Québec Curriculum Vitae (QCV): A Representative Large-Scale Bilingual Synthetic Résumé Dataset

Anonymous ACL submission

Abstract

Publicly available résumé datasets are scarce due to privacy constraints, limited linguistic coverage, and the sensitive nature of employment data, hindering reproducible research in résumé understanding and labour-market modeling. Existing synthetic résumé datasets are typically small-scale, focused on fairness or career progression, or generated through unconstrained prompting, which limits population representativity and linguistic diversity at scale. We introduce Québec Curriculum Vitae (QCV), a large-scale, privacy-free synthetic résumé dataset designed simulating a realistic labour-market population without relying on real personal data. QCV contains 30k workforce profiles grounded in Québec-specific distributions, including 19,843 French and 10,157 English résumés. Our generation methodology enforces macro-level demographic and occupational representativity while promoting linguistic and structural diversity beyond naïve LLM-based generation. Quantitative diversity metrics, embedding-based distributional analysis, and human and LLM-based realism assessments demonstrate that QCV achieves high structural and linguistic diversity while preserving population-level consistency. QCV is publicly released under an open license to support multilingual and workforce-oriented research.

1 Introduction

Resumes constitute a key document type for a wide range of machine learning and natural language processing applications, including information extraction, semantic retrieval, document classification, and decision-support systems in recruitment and workforce analytics. Unlike fully unstructured text, resumes contain a hybrid structure that combines free-form language with implicit section section organization, domain-specific conventions, and temporal information. They encode complex relationships between skills, roles, organizations, and career trajectories, often expressed through highly variable formats. As a result, resume understanding tasks are particularly demanding and needs representative benchmarks for real-world document understanding. The increasing integration of automated tools in hiring and workforce management further amplifies the need for robust, transparent, and well-evaluated resume processing models (FraiJ and László, 2021; Van Esch and Black, 2019; Hosain et al., 2025). Additionally, workforces are filled with biases and disproportions (Riach and Rich, 2002; Bertrand and Mullainathan, 2004; Azmat and Petrongolo, 2014) further justifying the need of a dataset representing this reality to help researchers find methods overcoming these issues.

Despite their importance, large-scale resume datasets remain scarce. This is an issue that has already been known and identified (Alzubaidi et al., 2023; Xu, 2022; Nikolenko et al., 2021). Privacy regulations, ethical constraints, and the difficulty of safely anonymizing personal information severely limit data sharing (Saldivar et al., 2025). Existing publicly available corpora are either small, English-centric, web-scraped with unclear consent, or designed for narrow benchmarking purposes. These challenges are further exacerbated for non-English languages: to date, publicly available resume datasets are overwhelmingly in English. To the best of our knowledge, there is no large-scale publicly released French resume dataset for open research. This lack of linguistic and regional coverage hinders the development and evaluation of resume processing systems in multilingual labour markets (Joshi et al., 2020; Hedderich et al., 2021).

Synthetic resume generation offers a viable alternative (Nadăş et al., 2025; de Gibert et al., 2025; Chen et al., 2025) by enabling the release of realistic, non-identifying documents with controlled structure and content. Such datasets can support core tasks including resume parsing, skill and experience extraction, and job-resume matching, while avoiding the risks associated with real personal

data (Alonso et al., 2025). Privacy-preserving text generation methods based on sensitive data can be used to generate large amounts of synthetic text with reasonable differential privacy guarantees and inevitably affect output quality (Amin et al., 2024). This motivates approaches that avoid the use of confidential data altogether by relying exclusively on aggregate or public sources.

Recent advances in large language models enable synthetic document generation to address limited availability of datasets (Villalobos et al., 2022). However, unconstrained prompt-based generation often fails to preserve occupational distributions, demographic proportions, and structural variability when scaled, leading to distributional collapse or unrealistic population statistics (Casco-Rodriguez, 2024). This highlights the need for generation frameworks that enforce macro-level representativity while maintaining linguistic diversity.

In this work, we focus on the generation of a realistic, synthetic dataset of resumes representative of the Québec labour market. We present *Québec Curriculum Vitae* (QCV), a fully synthetic, bilingual resume dataset grounded in publicly available labour-market statistics. To ensure occupational realism and consistency, resumes are aligned with the Canadian National Occupational Classification (NOC), a standardized taxonomy that defines occupations and skill requirements across the Canadian workforce. By combining population-level grounding with controlled diversity enforcement, QCV simulates a regional labour market without relying on any personal data. To the best of our knowledge, QCV is the first publicly available large-scale dataset of synthetic French and bilingual resumes designed to reflect a real regional labour market while remaining entirely privacy-free.

2 Related Work

Research on automated resume processing has largely focused on tasks such as resume parsing (Sajid et al., 2022), skill extraction (Konstantinidis et al., 2022), resume ranking (Dhobale et al., 2025), and job–resume matching (Guo et al., 2016). Progress in these areas depends critically on the availability of datasets. However, resumes contain highly sensitive personal information and identifiable details, which significantly restricts public data release. As a result, most resume datasets used in research remain either private, small-scale, or web-scraped under unclear consent conditions limiting their reuse and extensibility. Recent efforts include RésuméAtlas (Heakl et al., 2024), which curated 13,389 resumes across 43 categories scraped from public sources. While larger than earlier datasets, such approaches still rely on real personal documents and require filtering and anonymization.

To address data scarcity, several works explore synthetic resume generation. Skondras et al., 2023 generate synthetic resumes using large language models to augment training data for resume classification, demonstrating improved performances. Beyond data generation, recent work leverages LLMs to structure resumes and job postings for semantic similarity–based matching in zero-shot settings, demonstrating the feasibility of such approaches without task-specific training (Skondras et al., 2025). OpenResume (Yamashita et al., 2024) provides the first openly accessible, anonymized resume collection with ESCO-standardized titles and differential-privacy-protected temporal features and includes a 3k synthetic subset generated via graph construction, k-anonymity, and random pruning. Bruera et al., 2022 proposes a privacy-preserving framework for synthetic CV generation that combines differentially private Bayesian networks with prompt-based text generation, primarily targeting anonymization and classification utility. Saldivar et al., 2025 introduce a controlled, rule-based pipeline that recombines fragments of donated resumes to construct a synthetic dataset for fairness benchmarking in algorithmic hiring. Their approach emphasizes demographic control and distributional fidelity and is designed primarily for fairness auditing. Afzal et al., 2025 generate synthetic resumes from a small set of anchor CVs and applying a Retrieval-Augmented Generation (RAG) framework for candidate retrieval and summarization. Similarly, Patra et al., 2024 propose a RAG-based system for tailoring resumes to job descriptions using vector retrieval and generative refinement.

Beyond resume-specific datasets, recent work on synthetic dataset construction emphasizes the importance of producing high-quality resources that can serve as reliable standalone benchmarks rather than mere augmentation corpora. For instance, (Gabburo et al., 2024) introduce multilingual answer sentence selection datasets derived from supervised translation of English resources, demonstrating that automatically constructed datasets must be carefully validated to ensure cross-lingual consistency and downstream robustness. This mo-

tivates the use of diversity-aware metrics operating in the embedding space rather than employing surface statistics alone (Zhu et al., 2025). These principles have been shown to be particularly important when synthetic data is used as a standalone benchmark rather than purely as an augmentation technique, aligning with recent trends in other synthetic dataset releases across NLP and multilingual settings (Thakur et al., 2024).

3 Dataset Construction

3.1 Québec Labour-Market Statistics Database

To ground resume generation in regionally plausible demographic and labour-market patterns while avoiding any use of individual and personal data, we constructed a Québec labour-market statistics database from official aggregate *Census metropolitan area* (CMA) and *Census agglomeration* (CA) tables. The created database is stored as a versioned JSON file and is used downstream to sample structured synthetic profiles (described in the following section) that will guide résumé synthesis.

Starting from publicly available official tabular government statistics covering multiple geographic levels, we retain only rows corresponding to Québec CMA/CA regions. All subsequent processing is performed independently for each CMA/CA, yielding region-specific marginal distributions that reflect local labour-market structure.

For each Québec CMA/CA, we compute normalized marginal distributions over a set of key sociodemographic and labour-market variables. Each variable is defined by a predefined set of statistical categories specified in configuration, from which rates or count-based proportions are extracted and renormalized to form valid probability distributions. The resulting CMA/CA-level statistics cover:

- **Geographic location**: region, city.
- **Demographics**: age groups, gender, work status, legal status, period of immigration and country of birth.
- **Language background**: knowledge of official languages and mother tongue.
- **Occupation/education**: NOC TEER 1 and level 5 categories, augmented with specific job titles, highest attained education level and field of study.

These distributions are later combined to sample coherent synthetic profiles that reflect regional variability and population diversity across Québec.

Privacy and reproducibility. The database contains only aggregate statistics and derived probability distributions and does not include any individual-level records. All category definitions, mappings, and extraction rules are controlled via configuration files, ensuring full reproducibility and enabling straightforward updates when upstream statistics change.

