1. Mateusz Zenon ŁĘPICKI, 2. Jarosław KUREK

Department of Artificial Intelligence, Institute of Information Technology, Warsaw University of Life Sciences
ORCID: 1. 0000-0001-7154-2630; 2. 0000-0002-2789-4732

doi:10.15199/48.2025.03.44

Efficient information extraction from resumes using small language models for SMEs based on Zero-Shot learning approach

***Abstract**. In today's recruitment field, accurately extracting information from resumes is crucial. This paper looks at how three small language models—Llama 2, Llama 3, and Phi-3—can help with this task using a Zero-Shot approach. We checked how well these models perform by comparing their results with a hand-made dataset, focusing on accuracy and the time each model takes to run on computers that small businesses typically use. Our tests showed that even with different rules for resume data in various countries, these small, local models work well and can be used by small companies on their own equipment. We used a simple prompt in our tests, and the models performed reliably, proving their usefulness in real-world hiring situations. Our results show that small language models like Llama 2, Llama 3, and Phi-3 can accurately and efficiently extract information from resumes, helping small businesses handle resume data according to local regulations. This study highlights how these models can improve the job-matching process for smaller companies.*

***Streszczenie.** W dzisiejszym procesie rekrutacyjnym dokładne wyodrębnianie informacji z CV jest kluczowe. W niniejszym artykule analizujemy, w jaki sposób trzy małe modele językowe — Llama 2, Llama 3 i Phi-3 — mogą pomóc w tym zadaniu, korzystając z podejścia Zero-Shot. Sprawdziliśmy, jak dobrze te modele radzą sobie poprzez porównanie ich wyników z ręcznie stworzonym zbiorem danych, koncentrując się na dokładności oraz czasie potrzebnym na przetwarzanie danych na komputerach typowo używanych przez małe przedsiębiorstwa. Nasze testy wykazały, że nawet przy różnych zasadach dotyczących danych z CV w różnych krajach, te małe, lokalne modele działają dobrze i mogą być używane przez małe firmy na ich własnym sprzęcie. W naszych testach użyliśmy prostego promptu, a modele działały niezawodnie, potwierdzając swoją użyteczność w rzeczywistych sytuacjach związanych z rekrutacją. Wyniki pokazują, że małe modele językowe, takie jak Llama 2, Llama 3 i Phi-3, mogą dokładnie i efektywnie wyodrębniać informacje z CV, pomagając małym firmom w zarządzaniu danymi zgodnie z lokalnymi przepisami. To badanie podkreśla, w jaki sposób te modele mogą poprawić proces dopasowywania kandydatów do ofert pracy dla mniejszych firm. (**Efektywne wydobywanie informacji z CV przy użyciu małych modeli językowych SME w oparciu o podejście Zero-Shot Learning**)*

Keywords: information extraction, pretrained NLP models, small LLM, recruitment process, Zero-Shot Learning, resume data processing, Llama 2, Llama 3, Phi-3, natural language processing
Słowa kluczowe: ekstrakcja informacji, wstępnie trenowane modele NLP, małe modele LLM, proces rekrutacji, Zero-Shot Learning, przetwarzanie danych z CV, Llama 2, Llama 3, Phi-3, przetwarzanie języka naturalnego

Introduction

The recruitment process, a critical component of human resource management, has been significantly transformed in the digital age. Traditional methods of talent acquisition, often slow and reliant on subjective human judgment, are increasingly being enhanced and, in some cases, replaced by artificial intelligence (AI)-driven solutions [1]–[3]. This paper explores a specific facet of this technological evolution: the use of small language models (LLMs) such as Llama 2, Llama 3, and Phi-3 [7] for extracting information from candidate resumes using a Zero-Shot approach.

This study serves as an initial pilot investigation, with plans for further fine-tuning and refinement in subsequent research phases. By applying a step-by-step methodology, we aim to progressively improve the models' accuracy and reliability, making them more adept at handling the complexities of the recruitment domain.

Recruitment is a complex process, requiring a balance between speed and the accurate, fair assessment of candidates. Traditional methods have struggled to keep up with the rapidly evolving job market, where new roles continually emerge, and the skill requirements for existing roles change. Additionally, the high volume of job applications, particularly in densely populated industries, can overwhelm human resources, leading to inefficiencies and potential biases in candidate selection. These challenges highlight the need for an intelligent, adaptive, and efficient recruitment process that small LLMs can help address effectively.

