Generative AI

From use cases to enterprise-wide scaling
and a new organizational paradigm

Working paper - October/2023

futurice

A note about this document

What you are “holding” is a working paper of sorts; a document we use to explore and also educate ourselves about generative AI, its potentials and the whiplash speed at which the field is developing and changing. What this means is it will always be incomplete, a work-in-progress. It will have omissions, some of which we are working to fix and some we probably won’t get around to. We will try to keep updating it frequently as we work to keep on top of it all ourselves.

Above all, we want this to be a document that inspires dialog and exchanges of ideas, so please get in touch with any feedback, comments, questions or ideas you may have!

Editor and primary contact: Tuomas Syrjänen, tuomas.syrjanen@futurice.com, +358505470386

Last updated: 27.10.2023

Table of contents

Summary

Generative AI has been at the heart of our work since ChatGPT launched late last year. As of October 2023, we've done over 25 generative AI projects with our clients, over 60 leadership & board live demo sessions – and we've been drinking our own champagne by using it to reinvent our own sales process and selected areas of our delivery process. We are extremely excited by the potential of this technology, but, as usual, the greatest challenge has been how to derive real business value from it. We've learned a thing or two from our successes so far, so we thought we'd share our experiences in an evolving document and help guide you forward on your Generative AI journey.

Generative AI, large language models, and GPT generate a lot of buzz. Significant technological progress is being made at a dizzying pace, but technology alone does not create an impact on enterprises. While there is no shortage of freely available and high-quality technical materials about this phenomenon, the same can't be said about the business perspective. This evolving working document seeks to mitigate this shortage and share our experiences implementing generative AI solutions at Futurice and with our clients.

First, we need to identify use cases and decide where to start. It makes sense to divide the journey into maturity phases and start by implementing use cases that help people on the frontline succeed with their work, ease their tasks, and save costs – and where humans are in the loop to manage risks. Then, gradually, the focus should be balanced toward use cases unlocking new value for and directly exposing the functionality to clients. After this come the use cases that lead to refactoring end-to-end processes - enabling performance level-ups.

The real value of generative AI comes when we challenge the conventions of our organizational capabilities (people, process, technology) and reinvent how we operate and serve our clients. This change should gradually progress towards end-to-end process refactoring. This means the change is as much human and behavioral as technological development.

Enterprise-wide scaling also requires various enablers from IT to HR and Legal to enable the full potential of GenAI, bring efficiencies, manage risks, and, most importantly, help people get on board with the change.

Finally, Generative AI brings completely new questions to leadership teams' agendas: Where does our competitive advantage arise in the future? What if software development cost is drastically reduced? What role does our proprietary data play in this equation? And the hardest one: What does our workforce and talent look like in 5-10 years?

This document summarizes our generative AI learnings, mainly from a business point of view. The purpose is also to consider the long-term implications of this and similar technologies to the organization's strategy, competitive factors, organizational design, talent management, and culture.

Chapter 1 focuses on concrete use cases, many of which we can demonstrate live, so please do not hesitate to contact us if you want to see some of this stuff in action. Many of the considerations in chapters 2, 3, and 4 also apply to other AI technologies, and LLMs are not the only technologies used to achieve the described benefits. Still, for the sake of simplicity, this document refers mainly to generative AI.

A tiny bit of historical background

In 2018, Futurice grew to 400-500 people. Complexity related to the fundamental challenges of knowledge work started creeping in – in an unmanageable way. We had minimal visibility into our organizational knowledge and information exchange between frontline action and management agendas was lacking. Instead of leading success, we were spending too much time and energy addressing problems.

In 2019, we set up a data & AI renewal team, Futurice Exponential, to focus on data & AI-based knowledge work paradigm to address challenges in knowledge work, such as capturing organizational knowledge, visibility into our own organization, and the markets, flow efficiency of knowledge work, etc. The knowledge gained through seeking internal improvement has also been widely applied in client work.

When GPT 3.5 was launched in late 2022 and GPT 4 in Q1/2023, the transformation faced an inflection point and became mainstream. In 2023, we've been actively implementing generative AI solutions within Futurice and with our clients in over 20 projects - ranging from Media, to public sector , to legal and a global transportation and manufacturer.

Glossary

GEN AI is a subset of AI, focused on models that generate new content. The biggest impact is currently coming from text generation models like GPT. But generative AI includes image, video & audio generation too, with tools like DALL-E, MausicGen & Gen-2.

LLMs or Large Language Models are deep learning models, designed to understand & generate natural language. Leading models like GPT are trained on billions of words from the internet, books & code repos. See this great explainer article on how they work!

GPT-4, PALM 2, LLAMA 2 & CLAUDE 2 are the leading LLMs today - developed by OpenAI, Google, Meta & Antrophic, respectively. Most real business applications are built on GPT, but others are now catching up. See this practical guide to the evolution of LLMs.

RAG or retrieval augmented generation is the de-facto method for feeding LLMs with your own knowledge. You link your LLM with a database →search it for information that matches the user prompt → feed results to the LLM before it gives its answer. See this intro to RAG, from IBM.

FINE-TUNING is a process for taking LLMs and training them further on a smaller, specific dataset. Good for training cheaper LLMs to do specific tasks.

ALTERNATIVE DATA is data we're not traditionally using in business to generate insight - typically unstructured text data, which LLMs now help us put to use on a new scale . See our mini-blog series on using alt data to understand corporate sustainability.

1. Experience from use cases

Selected generative AI use cases we've worked on and can help you work with. Some have quantifiable outcomes we've seen, others not – or the outcomes are almost impossible to quantify without more context than we can provide in this working paper.

Organizational capability	Use cases
Customer service	● Access to correct information from a vast collection of knowledge-base articles, instruction, company policies, etc. Outcomes: Fast access to the right information/knowledge with ready-made formulated answers in seconds instead of minutes.
Market intelligence	● Market demand forecast improvements via analysis of news headlines and key sentences across a vast volume of articles ● Application review analysis, such as: why do some automotive companion apps receive 4+ ratings and some 2.x ratings?
Sales & marketing	● Content generation eg. cold emails, blogs, PoVs, rewriting materials to be more relevant for X profile ● Client understanding based on internal and external data ● Semi automatic customer material analysis ● Semi automated proposal process
Product development / R&D	● Customer need simulation ● Faster product and feature ideation cycles ● Custom engineering ● Dynamic customer feedback analysis that feeds to the product development prioritization
Finance	● Forecasting ● Situation awareness

Software development & production	● Writing new code ● Refactoring and optimizing existing code ● Tech stack upgrades ● Shifting roles: tech implementation by business experts Outcome: Code development Sample outcomes: 2 to 5x productivity improvement in refactoring and performance optimization tasks & 30-50% savings in cloud hosting costs Unlock a new way of working where business experts write the code they need.
Leadership	● Transparency to customer needs – positives and negatives ● Organizational alignment analysis
Legal	● Contract review: clause compliance to contracting policy ● Consolidated view into contract status, e.g., what kind of liabilities exist in our contracts ● Assessing customer claims related to legal risks and comparing with judgements ● Assessing new situations against prior court rulings Outcomes: Tech, data & AI-enabled legal capability offers unprecedented visibility to contracts , moves the work from fixing problems to avoiding them, streamlines workflows, and reduces wait times.
HR	● Target setting alignment analysis ● Organisational design benchmark analysis

A brief look at some internal Futurice use cases

Organizational capability	Use cases and outcomes
FutuCortex - our automated knowledge platform	● Easy access to relevant information from a vast collection of past proposals, presentations & thought leadership work - and enabling us to identify the right experts to involve. Outcome: Access to the right information in seconds instead of minutes. In practice, it helps our sales people write better client proposals, faster.
Data analysis & Software Development	● LLM-enabled data queries, analysis and LLM-generated code for data analysis saves time significantly and also makes it possible for non-tech people to perform more analytics themselves ● Summaries & insights from data ● LLM supports in writing new code and turns feature description into code ● Refactoring and optimizing existing code
Market intelligence	● Market demand forecast improvements via analysis of news headlines and key sentences across a vast volume of articles ● Application review analysis
Sales & Marketing	● Content generation ● Client understanding ● Customer-specific demonstration chatbots ● Domain-specific knowledge companion bots in parallel with our pre-sales activities & insights
Rethinking workshopping	● Customer need simulation ● Product and feature ideation
Leadership &	● Real-time situational awareness through automatically extracting knowledge directly from the tools we use and

Strategy	the 'digital footprint' we create ● Specific long & complex board material is provided as a GPT bot to allow board members to explore questions in a more intuitive and conversational manner
Project delivery reinvented	● Example: https://futuedge.com/ifrs-gptbot Outcome: Domain-specific knowledge base user interface. Project materials are delivered as an interactive bot in addition to the more traditional documentation. This enables the client to explore questions and get familiar with the outcomes in a conversational manner.

Technical approaches and learnings

In order for people and organizations to leverage generative AI, a user interface is required.I It takes a lot more under the hood than just adding a chat on top of data sources. For this there are several categorical approaches to implementing an LLM-based solution or harnessing LLMs. For a broader tech analysis, we recommend you google the latest developments.

To understand how the magic happens, we must separate generic and specific use cases.

The generic approach involves using out-of-the-box solutions like ChatGPT, Bingchat, and Github co-pilot via the chat interface and prompts.

More specific contexts involve two approaches - RAG and Fine tuning.

Retrieval augmented generation (RAG) is the current de facto approach and involves providing the context in the prompt. Before calling an LLM, the specific context-related material is searched using e.g., semantic similarity from a vector database, and this material is provided in the prompt, and LLM is asked to answer the user question only based on this information and not the generic training data.

For example, an IFRS bot has 1000 pages of IFRS standards, and when a user asks a question, relevant parts are searched from the document and provided to the LLM.
See, e.g., Retrieval Augmented Generation: Grounding AI Responses in Factual Data by Minhajul Hoque.

Fine-tuning is less widely used and requires vast volumes of high-quality data samples, frequent retraining, and is more likely to provide answers from outside the context-specific data set. Hallucinations are more likely, too.

Integrated solutions

All major tech vendors are integrating LLM into their products, making adoption more straightforward, but these solutions' actual performance and applicability are still to be validated. Due to the nature of LLMs, these solutions use an RAG approach to bring the proper context into the answers. Herein lies the most significant concern and performance risk: how we choose the context, which material we use, and how we present the right background information to define the answer is often a more critical factor in performance than the actual prompt. We can't adjust the parameters or approaches to how the context is built in these off-the-shelf approaches, and thus, the performance may be lackluster in specific use cases.

