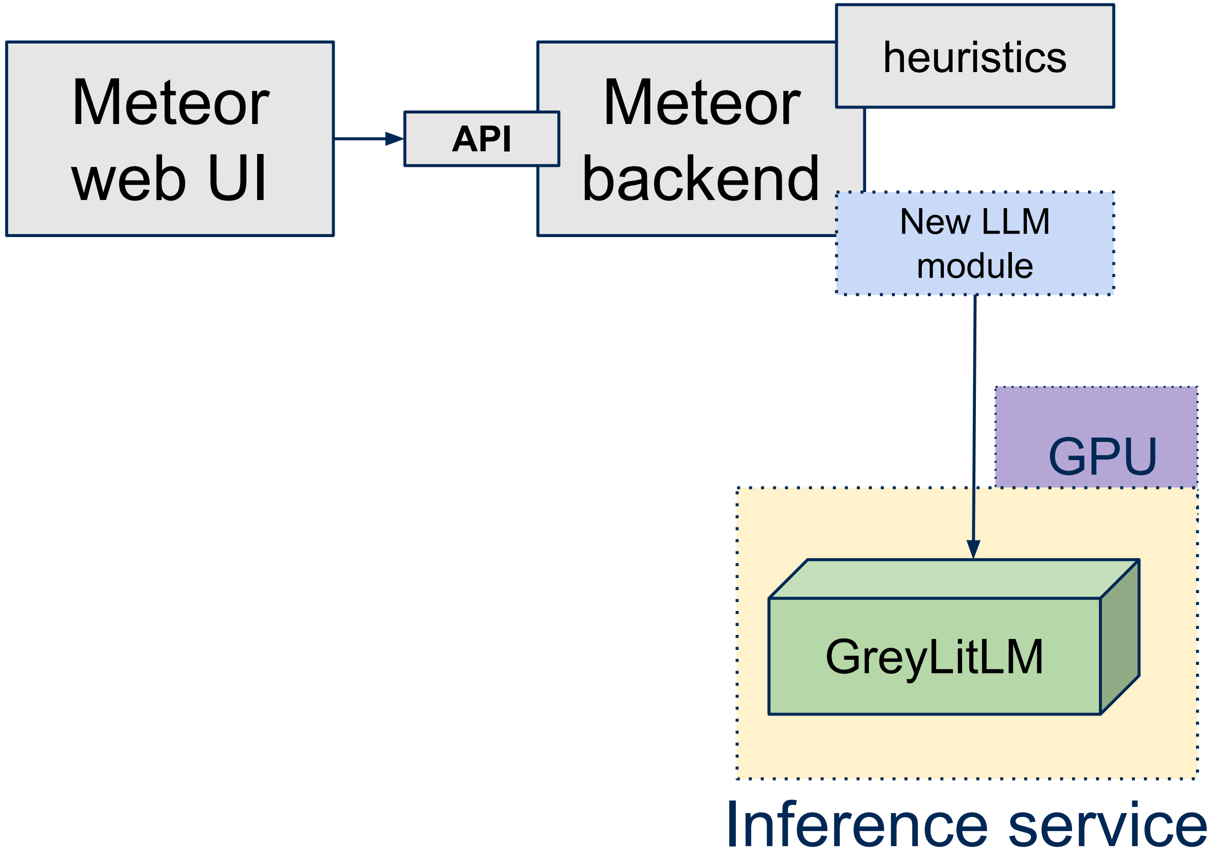
Meteor metadata extraction tool

- published on GitHub as Open Source by the National Library of Norway
- extracts title, authors, language, publisher, year, ISBN, ISSN
- main target is government / public sector reports in Norwegian and English
- based on hardcoded heuristics, no machine learning
- simple web UI + REST API for integration, used in production at NLN

1640
NATIONAL LIBRARY
OF FINLAND

Extending Meteor with an **external LLM**

- Meteor built-in heuristics didn't work well on Finnish grey literature
- We extended it to use an external LLM service with GreyLitLM



Demonstration

Extraction of metadata
from PDF publications
part 1

Prototype system

Small fine-tuned language models
(here based on Gemma3-4B)

Extension of the Meteor tool
by the National Library of Norway

Evaluation – does it work?