3.2 Synthetic Profiles Dataset

Synthetic data should aim to mimic properties of real data (Johansson et al., 2024). A central contribution of this work is the construction of a dataset of structured synthetic profiles designed to simulate a realistic and diverse Québec workingforce population without relying on any personal data. Rather than generating resumes directly from prompts or aggregate statistics, we explicitly model individual-level profiles as coherent combinations of socio-demographic, educational, and professional attributes. This representation enables controlled large-scale simulation of a regional labour market without any confidentiality risks and mitigates distributional collapse observed in unconstrained large-scale model generation, where high-probability patterns are overrepresented and tail configurations diminish (Casco-Rodriguez, 2024; Shumailov et al., 2024).

Each synthetic profile corresponds to plausible individual attributes sampled probabilistically from the Québec labour-market statistics database (Section 3.1). Sampling is performed at the individual level while respecting population-level distributions, allowing the dataset to capture both regional specificity and within-region diversity.

Profile generation follows a hierarchical and conditional sampling process where attributes are drawn sequentially under explicit dependency constraints (Figure 1). They encode domain knowledge about real workforce compositions. For example, age constrains admissible education levels, legal status conditions immigration-related attributes, and education fields are restricted based on industry and occupation. This prevents implausible profiles and maintains real distributions. The synthetic profiles attributes span multiple dimensions of sociodemographic and professional information, including region and city, age group, gender, legal status, language background, education level and field of study, employment status, industry sector, and a fine-grained occupation (NOC TEER 5) with a corresponding job title. While most attributes are

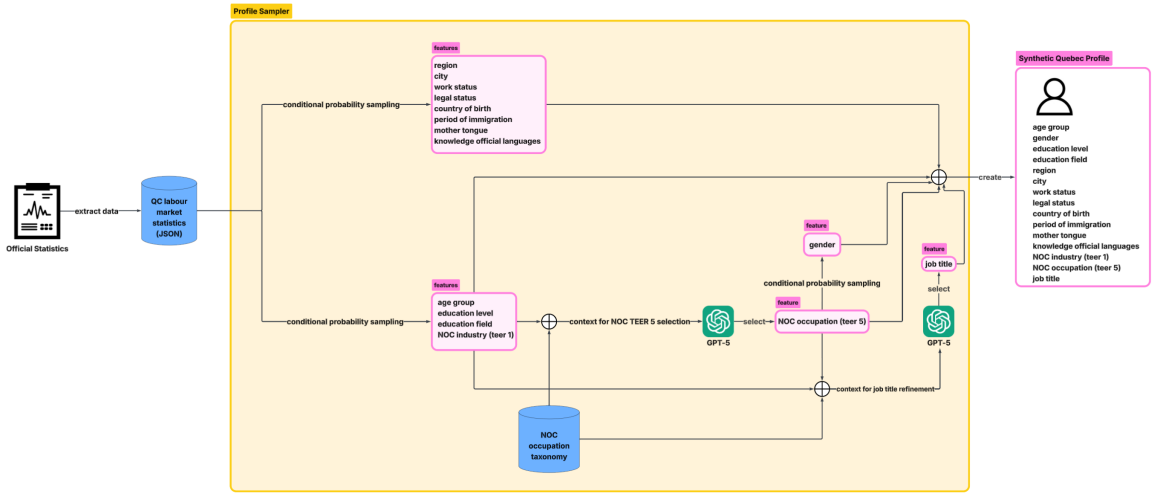


Figure 1: Overview of the synthetic profile generation process. Aggregate Québec labour-market statistics are transformed into structured synthetic profiles through hierarchical, constraint-aware sampling. These profiles are subsequently used as conditioning inputs for resume generation.

sampled directly from empirical or conditional distributions, the final occupation title and specific job title are selected by a LLM (GPT-5) from a constrained set of plausible occupations conditioned on the sampled industry, age, and education context. This step increases semantic specificity without altering the overall population structure.

The resulting dataset is stored as a tabular file with one row per synthetic profile. Each row includes both human-readable attributes and standardized categorical codes. These synthetic profiles form the foundation for resume generation: by grounding the language model on structured profiles rather than unconstrained prompts, we obtain synthetic resumes that exhibit realistic regional patterns, high diversity, and strong privacy guarantees, while remaining representative of the Québec labour market. This method of conditioned generation mitigates hallucination risks where language models may produce confident but factually unsupported outputs while promoting diversity and realism, in contrast to unconstrained large-scale LLM generation (Tonmoy et al., 2024; Reed, 2025).

3.3 Synthetic Resume Generation

Building on these synthetic profiles, a controlled, section-based generation pipeline is used to generate complete and detailed résumés, aiming at producing resumes that are realistic, internally coherent, and representative of Québec labour-market practices, strictly adhering to the attributes specified by each synthetic profile. The generation method relies on an LLM-based (GPT-5), controlled generation strategy that mirrors how real resumes are constructed (Figure 2).

For each synthetic profile, resume generation proceeds section by section. The generation process is conditioned on three elements: (i) a fixed system prompt encoding global formatting rules and Québec-specific constraints, (ii) specific information of the corresponding synthetic profile coupled with a section-specific prompt, and (iii) the content of previously generated sections. This allows the model to maintain consistency across sections while respecting demographic, educational, and occupational constraints imposed by the profile. Resume generation involves satisfying multiple interdependent constraints, such as temporal alignment, occupational consistency, language control. Decomposing such tasks into structured subtasks has been shown to improve reliability and controllability in LLM-based systems (Zhou et al., 2023, 2025) while reducing the risk of hallucination (Madhwal et al., 2026). In parallel, research on controllable text generation shows that explicit conditioning mechanisms improve adherence to high-level constraints compared to unconstrained prompting (Liang et al., 2024).

The resume is generated in a predefined order, starting with the header and education sections,

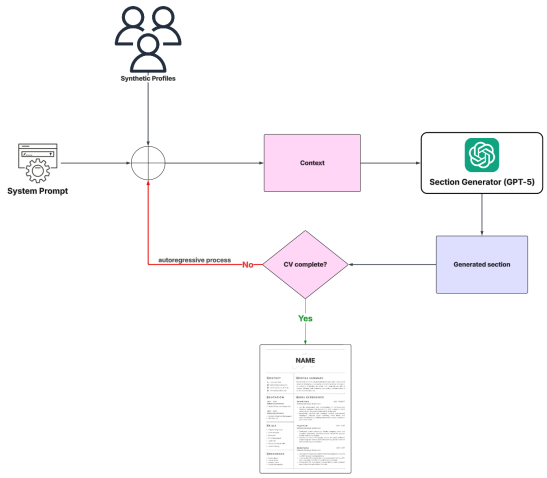


Figure 2: Controlled resume generation pipeline. Each resume is generated section-by-section conditioned on a global system prompt encoding Québec-specific formatting constraints, relevant synthetic profile attributes for the section, and previously generated sections.

followed by an experience outline, individual experience entries and skills. Each generated section is appended to the context and used to condition subsequent generations. This autoregressive process ensures temporal consistency (no overlapping dates), stylistic coherence, and alignment between occupations, responsibilities, and skills. It also allows control over optional sections (certifications and personal projects), which may be omitted depending on profile attributes such as age or education level. The choice of the written language for the given resume is explicitly controlled to reflect Québec's linguistic reality. For each resume, it is determined from the synthetic profile: profiles with French or English as a mother tongue generate resumes in the corresponding language, while bilingual or allophone profiles follow distributions derived from knowledge of official languages. This allows QCV to reflect the bilingual nature of the Québec labour market and language proportions similar to observed population distributions.

Each generated resume is stored in multiple complementary formats. A plain-text version represents a realistic, human-readable resume; a structured JSON version preserves section boundaries for automatic analysis; and a consolidated CSV file stores profile attributes alongside the full resume content.

4 QCV Dataset Overview

Our synthetic profile generation achieves broad occupational coverage, spanning 416 of the 516 occupations in the Canadian NOC and producing 1,987 distinct job titles across 30k profiles. After deduplication, 22,806 unique profiles remain, reflecting substantial combinatorial diversity. There are 19,843 French and 10,157 English résumés. The resulting dataset consists of paired profile–resume instances, enabling analysis at both the structured attribute level and the free-text level. It is organized into complementary components:

- **Synthetic resumes**: provided in three formats: (i) plain text (human-readable CV), (ii) structured JSON with explicit section boundaries (header, education, experience, skills, projects), and (iii) a rendered PDF for human evaluation and other downstream use.
- **Consolidated CSV file**: a single tabular file with one row per resume, containing all profile attributes, the resume language, full text, serialized JSON representation, and automated evaluation scores. This format supports large-scale statistical analysis and benchmarking.

5 Evaluation

We evaluate the synthetic CV generation framework across complementary dimensions. We put in place this multi-faceted evaluation to ensure that the generated resumes are not only useful for downstream applications, but also statistically representative, diverse, and realistic (explanations for the chosen datasets for comparison is in Appendix A).

5.1 Representativity

Representativity is evaluated by comparing the marginal and conditional distributions of key attributes between real and synthetic data. This includes demographic variables (age group, region), education levels, occupational categories, gender distributions per industry and more.

