# Adding Value with Foundation Models: a Balancing Act

Sietske Tacoma[1[0000-0002-9662-8489]], Matthieu Laneuville[2[0000-0001-6022-0046]], Marieke Peeters[1[0000-0002-9984-4910]] and Stefan Leijnen[1[0000-0002-4411-649X]]

[1] Utrecht University of Applied Sciences, Utrecht, The Netherlands
[2] SURF, Utrecht, The Netherlands

sietske.tacoma@hu.nl

**Abstract.** The rapid advancement of Generative AI (GenAI) has made AI a central focus for many organizations. They are keen to explore how GenAI can optimize their operations. Yet, they recognize that introducing GenAI influences their employees' jobs and customers' experience. Hence, implementing GenAI in organizations introduces a variety of new technical, practical, and ethical challenges, requiring careful decision-making. This study examines how three media organizations approach these challenges when developing and implementing applications based on Foundation Models: large generic GenAI-models that serve as a foundation for task-tailored models and applications. Through qualitative analysis of data collected in interviews and focus group sessions, organizations' considerations in decision-making concerning Foundation-Model based applications were identified. Five central themes were determined in this decision-making: use case selection, handling data, choices in models and infrastructure, evaluation, and integration and future-readiness in the organization. Adopting GenAI within organization-specific applications proved to be a multidimensional balancing act. Three dimensions stood out as especially challenging: 1) weighing short-term return on investment against future requirements for flexible and inclusive applications; 2) weighing ease of use and maintenance that large technology providers offer against transparency and control offered by open-source models; and 3) considering impact on human work, by addressing not only potential time savings but also possible job obsolescence and the creation of new tasks. The study underscores the importance of developing clear frameworks and best practices that help assess the diverse implications of adopting GenAI-based applications in real-world organizational settings.

**Keywords:** Foundation Models, Applied AI, AI in media.

## 1 Introduction

The recent and rapid rise of Generative AI (GenAI) has sparked much interest in AI within organizations. Media organizations, in particular, recognize the potential of these models and tools for editing and creating text, images, video and audio [1-3]. As a result, they are eager to explore how GenAI can enhance their workflows and operational processes.

There are, broadly speaking, three ways in which organizations can integrate GenAI into their operations. First, individual employees can interact directly with GenAI models through prompting interfaces provided by model developers (e.g. ChatGPT). Second, organizations can use GenAI models as building blocks for custom applications through API integrations. Third, organizations can fine-tune a GenAI model for an organization-specific task and then build custom applications based on this fine-tuned model. Hence, either generic or fine-tuned for specific use cases, GenAI models serve well as building blocks for custom applications. When used as such, they are therefore also referred to as Foundation Models (FMs) [4].

In exploring GenAI's potential organizations do not only see opportunities but also face concerns. Translating GenAI capabilities into organizational workflows involves technical questions about which FM to use, how to handle data, whether or not to fine-tune an FM, and how to evaluate task-specific performance. It also involves ethical, legal and societal issues, such as the risks of hallucinations that FMs may produce, data safety and privacy, and compliance with regulations [5].

As the field of AI, and GenAI in particular, is rapidly evolving, ways to deal with these challenges have not been crystallized yet. Now that the field is maturing, and organizations are exploring GenAI's potential, they need tools to weigh the added value and the associated costs of integrating these models into their operations. To develop such tools, more insight is needed into the specific challenges organizations face when building on FMs. Since this concerns challenges at an organizational level, we focus on FM implementations in which choices are made at an organizational level. This means that our focus is on organizations using FMs as building blocks for custom applications, rather than on employees who directly interact with prompting interfaces provided by FM developers. To develop such insight, the research question addressed in this paper is: what are media organizations' considerations in the development and implementation of FM-based applications?

## 2 MLOps and AI Governance as Guidance

In the past few decades, tools to guide the development of AI-based applications have emerged. Many of these are connected to the umbrella term Machine Learning Operations (MLOps), which is a paradigm for conceptualization, implementation, monitoring, deployment and scalability of machine learning (ML) products [6]. MLOps focuses on supporting the automation, integration and monitoring throughout the development and implementation of ML products. It provides structure by defining a workflow for developing ML products, including the stages of business problem understanding, data acquisition, ML methodology, ML training & testing, continuous integration, continuous delivery, continuous training, continuous monitoring, explainable AI and sustainability [7]. It does not only define these stages, but also provides guidance by defining principles and best practices concerning reproducibility, collaboration, metadata tracking and feedback loops.

