Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Des Prompts aux Profils: Evaluation de la qualité des données générées par LLM pour la classification des soft skills

Elena Rozera[2,3] Nédra Mellouli-Nauwynck[1, 2] Patrick Leguide[3] William Morcombe[3]
(1) De Vinci Higher Education, DVRC, Paris, France
(2) LIASD, Université Paris 8, Vincennes Saint-Denis, France
(3) Central Test, Paris, France
elena.rozera@devinci.fr, n.mellouli@iut.univ-paris8.fr

RÉSUMÉ

L'extraction automatique des soft skills à partir de CV constitue un enjeu central du Traitement Automatique du Langage Naturel (TALN) pour les ressources humaines. Toutefois, le manque de données annotées et les contraintes de confidentialité limitent le développement de modèles robustes. Cette étude préliminaire explore le potentiel des Grands Modèles de Langage (LLMs) pour générer des CV synthétiques dédiés à la classification des soft skills. Deux corpus sont proposés, un jeu de données de référence généré à partir de prompts explicites, et un corpus de CV complets produits selon une structure réaliste. Un cadre d'évaluation combinant des métriques avec et sans référence est mis en place, afin de mesurer la diversité, la redondance et la fidélité sémantique. Les résultats révèlent des compromis importants entre diversité lexicale et réalisme contextuel, apportant des pistes pour guider la génération future de données synthétiques pour la classification des compétences comportementales.

ABSTRACT

From Prompts to Profiles : Evaluation of the quality of LLM-generated Data for Soft Skills Classification

The automatic extraction of soft skills from resumes has emerged as a critical task in Natural Language Processing (NLP) for human resources applications. However, the scarcity of annotated datasets and privacy concerns over real CVs limit the development of robust models. In this preliminary study, we investigate the potential of Large Language Models (LLMs) for generating synthetic CV datasets tailored for soft skills classification. We introduce two corpora, a Baseline Dataset created via direct skill prompts, and a CV Dataset created through realistic, structured CV generation. To evaluate the quality of these datasets, we design a dual evaluation framework combining reference-free and reference-based metrics, with a focus on diversity, redundancy, and semantic fidelity. Our findings highlight important trade-offs between lexical diversity, contextual realism, and redundancy, providing insights for future synthetic data generation strategies in soft skill classification tasks.

MOTS-CLÉS : Extraction des Soft Skills, Génération de Données Synthétiques, Grands Modèles de Langage (LLMs), Classification de Texte, Évaluation de la Diversité des Données, Ingénierie de Prompts, Augmentation de Données pour le TALN.

KEYWORDS: Soft Skills Extraction, Synthetic Data Generation, Large Language Models (LLMs), Text classification, Data Diversity Evaluation, Prompt Engineering, Data Augmentation for NLP.

Actes de l'atelier Évaluation des modèles génératifs (LLM) et challenge 2025 (EvalLLM), pages 91–107.

1 Introduction

In the contemporary job market, soft skills —such as communication, teamwork, and adaptability— play a significant role in hiring decisions. The automatic extraction of soft skills from textual documents represents a task that has been widely explored in the fields of Natural Language Processing (NLP) and information retrieval. While most previous efforts frame this as a sequence labeling or information extraction problem (Fareri *et al.*, 2021; Zhang *et al.*, 2022; Wings *et al.*, 2021), our study approaches it as a classification task. Given a pre-defined taxonomy of soft skills, the objective is to assign the correct labels to a given CV. This formulation allows us to apply standard classification techniques, but it also presents a critical dependency : the need for a well-labeled dataset annotated according to our specific taxonomy. The absence of such a resource in the current literature and the specificity of our task motivated us to explore different, newer alternatives.

To face this challenge, we turn to Large Language Models (LLMs) for synthetic data generation. LLMs, such as GPT (Brown *et al.*, 2020), have demonstrated high proficiency in producing human-like text, offering a promising solution when real data is scarce. Moreover, using real CVs raises significant challenges related to data privacy and anonymization. Synthetic CVs generated by LLMs bypass this issue, offering a privacy-preserving alternative that eliminates the need for manual anonymization while still providing realistic training data. However, the efficacy of LLM-generated synthetic data is contingent upon its quality and representativeness. Past studies have highlighted the importance of evaluating synthetic data to ensure its reliability for NLP downstream NLP applications (Long *et al.*, 2024).

In this context, our paper presents a short preliminary study on using LLMs to generate synthetic data for soft skill classification. We make three contributions :

- The creation of a Baseline Dataset - Generated from explicit soft skills prompts.
- The creation of a CV Dataset - A more sophisticated corpus generated through more naturalistic prompting, simulating realistic CVs.
- An evaluation framework assessing the quality of these datasets, focusing on textual diversity and redundancy.

We aim to answer the following research question : How can we systematically evaluate LLM-generated soft skill data and use those insights to guide the creation of more diverse and realistic datasets for soft skill classification ?

2 Related Work

Our study intersects three growing areas of research : soft skill extraction from unstructured content, the use of LLMs for synthetic dataset generation, and their evaluation. In this section, we review prior work on both fronts to place our contributions within the broader literature.