As shown in Figures 5a–6e in Appendix B, the synthetic data closely matches the real Québec labour-market distributions for all major attributes.

5.2 Diversity

We assess the diversity of QCV using complementary semantic and lexical measures. Semantic diversity is evaluated in the embedding space by encoding each resume with a pretrained sentence em-

bedding model and computing the average pairwise distance between each resume and all others. The final score is obtained by averaging these per-sample mean distances, providing a measure of overall semantic dispersion in the latent space. Lexical diversity is measured using n-gram–based statistics, including the number of distinct n-grams, the cumulative probability mass of the top-k most frequent n-grams, and the n-gram entropy. Together, these metrics capture both lexical richness and stylistic variability within the dataset.

Dataset	Distinct n-grams↑	Top-k mass↓	n-gram entropy↑	Semantic dispersion↑
Kaggle	0.72	0.004	0.97	0.573
RésuméAtlas	0.48	0.019	0.95	0.613
QCV	0.53	0.007	0.94	0.539

Table 1: Diversity results

Table1 shows that QCV resumes achieve diversity levels comparable to real datasets. Lexically, QCV has intermediate distinct n-grams and entropy, with low top-k n-gram mass, indicating limited repetition. Semantically, its embedding dispersion is slightly lower but within the real-data range. The two real datasets differ due to their sources: Kaggle is more standardized, yielding higher lexical consistency and lower semantic dispersion, while RésuméAtlas is more heterogeneous, increasing semantic variability. Overall, QCV preserves both lexical richness and semantic diversity, positioning it between the extremes of real datasets.

5.3 Divergence

In synthetic resume generation, divergence quantifies how closely the distribution of synthetic resumes matches that of real resumes. It evaluates whether synthetic data captures the statistical structure and variability of real data. To measure this, we use the Fréchet distance computed over semantic embeddings, assuming that resume representations in the latent space follow multivariate Gaussian distributions. The distance is calculated between the mean and covariance of real and synthetic embeddings as shown in the equation below:

$$\mathrm{FD}(R,S) = \|\mu_R - \mu_S\| + \mathrm{Tr}\Big(\Sigma_R + \Sigma_S - 2\big(\Sigma_R\Sigma_S\big)^{1/2}\Big)$$

Using RésuméAtlas and Kaggle as real reference datasets, our analysis (Table 2) shows that QCV more closely approximates the distribution of real resumes than other synthetic datasets. This lower divergence likely stems from QCV's use of candidate profiles based on real job market statistics.

	RésuméAtlas	Kaggle
Skondras et al., 2023	0.35	0.45
Bruera et al., 2022	0.66	0.78
QCV	**0.26**	**0.31**

Table 2: Divergence results for the baseline and QCV. **Bold** represents the best result.

5.4 Utility

To evaluate utility, we measure the impact of synthetic resumes on a downstream task: predicting a candidate's professional field. Using RésuméAtlas as the real reference, we follow its original splits and re-label both real and synthetic resumes with GPT-5 to ensure consistent categories.

A BERT-based classifier trained only on real data serves as the baseline. We then train multiple classifiers with varying real–synthetic mixtures, progressively replacing real samples with synthetic ones. Since profession distributions differ between RésuméAtlas and QCV, synthetic sample availability varies by category. To account for this imbalance and explore diverse proportions, we train 350 BERT models using different random mixtures.

For each model, we compute per-category accuracy on the held-out test set. Results are grouped into 1% bins of synthetic proportion (e.g., 44% corresponds to mixtures between 44% and 45%) and averaged within each bin. Performance is compared to the real-only baseline to quantify relative degradation or improvement as a function of synthetic data proportion (Figure 3).

5.5 Structural Semantic Analysis

To assess whether QCV preserves meaningful semantic organization, we embed all 30k CVs using the multilingual model BGE-M3 (Chen et al., 2024) and visualize the embedding space using PCA (Figure 4). The projection reveals coherent regions rather than uniform dispersion, indicating non-trivial semantic structure. When colored by occupational variables (Appendix D.1), clear spatial organization emerges, whereas demographic attributes show no distinct separation, suggesting that the embedding space is primarily structured by professional semantics.

We further perform clustering using Spherical k-Means, selecting the number of clusters based on silhouette scores. To quantify alignment be-

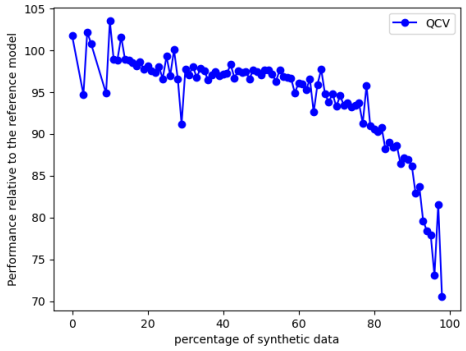


Figure 3: Relative classification performance of BERT models trained on varying mixtures of real and synthetic data, compared to a real-only baseline. Each point denotes the average per-category accuracy within a 1% bin of synthetic data proportion, illustrating how increasing synthetic content affects domain classification accuracy.

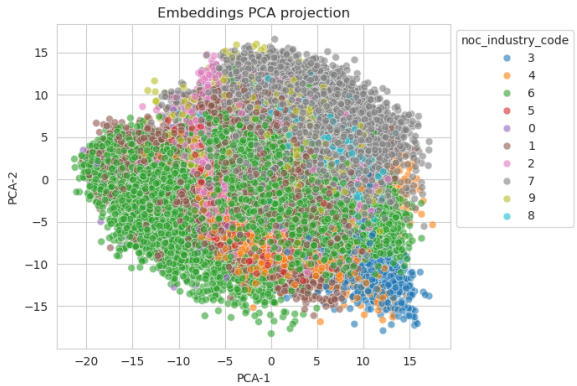


Figure 4: PCA projection of BGE-M3 embeddings for entire QCV, colored by the industry code variable.

tween clusters and profile attributes, we compute Adjusted Mutual Information (AMI) between cluster assignments and variables such as occupation, industry, education field, and demographic factors (Lazarenko and Bonald, 2021). AMI measures agreement with ground-truth labels while correcting for chance, enabling evaluation of how well the embedding space reflects underlying profile structure.

Results in Table 3 show that clustering is primarily aligned with professional attributes. Occupation achieves the highest AMI, followed by industry, while education field shows moderate alignment. In contrast, demographic variables display near-zero AMI and region shows only weak alignment, indicating limited influence of non-professional factors. For combined variables, occupation remains the dominant organizing factor, with education providing only minor refinement.

Variable	Type	AMI
industry	prof.	0.48
occupation	prof.	0.68
education field	prof.	0.41
noc occupation + education field	prof.	0.65
noc industry + education field	prof.	0.56
Average professional (prof.)		**0.56**
period of immigration	non-prof.	0.01
country of birth	non-prof.	0.01
region	non-prof.	0.05
age group	non-prof.	0.02
Average non-professional (non-prof.)		**0.02**

Table 3: Adjusted Mutual Information (AMI) between cluster assignments and candidate profile variables.

5.6 Realism

5.6.1 Human Evaluation

We assess perceived realism as the degree to which a synthetic CV is structurally plausible, internally coherent, and consistent with realistic career trajectories, as judged by human evaluators. 223 unique CVs were evaluated. Each evaluator reviewed 10 CVs randomly sampled from QCV and RésuméAtlas (7 synthetic and 3 real), without knowing their origin. Resumes were rated on a 5-point ordinal scale across 4 criteria: overall realism, coherence, realistic progression, and style & writing quality (see Appendix D.2.1 for full guidelines). These ratings serve as the primary human-grounded assessment of quality.

Table 4 reports the human evaluation results. The inter-human agreement was 80.2% on 61 CVs. In the English setting, QCV achieves high realism across all criteria, closely matching the scores obtained for real RésuméAtlas CVs. In the French setting, QCV similarly attains strong scores, remaining comparable to real French CVs across structural and stylistic dimensions. Across both languages, the small gap between synthetic and real CVs indicates that human evaluators perceive QCV resumes as highly plausible, internally coherent, and consistent with realistic career trajectories.

