



Generative AI-Enhanced
Knowledge Management in Business

Generative AI for Smarter Knowledge Work: From Ideas to Business Value

Haaga-Helia campus + online
September 17th, 2025

[Check the whole agenda \(link attached\)](#)



Co-funded by
the European Union



Haaga-Helia



UNIVERSITY OF HELSINKI



Tampere University

In collaboration with





Generative AI-Enhanced
Knowledge Management in Business

Analyzing structured business documents and audio streams using LLMs

Lidia Pivovarova, Yiheng Wu, Roman Yangarber

Helsinki University



Co-funded by
the European Union



Haaga-Helia



UNIVERSITY OF HELSINKI

 Tampere University

AI-Powered Document Parsing

Introduction & Motivation

Need for Structured Data

Unstructured documents dominate business workflows:

- doc, PDF, ...
- emails
- ...

extracting structured data enables automation and analytics.

Limitations of Manual Processing

- slow,
- error-prone,
- expensive
- not scalable.

AI as Solution

Modern pipelines integrate

- layout analysis,
- OCR,
- large language models

→ scalable document understanding.

From PDF Input to Structured Data

PDF Ingestion:

Raw documents enter system as PDF files, with mixed content:

- text,
- tables,
- images.

Layout Parsing & Classification:

Machine learning models identify

- text blocks, headings,
 - tables,
 - figures
- for structure segmentation.

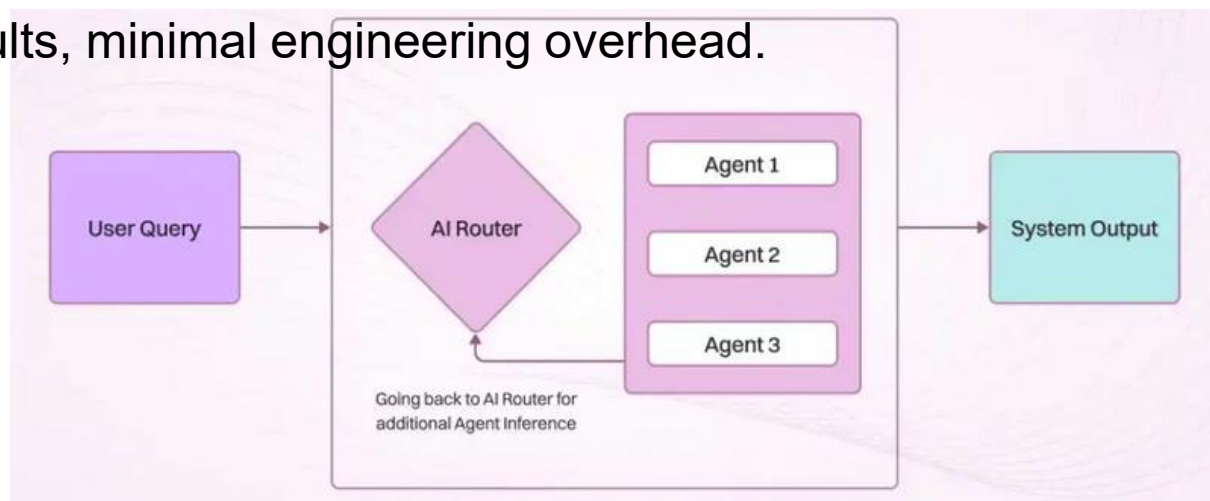
Final Data Formatting:

Reorganize segmented and recognized content into structured formats
→ for downstream analysis.

Method A: End-to-End Processing

Single Commercial Model with Agent Orchestration

- **Pipeline:**
 - Input PDF sent directly to powerful commercial model, no intermediate processing.
- **Agent-Controlled Flow:**
 - Orchestrator agent manages API calls, error handling, and structured result formatting.
- **Rapid Output:**
 - Fast results, minimal engineering overhead.