2.1 Soft Skills Extraction

Soft Skills Extraction has become an increasingly relevant task in Natural Language Processing, particularly for applications in human resource automation and talent profiling. Early approaches relied on keyword matching and rule-based methods (Fareri *et al.*, 2021), but these often struggled to capture the contextual, and multi-word nature of soft skills. More recent work reframes the task as sentence-level classification or sequence labeling, leveraging contextual embeddings and syntactic cues such as POS and dependency tags to improve accuracy (Ul Haq *et al.*, 2024). Despite these advances, annotated datasets remain scarce—especially for long or context-sensitive entities—limiting model performance. While publicly available datasets, like SKILLSPAN (Zhang *et al.*, 2022) offer valuable benchmarks, they are among the few focused on soft skills. To mitigate data scarcity, methods such as data augmentation and weak supervision have been explored (Ul Haq *et al.*, 2024), though token-label misalignment and entity length continue to pose challenges in sequence labeling tasks. Despite some success using domain-adapted models or classification pipelines, many of these solutions are built around job postings and structured taxonomies like ESCO (Clavié & Soulié, 2023), leaving the linguistic variability and informality of CVs underexplored. As a result, current approaches lack robustness when applied to free-form CV text—a gap this study aims to address by generating and evaluating synthetic CV-like data.

2.2 Dataset Generation with LLMs

The lack of labeled training data is a recurring challenge for different fields in NLP (Filice *et al.*, 2023), which has led to increasing interest in LLMs as generative tools for dataset creation. Early studies synthesized labeled data using class-conditional prompts, effectively reframing zero-shot learning as a data generation task (Meng *et al.*, 2022; Ye *et al.*, 2022b). As this paradigm evolved, a number of methods focused on improving the generation pipeline. PROGEN introduced iterative feedback loops between a small model and an LLM (Ye *et al.*, 2022a), while REGEN integrated dense retrieval to improve label consistency and topic relevance (Yu *et al.*, 2023b). Other methods tackled quality and diversity through filtering and re-weighting, or added controllable attributes such as topic and tone to reduce label imbalance (Meng *et al.*, 2022; Gao *et al.*, 2022; Yu *et al.*, 2023a). Human-in-the-loop strategies such as label filtering and logit suppression further improved diversity (Chung *et al.*, 2023), though most work remains focused on general tasks like sentiment or topic classification (Li *et al.*, 2024a). Our work extends these strategies to a novel domain : soft skill extraction from CVs. This domain lacks publicly available training data, and our goal is to evaluate both direct and attribute-guided LLM-based generation for creating diverse, realistic CV-like corpora.

2.3 Evaluating the Quality of LLM-generated Data

Evaluating text generated by LLMs has evolved from simple surface-level comparisons to more nuanced, multidimensional frameworks. Traditional metrics like BLEU (Papineni *et al.*, 2002) and ROUGE (Lin, 2004), based on n-gram overlap, often poorly correlate with human judgment and struggle with semantic or contextual accuracy (Gao *et al.*, 2024). Recent methods address these gaps through embedding-based (e.g., BERTScore, GPTScore), probability-based, and prompting-based metrics, which assess semantic similarity or leverage LLMs themselves for scoring and ranking (Li *et al.*, 2024b). Reference-free approaches, such as TrustScore (Zheng *et al.*, 2024) further allow quality

estimation without gold standards. Diversity is another critical dimension, captured through metrics like self-BLEU, n-gram uniqueness, compression ratios, and syntactic repetition patterns(Shaib *et al.*, 2024b,a). Tools like the *diversity* package and LLM Cluster-agent offer standardized ways to compute and compare these indicators (Chen *et al.*, 2024; Shaib *et al.*, 2024a). Frameworks such as SynEval (Yuan *et al.*, 2024) now integrate multiple perspectives—fidelity, utility, and privacy—offering a holistic approach to evaluating synthetic data quality. This multifaceted view is essential for applications such as ours, where semantic alignment, diversity, and realism all matter for downstream model performance.

3 Methodology

In the context of our research, the automatic extraction of soft skills from CVs is challenged by the lack of annotated data. Recent work has explored both extraction models and synthetic data generation using LLMs, and our study places itself in the line of research that looks into the evaluation of the quality of such synthetic data in context-rich domains like CVs. Our work assumes prompt quality and focuses instead on assessing the resulting data itself—measuring its semantic similarity, diversity, and redundancy to inform downstream use in soft skill classification.

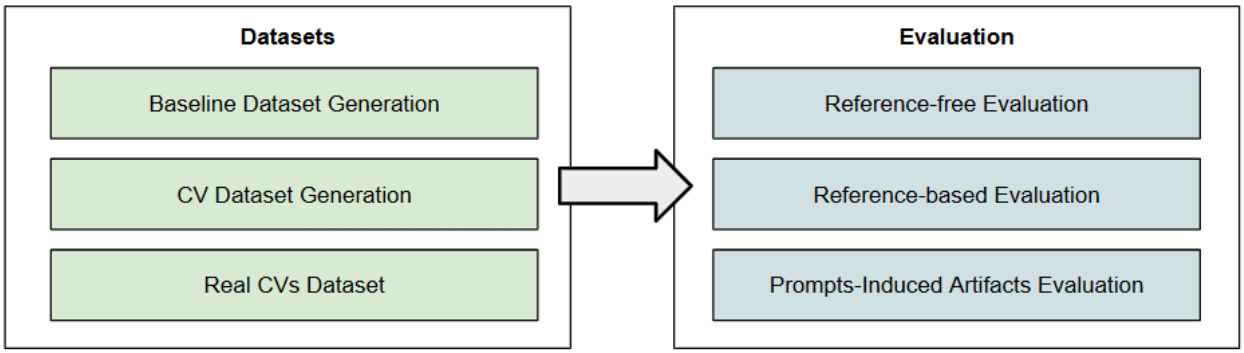

FIGURE 1 – *The Methodology pipeline*

3.1 Datasets Generation

3.1.1 The Baseline Dataset

The baseline dataset was generated using a structured single-prompting pipeline aimed at producing explicit, labeled examples of predefined soft skills [Figure 2]. The input to the generation process was a curated taxonomy of S soft skills (e.g., adaptability, leadership, collaboration), each representing a distinct target class for downstream multi-label classification.

