Transforming IT Consulting with Machine Learning Real-World Applications, Frameworks, and Strategies

Léonard Gonzalez

French School of Electronics and Computing Sciences, Paris, EFREI 94800
Sopra Steria - Machine Learning Engineer - ICS I2S - 2024
leonard.gonzalez@outlook.fr

Acknowledgements

I would like to express my deepest gratitude to my tutor, Anthony Pinto, Leader NextGen Tech for Ops at Sopra Steria, for his invaluable guidance and support throughout the construction of this work. My sincere thanks also go to Benjamin Mirande, Head of operations - ICS at Sopra Steria, for his assistance and encouragement during this project. I wish to thank Laurent Cetinsoy, Pedagogical Engineer and Machine Learning Expert, for his help in answering my theoretical questions. Lastly, I am grateful to Rostom Kachouri, Phd-Engineer in signal-image processing, for evaluating my work.

Abstract

From 2017 forward, the rise of artificial intelligence technologies is reshaping industries on a global level, including the consulting industry. This work analyzes the impact of AI on the IT consulting field, by demonstrating how machine learning tools are revolutionizing traditional consulting models, and how the integration of AI is remodeling the value proposition of consulting services. These subjects are addressed through two corporate case studies implementing machine learning technologies showcasing great promises, RAG/Fine-Tuning and AI-Agents, and are completed with afterthought about future implications for both consultants and clients. Additionally, in order to support these ideas, this paper introduces a new machine learning concept based on the two case studies : the Mixture of Dynamic Agents (MoDA). By studying these, this paper investigates the impacts of machine learning on IT consulting, focusing on potential strategic changes, operational challenges and process evolutions. Ultimately, it provides insights into how consulting practices can adapt to use machine learning technologies effectively and sustainably.

Table of contents

1. Introduction

a. Context and Importance

The IT consulting field is, nowadays, at a pivotal juncture in terms of digital disruption. We define this as the impact of new Artificial Intelligence (AI) technologies and AI business models on existing goods, services, and existing business models.[1]

Over the past decade, literature and businesses have highlighted the potential of AI and Machine Learning (ML), leading the IT consulting industry to acknowledge that moving and evolving symbiotically with AI solutions will soon be conditional to avoid market loss or even disappearance, announcing the emergence of AI solutions as a crucial aspect beyond a mere technological shift.[2]

As a result, AI, ML and deep learning (DL) technologies are becoming increasingly popular in companies, enabling them to leverage large data quantities to improve system performances and accelerate business development. [3]

According to a Forbes Advisor survey[4], key applications according to business owners include customer service (56%), cybersecurity and fraud management (51%), customer relationship management (46%), digital personal assistants (47%), inventory management (40%), and content production (35%). Other uses involve product recommendations (33%), accounting (30%), supply chain operations (30%), recruitment and talent sourcing (26%), and audience segmentation (24%). As we can see, all of these fields fall into the spectrum of activities of the IT consulting industry.

It is important to underline that this survey[4] invites us to remember that significant concerns remain for businesses willing to integrate AI. Data shows that 30% of respondents worry about misinformation, and 24% fear it could harm customer relationships. Privacy issues are also very notable, with 31% of businesses apprehensive about data security and data privacy. Additionally, technology dependence is a major concern for 43% of business owners, followed by the need for technical skills to use AI (35%) and the potential reduction of the human workforce (33%).

b. Objectives

The primary objective of this work is to explore and understand the current transformative impact of AI on the consulting industry, and secondly to introduce several tools, platforms and concepts that are altering the way consulting services are delivered and valued.

Furthermore, through comprehensive analysis and case studies, this paper provides insights about how consulting practices can effectively and sustainably evolve to leverage AI technologies, and insights concerning broader future implications for the IT consulting sector, both highlighting several potential strategic shifts that firms may need to adopt in response to AI development.

The ultimate goal is to offer a roadmap for consulting firms aiming to transform their processes through AI, ensuring they remain competitive and capable of delivering enhanced value to their clients in a rapidly evolving field.

c. Company and Missions

This production is conducted under the supervision of Sopra Steria SAI, leading consulting, digital services, and software development company, currently in the top 25 IT services companies in the world [5], top 5 in Europe, with 5.8 Mrd € turnover in 2023 and 56 000 employees in 30 different countries [6]. The current operations of Sopra Steria ICS I2S

are perfect examples of encounter points between AI and consulting, thus providing the ideal environment for conducting this study.

Notable projects and initiatives involving AI at Sopra Steria include predictive analytics tools, automated customer service platforms, and intelligent data processing systems. These projects will be presented and studied later on in this paper. These constitute a strong base to demonstrate the practical applications of AI in consulting and to highlight the strategic importance of integrating advanced technologies to enhance service delivery and client value.

d. Research methodology

The research design for this work uniquely employs a qualitative approach, to provide an analysis of how AI will impact the IT consulting field.

This paper ambition is to answer the following questions :

- How may the incorporation of AI in IT consulting impact the traditional processes and project management within the industry ?
- In what ways could the necessary skill sets for IT consultants evolve as machine learning technologies are increasingly adopted ?
- What strategies changes and operational challenges do IT consulting firms will have to operate and face when integrating AI into their services, and how to mitigate them ?

The data collection method used in order to answer the previous questions will lie on personal research through academic papers, academic journals, market research reports, industry publications, websites and books.