Examples: Microsoft co-pilots, Salesforce, ServiceNow, SAP, Google Duet AI

Off-the-shelf solutions

There is a Cambrian explosion of generative AI solutions afoot and even coming to grips with what is on offer out there and the countless areas they can applied to requires skill. According to the HBR podcast, some investors report the increase of AI companies in Silicon Valley from 800 to 8000. These tools implement individual use cases and can be very useful in the big picture, but there are also challenges:

- How to recognize the best tool or solution?
- How do we ensure the business viability that guarantees continuous improvement and development of a specific solution?
- Data privacy, confidentiality, and compliance aspects
- Integration into other systems
- Performance: quite often, LLM and in-context learning parameters, e.g., relevance metrics, number of samples, etc., define the system performance. Typically, these are not adjustable.
- How to integrate into enterprise architecture?

APIs

The hosted models and APIs of all major cloud providers are continuously improving. This is currently the de facto approach to building custom enterprise applications. Still, there are several aspects to consider:

- GPT4 APIs are slow and costly. Using GPT4 alone to analyze, e.g. 1000s of legal contracts, would be too slow and expensive.

- Referring to the previous bullet point, typical solutions are combinations of different technologies and models, such as GPT4, GPT 3/3.5, BERT, and Whisper
- Building the proper context is non-trivial and impacts performance even more than the prompt.
- Implementations require new kinds of database technologies, such as vector database

Examples: Azure, AWS, OpenAI

Open-source and commercial fine tuned models

There is a massive growth of various open-source LLMs, of which Llama 2 is currently getting the most attention. It's a fast-developing field (see below). There's a significant buzz, but several frictions make this approach quite tricky. The RAG explanation above sheds some light on this.

There is growing evidence that certain types of enterprise cases are valid for, e.g., Llama2-based solutions. The use case characteristics are:

- A narrow and well-defined problem, e.g., extracting specific information from similar documents
- High volumes that make Azure API, etc., usage very costly
- High-quality training data with input-output examples (e.g., full-text & summary, question & answer, etc.)is available. With enterprises, this high-quality unstructured training data is sometimes the most significant source of friction.

See the appendix for an example of analyzing and extracting information from job listings with Llama2.

Creating custom LLMs can be an option, depending on the level of investment.

- Fine-tuned on existing models: Available models like GPT 3.5 can be fine-tuned for a particular task (like classification, Q&A)via a user-specified high-quality custom training dataset.
- Foundational models: Open source models that rival GPT 3 and 4 are available and can be used as a foundation for additional training on your own custom dataset. Examples: Falcom LLM, LLAMA, GPT-J
- Training an LLM from scratch: This is the most expensive approach.

It's essential to remember that enterprise applications are typically based on several technologies, and LLMs are a critical part of a long chain of different technologies and models. Sometimes, GenAi is the central part; other times it can cover a specific functionality that complements the overall solution.

Current frictions

Although significant results are to be had with the outlined use cases, there are challenges, too.

Privacy and data confidentiality

OpenAI has already lost some credibility in privacy and data confidentiality. They explicitly state that free/chat version prompts can be used to train the system. Their systems have leaked prompts, and they also shared an estimate that 2% of the data has leaked. There are other widely publicized incidents of sharing personal and professional information with public AI chatbots, like Samsung. We recommend not harnessing the OpenAI implementation of GPT for enterprise usage.

Mitigation:
1. Using Microsoft, AWS, and Google-hosted APIs in Europe that comply with GDPR and enterprise confidentiality.
2. Setup locally hosted LLM, such Llama2.

IPR issues of generated content

Many large IP-driven companies have concerns over the code/content produced, whether it contains open source code causing its own code contamination. The legislation is also still immature regarding which data can be used legally for training purposes and what are the legal rights of the output.

In September 2023, Microsoft released a Copilot Copyright Commitment. Please note that this Commitment does not seem to cover Azure APIs.

Costs

Current hosted APIs are billed per token; e.g., GPT4 API is ten to thirty times more costly than GPT3.5. Careful cost management is essential: how to build solutions that perform well with GPT3.5, how to limit API requests, how to avoid cost surprises, and so on. Cost optimization is a key design driver: just like with all tech, you don't always need the best and most expensive and feature-rich version. Setup costs versus running costs may vary greatly from one provider to another.

Nondeterministic, volatility & hallucinations

LLMs are not deterministic, meaning the same question may produce different answers. The longer the prompt and more complex the query, the more volatile the results are. This needs to be taken into account in real-world solution creation.

Unraveling the long-term value capture in GenAI

All big tech companies and many startups are pushing hard on LLMs. From a business point of view, who will capture the value in the long run? Do technology providers (OpenAI, Anthropic, Google, etc.) have a defendable position to monetize aggressively?

Initially, OpenAI was in a league of its own with a 10B$ investment from Microsoft, which enabled Microsoft to provide commercial APIs quickly and start integrating LLMs into various products for a nice head start, but the playing field soon started to level, with new models from many vendors arriving almost weekly. Nobody is in a position to dominate and the competition is becoming increasingly heated. It's likely that companies will offer a variety of equally capable models – or different models focusing on specific problems. One of the first decision organizations looking for generative AI solutions wil have to make is a golden oldie: should we look for a holistic partner or by from different providers for different needs?

The questions remain to be answered, but there are solid indications that technology may not be the location of most value capture due to the progress of, e.g., open source.

- Google "We Have No Moat, And Neither Does OpenAI"
- Special Series: How Generative AI Changes Everything

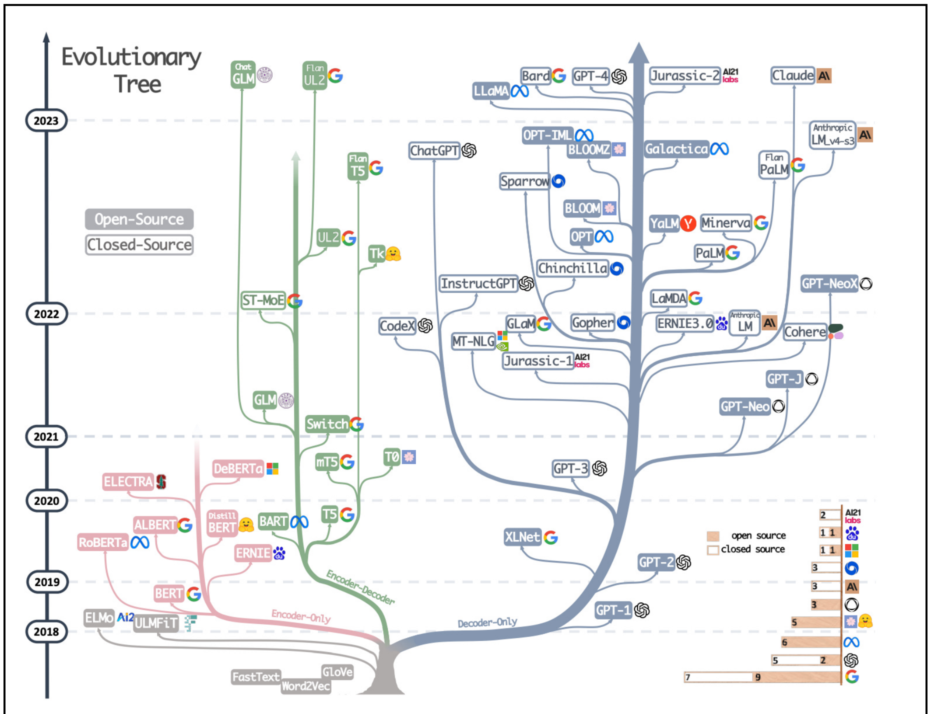
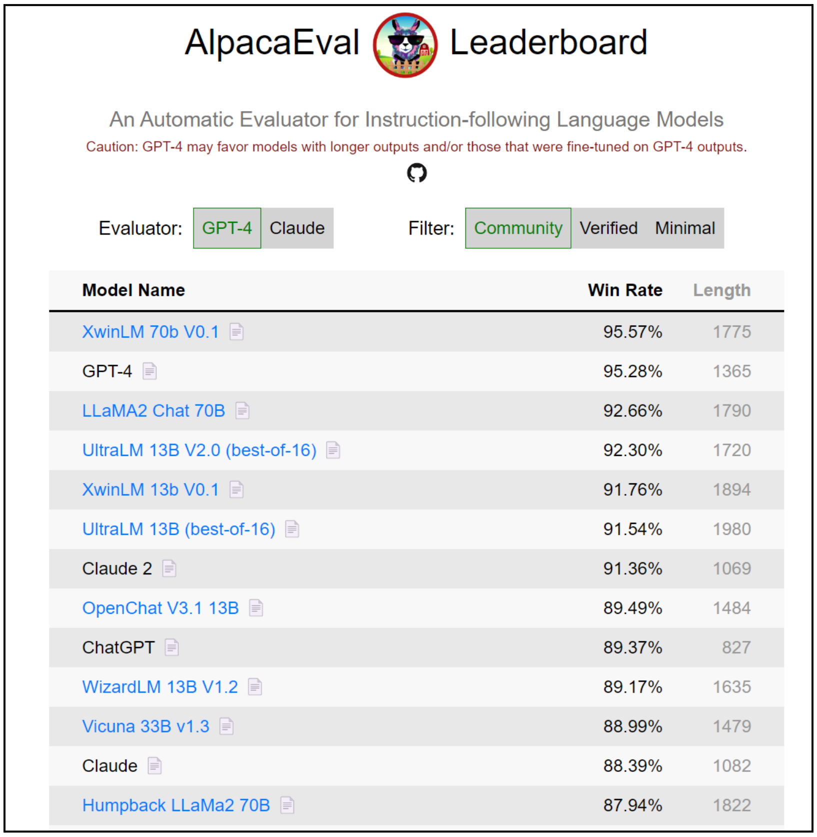


Fig. 1. The evolutionary tree of modern LLMs traces the development of language models in recent years and highlights some of the most well-known models. Models on the same branch have closer relationships. Transformer-based models are shown in non-grey colors: decoder-only models in the blue branch, encoder-only models in the pink branch, and encoder-decoder models in the green branch. The vertical position of the models on the timeline represents their release dates. Open-source models are represented by solid squares, while closed-source models are represented by hollow ones. The stacked bar plot in the bottom right corner shows the number of models from various companies and institutions.