Since FMs are advanced ML models, many MLOps principles and stages apply to the development of FM-based applications. The training methodologies, however, are

different than for traditional ML products. Training FMs from scratch is expensive – the costs of training a single model easily run into tens of millions of euros [8]. Hence, FM training is generally not done by (media) organizations themselves, but by large technology firms such as Google, Meta and Microsoft and by AI companies such as DeepSeek, Mistral and OpenAI. Organizations can use these companies' FMs directly, or use fine-tuned models that other parties have developed based on these FMs. In both cases, they typically experiment with fine-tuning and prompting FMs using their organizational data and aiming to carry out organization-specific tasks. This requires new ways of working and incurs new dependencies on external parties: organizations use these parties' FMs, but do not have full insight into or control over these FMs. Moreover, they share organizational data with these parties, not knowing for what other purposes these data may be used. These new ways of working and new dependencies are likely to impact the applicability of MLOps principles.

While MLOps provides guidance for technical development of FM-based applications, it does not fully accommodate for the ethical, legal and societal concerns that organizations face when implementing such applications. These themes are better covered by the paradigm of AI Governance: a system of rules, practices, processes and technological tools that aid organizations on aligning their use of AI technologies with their strategies and values, ethical principles, and to fulfill legal requirements [9]. Within organizations, AI Governance focuses on responsible technology use (data, algorithms and AI systems), stakeholders & context (internal stakeholders, external stakeholders, and the sociopolitical context), regulation (both hard and soft) and processes of oversight, auditing and impact assessments [10]. Like MLOps, AI Governance provides principles and best practices on how to address these themes within organizations. Similar to MLOps, the question is to what extent the AI Governance principles and best practices apply to the relatively new practice of building FM-based applications. This new context may ask for different operationalizations and concretization of AI Governance principles than more traditional AI-based developments.

An important difference between AI Governance and MLOps concerns the stakeholders involved. While MLOps involves mainly technical roles such as solution architects, data engineers, and software engineers, AI Governance asks for involvement of a much larger variety of stakeholders, including management, decision-makers, users, and regulators. Researchers on AI Governance therefore call for integration of AI Governance principles into the design, production and implementation stages of AI development and for clearer definition of stakeholder roles in this integration [10]. By identifying organizations' needs and considerations in developing FM-based applications, this paper aims to provide a starting point for connecting MLOps and AI Governance principles and operationalizing both for this new form of AI development.

## 3 Methods

Exploring organizations' considerations when building FM-based applications asks for qualitative research. The setting for this study consisted of the development of FM-based applications within three Dutch media organizations that were part of a larger

research consortium on the topic of FM use. During these development processes, data was collected through interviews and focus group sessions. Below, we describe participants, instruments and the data collection and analysis methods we employed.

### 3.1 Participants

The three media organizations differed in scope and organization type. One is a large commercial press agency (approximately 250 employees). The second is a national public broadcasting organization (approximately 800 employees) focusing mainly on news, sports and events. The third is a smaller commercial business (approximately 40 employees), providing a subtitling platform for video content.

The organizations' use cases for which they developed FM-based applications varied as well. The press agency keeps a large database of experts who can interpret news events. The application they developed uses an FM to match experts with news articles they deliver to their clients. This enables them to suggest appropriate experts who can help journalists interpret and reflect on news events. The public broadcasting company records parliamentary debates and press conferences and archives these recordings. The application they developed uses an FM to automatically transcribe these recordings. These transcriptions greatly enhance the archive's searchability. The company providing a subtitling platform also used an FM to automatically transcribe video content. In their case, it concerned Flemish television programs that needed Belgian Dutch subtitling to increase accessibility for a more diverse audience. To accomplish this, they fine-tuned a Dutch speech-to-text model to also perform well on Belgian Dutch audio.

In each organization, two interviews were conducted during the development of their FM-based application. In each organization, the first and second interview were conducted with the same participants. To obtain information from multiple perspectives, participants were selected such that both technical as well as managerial knowledge and experience was represented in the interviews. Where possible, interviews were conducted with one interviewee possessing technical as well as managerial expertise and roles. Otherwise, interviews were conducted with two participants. Table 1 summarizes the number and roles of interviewees per organization. All in all, four participants were interviewed. All four participants were invited to the focus group sessions as well.