Method B: Multi-Stage Processing

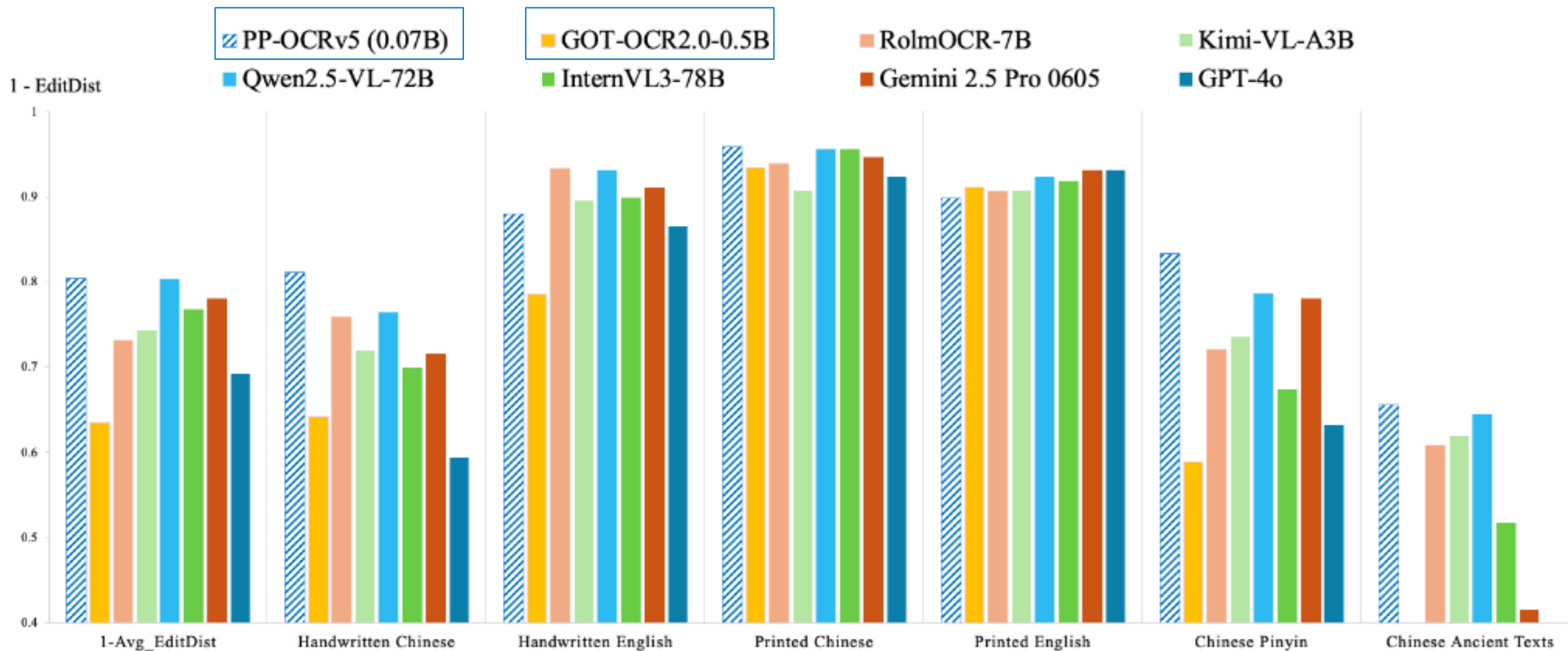
Small Layout Model → OCR → LLM

1. **Small Layout Model:**
 - Lightweight model identifies document layout / structure → headings, tables, sections.
2. **OCR:** optical character recognition
 - Extract text from identified regions, using OCR — optical character recognition.
3. **Large Language Model:**
 - Pass extracted text → LLM for contextual understanding → **structured** output

OCR evaluation

smaller models

Text OCR Evaluation



THE WALL STREET JOURNAL.

English Edition | Print Edition | Video | Audio | Latest Headlines | More

Subscribe

Sign In

LIMITED-TIME SALE

Latest World Business U.S. Politics Economy Tech Markets & Finance Opinion Arts Lifestyle Real Estate Personal Finance Health Style Sports

Advertisement



Track your stocks, ETFs, dividends & more all in one place.
Sharesight Investment Tracker

Read More >

2. Advertisement Section

- **Location:** Below main navigation
- **Content:** Sharesight Investment Tracker ad - "Track your stocks, ETFs, dividends & more all in one place"
- **Call-to-action:** "Read More" button

3. Main Content Area

Left Column - Primary News Stories

Lead Story 1

- **Headline:** "South Korea Says U.S. Agrees to Release Workers Detained in Hyundai Raid"
- **Subheading:** "Seoul's presidential office said a plane will be sent to repatriate the workers taken into custody in Georgia"

South Korea Says U.S. Agrees to Release Workers Detained in Hyundai Raid

Seoul's presidential office said a plane will be sent to repatriate the workers taken into custody in Georgia.