[Different development paths of LLMs - by Nathan Lambert]



AlpacaEval Leaderboard

An Automatic Evaluator for Instruction-following Language Models

Caution: GPT-4 may favor models with longer outputs and/or those that were fine-tuned on GPT-4 outputs.

Model Name	Win Rate	Length
XwinLM 70b V0.1	95.57%	1775
GPT-4	95.28%	1365
LLaMA2 Chat 70B	92.66%	1790
UltraLM 13B V2.0 (best-of-16)	92.30%	1720
XwinLM 13b V0.1	91.76%	1894
UltraLM 13B (best-of-16)	91.54%	1980
Claude 2	91.36%	1069
OpenChat V3.1 13B	89.49%	1484
ChatGPT	89.37%	827
WizardLM 13B V1.2	89.17%	1635
Vicuna 33B v1.3	88.99%	1479
Claude	88.39%	1082
Humpback LLaMa2 70B	87.94%	1822

Leaderboard October 11th, 2023: One way to understand the volatility of the generative AI field is to keep tabs on things like the [Alpaca Eval Leaderboard].

2. Enterprise-wide scaling

Individual use cases bring value to individual processes or tasks. Organization-wide impacts – such as elevated productivity levels and new competitive advantage – require enterprise-wide scaling that takes a much broader view than individual use cases.

The three main areas are

1. Maturity phases & portfolio management
2. Organizational enablers
3. Managing change

Maturity phases & portfolio management

The most common question we hear is, “Where to start?” The journey is easier to understand when divided into different maturity phases.

1. Support individual steps in existing processes
    a. Cost efficiency: e.g., service agents find information faster
    b. Added value to existing processes: helping sales to cross-sell/up-sell better
2. Redesign workflows, tasks, and processes
3. Paradigm change: end-to-end processes refactored in a completely novel way on top of data & AI capabilities
4. Optional: from *company-using-AI* to *AI company*. They are two different things.

The last phase is holistic paradigm change to reach the *AI company* level, which includes end-to-end redesigns, cultural change, and, in most cases, a holistic business model change. Reaching this maturity level takes years and may never be reached. The good thing is it's not necessary for all companies.

There are also other aspects to consider. People are naturally concerned about this technology, so it makes sense to start with use cases that focus on helping people succeed in their work by, e.g., easing some painful process steps. We also would not recommend exposing LLM/GPT functionality to customers before there is more experience and best practices to manage various risks better.

So the answer to the question is:

Start with **cost-efficiency use cases that support current processes** where **users are internal, humans make the final call** (human-in-the-loop)**,** and the focus is on **helping people succeed**, not enabling better management control**.**

After initial use cases, managing the whole enterprise-wide portfolio becomes crucial for a bigger impact. There are dozens of different ways to categorize portfolios, but the following aspects should be considered, at the very least

- Strategic alignment: What are organizational value levers, e.g., flow efficiency, customer experience, etc.? Define and support them.
- Desired portfolio balance
- Business case evaluation: Benefits and impact of use case and costs
- Use case risks & frictions
- Data access synergies: Quite often, the most challenging and costly issue is making data accessible for the system. Prioritization of use cases that use the same data sources is advisable.
- Technical similarities/synergies

Replicating success to similar needs across organizations.

Let’s say we have successfully validated information summarizing in customer service. Naturally, it makes sense to find similar needs in other processes and functions, such as IT helpdesk, shared services center, maintenance, etc.

Balancing the portfolio

In the early stages, the portfolio should be biased toward use cases that provide easy and concrete benefits with limited short-term risks. However, the portfolio should not consist solely of these low-hanging fruits.

A concerted effort should be made to unlock new value use cases and start the journey towards end-to-end refactoring. If existing use cases do not fulfill this requirement, a few should be chosen to kickstart the journey.

One starting point for portfolio balancing would be 60:30:10 prioritization in the early stages

- 60% for cost efficiency
- 30% unlocking new value
- 10% for use cases that open access to end2end process refactoring

The second most frequently asked question we encounter is, “Where should we aim?” We answer that organizations should aim for end-to-end process redesigns because they lead to **performance level-ups** and even **unfair competitive advantage**.

Enablers and foundations by function/domain

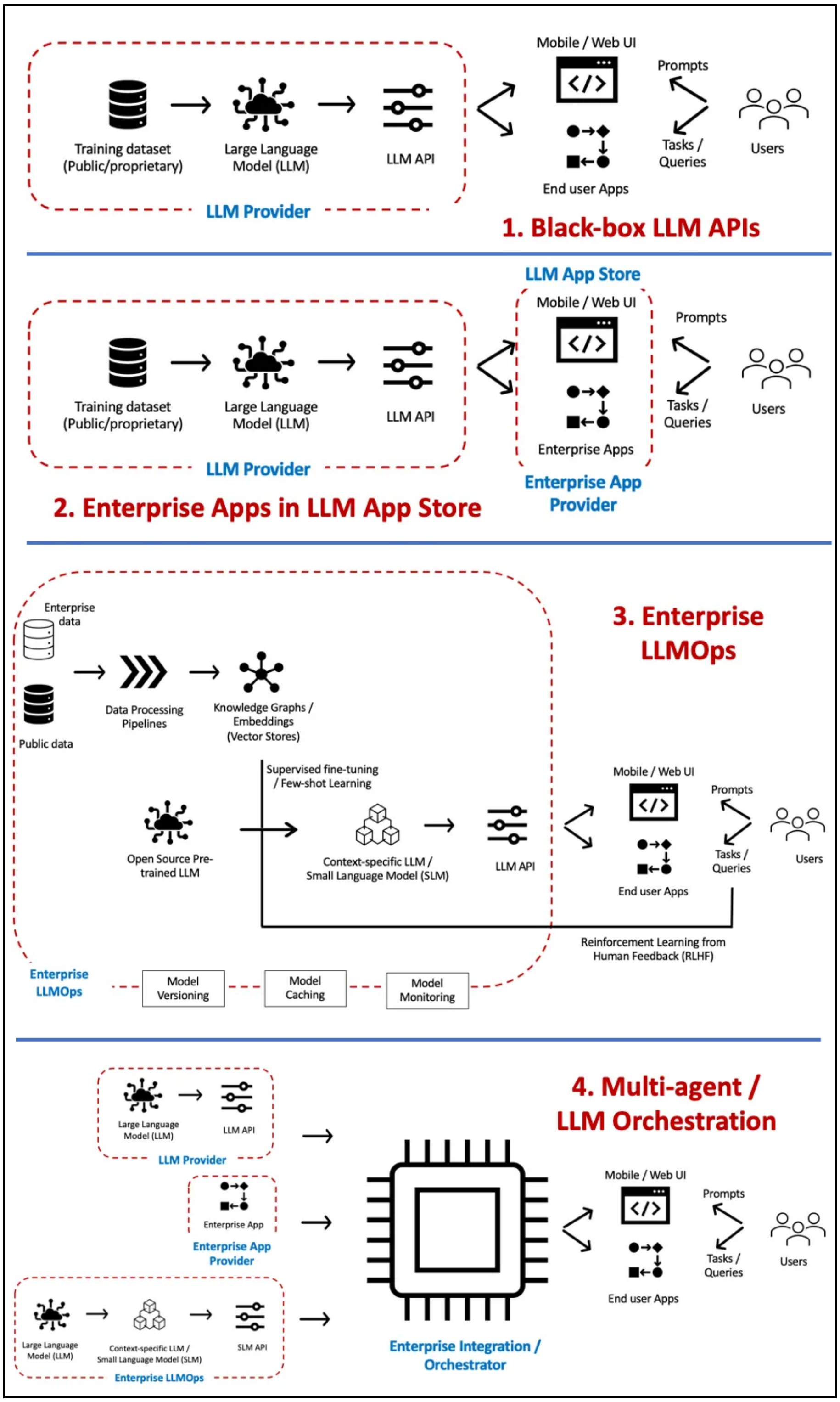
During 2023 in our client projects , we’ve learned that advancing generative AI in organizations through use cases alone is insufficient. We also need to build foundations that enable scaling, ensure security, keep people in the loop, etc.

Function/domain	Various enablers
IT	Enterprise architecture, governance, and technical enablers ● LLM Gateway: Access control, cost control, telemetry, … ● MLOps: managing model changes, etc. ● Managing & prioritization of various models from MS, AWS, .. ○ Platform LLMs vs. Niche SaaS solutions vs. Custom (including co-pilots vs. custom) ○ Hosted APIs vs. Open Source LLMs such Llama 2 ● Other new systems needed, such as Vector databases and related ● Governance practices ○ Cost ○ Quality ○ Telemetry ○ Explainability and repeatability ○ Risk assessment ○ Validation and managing risky use cases ● Security / Privacy / Compliance
Technology	● Constant evaluation of new technologies ● Guiding with model and how is used in which cases
Data	● What data is needed? Information portfolio management ● New alternative data, especially external data ● Data access to people
HR - skills & talent	● Competence development ○ AI & data science role ○ All roles with AI knowledge ● New division of labor → tech becomes everybody’s toolbox ● Prompt engineering - the ability to use & guide the systems ● Shaping culture ● Thinking different - challenging the convention in e2e process ● Thinking different in individual processes ● Sharing learnings across the organization

Leadership	● Expecting process/workflow changes in business - not just new technology ● Expecting visible results in operations plans: e.g., increasing sw roadmap items due to productivity, changing roles of customer service, change in # of people involved in a process, process metrics e.g., cycle time ● Managing conflict: changing how we work will inevitably lead to conflicts inside the organization. Leadership should build skills in managing and facilitating these conflicts.
Legal and risk management	● IP of generated content ● Which are high-risk use cases ● Customer exposed functionality risks: brand, legal, … ● Compliance to various regulations
Organization design	● New process templates/guides to drive desired strategic change, e.g., flow efficiency, customer experience (ultimate personalization), knowledge centricity…
Cyber security - understanding risks and conscious decisions	For example, to harness generative AI properly we need to provide access to organizational data to everybody. Already now, data breaches are the fastest growing cyber security incident class. How should we balance competitiveness and risk?

