



Region Halland

The AI Implementation Spectrum

Strategies for Sustainable and Scalable Adoption

White Paper within the project Data-Driven Organizations / SAIL

Version: 1.0 | **Date:** 2025-12-16 | **Author:** Aixia AB and Region Halland



Aixia AB | +46 31 762 02 40



Region Halland | +46 35 13 48 00

Table of contents

1. Executive Summary	3
2. Background	3
3. Methodology	4
4. Training AI Models From Scratch	4
5. Fine-Tuning Existing Models	5
5.1. <i>Data</i>	5
5.2. <i>Compute Requirements</i>	6
5.3. <i>Skilled Practitioners</i>	6
5.4. <i>Example Use Case: Fine-Tuning Vision Models for Object Detection Tasks</i>	7
6. Leveraging Pre-Built AI Solutions	7
6.1. <i>Vision Tasks</i>	7
6.2. <i>Language Models (APIs and Open Source LLMs)</i>	8
6.3. <i>Domain-Specific AI and Data-Driven Tools</i>	8
7. Alternative Approaches Beyond Training	9
7.1. <i>Retrieval-Augmented Generation (RAG)</i>	9
7.2. <i>Agentic AI and Prompt Engineering</i>	9
7.3. <i>Hybrid Intelligence</i>	10
8. Choosing the Right Path For Your Organization	10
8.1. <i>Decision Framework: When to Train, Fine-Tune, or Reuse</i>	10
8.2. <i>Making the Decision: Costs, Capabilities, and Sustainability</i>	10
9. Conclusion	11
Appendix A – The project “Data-driven organisations – Best practices for operationalisation of AI in Sweden	12
Appendix B – Resources	14
Appendix C – Glossary	15



1. Executive Summary

Artificial Intelligence has become a strategic priority for Swedish organizations, yet many waste resources building solutions from scratch that already exist or purchasing costly tools for problems solvable in-house.

This whitepaper presents a fundamental reframing: AI implementation is not a single choice, but a spectrum of approaches. At one end lies building new models from the ground up—resource-intensive and rarely necessary. At the other end are integration techniques that orchestrate existing AI systems. In between lie fine-tuning, pre-built solutions, and hybrid approaches, each with distinct trade-offs.

The core insight is that sustainable AI adoption depends on knowing where to position yourself on this spectrum. Organizations that understand these options can avoid duplication, focus resources where they create value, and build capabilities that evolve rather than require constant rebuilding.

Key principles are: reuse existing models and tools first, fine-tune strategically when domain specificity demands it, train from scratch only for novel research problems, and integrate intelligently using techniques like RAG and agentic AI to enhance existing systems without additional training.

Developed through collaboration between Aixia and Region Halland within the national Data-Driven Organizations initiative, this framework provides leaders with a strategic foundation for deciding when to build, adapt, or reuse.

2. Background

Artificial Intelligence is no longer a distant frontier, yet the road from idea to production is often longer and more complex than expected. Many initiatives stall because resources are not used efficiently – organizations spend time and effort developing solutions from scratch that already exist elsewhere or buy costly solutions that can easily and cost-effectively be developed in-house.

This white paper, developed within the national initiative Data-Driven Organizations, aims to show that AI landscape is a myriad of open-source and proprietary solutions that organizations can reuse or adapt. Further details about the project can be found in *Appendix A*.

AI implementation should be seen as a spectrum of approaches. At one end lies the development of entirely new models from scratch, where every aspect of the implementation is crafted in-house. Further along the spectrum, organizations may fine-tune existing models to their own needs, or adopt ready-made solutions such as vision models, language models, or specialized AI tools. At the other end are techniques that go beyond training, such as retrieval-augmented generation or agent-based orchestration, which allow existing AI systems to be repurposed and combined for new tasks. Each of these approaches comes with different trade-offs. Building models from scratch offers full control but demands enormous resources. Fine-tuning enables adaptation but requires careful data preparation and evaluation. Pre-built models can accelerate adoption but may be less transparent or create dependencies on external vendors. And techniques like RAG and orchestration reduce training needs but shift the complexity into integration, data quality, and governance. By being aware of these trade-offs, organizations can make informed choices along the spectrum.