The extraction of basic information from candidate profiles is often a long and tedious process for HR experts [8]. This repetitive task not only consumes valuable time but also distracts HR professionals from more strategic activities that require human insight and judgment. Automating such routine tasks can significantly accelerate the recruitment process, allowing HR experts to focus on more complex decision making. Furthermore, automation can enhance the accuracy of identifying relevant skills and qualifications, which HR experts might miss due to a lack of domain-specific knowledge.

However, resume data is subjected to various regulations in different countries, complicating the use of automated solutions. In many cases, processing this data through external providers necessitates a compliance with stringent legal requirements, which can be a significant hurdle. By applying small, local LLMs that can be run on-premise, smaller companies can manage resume data processing internally, ensuring compliance with local regulations while maintaining control over sensitive candidate information. This approach not only streamlines the recruitment process but also mitigates the risks associated with data privacy and regulatory compliance.

Data Set

The dataset comprises images of holes drilled during the experiment. The images were collected in collaboration with the Institute of Wood Sciences and Furniture at the Warsaw University of Life Sciences. A standard CNC vertical machining centre, Busellato Jet 100, Thiene, Italy, was used for the drilling process. The material drilled was a standard laminated chipboard (U511SM – Swiss Krono 88 Group), typically used in the furniture industry, with dimensions of 2500x300x18. A 12mm Faba WP-01 drill with a tungsten carbide tip was utilized.

We manually selected 20 resumes [14], each with different layouts and representations of information. All the resumes were valid PDF files containing both text and images. The set is a representation of the bigger dataset of

which we have consents to work on. Table I shows small summarization of the data. It is worthwhile to mention that 2-column layout is slightly more challenging to extract information, so we focused on having significant representation of this kind. Also, most of the LLM models excel in the English language, while have some problems with the Polish one, but due to the expertise of our candidates, the representation of Polish-only resumes is very small, so we added one for proper representation.

Table 1. Dataset overview.

Category	Count
1-column layout resumes 11	11
2-columns layout resumes 9	9
English language resumes 19	19
Polish language resumes	1

Zero-Shot Learning

Zero-Shot Learning (ZSL) [4], [9] is a cutting-edge approach in machine learning where a model predicts or recognizes classes it has not encountered during training. Unlike traditional models that need extensive labeled datasets, ZSL models use relationships and attributes from seen classes to generalize to unseen ones. This method bridges the gap between training data and real-world scenarios by using semantic representations like attribute vectors or textual descriptions [4]–[6], [9]–[12].

In recruitment, ZSL is particularly beneficial as it adapts to diverse candidate profiles without constant retraining. By understanding the semantic representation of candidate qualifications, ZSL models can effectively extract candidates information they weren't explicitly trained on, enhancing the recruitment process's efficiency and adaptability.

Applying ZSL in recruitment also helps reduce biases inherent in traditional models that depend on historical data. By focusing on the attributes and capabilities possessed by candidates, ZSL promotes a more meritocratic and inclusive talent acquisition process. Thus, ZSL offers a flexible, efficient, and fair mechanism for candidate data extraction.

Large Language Models (LLMs)

In the recent years, the development of Large Language Models (LLMs) has revolutionized natural language processing tasks. These models leverage deep learning techniques to understand, generate, and manipulate human language with a high degree of accuracy. In our study, we specifically selected three LLMs: Llama 2, Llama 3, and Phi-3 (Table II), due to their distinctive characteristics and performance capabilities, which make them well-suited for the task of resume information extraction using a Zero-Shot approach [15], [16].

Table 2. Characteristics of LLAMA 2, LLAMA 3 and PH-3 Models.

Characteristic	Llama 2	Llama 3	Phi-3
Model Size	7B	8B	14B
Architecture	Transformer	Transformer	Transformer
Context Length	4096 tokens	8192 tokens	4096 tokens
Special Features	Optimized for inference	Improved accuracy and inference speed	High accuracy with large scale text and code data
Typical Use Cases	Text summarization, classification	Text summarization, classification, Q&A	Advanced text and code understanding
Availability	Open source	Open source	Open source

The choice of Llama 2, Llama 3, and Phi-3 for our study was driven by several key factors [15]–[17]:

- Performance and Accuracy - Llama 2 and Llama 3 are known for their high performance in natural language understanding and generation tasks. Llama 2, with its 7 billion parameters, offers a balance between computational efficiency and accuracy, making it ideal for environments with limited resources. Llama 3, slightly larger with 8 billion parameters, provides enhanced accuracy and faster inference speeds, which are crucial for real-time applications. Phi-3, the largest among the three with 14 billion parameters, excels in understanding complex texts and code, offering superior accuracy. Its large-scale architecture allows it to process and analyze extensive textual data effectively, which is beneficial for extracting detailed information from resumes.
- Context Length and Usability - The context length of these models is another critical factor. Llama 2 and Phi-3 both support a context length of 4096 tokens, which is adequate for most text processing tasks. Llama 3, with an extended context length of 8192 tokens, can handle more extensive inputs, making it particularly useful for summarizing long resumes or documents without losing important information.
- Open Source Availability - All three models are open source, which is a significant advantage for small and medium-sized enterprises (SMEs) that may not have the resources to invest in proprietary solutions. Open source models provide the flexibility to customize and adapt the models to specific needs without significant financial investment.

Prompt Specific Information From Resumes

To evaluate the performance of the selected models, we conducted three runs with the same prompt to extract specific information from the resumes.

The following prompt was used in numerical experiments:

"role": "system",

"content": """You are a HR assistant.

The candidate resume is at the end and is

unstructured. Your job is to fill <blank>

in the example form. Do not answer with the

whole resume. Fill only the form between tags

<form> and </form>, do not add any other

description. If you can't find the information,

leave the <blank> not filled. Response only with the filled form:

<form>

Candidate Name: <blank_name>

Candidate Surname: <blank_surname>

Candidate Email: <blank_email>

Candidate Phone Number: <blank_phone_number>

Candidate Skills: <blank_skills>

Candidate Experience: <blank_experience>

</form>

"""

Numerical Experiments

Using the Tesseract toolset [13], we extracted the text while preserving the original layout and loaded it directly into a DataFrame using Python's Pandas library. After loading, the resumes became an unstructured collection of information, including some artifacts.

Next, we created a second dataset by manually extracting key information from the resumes processed in the previous step. This dataset included the following fields: "name", "surname", "email address", and "phone". All entries were standardized to lowercase letters, and spaces were removed from phone numbers to ensure consistency.

As the next step, we ran all three models—Llama 2, Llama 3, and Phi-3—in a loop for each resume entry. For this purpose, we utilized the Transformers library from Hugging Face, ensuring that each model was instantiated afresh for every call. To maintain consistency and accuracy, we conducted three separate runs for each model. This thorough approach allowed us to capture a comprehensive set of results and minimize any potential anomalies.

The computational environment for our experiments was a server equipped with two Nvidia A40 graphics cards, providing robust processing power. The server operated on Ubuntu 22.04 LTS, a stable and reliable Linux distribution well-suited for machine learning tasks. By leveraging this high-performance setup, we ensured that the models operated efficiently, and we could measure their performance accurately under realistic conditions. This setup is representative of the type of infrastructure accessible to small and medium-sized enterprises, highlighting the practicality of deploying such models in real-world recruitment scenarios.

Method for Calculating Accuracy

In the context of evaluating language models for extracting information from resumes, accuracy is a critical metric.

A. Similarity Calculation

To compare the extracted data with the reference data, we use a similarity measure. The similarity between two strings is computed using the ratio of matches between the two sequences, which can be mathematically represented as follows:

$$\mathrm{s}(a,b) = \frac{2 \cdot \mathrm{matches}(a,b)}{\mathrm{len}(a) + \mathrm{len}(b)} \quad (1)$$

where matches(a, b) is the number of matching characters between the two strings a and b, and len(a) and len(b) are the lengths of the strings a and b, respectively.

For a more detailed comparison, the similarity scores are computed for each field (name, surname, email, phone number) individually. This allows us to evaluate the model's performance on each specific type of information extracted from the resumes.

B. Accuracy Calculation

The overall accuracy for a given set of fields is calculated as the average of the similarity scores of the individual fields. This can be expressed with the following formula:

$$\mathrm{acc}(e) = \frac{\mathrm{s}(e_1,r_1) + \mathrm{s}(e_2,r_2) + \mathrm{s}(e_3,r_3) + \mathrm{s}(e_4,r_4)}{4} \quad (2)$$

where e represents the extracted data, r represents the reference data, and e1, e2, e3, and e4 are the fields for name, surname, email, and phone number, respectively.

