Large Language Models versus Foundation Models for Assessing the Future-Readiness of Skills

Norman Süsstrunk, Albert Weichselbraun, Roger Waldvogel

University of Applied Sciences of the Grisons, Chur, Switzerland
{norman.suesstrunk, albert.weichselbraun, roger.waldvogel}@fhgr.ch

Abstract

Automatization, offshoring and the emerging "gig economy" further accelerate changes in the job market leading to significant shifts in required skills. As automation and technology continue to advance, new technical proficiencies such as data analysis, artificial intelligence, and machine learning become increasingly valuable. Recent research, for example, estimates that 60% of occupations contain a significant portion of automatable skills.
The "Future of Work" project uses scientific literature, experts and deep learning to estimate the automatability and offshorability of skills which are assumed to impact their future-readiness. This article investigates the performance of two deep learning methods towards propagating expert and literature assessments on automatability and offshorability to yet unseen skills: (i) a Large Language Model (ChatGPT) with few-shot learning and a heuristic that maps results to the target variables, and (ii) foundation models (BERT, DistilBERT) trained on a gold standard dataset. An evaluation on expert data provides initial insights into the systems' performance and outlines the strengths and weaknesses of both approaches.

1 Introduction

Forces such as automatization, offshoring and the emerging "gig economy" create and accelerate disturbances in labor markets. As automation and technology continue to advance, new technical skills such as data analysis, artifi-

cial intelligence, and machine learning become increasingly valuable. With offshoring intercultural communication and global collaboration skills gain in significance. Finally, the emergence of the gig economy increases the importance of adaptability, self-management, and entrepreneurship, as individuals navigate through various short-term projects and roles. Offshoring and gig economy platforms such as Fiverr, UpWork and Freelancer further disrupt labor markets by forcing employers and employees to adapt to new, more flexible working structures. However, as certain tasks become automated or outsourced, traditional routine-based skills experience a decline in demand, highlighting the need for individuals to continually upskill and reskill to remain competitive in the evolving job market.

The presented research has been conducted within the "Future of Work" project[1] which aims at assessing the future-readiness of skills by gathering insights on their *automatability* and *offshorability*. Given an extensive ontology of approximately 51,000 skills across diverse professional domains provided by our research partner, the project develops deep learning components to automatically calculate these metrics for each skill. Manual evaluation is not a feasible option due to the sheer volume of data, making it more sensible to leverage machine learning techniques for this purpose.

We draw upon scientific literature, domain experts and evidence mined from the web to create a gold standard dataset which classifies skills based on these two dimensions. Afterwards, two approaches for distributing gold standard assessments to related skills are deployed:

1. Large Language Models yield information on a skill's properties which are then combined with a heuristic that incorporates expert knowledge to assess their automatability and offshorability.
2. Foundation Models trained on the gold standard dataset that are then used for classifying yet unseen skills.

The presented method provides decision makers with insights into the future-readiness of skills and job profiles. It, therefore, addresses the United Nation Sustainable Development Goals on

- "Quality education" (SDG 4) which aims at increasing the number of people with relevant skills to succeed in their professional life [19]; and
- "Decent work and economic growth" (SDG 8) promoting sustained, inclusive and sustainable economic growth, full and productive employment, and decent work for all [19].

1 https://www.fhgr.ch/en/uas-grisons/angewandte-zukunftstechnologien/swiss-institute-for-information-science-sii/projekte/future-of-work/

The rest of this paper is structured as follows: Section 2 provides a compact overview of related work which is followed by a description of the applied methods (Section 3). Afterwards, Section 4 describes the evaluation setting and introduces our initial results. The paper concludes with an outlook and conclusion presented in Section 5.

2 State of the Art

The following discussion of the state of the art focuses on (i) literature discussing criteria for automatization, (ii) offshorability and (iii) relevant developments in the area of deep learning such as foundation models and language models.