Ultimately, this perspective supports the DDO project's goal of making AI implementation more sustainable and scalable. By considering the full spectrum, organizations can avoid unnecessary duplication of effort, reuse what already works, and focus their energy on where AI creates the most value in their context.

This whitepaper is intended for leaders, strategists, and practitioners involved in digital transformation and AI adoption within Swedish organizations — across both public and private sectors. It is written for those who may not be AI researchers but who are responsible for shaping how AI is implemented and maintained in practice. The goal is to provide a strategic and practical framework for making informed decisions about when to build, adapt, or reuse AI solutions in a sustainable way.

3. Methodology

The methodology behind gathering the principles presented in this whitepaper is grounded in practical experience of AI specialists at Aixia from developing AI applications across multiple domains, including both computer vision and text-based generative AI. Through close collaboration with Region Halland, the work has been anchored in their concrete needs and organizational challenges, which helped develop a broader understanding of the roadblocks on implementing AI. Experimentation and hands-on implementation have been central, ensuring that the lessons learned are based on tested practice rather than theory alone, and additional supplement to the lessons learned came from many interviews with the researchers and practitioners at Region Halland and beyond.

4. Training AI Models From Scratch

Developing an AI model entirely from scratch is the most demanding form of AI implementation. It involves designing model architecture, collecting and cleaning large amounts of training data, defining loss functions, setting hyperparameters, and running extensive training loops — often iteratively, to reach acceptable performance. While this approach offers full control over the system's design and behaviour, it also requires substantial resources in terms of data, compute, expertise, and time. In practice, training models from scratch is most relevant in research contexts or in highly specialized domains where no suitable pre-trained models exist. The resulting models are typically smaller in scale and narrowly focused on a specific task, such as identifying rare medical conditions or analysing data from unique industrial sensors. It is worth mentioning that even in these unique cases there usually are pre-made models from other domains that could be reworked to fit these highly specialised tasks.

Training from scratch also comes with challenges that extend beyond the initial build phase. Data bias is a constant concern, as training data often reflects real-world imbalances that the model can inadvertently learn. Data drift — the gradual change in how real-world data looks or behaves — can cause performance to degrade over time, making retraining essential. Effective monitoring and governance are therefore needed to maintain reliability and safety once the model is deployed. Finally, large-scale training consumes significant computational resources and energy, raising both cost and sustainability concerns.

Because of these factors, developing models entirely from scratch is rarely feasible or necessary for most organizations. It remains primarily the domain of academic research and large AI labs, while



practical AI initiatives more often focus on adapting or reusing existing models. A notable example is Region Halland's development of a custom T5 model for ICD-10 code classification, a project that required substantial data preparation and experimentation over an extended period. The outcome demonstrated that training specialized models is possible, but also that it demands long-term commitment and is rarely the fastest or most sustainable route to production. For that reason, this whitepaper will not cover model training in detail but rather focus on the approaches that are more attainable — fine-tuning, reuse, and orchestration — which together make up the core of sustainable and scalable AI implementation.

5. Fine-Tuning Existing Models

Fine-tuning means adapting a pre-trained model to a specific task or domain instead of training from scratch. The process begins by selecting a suitable base model — one that already understands general patterns relevant to the problem. In language tasks, this might be a large language model trained on broad text corpora that already knows how to reason, follow instructions, and capture linguistic structure. In vision tasks, it could be a model that has learned to recognize edges, shapes, textures, and colours. By building on this prior knowledge, fine-tuning allows organizations to specialize a model for their own data with far fewer resources than full model training.

This technique significantly lowers the required time and effort for creating a model, but still requires significant labelled data, GPU resources, and skilled practitioners.

5.1. Data

One of the main challenges in fine-tuning is obtaining the right data. The quality and quantity of training data can make or break an AI project, and the process of collecting it rarely ends — models must continually be retrained to stay relevant as data and real-world conditions evolve. Training on too little or insufficiently varied data can lead to overfitting, where a model performs well on training examples but fails to generalize to new cases. To achieve robust performance, models need to be exposed to enough diverse and nuanced examples.