**Table 1.** Number and role of interviewees for each media organization.

| Organization | # Interviewees | Role and background interviewees |
|---|---|---|
| Press agency | 1 | Technical as well as managerial |
| Public broadcasting organization | 2 | One technical, one managerial |
| Commercial business providing subtitling platform | 1 | Technical as well as managerial |

### 3.2 Instruments

The interviews were semi-structured, guided by a question matrix [11]. The question matrix was aimed at identifying organizations' considerations when developing and implementing FM-based applications. This asked for breaking down the development and implementation process into pieces and identifying choices made in each of these pieces. Model cards, an approach to document choices made in the design of AI models [12], provided a useful starting point. Adjustments were made to the model card structure, to (1) accommodate for the development and implementation of AI-based applications rather than AI models, and (2) identify considerations underlying the choices made during development and implementation.

The model card structure covers categories including intended model use, model properties, training procedures, model performance, and contextual factors [12]. To address the full development and implementation of FM-based applications, additional themes were drawn from MLOps literature [6]. These included intended use and context of the application, integration of the application into users' daily work, performance evaluation and monitoring, deployment, maintenance and future development. For validation purposes, the list of categories and intended items was presented to two experts in the field of Human-Centered AI, who confirmed the relevance of the items and suggested to also address decision-making in the implementation.

The final question matrix contained 48 rows, grouped into five overarching categories: 1) Intended use, 2) Model properties, 3) Training, model performance and application performance, 4) Scope of the application (contextual factors) and 5) Implementation, maintenance and development. Each category consisted of three sub-categories, containing two to five rows each. The full list of categories, sub-categories and rows is presented in [11].

Asking organizations to explicate their choices for each piece of the development and implementation process has the potential to foster more deliberate choices in the designed applications. During the interviews, interviewees reflected on these choices from technical as well as managerial perspectives. By involving the management perspective as well, these reflections were not only informed by MLOps principles, but also by themes related to AI Governance. The columns of the question matrix provided an additional aid to address AI Governance themes and to identify organizations' considerations. For each item, the question matrix asked for the choices that the organization made (first column), considerations or arguments that informed these choices (second column), and alternatives that the organization had considered (third column).

The interview protocol consisted of addressing the items of the question matrix one by one and asking for choices, considerations underlying these choices and alternatives that had been considered. The order of the items was not fixed; when interviewees addressed a related item in their answers, that item was addressed next. To evaluate whether the question matrix fully covered the development of FM-based applications, the interview protocol ended with an open question asking for considerations that had not been mentioned in one of the items.

### 3.3 Data Collection and Analysis

In each organization, the first interview was conducted close to the start of the development and the second approximately three months later. All interviews lasted approximately one hour and were conducted by the first author. The first interviews with the press agency and the provider of the subtitling platform were conducted face-to-face, the other four through video calls. All interviews were recorded and transcribed using AmberScript. Participants gave permission for recording each interview.

During the first interviews within each organization, most attention was paid to the first three categories of the question matrix: 1) Intended use, 2) Model properties and 3) Training, model performance and application performance. The recordings and transcriptions were used to fill in the question matrix as completely as possible. The completed question matrix from the first interview served as basis for the second interview. The second interviews focused on developments since the first interviews and on questions from the question matrix that had not been discussed in the first interview. Consequently, in the second interviews, more attention was paid to the last two categories of the question matrix: 4) Contextual factors (the scope of the applications) and 5) Implementation, maintenance and development.

Three focus group sessions were organized, one before, one during and one after the development processes within the organizations. They were led by two researchers, one of which represented an IT cooperative of Dutch education and research institutions. The main aim of these sessions was to facilitate discussion and collaboration among the participants. The focus group sessions were not recorded. Notes of these sessions served as validation for the results obtained from the interviews.

Data analysis was carried out mainly by the first author and validated by the second author. The recordings and transcriptions from the interviews were used to fill out the question matrix as complete as possible for each organization, including design and implementation decisions as well as considerations that led to these decisions. This resulted in three filled out question matrices, one for each organization. As the answers that the organizations gave were highly use-case-specific, in the next step these answers (both decisions and considerations) were described in more generic terms. Next, open coding was carried out on the three question matrices to obtain a first categorization of considerations. For each consideration category, a code description was drafted, containing an organization's need or desire (e.g., "use developed knowledge and infrastructure as a base for future developments"), a condition to satisfy the desire (e.g., "the need to store data and any metadata in a future-proof way") and examples of implementation (e.g., "storing handmade adjustments to the outputs of an FM in a database").

