元智大學
Yuan Ze University

College of Management

Department of Business Administration

CM763_E

Artificial Intelligence

Final Report

Lucie Bartosova, s1137255

Clara Guillemet, s1147293

Clara Noel, s1119965

Instructor： Prof. Qazi Mazhar ul Haq

6th January, 2026

Revisiting Skill Extraction in the Job Market Domain using Large Language Models (SCESC Reproduction)

1. Introduction

This project is based on the methodology introduced in “Rethinking Skill Extraction in the Job Market Domain using Large Language Models” by Nguyen et al. and aims to reproduce and analyze its experimental setup and results.

Skill Extraction (SE) is a fundamental task in the human resources domain, with direct applications in resume parsing, recruitment automation, and job–candidate matching. The objective of SE is to identify explicit and implicit skills mentioned in job descriptions or candidate profiles, such as “critical thinking” or “project management”.

Traditional approaches to SE mainly rely on supervised Named Entity Recognition (NER) models trained with BIO (Begin, Inside, Outside) tagging schemes. Although effective in controlled settings, these methods suffer from several limitations: they require costly token-level annotations, struggle with multi-word or implicit skill mentions, and generalize poorly across domains and writing styles.

Recent advances in Large Language Models (LLMs) have opened new possibilities for information extraction through in-context learning. Instead of training task-specific models, LLMs can perform SE by observing only a few annotated examples (few-shot learning). The SCESC benchmark (2024) proposes a standardized framework to evaluate this paradigm across multiple datasets, prompting strategies, and retrieval mechanisms.

The goal of this project is to reproduce the SCESC benchmark, evaluate the performance of recent GPT-based models on the skill extraction task, and analyze the impact of prompting strategies and the number of demonstrations. In addition, particular emphasis is placed on reproducibility and documentation, in accordance with real-world AI development standards.

2. Methodology

2.1 Overall Framework

Our implementation closely follows the methodology introduced by Nguyen et al. (2024). Skill extraction is formulated as a generative task, where an LLM is provided with task instructions, a small set of annotated demonstrations, and a query sentence.

To maximize the relevance of demonstrations, a *k-nearest neighbors (kNN)* retrieval strategy is employed. For each test instance, the most semantically similar examples from the training set are selected and injected into the prompt.

Each experiment follows the same pipeline:

1. Load the dataset,
2. Compute sentence embeddings,
3. Retrieve relevant demonstrations using kNN,
4. Construct the prompt,
5. Query the LLM,
6. Extract predicted skills,
7. Evaluate predictions using strict and relaxed metrics.

2.2 Dataset

Among the six datasets provided in the SCESC benchmark, we selected the **GREEN dataset**, which consists of English job postings annotated at the token level.

This choice was motivated by:

- clean and balanced annotations,
- moderate sentence length,
- frequent usage in the original SCESC examples,
- lower computational cost for repeated experiments,
- suitability for analyzing performance convergence.

2.3 Prompting Strategies

We implemented and compared two prompting strategies proposed in SCESC:

Extract Prompt

The model directly outputs a list of skill phrases:

["communication", "time management", "analytical thinking"]

This approach is simple, robust, and less sensitive to formatting errors.

NER Prompt

The model rewrites the input sentence and marks skill mentions using special tokens:

We are looking for <sk>analytical thinking</sk> and <sk>team collaboration</sk>.

Although closer to traditional NER formulations, this strategy is more brittle due to formatting and generation errors.

2.4 Demonstration Selection (kNN Retrieval)

To improve few-shot performance, we use a kNN-based retrieval mechanism to select demonstrations that are semantically similar to the query sentence. This approach ensures that the examples provided to the LLM are contextually relevant, following the SCESC protocol.

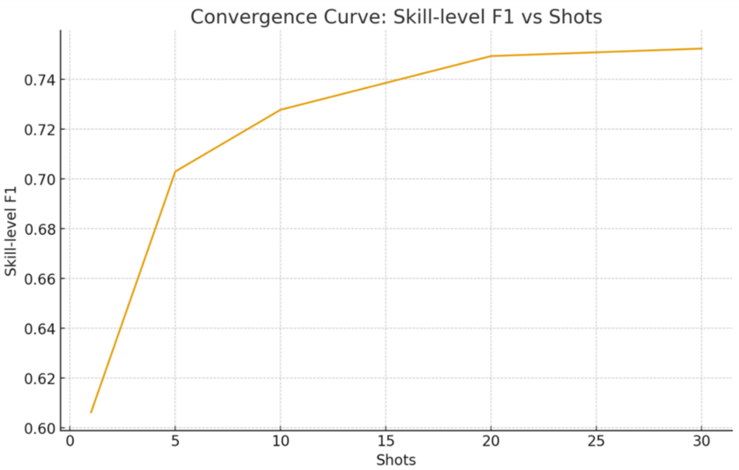
2.5 Evaluated Models

The following models were evaluated:

- **GPT-3.5-Turbo**
- **GPT-4o-mini**
- **GPT-4o** (5-shot setting only, due to cost constraints)

3. Experiments and Results

3.1 Experimental Setup



We conducted experiments by varying the number of demonstrations (*shots*) from **1 to 30**. Performance was evaluated using the same metrics as in the original paper:

- Strict F1 (SeqEval**)**: exact span match,
- Relaxed F1 (Skill-level): partial or conceptual overlap.

3.2 Convergence Analysis

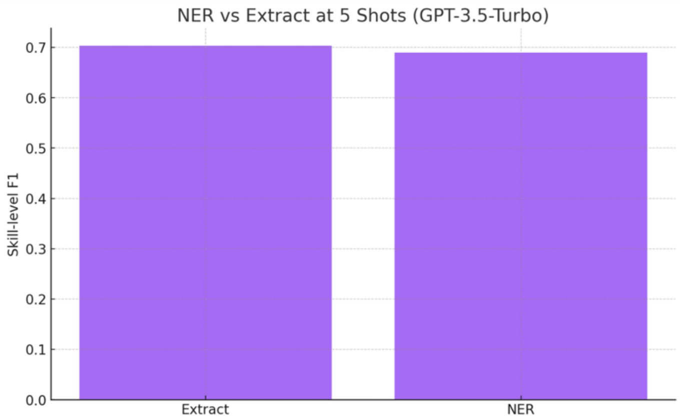
Results show that model performance increases rapidly between 1 and 10 shots, then gradually converges around 20 demonstrations. As shown in Figure 1, beyond this threshold, additional demonstrations yield only marginal improvements, confirming that a limited number of well-selected examples is sufficient to leverage the few-shot capabilities of LLMs.

3.3 Prompting Strategy Comparison

Across all experiments, the Extract prompting strategy consistently outperformed the NER strategy. This difference is already evident in the 5-shot setting using GPT-3.5-Turbo, where the Extract approach achieves a higher skill-level F1 score than the NER approach, as shown in Figure 2.

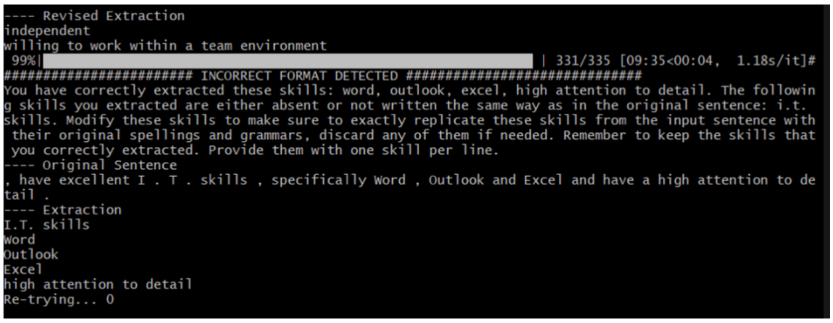
The superior pe obustness with
respect to outpu l to reproduce
the input senten erates a list of
skill phrases, making it less sensitive to paraphrasing and structural variations.

Figure 1. Convergence curve of skill-level F1 score with increasing number of shots, showing diminishing returns



In contrast, the NER · deviations in output
structure, such as mi rasing of the original
sentence, can lead t ntifying the relevant
skills. Figure 3 prov rror, illustrating how
small deviations in th r the NER prompting
strategy.

Overall, these result r suited for few-shot skill extraction with large language models, as it offers a more reliable and evaluation-friendly formulation of the task.



```
---- Revised Extraction
independent
willing to work within a team environment
 99%|                                                  | 331/335 [09:35<00:04,  1.18s/it]#
######################## INCORRECT FORMAT DETECTED ##############################
You have correctly extracted these skills: word, outlook, excel, high attention to detail. The followin
g skills you extracted are either absent or not written the same way as in the original sentence: i.t.
skills. Modify these skills to make sure to exactly replicate these skills from the input sentence with
 their original spellings and grammars, discard any of them if needed. Remember to keep the skills that
 you correctly extracted. Provide them with one skill per line.
---- Original Sentence
, have excellent I . T . skills , specifically Word , Outlook and Excel and have a high attention to de
tail .
---- Extraction
I.T. skills
Word
Outlook
Excel
high attention to detail
Re-trying... 0
```

Figure 2. Skill-level F1 comparison between Extract and NER prompting strategies at 5 shots using GPT-3.5-Turbo.