The amount of data needed also depends on the size of the model and the scope of its task. Larger models and broader problem definitions typically require much more data to reach stable and generalizable results. Conversely, narrower and well-defined use cases can often perform well with smaller, more focused datasets. Choosing to design multiple specialized models instead of one general-purpose system can therefore reduce the overall data requirements, as each model only needs examples relevant to its specific task.

In most cases, data also needs to be labelled — meaning that each piece of data (text, image, or tabular entry) must be accompanied by information about what it represents. In computer vision, this could mean marking the location of defects on a product image; in natural language processing, it could mean identifying whether a piece of text expresses a diagnosis, a symptom, or a treatment. Proper annotation is key, and it often requires domain expertise in addition to technical AI skills. This is what makes AI development inherently interdisciplinary — success depends on collaboration between AI specialists and subject-matter experts.

There are many tools and platforms available to assist with data annotation, but it remains a time-consuming process. To accelerate it, teams sometimes start with publicly available datasets, such as



those found on Kaggle or other open repositories. Another approach is to train an early version of the model and use it to assist in annotating additional data. However, strong human supervision is crucial — models should never be trained directly on labels they have generated themselves. This approach can speed up workflows, especially in vision tasks, but the final responsibility for quality always rests with human reviewers. In text-based projects, it is also common to use a larger, pre-trained model to support the labelling process, again under close human oversight.

5.2. Compute Requirements

Fine-tuning models also places significant demands on computing infrastructure. While training can technically be done on CPUs, the process becomes prohibitively slow. Modern AI models are designed to leverage GPUs, which can perform thousands of parallel operations and drastically shorten training times. The choice of GPU matters: larger models require cards with high memory capacity (VRAM), often 24 GB or more, while smaller or parameter-efficient models can be trained on mid-range GPUs with less memory. Insufficient GPU memory can limit the batch size or even prevent training altogether.

It is also important to note that the CPU and system memory still play a supporting role. A weak CPU or limited RAM can become a bottleneck, slowing down data loading and preprocessing even if the GPU itself is powerful. For more complex training pipelines, such as those involving multiple GPUs or distributed setups, careful coordination of CPU, GPU, and storage throughput is required to maintain efficiency.

These hardware requirements translate directly into costs and environmental impact — which is why efficient use of resources, such as through transfer learning or lightweight fine-tuning methods, is essential for sustainable AI implementation.

5.3. Skilled Practitioners

Fine-tuning requires a combination of AI-specific and general data engineering skills. At its core, it demands practitioners who can select a suitable base model, design an effective training script, and interpret key metrics to ensure that the model learns effectively. These tasks call for a strong understanding of machine learning fundamentals, model architecture, and optimization techniques.

Beyond pure AI expertise, success also depends on the team's surrounding technical environment. Efficient data pipelines, preprocessing workflows, and database management are essential — whether the data resides in relational systems like PostgreSQL and MySQL or in document-based solutions such as MongoDB and Elasticsearch. In many cases, fine-tuning work also involves integration with existing IT infrastructure and collaboration with data engineers who can ensure smooth data flow and version control. Additionally, knowledge of how to best utilize the available hardware is needed for a smooth and efficient process.

Choosing to use cloud environments, such as Azure or AWS, simplifies access to GPU resources and pre-configured training environments. However, they introduce their own learning curve and can be costly if not carefully managed. While cloud tools can reduce the need for low-level programming, they still require practitioners who understand both the AI workflows and the underlying cost-performance trade-offs.



5.4. Example Use Case: Fine-Tuning Vision Models for Object Detection Tasks

The common thread in choosing the fine-tuning strategy is specialization: adapting general-purpose intelligence to the nuances of a specific context, dataset, or organization.

Many practical AI applications are built around object detection, which means that the model locates and labels specific items within an image. Fine-tuned models perform these tasks faster and more consistently than manual inspection, improving efficiency and quality control.

One practical illustration of this principle comes from Aixia's work. A pre-trained object-detection model was fine-tuned on labelled images of used laptops to identify surface scratches, cracks, missing keys or ports, and screen defects. Deployed at the inspection station, the model flags issues in real time and records structured findings in the grading application. The result is more consistent condition grading and higher throughput, when compared with a human quality assessor.