Selected topics from the table

Generative AI LLMOps deployment architecture patterns


Training dataset (Public/proprietary)
Large Language Model (LLM)
LLM API
LLM Provider
Mobile / Web UI
End user Apps
Prompts
Tasks / Queries
Users
1. Black-box LLM APIs
LLM App Store
Training dataset (Public/proprietary)
Large Language Model (LLM)
LLM API
LLM Provider
Mobile / Web UI
Enterprise Apps
Enterprise App Provider
Prompts
Tasks / Queries
Users
2. Enterprise Apps in LLM App Store
Enterprise data
Public data
Data Processing Pipelines
Knowledge Graphs / Embeddings (Vector Stores)
3. Enterprise LLMOps
Supervised fine-tuning / Few-shot Learning
Open Source Pre-trained LLM
Context-specific LLM / Small Language Model (SLM)
LLM API
Mobile / Web UI
End user Apps
Prompts
Tasks / Queries
Users
Reinforcement Learning from Human Feedback (RLHF)
Enterprise LLMOps
Model Versioning
Model Caching
Model Monitoring
4. Multi-agent / LLM Orchestration
Large Language Model (LLM)
LLM API
LLM Provider
Enterprise App
Enterprise App Provider
Large Language Model (LLM)
Context-specific LLM / Small Language Model (SLM)
SLM API
Enterprise LLMOps
Enterprise Integration / Orchestrator
Mobile / Web UI
End user Apps
Prompts
Tasks / Queries
Users

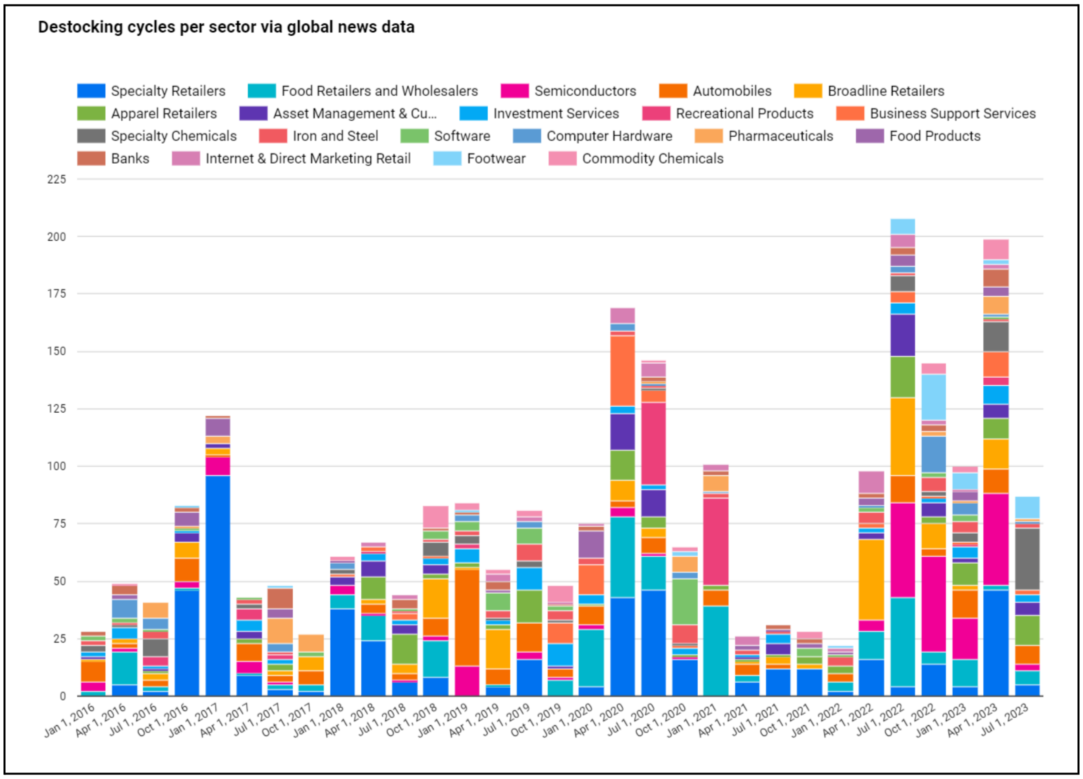
https://medium.datadriveninvestor.com/generative-ai-llmops-deployment-architecture-patterns-6d45d1668aba

GenAI/LLM Companion - Alternative data

Generative AI - especially enterprise use cases - needs data to perform company-specific and relevant use cases. Organizations typically consider data as structured and numerical data from e.g ERP, CRM or Finance systems.

In many use cases, we've found out that so-called alternative data can bring valuable insights such as, e.g., access control gate in construction predicting site success, forming knowledge profiles from, e.g., hour marking, or understanding clients, competitors, and markets via news data, job listings, patents, reviews, investor data, social media...

Organizations need to develop a way to manage their information portfolio and acquire new datasets both from their own operations and external sources. The external alternative data market is developing continuously, especially to serve investors, but we've found out that those data sets are highly valuable to company use cases as well.



Example of using news data to understand and forecast "destocking" phenomena.

Leadership & Conflicts

Driving impact from generative AI means changing workflows, ways of working and end-to-end processes. However, current ways of working contain implicit mental models of how business should be conducted and decades of ingrained habits and beliefs. Conflict is inevitable now that there is a need to change assumptions, beliefs, and mental models to enable new ways of working. This conflict is a necessary part of the change but needs to be managed. Leadership must approach conflicts directly and help people build new thinking. This takes time and dialogue. Leadership needs to patiently go through the same thinking process they went through.

Achieving impact

Expectations

Expectation setting by the leadership is crucial in driving organizational change, and generative AI is no exception. Leadership needs to clearly articulate clearly how they want this enterprise-wide adoption or even paradigm change to be visible in organizational KPIs:

- Cost KPIs: e.g., FTEs per business volume in customer service, etc.
- Software developer FTEs per product roadmap(s)
- Improved conversion rates across the customer journey via more personal messaging
- New high-value metrics like cycle time across the whole organization

A clear agenda/value lever vs. innovation at random

In LLM and other transformations, a clear agenda and very clear, simple metrics to drive it help. For example

- “Flow efficiency”: Whatever is innovated and done should reduce lead times with an overall strategic goal of reducing the cycle time initially by 25%, then 50%, and finally 60-70% end-to-end. In this case, the 50% cycle time reduction is enabled by a phase 3 maturity level, i.e., redesigned end-to-end processes on top of new technical capabilities.
- Customer experience is visible in Customer-Lifetime-Value, cross-sell/up-sell metrics, and so on, achieved via extreme personalization, customer understanding, etc.

Renewal Team

As with any major transformation, our experience highlights the need for a dedicated team to drive the change. We do not recommend traditional top-down PMO. What’s needed is a balanced “renewal team” approach to drive use cases and build the bottom-up culture, activity, and skills to advance the agenda, both with top-down initiatives and bottom-up activity. The renewal team works with the people, co-creates, communicates, supports, and makes activities transparent.

Phasing

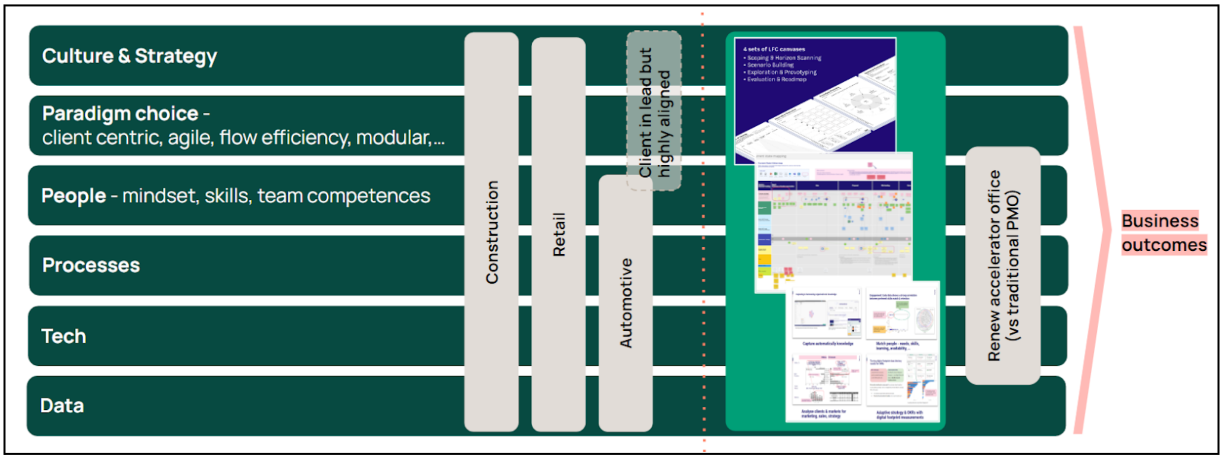
In some cases, we've started enterprise-wide scaling immediately with first proof-of-value use cases, and in others, we've started from PoCs and PoVs. In case a separate PoC phase is used, it should ideally last at most three months and never more than half a year.

We do not recommend concentrating solely on the foundations in the beginning. This is because use cases will guide what foundations are needed and how they should be implemented in detail.

3. Rethinking organizational capabilities

After initial use cases and starting an organization-wide transformation, we've usually initiated work towards the next maturity level – in other words, rethinking broader organizational capabilities and end-to-end processes to achieve a level-up.

A full-stack approach to organizational capabilities


Culture & Strategy
Paradigm choice - client centric, agile, flow efficiency, modular,...
People - mindset, skills, team competences
Processes
Tech
Data
Construction
Retail
Automotive
Client in lead but highly aligned
4 sets of LFC canvases
• Scoping & Horizon Scanning
• Scenario Building
• Exploration & Provotyping
• Evaluation & Roadmap
Capture automatically knowledge
Match people - needs, skills, learning, availability,...
Analyse clients & markets for marketing, sales, strategy
Adaptive strategy & OKRs with digital footprint measurements
Renew accelerator office (vs traditional PMO)
Business outcomes

The performance level up comes when we start redesigning end-to-end processes and rethinking how our organizational capabilities work with the help of new technologies.

Refactoring end-to-end processes

The most significant data and AI impacts we've witnessed have arisen from solutions that redesign end-to-end processes:

- German automotive car design process across marketing, product development, and production to ensure that cars can be emission-certified

- Construction processes across design and construction are transformed using data and AI from the project paradigm to an industrial assembly paradigm.
- Challenging traditional industry trade-off between global synergies vs. local autonomy. Instead of choosing either or, one company built an operating model around data that provided both.