We designed a set of P fixed prompt templates intended to simulate how different soft skills might be expressed in real-world CVs or cover letters. For each soft skill in the taxonomy, we applied all P prompts, resulting in P separate generations per skill. Each prompt asked the model (GPT-3.5) to produce a list of N short examples, typically phrased as :

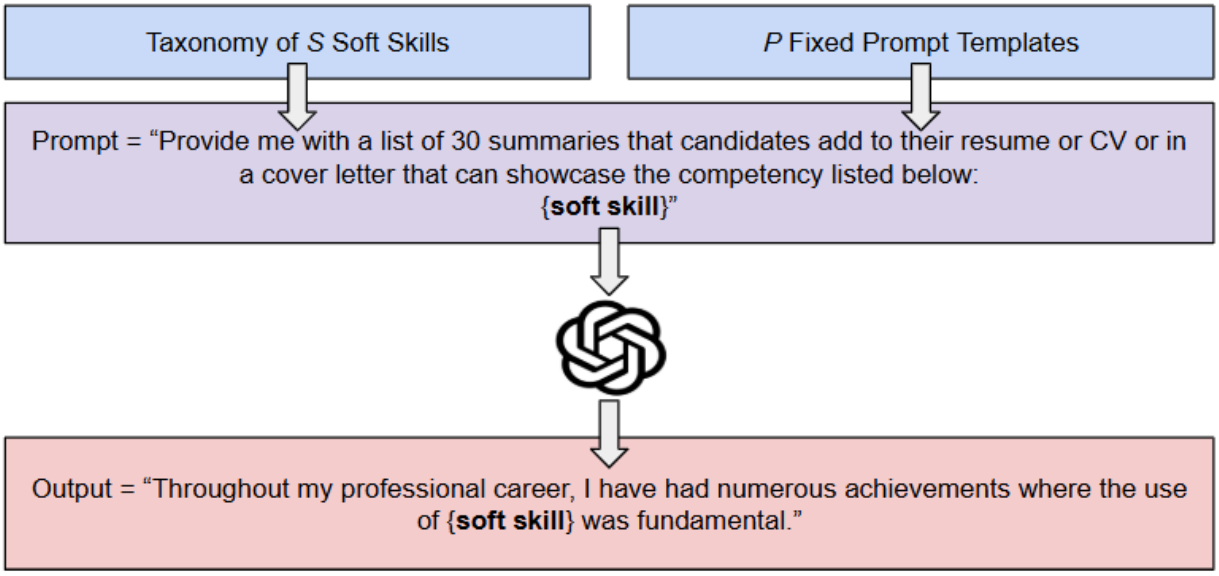

FIGURE 2 – *The Baseline dataset generation pipeline*

"Provide *N* examples of short resume or cover letter summaries that demonstrate the skill : {**soft skill**}."

This resulted in *P x N* labeled examples per soft skill, and a total dataset size of $S \times P \times N$ examples. Each generation returned sentences or paragraphs explicitly demonstrating the requested skill. For example, for adaptability, one output might be :

"Throughout my career, I have had numerous achievements where the use of {**soft skill**} was fundamental."

3.1.2 The CV Dataset

With the idea of overcoming the limitations of the baseline dataset and producing more naturalistic and contextually varied data, we developed a second generation pipeline aimed at simulating full-length CV-style documents [Figure 3]. The process was guided by a curated taxonomy of *J* job titles (e.g., *Accountant, Data Scientist*), each serving as input for a compositional prompting framework. For each job title, we defined a set of *C* CV sections—such as *About Me, Work Experience, Education, and Interests*—to mimic the structure of real-world resumes. For each section, a bank of *P* distinct prompt templates was designed to elicit varied and realistic outputs.

The prompting framework combined :

1. A **structured system prompt** that positioned the LLM (GPT-4o) as an AI specialized in CV writing, and
2. A **generic user prompt template** that instantiated the job title and section-specific instruction.

To further enhance stylistic diversity, each generation included a randomly selected writing style constraint (e.g., *formal, conversational, achievement-oriented*), validated in collaboration with an HR expert.

Each generation -carried out with a GPT-4o-mini model- produced a fragment of a CV corresponding to a specific job title and section. The final dataset was composed by concatenating these fragments to form full-length, CV-like documents. This resulted in $J \times C \times P \times N$ text samples, where N is the number of examples requested per prompt.

Unlike the baseline dataset—characterized by short, label-aligned sentences—this dataset consists of longer documents where soft skills are implicitly embedded across narrative context. As such, it offers a more challenging yet realistic substrate for training models that rely on contextual or latent representation learning for soft skill inference.