2.1 Automatization

Autor and Dorn [1] consider jobs which are high in routine tasks as likely candidates for automatization. Work by Josten and Lordan [10] identifies the following additional indicators for occupations and skills which impact automatization:

(1) "people", i.e., whether the job requires interaction with people on a day-to-day basis,
(2) "brains", i.e., whether abstract thinking is required, and
(3) "brawn", i.e., whether a physical interaction with objects is required.

The authors draw upon prior work [9] that classified O*Net[2] occupations into "automatable", "non-automatable" and "partly automatable", to align their model with the European Labour Force Survey 2013–2016. A regression analysis outlined, that occupations requiring brain (i.e, abstract thinking) have the highest protection from displacement due to automatization, followed by occupations involving people skills ("people") and physical interaction ("brawn"). Combining these factors can provide an even higher protection from automatization.

Recent research by Eloundou et al. [8] investigates the impact of Large Language Models (LLM) such as ChatGPT and BLOOM [18] on the labor market. They note that particularly routine and repetitive tasks have a high

2 https://www.onetonline.org

risk of technology-driven displacement. Brynjolfsson et al. [4] distinguish between labor augmenting and labor displacing effects of automatization. Eloundou et al. [8], in contrast, classify an occupation's exposure towards replacements by LLMs into three classes:

- no exposure, since LLMs do provide only minimal or no reduction in the time required for completing work tasks,
- direct exposure, where the time required for completing a task is cut at least in half, and
- indirect exposure, if productivity could be doubled with additional software or tooling that is not available yet.

Based on this classification, particularly routine tasks in data processing, information processing and hospitals exhibit a high expose to displacement while tasks in the area of manufacturing, agriculture and mining seem to be relatively safe.

Nevertheless, as with any general purpose technology (e.g., printing press, steam engine, etc.) the impact of artificial intelligence will unfold over decades and is difficult to assess, since the realization of its full potential requires extensive and time-consuming co-invention and the discovery of new business models and processes [6, 12].

2.2 Offshorability

Research by Wagner et al. [22] on digital platforms for knowledge work such as Freelancer, Upwork and Fiveer has been instrumental in providing gold standard data on offshorable tasks. These platforms enable companies to

(1) meet ad-hoc demand for knowledge work services [14],
(2) access specialized skills without creating permanent internal positions [20], and
(3) fill staffing needs that cannot be addressed by traditional labor markets [11].

Work by Dunn et al. [7] distinguishes between gig economy platforms that provide (i) low-skill location dependent services (e.g., Uber, Lyft, TaskRabbit), (ii) high-skill location dependent services (e.g., Outschool for private lessons and Thumbtack for craftspeople), (iii) low-skill location independent services (e.g., Amazon Mechanical Turk and Cloudflower), and (iv) high-skill location independent services (e.g., Fiverr and Upwork).

The services listed on these platforms provide valuable insights into a task's offshorability since they provide catalogs of work activities that have already been successfully offshored.

2.3 Deep Learning for Text Classification

Text classification methods have benefited tremendously from recent advances in the area of machine learning, particularly the development of the transformer architecture.

Transformers capture long-range dependencies in sequential data effectively, making them ideal for tasks involving contextual understanding, such as sentiment analysis, document classification, and question answering [16]. Unlike traditional approaches that rely on recurrent or convolutional neural networks, transformers utilize a self-attention mechanism to model interdependencies between words. This self-attention mechanism allows transformers to weigh the importance of different words in a sentence, giving more weight to semantically relevant ones and discarding noise [21].

Language models such as BERT [5], DistilBERT [17] and RoBERTa [13] draw upon the transformer architecture and achieve a remarkable performance on various natural language processing (NLP) benchmarks, surpassing previous techniques by a significant margin. The attention-based architecture of transformers not only facilitates better representation learning but also enables effective transfer learning by fine-tuning pre-trained language models to specific tasks such as text classifications.