C. Generating Statistics

To evaluate the performance of the models, we compute the similarity and accuracy for each entry in the dataset across different models (Llama 2, Llama 3, and Phi-3). For each model, we calculate the similarity scores for each field and then the overall accuracy as described above.

Results

The primary goal of this study was to assess the time and accuracy of information extraction from real-world candidate resumes using a Zero-Shot Learning approach. Accuracy was determined by comparing the word similarity between the information extracted by each model and a manually created dataset. Here, we discuss the results of our experiments.

Table 3. Performance of LLAMA 2, LLAMA 3 and Phi3 Models.

Run	Model	Accuracy	Execution Time (s)
1	Llama 2	78.47 %	18.34
1	Llama 3	98.06 %	20.87
1	Phi-3	96.79 %	20.86
2	Llama 2	74.75 %	17.48
2	Llama 3	95.48 %	19.82
2	Phi-3	96.79 %	20.44
3	Llama 2	81.07 %	18.55
3	Llama 3	95.79 %	20.26
3	Phi-3	96.79 %	20.50

We found that the accuracy of Llama 2 was reduced due to it incorrectly filling the "name" parameter with the candidate's full name. Llama 3 and Phi-3 understood this task better. Future improvements to the prompt could significantly enhance Llama 2's accuracy.

The execution times for these models were very similar, ranging from 17 to 21 seconds. This speed allows for the models to be integrated as background tasks in day-to-day recruitment activities, enabling the recruitment team to focus on other aspects of the candidate process.

Discussion

This study presented an in-depth comparison of several AI architectures The results of our study underscore the potential of small Large Language Models (LLMs) in the recruitment process, particularly for information extraction from resumes using a Zero-Shot Learning approach.

Our experiments demonstrated that small LLMs like Llama 2, Llama 3, and Phi-3 are capable of accurately extracting key information from resumes without prior training on the specific data sets. The success of these models, particularly Llama 3 and Phi-3, suggests that even resource-constrained environments can leverage advanced AI technologies for enhancing the recruitment process.

The performance of Llama 3, with its balance of speed and accuracy, positions it as a highly viable option for small and medium-sized enterprises (SMEs) looking to integrate AI into their HR processes. Its near-parity with the larger Phi-3 model in terms of accuracy, combined with its relatively lower computational requirements, highlights the efficiency gains that can be achieved with modest infrastructure investments.

One of the key strengths of our approach lies in the practical applicability of the models used. All three models demonstrated the ability to process and extract relevant information from resumes quickly enough to be integrated into daily recruitment workflows.

However, the study also highlighted some limitations. The Llama 2 model, while generally effective, occasionally misinterpreted the prompt, leading to reduced accuracy. This suggests that further refinement in prompt engineering is

necessary to optimize the performance of smaller models. Additionally, the variability in accuracy across different runs indicates that consistency is an area that needs addressing, possibly through ensemble methods or additional fine-tuning.

Conclusion

In this paper, we explored the potential of small LLM models and the Zero-Shot Learning approach for information extraction from candidate resumes. The results demonstrated very promising applications of small models executed entirely in on-premise environments, making them accessible to smaller companies.

The "Llama 2" model, although the smallest, exhibited commendable performance and accuracy in data extraction. Crafting an effective prompt for this model can be more challenging, but it already shows significant potential for practical use. Despite its size, Llama 2's efficiency highlights the viability of small models in real-world applications.

The "Llama 3" model, an advancement over the "Llama 2" architecture, demonstrated superior accuracy. Its execution time was only marginally slower than its predecessor, and in some runs, it even outperformed the nearly twice-as-large "Phi-3" model. This balance of size and performance makes Llama 3 a robust choice for efficient information extraction.

The "Phi-3" model, being the largest of the trio, generally delivered the best results. However, its performance was not significantly better than that of Llama 3, suggesting that size alone does not determine effectiveness. The consistent accuracy of Phi-3 underscores its reliability, though it might not always justify the additional computational resources required.

We believe that further improvements in prompt engineering and the introduction of Few-Shot techniques could enhance the accuracy of these models even more. Such advancements would solidify the role of small LLMs and Zero-Shot Learning in streamlining the recruitment process, offering scalable and efficient solutions for small and medium-sized enterprises.