Extraction methods to compare:

1. Empty baseline – predict nothing; effective for seldom used fields, e.g. DOI.
2. Meteor – hardcoded logic and heuristics in the Meteor tool.
3. Qwen2.5 0.5B – extremely small language model, fine-tuned by us.
4. Gemma3 4B – very small language model, fine-tuned by us.
5. Mistral NeMo 12B – small language model, fine-tuned by us.

Main metric: **Average per-field F1 score across test set documents (n=400)**

Separate overall averages calculated for:

1. 7 fields extracted by Meteor tool
2. 13 fields extracted by the fine-tuned language models

Table 1

Per-field evaluation results (average F1 scores) for different metadata extraction methods. We have calculated the averages separately for the seven fields supported by Meteor and all 13 fields extracted by language models.

	baseline	Meteor	Qwen2.5 0.5B	Gemma3 4B	Mistral NeMo 12B
language	0%	97%	99%	99%	**100%**
title	0%	40%	74%	87%	**92%**
alt_title	76%	-	83%	**87%**	**87%**
creator	17%	65%	79%	87%	**89%**
publisher	8%	8%	76%	80%	**86%**
year	14%	72%	89%	91%	**93%**
e-isbn	63%	83%	89%	**92%**	90%
e-issn	78%	86%	92%	92%	**94%**
p-isbn	72%	-	92%	**92%**	**92%**
p-issn	83%	-	96%	**96%**	**96%**
doi	91%	-	99%	**100%**	**100%**
type_coar	0%	-	80%	82%	**89%**
average (7 fields)	26%	65%	86%	90%	**92%**
average (13 fields)	42%	-	87%	90%	**92%**

Saturation point reached?
Larger LLMs won't help much

Beyond Meteor: BIBRA

Meteor provides a platform for quick experimentation, but we need to do more:

- User interface for metadata experts
- Support for basic extraction + turning the result into a bibliographic record (e.g. MARC, BIBFRAME, Dublin Core)
- REST API with the above functionality
- Modular backends (many different methods), as in Annif
- Test and evaluate methods using CLI, as in Annif

→ First step: BIBRA toolkit that we just started to develop

Demonstration

Extraction of metadata from PDF publications part 2

Prototype system

BIBRA tool, new user interface

Behind the scenes:
- REST API
- modular toolkit, to be extended

🤔 Is this civilized AI?

5 points

for building civilized AI

1. Use AI to make the world better

AI is not the goal

Human in the center

What positive changes can we make via AI?

2. Use the smallest AI that works

AIs can solve data-oriented challenges

Data in

- **text**
- code
- JSON
- image
- sound
- video

system

Data out

- **text**
- code
- JSON
- image
- sound
- video

Try lighter solutions first!

1. Rules, heuristics, good old code
2. Traditional machine learning - lots of training data
3. Small, specialized language models
4. General purpose LLMs

Rise of "small" language models (~10B or less)