Bommasani et al. [2] call these pre-trained language models foundation models to underscore (i) their central role for NLP and (ii) incomplete nature which requires adaptation and fine-tuning to specific tasks such as text classification. Large Language Models (LLMs) exceed foundation models in size (i.e., over 10 billion parameters [24]), and apply training strategies such as instruction tuning and adaptation tuning to enable instructing following and zero-shot capabilities. The resulting models (e.g., GPT-3 [3], GPT-4 [15] and ChatGPT[3]) inhibit so-called emerging capabilities which further improve their capability to correctly interpret human language and, therefore, pave the way for even more advanced text classification systems [24].

3 https://chat.openai.com

3 Method

This section introduces the datasets used within this paper, the baseline classifier, the foundation model classifiers, and the LLM-based skill assessment method.

3.1 Datasets

The paragraphs below discuss datasets used for domain adaptation, and the gold standard deployed for fine-tuning and testing.

Domain adaptation dataset. This study deploys domain adaptation to enhance the performance and efficiency of the foundation models in the application domain. We utilized a dataset of 150,366 job postings sourced from Switzerland, encompassing diverse industries and job roles which have been converted to text with the Inscriptis HTML to text conversion library [23]. The data collection process was carried out by x28, a company with expertise in web crawling and data aggregation from job boards and corporate websites.

Gold standard dataset. The baseline data for this study comes from the x28 company ontology and represents a collection of skills necessary for various occupations. Each skill is characterized by a predicate, and a topic, where the predicate represents an action and the topic defines the context or object of that action. The skill "Control production machine", for example, is composed of the topic "production machine" and the predicate "control". In total, the gold standard dataset comprises 434 such skills.

Two experts from x28 drew-upon the following guidelines to manually assess these 434 skills for their offshorability and automatability, using binary ratings:

(1) *Offshorability* indicates whether a task can currently be performed entirely in the absence of the person performing it. The assessment is based on current corporate practices in Switzerland and only considers tasks fully offshorable, where all elements can be implemented regardless of physical presence. Barriers to outsourcing may include face-to-face interaction, task commitment to a specific location, working with large objects, and cultural preferences for personal presence.

(2) *Automatability* assesses whether a task can currently be performed entirely by technology. To do this, it considers a variety of automation technol-

ogies, including hardware such as robots and drones, and software such as predictive systems and generative algorithms. Activities that require human involvement are too complex, outside the scope of machines, or have unclear goals are generally considered non-automatable.

The experts considered 69.4% of the assessed skills as offshorable, and 55.1% as automatable. Table 1 lists some example expert assessments.

Table 1: Example expert assessments on offshorability and automatability

Predicate	Topic	Offshorability	Automatability
create	dossier	1	1
correct	jaw malposition	0	0
program	user interface	1	0
clean	object	0	1

Both experts autonomously reviewed and annotated the entire dataset. Following this phase, manual amalgamation and consensus were achieved to create the final gold standard dataset.

3.2 Baseline Classifier

The evaluation presented in Section 4 draws upon a strong baseline that uses Word Embeddings in conjunction with Support Vector Machines (SVM).

To convert the textual data into a format suitable for machine learning algorithms, we first tokenize sentences with the Natural Language Toolkit (NLTK), and then convert the tokens into embeddings using the pre-trained *fasttext-wiki-news-subwords-300* embedding model in conjunction with the Gensim library[4]. Combining the all sentence tokens with an averaging algorithm yields sentence-level embeddings which serve as input into the Support Vector Machine.

Both sentence embeddings and gold standard labels are converted into NumPy arrays which are then used as inputs for training the SVM implementation provided with scikit-learn[5]. We use a four-fold cross-validation strategy for training and evaluating the created classifier.

4 https://radimrehurek.com/gensim

5 https://scikit-learn.org

3.3 Foundation Model Classifiers

The presented research draws upon the Hugging Face library's implementation[6] of foundation models. Hugging Face provides a comprehensive open-source framework for natural language processing that contains a rich collection of pre-trained language models, tokenizers, and utilities. It offers a unified API for various transformer-based architectures and standard tasks such as text classification, question answering and text summarization. The library allows researchers to load pre-trained models, adapt them to a target domain and fine-tune the model on various tasks such as text classification.