Sales & marketing

There are countless opportunities to rethink sales and marketing. Client insights for decision-making is a good one! Improved understanding of clients is paramount to drive meaningful interactions and foster lasting relationships. One option is to construct a comprehensive client profile by collecting a multitude of data sources – both internal such including CRM systems, service ticket data, internal dialog, legal, etc., and external alternative datasets like news, investor data, job listings, product releases, social media, patents, reviews – and harnessing generative AI to make sense of it all. This creates a kind of a digital twin of our customer/client.

Parallel to this, creating a digital twin of the offering empowers us with a mirror reflection of skills, capabilities, and future trajectories. This dual insight lays the foundation for a tailored approach to client engagement. By leveraging this information, businesses can sculpt unique messages and strategies tailored to each client's unique needs and characteristics. We can also harness the customer digital twin for a variety of other decision-making points, such as the leadership team entertaining decisions regarding customer service approach and levels or product/service portfolio decisions.

Possessing robust data assets about offerings and configurations revolutionizes our approach to client needs. We can actively iterate and refine solutions in real time during client interactions. This dynamic method contrasts starkly with the conventional approach of waiting weeks for RFP responses. Adopting a real-time iteration process expedites decision-making and fosters a more collaborative and responsive relationship with clients, optimizing the solution for both parties.

More critically, a data-centric approach pivots the nature of client relationships from a traditionally reactive stance to a proactive one. This proactive paradigm enhances client satisfaction and ensures a deeper alignment with their evolving needs and aspirations.

Organizational alignment, coordination, and collaboration

Today, this process is typically carried out via cross-organizational meetings and presentations as artifacts. Cross-organizational meetings are an integral part of ensuring alignment across an organization. These meetings provide a platform for leaders, teams, and individuals from different departments to connect, collaborate, and communicate, fostering a shared understanding of organizational goals and strategies.

This approach typically creates major frustrations due to the number of meetings needed, inefficiencies in meetings, topics falling between silos due to lack of participation and engagement, etc.

One way to eliminate many sources of frustration is to use Amazon Six Pagers & RFCs supercharged with generative AI. Amazon's "six-pager" approach refers to a practice in which a detailed, six-page narrative is prepared ahead of meetings. The memo outlines the topic and presents relevant analysis, arguments, and proposed actions.

Key benefits of this approach include clarity of thought, shared understanding, promotion of deep discussion, respect for time, and thorough documentation.

Request for Comments (RFC) is used in the IT and software development industry to propose changes, gather input, and facilitate decision-making. The process typically involves drafting an RFC document outlining a problem and a proposed solution, inviting stakeholders to review and discuss, and then deciding based on the feedback received.

Key benefits of this approach include clear communication, inclusive decision-making, transparency and accountability, learning and improvement, documenting decisions and their rationale, and promoting innovation. The RFC process allows for a safe environment to propose and discuss innovative ideas or substantial changes, thus fostering a culture of continuous improvement.

Generative AI and similar technologies take these approaches to the next level:

- Automatic alignment of organization-wide action and thinking with strategic goals
- Detection of misalignment between different organizational units
- Automatic alerts in one unit to detect if another unit is preparing something that impacts its work
- Check and simulate thinking and drafts against historical decisions and current action for potential misalignment
- We can form an automatic consolidate view of the organizational status

A new division of labor between business and IT/R&D

Tools like GPT, code-interpreters, co-pilots and similar empower non-technical people to write software, perform data analysis and similar traditionally highly technical tasks. Business/process experts already understand the domain, which is a great advantage. If these people can think through software and data, they can implement the required solutions or changes themselves.

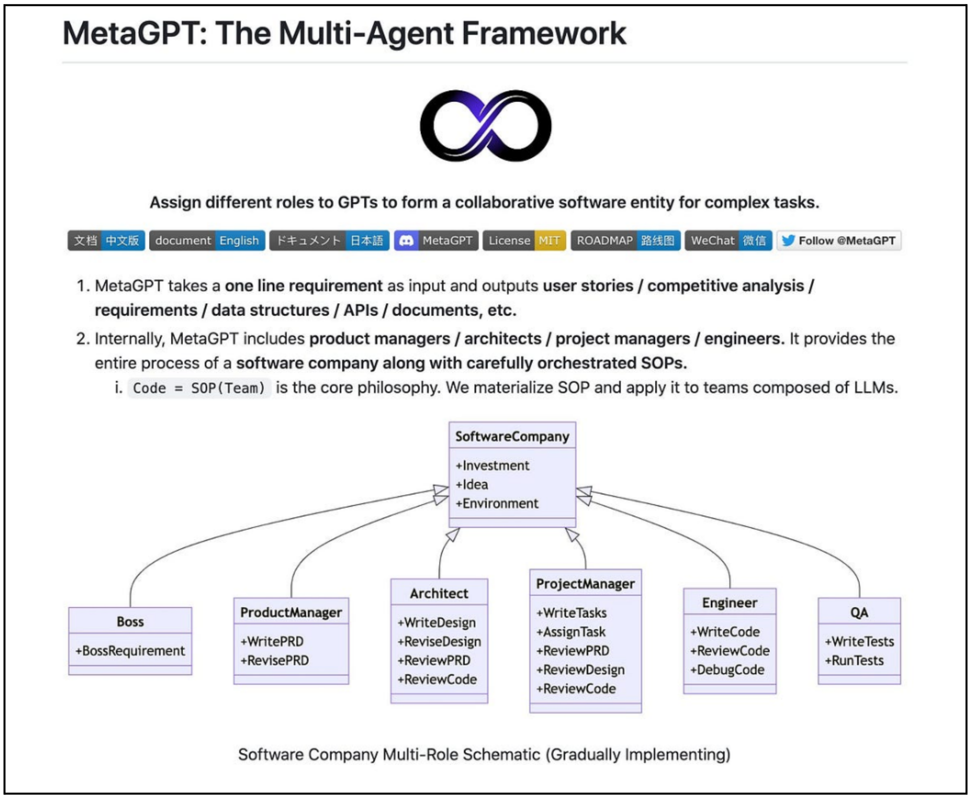
This leads to interesting benefits as technology work moves closer to business problems.

This also leads to interesting risks and challenges. How is this kind of emergent software development governed? What kind of environments are provided for people? How do we build effective guardrails?

Naturally, this does not only concern non-technical people, but also junior developers can also perform at a higher level with these technologies.

Software development

Various emerging frameworks seek to move the paradigm in software development towards automatically turning specifications into code.



MetaGPT: The Multi-Agent Framework

Assign different roles to GPTs to form a collaborative software entity for complex tasks.

1. MetaGPT takes a **one line requirement** as input and outputs **user stories / competitive analysis / requirements / data structures / APIs / documents, etc.**
2. Internally, MetaGPT includes **product managers / architects / project managers / engineers.** It provides the entire process of a **software company along with carefully orchestrated SOPs.**
   i. `Code = SOP(Team)` is the core philosophy. We materialize SOP and apply it to teams composed of LLMs.

Software Company Multi-Role Schematic (Gradually Implementing)

- MetaGPT: a Multi-Agent Framework to Automate Your Software Company | by Peter Xing | DataDrivenInvestor
- Introducing gpt-engineer: Streamlining Code Generation and Technical Specifications | by Eugene B | Medium
- ChatGPT Code Interpreter: What Is It and How It Works? | Beebom

Renewed organizational capabilities

Management and leadership clock speed

Many organizational topics are managed or led very infrequently due to slow, manual, and complex information-gathering processes:

- How is strategy progressing?
- How are clients responding to our service changes?
- What is the current status of our construction site?

Automatic data flows, automatic sense-making of large amounts of data, and the availability of new data sources – such as service line recording and external alternative data, e.g., job listings and news articles – all provide opportunities to rethink the clock speed. In our experience, increasing management clock speed from months to weeks to days to hours, typically up to ten times clock speed, is one of the more effective capability changes.

Leading success instead of fixing problems

Traditional management often focuses on problem-solving, reacting to issues once they've arisen. The rise of data analytics and AI is transforming this approach. By leveraging these technologies, we can anticipate and prevent issues before they occur, shifting from reactive to proactive. This enables us to avoid disruptions, minimize costs, and increase efficiency. Moreover, data and AI provide insights to drive improvement and innovation, fostering more intelligent decision-making and strategic planning. In short, data and AI are empowering us to move from merely fixing problems to actively avoiding them and driving success.

Knowledge-centric organization

Due to a lack of access to organizational knowledge, processes, in many cases, are executed without a client's knowledge, previous solutions, own organizational capabilities, etc. This results in suboptimal solutions, reinventing the wheel, slowness, and slow iteration. The ability to harness organizational knowledge into processes offers numerous novel opportunities to improve results.

Organizational connectivity - The Connected Company.

As organizations grow, complexity increases; therefore, to keep operating, we need to simplify through units, functions, etc. These organizational units, functions, and silos are disconnected, resulting in numerous issues. For example, one unit makes decisions that make sense from their point of view but are detrimental to other units and organizations.

With data & AI, we can rebuild organizational connections, tap into organizational knowledge and align our actions, simulate impact to other units, etc.

From solving problems to leading success

The knowledge-centric organization and the Connected Company are close to enabling front-line people with unprecedented knowledge to handle tasks autonomously without tapping into 2nd, 3rd, or other back office support.

Typically, it takes time to create the support from, e.g., legal, HR, technical, engineering or finance, and that time is wasted. As we know from Lean, every handover is also a source of waste and issues. For clients, frontloading knowledge means instant serving, instant offers, and instant answers.

If problems can be detected when they are still being created, they can be avoided altogether.

Realtime organization

Using data & AI to create instant visibility on the frontline, customers, internal stakeholders, and their relationships via simulation leads to a real-time organization. We can instantly see how strategy is progressing or what our clients need. This leads to real-time characteristics where every single conversation, meeting, and interaction can enrich the dialogue using real-time knowledge assets.

How to screw it all up

What if we don’t rethink how we work, operate, and create value? In other words, what happens if we introduce the latest technology but neglect to change the environment?

This is what a stereotypical tech-focused AI solution looks like: focus on technology, not change anything else, such as processes, operating model, thinking, mindset or metrics. The outcomes will be marginal – at best.

The above image was made with DALL·E 3 using the following prompt:

Picture of a modern looking robot dragging an old fashioned looking carriage, as if it is a horse. This is to be used in management consulting, to make the point that you should not use modern tools to drag forward your outdated processes and ways of working.

It is a variation on the idea in this YouTube video: Adam Savage's Spot Robot Rickshaw Carriage! Savage obviously neglected to change the environment.