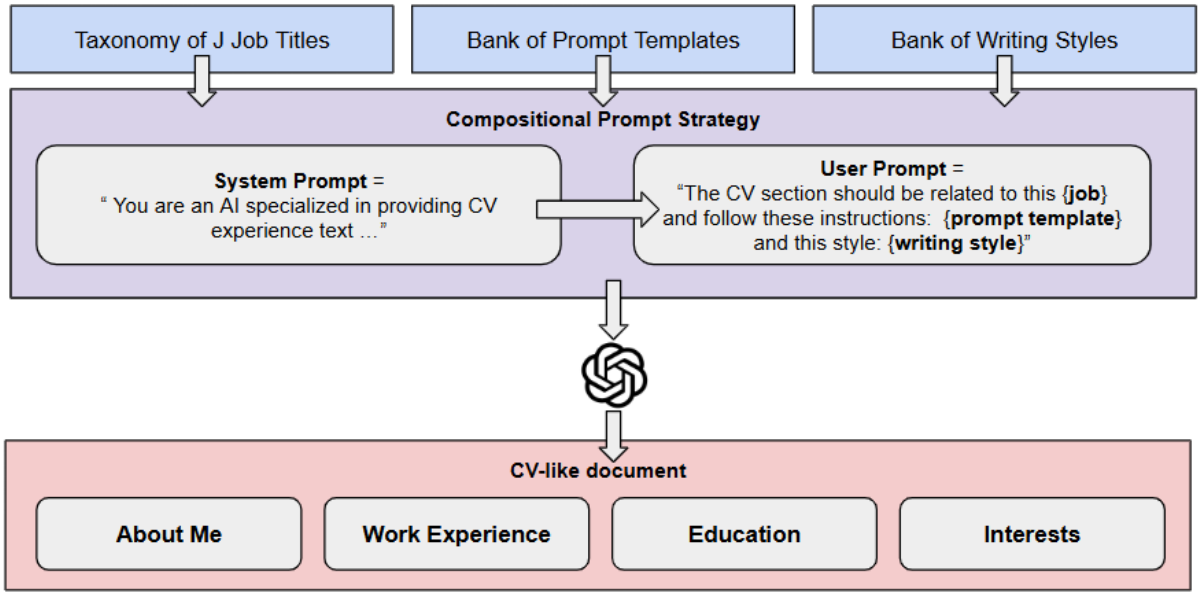

FIGURE 3 – *The CV dataset generation pipeline*

3.2 The Evaluation Framework

To assess the quality of the LLM-generated datasets we propose a reference-free evaluation framework based on the diversity python package developed by Shaib (Shaib *et al.*, 2024a), and a reference-based framework. These aim to capture key textual properties such as diversity, redundancy, and semantic fidelity. The different metrics are explained extensively in the Appendix A.

3.2.1 Reference-free metrics

For reference-free evaluation, we use standard metrics to measure :

— **Lexical and semantic diversity** - using n-gram diversity scores and BERT-based similarity.
— **Redundancy** - via Compression Ratio (CR), CR using Part-of-Speech (POS) tags for syntactic repetition, and intra-sample self-repetition.

3.2.2 Reference-based metrics

To further evaluate the quality of the LLM-generated CVs, we adopted a reference-based evaluation strategy. Each real CV was paired with a synthetic CV corresponding to the same, or very similar, job title. The synthetic CVs were manually selected and matched to ensure alignment in the role. We used three well-established metrics for reference-based evaluation : **BLEU**, **ROUGE-L**, and **BERTScore**. Each metric captures different aspects of similarity, providing a multifaceted understanding of the correspondence between real and generated CVs.

In addition, we introduced a perplexity-based evaluation to assess the stylistic proximity of generated CVs to real ones. Perplexity, in our use case, quantifies how likely a text is under a language model trained on real CVs, thus reflecting how 'natural' the generated text appears from a linguistic standpoint. We fine-tuned a lightweight GPT-2 model on our corpus of real CVs, and computed the perplexity of each synthetic CV under this model. This approach provides a complementary lens to BLEU and BERTScore by capturing writing conventions and surface-level fluency.

Together, these metrics offer complementary perspectives on how well the generated CVs replicate the form, content, semantics, and linguistic style of real-world CVs.

3.2.3 Prompt-Induced Artifactuality

To assess potential prompt-induced artifacts in the LLM-generated baseline dataset, we conducted a frequency-based comparison of soft skill phrase occurrences between baseline synthetic data and a small reference set of real CVs. Due to the imbalance in dataset sizes - 700 Baseline samples compared to the 20 real CVs - we employed a Monte Carlo sampling to ensure a fair comparison. We repeatedly sampled subsets from the Baseline dataset to match the size of the real CV set, computing the frequency of each soft skill across both datasets in each iteration. This enabled us to identify over-represented phrases in the baseline data that may indicate prompt-induced generation patterns.

4 Results and Discussion

We applied our evaluation framework to two datasets : the Baseline Dataset, generated via direct prompting with soft skills names, and the CV Dataset, generated using more naturalistic prompts simulating full resumes. Below, we discuss the main findings based on diversity, redundancy, semantic similarity, and alignment with real CVs.

4.1 Reference-free Evaluation

4.1.1 Lexical and Semantic Diversity

The Baseline dataset shows higher N-gram Diversity (2.429 vs. 1.832), as shown in Table 1 reflecting more varied lexical choices, likely due to skill-specific prompts. In contrast, the CV dataset favors longer, structured texts, which tend to reuse standard resume phrasing.
The CV dataset, despite presenting a lower lexical diversity, presents higher semantic similarity scores, with BERTScore (0.626 vs. 0.504) and Sentence-BERT (0.398 vs. 0.207). This suggests

	Baseline	CV
N-gram Diversity	2.429	1.832
BERTScore	0.504	0.626
Sentence-BERT	0.207	0.398
CR	3.646	3.797
CR :POS	3.797	3.916
Self-Repetition	1.28	7.42

TABLE 1 – *Reference-free metrics : Baseline vs CV datasets*

that while Baseline samples exhibit more varied surface forms, the CV samples may embed soft skills information within richer and more coherent narrative contexts. Such semantic density can be crucial for downstream tasks that rely on context-aware representations rather than a simpler keyword extraction.
Therefore, the CV dataset, despite its reduced surface variability, offers higher-quality input for training robust soft skill classifier.