5.6.2 LLM-as-a-Judge

To scale realism evaluation beyond the human-labeled subset, we employ an LLM-as-a-Judge framework following recent work on automated LLM-based evaluation (Gu et al., 2024) which

	LLM-as-a-Judge				Human Evaluation			
Dataset	Realism	Coherence	Progression	Style	Realism	Coherence	Progression	Style
Kaggle	2.28±0.84	2.51±0.97	2.53±1.01	2.17±0.83	–	–	–	–
RésuméAtlas	2.05±0.89	2.48±1.03	2.57±1.07	2.14±0.88	3.72±1.04	3.80±0.96	3.96±0.98	3.66±1.06
QCV (English)	4.86±0.40	4.54±0.86	4.38±1.05	4.73±0.47	4.26±0.91	4.26±1.02	4.26±0.99	4.25±0.80

(a) English CVs (Human evaluation: 113 unique CVs from QCV; 5 unique CVs from RésuméAtlas)

	LLM-as-a-Judge				Human Evaluation			
Dataset	Realism	Coherence	Progression	Style	Realism	Coherence	Progression	Style
Kaggle (translated)	2.59±0.75	2.65±0.96	2.62±1.03	2.40±0.78	–	–	–	–
Real French CVs	4.33±0.52	4.17±0.75	4.00±0.89	4.17±0.41	4.09±0.62	4.09±0.71	4.14±0.73	3.89±0.83
QCV (French)	4.77±0.50	4.47±0.91	4.30±1.11	4.67±0.50	4.15±0.66	4.13±0.84	4.08±0.75	3.95±0.74

(b) French CVs (Human evaluation: 103 unique CVs from QCV; 5 unique real CVs)

Table 4: Average realism scores and their standard deviation (LLM-as-a-Judge and Human evaluation) across datasets (scale 1-5). The judge agreement score after prompt optimization was 83.1%, used on the entire datasets.

is effective when aligned with well calibrated using human-labeled data (Zheng et al., 2023). We use Gemini 2.5 Flash - another model family as our main model to avoid self-biais. (Panickssery et al., 2024). Using a distinct judge mitigates this risk. The judge scores CVs using the same criteria as human annotators. Human ratings serve as gold references to iteratively refine the prompt, with agreement measured using standard correlation metrics. After optimization, the judge achieves 83.1% agreement with human annotations, exceeding inter-human agreement (80.2%). The calibrated judge is then applied to the full dataset (see Figure 11, Appendix D.2.2).

Table 4 shows that the LLM judge assigns consistently high realism scores to QCV in both English and French. Real-world datasets receive lower automatic scores than QCV and lower than their human evaluations. This discrepancy likely reflects the presence of abbreviated formats, omitted sections, or stylistic variability in real CVs that deviate from the standardized rubric. While such variation is natural in real CVs, the LLM judge—aligned with a structured evaluation rubric—tends to penalize abbreviated formats or missing sections more strongly. As a result, datasets with more standardized structure, such as QCV, receive higher automatic scores. This reflects limitation of rubric-aligned LLM evaluation rather than an inherent superiority of the generated data. While such variation is natural, the LLM judge penalizes structural incompleteness more strictly. This distinction highlights the structural consistency enforced by our profile-conditioned generation.

6 Conclusion

We introduced *Québec Curriculum Vitae*, a large-scale, statistically grounded and multilingual synthetic résumé dataset designed to simulate the Québec labour market without relying on personal data. By combining hierarchical profile sampling from official labour statistics with controlled, section-based LLM generation, QCV produces résumés that are regionally plausible, internally coherent, and diverse across demographic and occupational dimensions. Our evaluation demonstrates strong representativity with respect to real-world distributions, as well as high perceived realism validated through both human assessment and a calibrated LLM-as-a-Judge framework. The results show that structured, profile-conditioned generation enables scalable résumé synthesis while preserving privacy and maintaining statistical fidelity. We view QCV as both a dataset and a methodological contribution: it provides a reproducible framework for workforce-aligned synthetic profile construction and controllable résumé generation. We hope this publicly available dataset will support future research in the broad range of potential downstream tasks such as résumé understanding, labour-market modeling, fairness analysis, and privacy-preserving data generation.

Limitations

Despite its scale and statistical grounding, QCV has several limitations. Representativity is enforced pri-

marily at the level of marginal and selected conditional distributions; higher-order dependencies and rare real-world correlations may not be fully captured. Although profile-conditioned, section-based generation reduces inconsistencies and hallucinations and forces diversity, the textual content is still produced by a large language model and may exhibit stylistic similarities or subtle homogenization effects at scale.

Moreover, the dataset reflects a static snapshot of Québec labour-market statistics and does not model temporal shifts in employment or migration patterns. Although the LLM-as-a-Judge shows strong agreement with human evaluation, automated evaluation remains dependent on prompt calibration and criteria design. In particular, judges aligned with structured criteria may favor standardized CV formats over the stylistic variability commonly found in real resumes.

However, our dataset shows that combining statistical conditioning with structured LLM generation can produce a large-scale, realistic, and privacy-preserving résumé dataset. We view this work as a step toward controllable, representative synthetic workforce simulation rather than a definitive model of real labour-market complexity.

References

Anum Afzal, Ishwor Subedi, and Florian Matthes. 2025. Candidate profile summarization: A rag approach with synthetic data generation for tech jobs. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pages 22–31.

Ruben Alonso, Danilo Dessi, Antonello Meloni, and Diego Reforgiato Recupero. 2025. A novel approach for job matching and skill recommendation using transformers and the o* net database. *Big Data Research*, 39:100509.

Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, Jose Santamaría, Ahmed Shihab Albahri, Bashar Sami Nayyef Al-Dabbagh, Mohammed A Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali H Al-Timemy, and 1 others. 2023. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, 10(1):46.

Kareem Amin, Alex Bie, Weiwei Kong, Alexey Kurakin, Natalia Ponomareva, Umar Syed, Andreas Terzis, and Sergei Vassilvitskii. 2024. Private prediction for large-scale synthetic text generation. *arXiv preprint arXiv:2407.12108*.

Ghazala Azmat and Barbara Petrongolo. 2014. Gender and the labor market: What have we learned from field and lab experiments? *Labour economics*, 30:32–40.

Marianne Bertrand and Sendhil Mullainathan. 2004. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013.

Andrea Bruera, Francesco Alda, and Francesco Di Cerbo. 2022. Generating realistic synthetic curricula vitae for machine learning applications under differential privacy. In *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*, pages 53–63.

Josue Casco-Rodriguez. 2024. Self-consuming generative models go mad. Master's thesis, Rice University.

Haonan Chen, Liang Wang, Nan Yang, Yutao Zhu, Ziliang Zhao, Furu Wei, and Zhicheng Dou. 2025. mme5: Improving multimodal multilingual embeddings via high-quality synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8254–8275.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Ona de Gibert, Joseph Attieh, Teemu Vahtola, Mikko Aulamo, Zihao Li, Raúl Vázquez, Tiancheng Hu, and Jörg Tiedemann. 2025. Scaling low-resource mt via synthetic data generation with llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27662–27680.

Prasad Dhobale, Vinayak Bhoir, Abhijeet Vyavhare, Pratik Yelkar, and Pramod A Dharmadhikari. 2025. Resumatcher: An intelligent resume ranking system. In *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, pages 1778–1783. IEEE.

JihaD FraiJ and Várallyai László. 2021. A literature review: artificial intelligence impact on the recruitment process. *International Journal of Engineering and Management Sciences*, 6(1):108–119.

Matteo Gabburo, Stefano Campese, Federico Agostini, and Alessandro Moschitti. 2024. Datasets for multilingual answer sentence selection. *arXiv preprint arXiv:2406.10172*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Shiqiang Guo, Folami Alamudun, and Tracy Hammond. 2016. Résumatcher: A personalized résumé-job matching system. *Expert Systems with Applications*, 60:169–182.

Ahmed Heakl, Youssef Mohamed, Noran Mohamed, Aly Elsharkawy, and Ahmed Zaky. 2024. Resuméatlas: Revisiting resume classification with large-scale datasets and large language models. *Procedia Computer Science*, 244:158–165.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.

Md Sajjad Hosain, Mohammad Bin Amin, Gouranga Chandra Debnath, and Md Atikur Rahaman. 2025. The use of artificial intelligence (ai) in the hiring process: Job applicants' perceptions of procedural justice. *Computers in Human Behavior Reports*, 19:100713.

Pica Johansson, Jonathan Bright, Shyam Krishna, Claudia Fischer, and David Leslie. 2024. Exploring responsible applications of synthetic data to advance online safety research and development. *arXiv preprint arXiv:2402.04910*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6282–6293.

Ioannis Konstantinidis, Manolis Maragoudakis, Ioannis Magnisalis, Christos Berberidis, and Vassilios Peristeras. 2022. Knowledge-driven unsupervised skills extraction for graph-based talent matching. In *Proceedings of the 12th Hellenic conference on artificial intelligence*, pages 1–7.

Denys Lazarenko and Thomas Bonald. 2021. Pairwise adjusted mutual information. *arXiv preprint arXiv:2103.12641*.

Xun Liang, Hanyu Wang, Yezhaohui Wang, Shichao Song, Jiawei Yang, Simin Niu, Jie Hu, Dan Liu, Shunyu Yao, Feiyu Xiong, and 1 others. 2024. Controllable text generation for large language models: A survey. *arXiv preprint arXiv:2408.12599*.

Dhruv Madhwal, Lyuxin David Zhang, Dan Roth, Tomer Wolfson, and Vivek Gupta. 2026. Decomposed prompting does not fix knowledge gaps, but helps models say "i don't know". *Preprint*, arXiv:2602.04853.

Mihai Nadăş, Laura Dioşan, and Andreea Tomescu. 2025. Synthetic data generation using large language models: Advances in text and code. *IEEE Access*, 13:134615–134633.

Sergey I Nikolenko and 1 others. 2021. *Synthetic data for deep learning*, volume 174. Springer.

Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.

Komal Prakashchandra Patra, Manoj Diwakar, and Chandrakala Arya. 2024. Tailored resume generation using rag with llm as per job specifications. In *2024 Eighth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 848–853. IEEE.

Scott M Reed. 2025. Augmented and programmatically optimized llm prompts reduce chemical hallucinations. *Journal of Chemical Information and Modeling*, 65(9):4274–4280.

Peter A Riach and Judith Rich. 2002. Field experiments of discrimination in the market place. *The economic journal*, 112(483):F480–F518.

Hira Sajid, Javeria Kanwal, Saeed Ur Rehman Bhatti, Saad Ali Qureshi, Amna Basharat, Shujaat Hussain, and Kifayat Ullah Khan. 2022. Resume parsing framework for e-recruitment. In *2022 16th international conference on ubiquitous information management and communication (IMCOM)*, pages 1–8. IEEE.

Jorge Saldivar, Anna Gatzioura, and Carlos Castillo. 2025. Synthetic cvs to build and test fairness-aware hiring tools. *arXiv preprint arXiv:2508.21179*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Panagiotis Skondras, Panagiotis Zervas, and Giannis Tzimas. 2023. Generating synthetic resume data with large language models for enhanced job description classification. *Future Internet*, 15(11):363.

Panagiotis Skondras, Panagiotis Zervas, and Giannis Tzimas. 2025. Zero-shot resume–job matching with llms via structured prompting and semantic embeddings. *Electronics*, 14(24):4960.

Nandan Thakur, Jianmo Ni, Gustavo Hernández Ábrego, John Wieting, Jimmy Lin, and Daniel Cer. 2024. Leveraging llms for synthesizing training data across many languages in multilingual dense retrieval. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7699–7724.

SMTI Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

Patrick Van Esch and J Stewart Black. 2019. Factors that influence new generation candidates to engage with and complete digital, ai-enabled recruiting. *Business horizons*, 62(6):729–739.

Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbhahn, and Anson Ho. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*, 1(1).

Tammy Xu. 2022. We could run out of data to train ai language programs. *MIT Technology Review*.

Michiharu Yamashita, Thanh Tran, and Dongwon Lee. 2024. OpenResume: Advancing Career Trajectory Modeling with Anonymized and Synthetic Resume Datasets. In *2024 IEEE International Conference on Big Data (BigData)*. IEEE.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2023. Least-to-most prompting enables complex reasoning in large language models. *Preprint*, arXiv:2205.10625.

Tianle Zhou, Jiakai Xu, Guanhong Liu, Jiaxiang Liu, Haonan Wang, and Eugene Wu. 2025. An approach for systematic decomposition of complex llm tasks. *arXiv preprint arXiv:2510.07772*.

Yuchang Zhu, Huizhe Zhang, Bingzhe Wu, Jintang Li, Zibin Zheng, Peilin Zhao, Liang Chen, and Yatao Bian. 2025. Measuring diversity in synthetic datasets. *Preprint*, arXiv:2502.08512.

A Datasets and Baselines

We provide more details about the datasets and data generation methods referenced in the related work.

A.1 Real Datasets Used in Experiments

RésuméAtlas (Heakl et al., 2024) RésuméAtlas is a curated dataset containing 13,389 resumes across 43 occupational categories collected from publicly available sources. In our experiments, RésuméAtlas is used as a real-resume reference dataset for evaluating linguistic diversity and structural properties of resumes. Although it varies in completeness, they provide additional real-world resume samples useful for diversity comparisons.

Kaggle Resume Dataset We also evaluate against publicly available CVs obtained from Kaggle. This dataset contains web-scraped CVs in raw text format.

An anonymization pipeline combined curated regular expressions with a local LLM (Qwen3) to detect and remove both structured and residual personal information.

A.2 Synthetic Datasets Used in Experiments

LLM-Based Synthetic Resume Generation (Skondras et al., 2023) Skondras et al. propose generating synthetic resumes using large language models to augment resume classification datasets. We implement a similar prompt-based generation setup as a baseline to evaluate the diversity and realism of unconstrained LLM-generated resumes.

Privacy-Preserving CV Generation (Bruera et al., 2022) Bruera et al. introduce a synthetic CV generation framework combining differentially private Bayesian networks with prompt-based text generation. We include this approach as a baseline to compare our generation method with a privacy-preserving synthetic data pipeline designed to maintain statistical properties of resume datasets.

A.3 Datasets Not Used

OpenResume (Yamashita et al., 2024) provides an anonymized resume dataset with ESCO-standardized job titles and differentially private temporal employment features. But, the dataset contains only job titles and timelines, not full resume text or experience descriptions. Since our evaluation relies on textual diversity and semantic metrics computed from resume content and OpenResume does not provide sufficient textual information for meaningful comparison, it wasn't used.

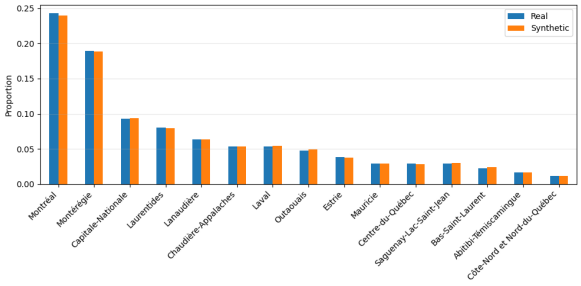
Fragment Recombination: Saldivar et al., 2025 generate synthetic resumes by recombining fragments from donated resumes for fairness benchmarking. The resulting resumes are very short, typically containing very brief education, experience, and skill entries. Since our evaluation relies on richer textual content, this dataset was not used.

RAG-Based Resume Generation (Afzal et al., 2025; Patra et al., 2024) These two papers propose RAG-based systems for CV generation and tailoring using anchor CVs and vector retrieval. They target candidate retrieval or resume customization rather than large-scale dataset construc-

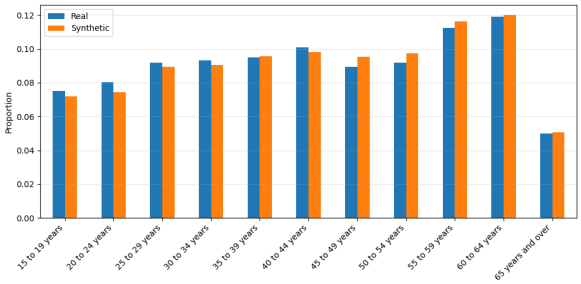
tion. Their datasets and pipelines were not available, preventing reproducible comparison.

B Synthetic Quebec Working Force Profiles Dataset

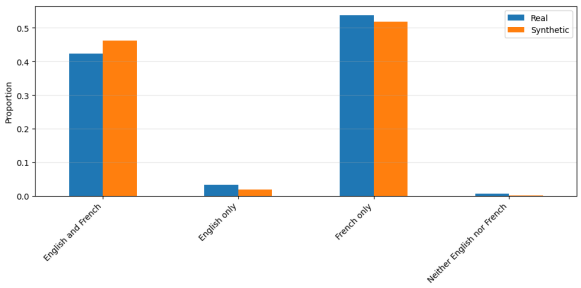
Here are the results of the evaluation of the representativeness of the synthetic profiles dataset. We can see that we have a representative population of synthetic profiles that simulate the Quebec market with distributions across several variables being almost identical to the real distributions. Our synthetic profile generation achieves broad occupational coverage, spanning 416 of the 516 occupations in the Canadian NOC and producing 1,987 distinct job titles across 30k profiles. After deduplication, 22,806 unique profiles remain, showing important combinatorial diversity.



(a) Administrative Region

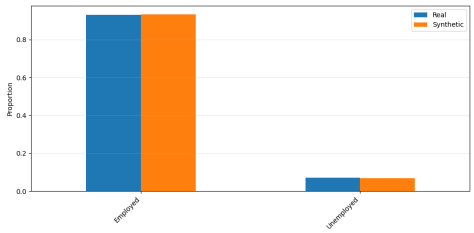


(b) Age Group

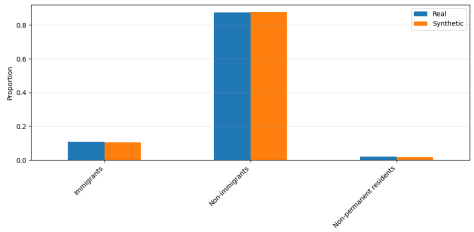


(c) Knowledge of Official Languages

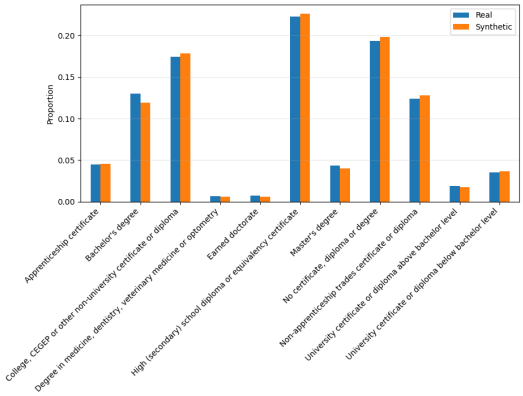
Figure 5: Demographic distribution comparison between synthetic and real Québec profiles.