REFERENCES

[1] Patil, A.; Suwalka, D.; Kumar, A.; Rai, G.; Saha, J. A Survey on Artificial Intelligence (AI) based Job Recommendation Systems. In Proceedings of the 2023 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 23–25 March 2023; pp. 730–737. https://doi.org/10.1109/ICSCDS56580.2023. 10104718.

[2] Thali, R.; Mayekar, S.; More, S.; Barhate, S.; Selvan, S. Survey on Job Recommendation Systems using Machine Learning. In Proceedings of the 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Coimbatore, India, 10–11, January 2024; pp. 453–457. https://doi.org/10.1109/ICIDCA56705. 2023.10100122.

[3] Kamble, A.; Tambe, S.; Bansode, H.; Joshi, S.; Raut, S. Job Recommendation System for Daily Paid Workers using Machine Learning. Int. J. Res. Appl. Sci. Eng. Technol. 2023, 11, 3086–3089. https: //doi.org/10.22214/ijraset.2023.50876.

[4] Xian, Y.; Lampert, C.H.; Schiele, B.; Akata, Z. Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly. IEEE Trans. Pattern Anal. Mach. Intell. 2019, 41, 2251–2265. https://doi.org/10.1109/TPAMI.2018.2857768.

[5] Socher, R.; Ganjoo, M.; Sridhar, H.; Bastani, O.; Manning, C.D.; Ng, A.Y. Zero-Shot Learning Through Cross-Modal Transfer. Adv. Neural Inf. Process. Syst. 2013, 26, 1–10.

[6] Wang, J.; Krishnan, A.; Sundaram, H.; Li, Y. Pre-trained Neural Recommenders: A Transferable Zero-Shot Framework for Recommendation Systems. arXiv 2023, arXiv:2309.01188.

[7] Zheng, Z.; Qiu, Z.; Hu, X.; Wu, L.; Zhu, H.; Xiong, H. Generative Job Recommendations with Large Language Model. arXiv 2023, arXiv:2307.02157.

[8] Kumari, S.V. Job Recommendation System Using NLP. Int. J. Eng. Sci. 2023, 11, 2721–2727. https://doi.org/10.22214/ijraset.2023.52183.

[9] Tiong, A.M.H.; Li, J.; Li, B.; Savarese, S.; Hoi, S.C.H. Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training. arXiv 2022, arXiv:2210.08773.

[10] Phan, T.; Vo, K.; Le, D.; Doretto, G.; Adjeroh, D.; Le, N. ZEETAD: Adapting Pretrained Vision-Language Model for Zero-Shot End-to-End Temporal Action Detection. arXiv 2023, arXiv:2311.00729.

[11] Öztürk, E.; Ferreira, F.; Jomaa, H.S.; Schmidt-Thieme, L.; Grabocka, J.; Hutter, F. Zero-Shot AutoML with Pretrained Models. arXiv 2022, arXiv:2206.08476.

[12] Kang, H.; Blevins, T.; Zettlemoyer, L. Translate to Disambiguate: Zeroshot Multilingual Word Sense Disambiguation with Pretrained Language Models. arXiv 2023, arXiv:2304.13803.

[13] Kay, A. Tesseract: an open-source optical character recognition engine. Linux J.. 2007, 2 (2007,7)

[14] Kurek, J., Latkowski, T., Bukowski, M., ´ Swiderski, B., Ł˛epicki, M., Baranik, G., Nowak, B., Zakowicz, R. & Dobrakowski, Ł. Zero-Shot Recommendation AI Models for Efficient Job–Candidate Matching in Recruitment Process. Applied Sciences. 14 (2024), https://www.mdpi.com/2076-3417/14/6/2601

[15] Lin, Y., Tang, H., Yang, S., Zhang, Z., Xiao, G., Gan, C. & Han, S. QServe: W4A8KV4 Quantization and System Co-design for Efficient LLM Serving. ArXiv Preprint ArXiv:2405.04532. (2024)

[16] Abdin, M., Jacobs, S., Awan, A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H. & Others Phi-3 technical report: A highly capable language model locally on your phone. ArXiv Preprint ArXiv:2404.14219. (2024)

[17] Ferrag, M., Alwahedi, F., Battah, A., Cherif, B., Mechri, A. & Tihanyi, N. Generative AI and Large Language Models for Cyber Security: All Insights You Need. ArXiv Preprint ArXiv:2405.12750. (2024)