Automated Hyperparameter Optimization. We draw upon the Optune framework[7] for hyperparameter optimization. The framework supports various search algorithms, such as TPE (Tree-structured Parzen Estimator) and CMA-ES (Covariance Matrix Adaptation Evolution Strategy). Optuna optimizes the hyperparameters of the text classification model by iteratively evaluating different combinations and locating the hyperparameter set that maximizes performance. This process significantly reduces the manual effort required for hyperparameter tuning and improves the overall efficiency of the classification pipeline.

Foundation Language Models. The experiments presented in this paper use the following pre-trained language models:

- BERT multilingual base model (cased)
- DistilBERT base multilingual (cased)
- XLM-RoBERTa

These models have been trained on large-scale multilingual corpora which facilitate the acquisition of cross-language and language-specific patterns and nuances. Moreover, they are capable of effectively handling German vocabulary, idiomatic expressions, and syntactic structures.

Domain Adaption. Adapting foundation models to a particular target domain offers numerous advantages. By exposing the model to domain-specific documents, prior to fine-tuning it for a specific task, the model gains several benefits.

Firstly, domain adaptation enables the model to closely align itself with the characteristics, linguistic nuances, and domain-specific features of the target text corpus. This alignment ensures that the model becomes more pro-

6 https://huggingface.co

7 https://optuna.readthedocs.io

ficient in understanding the specific vocabulary, phraseology, and contextual nuances required for accurate predictions in the target domain. Consequently, the model's pre-existing knowledge is effectively integrated with the domain-specific requirements, reducing the likelihood of encountering semantic mismatches or misinterpretations during fine-tuning. To evaluate the effectiveness of domain adaptation, the fine-tuned models underwent comparison both with and without domain adaptation. In this study, a corpus of 150,366 job description documents (Section 3.1) was utilized in the domain adaptation phase. Figure 1 outlines the adaptation and fine-tuning process with and without domain adaptation.

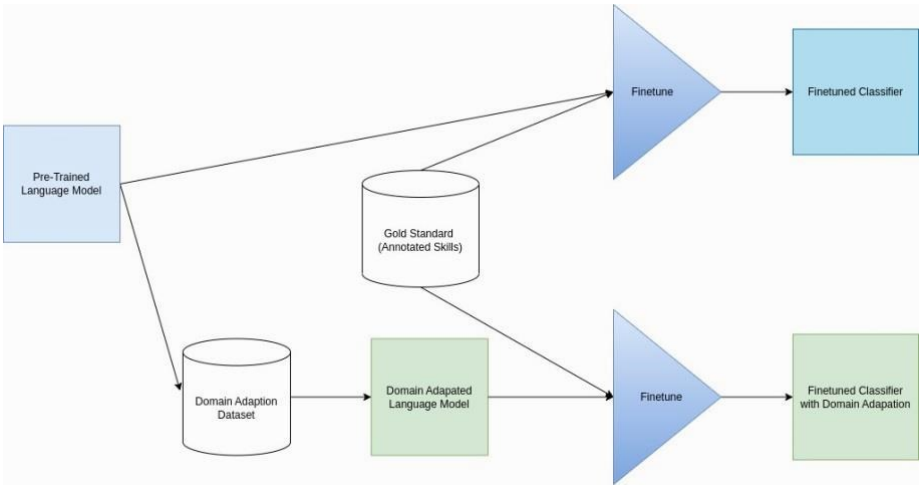

Fig. 1 Fine-tuning foundation models with and without domain adaption

Model fine-tuning. A common approach towards fine-tuning foundation models is freezing certain layers during the training process. Freezing layers refers to the practice of preventing their weights from being changed during the fine-tuning stage, while allowing the remaining layers to be updated based on the task-specific data. This approach serves two primary purposes. First, it helps to retain the knowledge obtained during creation of the foundation model. By freezing lower-level layers, which capture more general linguistic features, the model maintains the ability to extract meaningful representations from input text. Second, freezing layers reduces the computational burden and accelerates the training process, as updating the weights of all layers would require considerably more resources and time. Typically, the

initial layers, which learn lower-level features, are frozen, while the subsequent layers are fine-tuned on the task-specific data.