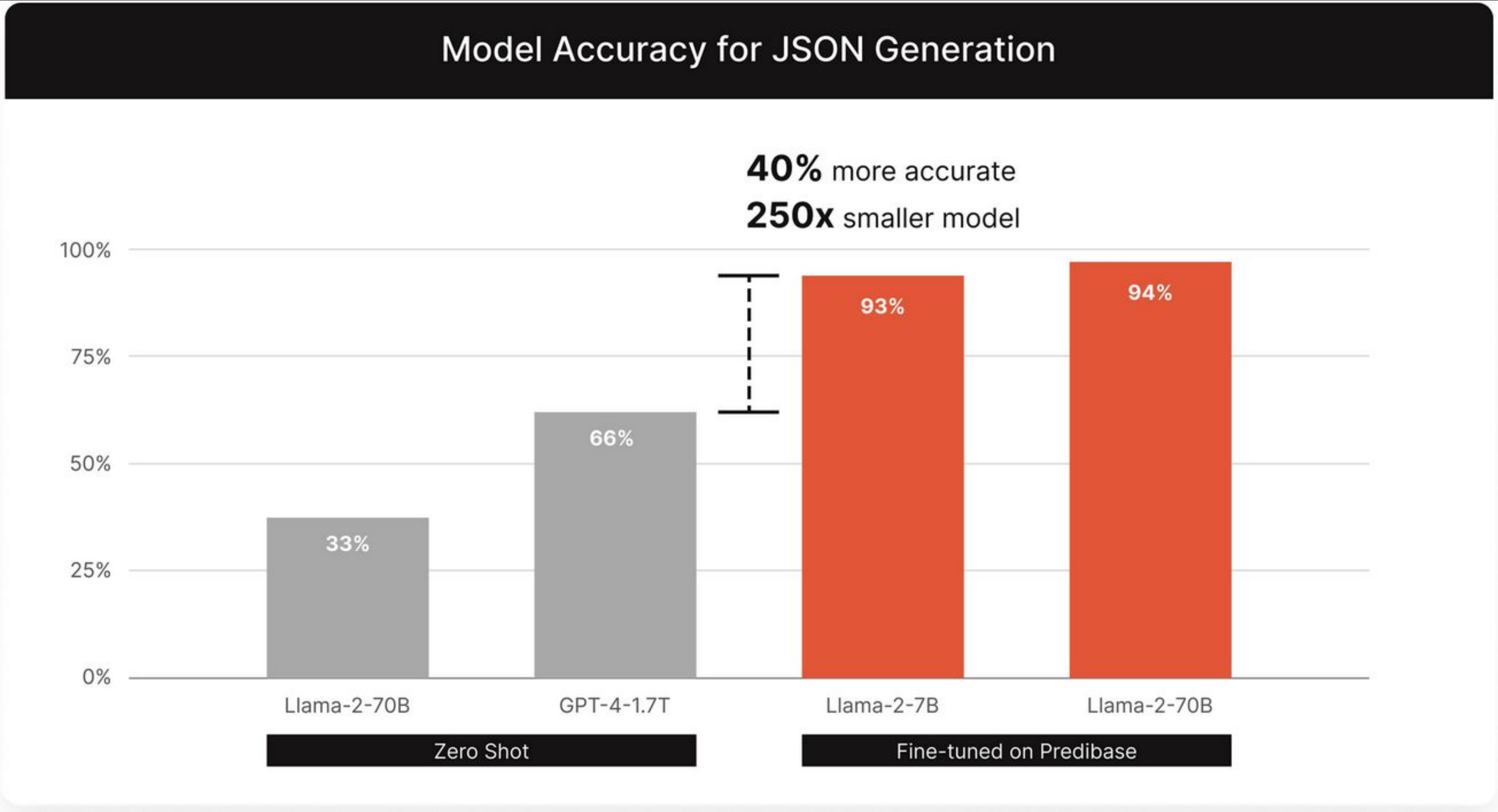
Gemma

PORO

MISTRAL
AI_

PHI-3
Microsoft

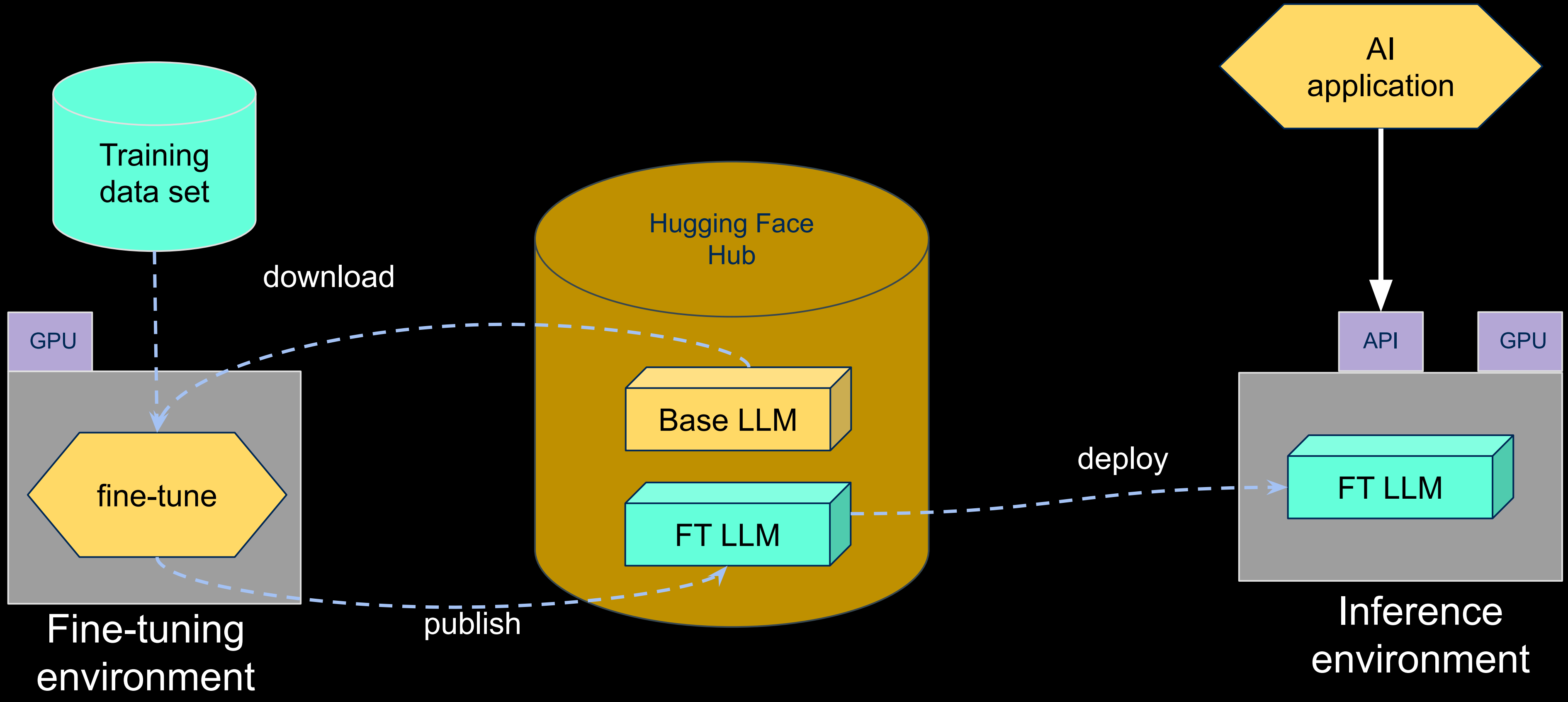
Smaller can be better and faster