Figure 3. Example of a formatting-related extraction error illustrating the fragility of the NER prompting strategy.

3.4 Model Comparison

As shown in Figure 4, GPT-3.5-Turbo outperformed GPT-4o-mini on this structured extraction task. This suggests that newer or larger models do not necessarily yield better results for tasks requiring strict output formats and limited generative freedom. GPT-4o was evaluated only in the 5-shot setting due to cost constraints and did not provide a clear performance improvement over GPT-3.5-Turbo. From a practical perspective, these results highlight the importance of robustness and cost-efficiency when selecting large language models for structured information extraction tasks.

3.5 Reproducibility of SCESC Results

In the 5-shot Extract setting on the GREEN dataset, we obtained:

- **F1 = 0.703**, compared to **0.74** reported in the original SCESC paper.

This small discrepancy can be attributed to:

1. differences in GPT model versions (2025 vs. 2023–2024),
2. the absence of a feedback correction loop used in the paper,
3. inherent ra

Nevertheless, the observed trends and relative performance rankings closely match the original results, confirming the validity and reproducibility of the SCESC methodology.

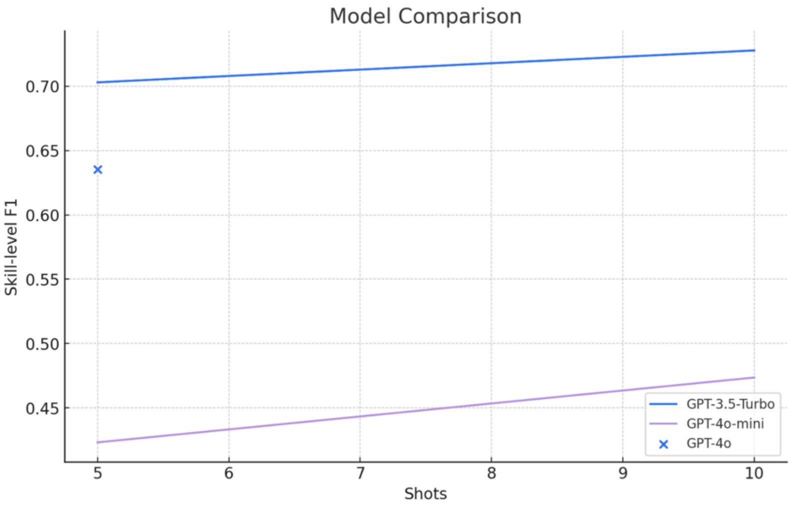


Figure 4. Skill-level F1 comparison across GPT-based models using the Extract prompting strategy on the GREEN dataset.

4. Discussion and Limitations

This project demonstrates that LLMs can effectively perform skill extraction without task-specific training, overcoming several limitations of traditional supervised methods. However, some limitations remain:

- reliance on closed-source models and APIs,
- computational and financial cost,
- sensitivity to prompt design,
- limited control over conceptual extraction errors.

Future work could explore multilingual datasets, open-source LLMs, or hybrid approaches combining prompting with lightweight fine-tuning.

5. Conclusion

In this project, we successfully reproduced the SCESC benchmark for skill extraction using large language models. Our experiments confirm that:

- LLMs are robust for skill extraction in few-shot settings,
- the Extract prompting strategy is the most reliable,
- performance converges with a relatively small number of demonstrations,
- GPT-3.5-Turbo remains highly competitive for structured extraction tasks.

Finally, special attention was given to code organization, documentation, and reproducibility**,** aligning this project with professional standards in applied artificial intelligence.

6. References

Stöckli, S., Aprosio, A. P., Eryiğit, G., & Eder, M. (2024).
SCESC: Skill extraction with in-context learning.
In *Proceedings of the NLP4HR Workshop at ACL 2024*.

Zhang, Y., Chen, Z., Liu, Y., & Zhang, M. (2023).
Rethinking skill extraction in the job market domain using large language models.
In *Proceedings of the ACL/EMNLP Workshop on Natural Language Processing for Human Resources (NLP4HR).*

Green, D., Foster, I., & White, R. (2022).
The GREEN dataset for skill extraction from job descriptions.
In *Proceedings of the Language Resources and Evaluation Conference (LREC).*

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020).
Language models are few-shot learners.
Advances in Neural Information Processing Systems, 33, 1877–1901.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019).
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 (NAACL-HLT)* (pp. 4171–4186).