After this first round of open coding, the three question matrices were merged into one question matrix, by merging all organizations' answers to each question. Next, a second round of open coding was conducted, this time not by organization but by question in the question matrix. This led to a final set of 14 consideration categories.

In the next step, the first and second author discussed the final set of consideration categories and plotted relationships between them. Some considerations were conditional for other considerations, whereas other considerations conflicted with each other. By identifying these relationships, we determined five overarching themes in the

development of FM-based applications that organizations need to make decisions about. These themes, the considerations within each theme and the examples used to illustrate these considerations, were presented to the participants, who confirmed to recognize their considerations and examples. The results section is structured along the five overarching themes and outlines the options that organizations have within these themes as well as the considerations that organizations used to decide between options.

## 4 Results

The five themes that categorize the organizations' considerations are 1) Selecting a use case and data, 2) Handling data, 3) Choices in models and infrastructure, 4) Evaluation, and 5) Integration and future-readiness within the organization. While themes 1 and 5 largely overlap the first and fifth question matrix category, the other three themes are structured differently than the question matrix. This illustrates that the question matrix provided a useful structure, but also offered enough freedom to address FM-specific considerations not fitting into a single category. Some considerations concerned clear preferences of one option over the other (e.g. "for this use case we prefer models with multilingual capabilities to models without"). In many more cases, deciding between options involved balancing competing factors. Table 2 summarizes these competing factors, grouped by theme. The next sections discuss these competing factors in more detail, illustrating their meaning in the context of FM-based application development.

### 4.1 Selecting a Use Case and Data

A central desire that organizations expressed when starting to work with FMs is that they add value in their organization. This was an issue of matching supply and demand. The supply that FMs offered was the capability to perform tasks concerning text, image, and audio generation. The demand consisted of the operational processes that organizations wanted to carry out to obtain their goals. Added value could be obtained for processes on which FMs were expected to perform well, that is, for processes that involved text, image and audio generation. According to the interviewees, nominee processes should also involve work that was labor-intensive and not satisfying for humans. The processes should be prominent enough that investing in enhancing them was expected to add value for the organization, but not so prominent that they risked harming the organization through errors that could result from the application. Organizations expected that outsourcing processes to FM-based applications could result in new human tasks, namely checking and verifying the produced results. These new tasks should be significantly less labor-intensive and, simultaneously, not less satisfying for humans than the human tasks that the application substituted. For example, the press agency did not want to replace the journalists' time-consuming though satisfying work of writing news articles by the less satisfying and possibly still time-consuming task of checking news articles for inconsistencies and factual correctness. Hence, even though writing news articles was a task that FMs were expected to perform quite well, the press agency did not deem this task suitable for replacement by FMs.

**Table 2.** Competing factors for topics in the design and implementation of FM-based applications. Factors that were clearly decisive in the organization's considerations are italicized.

| Topic | Factor 1 | Factor 2 |
|---|---|---|
| **Theme: Selecting a use case and data** | | |
| Process prominence | Expected added value | Associated risks |
| Human tasks | Tasks substituted by FM | New tasks resulting from FM use |
| Scope | Diversity and completeness | *Quick successes* |
| **Theme: Handling data** | | |
| Data preprocessing | *Reusability* | Time investment |
| Performance goals | (Perfect) quality | *Time investment* |
| Local or cloud storage | Data safety | Ease of use |
| **Theme: Choices in models and infrastructure** | | |
| Model provider | Open-source: control, data safety, transparency | *Proprietary: state-of-the-art, support, ease of use* |
| Infrastructure provider | Open-source: integration with own systems | Proprietary: integration with other components |
| Model size | Performance | *Low costs and fast training and inference* |
| Development | Flexible pipelines for experimentation | Time and money investment |
| **Theme: Evaluation** | | |
| Scope | *Diversity context use case* | (Technical) benchmarks |
| Methods | Human informed | Automated |
| Performance goals | *Impact on people's activities* | Model performance |
| **Theme: Integration and being future-proof in the organization** | | |
| Future developments | *Closely follow developments* | Costs of retraining |
| Expanding use case | *Performance* | Effort in new preprocessing |

Another important driver in use case selection was the desire to get off to a good start. Organizations felt a need to quickly achieve small successes, for obtaining organizational support and justifying larger investments. They also desired to create an environment for experimentation, by selecting use cases that involved less prominent, yet possibly value-generating processes. To get off to a good start without larger investments first, they chose to start with clean, well-defined data, such as single male voices in video recordings that needed to be transcribed. They acknowledged that these scoping choices did not do justice to the full diversity within their use cases.