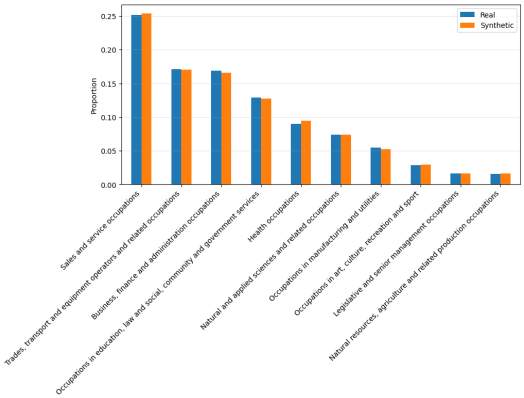
(a) Work Status



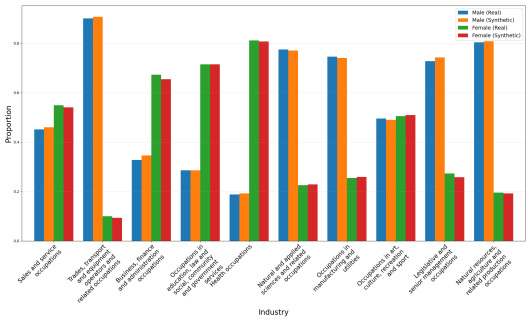
(b) Legal Status



(c) Education Level



(d) Industry (NOC)



(e) Gender per Industry

Figure 6: Professional and structural distribution comparison between synthetic and real Québec profiles.

Regional and age-group proportions exhibit strong alignment, preserving both dominant regions and peak working-age cohorts. Education and industry distributions maintain comparable relative weights across vocational, college, and university credentials as well as across major NOC sectors. Language knowledge, work status, and legal status proportions are nearly identical between real and synthetic data. It is also preserved for gender distributions within industries and replicates the observed sectoral imbalances, indicating that the hierarchical sampling procedure maintains realistic cross-attribute dependencies.

C Québec Curriculum Vitae (QCV) Dataset

C.1 CV PDF Layouts

To increase realism and avoid template homogeneity, PDFs are generated using multiple layout styles that vary in section ordering, typography, header structure, and spacing. The textual content is produced through the sectioned generation pipeline, while visual rendering is handled by a template system that assembles sections into predefined document layouts selected at random (Figure 7). These layouts simulate common résumé formats observed in real-world CV collections while introducing variation in presentation and structure. The objective is not to replicate a specific résumé template, but to introduce realistic visual diversity. This ensures that downstream applications interacting with the dataset, such as document parsing, information extraction, or résumé ranking models, are exposed to heterogeneous document presentations rather than a single standardized format.

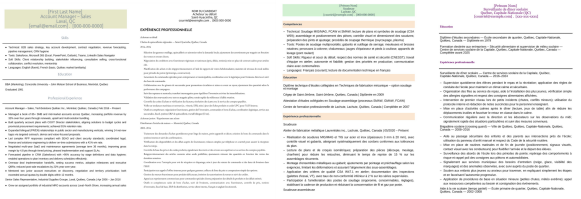


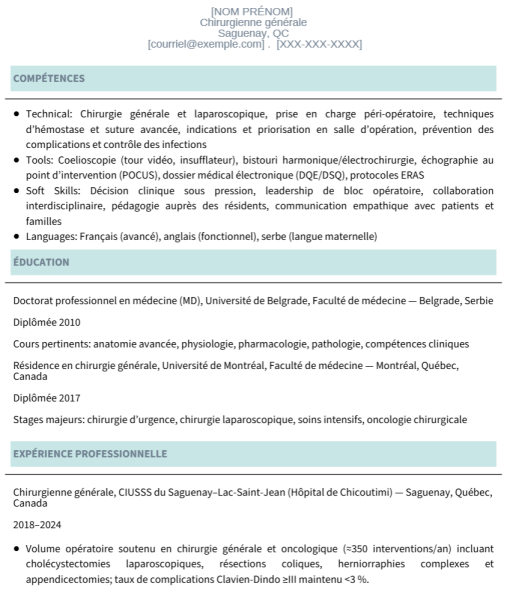
Figure 7: Diversity of first-page résumé layouts in QCV.

C.2 Profile to CV

To illustrate how profiles translate into full résumés, we show an example generated from one synthetic profile to demonstrate how demographic attributes, education, and occupational information are transformed into a coherent CV (Figure 8).

Profile 9698 attributes

- **region:** Saguenay–Lac-Saint-Jean
- **city:** Saguenay
- **age group:** 50–54 years
- **gender:** Female
- **mother tongue:** Serbian
- **knowledge of official languages:** French only
- **education level:** Degree in medicine / dentistry / veterinary medicine / optometry
- **education field:** Health professions and related programs
- **work status:** Employed
- **NOC industry:** Health occupations
- **NOC occupation:** Specialists in surgery
- **NOC title:** General surgeon
- **legal status:** Immigrant
- **country of birth:** Serbia
- **period of immigration:** 2011–2015



[NOM PRÉNOM]
Chirurgienne générale
Saguenay, QC
[courriel@exemple.com] . [XXX-XXX-XXXX]

COMPÉTENCES

- Technical: Chirurgie générale et laparoscopique, prise en charge péri-opératoire, techniques d'hémostase et suture avancée, indications et priorisation en salle d'opération, prévention des complications et contrôle des infections
- Tools: Coelioscopie (tour vidéo, insufflateur), bistouri harmonique/électrochirurgie, échographie au point d'intervention (POCUS), dossier médical électronique (DQE/DSQ), protocoles ERAS
- Soft Skills: Décision clinique sous pression, leadership de bloc opératoire, collaboration interdisciplinaire, pédagogie auprès des résidents, communication empathique avec patients et familles
- Languages: Français (avancé), anglais (fonctionnel), serbe (langue maternelle)

ÉDUCATION

Doctorat professionnel en médecine (MD), Université de Belgrade, Faculté de médecine — Belgrade, Serbie

Diplômée 2010

Cours pertinents: anatomie avancée, physiologie, pharmacologie, pathologie, compétences cliniques

Résidence en chirurgie générale, Université de Montréal, Faculté de médecine — Montréal, Québec, Canada

Diplômée 2017

Stages majeurs: chirurgie d'urgence, chirurgie laparoscopique, soins intensifs, oncologie chirurgicale

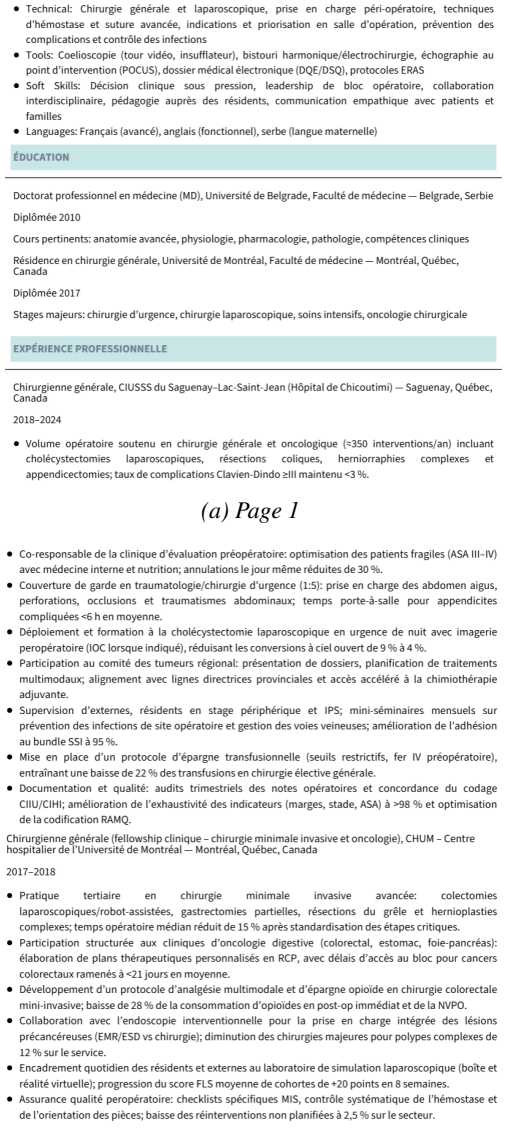
EXPÉRIENCE PROFESSIONNELLE

Chirurgienne générale, CIUSSS du Saguenay–Lac-Saint-Jean (Hôpital de Chicoutimi) — Saguenay, Québec, Canada

2018–2024

- Volume opératoire soutenu en chirurgie générale et oncologique (≈350 interventions/an) incluant cholécystectomies laparoscopiques, résections coliques, herniorraphies complexes et appendicectomies; taux de complications Clavien-Dindo ≥III maintenu <3 %.