Table 2 provides a summary of the model parameters used in the presented experiments.

Table 2: Model specifications

Base Language Model	Bert Base Multilingual Cased	Bert Base Distilbert	XLM Roberta
Solver (learning rate)	AdamW	AdamW	AdamW
Activation	Gelu	Gelu	Gelu
Attention dropout	0.1	0.1	0.1
Dimension	768	3072	1024
Dropout	0.1	0.1	0.1
Hidden layer dimensions	12	(n.a.)	24
Initializer range	0.02	0.02	0.02
Max position embeddings	512	512	514

3.4 Large Language Model with Heuristic Classifier

Figure 2 provides an overview of the LLM-based approach for classifying the automatability and outsourcability of skills. The system queries a Large Language Mode (ChatGPT) for the following basic characteristics which draw upon and extend the indicators identified by Josten and Lordan [10]:

(1) *brawn*: Is there interaction with a physical object?
(2) *people*: Is any interaction with humans required?
(3) *brain*: Is abstract thinking required?
(4) *location*: Does the activity need to be performed on-site at the customer's location?
(5) *digitalization*: Can the activity be performed digitally?
(6) *routine*: Can the activity be broken down into standardized processes that can be performed equally anywhere in the world?

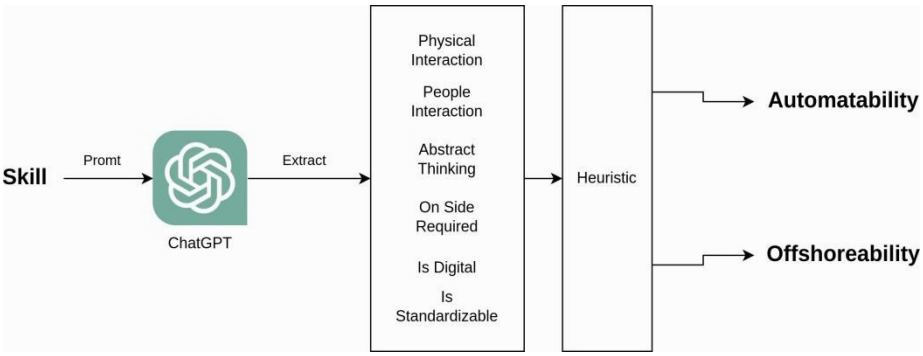

Fig. 2 Obtaining information on a skill's automatability and outsourcability from ChatGPT and heuristics

The experiments query ChatGPT's application programming interface (API) with the gpt-3.5-turbo model, using a prompt which provides view-shot examples for contextualizing the request as illustrated below:

> **Kann die Tätigkeit [java programmieren] zwingend nur vor Ort beim Kunden ausgeführt werden? Beispiele Ja: [Fenster putzen], [Wasserleitung reparieren], [Tontechnik einstellen], [Haare färben]. Beispiele Nein: [Ausflug organisieren], [Spiegel fertigen], [Webseite erstellen], [Team leiten]. Antworte nur mit Ja oder Nein.**

Each prompt only considered one of the six characteristics since combining them yielded worse results. Afterwards, we used the following heuristic to retrieve the final classifications from the ChatGPT response:

(1) If a skill needs to be performed at the customer's location, it was postulated that it was unsuitable for automation or outsourcing.
(2) For a skill to be categorized as automatable, it had to meet certain conditions: It could not depend on interactions with physical objects or require interactions with humans or abstract reasoning. However, this rule was not valid if the skill was digitally executable.
(3) Skills were considered outsourceable if they were either digitally or globally implementable as a standardized process.

4 Evaluation

The evaluation assesses the models' accuracy in predicting automatability and offshoreability on the gold standard dataset introduced in Section 3.1.

Stratified k-fold cross-validation with four folds was used to validate the SVM baseline and foundation models. The test data was divided into 80% training data and 20% evaluation data in each run. Once all runs were executed, the results were summarized. This procedure was not applied to the ChatGPT/heuristic classifier since it only used view-shot training with examples that have not been part of the evaluation dataset.