A final aspect that played a large role in use case selection was the wish to make better use of the data that was available within the organization. Two organizations selected a process that involved databases that they collected and maintained as part of their core business. The third organization was handed the data by their clients.

Availability of, acquaintance with, and expected value of the data hence played a decisive role in use case selection.

### 4.2 Handling Data

When working with FMs, interviewees considered data quality a central issue. FMs, like traditional AI models, perform better for higher quality data. Consequently, organizations were willing to invest in preprocessing the data, but in a balanced way. The effort in preprocessing needed to weigh up against the work that the application took over.

Preprocessing and cleaning data for example concerned using speech detection software to only pass fragments containing speech to a speech-to-text FM. Fragments without (understandable) speech did not only result in bad-quality transcripts, but also required more time and computing power, because they were harder to transcribe than fragments with clear speech. Data cleaning for example concerned removing pieces of information that led to incorrect associations by the FM, such as associations based on experts' surnames instead of their expertise.

Generally, organizations did not find preprocessing and data-cleaning for using FMs complex, but it did tend to become time-consuming. They did, therefore, strive for higher data-quality, but not for perfection. Instead, they strived for data for which the FM could perform well enough in the selected use case. They also looked for ways to become future-proof, for example by creating databases of clipped fragments with metadata, or of handmade adjustments to model outputs. This enabled easy reuse of data in new training and development processes.

As mentioned before, organizations used valuable data, which sometimes contained privacy-sensitive information. Hence, to avoid data breaches and other damage, data safety was a crucial issue. The question of whether these data could be stored locally or had to be transferred to servers outside the organization was an important consideration, both in storage decisions as well as in choices concerning the FMs to use. This involved weighing data safety that open-source solutions offer against the ease of use that large tech companies' solutions provide.

### 4.3 Choices in Models and Infrastructure

Besides data safety, many other factors were at play in choosing FMs and deciding how to build applications upon them. Organizations balanced between quick successes and possibilities for longer-term extensions. They weighed the consequences of errors potentially made by the models and applications, and considered potential dependence on large tech providers and AI companies that offer FMs. And, naturally, money played a role as well.

Finding candidate FMs that suited the intended task was straightforward for the organizations. This involved identifying the required input and output modality and using platforms like HuggingFace to find FMs that served these modalities. The selection of candidate FMs was often informed by leaderboards, available at HuggingFace or similar platforms, that compare model performance against benchmarks. Besides these

general performance metrics, other model properties also weighed heavily in FM selection. For example, metadata that FMs delivered with their output, such as exact timings of audio transcripts, was regarded essential in some use cases. Not only the current use case, but also future extensions of the use cases were taken into account when judging the usefulness of model properties. For example, even when their current use case only involved a single language, organizations preferred FMs with multilingual capabilities.

Organizations often also needed to decide on model size. Generally, they preferred smaller FMs, mainly because of the associated costs. Before investing in larger FMs, organizations desired to build confidence by successful experiences with smaller, less expensive FMs. Although smaller FMs typically performed worse than larger FMs, this was compensated by faster training, embedding and inference. Depending on the use case, especially faster inference could weigh up against possible quality loss. An additional benefit of smaller FMs was that they calculated and stored data (e.g. embeddings) in fewer dimensions, which not only reduced the required computational power, but also the required storage capacity.

A theme that was regarded as complex was trust in the FMs' providers. Organizations used models developed by large tech and AI companies, such as Microsoft, Google and OpenAI. These providers offered well-performing, state-of-the-art FMs with solid documentation, support and maintenance. Organizations expected that these FMs, or improved versions, would remain supported for quite some time. Yet, organizations were also concerned about the limited control they have over these proprietary models, given the opacity of data collection and training procedures. Organizations were worried that model updates might lead to unexpected and unpredictable application behavior. They were also concerned about the safety of their organizational data. While analyzing the companies' licenses and contracts offered some insight, organizations found it difficult to oversee the consequences of these agreements.