958 4 min read

Trump Said RFK Jr. Could 'Go Wild.' The Health Secretary Is Testing the Limits.

Kennedy's disruption of the CDC, vaccine recommendations and other policies drew a Senate rebuke. Trump is standing by him, for now.

953 6 min read



The Billionaires Fueling the Quest for Longer Life

A network of ultrawealthy investors including Peter Thiel and Sam Altman are infusing the longevity industry with cash and making big bets on where the science is headed.

Opinion



America's 'Buy Now, Pay Later' Economy

By Allysia Finley



At Long Last, Clarity on Climate

By Steven E. Koonin



Can Trump's 'War Department' Win a War?

By The Editorial Board



The Trump Summer Jobs Stall

By The Editorial Board



Putin Sends Trump Two More Bad Signals

By The Editorial Board

THE WALL STREET JOURNAL.

English Edition | Print Edition | Video | Audio | Latest Headlines | More

Latest World Business U.S. Politics Economy Tech Markets & Finance Opinion Arts Lifestyle Real Estate Personal Finance Health Style Sports

Advertisement



Track your stocks, ETFs, dividends & more all in one place.
Shareright Investment Tracker

Read More

South Korea Says U.S. Agrees to Release Workers Detained in Hyundai Raid

Seoul's presidential office said a plane will be sent to repatriate the workers taken into custody in Georgia.
958 4 min read

Trump Said RFK Jr. Could 'Go Wild.' The Health Secretary Is Testing the Limits.

Kennedy's disruption of the CDC, vaccine recommendations and other policies drew a Senate rebuke. Trump is standing by him, for now.
958 6 min read



The Billionaires Fueling the Quest for Longer Life

A network of ultrawealthy investors including Peter Thiel and Sam Altman are infusing the longevity industry with cash and making big bets on where the science is headed.

Opinion

America's 'Buy Now, Pay Later' Economy
By Alyssa Finley

At Long Last, Clarity on Climate
By Steven E. Koonin

Can Trump's 'War Department' Win a War?
By The Editorial Board

The Trump Summer Jobs Stall
By The Editorial Board

Putin Sends Trump Two More Bad Signals
By The Editorial Board

THE WALL STREET JOURNAL.

English Edition | Print Edition | Video | Audio | Latest Headlines | More

Latest World Business U.S. Politics Economy Tech Markets & Finance Opinion Arts Lifestyle Real Estate Personal Finance Health Style Sports

Advertisement



Track your stocks, ETFs, dividends & more all in one place.
Shareright Investment Tracker

Read More

South Korea Says U.S. Agrees to Release Workers Detained in Hyundai Raid

Seoul's presidential office said a plane will be sent to repatriate the workers taken into custody in Georgia.
958 4 min read

Trump Said RFK Jr. Could 'Go Wild.' The Health Secretary Is Testing the Limits.

Kennedy's disruption of the CDC, vaccine recommendations and other policies drew a Senate rebuke. Trump is standing by him, for now.
958 6 min read



The Billionaires Fueling the Quest for Longer Life

A network of ultrawealthy investors including Peter Thiel and Sam Altman are infusing the longevity industry with cash and making big bets on where the science is headed.

Opinion

America's 'Buy Now, Pay Later' Economy
By Alyssa Finley

At Long Last, Clarity on Climate
By Steven E. Koonin

Can Trump's 'War Department' Win a War?
By The Editorial Board

The Trump Summer Jobs Stall
By The Editorial Board

Putin Sends Trump Two More Bad Signals
By The Editorial Board

THE WALL STREET JOURNAL.

English Edition | Print Edition | Video | Audio | Latest Headlines | More

Latest World Business U.S. Politics Economy Tech Markets & Finance Opinion Arts Lifestyle Real Estate Personal Finance Health Style Sports

Advertisement



Track your stocks, ETFs, dividends & more all in one place.
Shareright Investment Tracker

Read More

South Korea Says U.S. Agrees to Release Workers Detained in Hyundai Raid

Seoul's presidential office said a plane will be sent to repatriate the workers taken into custody in Georgia.
958 4 min read

Trump Said RFK Jr. Could 'Go Wild.' The Health Secretary Is Testing the Limits.

Kennedy's disruption of the CDC, vaccine recommendations and other policies drew a Senate rebuke. Trump is standing by him, for now.
958 6 min read



The Billionaires Fueling the Quest for Longer Life

A network of ultrawealthy investors including Peter Thiel and Sam Altman are infusing the longevity industry with cash and making big bets on where the science is headed.