4.1.2 Redundancy and Repetition

Baseline	**CV**
Frequent N-grams	
"i was able to" — 80 times "in my role as" — 75 times "my role as a" — 65 times	"bachelor of science in" — 320 times "about me as a" — 246 times "education bachelor of science" — 208 times
Top POS N-grams	
NNP NNP NNP NNP — 4675 times 'Sales Customer Relationship Client', 'Customer Relationship Client Understanding', 'Decision Making Problem Solving Critical'	**NNP NNP NNP NNP — 4517 times** 'Certified Professional Engineer PE', 'Public Health Graduated Magna', 'Health Graduated Magna Cum'

TABLE 2 – *Top repeated N-grams and POS n-grams in Baseline and CV datasets*

Redundancy metrics confirm higher repetition in the CV dataset across all indicators : Compression Ratio (3.797 vs. 3.646), POS Compression (3.916 vs. 3.797), and Self-Repetition (7.28 vs. 1.28), also presented in Table 1. This can be attributed to the structured, section-based prompting strategy, which, while promoting fluency and realism, inadvertently leads to consistent phrasing across and within samples. In contrast, the Baseline dataset—driven by varied prompts—exhibits a less repetitive structure. While this repetition supports fluency and realism in the CV format, it may reduce diversity and increase overfitting risks for downstream models. Future generation pipelines might benefit from balancing structural coherence with better stylistic variation.

Frequent N-grams observed in each dataset are summarized in Table 2. In the Baseline dataset, common phrases such as "I was able to" and "in my role as" appear with high frequency, reflecting the prompt-driven generation centered on specific skills. Conversely, in the CV dataset, frequent n-grams like "Bachelor of Science in" and "Education Bachelor of Science" illustrate the structured, academic-oriented language typical of real resumes, but also the tendency of the LLM to over-represent a specific educational field.

4.2 Reference-based Evaluation

Job Title	**BLEU**	**ROUGE-L**	**BERTS**
Radio/Television Presenter	0.0096	0.1492	0.8118
Full Stack Developer	0.0019	0.1148	0.8095
Communication Officer	0.0011	0.1503	0.8014
Civil Engineer	0.0006	0.1083	0.8010
Scientific Researcher	0.0019	0.1429	0.8006

TABLE 3 – *Reference-based metrics on matched CVs*

4.2.1 Lexical and Semantic Diversity

Table 3 shows the evaluation results for the five most semantically aligned CV pairs, based on their BERTScore F1 values. Despite low **BLEU** (≤ 0.0096) and **ROUGE-L** (≤ 0.1503) scores, the CV dataset shows strong semantic alignment with real CVs, as reflected in high **BERTScores** (≥ 0.8006). This confirms that lexical and structural variation—common in CVs—limits the usefulness of overlap-based metrics, while embedding-based scores capture deeper similarities. Notably, the highest BERTScores occur in technical or formal roles (e.g., developers, researchers), where LLMs likely mimic domain-specific conventions more reliably. This suggests that generation quality varies by job type, depending on how well the model can anchor to recognizable patterns.

In addition to the semantic alignment, we also examined document length. The mean length of real CVs was found to be 2,244.84 tokens, while the generated CVs averaged 5,387.55 tokens. This is a significant difference that suggests that the generated documents are more verbose, potentially due to the compositional prompting used. While longer documents may embed more contextual clues for soft skill inference, they can also introduce noise and redundancy. These numbers further contextualize also the redundancy metrics discussed earlier.

4.2.2 Style-Based Evaluation

We complemented our lexical and semantic evaluation by computing **perplexity** as a style-based analysis. We fine-tuned a GPT-2 model on the 20 real CVs and used it to evaluate the perplexity of both real and generated documents.The results show that synthetic CVs have a slightly lower mean perplexity (34.78) compared to real CVs (37.19), and significantly lower variance (6.41 vs. 16.14), indicating more consistent language usage. This suggests that while real CVs are more stylistically diverse—possibly due to personal idiosyncrasies, formatting, or varied writing styles—the LLM-generated CVs effectively learn and reproduce a standardized resume style.

These findings reinforce the conclusion that LLMs can generate not only semantically relevant but also stylistically credible synthetic data.

4.2.3 Prompt-Induced Artifactuality Results

To assess the semantic focus of each dataset, we analyzed the frequency distribution of soft skill expressions across both the Baseline and Real CV datasets.

As shown in Figure 4, the Baseline dataset exhibits a relatively uniform distribution, with many soft skills mentioned infrequently but evenly. This pattern reflects its design, where each generated sample was explicitly conditioned on a distinct soft skill concept. In contrast, the Real CV dataset displays a highly skewed distribution : a small subset of soft skills accounts for the majority of mentions, while most others are either underrepresented or absent. This concentration suggests that real-world CVs tend to emphasize a narrow range of high-salience competencies.