Open-source FMs offered more transparency and control possibilities. These FMs allowed better for tweaking towards specific organizational needs and did not require sharing data with external partners. This did, however, require technical expertise, and hence did not help organizations to quickly get started. Organizations did express the ambition to use open-source FMs in the future, but at the same time obtained such promising results with proprietary FMs that investing the extra effort needed for open-source FMs was not a priority.

Besides choosing which FM to work with, organizations also had to make decisions on how to implement these FMs. These concerned storing capacity, facilities for data processing and sharing, cloud computing and training and test facilities. Again, dependence on large tech companies was an issue for organizations. Using multiple components provided by the same company enabled easy integration between components, but also increased the organization's dependence on this company and its choices. Open-source components, meanwhile, allowed for easier integration with the organization's own architecture components, but required more technical expertise, hosting and maintenance. Whether using open-source or proprietary components, organizations valued flexibility in infrastructure. They invested in developing flexible pipelines that allowed for easy model exchange and comparison. While this required larger time and

monetary investments, this helped them experiment and keep up with rapid development of new models.

A final consideration regarding models concerned implementing guardrails. Organizations used filters and hard-coded rules to avoid severe errors by the application that could be caused by a FM's hallucinations.

### 4.4 Evaluation

An application's added value for the organization did not only depend on the performance of the FM itself, but also on how the organization exactly deployed the model. Therefore, organizations desired to evaluate FM-based applications in the context of their use case. Off-the-shelf evaluation methods, such as the benchmarks that are generally used to evaluate AI-model performance, were not regarded sufficient to evaluate use-case-specific performance. Consequently, organizations developed custom evaluation methods. Importantly, organizations noticed a need to evaluate the application on the full range of categories or varieties for which the application was supposed to be used (e.g., various news categories, various voice types). Identifying categories for which the application did not perform well yet, did not only help organizations to know where to improve their models and applications, but also gave them insight into the completeness and quality of their databases.

In their custom evaluation methods, organizations typically involved domain experts and (intended) users of the application in crucial roles. Where possible, they tried to automate their evaluations methods, for example by keeping track of useful evaluation data. For example, when humans checked and corrected AI-generated transcripts, these corrections were stored in a database as evaluation data and training material for a next model version. When a model's outputted suggestions were presented to end users, the end users were asked to rate these suggestions for their usefulness (e.g., by indicating thumbs up or thumbs down, or on a 1-5 Likert scale).

Lastly, an important but difficult question for organizations was what application performance would be good enough. Interviewees explained that they were, generally, not looking for the best possible result, but for an improvement of the current situation by using FMs. They envisioned that in defining good performance, an important aspect would be how the application changed the users' work activities. For example, slightly reducing the number of errors an application made would not make much of a difference when, in the given use case, the end user still would have to check every result manually.

### 4.5 Integration and Being Future-proof in the Organization

To achieve added value, organizations realized that they needed to take measures to ensure that the FM-based application would indeed be used within the organization. They desired that the application would keep functioning and would not cause damage to the organization in the long term.

To make sure the application would be used, organizations often chose to implement it as addition to an existing system or product. In this implementation, they envisioned how the application could fit into the end user's working activities.

To be future-proof, organizations took into account that the environment in which the application was supposed to be used could change. In news contexts, for example, the frequent appearance or spread of new words (e.g., covid, ChatGPT) posed a challenge. Organizations accounted for these possible future developments by setting up procedures for regular model retraining, database re-embedding, and regular evaluation to monitor their application's performance in this changing environment. When expanding to new variants of the use case the application was intended for (e.g. audio with multiple voices, different languages or dialects), organizations considered whether new preprocessing steps and evaluation methods were needed to guarantee that application performance for these new perspectives met their end users' expectations. Lastly, organizations desired to be transparent about their AI use and, for example, added disclaimers to their application to explain how it used AI.

## 5 Discussion

The results illustrate that in developing FM-based applications organizations carry out a multidimensional balancing act, involving both technical and strategic considerations. The dimensions of this balancing act touch on several of the previously identified five themes. A first dimension, touching on all five themes, is the balance between obtaining quick successes and developing flexible, future-proof and inclusive applications that serve diverse use cases. This balance often tips towards obtaining quick successes: organizations tend to choose simpler, smaller models and use cases with narrow scopes. This aligns with early-stage MLOps practices, where organizations prioritize agility and experimentation over long-term operational stability. By enabling fast deployment cycles and iterative feedback and evaluation, organizations create environments for experimentation and learning about the potential value of FMs while reducing risk. In evaluating developed applications, this focus on quick success translates into a strive for improvement of the current situation rather than for perfection. Moreover, even without improvement to the current situation, organizations value this experimental setting because it helps them initiate discussions with multiple stakeholders about ethical, social and legal consequences of FM use for their organization. As such, this focus on experimentation supports growth in AI Governance as well.