(a) Page 1



- Co-responsable de la clinique d'évaluation préopératoire: optimisation des patients fragiles (ASA III–IV) avec médecine interne et nutrition; annulations le jour même réduites de 30 %.
- Couverture de garde en traumatologie/chirurgie d'urgence (1:5): prise en charge des abdomen aigus, perforations, occlusions et traumatismes abdominaux; temps porte-à-salle pour appendicites compliquées <6 h en moyenne.
- Déploiement et formation à la cholécystectomie laparoscopique en urgence de nuit avec imagerie peropératoire (IOC lorsque indiqué), réduisant les conversions à ciel ouvert de 9 % à 4 %.
- Participation au comité des tumeurs régional: présentation de dossiers, planification de traitements multimodaux; alignement avec lignes directrices provinciales et accès accéléré à la chimiothérapie adjuvante.
- Supervision d'externes, résidents en stage périphérique et IPS; mini-séminaires mensuels sur prévention des infections de site opératoire et gestion des voies veineuses; amélioration de l'adhésion au bundle SSI à 95 %.
- Mise en place d'un protocole d'épargne transfusionnelle (seuils restrictifs, fer IV préopératoire), entraînant une baisse de 22 % des transfusions en chirurgie élective générale.
- Documentation et qualité: audits trimestriels des notes opératoires et concordance du codage CIIU/CIHI; amélioration de l'exhaustivité des indicateurs (marges, stade, ASA) à >98 % et optimisation de la codification RAMQ.

Chirurgienne générale (fellowship clinique – chirurgie minimale invasive et oncologie), CHUM – Centre hospitalier de l'Université de Montréal — Montréal, Québec, Canada

2017–2018

- Pratique tertiaire en chirurgie minimale invasive avancée: colectomies laparoscopiques/robot-assistées, gastrectomies partielles, résections du grêle et hernioplasties complexes; temps opératoire médian réduit de 15 % après standardisation des étapes critiques.
- Participation structurée aux cliniques d'oncologie digestive (colorectal, estomac, foie-pancréas): élaboration de plans thérapeutiques personnalisés en RCP, avec délais d'accès au bloc pour cancers colorectaux ramenés à <21 jours en moyenne.
- Développement d'un protocole d'analgésie multimodale et d'épargne opioïde en chirurgie colorectale mini-invasive; baisse de 28 % de la consommation d'opioïdes en post-op immédiat et de la NVPO.
- Collaboration avec l'endoscopie interventionnelle pour la prise en charge intégrée des lésions précancéreuses (EMR/ESD vs chirurgie); diminution des chirurgies majeures pour polypes complexes de 12 % sur le service.
- Encadrement quotidien des résidents et externes au laboratoire de simulation laparoscopique (boîte et réalité virtuelle); progression du score FLS moyenne de cohortes de +20 points en 8 semaines.
- Assurance qualité peropératoire: checklists spécifiques MIS, contrôle systématique de l'hémostase et de l'orientation des pièces; baisse des réinterventions non planifiées à 2,5 % sur le secteur.

(b) Page 2

préparation au bloc; participation active à >600 admissions chirurgicales par année avec documentation conforme au Dossier santé Québec.

- Assistance opératoire et premiers opérateurs supervisés sur procédures courantes (appendicectomies, cholécystectomies, herniorraphies, colectomies segmentaires); courbe d'apprentissage validée par carnet opératoire (>900 cas cumulés, complexité croissante PGY1→PGY5).
- Prise en charge postopératoire en salle d'hospitalisation et soins intensifs: protocoles d'antibiothérapie, prévention des thromboses, gestion de la douleur multimodale; réduction des réadmissions 30 jours pour plaies superficielles sur l'unité de 8 % à 5 % grâce à un suivi téléphonique structuré.
- Garde sur appel (1:4 à 1:6 selon rotations): consultation rapide aux urgences, priorisation des imageries, coordination avec anesthésie et radiologie interventionnelle; délais décision–incision pour appendicites compliquées maintenus <8 h.
- Endoscopie diagnostique de base sous supervision (gastroscopies, coloscopies courtes) et gestion des hémorragies digestives hautes simples avec adrénaline/clip; taux de succès hémostatique primaire >90 % dans le registre de service.
- Participation aux réunions multidisciplinaires (colorectal, sein, HPB): présentation de cas, revue de littérature et alignement sur lignes directrices provinciales; amélioration du délai médian de complétion d'investigations préopératoires de 5 à 3 semaines pour les patients désignés «voies rapides».
- Projet d'amélioration continue: mise à jour de la check-list de préparation de salle et standardisation du set de laparoscopie de garde, entraînant une diminution des changements de matériel en cours de cas de 23 %.
- Encadrement des externes en stage de base: mini-enseignements au lit du malade (drainages, pansements, consentement éclairé) et rétroaction formative; progression des évaluations EPA «soins périopératoires» des externes de niveau 2 à 3 en moyenne sur la rotation.

Médecin généraliste – service de chirurgie (junior doctor), Clinical Centre of Serbia — Belgrade, Serbie

2010–2011

- Admissions et évaluations préopératoires de patients chirurgicaux sous supervision senior; anamnèse ciblée, examens de base et demandes d'imageries/labo complètes, réduisant les retours pour dossiers incomplets au bloc.
- Suivi quotidien en service de chirurgie générale: pansements, gestion IV, prévention thromboembolique et ajustement simple d'analgésie selon protocoles; amélioration de la conformité aux ordonnances standardisées à >90 %.
- Assistance au bloc sur cas de base (appendicectomies, hernies, cholécystectomies ouvertes) avec préparation de salle et instruments; temps d'installation préopératoire raccourci d'environ 10 minutes grâce à une check-list interne.

(c) Page 3

Figure 8: Example synthetic profile and corresponding generated French résumé.

The generated CV illustrates strong regional and temporal plausibility with respect to the attributes of the profile. The professional trajectory is consistent with a 50–54 year-old specialist in surgery, including an appropriate sequence of medical education, residency training, and progressively senior clinical responsibilities. The country of birth and immigration period (2011–2015) are coherently reflected in the narrative, capturing the transition from prior medical training abroad to professional integration within Québec's healthcare system, including French-language proficiency aligned with the profile's official language constraint. Institutional references (e.g., regional hospitals and Québec healthcare structures) are geographically consistent with Saguenay–Lac-Saint-Jean, reinforcing regional realism and grounding the résumé within the Québec labour market rather than a generic international context.

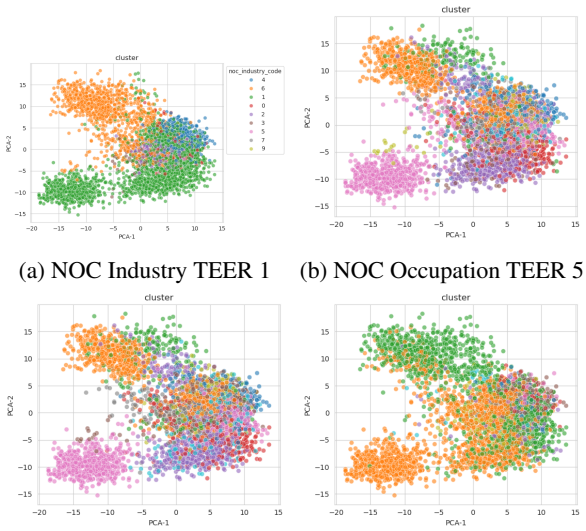
The overall document length and level of detail are also appropriate for a mid-to-late career medical professional. Senior profiles typically include extensive clinical experience, leadership responsibilities, research or quality-improvement initiatives, and detailed technical competencies, which naturally result in multi-page résumés. In this case, the three-page format reflects accumulated professional experience rather than verbosity, aligning with common practices in medical and specialist CVs in Québec. Together, these elements demonstrate that profile-conditioned generation produces résumés that are structurally coherent, regionally grounded, and temporally consistent with realistic career development patterns.

D Evaluation Metrics

D.1 Structural Semantic Analysis

All embeddings are obtained using the BGE-M3 multilingual embedding model without fine-tuning. We compute embeddings for the full set of generated resumes and analyze their structure through a PCA projection of one representative cluster of the embedding space.

When coloring the embedding space by occupation-related variables, structured spatial organization emerges. Distinct occupational families occupy contiguous regions of the embedding space, suggesting that professional semantics are preserved in the synthetic data (Figure 9). This alignment arises despite the embedding model not being trained specifically on occupational taxonomies.



(a) NOC Industry TEER 1 (b) NOC Occupation TEER 5

(c) Job Titles (d) Education Field (CIP)

Figure 9: PCA projection of Cluster 3 colored by profession-related attributes.

Interestingly, the organization appears consistent across different levels of professional granularity. Broad sectors such as NOC industry categories already exhibit partial spatial separation, while finer labels such as specific occupations and job titles further refine the cluster structure. Education fields (CIP codes) show a similar alignment, indicating that the generated resumes encode coherent relationships between training background and profes-

sional roles. These observations suggest that the embedding space captures meaningful professional identities rather than isolated attributes.

In contrast to occupation-related attributes, demographic variables do not exhibit clear geometric separation in embedding space. Categories appear interspersed across clusters, indicating that latent structure is not trivially driven by protected or demographic characteristics (Figure 10).