Opinion

America's 'Buy Now, Pay Later' Economy
By Alyssa Finley

At Long Last, Clarity on Climate
By Steven E. Koonin

Can Trump's 'War Department' Win a War?
By The Editorial Board

The Trump Summer Jobs Stall
By The Editorial Board

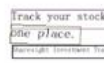
Putin Sends Trump Two More Bad Signals
By The Editorial Board

THE WALL STREET JOURNAL.

English Edition | Print Edition | Video | Audio | Latest Headlines | More

Latest World Business U.S. Politics Economy Tech Markets & Finance Opinion Arts Lifestyle Real Estate Personal Finance Health Style Sports

Advertisement



Track your stocks, ETFs, dividends & more all in one place.
Shareright Investment Tracker

Read More

South Korea Says U.S. Agrees to Release Workers Detained in Hyundai Raid

Seoul's presidential office said a plane will be sent to repatriate the workers taken into custody in Georgia.
958 4 min read

Trump Said RFK Jr. Could 'Go Wild.' The Health Secretary Is Testing the Limits.

Kennedy's disruption of the CDC, vaccine recommendations and other policies drew a Senate rebuke. Trump is standing by him, for now.
958 6 min read



The Billionaires Fueling the Quest for Longer Life

A network of ultrawealthy investors including Peter Thiel and Sam Altman are infusing the longevity industry with cash and making big bets on where the science is headed.

Opinion

America's 'Buy Now, Pay Later' Economy
By Alyssa Finley

At Long Last, Clarity on Climate
By Steven E. Koonin

Can Trump's 'War Department' Win a War?
By The Editorial Board

The Trump Summer Jobs Stall
By The Editorial Board

Putin Sends Trump Two More Bad Signals
By The Editorial Board

Trade-offs in Document AI: Balancing Speed, Accuracy, and Cost

End-to-End Models:

Fast and simple to deploy, but expensive and less customizable.

Multi-Stage Pipelines:

More interpretable, cost-efficient, but for some tasks small models are not as good.

Hybrid Strategy:

Combine strengths at different stages:

— use smaller models when performance is adequate, commercial models otherwise

Deployment Scenarios & Use Cases

Examples of Industry Applications of Document AI

- **Finance:**
 - Automate extraction of invoices, receipts, regulatory filings, ...
→ for faster auditing and compliance.
- **Healthcare:**
 - Parse medical records, prescriptions, lab reports
→ for patient data integration.
- **Legal:**
 - Contract analysis, clause extraction, due diligence support.

Analysis of Spoken Data

Challenges

- Existing models are black boxes,
- They are hard to improve if something not working,
- Many are commercial

ASR: Automatic Speech Recognition

Finnish Landscape

1. Increasing demand for high-quality Finnish ASR across domains.
2. Large-scale multilingual and monolingual datasets enable progress.
3. Rapid improvement in open-source and commercial solutions

Finnish Speech Resources: domain-specific

- Broadcast & Parliament Corpora:
 - YLE [Broadcast](#) Corpus, RTVA archives, Aalto Finnish [Parliament](#) ASR Corpus (3K+ hours, 449 speakers)
- MOSEL & Subsets:
 - [Wikipedia](#), [audiobooks](#), VoxPopuli (14K+ hours [parliament](#) speech), CSS10, FLEURS (Google), [YouTube](#).
- Lahjoita Puhetta:
 - 3K hours colloquial Finnish, 20K+ speakers, [conversational domain](#).

Finnish Speech Resources: domain-specific

- Training domains can be very far from particular business needs
 - results reported on general tasks
 - may not be valid for your specific domain
 - In your domain: domain-specific terminology, phraseology, proper names (companies, addresses)
- Models pre-trained on general resources will not perform well
- Need domain-specific work:
 - evaluation !!!
 - data
 - fine-tuning

Model Performance on Finnish Speech Recognition

1. **Wav2Vec2 Variants:**

Base/large/xlarge models pretrained on VoxPopuli, Lahjoita Puhetta

- Word error rate (WER) $\approx$ 16–25%

2. **Whisper Models:** Whisper v2/v3 — largest multi-lingual model

- WER $\approx$ 14–24%
- Finnish fine-tuned Whisper-large-finnish-v3
- WER $\approx$ 8.2%