Although obtaining quick successes is valued highly, organizations keep future-readiness in mind too. From an MLOps perspective this is, for example, apparent from how they store data and annotations in reusable ways and set up flexible pipelines allowing for easy model exchange. From an AI Governance perspective, when choosing the quick and less complex options, organizations do acknowledge that for future expansion, and specifically for attending to user diversity, they may need to consider more complex options.

A second dimension of the balancing act is perceived as more complex by organizations: the choice of providers, and especially the choice between open-source and

proprietary models and infrastructure components. This dimension closely relates to the theme C*hoices in models and infrastructure*, but also touches on *Handling data* and *Integration and being future-proof in the organization*. From an MLOps perspective, proprietary models tend to be favored over open-source models, for providing better support and maintenance, offering better opportunities for integration between various infrastructure components, and being generally less expensive. Concerning AI Governance themes, however, proprietary models are criticized for offering their users less control over outcomes and inducing more risks of hallucinations, which may cause damage to the organization. Data safety is also regarded a potential downside of proprietary models – although regulations are in place, organizations do not always trust that these regulations will keep their data safe in the near and distant future. Lastly, although implementing open-source models requires more technical expertise, it also generally allows for easier integration with the organization's own systems. This dimension illustrates a tension between MLOps and AI Governance perspectives, or a need to choose between technological performance and ethical values.

A final dimension of the balancing act concerns the human tasks that an FM-based application substitutes and the new human tasks its implementation creates. This dimension touches on the themes *Selecting a use case and data* and *Evaluation*, and is strongly related to AI Governance. In selecting use cases, organizations look for processes involving work that employees either do not perceive as rewarding, or that they do not have time for. At the same time, organizations notice that new tasks need to be carried out to obtain satisfactory performance of the FM-based application, such as data annotation, data preprocessing, and evaluating and checking outputs. Moreover, they observe that a significant gain in time or job satisfaction may weigh up against a slight quality loss. However, quantifying and evaluating the new work and the substituted work is challenging. Since this concerns people's daily activities in their jobs, this dimension emphasizes that implementing FM-based applications is not just a technical challenge, but a socio-technical one.

It is important to note that the participating organizations were starting with their development of FM-based applications and did not finish this development before the final interviews. Consequently, for themes appearing later in an application's lifecycle, such as evaluation, widening the scope to more users, and maintenance, they expressed intentions rather than decisions that were set in stone. Notably, compliance with regulations such as the AI act was barely mentioned in the interviews. We expect that in later stages, when organizations have their MLOps structures in place and have gained more knowledge about and experience with FMs, the balance may shift to a stronger focus on AI Governance themes. These include better alignment with desires such as reducing risks of damage, serving the full diversity of use cases, compliance with regulations, and considering environmental costs such as CO2 emissions [13]. Further research is needed to address the full development cycle, and to evaluate how organizations' intentions early in the development process are translated into decisions later on.

Another limitation to mention is the limited scope of this study, which focused on three organizations within a single sector, and with only one or two interviewees per organization. We expect that the organizations' considerations are articulated appropriately by the interviewees and that they are not, or to a very limited extent, sector-

specific. Further research in other sectors is needed to validate these results. Such future research should take into account that the FM and AI landscape shifts rapidly, implying that considerations that are complex today, may be straightforward in the near future.

## 6 Conclusion

This study highlights that successful implementation of FM-based applications asks for many considerations, and undeniably not just technical ones. In fact, technical decisions such as choosing appropriate model versions turned out to be the less complex ones. We found three main complex decisions: choosing between obtaining quick successes and serving wide, diverse use cases, choosing between proprietary models and open-source models, and weighing human tasks that FM-based applications take over against new human tasks that their implementations creates.