2. Frameworks and State of the Art

a. Machine Learning Fundamentals

Please note that this section will only cover basic notions needed to understand the work produced in this paper.

We can define artificial intelligence systems as a set of mathematical and algorithmic techniques, often applied through computer tools, aimed at mimicking human thinking.

Within AI, there is a subset called Machine Learning, and within it, there is a further subset called Deep Learning. We can see DL as a more complex form of ML. These systems learn and improve from experience without being explicitly programmed. [7]

The notion of embeddings designate the process where an input object, comprehensible by a human (text, image, sound, video, etc) is transformed through mathematical functions into a mathematical object usable by an AI algorithm . Typically, this mathematical object is a matrix or a vectorial space in n dimensions that aims to accurately represent the meaning of the initial object and allow mathematical manipulations.

An AI model is simply a mathematical function that takes an input and modifies it to give an output. Let's take a look at two fundamental equations in order to approach basic AI models concepts.

Linear regression analysis is used to predict the value of a variable based on the value of another variable. For example, the price of a house based on its characteristics. The given formula is :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

- y: Dependent variable (target variable).
- β_0: Intercept (the value of y when all x are 0).
- $\beta_1, \beta_2, \ldots, \beta_n$: Coefficients for each independent variable $x_1, x_2, \ldots, x_n$ (representing the change in y for a one-unit change in each x).
- $x_1, x_2, \ldots, x_n$: Independent variables (predictors).
- ϵ: Error term (residuals).

On the other hand, logistic regression analysis is used to classify an input. For example, recognize if a picture is a cat or a human. The given formula is :

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

- $P(y = 1|x)$: Probability that the dependent variable y is 1 given the independent variables $x_1, x_2, \ldots, x_n$
- β_0: Intercept (the log-odds of the outcome when all x are 0).
- $\beta_1, \beta_2, \ldots, \beta_n$: Coefficients for each independent variable $x_1, x_2, \ldots, x_n$ (log-odds change in y for a one-unit change in each x).
- $x_1, x_2, \ldots, x_n$: Independent variables (predictors).
- e: Base of the natural logarithm

These two simple examples introduce the concepts of models, features and weights. The model is the given function, the features ($x_1, x_2, \ldots, x_n$) are what you give as input to your model (ex: a sentence embedded as a vector), and the weights ($\beta_1, \beta_2, \ldots, \beta_n$) are what will be set during training in order to correctly set the model's function.

Gradient Descent is given in the form of :

$$\nabla f(x) = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right]$$

$$\theta_{\text{new}} = \theta_{\text{old}} - \alpha \nabla_\theta f(\theta)$$

With θ_{new} the updated parameter value. We will simply assume here that Gradient Descent is used inside the training process of a model in order to make it better itself iteratively.

Given these definitions, we define in this study that a Large Language Model (LLM) is a mathematical function defined by its weights, taking as input objects defined by features, and results of an auto-ameliorative, iterative training process.

Lastly, defining knowledge as a concept isn't in the scope of this study, and the philosophical aspect of it will not be discussed. However, we will define what knowledge is for a Large Language Model[8]. If a LLM acknowledges a fact, it can accurately and consistently answer questions about it and reliably separate true and false statements related to it. This definition can be extended to an entire knowledge base, not just some individual facts.

Mathematically, let $Q = \{q_n\}_{n=1}^N$ be a set of N multiple choice factual questions, where each question has L possible answers and exactly one correct answer.

Let $A = \{(a_1^n, \ldots, a_L^n)\}_{n=1}^N$ be the corresponding set of possible answers, and $C = \{c_n\}_{n=1}^N$ be the correct answers.

Let M be a language model. We denote by $M(q_n) \in \{a_1^n, \ldots, a_L^n\}$ the predicted answer of the model to the n-th question.

We define the knowledge score L of M in relation to Q to be the standard accuracy score:

$$L_{M,Q} := \frac{\#\{q_n | M(q_n) = c_n\}}{N}.$$

We say that the model M possesses any knowledge regarding the set of questions Q if the following holds:

$$L_{M,Q} > \frac{1}{L}.$$

We understand the model can consistently outperform a simple random guessing baseline. Naturally, if the knowledge score $L_{M,Q}$ is higher for one model compared to another, then we assert that we can rank them based on accuracy score $L_{M,Q}$.

b. Literature Review

In this literature review, this paper will provide information about three key concepts of machine learning usages in order to understand the case studies later on. The three following concepts are : Retrieval Augmented Generation (RAG), Fine-Tuning, and AI Autonomous Agents.

Retrieval-Augmented Generation (RAG) is an advanced technique in natural language processing that combines retrieval mechanisms with generation capabilities to enhance the performance of language models. RAG consists of two main components: the retriever and the generator. [9]

The retriever component encodes all text chunks in a dataset into vectors using embedding techniques. It also encodes the query or prompt using the same techniques. As a result, we are left with a vectorial space containing the documents (each one stored as a vector) and separately the vectorized user's query. The retriever then finds the most similar text chunks in the dataset by using linear algebra. For example, a simple cosine similarity comparison applied between the user's query and the vectorial space results in a ranking of the most similar documents. These best fitting documents are then retrieved to provide additional context to a Large Language Model (LLM) for answering questions more precisely.

The generator component is typically a large language model (LLM). It generates answers based on both the query and the retrieved information.

To simplify, RAG allows you to automatically add to your queries relevant context, making an LLM educated about a specific set of documents.