(a) Age Group (b) Gender

(c) Legal Status (d) Administrative Region

Figure 10: PCA projection of Cluster 3 colored by demographic attributes.

This pattern is consistent across age groups, gender, legal status, and administrative regions. While minor local variations may appear due to correlations between demographics and certain professions, no dominant geometric partition emerges along these variables. This suggests that the embedding organization primarily reflects professional content encoded in the resumes rather than demographic information.

D.2 Realism

D.2.1 Human Evaluation

Evaluation Setup. Evaluators were instructed to assess each synthetic CV independently (in pdf form), assuming the reference year to be 2025. All resumes were evaluated using the same criteria and scoring scale. Annotators were asked evaluate the resumes in the order of the presented criteria. They were also asked to focus on realism and coherence rather than candidate quality.

The evaluation criteria are the following:

1. **Overall Realism:** Evaluates whether the resume could plausibly be submitted to an employer based on structure, formatting, and surface credibility, without inspecting detailed content.

2. **Coherence:** Assesses logical consistency between education, experience, skills, and timelines.

3. **Realistic Progression:** Measures whether job titles and degrees follow a plausible chronological and seniority progression.

4. **Style and Writing Quality:** Evaluates clarity, professionalism, and appropriateness of language relative to the candidate's background and role level.

5. **Scoring and Justification:** Each criterion was rated on a 5-point scale (1 = poor, 5 = excellent). Evaluators provided short textual justifications citing specific evidence.

The evaluation was conducted using the Amazon Mechanical Turk platform. Each evaluator was presented with a set of 10 résumés, comprising 3 real résumés (manually anonymized) and 7 synthetic résumés generated by QCV. For each category (real and synthetic), one résumé was shared across all evaluators to ensure consistency in cross-participant comparison. In total, 21 participants evaluated English résumés and 18 participants evaluated French résumés. Each participant was compensated 20 USD for completing a task that required approximately 60 to 80 minutes.

D.2.2 LLM-as-a-Judge

The judge is prompted to score synthetic CVs using the same criteria as human annotators. Human evaluation scores are used as gold references to iteratively refine the judge prompts, with agreement measured using standard correlation and agreement metrics. After optimizing the prompt, we obtain an agreement rate of 83.1%, indicating strong alignment between automated and human judgments. Once satisfactory agreement is achieved, the calibrated LLM judge is applied to the remainder of the synthetic dataset. Avisualization of the framework can be seen in Figure 11.

Table 5 shows that the LLM-as-a-Judge assigns consistently high realism scores to QCV across both English and French datasets. For English CVs, QCV obtains strong automatic scores (Realism: 4.86, Coherence: 4.54, Progression: 4.38, Style: 4.73). In French, QCV similarly achieves high scores (Realism: 4.77, Coherence: 4.47, Progression: 4.30, Style: 4.67). Real-world datasets

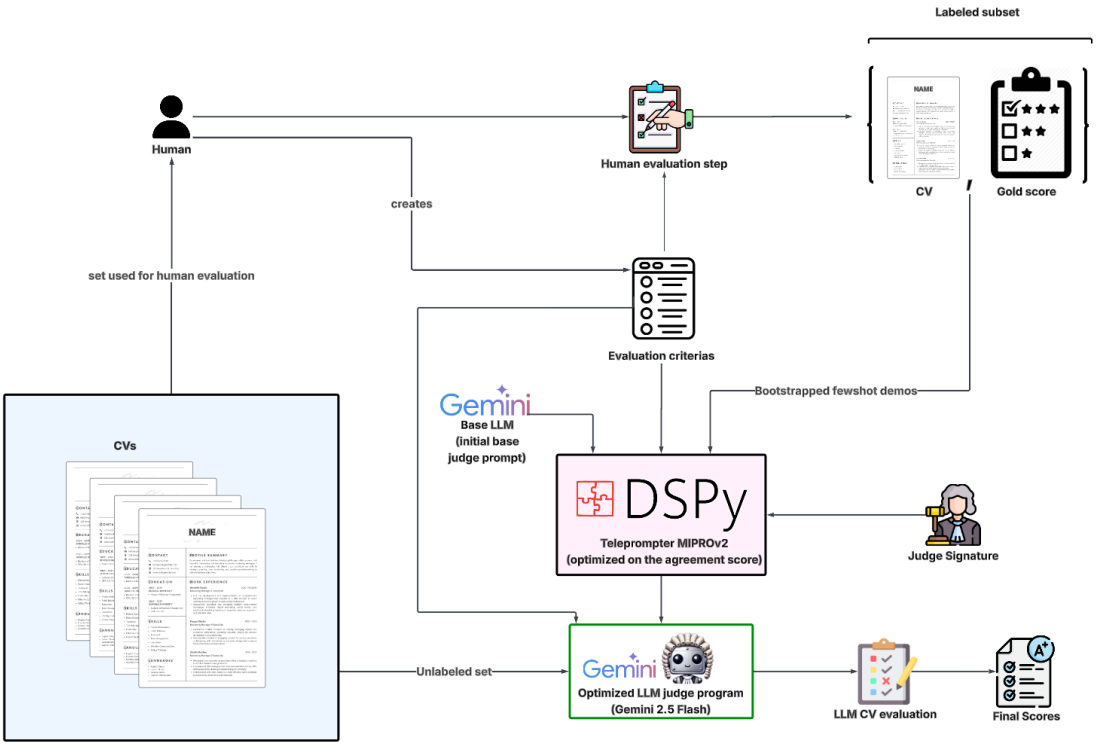


Figure 11: LLM-as-a-Judge evaluation pipeline. A human-labeled subset of CVs is first scored according to predefined evaluation criteria to obtain gold reference scores. These labeled examples are used within DSPy (Teleprompter MIPROv2) to optimize the judge prompt by maximizing agreement with human annotations. The calibrated judge program, implemented using Gemini 2.5 Flash, is then applied to the unlabeled CV set to produce large-scale automated realism scores.

obtain noticeably lower automatic scores than QCV and lower than their corresponding human evaluations. This discrepancy is expected given that many of these real CVs include abbreviated formats, omitted sections, or stylistic variability that deviates from the standardized rubric used for evaluation. While such variation is indeed found in real-world contexts, the LLM judge penalizes structural incompleteness or missing elements more strictly. This highlights an important distinction between natural variation in real documents and structural completeness, which our profile-conditioned generation explicitly enforces.

Beyond scalability, the LLM-as-a-Judge framework enables a consistent evaluation protocol across large synthetic datasets where exhaustive human annotation would be impractical. Once calibrated against human judgments, the automated evaluator provides a reproducible scoring mechanism that can be applied uniformly to thousands of documents. This allows realism assessment to move from small manually annotated samples to dataset-wide analysis while maintaining alignment with human perception. However, automated judging should not be interpreted as a replacement for human evaluation in this context. It functions as a complementary tool that enables large-scale comparative analysis while human annotations remain essential for calibration and qualitative validation. In our setting, the high agreement rate (83.1%) suggests that the judge captures the main criteria used by human annotators, making it suitable for scalable realism evaluation across the full dataset.

	LLM-as-a-Judge				Human Evaluation			
Dataset	Realism	Coherence	Progression	Style	Realism	Coherence	Progression	Style
Skondras et al., 2023	1.43±0.83	1.68±1.13	1.79±1.18	1.45±0.87	–	–	–	–
Bruera et al., 2022	1.12±0.44	1.08±0.27	1.07±0.27	1.11±0.41	–	–	–	–
Kaggle	2.28±0.84	2.51±0.97	2.53±1.01	2.17±0.83	–	–	–	–
RésuméAtlas	2.05±0.89	2.48±1.03	2.57±1.07	2.14±0.88	3.72±1.04	3.80±0.96	3.96±0.98	3.66±1.06
QCV (English)	4.86±0.40	4.54±0.86	4.38±1.05	4.73±0.47	4.26±0.91	4.26±1.02	4.26±0.99	4.25±0.80

(a) English CVs (Human evaluation: 113 unique CVs from QCV; 5 unique CVs from RésuméAtlas)

	LLM-as-a-Judge				Human Evaluation			
Dataset	Realism	Coherence	Progression	Style	Realism	Coherence	Progression	Style
Kaggle (translated)	2.59±0.75	2.65±0.96	2.62±1.03	2.40±0.78	–	–	–	–
Real French CVs	4.33±0.52	4.17±0.75	4.00±0.89	4.17±0.41	4.09±0.62	4.09±0.71	4.14±0.73	3.89±0.83
QCV (French)	4.77±0.50	4.47±0.91	4.30±1.11	4.67±0.50	4.15±0.66	4.13±0.84	4.08±0.75	3.95±0.74

(b) French CVs (Human evaluation: 103 unique CVs from QCV; 5 unique real CVs)

Table 5: Average realism scores and their standard deviation (LLM-as-a-Judge and Human evaluation) across different datasets (scale 1-5). LLM agreement score after prompt optimization: 83.1%. The LLM-as-a-Judge was used on the entire datasets each time.