Table 3: Classification performance for the "offshorable" indicator

model	domain adaptation	layer freeze	f1	precision	recall	accuracy
SVM baseline			0.81	0.70	**0.97**	0.69
bert-base-multilingual-cased			0.83	0.80	0.87	0.76
bert-base-multilingual-cased		✓	0.83	0.76	0.91	0.73
bert-base-multilingual-cased	✓		0.83	0.80	0.87	0.76
bert-base-multilingual-cased	✓	✓	**0.84**	0.78	0.91	0.76
distilbert-base-multilingual-cased			**0.84**	**0.81**	0.88	0.77
distilbert-base-multilingual-cased		✓	0.80	0.76	0.86	0.70
distilbert-base-multilingual-cased	✓		0.82	0.77	0.88	0.84
distilbert-base-multilingual-cased	✓	✓	0.80	0.73	0.88	0.81
xlm-roberta-large			0.83	**0.81**	0.86	0.76
xlm-roberta-large		✓	0.81	0.78	0.84	0.72
xlm-roberta-large	✓		0.77	0.69	0.88	0.79
xlm-roberta-large	✓	✓	0.75	0.69	0.83	**0.82**
ChatGPT combined with Heuristic			0.80	**0.81**	0.78	0.71

The results presented in Table 3 and Table 4 show that the transformer classifier outperforms both ChatGPT and the baseline. This outcome can be attributed to several key factors. Firstly, transformers employ self-attention mechanisms, enabling them to capture intricate dependencies among input features. This attention mechanism allows transformers to effectively exploit

the limited training data, extracting meaningful relationships even from a small sample size. Secondly, transformers are pre-trained on large-scale corpora, leveraging vast amounts of unlabeled data to learn generic representations. Consequently, these pre-trained models can effectively transfer knowledge to downstream tasks, which proves particularly beneficial when training data is limited. Moreover, transformers have proven to be highly adaptable and adept at handling complex patterns and non-linear relationships, making them well-suited for challenging classification tasks. In contrast, SVMs heavily rely on kernel functions, which can struggle with high-dimensional and non-linear feature spaces, often leading to reduced performance in the presence of limited training samples.

Table 4: Classification performance for the "automatable" indicator

model	domain adaptation	layer freeze	f1	precision	recall	accuracy
SVM baseline			0.71	**0.94**	0.58	0.59
bert-base-multilingual-cased			0.71	0.73	0.69	0.69
bert-base-multilingual-cased		✓	0.73	0.72	**0.79**	0.68
bert-base-multilingual-cased	✓		**0.75**	0.76	0.74	0.73
bert-base-multilingual-cased	✓	✓	0.73	0.72	0.74	0.69
distilbert-base-multilingual-cased			0.69	0.70	0.69	0.66
distilbert-base-multilingual-cased		✓	0.69	0.68	0.72	0.65
distilbert-base-multilingual-cased	✓		0.67	0.58	0.78	0.75
distilbert-base-multilingual-cased	✓	✓	0.68	0.60	**0.79**	**0.76**
xlm-roberta-large			0.74	0.75	0.75	0.72
xlm-roberta-large		✓	0.69	0.66	0.74	0.65
xlm-roberta-large	✓		0.67	0.62	0.74	0.73
xlm-roberta-large	✓	✓	0.67	0.64	0.71	0.71
ChatGPT combined with Heuristic			0.62	0.68	0.57	0.54

The lower performance of ChatGPT can be attributed to multiple factors: Firstly, ChatGPT is a general language model trained on a diverse range of texts, which includes a vast array of topics and genres. As a result, it lacks the specific knowledge and biases that may be helpful for accurately classifying texts within a particular domain. Without further fine-tuning, it struggles to correctly interpret the questions in the context of the provided task. Secondly, ChatGPT's training objective is to generate coherent and contextually appropriate responses in conversational settings. It is not explicitly optimized for text classification tasks, particularly not in the human resource domain. Consequently, it may not possess the necessary specialized mechanisms to identify and extract the most relevant features for classification purposes.