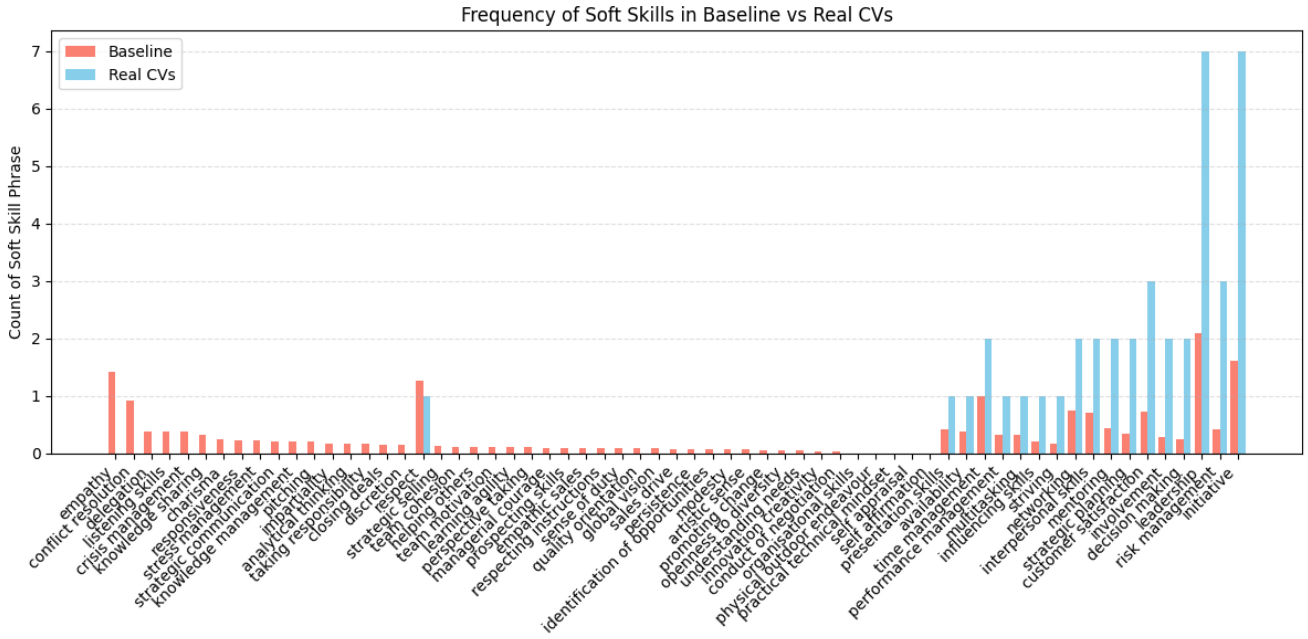

FIGURE 4 – *Soft Skills frequency for Baseline vs Real CVs*

5 Limitations and Future Perspectives

While our study provides valuable insights into LLM-generated data for soft skill classification, several limitations remain. First, the behavior of LLMs is model-dependent, and the results obtained with one model, such as GPT-3.5, may differ significantly from other models. This study may be expanded in the future to compare various LLMs. Second, we focused on evaluating the data itself but did not systematically assess or compare the prompts, which are known to strongly influence generation quality. Third, domain drifts are a valid concern. LLMs, primarily trained on broad internet corpora, might introduce hallucinated elements not representative of actual CVs. Future work should explore domain-adaptive fine-tuning, retrieval-augmented generation, and hallucination detection to mitigate such risks. Integrating hallucination detection metrics would allow us to quantify factual inaccuracies in generated CVs, and, more importantly, detect internal contradictions. Fourth, while we computed a wide range of textual measures, which enable relative comparison between data generation strategies, we did not benchmark these metrics against existing datasets. This absence of external reference points makes it difficult to contextualize our results in relation to real-world or human-authored datasets. Although no publicly available benchmark datasets for soft skill CV classification currently exist to our knowledge, this remains an important direction for future research.

Moreover, while our evaluation focused on textual properties such as diversity and semantic similarity,

it did not directly assess the datasets' effectiveness in downstream classification tasks. A definitive assessment of dataset utility requires training classification models exclusively on synthetic data and evaluating their generalization performance on real-world CVs. The best dataset ultimately depends on the model used, highlighting a trade-off between content quality and task-specific performance.

Looking ahead, future work should explore more sophisticated prompt engineering and integrate evaluation metrics into the generation process. Incorporating human-in-the-loop validation, where synthetic outputs are lightly reviewed and corrected, could significantly enhance data quality. We also plan to refine our generation strategy by focusing on key CV sections that better capture soft skills, balancing generation cost with quality. Finally, mixing real and synthetic data, and testing across models, will be crucial for building robust and generalizable systems.

Références

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

CHEN H., WAHEED A., LI X., WANG Y., WANG J., RAJ B. & ABDIN M. I. (2024). On the diversity of synthetic data and its impact on training large language models. *arXiv preprint arXiv :2410.15226*.

CHUNG J. J. Y., KAMAR E. & AMERSHI S. (2023). Increasing diversity while maintaining accuracy : Text data generation with large language models and human interventions. *arXiv preprint arXiv :2306.04140*.

CLAVIÉ B. & SOULIÉ G. (2023). Large language models as batteries-included zero-shot esco skills matchers. *arXiv preprint arXiv :2307.03539*.

FARERI S., MELLUSO N., CHIARELLO F. & FANTONI G. (2021). Skillner : Mining and mapping soft skills from any text. *Expert Systems with Applications*, **184**, 115544.

FILICE S., CHOI J. I., CASTELLUCCI G., AGICHTEIN E. & ROKHLENKO O. (2023). Evaluation metrics of language generation models for synthetic traffic generation tasks. *arXiv preprint arXiv :2311.12534*.

GAO J., PI R., LIN Y., XU H., YE J., WU Z., ZHANG W., LIANG X., LI Z. & KONG L. (2022). Self-guided noise-free data generation for efficient zero-shot learning. *arXiv preprint arXiv :2205.12679*.

GAO M., HU X., RUAN J., PU X. & WAN X. (2024). Llm-based nlg evaluation : Current status and challenges. *arXiv preprint arXiv :2402.01383*.

LI Y., BONATTI R., ABDALI S., WAGLE J. & KOISHIDA K. (2024a). Data generation using large language models for text classification : An empirical case study. *arXiv preprint arXiv :2407.12813*.

LI Z., XU X., SHEN T., XU C., GU J.-C., LAI Y., TAO C. & MA S. (2024b). Leveraging large language models for nlg evaluation : Advances and challenges. *arXiv preprint arXiv :2401.07103*.

LIN C.-Y. (2004). Rouge : A package for automatic evaluation of summaries. In *Text summarization branches out*, p. 74–81.

LONG L., WANG R., XIAO R., ZHAO J., DING X., CHEN G. & WANG H. (2024). On llms-driven synthetic data generation, curation, and evaluation : A survey. *arXiv preprint arXiv :2406.15126*.

MENG Y., HUANG J., ZHANG Y. & HAN J. (2022). Generating training data with language models : Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, **35**, 462–477.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, p. 311–318.