During training, both the retriever and generator are trained end-to-end. Given a query, the retriever identifies relevant documents from a document index. The generator then uses this retrieved information along with the query to generate an answer. By retrieving relevant information from external sources, RAG can incorporate new knowledge that was not part of its initial training data. Experiments have shown that RAG improves performance on tasks like question answering and fact verification by providing more accurate and contextually relevant answers.

Training both components end-to-end can be computationally expensive. To mitigate this, some approaches use pre-trained retrievers or asynchronous re-indexing during training. [10] For specialized domains, tuning both document and query encoders can improve performance significantly. Techniques like reconstruction-based domain adaptation [11] help ensure retrieved passages are relevant to specific domains. Some papers propose methods for better domain adaptation by asynchronously re-indexing datasets during training [12]. This approach enhances LLMs' ability to incorporate retrieved knowledge without end-to-end training, making it more scalable for larger models.

RAG represents a significant advancement in leveraging external knowledge sources to improve LLMs' performance on various NLP tasks. By combining retrieval mechanisms with generation capabilities, it addresses key limitations related to incorporating new information and adapting to specific domains. Future research revolves around Knowledge Graphs (KG) [13]. KG allows to create contextual relationships between stored vectors, thus improving the retrieval process and adding another comprehensive layer to the RAG process.

Let's now approach the concept of Fine-Tuning. Fine-tuning is a critical concept in machine learning where pre-trained models are adjusted using new datasets for improved task-specific performance. Fine-tuning involves taking a pre-trained model and making small adjustments using a new dataset—leveraging existing knowledge captured during initial training phases allows efficient adaptation compared to starting from scratch.

Fine-Tuning a model includes several steps, such as ; Pre-training, transfer learning, layers adjustment and optimization. Applications span across various domains including Natural Language Processing (NLP), Computer Vision, and Speech Recognition. There are a lot of advantages using these technologies, enhanced efficiency, improved performance, and greater flexibility in handling different tasks. However, there are also challenges that need to be addressed, such as overfitting, maintaining data quality, and the complexities of hyperparameter tuning.

Recent developments focus on improving techniques via Layer-wise adaptive rate scaling [14], differential privacy techniques [15], and meta-learning approaches [16]. Fine-tuning remains an essential modern machine learning workflow due to its efficiency and adaptability.

The future of AI-enhanced knowledge retrieval lies in the combination of both RAG and Fine-Tuning, in order to capture the best of these two technologies.

Now that we have covered knowledge and skills retrieval with RAG and Fine-Tuning, we will educate ourselves about AI agents, or how to use learned knowledge and skills in order to complete complex tasks. AI agents are autonomous entities designed to perform tasks requiring human intelligence such as learning, reasoning, problem-solving, perception, and language understanding. Key concepts of AI agents revolve around four main pillars ; operating without human intervention, learning from their experiences and adapting their behavior accordingly, interacting with their environment (including other agents), and behaving in a goal-oriented manner in order to achieve specific objectives.

An AI agent team is constituted of several agents, each one specialized in a limited number of tasks. These agents are considered "workers" because of their role. These workers are coordinated towards a goal by a "proxy" (human or AI). Each agent, workers or proxy, is given a LLM model to operate and a specific set of "tools", which are coded functions for them to use. [17] There are several types of AI agents, such as reactive, deliberative, hybrid and learning. [18]

Reactive agents respond to stimuli from the environment without using internal states or history, and follow simple rules without planning ahead. Deliberative agents use internal models of the world for decision-making based on planning and reasoning about future actions. Hybrid agents combine reactive and deliberative approaches to leverage strengths of both methods. Learning agents improve their performance over time through learning mechanisms such as reinforcement learning or supervised learning.

AI agents represent significant technological advancements capable of transforming various industries by automating complex tasks traditionally performed by human teams.

Now let's assume that any Agents (including an AI proxy) can be given :

- A (dynamically) Fine-Tuned model as LLM
- Autonomous Fine-Tuning capacity as a tool
- RAG + KG system as a tool
- Internet Scraper as a tool

This paper will now refer to this concept when speaking of a Mixture of Dynamic Agents (MoDA).

It is important to specify that the MoDA concept is very close to a Mixture of Experts (MoE) [19], such as the supposed architecture of GPT-4o. Briefly, MoE is the concept of addressing user's inputs by using different models trained in specific fields, and then merging the outputs. The main difference of MoDA is the capacity to retrieve information, knowledge and behavior evolutions dynamically.

c. Technologies and Tools

In this section, we will present the different technologies used in the following case studies.

For databases and containers, PostgreSQL with the PGvector extension are used for vector storing and embedded vector manipulation (such as cosine similarity search). PostGreSQL is the standard database in Sopra Steria proof of concept development. PostgreSQL is an object-relational database management system (ORDMBS), which means that it has relational capabilities and an object-oriented design. The PGvector extension allows vector database storing, and vector manipulation from python code.

Neo4j serves as a graph database for complex relationship mapping, allowing RAG models to automatically generate Knowledge Graphs. Neptune offers managed graph database services in the same manner, but will be used in a cloud deployment, when Neo4j will be used as a local tool.

OVH provides cloud infrastructure solutions, allowing us to Deploy LLMs, Fine-Tune them, and deploy them as API. The reason for the OVH choice as AI cloud services provider is linked to data governance, and location of their infrastructures (Europe).

Docker is a container solution used here as a base for OVH deployment, in order to ensure consistent environments across various stages of development and deployment.