While the transformer classifier for offshorable demonstrates robust performance, the classifier for the automatable label consistently underperforms. This discrepancy indicates the presence of underlying complexities within the automatable classification problem, making it considerably more challenging for the classifier to draw meaningful conclusions from the training data. This observation also holds for the ChatGPT-based classifier which yields significantly lower scores for the automatable category.

It becomes apparent that the automatable classification problem is inherently intricate, primarily due to the complexity associated with drawing meaningful conclusions from the available training data. Unlike the offshorable label, which may have relatively clearer patterns or explicit indicators, the automatable label encompasses a multitude of underlying factors that are often nuanced and context-dependent. These multidimensional aspects make it considerably more difficult to discern and extract relevant features.

5 Outlook and Conclusions

The presented work assesses the future-readiness of skills and working activities based on two criteria: (i) offshorability which determines whether a skill can be performed overseas, and (ii) automatability indicating the potential for a skill to be automated.

Annotators drew upon scientific literature and domain experts to create a gold standard dataset of skills that have been classified according to these metrics. Afterwards, we tested three different approaches towards automatically classifying yet unseen skills:

- a strong baseline that uses Support Vector Machines (SVM) in conjunction with word embeddings,
- foundation models such as BERT, DistilBERT and RoBERTa, and
- ChatGPT which assessed skills in regard to six proxy metrics. Experts then crafted a heuristic that maps these proxy metrics to the target indicators.

An evaluation based on the expert gold standard dataset revealed that the BERT model provided the most accurate predictions. The ChatGPT model, in contrast, was beat both by the SVM baseline and the investigated foundation models. This result is not surprising, given the fact that these models received a fraction of the gold standard dataset (i.e., the training partition) during fine-tuning. ChatGPT, in contrast, only benefited from few-shot learning based on four examples provided with the query prompt. Considering this systematic disadvantage, its results are still commendable. Nevertheless, they also clearly reflect the limitations of prompt engineering and few-shot learning in complex settings. Future work will explore further options for improving the classification process. We plan to develop an ensemble method that combines foundation models with Large Language Models, particularly for skills which are dissimilar with existing gold standard data. Another interesting strategy would be the incorporation of background knowledge from ontologies such as O*NET and ESCO[8] into the classification process.

Acknowledgements

The research presented in this paper has been conducted within the Future of Work project (https://www.fhgr.ch/en/uas-grisons/angewandte-zukunftstechnologien/swiss-institute-for-information-science-sii/projekte/future-of-work/) funded by Innosuisse.

References

[1] David H. Autor, & David Dorn. 2013. The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market. *American Economic Review 103*, (5), 1553–1597. https://doi.org/10.1257/aer.103.5.1553

8 European Skills, Competences, Qualifications and Occupations; https://ec.europa.eu/esco/lod/static/model.html

[2] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, … Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258 [cs] (Aug. 2021). http://arxiv.org/abs/2108.07258

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, … Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005. 14165 arXiv:2005.14165 [cs].

[4] Erik Brynjolfsson, Tom Mitchell, & Daniel Rock. 2018. What Can Machines Learn, and What Does It Mean for Occupations and the Economy? *AEA Papers and Proceedings 108* (May), 43–47. https://doi.org/10.1257/pandp.20181019

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, & Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Minneapolis, Minnesota (pp. 4171–4186). https://www.aclweb.org/anthology/N19-1423 arXiv: 1810.04805

[6] Jay Dixon, Bryan Hong, & Lynn Wu. 2021. The Robot Revolution: Managerial and Employment Consequences for Firms. *Management Science 67*(9), 5586–5605. https://doi.org/10.1287/mnsc.2020.3812

[7] Michael Dunn, Isabel Munoz, & Mohammad Hossein Jarrahi. 2023. Dynamics of flexible work and digital platforms: Task and spatial flexibility in the platform economy. *Digital Business 3*(1), 100052. https://doi.org/10.1016/j.digbus.2022.100052

[8] Tyna Eloundou, Sam Manning, Pamela Mishkin, & Daniel Rock. 2023. GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models. https://doi.org/10.48550/arXiv.2303.10130 arXiv:2303.10130 [cs, econ, q-fin].