4. Long-term organizational implications

With the landscape changing almost weekly, most of us are running just to keep up and concentrating on details that are vital right now. It's vital that we keep an eye on the big picture, too. The rapid change we are undergoing right now will have significant long-term implications for organizations in countless areas like competitive factors, organizational design, talent, and more.

How to structure the long-term strategic implications

Team/People	Topics
Leadership team / executive team/board	● Where is our competitive advantage in 5 years' time? ● Organizational design principles ● Judgement calls between driving competitive advantage & data/tech risks, making choices between pareto-optimal options prepared by e.g., IT
Human resources prepares and decides with the executive team	If a “strategic decision” is defined by a simple rule: “decisions that are either costly or difficult to change in the future strategic” - then people topics are very much such issues. ● How do we make our people future capable - ability to harness latest technologies, and ability to challenge the convection. ● How does our workforce look like in 5-10 years time? Which talent we need more, which talent we need less
IT	● Where is our competitive advantage in 5 years time ● Organisational design principles
IT & Risk management	● Organizational design principles

Source of competitive advantage and barriers to entry

Many organizations see their product’s complexity and proprietary software as a competitive advantage and barrier to entry. If the so-called from-spec-to-software automation becomes a reality, are millions of lines of code still a competitive advantage? What happens when generally available information no longer offers a competitive advantage in the future?

The democratization of information and knowledge took a huge step forward with Google 15 years ago, making many traditional competitive advantages obsolete almost overnight. Generative AI takes it to the next level.

Where do the future competitive advantage and barriers to entry arise from? Proprietary data is one answer. This means companies should look at data from operations, codifying organizational knowledge, acquiring proprietary datasets, and creating derivative insights specifically for our business to strengthen their position in the long term. The future belongs to proprietary data over proprietary software.

We must start leading our knowledge work in a completely different way.

Leading information and knowledge portfolios

Organizations are very familiar with leading portfolios like product, solution, and initiative, but very few are leading information and knowledge using the same portfolio approach. The reason is that traditionally, knowledge has been hard to make tangible, so applying portfolio management to it has been practically impossible. Data, AI, and especially generative AI have changed this.

We should design information and knowledge portfolios that enable organizations to make the right decisions at the right time about the right topics/products, etc. Barry O'Reilly, among others, provides insights into this.

(Super)Talent in the future

This is an area where questions currently outnumber answers by an order of magnitude.

- How do we ensure that all our talent keeps up with these latest technologies? We are already on a path where individual productivity differences are growing and generative AI may expedite this even further. People who can harness the power of GenAI become super productive while others stagnate.
- How do we ensure that all our people have a developer and data scientist mindset in the future?
- How must organizations change when technology comes closer to every single individual?
- How should we think about super talents? Can we scale them? Can we codify their thinking?

Cost of technology goes down

Every time the cost of technology has gone to 0.1X, the impact structurally, market-wise, and more has been something special. Now, we are on a journey where software cost is going down radically. What are the systemic implications of this?

Somebody stealing organizational knowledge

If we can codify our knowledge into LLMs, then somebody can really steal our whole knowledge. This has not been a threat traditionally. Yes, one has had the opportunity to steal code or product designs but not really the organisational knowledge that created to designs. Maybe this is a threat in the future as well?

Systemic change in the operating model opens options for strategic repositioning

Companies who challenge their operating paradigm with the help of data & AI go through an interesting journey. It progresses in three phases

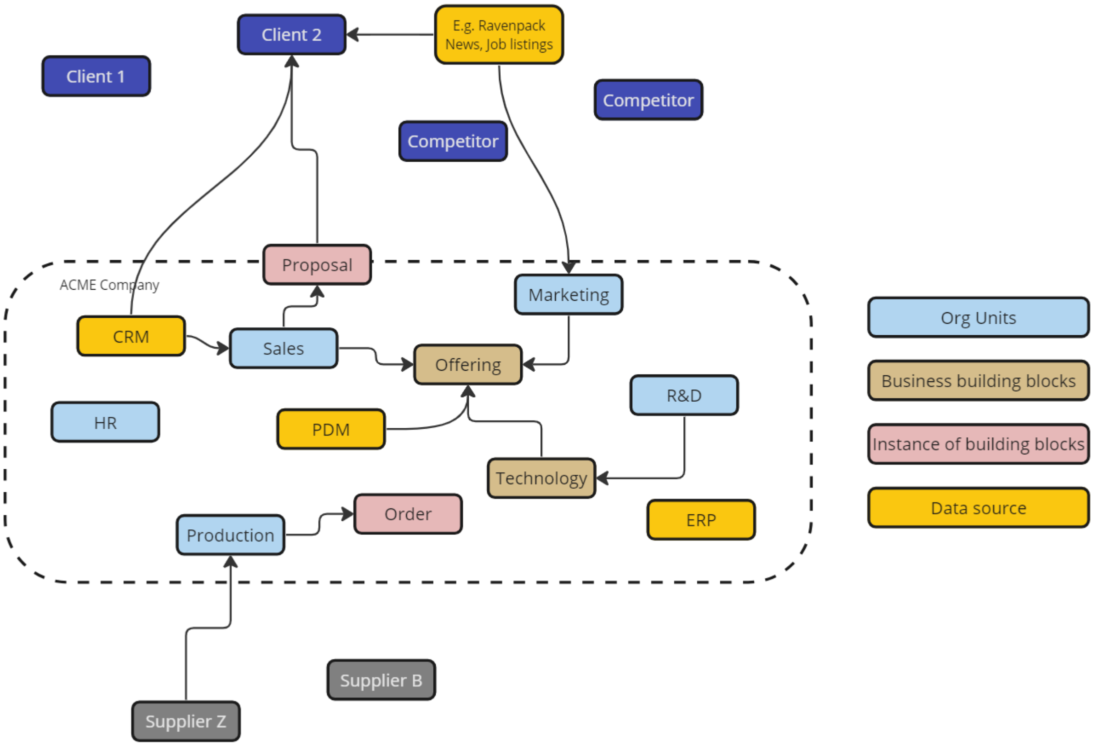
1. *Data & AI applied to existing processes and operating models.* This phase provides some ROI but nothing special, and the purpose is to become familiar with the technology and develop the organization's maturity towards phase 2.
2. *Systemic change.* During this phase, the company challenges convention and drives through a systemic change in its operating model. In the case of construction, we have seen the shift from a chaotic project paradigm to a flow-efficient industrial assembly paradigm, or in retail, from the trade-off between global synergies and local autonomy and company to building an operating model that offers both. The outcome of this phase is an unfair competitive advantage - something that produces superior customer experience/cost efficiency and is difficult for others to replicate because the change covers all areas of the organization: strategy, operating model paradigm, thinking, culture, processes, tech, and data.
3. *Strategic options/repositioning.* The systemic change and unfair competitive advantage created in phase 2 allow the company to choose how to proceed with this new capability: keep it internal and drive organic competitive advantage, or maybe start an M&A campaign to acquire competitors with traditional playbooks to upgrade them. This results in some exciting valuation growth calculations.

   We've seen organizations build a whole new tech business around their new capabilities. In practice, it means selling the same technology to other companies – including the competition – in the same industry. We've also seen data-based business models where the product is *insight*. For example, Finnish construction company Fira has gone through this journey and built an industrial assembly-based construction operating model, validated it in their own business, and de-merged Flow Technologies to be a global tech company selling the same tech to other companies such as Skanska and AF Gruppen.

How do we build organizations in the future?

Organizations are about collaboration, coordination, managing focus, and performance, and traditional toolboxes consist of people, structures, and processes. Mainly, though, organizations are about managing complexity. So, how do we build organizations in the future when the toolbox is so different? We can manage complexity with data, AI and tech very effectively. Do we still need structures? Do we still need rigid processes or do we guide the work via data?

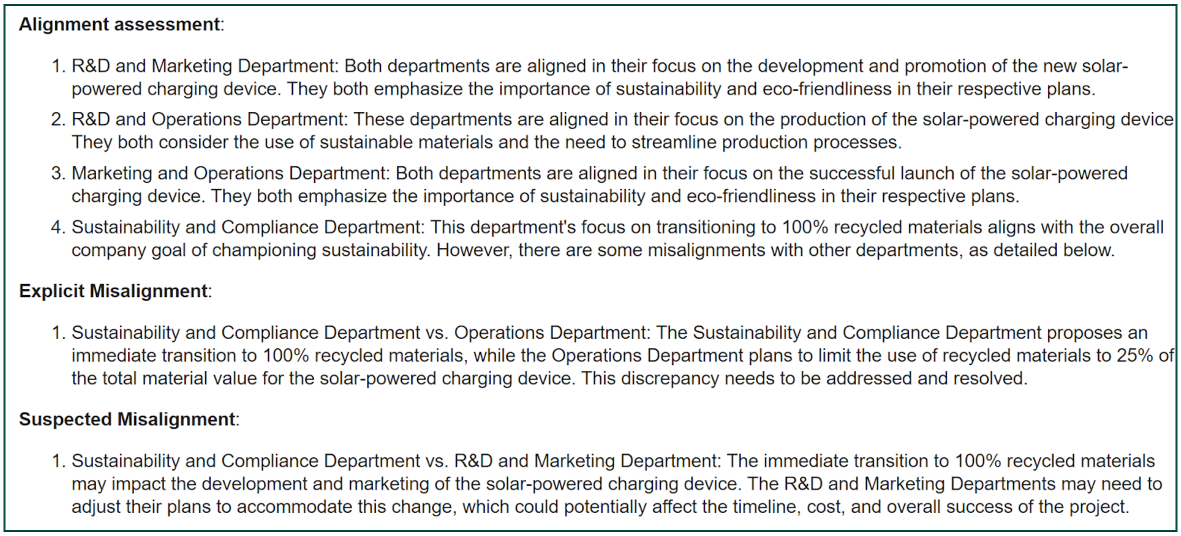
Organization design paradigm vision - digital twin mesh



One approach is to create an LLM-powered intelligence digital twin for every unit and logical entity, such as offering, customer, market segment, product, etc., in the organization. The focused digital twin is built using curated datasets (both internal and, potentially, external) and curated business logic and priorities. If. one organizational unit plans a new initiative, it can query all other units on whether the initiative is aligned and where the misalignments are.

A customer/market-specific digital twin is curated via CRM, offering, customer service data, etc., and external curated datasets like news analytics, job listing insights, patents, and investors. Then the Sales, Marketing, and Offering units can check how specific ideas match specific clients.