The same underlying principle applies in other domains. In healthcare, models can assist in detecting irregularities in medical images, supporting clinical decision-making rather than replacing it. In administrative workflows, similar techniques can be used to automatically categorize scanned forms, invoices, or patient documents — reducing repetitive manual work while maintaining accuracy.

6. Leveraging Pre-Built AI Solutions

Not every AI project requires model training. In many cases, organizations can achieve strong results by reusing pre-built AI models and tools that already capture general capabilities, such as visual recognition, speech processing, or natural language understanding. These solutions can be accessed through open-source libraries, public model hubs, or cloud APIs, offering a faster and more cost-efficient path to implementation.

The advantage of using pre-built models lies in immediate functionality and reduced complexity. Instead of building a dataset, setting up GPU infrastructure, and managing training loops, organizations can integrate an existing model through an API or SDK, focusing their efforts on data flow, business logic, and system integration. This approach often allows AI to be deployed in weeks rather than months, making it highly suitable for organizations taking their first operational steps with AI.

It needs to be stated, however, that pre-built solutions have a performance ceiling. Because they are optimized for broad, generic use cases, they often struggle with domain-specific data, edge cases, or local-language nuances. As maturity grows, returns diminish until a more hands-on approach is needed.

6.1. Vision Tasks

Several pre-trained vision models can be used directly without retraining in scenarios where the task is general and there is no high demand for impeccable accuracy. We present a few examples of types of models in order to build an understanding of the types of tasks that can be performed by these out-of-the-box models:



YOLO (You Only Look Once) detects and classifies multiple objects in real time, useful for monitoring equipment, counting items, or verifying that safety gear is in place.

SAM (Segment Anything Model) separates objects from their background at pixel level, helping with annotation, visual documentation, or highlighting specific regions in images.

CLIP (Contrastive Language–Image Pretraining) links images and text, enabling visual search or automatic tagging of photographs and illustrations through simple text queries.

It is worth noting that these models might also be fine-tuned if the needs are more specific or the demand for accuracy is higher.

Running inference — that is, using an already trained model to make predictions — requires far fewer computational resources than training but still benefits from capable hardware. Even pre-built models perform best on systems equipped with modern GPUs, particularly for image-heavy or real-time applications. The topic of hardware requirements and performance trade-offs for vision inference is explored in more detail in white paper *Sustainable Image Inference In Practice*, developed within the national initiative *Data-Driven Organizations*.

6.2. Language Models (APIs and Open Source LLMs)

A wide range of general-purpose and specialized language models (LLMs) is now available, allowing organizations to use advanced text analysis and generation without developing models themselves. Some are broad in scope, while others are already fine-tuned for narrower tasks such as legal or medical text processing.

Models can be accessed through APIs from providers like OpenAI, Anthropic, or Cohere, offering immediate functionality for summarization, reasoning, and translation. Services like Perplexity combine LLMs with live web search, enabling up-to-date, reference-based insights. These solutions are quick to adopt but rely on external infrastructure, raising questions of data governance and cost.

Alternatively, open-source ecosystems such as Hugging Face and Ollama provide hundreds of models that can be hosted locally, ranging from lightweight systems with around 130 million parameters to large-scale models with hundreds of billions. Similarly to the vision inference described earlier, LLM inference requires far less compute power than training but still benefits from modern GPUs, especially for longer or multi-turn interactions.

Common use cases include generating summaries, classifying or structuring documents, simplifying administrative text, and supporting internal Q&A tools that make knowledge more accessible. These ready-made language models enable quick, low-barrier adoption while giving organizations the flexibility to later fine-tune or combine them with their own data sources.