[9] Cecily Josten, & Grace Lordan. 2020. Robots at Work: Automatable and Non-automatable Jobs. In Klaus F. Zimmermann (Ed.), *Handbook of Labor, Human Resources and Population Economics*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-57365-6_10-1

[10] Cecily Josten, & Grace Lordan. 2022. Automation and the changing nature of work. *PLOS ONE 17*(5), e0266326. https://doi.org/10. 1371/journal.pone.0266326

[11] Naufal Khan, & Johnson Sikes. 2014. IT under pressure | McKinsey. Technical Report. https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/it-under-pressure-mckinsey-global-survey-results

[12] Richard G. Lipsey, Kenneth Carlaw, & Clifford Bekar. 2005. *Economic transformations: general purpose technologies and long-term economic growth*. Oxford, New York: Oxford University Press. OCLC: ocm60931387.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (July 2019). http://arxiv.org/abs/1907.11692 arXiv: 1907.11692

[14] Dorit Nevo, & Julia Kotlarsky. 2020. Crowdsourcing as a strategic IS sourcing phenomenon: Critical review and insights for future research. *The Journal of Strategic Information Systems* 29(4), 101593. https://doi.org/10.1016/j.jsis.2020.101593

[15] OpenAI 2023. GPT-4 Technical Report. http://arxiv.org/abs/2303.08774 arXiv:2303.08774 [cs].

[16] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, & S. S. Iyengar. 2018. A Survey on Deep Learning: Algorithms, Techniques, and Applications. *ACM Computing Surveys 51*(5), 92:1–92:36. https://doi.org/10.1145/3234150

[17] Victor Sanh, Lysandre Debut, Julien Chaumond, & Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*. http://arxiv.org/abs/1910.01108 arXiv: 1910.01108.

[18] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, … Thomas Wolf. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. https://doi.org/10.48550/arXiv.2211.05100 arXiv:2211.05100 [cs].

[19] UN General Assembly (70th Sess.: 2015–2016) and UN Department of Economic and Social Affairs Division for Sustainable Development Goals. 2015. Transforming our world: the 2030 Agenda for Sustainable Development. Technical Report. https://digitallibrary.un.org/record/1654217

[20] Elham Shafiei Gol, Mari-Klara Stein, & Michel Avital. 2019. Crowdwork platform governance toward organizational value creation. *The Journal of Strategic Information Systems 28*(2), 175–195. https://doi.org/10.1016/j.jsis.2019.01.001

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, & Illia Polosukhin. 2017. Attention is All you Need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, & Roman Garnett (Eds.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017, Long Beach, CA, USA* (pp. 5998–6008). http://papers.nips.cc/paper/7181-attention-is-all-you-need

[22] Gerit Wagner, Julian Prester, and Guy Paré. 2021. Exploring the boundaries and processes of digital platforms for knowledge work: A review of information systems research. *The Journal of Strategic Information Systems* 30, 4 (Dec. 2021), 101694. https://doi.org/10.1016/j.jsis.2021.101694

[23] Albert Weichselbraun. 2021. Inscriptis – A Python-based HTML to text conversion library optimized for knowledge extraction from the Web. *Journal of Open Source Software 6*(66), 3557. https://doi.org/10.21105/joss.03557

[24] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, & Ji-Rong Wen. 2023. A Survey of Large Language Models. http://arxiv.org/abs/2303.18223 arXiv:2303.18223 [cs].

In: W. Semar (Hrsg.): Nachhaltige Information — Information für Nachhaltigkeit. Tagungsband des 17. Internationalen Symposiums für Informationswissenschaft (ISI 2023), Chur, Schweiz, 7.—9. November 2023. Glückstadt: Verlag Werner Hülsbusch, S. 294—311. DOI: https://dx.doi.org/10.5281/zenodo.10009338