To facilitate decision-making in these complex choices, FM-specific operationalizations for MLOps and AI Governance principles need to be developed. In FM-specific MLOps, operationalizations of the principles should focus on those elements of FM use that organizations can control. This means that the core training procedure is less important than in traditional ML, but more tools are needed to facilitate experimentation with own data and building flexible pipelines to easily exchange and compare FMs. Concerning FM-specific AI Governance, there is a need for more specific elaboration and quantification of the ethical, legal and societal consequences of using FMs (either proprietary or open-source). This includes the impact on people's jobs and risks that errors or hallucinations produced by FMs may incur. Tools to quantify such consequences could help to evaluate the impact of ethical, legal and societal consequences together with technical performance in application development, and hence, to connect MLOps practices and AI Governance principles. Such tools should also focus on involving multiple stakeholders, including management, in early development stages.

The question matrix that was used as interview guide in this study can serve as starting point for the development of such tools. When a team containing people with multiple perspectives, including technical, strategical, ethical and managerial roles and expertise, uses this question matrix as guideline, the development of an FM-based application is informed by considerations covering these multiple perspectives. Following the discussion above, the question matrix could be expanded with more AI Governance themes, such as laws and compliance.

While tools for quantifying ethical, legal and societal consequences and approaches for involving multiple stakeholders early on may help streamline certain decisions, they do not untangle the tension between the technological superiority of proprietary models and the preference for open-source models from the governance perspective. Untangling this tension asks for the development of alternative FMs that do not only perform well on a technological level, but also offer the transparency and control possibilities that organizations need to use Generative AI confidentially.

**Acknowledgments.** This study was funded by Taskforce of Applied Research SIA (grant number KIEM.K23.01.177).

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References


[1] Singla, A., Sukharevsky, A., Yee, M., Chui, M., and Hall, B.: The state of AI in early 2024 Accessed: Sept. 12, 2024. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[2] Nah, F. F.-H., Zheng, R., Cai, J., Siau, K., and Chen, L.: Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. Journal of Information Technology Case and Application Research **25**(3), 277–304 (2023). doi: 10.1080/15228053.2023.2233814.

[3] Head, C. B., Jasper, P., McConnachie, M., Raftree, L., and Higdon, G.: Large language model applications for evaluation: Opportunities and ethical implications. New Directions for Evaluation **2023**(178–179), 33–46 (2023). doi: 10.1002/ev.20556.

[4] Bommasani, R. et al.: On the Opportunities and Risks of Foundation Models. ArXiv:2108.07258 (2021). doi: 10.48550/arxiv.2108.07258.

[5] Leijnen, S., Aldewereld, H., van Belkom, R., Bijvank, R., and Ossewaarde, R.: An agile framework for trustworthy AI. NeHuAI@ ECAI, 75–78 (2020).

[6] Kreuzberger, D., Kuhl, N., and Hirschl, S.: Machine Learning Operations (MLOps): Overview, Definition, and Architecture. IEEE Access **11**, 31866–31879 (2023). doi: 10.1109/ACCESS.2023.3262138.

[7] Testi, M. et al.: MLOps: A Taxonomy and a Methodology, IEEE Access **10**, 63606–63618 (2022). doi: 10.1109/ACCESS.2022.3181730.

[8] Dickson, B.: The GPT-3 economy. https://bdtechtalks.com/2020/09/21/gpt-3-economy-business-model/.

[9] Mäntymäki, M., Minkkinen, M., Birkstedt, T., and Viljanen, M.: Defining organizational AI governance. AI Ethics **2**(4), 603–609 (2022). doi: 10.1007/s43681-022-00143-x.

[10] Birkstedt, T., Minkkinen, M., Tandon, A., and Mäntymäki, M.: AI governance: themes, knowledge gaps and future agendas, Internet Research **33**(7), 133–167 (2023). doi: 10.1108/INTR-01-2022-0042.

[11] Tacoma, S., Mulder, J., Laneuville, M., and Leijnen, S.: A Thought-provoking Question Matrix to Guide the Development of Foundation-Model-based Applications. In: Ericson, P., Khairova, N., De Vos, M. (eds.): Proceedings of the Workshops at the Third International Conference on Hybrid Human-Artificial Intelligence 2024. CEUR-WS **3825**, 193–201 (2024).

[12] Mitchell, M. et al.: Model cards for model reporting. In: FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, 220–229 (2019). doi: 10.1145/3287560.3287596.

[13] Tu, X., He, Z., Huang, Y., Zhang, Z.-H., Yang, M., and Zhao, J.: An overview of large AI models and their applications. Visual Intelligence **2**(34) (2024). doi: 10.1007/s44267-024-00065-8.