Observations: Getman models (Aalto U) transcribe verbatim;
Whisper tends to normalize transcription into written form

Ground truth	Tänään meillä on teemana menetetty hammas. Kumppanina meillä tässä tässä ensimmäisessä teemassa on Dentsply Sirona, ja ja näissä videoissa ja kuvissa implanttimerkki on Osseospeed EV, mitä käytetään. Osa varmaan on katsonutkin jo pohjustuskeskustelun tähän tähän vodcastiin, joka käytiin hammasteknikkomestari Aki Lindenin kanssa. Jos ei ole katsonut, niin kannattaa käydä se se vielä vilkaisemassa. Siinä keskustellaan aika paljon implanttiprotetiikasta, näin yleishammaslääkärin ja sitten hammasteknikon näkökulmista.
Getman	tänään meli on teemana menetetty hammas kumppanina meillä tässä tässä ensimmäisessä teemassa on dentsplai sirona ja ja näissä videoissa ja kuvissa inplanttimerkki on osseusbid eevee mitä käytetään osa varmaa on kattonukki jo pohjuistuskeskustelun tähän tähän vodkästiin joka käytti hammastekniikkomestari akilindeenin kanssa jos ei oo kattonu nii kannattaa käydä se se vielä vilkasemassa siinä keskustellaa aika paljo inimplanttiprotetiikasta näi yleishammaslääkäri ja sitte hammastekniikon näkökulmista
Wispe r	Tänään meillä on teemana menetetty hammas. Kumppanina meillä tässä [MISSING tässä] ensimmäisessä teemassa on Dentsply Sirona ja [MISSING ja] näissä videoissa ja kuvissa implanttimerkki on Osseospid EV mitä käytetään. Osa varmaan on katsonutkin jo pohjoistuskeskustelun tähän [MISSING tähän] vodcastiin joka käytiinhammastekniikkomestari Aki Lindenin kanssa. Jos ei ole katsonut, niin kannattaa käydä se [MISSING se] vielä vilkasemassa. Siinä keskustellaan aika paljon implaantiprotetiikasta näiden yleishammaslääkärien ja sitten hammastekniikon näkökulmista.

Future directions

1. Strengths:

- Broad dataset coverage and availability of open-source benchmarks.

2. Challenges:

- Name recognition, colloquial language, domain terminology

3. Future directions:

- Domain adaptation
- Data augmentation: e.g., synthetic data using TTS — text-to-speech
- Fine-tuning of multilingual models

References

- [1] PaddleOCR 3.0 Technical Report. Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, Yanjun Ma. <https://arxiv.org/abs/2507.05595>. 2025.
- [2] Qwen. An, Y., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., & Qiu, Z. (2024). *Qwen2.5 Technical Report* (arXiv preprint arXiv:2412.15115). <https://arxiv.org/abs/2412.15115>
- [3] OpenAI. Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A. J., Welihinda, A., Hayes, A., Radford, A., Ćwajda, A., Baker-Whitcomb, A., Beutel, A., Borzunov, A., Carney, A., Chow, A., Kirillov, A., Nichol, A., ... Malkov, Y. (2024). *GPT-4o System Card* (arXiv preprint arXiv:2410.21276). <https://arxiv.org/abs/2410.21276>
- [4] <https://claude.ai>.

Advancing AI Pipelines

- **Hybrid Models:**
 - Combining modular and end-to-end approaches for efficiency and accuracy.
- **Reinforcement Learning:**
 - Leveraging RL-based methods like Infinity-Parser to optimize layout fidelity.
- **Multilingual & Complex Layouts:**
 - Expanding OCR and parsing to support more languages, more complex layouts

Thank you!

Questions?

Contacts:

Roman.Yangarber@helsinki.fi

Lidia.Pivovarova@helsinki.fi



Co-funded by
the European Union



How to join right now?

- Join our LinkedIn group <https://www.linkedin.com/company/gen-ai-enhanced-km/>
- Read our newsletter
- ... and follow the updates



If you represent small or medium-sized company and you are interested in the following:

- AI advisory
- Proof-of-Concept development for your AI idea
- Master students
- Hackathon

Take a look at the Finnish AI Region EDIH (European Digital Innovation Hub) services

<https://www.fairedih.fi/en/services/> and apply <https://www.fairedih.fi/en/contact/>



FAIR
Finnish AI Region | EDIH



Co-funded by
the European Union



Haaga-Helia



Tampere University