Example from Predibase

3. Don't depend on corporate AI

DIY, modular AI systems development



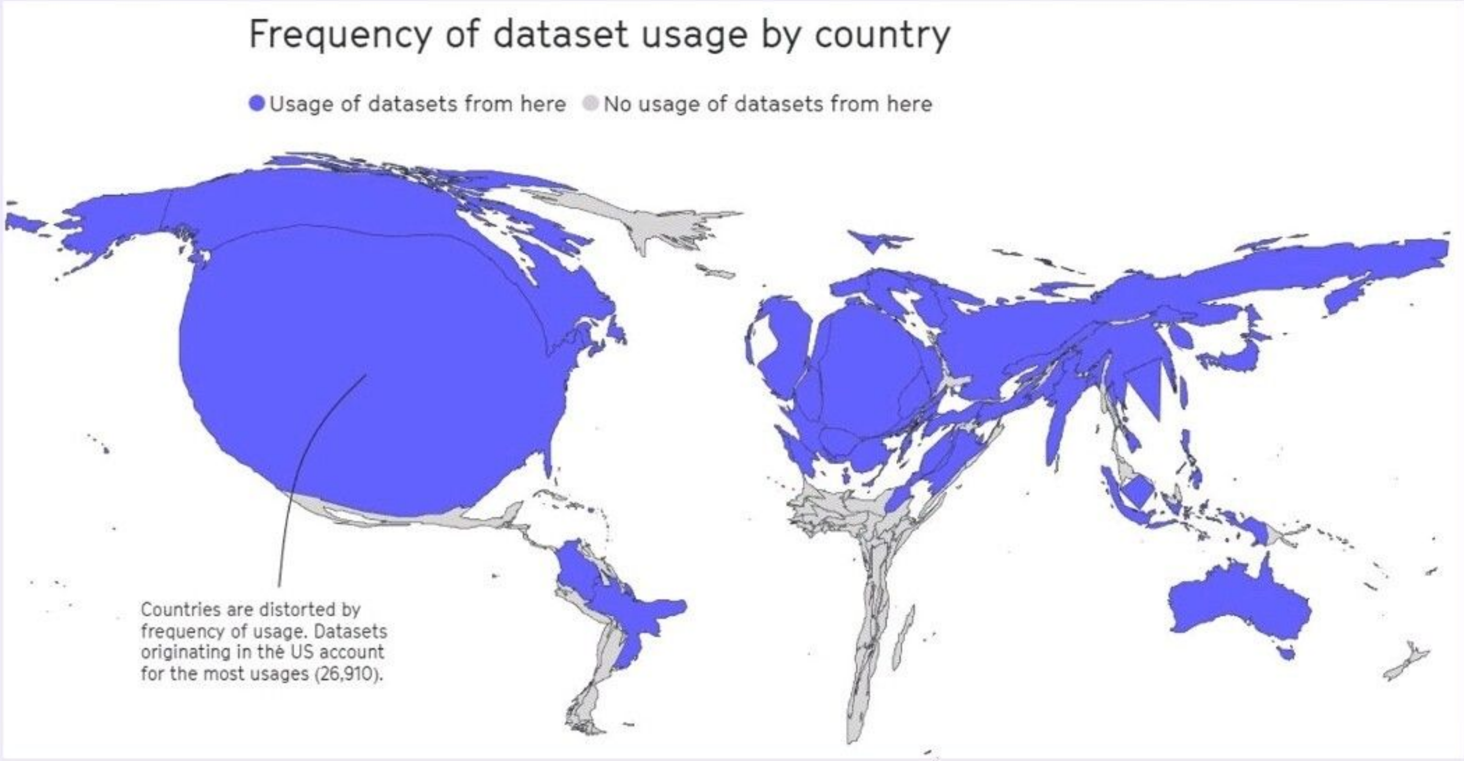
All pieces are interchangeable,
including the base model

4. Evaluate & Create data sets

Common questions about AI

- is it biased?
- does it hallucinate?
- can you trust it?
- does it know my language and culture?