6.3. Domain-Specific AI and Data-Driven Tools

Beyond general-purpose models, many tools focus on a single, well-defined job and are ready to use. Some tools are model-centred, meaning they include the core AI technique themselves—for instance, speech-to-text engines such as Whisper or Vosk that turn audio into written text, anomaly-detection libraries like PyOD or Merlion that spot unusual patterns in sensor or log data, and text-extraction software such as Tesseract that reads printed text from scans (optical character recognition, OCR). Other tools sit at the solution layer and act as wrappers: instead of exposing the



underlying models, they offer a simple service or API that combines several building blocks—for example, invoice-processing services that read a scanned document, understand its layout, and file the right numbers automatically; or managed speech services that take an audio file and return a transcript with speakers separated.

These tools can be deployed locally or in existing IT environments with moderate compute requirements. They provide an efficient way to automate recurring operational tasks — from transcribing meetings to monitoring system health — without the need for large datasets or model training, making them a practical entry point for organizations starting their AI journey.

7. Alternative Approaches Beyond Training

The last approach to be discussed is not about training new models, but about intelligent integration — connecting existing AI components with data sources, tools, and human expertise. These methods shift the focus from model accuracy to system effectiveness, enabling scalable, maintainable, and context-aware AI solutions.

7.1. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation, or RAG, is an approach that allows language models to access external information in real time instead of relying only on what they were trained on. Before generating a response, the model retrieves relevant documents or text passages from a connected database, knowledge base, or file system. This makes it possible to deliver accurate and up-to-date answers without retraining the model.

A practical example comes from Region Halland’s AI Portal, where a RAG setup is used to connect several cloud-based large language models to internal manuals, steering documents and other sources of information. This enables quick and efficient deployment while keeping data management simple and transparent. At the same time, Region Halland is exploring ways to extend the approach with fully on-premise RAG and inference systems for more sensitive or confidential information. This combination demonstrates the full strength of RAG: fast results from cloud infrastructure where appropriate, and tighter control when integrating more critical data sources.

7.2. Agentic AI and Prompt Engineering

Agentic AI refers to systems in which multiple AI components — often smaller models — cooperate to complete tasks through reasoning, planning, and interaction with external systems. Instead of one large model handling everything, each agent focuses on a specific capability, such as retrieving information, writing summaries, or updating a database. Through orchestration, these agents can call tools, query APIs, or submit data, moving beyond text generation to concrete actions.

In this setup, prompt engineering plays a key role. Rather than fine-tuning, smaller models can perform exceptionally well when given well-crafted instructions and the right context. Information can be injected dynamically into prompts — for example through RAG or API calls — ensuring that the model always has access to the most relevant data. While modern language models support large context windows, providing too much information at once often reduces focus and precision.



Dynamically injecting only what is needed keeps the model's attention concentrated on the current step, resulting in more accurate, interpretable, and reliable behaviour.

This approach makes agentic systems modular and maintainable: each model or agent is specialized, context-aware, and easier to oversee than one large, general-purpose model that tries to handle everything at once.

7.3. Hybrid Intelligence

In many successful AI projects, the AI component is just one part of a broader solution that also includes traditional software, rule-based logic, and human oversight. This form of hybrid intelligence leverages AI only where it adds clear value — for instance, handling ambiguous or variable inputs — while letting established systems manage predictable processes.

Examples include workflows where AI extracts or classifies data, and existing rule-based systems handle validation and decision-making; or maintenance systems where conventional predictive models monitor sensor data, and AI assists technicians by explaining or prioritizing alerts. In such setups, the focus shifts from replacing systems to enhancing them with adaptive intelligence, improving both performance and sustainability.

8. Choosing the Right Path For Your Organization

There is no single best approach to AI implementation. The right strategy depends on the organization's goals, resources, and level of digital maturity. This section outlines a simple decision framework to help identify when to train, fine-tune, or reuse models — and how to balance cost, governance, and sustainability along the way.

8.1. Decision Framework: When to Train, Fine-Tune, or Reuse

- Train from scratch only when tackling novel research problems or creating a model unavailable elsewhere. This path requires extensive data, compute resources, and expertise, and is generally relevant for research institutions or technology providers.
- Fine-tune when a general model exists but needs to adapt to your domain — for example, recognizing specific defects in production or understanding specialized language. It strikes a balance between customization and cost but demands labelled data and technical skill.
- Reuse or integrate when the task is already well covered by existing models or APIs, using, for example, RAG. This is the most time- and cost-efficient route for organizations looking for operational results.