SHAIB C., BARROW J., SUN J., SIU A. F., WALLACE B. C. & NENKOVA A. (2024a). Standardizing the measurement of text diversity : A tool and a comparative analysis of scores. *arXiv preprint arXiv :2403.00553*.

SHAIB C., ELAZAR Y., LI J. J. & WALLACE B. C. (2024b). Detection and measurement of syntactic templates in generated text. *arXiv preprint arXiv :2407.00211*.

UL HAQ M. U., FRAZZETTO P., SPERDUTI A. & DA SAN MARTINO G. (2024). Improving soft skill extraction via data augmentation and embedding manipulation. In *Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing*, p. 987–996.

WINGS I., NANDA R. & ADEBAYO K. J. (2021). A context-aware approach for extracting hard and soft skills. *Procedia Computer Science*, **193**, 163–172.

YE J., GAO J., FENG J., WU Z., YU T. & KONG L. (2022a). Progen : Progressive zero-shot dataset generation via in-context feedback. *arXiv preprint arXiv :2210.12329*.

YE J., GAO J., LI Q., XU H., FENG J., WU Z., YU T. & KONG L. (2022b). Zerogen : Efficient zero-shot learning via dataset generation. *arXiv preprint arXiv :2202.07922*.

YU Y., ZHUANG Y., ZHANG J., MENG Y., RATNER A. J., KRISHNA R., SHEN J. & ZHANG C. (2023a). Large language model as attributed training data generator : A tale of diversity and bias. *Advances in Neural Information Processing Systems*, **36**, 55734–55784.

YU Y., ZHUANG Y., ZHANG R., MENG Y., SHEN J. & ZHANG C. (2023b). Regen : Zero-shot text classification via training data generation with progressive dense retrieval. *arXiv preprint arXiv :2305.10703*.

YUAN Y., LIU Y. & CHENG L. (2024). A multi-faceted evaluation framework for assessing synthetic data generated by large language models. *arXiv preprint arXiv :2404.14445*.

ZHANG M., JENSEN K. N., SONNIKS S. D. & PLANK B. (2022). Skillspan : Hard and soft skill extraction from english job postings. *arXiv preprint arXiv :2204.12811*.

ZHENG D., LIU D., LAPATA M. & PAN J. Z. (2024). Trustscore : reference-free evaluation of llm response trustworthiness. *arXiv preprint arXiv :2402.12545*.

A Explanation of the Evaluation Metrics

A.1 Reference-free metrics

A.1.1 Diversity

We evaluate lexical and semantic diversity using the following metrics :

— **N-gram Diversity Score** : Captures the ratio of unique n-grams to total n-grams in the dataset.

$$\text{Diversity}_n = \frac{|\text{Unique } n\text{-grams in } D|}{\text{Total } n\text{-grams in } D} \tag{1}$$

We compute this score for n=4 to assess phrasal diversity beyond individual words.

Semantic similarity :

— **BERTScore** : Computes the semantic similarity between generated texts using contextualized embeddings from a pretrained BERT model. Given two sentences A and B, BERTScore aligns tokens based on the cosine similarity of embeddings and returns a precision, recall, and F1 score that reflects how semantically similar A is to B. We use the F1 score for analysis.

$$\text{BERTScore}_{F1} = 2 \cdot \frac{P \cdot R}{P + R} \tag{2}$$

where P and R are the precision and recall based on maximum cosine alignment.

— **Sentence-BERT** : Another method that can be used to assess the semantic diversity of the generated dataset is computing the average pairwise cosine similarity between Sentence-BERT embeddings. While BERTScore provides fine-grained alignment at the token level, this metric complements it by capturing global semantic redundancy across the dataset across the dataset. Formally, given N embedded paragraphs $v_1, v_2, \ldots, v_N$, the homogenization score H is computed as :

$$H = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \cos(v_i, v_j) \tag{3}$$

A.1.2 Redundancy

To detect repetition both within and across samples, we use the following :

— **Compression Ratio** : Measures how compressible the text is using standard algorithms (e.g., gzip).

$$\text{CR} = \frac{\text{Original Size (bytes)}}{\text{Compressed Size (bytes)}} \tag{4}$$

A higher ratio indicates more redundancy and repeated phrasing.

— **POS Compression Ratio** : Applies the same principle to part-of-speech sequences, revealing syntactic repetition even when surface words vary.

$$\text{CR}_{\text{POS}} = \frac{\text{Length}(T)}{\text{CompressedLength}(T)} \tag{5}$$

— **Self-Repetition Score** : Quantifies the internal repetition of long n-grams within each individual sample, flagging looped or recycled phrases.

$$\text{SRS}(t) = \frac{|\text{Repeated } n\text{-grams in } t|}{\text{Total } n\text{-grams in } t} \tag{6}$$

A.2 Reference-based metrics

We used three well-established metrics for reference-based evaluation : **BLEU**, **ROUGE-L**, and **BERTScore**. Each metric captures different aspects of similarity, providing a multifaceted understanding of the correspondence between real and generated CVs.

— **BLEU** : The Bilingual Evaluation Understudy (BLEU) score is a precision-oriented metric commonly used in machine translation. It calculates n-gram overlaps between a generated text and a reference text. While BLEU can underestimate performance for longer, paraphrased text like CVs, it still provides insight into lexical overlap and phrasing consistency.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{7}$$

where p_n is the modified n-gram precision, w_n are weights (usually uniform), and BP is the brevity penalty.

— **ROUGE-L** : ROUGE-L evaluates the longest common subsequence (LCS) between the generated and reference texts. This makes it sensitive to word order and structural similarity, which is especially relevant for CVs where sections such as education, experience, and skills often follow a predictable sequence. ROUGE-L is recall-oriented, emphasizing how much of the real CV is reflected in the generated one.

$$\text{ROUGE-L} = \frac{(1+\beta^2) \cdot R \cdot P}{R + \beta^2 \cdot P} \tag{8}$$

where R and P are LCS-based recall and precision, and β controls the balance (often $\beta = 1.2$ for ROUGE-L).