The World Map according to the data AI sees

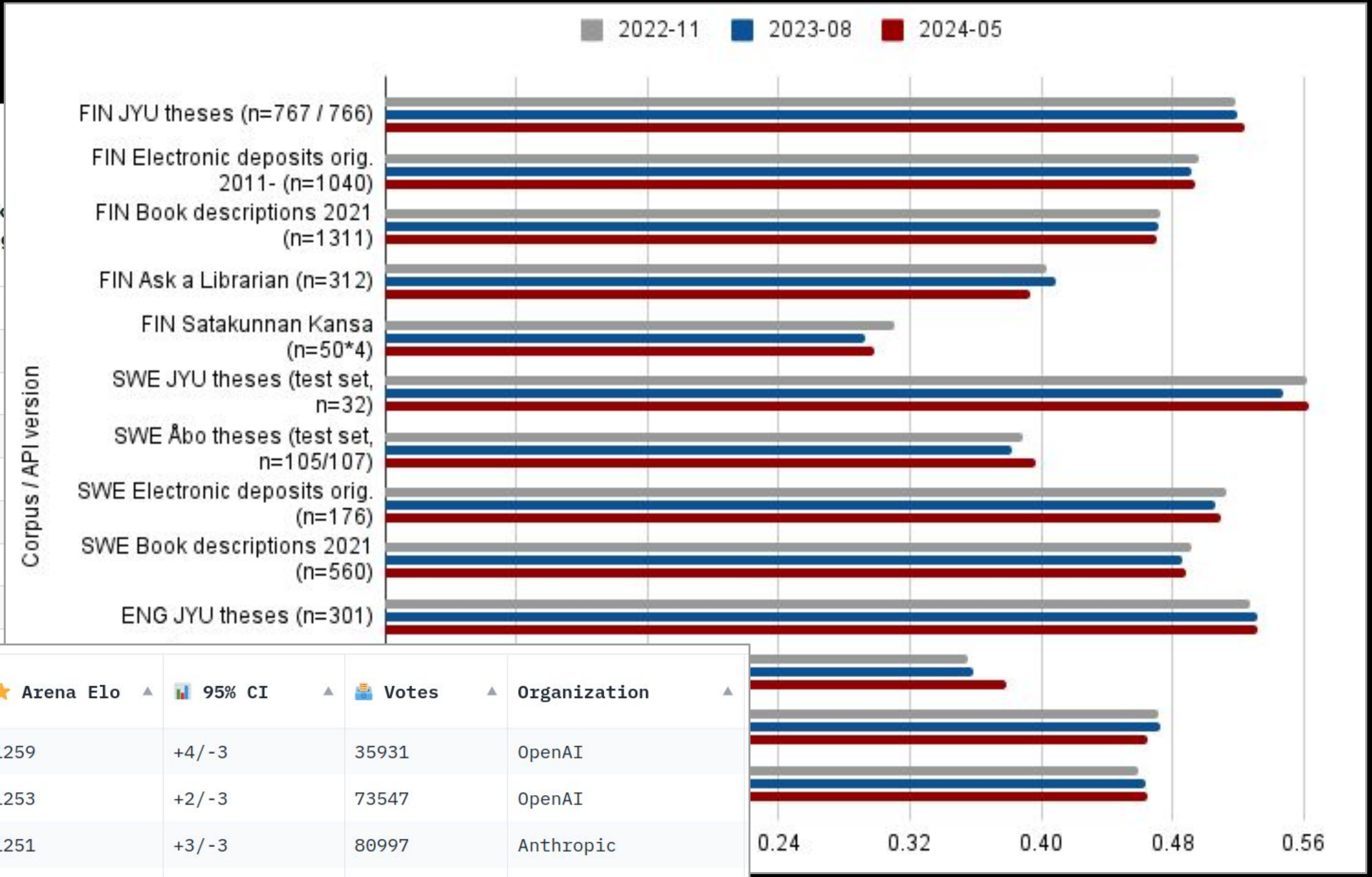


Sources
Research by: Koch, Denton, Hanna, and Foster (2021)
Visual by: The Mozilla Internet Health Report 2022

LLM Metadata Extraction model evaluation

language	field	axolotl-fine-tune-MiniChat-1	axolotl-fine-tune-Nous-Hermes-2-Mistral-7B-DPO	axolotl-fine-tune-stablelm-2-zephyr-1_6b	baseline-null	ludwig-fine-tune-zephyr-7b	meteor	openai-gpt35-turbo-16k promptin
eng	dc.contributor.author	0.7458	**0.8136**	0.7797	0.0508	0.7966	0.5763	0.7966
eng	dc.date.issued	0.9153	**0.9492**	0.8983	0	0.9322	0.7119	0.8136
eng	dc.identifier.isbn	0.6271	0.8305	0.8136	0.4746	**0.8814**	0.7966	0.6102
eng	dc.language.iso	0.8983	**1.0000**	0.9661	0	0.9831	**1.0000**	0.5254
eng	dc.publisher	0.5593	**0.7627**	0.6780	0.0169	0.6780	0.0508	0.5254
eng	dc.relation.eissn	0.7627	**0.9492**	0.9322	0.7288	0.9153	0.8475	0.8983
eng	dc.title	0.8644	**0.8983**	0.8475	0	0.8644	0.5763	0.8475
fin	dc.contributor.author	0.7538	**0.9077**	0.7692	0.2	0.7692	0.7077	0.7385
fin	dc.date.issued	0.9231	0.9231	0.923				
fin	dc.identifier.isbn	0.7077	**0.9077**	0.876				
fin	dc.language.iso	0.9846	0.9846	0.984				
fin	dc.publisher	0.6	**0.8615**	0.815				
fin	dc.relation.eissn	0.8	0.9385	0.907				
fin	dc.title	0.5538	0.7692	0.630				