8.2. Making the Decision: Costs, Capabilities, and Sustainability

The right place on the AI implementation spectrum depends on what the organization wants to achieve and what resources it can commit. Training, fine-tuning, reuse, and integration each come with different demands — not only in money, but also in time, data, and people.



A good starting point is to evaluate the available capacity. Do you have the data, compute power, and in-house skills to develop and maintain a fine-tuned model, or will an existing solution already cover most of your needs? Even small AI initiatives require ongoing effort to keep data organized, test performance, and update components.

In practice, a gradual approach often works best. Start by reusing or integrating existing models to test ideas and measure value. If performance gaps remain — for example, when the language or data is very specific — fine-tuning can then be introduced to improve results. This stepwise strategy limits cost and risk while helping teams build experience along the way.

Regulations, data quality, and internal processes will influence what approaches are possible. The goal is to find a setup that matches the organization's current capabilities while leaving room to grow as skills, infrastructure, and data maturity improve.

Finally, consider sustainability from the outset. Choose solutions that can evolve with minimal rework: smaller, modular models where possible; retraining only when data changes significantly; and architectures that allow individual components to be swapped or updated easily. Sustainable AI is not just energy-efficient — it's also simple, adaptable, and maintainable over time.

9. Conclusion

The journey toward sustainable AI adoption is not about choosing one approach, but about learning how to combine them. The examples and practices discussed in this paper show that real progress happens when organizations align ambition with capacity — reusing where possible, fine-tuning where needed, and integrating intelligently.

In doing so, AI becomes less of a project and more of a capability: something that evolves with the organization rather than being rebuilt for every new need. This is the essence of a data-driven organization — one that treats AI not as a destination, but as a continuously improving system of learning, adaptation, and shared value.

Appendix A – The project “Data-driven organisations – Best practices for operationalisation of AI in Sweden

This material has been produced as part of the Vinnova-funded project Data-driven organisations – Best practices for operationalisation of AI in Sweden (DDO), a project lasting just under two years with twenty participants from private sector, public sector, and academia. Together, they tackled issues concerning large- and small-scale operation of AI solutions and how to enable and use AI broadly across an organisation.

The work focused on three specific use cases: Local sustainable operation of AI, legal and technical prerequisites for effective infrastructure, and how to create the best conditions for keeping thousands of AI models in operation.

A compilation of all material produced within the framework of DDO is available on AI Sweden's website – <https://www.ai.se/en/project/data-driven-organizations-best-practices-ai-operationalization-sweden>.

The organisations that participated in DDO were:

- Aixia <https://aixia.se>
- Hewlett Packard Enterprise <https://www.hpe.com>
- Hopsworks <https://www.hopsworks.ai>
- IBM <https://www.ibm.com>
- Linköpings Universitet <https://liu.se>
- NetApp <https://www.netapp.com>
- Predli <https://www.predli.com>
- Proact <https://www.proact.se>
- RISE <https://www.ri.se>
- RedHat <https://www.redhat.com>
- Region Halland <https://www.regionhalland.se>
- Sahlgrenska University Hospital <https://www.sahlgrenska.se>
- Statistics Sweden (Statistiska Centralbyrån) <https://www.scb.se>
- The Swedish Tax Agency (Skatteverket) <https://www.skatteverket.se>
- Stormgrid <https://www.stormgrid.ai>
- The Swedish Transport Administration (Trafikverket) <https://www.trafikverket.se>
- Volvo Parts <https://www.volvogroup.com>
- Region Västra Götaland <https://www.vgregion.se>
- Santa Anna <https://www.santa-anna.se>
- AI Sweden <https://www.ai.se/en>

The project was funded by the participating organisations and Vinnova. AI Sweden is in part financed by the EU.

SAIL (Sustainable AI Infrastructure Lifecycle)

This white paper is produced within the SAIL use case, which is part of the DDO project. SAIL focuses on building cost-efficient, environmentally sustainable, and operationally effective infrastructure that supports the entire AI lifecycle – from exploration and development to training, deployment, inference, and long-term operations.