Now that we've seen the database and container section, let's take a look at LLMs used.

Large Language Models (LLMs) are AI mathematical models designed to understand and generate human-like text based on vast amounts of data. The most famous ones are the OpenAI's GPT (Generative Pre-trained Transformer) LLMs family.

OpenAI's API provides access to their LLMs family, enabling the integration of advanced language understanding and generation capabilities into applications. OpenAI's API is currently the easiest and fastest solution to implement AI in systems.

Hugging Face is a platform offering tools and libraries for building and deploying machine learning models with a focus on natural language processing. Where OpenAI's do not release open sourced models, Hugging Face is a platform that shares all open sourced LLMs. Hugging Face also provides insights about models and their capabilities, by testing each LLMs through different world-famous benchmarks, recognized and validated by the AI community.

Both OpenAI and Hugging Face platforms provide pre-trained models and easy-to-use interfaces that help developers implement sophisticated NLP tasks without needing extensive expertise in machine learning.

In the following case studies, the models cited are : GPT-3.5, GPT-3.5 fine-tuned, GPT-4, GPT-4o, Llama2 7B, TinyLlama-1.1B and Qwen 1,5/1,8B. The choices of the models are conditioned by Sopra Steria decision, and limit the possibilities to use and test different models.

After understanding LLMs choices, let's discover the frameworks used for both RAG and Fine Tuning.

LlamaIndex, is an open-source specialized data framework designed to support the development of applications based on LLMs. It allows developers to integrate various data sources, including many different file formats and databases. The framework includes multiple connectors that streamline data ingestion for seamless interaction with LLMs.

LlamaIndex features an efficient RAG system that enables developers to input any LLM prompt and receive context-rich, knowledge-augmented outputs. It manages interactions with an LLM by creating an index from the input data, which is then used to respond to related queries. The framework can generate different types of indexes—vector, tree, list, or keyword—but in our case studies, only vectors.

Overall, LlamaIndex provides a set of tools for data ingestion, RAG and integration with various other frameworks.

On the Fine-Tuning side, the chosen framework is Llama-Factory. LLaMA-Factory is again an open-source project. It offers tools and scripts for fine-tuning, serving, and benchmarking any LLMs models locally. This framework give the advantage to be updated fastly and follow the latest techniques improvement in the Fine-Tuning research field, such

as LoRa (Low-Rank Adaptation), QLoRa (Quantization Low-Rank Adaptation) or QaLoRa (Quantization-aware Low-Rank Adaptation). [20]

Lastly, let’s focus on the frameworks used for AI agents automation.

AutoGen is an open-source library developed by Microsoft research lab. It aims to address complex problem resolution through multi-agent collaborations. AutoGen put the emphasis on modularity and NLP-based systems to promote reuse. This framework allows multiple agents to learn and collaborate through a shared context.

Key benefits of AutoGen's multi-agent approach include support for various LLM configurations, native tool usage (code generation and execution), and (optionally) a Human Proxy Agent for integrating human feedback at different levels.

Originating as a spinoff from FLAML (a fast library for automated machine learning and tuning), AutoGen is actively developed as a community-driven project. Collaborators from Microsoft Research, Pennsylvania State University, the University of Washington, Microsoft Fabric, and ML.NET have significantly contributed to the project.

3. First case study - Retrieval Augmented Generation / Fine-Tuning : How to value existing data

a) Mission Description and Objectives

One of the main promises of RAG and Fine-Tuning, as seen previously, is to change the way we retrieve, understand and value data. LLMs store extensive factual information within their pre-trained weights, enabling them to answer a wide range of questions across various domains. However, as established precedently, this knowledge is limited and depends on both training process and training data. As we know, the training process for LLMs is extremely demanding for both computational power and time, to the extent that only large companies, with tremendous means, are today able to create them from scratch. Integrating new information or refining LLMs capabilities using external datasets, for example relative to customers and clients, presents a significant challenge.

The present case study has been a leading topic in the Sopra Steria ICS team in the last year. We easily acknowledge that for any given client, their data have to be valued and integrated in the AI development processes. As a result, the goal of the mission was to be able to integrate any client-specific knowledge into generic LLMs.

b) Challenges

Understanding the subtleties between fine-tuning and RAG could be, at first glance, a difficult task. The theory behind modern techniques of Fine-Tuning was fairly complicated to grasp [22], where RAG logic as a global technique was much more easy to understand.

As defined before, Fine-Tuning is the process of modifying the weights of a given pre-trained model after its initial training. RAG, on the other hand, adds external knowledge into generating models and retrieval procedures in order to increase the relevance and quality of produced replies. As seen previously, RAG intervenes on the features given to a model.

The outputs from both solutions may seem similar, but are not. The nuances lie in how knowledge is ingested by the model, and what purpose the architecture serves. To resume, Fine-Tuning modifies the weights of a model, where RAG modifies the features given to a model.

Fine-tuning necessitates a carefully built dataset. We created a proof of concept about Streamlit code generation, and created around 1000 lines of JSON format in question-response pairs. The replies were gathered from GitHub repositories that include Streamlit apps, using internet scraping automated techniques. The questions were produced automatically using the Qwen LLM, based on documentation and specification also scraped. Then, we used the openAI playground, a cloud based fine-tuning solution, to train a GPT-3.5 basis on our dataset. The resulting model outperformed GPT-4 for the same set of Streamlit questions.