In other words, we would model the organization, make relevant structures, and mirror structures, entities, units, and relationships via data and use LLM to add intelligence and flexibility of harnessing data and business logic automatically to new queries and needs. The LLM layer makes this digital twin mesh flexible and able to adapt to different situations.



Alignment assessment:

1. R&D and Marketing Department: Both departments are aligned in their focus on the development and promotion of the new solar-powered charging device. They both emphasize the importance of sustainability and eco-friendliness in their respective plans.
2. R&D and Operations Department: These departments are aligned in their focus on the production of the solar-powered charging device They both consider the use of sustainable materials and the need to streamline production processes.
3. Marketing and Operations Department: Both departments are aligned in their focus on the successful launch of the solar-powered charging device. They both emphasize the importance of sustainability and eco-friendliness in their respective plans.
4. Sustainability and Compliance Department: This department's focus on transitioning to 100% recycled materials aligns with the overall company goal of championing sustainability. However, there are some misalignments with other departments, as detailed below.

Explicit Misalignment:

1. Sustainability and Compliance Department vs. Operations Department: The Sustainability and Compliance Department proposes an immediate transition to 100% recycled materials, while the Operations Department plans to limit the use of recycled materials to 25% of the total material value for the solar-powered charging device. This discrepancy needs to be addressed and resolved.

Suspected Misalignment:

1. Sustainability and Compliance Department vs. R&D and Marketing Department: The immediate transition to 100% recycled materials may impact the development and marketing of the solar-powered charging device. The R&D and Marketing Departments may need to adjust their plans to accommodate this change, which could potentially affect the timeline, cost, and overall success of the project.

Competitive advantage option: the ability to change processes with tech, data & AI

For a while now, one of the biggest bottlenecks in extracting business impact from technology adoption has been the ability to change processes to take full advantage of the available technology. The progress of technology, especially now with Generative AI is not going to slow down so probably the biggest enterprise bottleneck to results will be even more the ability to adopt tech, adapt and change how processes & customer experience is created. This can be turned around, meaning that organizations that build capability to continuously adapt technology and change operations accordingly might be able to build competitive advantage.

Sustainability

Like almost every major technological breakthrough in history, generative AI, too, generates both positive and negative change.

Environmental impact

LLMs are very expensive and computationally intensive to train. The environmental impact is not insignificant.

- Risks and Benefits of Large Language Models for the Environment I Environmental Science & Technology
- Environmental impact I CS324
- Here Comes the Sun! Why Large Language Models Don't have to Cost the Earth

Societal Impact

We have yet to see how generative AI impacts the employment market, but it will probably be drastic. For discussion's sake, we can already look at a couple of current examples.

- *Gig economy*. What happens to e.g. visual artists at marketplaces like Fiverr? Will organizations use generative AI directly to generate designs or will they keep subcontracting from AI-enabled-artists? https://www.zdnet.com/article/this-is-how-generative-ai-will-change-the-gig-economy-for-the-better/
- *Knowledge sharing.* Generative AIs are trained with publicly available data, e.g., Stack Overflow, where people help each other via questions & answers. But there are already anecdotal comments in social media that people no longer contributing because the social rewards have shrunk and their contributions "only feed the AI"

Governance impact

We will delve into this area of generative AI's impact in future versions of our working paper. Stay tuned.

Closing remarks

In conclusion, the landscape of generative AI is both fascinating and fast-evolving, ushering in a new era of possibilities across diverse sectors. It is not merely a technology but a transformative force, reshaping the way we create, innovate, and interact with our digital environment and organizations. In order to harness the full benefits, our thinking should evolve from individual use cases to more systemic change and end-to-end thinking.

However, it is crucial to acknowledge that the pace of innovation in this field is staggering. The paper provides foundational insights at this point in time (October 2023). We encourage readers to view this work as a stepping stone, staying updated through ongoing research and emerging resources. Embrace the fluid nature of generative AI, stay curious and engage with the evolving landscape.

Thank you for reading! We will continue exploring this transformative domain, where the future is shaped by continual innovation, discovery and challenging the conventions.

Once again, please get in touch with any feedback, comments, questions or ideas you may have!

Tuomas Syrjänen, tuomas.syrjanen@futurice.com, +358505470386

Content credits: Teemu Toivonen, Joonas Nissinen, Christoffer Ventus, Jack Richardson, Rachhek Shestra, Heidi Voss, Santeri Vilos, Ville Takanen, Kaj Pyyhtiä, Seth Peters, ChatGPT, Azure OpenAI APIs, and AutoGPT

Appendices

Appendix 1. Futurice Connected Company

Size doesn't slow companies down. Disconnection does.

As companies grow, they also grow more complex. Work is divided into functional silos. Leaders are separated from frontline workers, and from clients and customers. And communication starts to break down. **The true problem isn't size. It's disconnection.**

Enter the Connected Company: where data & AI builds bridges between leaders, teams, clients and markets.

BENEFITS The power of Connected Companies

Using data to build a Connected Company can transform your organisation:

Knowing more, so you can do more. At the heart of the Connected Company is a key observation: most large organisations don't know what they know. If they did, they would be able to move and grow faster.

We learnt this ourselves when we built FutuCortex: a powerful tool for joining up our knowledge from different IT systems across the company. This allows us to quickly uncover knowledge & experts in, say, autonomous transportation. Or retail marketing. Or anything. And it happens automatically, without the need for manual taxonomy or tagging.

It's a simple way for our team to build connections that make them more productive and more successful – which has never been more important than right now. In a Connected Company, data isn't just noise. It's a way for you and your team to harness the knowledge you already have all around you.

Finding a better flow

Accelerating the flow of work can transform large organisations. It frees them to explore new ideas, change direction quickly, and become more responsive to clients and customers. At the same time, teams become more motivated, because they can quickly see the impact of their work.

The concept of flow isn't new: many manufacturing companies moved to just-in-time processes decades ago. But this flow seldom extends across the rest of an organisation. Luckily, data can help.

We worked with a client in the automotive industry, where sales and marketing were disconnected from R&D and R&D from manufacturing. This meant a critical piece of information – emissions impact analysis – was missing from the R&D process. The result? Failed certifications, wasted R&D effort, and lost sales. A new simulation tool connected different departments, and made the work flow and led to a higher certification pass rate for new cars.

In a Connected Company, better flow brings better efficiency and effectiveness, making your entire organisation more responsive.

CONCLUSION From reactive to resilient

At Futurice, we focus on resilience: the power of a company to adapt and grow in a changing world. Connected Companies are resilient by design. They use data to understand the past, present and future, and the differences between all three. In this way, they become more closely connected to their market, their customers, and themselves.

Connected leaders can lead from the front, armed with a better understanding of the organisation they are in charge of. And a connected team is empowered to achieve more, both individually and together. As a Connected Company, you can harness the power of data to become resilient. And when the world changes, you can change with it.

Appendix 2 : Amazon Six Pagers by ChatGPT

Amazon is known for its unique meeting culture, specifically, the six-page memo, or "six-pager." This approach is intended to foster deeper understanding, thoughtful discussion, and ultimately, better alignment. Here's how it works:

Preparation: Before the meeting, the person or team who has called the meeting prepares a detailed six-page memo. This memo lays out the topic of the meeting in a narrative format, presenting relevant information, analysis, and arguments. The goal is to provide a comprehensive, yet concise, overview of the topic, and to propose a course of action.

Reading Time: At the beginning of the meeting, everyone in the room spends the first 20-30 minutes silently reading the memo. This "study hall" approach ensures that everyone has understood the memo, and is ready to discuss it. It also respects the work put into the memo and ensures that everyone starts the discussion on the same page.

Discussion: After everyone has finished reading, the meeting proceeds with a discussion of the memo. The author does not present the memo; instead, it stands on its own, and the meeting is dedicated to discussing it. This fosters a deeper, more thoughtful discussion and helps to avoid groupthink.

The six-pager approach to meetings has several advantages:

It forces clarity of thought: Because the memo needs to be written in a narrative format and fit within six pages, the author needs to think deeply about the topic and present their ideas clearly and concisely.

It fosters understanding: By giving everyone the chance to read the memo at the start of the meeting, it ensures that everyone understands the topic at hand, even if they come from different parts of the organization or have different levels of prior knowledge.

It encourages thoughtful discussion: Because the meeting is devoted to discussing the memo, rather than presenting it, it encourages deeper, more critical thinking, and helps to uncover assumptions, challenge ideas, and explore different perspectives.

This approach aligns well with Amazon's leadership principles, particularly "Dive Deep," "Earn Trust," and "Have Backbone; Disagree and Commit." It's a tool for alignment because it encourages everyone to engage with the issues at a deep level, understand each other's perspectives, and reach a shared understanding or decision.

Appendix 3 : RFC based alignment - by ChatGPT

A Request for Comments (RFC) is a method used in the software and IT world to propose changes, gather input, and drive decision-making. RFCs originated within the internet development community but have become more popular in the software industry at large due to their ability to enhance alignment and communication.

The RFC process often involves the following steps:

Propose: An individual or team drafts an RFC document outlining a problem and a proposed solution. This could be a new feature, a change to an existing system, a technical strategy, or any other significant decision or change.

Review: The RFC is then circulated among relevant stakeholders, who may include other engineers, product managers, or anyone else who may be affected by or have insights into the proposal.

Discuss: Stakeholders provide feedback on the RFC, asking questions, raising concerns, or suggesting changes. This discussion may occur on a mailing list, a forum, a chat platform, or any other shared communication channel. It's important to have a culture of open and respectful discussion, to ensure all perspectives are considered.

Decide: After a period of discussion, a decision is made. This could be a consensus decision, a decision made by a designated authority (such as a tech lead or committee), or a decision made by the original author(s) taking into account the feedback received.

Implement: If the decision is to proceed, the change is implemented. The RFC then serves as a historical document, recording the decision-making process and rationale for future reference.

The RFC process can help enhance alignment in several ways:

Clear Communication: By documenting proposals in writing, RFCs can ensure that everyone has a clear, shared understanding of the problem, the proposed solution, and the rationale behind it.

Inclusive Decision-Making: RFCs invite feedback from a broad range of stakeholders, ensuring that multiple perspectives are considered, and helping to build consensus.

Transparency and Accountability: The RFC process is generally open and visible to all, ensuring that decisions are made in a transparent way, and providing a record of the decision-making process.

Learning and Improvement: By documenting decisions and their rationales, RFCs can provide valuable insights for future decision-making, helping to improve and refine processes over time.