— **BERTScore** : As mentioned before, BERTScore leverages contextual embeddings from BERT to assess semantic similarity. It can be used both as a reference-free and a reference-based metric. In this case, it computes the similarity between the two datasets' contextualized embeddings. We report the F1 score, which balances precision and recall of semantic overlap.

B Prompt Examples

The prompts displayed in this section are intended to serve as illustrative examples of the overall generation pipeline used in the study. For clarity and brevity, we present only one instance of the prompt architecture—specifically, the one used to generate the *Work Experience* section within the CV Dataset. Each CV section (*About Me, Education, Interests*) has its own tailored prompt templates and configurations, following the same compositional design principles. Therefore, the examples here are not exhaustive, but representative of the prompting methodology employed across all the sections.

User Prompt

`{exp_prompts[exp_idx]}`The CV that the experience section should be related to is for the job `{job}`. The writing style that I want you to follow is defined here : `{writing_styles[writing_idx]}`. Also follow these basic instructions : `{base_exp_instructions[base_inst_idx]}`.

Experience System Prompt

You are an AI specialized in providing CV experience text for a job classification model. Your goal is to generate **highly structured and realistic CV experience text** for different job roles. Each output must follow this format :
- **Experience** : A one or two-sentence piece of experience relating to the specific job role provided, including key achievements and tasks that the candidate might accomplish doing that job.
The structure should look like :

```
    [
    {
        job title : [job achievements]
    },
    {
        job title : [job achievements],
    },
    {
        job title : [job achievements]
    }
]
```

Hard Guidelines :
1. **Ensure realism** : Use actual job responsibilities and tools.
2. **Use JSON format** for structured output : Insure you only respond with the json string, no extra decoration around it, and no newlines in between objects. I should be able to use json.loads() on this response to get the data.
3. **Avoid placeholders** : Use common industry names when necessary.
4. **Randomness** : for each experience, make sure the length and tasks are randomised.
5. **Variation** : Each iteration, make sure that the experience is sligtly different from the previous one provided. They should all be unique and different, maybe getting at the same thing but they should be different from one another.
6. **Diversity** : Avoid using the same verbs or sentence starters across experiences.
7. **Avoid redundancy** : Do not repeat phrasing across experiences.
8. **Syntactic diversity** : Ensure that each bullet has a unique grammatical construction.
9. Always replace 'job title' with the actual job title. Avoid formats like 'job title' : 'Junior Accountant : I ...'
Your task is to return CV experience text for a given job title.

Experience Prompts

1. Generate a compelling experience section for a CV, ensuring clarity, relevance, and impact. Focus on showcasing career growth and key contributions.
2. Write a well-structured experience section for a CV, tailored to industry expectations.
3. Craft a CV experience section that highlights key contributions, making a strong case for the candidate's expertise and impact.
4. Generate a concise yet impactful CV experience section, ensuring each point adds significant value to the overall presentation.
5. Write a well-crafted experience section that reflects the candidate's career trajectory and unique strengths.
6. Develop a CV experience section that emphasizes measurable outcomes and tangible contributions, incorporating relevant data when applicable.
7. Generate a CV experience section that is optimized for recruiters, using clear structure and impactful wording.
8. Write an experience section for a CV, tailored specifically to the demands and expectations of the industry.
9. Craft a CV experience section that strongly conveys the candidate's professional strengths and career achievements.
10. Generate a CV experience section that is keyword-rich and ATS-friendly, ensuring high relevance for job applications.

Base Instructions Prompts

1. Randomly pick 3 jobs to put in the experience section that are related to the job provided. Each with their own key achievements/experience.
2. Randomly pick 2 jobs to put in the experience section that are related to the job provided. Each with their own key achievements/experience.
3. Randomly pick 2 jobs that are related to the job provided, ensuring strong alignment with the target domain. Then, select 1 additional job that is not related to the provided job at all—it should come from a completely different field, showcasing a contrasting background or unrelated experience. Each job has its own key achievements/experience.
4. Randomly pick 3 jobs that are related to the job provided, ensuring strong alignment with the target domain. Then, select 1 additional job that is not related to the provided job at all—it should come from a completely different field, showcasing a contrasting background or unrelated experience. Each job has its own key achievements/experience.

Writing Styles Prompts

1. Formal & Professional Style – Uses corporate language and maintains a professional tone. Use short sentences, resembling a bulleted list format. Example (specifically for accountant, with the job provided adjust the context) : "Prepared and analyzed financial statements in compliance with IFRS and GAAP regulations."
2. Conversational Style – More relaxed, engaging, and friendly, sometimes using first-person (I statements). Example (specifically for accountant, with the job provided adjust the context) : "I've always had a passion for numbers. Whether it's balancing budgets or optimizing financial processes, I enjoy helping businesses stay financially healthy and make data-driven decisions."
3. Achievement-Based Style – Focuses on results, metrics, and performance improvements. Use short, straightforward sentences. Example (specifically for accountant, with the job provided adjust the context) : "Reduced annual financial discrepancies by 30% through the implementation of a new reconciliation system, increasing efficiency and ensuring compliance with internal audit standards."
4. Storytelling Style – Narrative-driven and reflective. Follows a progression, including a beginning, a challenge, or learning phase. Includes a mini-career story that demonstrates growth. Example (specifically for accountant, with the job provided adjust the context) : "I started my career handling day-to-day bookkeeping, but over time, I took on complex financial analysis and tax planning. Today, I help companies streamline financial processes and optimize cash flow to support long-term growth."