Modern fine-tuning approaches like LoRa, QLoRa, and QaLoRa can be difficult to grasp, but when used effectively, they allow more exact control over the model's behavior. Currently, the next step is to use a more complex approach via Llama Factory, in order to boost the model's performance even more. We will try other open-source models, such as Tiny Llama and the Llama LLM family.

RAG, on the other hand, presents its own set of advantages and disadvantages. Understanding the operation of retrieval systems, particularly the vectorization and comparison algorithms, needs substantial feature engineering. RAG can retrieve particular knowledge effectively, but its inclination toward hallucinations may make it difficult to

generate intelligent replies.

In order to overcome this issue, specific prompting techniques help a lot, and can improve results quality up to 20% in some cases. Consequently, training and educating people become a core solution today to ensure proper functioning of RAG technologies. Another issue faced during this mission was how to address database clustering for secure queries. Given that the vectorial space for retrieval in RAG is unique to the LLM working on it, we are currently investigating multi-vectorial space solutions, where vectorial databases are merged in function of user's permissions.

Ultimately, we created a pipeline for this proof of concept in order to retrieve documents directly from any gitlab or github repository. Our solutions then became fully hosted on cloud by using OVH and AWS Neptune. The next steps include making the document sources more diversified, such as SharePoint and Google Drive.

c) Impact and Results

The current result of this mission is global across Sopra Steria. The RAG and Fine-Tuning techniques propose a new approach to data valuation, and are very widely adaptable across different projects. For clients, this means any datasets can be more accurately reflected in any AI models, offering more personalized and relevant outputs. The proof-of-concept we created demonstrated how tailored a solution could be achieved by fine-tuning a model with a carefully curated dataset. Our shift to using more complex techniques such as LoRa, QLoRa, and QaLoRa through Llama Factory and exploring other open-source models aim to make our solutions even more robust and versatile.

On another hand, this case study demonstrates how machine learning techniques reproducing specific knowledge and behavior can be, nowadays, easily implemented in a workflow. This observation will later be used to underline how MoDAs could change the face of IT consulting.

4. Second case study - AI Hypervisor : Autonomous self-remediation of IT solutions

a) Mission Description and Objectives

AI agents help achieve complex goals for large language models by managing tasks and optimizing performance through a shared context. These agents can automate repetitive tasks, maintain consistency, and allow the models to focus on higher-level problem-solving. This streamlines operations and enhances the capabilities of language models, making them more effective and reliable in providing solutions.

Sopra Steria wanted to be able to monitor, appreciate, detect and remediate server problems on the client side automatically. Naturally, with the recent emergence of highly reliable NLP techniques, the focus went on using LLMs.

The main objective of the Auto Remediation Project (ARP) was to implement a hypervisor able to adapt in real time to any anomaly on any client server park.

b) Challenges

AI-agents come nowadays in very efficient frameworks ready-to-use. As specified before, we choose to work with AutoGen, the microsoft open-source AI-agents solution. Given that this technology was fairly new when we began to work with it, several subtleties were difficult to grasp. Firstly, AutoGen allows a wide range between a lot of human action and full AI autonomy. Ai-agents constitute a very high-level framework, making adjustability and reproducibility difficult to guarantee. These issues have been addressed by trying multiple times on the same use-cases, and then verify if the settings generalize well on unseen use-cases.

The ARP process was a workflow composed of : retrieving any anomaly description, understanding what these error strings mean, generating commands or code to try solving the issue and ensure its cyber-security, monitor the result, generate logs of actions, store the code, and generate/store documentation about it.

In this process, one of the main questions was to ensure good understanding of the anomalies by the model from sometimes short and obscure error strings. We firstly thought of training a specific model to various sets of errors, but the diversity of client's server parks made it nearly impossible with our current means. The solution we settled for was to implement reinforcement learning (RL) and/or reinforcement learning with human feedback (RLHF) based on current human experts in this work field. This solution is currently still in progress.

Surprisingly, the code generated from clear anomalies description was good, even for specific or rare server configuration. Given that the produced code would be executed directly, the difficult part in that step was to ensure cyber vigilance in the coding process. We tried with the GPT family models (GPT-4 and GPT-4o) to verify each generated code and adapt it to be cyber-security compliant. The success was mitigated, obvious flaws and known vulnerability were correctly corrected, but specific vigilance points were not always taken into account.

In order to generate and store documentation from the created code, we simply make an API call on OpenAI's GPT-4o LLM. We investigate the use of Qwen LLMs in order to optimize the processing time for later implementation, but are still at the research stage.

The implementation part of this project doesn't fall into our business specialty spectrum, and will be carried on later by the specialized Sopra Steria polish teams.

The previous steps have been attributed to different agents. Each AI-agent was responsible for a part of the process, and even in some cases, multiple agents were needed for one step of it. The tuning of each AI-agent was linked to ultra-specific prompting (generic and system prompts), model settings (temperature, stop sequences, top P, frequency penalty and presence penalty), roles (AI-agent description, similar to a prompt), maximum iteration (replies) of AI-agent discussion and Proxy settings.

The lack of precise metrics due to the high-level abstraction of AutoGen make the feature engineering process a very difficult task. We addressed this issue by testing each agent separately, then altogether by adding them one at a time.

c) Impact and Results

This mission is currently still under development, but is already used in chatbots in order to automate the service desks processes, and the gestion of IT tickets through ServiceNow. The potential use of AI-agents is very wide, and we are in the beginning of presenting these tools internally inside Sopra Steria. We expect more use-cases to come in the next few months. It is important to underline that we encounter much more reserve from clients (internal or external) about AI-agents tha with other AI tools. The idea to delegate actions to AI seems to be much more difficult to accept, maybe due to the swift change from AI as an helper to a producer in this configuration.