However, it's important to note that the RFC process also requires careful management to ensure it is effective and does not become a bottleneck. This might include guidelines on when an RFC is required, clear timelines for feedback and decisions, and systems for managing and tracking RFCs.

Appendix 4: Structured data extraction by fine-tuned LLama 2

One major time-sink in working with large amounts of unstructured text is that in order to get some use from it - you first need to spend the time to extract and organise information contained within. Traditionally, this has been a challenging NLP problem often attempted to be solved with keyword extraction, or phrase extraction. However, with the introduction of LLMs, we can offload the reasoning of deciding which bits of the text to extract. We simply give the LLM a structure to follow (what types of information and in what format) and then allow it to make the decisions. This task can get expensive and time consuming when dealing with a large number of texts/samples.

In the current landscape there are a few ways to do this. Let’s look at a couple of the most common options:

1. **Few-shot prompting** - showing the LLM several examples of what you want it to do, then passing it a new sample and hoping it follows the same pattern you have shown it.

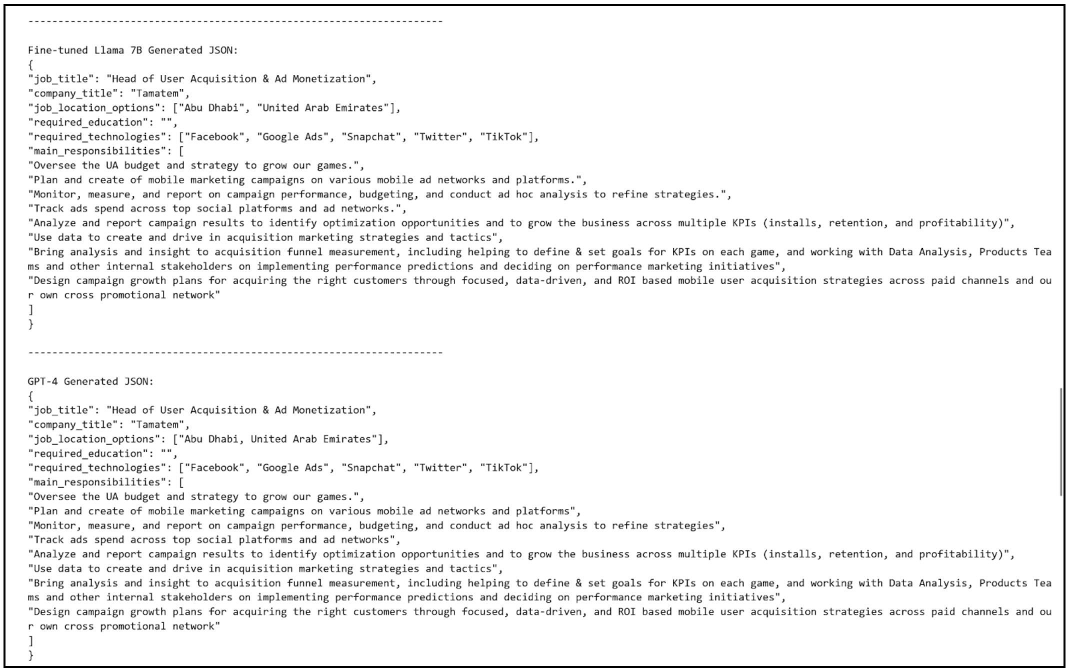
    a. *Pros*:
        i. Relatively low effort - mostly just structuring your prompt.
    b. *Cons*:
        i. High costs - larger context to show examples.
        ii. Reliant on Open AI APIs (also costly)
        iii. May run out of context tokens.
        iv. Unreliable output

2. ***Function calling with Open AI*** - passing a structured object to the Open AI API and forcing it to attempt to fill that structured object with the text that you pass to it.
    a. *Pros*:
        i. Low effort - just pass an object to the Open AI API
    b. Cons:
        i. Reliant on Open AI APIs (also costly)
        ii. Unreliable output, however typically better than Few-shot prompting.

We have found that fine-tuning a smaller, open source model can be a reasonable solution to this problem in many situations. We validated this idea by fine-tuning **LLama-7B chat** on a dataset of *job description / JSON extraction* pairs. Then when tasking our fine tuned Llama model to extract information from a job description that it had not previously seen, the resulting JSON object (format and content) was equivalent to if not better than results from GPT-4 in many cases. Let's review:

3. **Fine-tuning open source model** - creating high quality training data of input/expected output pairs for fine-tuning an open source model to learn more reliably what you expect.
    a. Pros:
        i. Lower costs when doing batch inference (spin up endpoint, batch inference, tear down endpoint).
        ii. Reliably output expected format for ingestion by another system.
        iii. Rate-limit is much higher than Open AI's APIs.
    b. Cons:
        i. Self-managed infrastructure.
        ii. Training time of a few hours.
        iii. Higher effort.

Example screenshots comparing fine-tuned LLama and GPT-4:

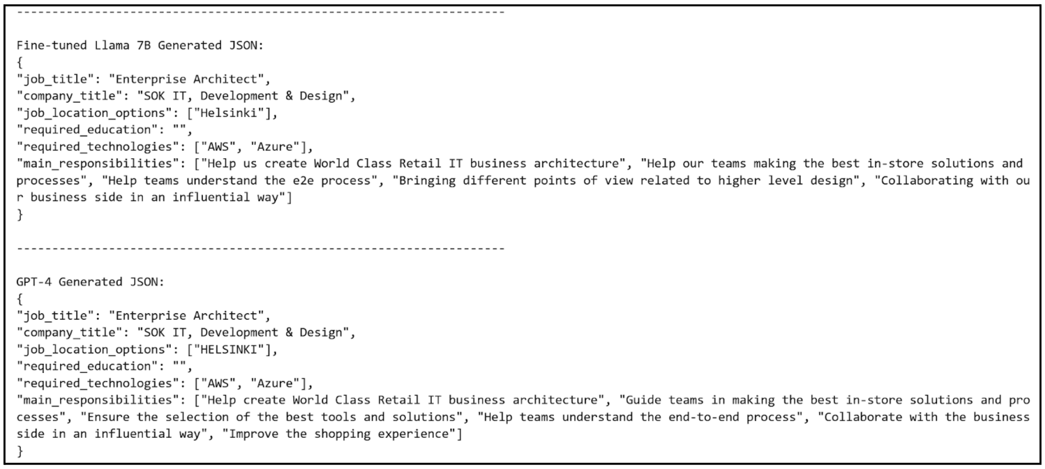


```
----------------------------------------------------------------------

Fine-tuned Llama 7B Generated JSON:
{
"job_title": "Head of User Acquisition & Ad Monetization",
"company_title": "Tamatem",
"job_location_options": ["Abu Dhabi", "United Arab Emirates"],
"required_education": "",
"required_technologies": ["Facebook", "Google Ads", "Snapchat", "Twitter", "TikTok"],
"main_responsibilities": [
"Oversee the UA budget and strategy to grow our games.",
"Plan and create of mobile marketing campaigns on various mobile ad networks and platforms.",
"Monitor, measure, and report on campaign performance, budgeting, and conduct ad hoc analysis to refine strategies.",
"Track ads spend across top social platforms and ad networks.",
"Analyze and report campaign results to identify optimization opportunities and to grow the business across multiple KPIs (installs, retention, and profitability)",
"Use data to create and drive in acquisition marketing strategies and tactics",
"Bring analysis and insight to acquisition funnel measurement, including helping to define & set goals for KPIs on each game, and working with Data Analysis, Products Tea
ms and other internal stakeholders on implementing performance predictions and deciding on performance marketing initiatives",
"Design campaign growth plans for acquiring the right customers through focused, data-driven, and ROI based mobile user acquisition strategies across paid channels and ou
r own cross promotional network"
]
}

----------------------------------------------------------------------

GPT-4 Generated JSON:
{
"job_title": "Head of User Acquisition & Ad Monetization",
"company_title": "Tamatem",
"job_location_options": ["Abu Dhabi, United Arab Emirates"],
"required_education": "",
"required_technologies": ["Facebook", "Google Ads", "Snapchat", "Twitter", "TikTok"],
"main_responsibilities": [
"Oversee the UA budget and strategy to grow our games.",
"Plan and create of mobile marketing campaigns on various mobile ad networks and platforms",
"Monitor, measure, and report on campaign performance, budgeting, and conduct ad hoc analysis to refine strategies",
"Track ads spend across top social platforms and ad networks",
"Analyze and report campaign results to identify optimization opportunities and to grow the business across multiple KPIs (installs, retention, and profitability)",
"Use data to create and drive in acquisition marketing strategies and tactics",
"Bring analysis and insight to acquisition funnel measurement, including helping to define & set goals for KPIs on each game, and working with Data Analysis, Products Tea
ms and other internal stakeholders on implementing performance predictions and deciding on performance marketing initiatives",
"Design campaign growth plans for acquiring the right customers through focused, data-driven, and ROI based mobile user acquisition strategies across paid channels and ou
r own cross promotional network"
]
}
```



```
----------------------------------------------------------------------

Fine-tuned Llama 7B Generated JSON:
{
"job_title": "Enterprise Architect",
"company_title": "SOK IT, Development & Design",
"job_location_options": ["Helsinki"],
"required_education": "",
"required_technologies": ["AWS", "Azure"],
"main_responsibilities": ["Help us create World Class Retail IT business architecture", "Help our teams making the best in-store solutions and
processes", "Help teams understand the e2e process", "Bringing different points of view related to higher level design", "Collaborating with ou
r business side in an influential way"]
}

----------------------------------------------------------------------

GPT-4 Generated JSON:
{
"job_title": "Enterprise Architect",
"company_title": "SOK IT, Development & Design",
"job_location_options": ["HELSINKI"],
"required_education": "",
"required_technologies": ["AWS", "Azure"],
"main_responsibilities": ["Help create World Class Retail IT business architecture", "Guide teams in making the best in-store solutions and pro
cesses", "Ensure the selection of the best tools and solutions", "Help teams understand the end-to-end process", "Collaborate with the business
side in an influential way", "Improve the shopping experience"]
}
```

Appendix 5: Links and Further reading

What	Links
LLM/GenAI community	https://huggingface.co/
HBR podcast how Generative changes everything from strategy to productivity	https://hbr.org/podcast/2023/04/special-series-how-generative-ai-changes-everything
Environmental impact	https://stanford-cs324.github.io/winter2022/lectures/environment/