Annif Subject Indexing evaluation



Rank* (UB)	🤖 Model	⭐ Arena Elo	📊 95% CI	🗳️ Votes	Organization
1	GPT-4-Turbo-2024-04-09	1259	+4/-3	35931	OpenAI
2	GPT-4-1106-preview	1253	+2/-3	73547	OpenAI
2	Claude 3 Opus	1251	+3/-3	80997	Anthropic
2	Gemini 1.5 Pro API-0409-Preview	1250	+3/-3	39482	Google
2	GPT-4-0125-preview	1247	+3/-2	67354	OpenAI
6	Llama-3-70b-Instruct	1210	+3/-4	53404	Meta
6	Bard (Gemini Pro)	1209	+5/-6	12387	Google
7	Claude 3 Sonnet	1201	+2/-3	78956	Anthropic
9	Command R+	1191	+3/-3	44988	Cohere
9	GPT-4-0314	1190	+3/-4	52079	OpenAI

LMSYS Chatbot Arena Leaderboard

5. Be open and transparent

Open code, open models, open and inclusive communication

Annif tutorial outline

This page is an overview of Annif tutorial contents. There are video-only lec
some coding, and those with 📖 are for reading only.

🎞️ Introduction and overview

annif tutorial

Introduction to the online hands-on tutorial

1640 THE NATIONAL LIBRARY OF FINLAND

ZBW

Annif tutorial

Annif code on GitHub

annif

DOI 10.5281/zenodo.8262313 | License Apache 2.0 | container image quay.io | CI/CD passing | codecov 100% | Scrutinizer 9.11 | maintainability A | openssf scorecard 8.2 | codebeat A | CodeQL passing | quality gate passed | docs passing | code style black

Annif is an automated subject indexing toolkit. It was originally created as a statistical automated indexing tool that used metadata from the Finna.fi discovery interface as a training corpus.

This repo contains a rewritten production version of Annif based on the prototype. It is a work in progress, but already functional for many common tasks.

Finto AI is a service based on Annif; see the source code for Finto AI.

Fulltext

Fulltext data sets contain both the actual contents (often digitized and OCRd), as well as metadata about the documents.

- Books — Copyright free books that the National Library has digitised from its collections.
- Classics Library — A collection of classic Finnish fiction from 19th and 20th centuries.
- Collection Catalogues — Digitized catalogues and card files of the National Library collections. Collections are not fully catalogued in the library databases, hence the old card files and catalogues can provide supplemental information on the collections.
- Digi collection texts and metadata — Metadata of digitized collections texts and metadata
- Digitalia data packages
- Dissertations of the Royal Academy of Turku — This collection contains 4173 digitized dissertations that were defended at the Royal Academy of Turku between 1642 and 1828. The collection also includes a number of Pehr Kalm's dissertations.
- Ephemera Collection — A digitised collection of ephemera from the legal deposit collections of the National Library of Finland. Subject matters include tourism, protection of animals, war-time rationing, women's movement, etiquette, sports, board games and vehicles. Publication dates range from early 19th century to 1944.

Datasets in NLF Data Catalog

Models 7 — Sort: Recently updated

- NatLibFi/FintoAI-data-YSO — Text Classification • Updated about 22 hours ago
- NatLibFi/FintoAI-data-KAUNO — Text Classification • Updated 1 day ago • ♡ 1
- NatLibFi/HogwartsSortingHat-fastText — Text Classification • Updated 14 days ago
- NatLibFi/FintoAI-data-YKL — Text Classification • Updated 23 days ago
- NatLibFi/stablelm-2-zephyr-1_6b-meteor — Updated Mar 28
- NatLibFi/Nous-Hermes-2-Mistral-7B-DPO-met… — Updated Mar 20
- NatLibFi/MiniChat-1.5-3B-meteor-GGUF — Updated Mar 1 • ⤓ 9

Datasets 4 — Sort: Recently updated

- NatLibFi/Finna-metadata — Updated 31 minutes ago • ⤓ 3
- NatLibFi/Finna-JOKA-images — Preview • Updated about 22 hours ago
- NatLibFi/Finna-HKM-images — Preview • Updated about 22 hours ago
- NatLibFi/Annif-corpora private — Updated Jan 10

Models and Datasets in Hugging Face Hub

KANSALLISKIRJASTO
NATIONALBIBLIOTEKET

Thank you!

Osma Suominen
osma.suominen@helsinki.fi

finto Skosmos fintoai annif