The goal is to enable organizations to adopt and scale AI sustainably by reducing costs, minimizing environmental impact, and ensuring that AI systems can operate and evolve over time without unnecessary complexity.

Key areas of focus include:

- Designing scalable and flexible AI infrastructure that grows with organizational needs
- Exploring hardware, cloud, and modular/shared platform options
- Creating practical guidelines for long-term AI operations
- Optimizing resources and reducing technical and administrative overhead

The outcome will be a validated model and actionable recommendations that help organizations of all sizes and maturity levels build, operate, and evolve AI solutions in a sustainable, efficient, and future-proof way.

In addition to this white paper, SAIL has also produced the following white papers:

- **MLOps on-prem without Kubernetes – A Faster Path to Production:** Demonstrates how an efficient on-prem MLOps pipeline can be implemented without Kubernetes, emphasizing simplicity, reproducibility, and rapid deployment to production.
- **Benchmarking Large Language Models for ICD-10 Code Generation:** Evaluates different hardware and software configurations to identify the most efficient and sustainable setup for running large language model inference when generating ICD-10 codes from clinical notes.
- **Sustainable Image Inference in Practice:** Investigates which hardware platforms deliver the best balance of performance, energy efficiency, and cost for running AI inference on radiology images in an on-premises healthcare environment.

Appendix B – Resources

AI Model Hubs and Ecosystems	URL
Hugging Face – Repository for open-source AI models, datasets, and tools for fine-tuning and deployment.	https://huggingface.co
Ollama – Framework for running and serving large language models locally on-premise.	https://ollama.ai
Kaggle – Platform for public datasets and AI competitions, useful for experimentation and benchmarking.	https://www.kaggle.com

Cloud AI Platforms and Providers	URL
Microsoft Azure AI – Cloud platform offering managed AI services, GPU compute, and deployment environments.	https://azure.microsoft.com/en-us/products/ai-services
AWS (Amazon Web Services) – Cloud infrastructure for machine learning, model hosting, and distributed training.	https://aws.amazon.com/machine-learning
Google Cloud Vertex AI – Platform for managing end-to-end machine learning workflows.	https://cloud.google.com/vertex-ai

Pre-Trained Vision Models and Tools	URL
YOLO (You Only Look Once) – Real-time object detection framework.	https://github.com/ultralytics/yolov5
SAM (Segment Anything Model) – Meta AI's model for universal image segmentation.	https://segment-anything.com
CLIP (Contrastive Language–Image Pretraining) – OpenAI's model linking images with text concepts.	https://github.com/openai/CLIP

Language Models (LLMs)	URL
OpenAI API (GPT models) – Text understanding and generation via hosted API.	https://platform.openai.com
Anthropic Claude – LLM optimized for safe and transparent reasoning.	https://www.anthropic.com
Cohere – APIs for text classification, embeddings, and generation.	https://cohere.com
Perplexity – RAG-enabled search assistant combining live retrieval with LLM reasoning.	https://www.perplexity.ai
AI Sweden GPT-SW3 (6.7B) – Swedish open-weight language model.	https://huggingface.co/AI-Sweden-Models/gpt-sw3
Gemma 3 (4B / 27B) – Google's open-weight language models for multilingual tasks.	https://ai.google.dev/gemma
Mistral-Nemo-12B – Open-weight multilingual LLM trained by Mistral AI.	https://mistral.ai/news/mistral-nemo/

Domain-Specific and Supporting Tools	URL
Whisper – Speech recognition model by OpenAI for multilingual transcription.	https://github.com/openai/whisper
Vosk – Lightweight offline speech recognition toolkit.	https://alphacephei.com/vosk
PyOD – Python toolkit for detecting anomalies in multivariate data.	https://pyod.readthedocs.io