We can note that this case study is representative of how machine learning begins to achieve task segmentation in order to achieve complex goals. In the same manner as RAG/Fine-Tuning case study, this observation will later be used to underline how MoDAs could change the face of IT consulting.

5. Analysis and Synthesis of Case Studies

a) Comparative Analysis

As seen previously, three key components demonstrate significant potential for IT consulting : RAG, Fine-Tuning, and Agent-based systems. RAG is able to retrieve existing data and adapt to new documents, enabling LLMs to fastly acquire knowledge. This allows for rapid adaptation to large datasets (such as client information) facilitating the rapid identification of specific details. However, current technologies are facing challenges concerning stability and reproducibility. Fine-Tuning focuses on learning specific skills and behavior tailored to particular tasks or domains. This method aims to adapt the precision and specificity of global models on a specific field. Agent-based systems represent the capacity to decompose tasks in a subset of tasks in order to achieve complex objectives, functioning similarly to a coordinated team.

It's important to understand that RAG addresses specific knowledge, Fine-Tuning addresses behavior and skills, and Agent-based systems address the capacity to achieve more complex goals.

b) Strategic Impact

Any competitive position in IT consulting will be significantly enhanced through the implementation of machine learning solutions, offering a new way to the proposition of flexibility, responsiveness, and adaptability. These improvements could provide substantial customer value, including the ability to quantify positive results effectively, tailor products more precisely than ever before, and leverage extensive customer experience data to continually enhance outcomes. We can project that these changes may lead to improved reproducibility and increased customer attractiveness and loyalty. [22]

Additionally, our services would be available around the clock (24/7/365) at a fraction of traditional costs since machines are less expensive than human labor. Automation will reduce costs without compromising output quality. These advancements would significantly reduce operational costs and would allow us to scale up the workforce easily anytime in order to expand our operations. [23]

It is important to take into consideration that the adoption of AI introduces new risks such as fears, misconceptions, and ethical concerns about replacing humans [24]. To address these, it is central to develop strategies in order to manage these risks effectively. The created systems will have to guarantee a certain percentage of success and failure rates, track client satisfaction more efficiently, and ensure a smooth transition by capitalizing on data insights.

Moreover, embracing results in duality : AI opens up new markets but presents an ultra-fast changing landscape—businesses must adapt continuously or risk obsolescence. [25] Any approach will have to ensure easy scalability and improvements with growing customer bases. Some projections give that within a few years, global or region-specific regulations will emerge to govern AI usage and limit workforce displacement due to automation. [26] Additionally, green IT and robust data governance practices will be needed to ensure compliance with future regulatory standards. Regulations compliance will be one of the main questions during the AI transition era for IT consulting.

Ultimately, we can add that the substantial cost reductions achieved through AI implementation will translate into significantly improved results, positively impacting stakeholders. [22] This strategic shift is expected to drive stock market performance upwards considerably in business making AI transitions.

6. Strategic Preparations for IT Consulting Firms

From the previous chapters, we can establish that the IT consulting field may tend to transition towards the wide implementation of the Mixture of Dynamic Agents architecture (MoDA). This transition will lead to several strategic impacts as seen before. To prepare for this evolution, firms might consider the following strategic actions.

a) Workflows and Infrastructures

First, it is inevitable to make a shift between human-centric workforce to AI-centric infrastructures. To support the shift towards augmented processes, it is essential to develop dedicated hardware parks and data centers able to withstand AI computational demands. Redesigning workflows to incorporate the AI tools is a major axis for augmenting human capabilities. This involves integrating machine learning algorithms, natural language processing systems, and many other AI technologies into all the existing processes. By doing so, IT consulting firms will be able to enhance productivity through automation of routine tasks and improve decision-making accuracy.

Equally, the IT consulting field has to effectively change management strategies to facilitate a smooth transition from human-centric to augmented processes. This means developing different training programs that will give the human workforce the skills and comprehension about this AI transition and maintaining a clear communication plan will ensure that employees understand the benefits of these changes and feel supported throughout the transition period. Implementing feedback mechanisms can also help in addressing employee's concerns and continuously improving this integration process.

In order to support the AI software requirements, firms will have to significantly develop robust hardware clusters capable of supporting large-scale data processing and machine learning tasks. This includes high-performance computing (HPC) systems equipped with powerful GPUs or TPUs to accelerate computationally intensive processes. Additionally, ensure adequate storage solutions are in place to handle vast amounts of data efficiently. Cloud-based infrastructure may also be considered for its scalability and flexibility advantages.

Due to the rapidly changing AI technology field, companies will have to adopt state-of-the-art tools for data analysis, model training, deployment automation, and performance monitoring. Parallely, they will also have to anticipate adapting the DevSecOps practices to support various stages of the AI lifecycle such as data ingestion, preprocessing frameworks like Apache Kafka or Hadoop; model training environments like TensorFlow or PyTorch; deployment tools such as Kubernetes for container orchestration; and monitoring systems like Prometheus or Grafana to track performance metrics in real-time. Seamless integration of these tools ensures a streamlined workflow from development to production.

b) Talent Acquisition and Development

Secondly, talent Acquisition and development is a core problematic to address in future directions to adopt. MoDA, as presented before, will require gathering rare skills for training purposes, and organizations will have to focus on recruiting individuals or small groups with high or rare skill sets in areas. We can imagine partnering with academic institutions to create pipelines for emerging talent and stay updated on cutting-edge research.

Furthermore, allocating resources towards hiring data experts, machine learning experts, software experts, and specialists in core fields relevant to AI implementation is one main axis for MoDA implementation, continuous development and maintenance. Recruitment

efforts should focus on individuals with proven experience in handling large datasets, designing complex algorithms, and deploying scalable AI solutions.

c) Data Governance Policies

Given the architectures of MoDA and their capacity to retrieve data autonomously from the internet as well as the environment, data governance questions are absolutely central to our topic. It is imperative to formulate and implement stringent data governance policies that emphasize the ethical use of data, robust privacy protection measures, and the maintenance of data integrity. These policies should also ensure transparency in data handling processes and establish clear accountability mechanisms for any breaches or misuse. Additionally, it is crucial to ensure that these policies are in full compliance with existing customer data protection regulations such as the General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA).

To maintain compliance with evolving legal standards and internal policies, organizations must conduct regular audits of their data management practices. These audits should be comprehensive, covering all aspects of data collection, storage, processing, and sharing. They should evaluate the effectiveness of current safeguards against unauthorized access or breaches. These will help identify areas for improvement, ensuring that any deficiencies are promptly addressed. Furthermore, audit findings should be documented meticulously and used to update training programs for staff involved in data handling.

d) Strategic Alliances And Partnerships

As IT consulting businesses progress, forming strategic relationships with major technology companies will be critical to acquiring access to sophisticated tools and platforms, hence expanding the scope and quality of service offerings. These alliances can act as critical points in facilitating the integration of cutting-edge technology into current systems, hence fostering innovation and operational efficiency

Firms should actively engage in industry consortia committed to creating and establishing standards for the future application of MoDA. Participation in these consortia not only assures alignment with industry best practices, but also positions the organization as a thought leader in the future MoDA sector.

IT consultants should try to create closer ties with customers through co-development projects, allowing them to actively participate in and design solutions that are uniquely customized to their unique needs and business circumstances. This collaborative approach may significantly increase client happiness and loyalty by ensuring that offered solutions are both relevant and effective.

Furthermore, IT consulting firms must provide full consultation services geared to help customers through the transition to MoDA frameworks. These services should include detailed evaluations of present digital capabilities, strategic planning for MoDA adoption, training programs, and continuing assistance to guarantee the effective implementation and long-term advantages of new digital initiatives.

e) Innovation and development

To ensure success, IT consulting businesses should create internal laboratories or innovation centers stocked with tools that allow their staff to experiment with new ideas without fear of failure. They should also establish organized programs such as hackathons, innovation challenges, and suggestion schemes to capture and develop new ideas in a methodical manner. Investing in ongoing learning opportunities like training programs, seminars, conferences, and online courses will keep the staff current on changing trends and technology. Finally, firms should form external collaborations with universities, research institutes, startups, and other businesses to get new perspectives and chances for co-innovation.

f) Green IT

It is important to discuss environmental questions, and to imagine what preparations these will bring with them.

First, detailed research on the environmental effect of AI technologies should be conducted to identify companies areas that require action. These studies should cover both direct and indirect implications, such as energy usage, resource use, and possible waste creation from AI research and implementation. Quantitative measures should be developed to systematically assess the environmental impact of AI activities in order to avoid future legislative constraints.

Furthermore, IT consulting businesses must verify that all activities adhere to current environmental rules while anticipating future legislative trends. This necessitates remaining current on national and international environmental legislation related to AI technology, actively participating in policy discussions, and proactively adjusting operational methods to meet or exceed regulatory criteria. Engaging with stakeholders, such as governmental authorities, non-governmental organizations (NGOs), and the general public, is critical for ensuring compliance and contributing to overall environmental sustainability goals.

g) Roadmap

- Retrieve data on the skills, knowledge, and positive behavior of employees.
- Automate and create processes for training models on these data.
- Store and implement them initially as support. Continue to train them as we progress.
- Develop some high or rare skilled individuals into business experts in their respective fields.
- Invest heavily in hardware, both in the cloud and on-premises.
- Invest in machine learning experts, data experts, business experts, and project leaders.
- Quantify the environmental impacts for green IT initiatives.
- Closely follow regulations and the evolution of legislation.
- Create the first MoDAs for emerging markets to test and improve.
- Transition from having many people working differently, to having many people working with AI standards.
- From the previous, increasingly switch to a business architecture involving more MoDA.

7. Conclusion

a) Synthesis

Today, the value proposition of IT consulting firms is to provide their clients with experts in various fields, able to work together in order to achieve complex goals. As seen previously, expertise and capacity to sequence tasks in an efficient way already began to be reproduced efficiently by machine learning technologies, and are conceptually replaceable by MoDAs. The question is no more “If ?” but “When ?”. MoDAs arrivals will bring changes comparable to the emergence of the internet in the 2000s, so preparations are needed to ensure a smooth transition. Furthermore, it is essential for firms to find a balance between leveraging AI technologies and preserving the important human element in their services. In finding the balance, the combination of machine learning capabilities with the unique expertise of several chosen experts will enable firms to deliver exceptional value to clients, as never seen before.

b) Limitations

It is important to remember that we are at the beginning of the emergence of AI. “Attention is all you need” [27], the founding paper that allowed the AI research field to reach a major breakthrough in 2017, and gave birth to the world-famous OpenAI’s GPT LLMs family, was only published 7 years ago. For comparison, the Web opened to the public in 1991. We can make a parallel between the rise of the internet and the future of AI. Consequently, it is early today to make predictions about AI and how business will be reshaped, so please remember that this paper is theoretical and has to be considered as so. The work produced is based on research, real-world cases and engineering level knowledge of the machine learning fields, but at the end, remain personal projections.