Domain-Specific and Supporting Tools	URL
Merlion – Time-series anomaly detection and forecasting library by Salesforce.	https://github.com/salesforce/Merlion
Tesseract OCR – Open-source optical character recognition engine by Google.	https://github.com/tesseract-ocr/tesseract
Techniques and Frameworks for AI Integration	URL
Retrieval-Augmented Generation (RAG) – Framework for connecting LLMs to external data sources.	https://arxiv.org/abs/2005.11401
LangChain – Framework for building applications with LLMs, RAG pipelines, and tool orchestration.	https://www.langchain.com
Agentic AI and Orchestration – Architectural approach to coordinating specialized AI agents for complex workflows.	(General concept; see https://arxiv.org/abs/2401.10968 for overview research.)
Collaborative Partners and Contributors	URL
Aixia AB – Contributor to the Data-Driven Organizations initiative, providing expertise in AI infrastructure and MLOps.	https://www.aixia.se
Region Halland – Public sector partner contributing practical use cases and validation within healthcare and administration.	https://www.regionhalland.se
AI Sweden – National center for applied AI and coordinator of the Data-Driven Organizations project.	https://www.ai.se

Appendix C – Glossary

Term	Description
Agentic AI	An approach where multiple AI components (“agents”) collaborate to complete tasks through reasoning, planning, and interaction with external systems. Each agent performs a specialized function, coordinated through orchestration.
API (Application Programming Interface)	A standardized interface that allows software applications to communicate and exchange data or functionality, often used to access cloud-based AI services.
Data Drift	The gradual change in the statistical properties or patterns of data over time, which can cause AI model performance to degrade if not monitored and retrained.
Data Governance	Policies and practices that ensure data is managed responsibly, including aspects of quality, security, privacy, and regulatory compliance.
Data Labeling / Annotation	The process of tagging raw data (such as text, images, or audio) with relevant information so it can be used to train or fine-tune AI models.
Fine-Tuning	The process of adapting a pre-trained model to a specific domain or task by training it further on a smaller, domain-specific dataset.
GPU (Graphics Processing Unit)	A type of processor designed for parallel computation, essential for efficient AI model training and inference.
Hybrid Intelligence	A system design combining AI with rule-based logic, traditional software, and human oversight — using AI where it adds value while retaining control and interpretability.
Inference	The process of using a trained AI model to make predictions, classifications, or generate outputs from new data.

Term	Description
LLM (Large Language Model)	A neural network trained on massive text datasets to understand and generate natural language, often capable of reasoning and following instructions.
MLOps (Machine Learning Operations)	A set of practices and tools that streamline the development, deployment, and monitoring of AI models in production environments.
Model Training	The process of teaching a model to recognize patterns in data by adjusting its internal parameters to minimize prediction errors.
Open Source	Software or models whose source code is publicly available, allowing anyone to use, modify, and distribute it.
Retrieval-Augmented Generation (RAG)	A method that connects language models to external data sources so they can access up-to-date or domain-specific information during text generation.
Sustainability (in AI)	Designing AI systems that minimize unnecessary computation, energy use, and redundancy — focusing on modularity, reuse, and long-term maintainability.
Transfer Learning	A technique that leverages knowledge from one task or dataset to improve performance on another, typically forming the basis for fine-tuning.
Vision Model	An AI model designed to process and interpret visual data such as images or video, often used for classification, detection, or segmentation.

Project Context and Contributors

This white paper was developed as part of the Sustainable AI Infrastructure Lifecycle (SAIL) use case within the national project Data-Driven Organizations (DDO), coordinated by AI Sweden.

The purpose of the SAIL use case is to explore financially and environmentally sustainable approaches to AI infrastructure that support the entire AI lifecycle — from research and development to deployment, inference, and long-term operation.

Project Partners

The work has been carried out in collaboration between: Aixia AB, Region Halland, and AI Sweden, with additional insights shared through the broader DDO consortium including industry, academia, and public-sector partners.

Authors

Milena Miernik, Aixia AB

Use Case SAIL

Aixia AB: Cecilia Millheim, Ellen Reinhardt, Jonas Nordin, Klas Ludvigsson, Milena Miernik, Olof Sandell, Simon Janeck
Region Halland: Georgios Bramis, Karin Westerberg, Lina Gårdemark, Stefan Bäckström, Torbjörn Olander

Aixia AB
Hälsingegatan 10
414 63 Göteborg

www.aixia.se

Region Halland
Box 517
301 80 Halmstad

www.regionhalland.se