8. Bibliography

1
OKUNLAYA, Rifqah Olufunmilayo, SYED ABDULLAH, Norris, et ALIAS, Rose Alinda. Artificial intelligence (AI) library services innovative conceptual framework for the digital transformation of university education. Library Hi Tech, 2022, vol. 40, no 6, p. 1869-1892.

JOHN, Meenu Mary, OLSSON, Helena Holmström, et BOSCH, Jan. Towards an AI‐driven business development framework: A multi‐case study. Journal of Software: Evolution and Process, 2023, vol. 35, no 6, p. e2432.

2
KITSIOS, Fotis et KAMARIOTOU, Maria. Artificial intelligence and business strategy towards digital transformation: A research agenda. Sustainability, 2021, vol. 13, no 4, p. 2025.

3
https://www.intuition.com/machine-learning-by-the-numbers-its-impact-on-business/

4
How Businesses Are Using Artificial Intelligence In 2024, Katherine Haan, Contributor, Rob Watts, Apr 24, 2023, https://www.forbes.com/advisor/business/software/ai-in-business/

5

https://brandirectory.com/rankings/it-services/table

6

https://www.soprasteria.com/fr/investisseurs/a-propos-de-sopra-steria/chiffres-cles

7

https://www.ibm.com/think/topics/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks

8

OVADIA, Oded, BRIEF, Menachem, MISHAELI, Moshik, et al. Fine-tuning or retrieval? comparing knowledge injection in llms. arXiv preprint arXiv:2312.05934, 2023.

9

FINARDI, Paulo, AVILA, Leonardo, CASTALDONI, Rodrigo, et al. The Chronicles of RAG: The Retriever, the Chunk and the Generator. arXiv preprint arXiv:2401.07883, 2024.

10

GUU, Kelvin, LEE, Kenton, TUNG, Zora, et al. Retrieval augmented language model pre-training. In : International conference on machine learning. PMLR, 2020. p. 3929-3938.

11

FARAHANI, Abolfazl, VOGHOEI, Sahar, RASHEED, Khaled, et al. A brief review of domain adaptation. Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020, 2021, p. 877-894.

12

SIRIWARDHANA, Shamane, WEERASEKERA, Rivindu, WEN, Elliott, et al. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. Transactions of the Association for Computational Linguistics, 2023, vol. 11, p. 1-17.

13

EHRLINGER, Lisa et WÖß, Wolfram. Towards a definition of knowledge graphs. SEMANTiCS (Posters, Demos, SuCCESS), 2016, vol. 48, no 1-4, p. 2.

14

GINSBURG, Boris, GITMAN, Igor, et YOU, Yang. Large batch training of

convolutional networks with layer-wise adaptive rate scaling. 2018.

15

DU, Miao, WANG, Kun, XIA, Zhuoqun, et al. Differential privacy preserving of training model in wireless big data with edge computing. IEEE transactions on big data, 2018, vol. 6, no 2, p. 283-295.

16

TIAN, Yingjie, ZHAO, Xiaoxi, et HUANG, Wei. Meta-learning approaches for learning-to-learn in deep learning: A survey. Neurocomputing, 2022, vol. 494, p. 203-223.17

17

https://medium.com/@nageshmashette32/autogen-ai-agents-framework-3ee68bab6355

18

https://medium.com/@niall.mcnulty/an-intro-to-ai-agents-fbf50e6c80e5

19

RAJBHANDARI, Samyam, LI, Conglong, YAO, Zhewei, et al. Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In : International conference on machine learning. PMLR, 2022. p. 18332-18346.

20

XU, Yuhui, XIE, Lingxi, GU, Xiaotao, et al. Qa-lora: Quantization-aware low-rank adaptation of large language models. arXiv preprint arXiv:2309.14717, 2023.

21

PAPAGEORGIOU, Elpiniki I., STYLIOS, Chrysostomos, et GROUMPOS, Peter P. Unsupervised learning techniques for fine-tuning fuzzy cognitive map causal links. International Journal of Human-Computer Studies, 2006, vol. 64, no 8, p. 727-743.

22

AMEEN, Nisreen, TARHINI, Ali, REPPEL, Alexander, et al. Customer experiences in the age of artificial intelligence. Computers in human behavior, 2021, vol. 114, p. 106548.

23

https://indatalabs.com/blog/ai-cost-reduction#:~:text=The%20sectors%20that%20seem%20to,reductions%20of%2010%20to%2019%25.

24

https://www.eit.edu.au/artificial-intelligence-decoding-human-fears/

25

https://hyperight.com/how-to-prepare-for-fast-changing-ds-ml-ai-landscape-trends-challenges-and-opportunities/

26

https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

27

VASWANI, Ashish, SHAZEER, Noam, PARMAR, Niki, et al. Attention is all you need. Advances in neural information processing systems, 2017, vol. 30.