JobHop v2: A Large-Scale Career Trajectory Dataset from Unstructured Resumes

Iman Johary*, Guillaume Bied, Alexandru C. Mara and Tijl De Bie

AIDA-IDLab, Department of Electronics and Information Systems, Ghent University, Ghent, Belgium

Abstract

Large-scale, richly annotated career trajectory data underpins workforce planning, job recommendation, and labour market analysis, yet publicly available datasets are either small, closed to independent use, or built from pre-standardized occupational codes with LLM-synthesized rather than authentic free text. We present **JobHop v2**, an improved version of the publicly available JobHop dataset, constructed through end-to-end large language model (LLM) extraction from a corpus of ∼440,000 pseudonymized, multilingual resumes provided by VDAB, the Flemish Public Employment Service. The released dataset comprises 355,315 career trajectories annotated with ESCO occupational codes, quarter-level temporal information, and normalized five-level education attainment, broadening both the coverage and the annotation richness of the original release. Relative to v1, JobHop v2 introduces a redesigned extraction pipeline based on reasoning-controlled LLM inference with a retry mechanism (achieving a 100% JSON parse rate), a richer extraction schema, and a revised evaluation protocol scored against three complementary annotation baselines. Evaluated against these baselines, our best extractor comes closest to the inter-annotator agreement reference among all compared models, trailing it by only 1.1–2.7 percentage points. The dataset and code are publicly released to support reproducible career-trajectory research.

Keywords

career trajectories, dataset, ESCO, information extraction, labour market analysis, large language models, job recommendation

1. Introduction

Understanding how careers evolve (which roles people transition into, when, and from what educational background) has broad practical value: it supports evidence-based labour market policy, powers job recommendation systems, and enables career counseling at scale [1, 2]. The limiting factor for computational career analysis is data. Resumes are the richest naturally occurring source of career-trajectory information: unlike job postings, which capture only open roles, or administrative records, which rarely include job details, resumes document the full arc of a working life (job titles, responsibilities, education, and the timing of transitions) in a single document. Yet their unstructured, multilingual, and heterogeneous nature has long prevented large-scale systematic use.

Recent large language models (LLMs) have changed this calculus. By combining open-ended text comprehension with structured reasoning, LLMs can normalize job titles to occupational taxonomies, resolve temporal ambiguities, and extract skill inventories from free-form descriptions at a scale and accuracy that rule-based and sequence-labeling pipelines could not approach [3, 4]. It is now possible to source rich career trajectory data containing temporal and educational signals from unstructured resumes, enabling large resume corpora to be used for career-path analysis and modeling.

Despite this opportunity, the field still lacks a suitable public benchmark. Existing public datasets are either small (Decorte et al. [1] release 2,164 career histories and OpenResume [5] only 301 real resumes), built from pre-standardized occupational codes rather than raw text [2], or derived from proprietary platforms [6, 7, 8] that are not publicly available for independent use. JobHop [4] was, to our knowledge, the first large-scale public dataset to pair ESCO occupation codes with both temporal information and educational attainment, extracted end-to-end from real unstructured resumes. However, its extraction pipeline suffered from two practical limitations: degraded extraction quality on noisy multilingual inputs, and failure to produce consistently machine-parseable output.

We present **JobHop v2**, an improved version of the publicly available JobHop dataset [4]. Starting from ∼440,000 pseudonymized, multilingual resumes provided by VDAB, reasoning-controlled LLM extraction yields 355,315 career trajectories annotated with ESCO taxonomy codes [9], quarter-level temporal information, and five education levels. The revised pipeline improves on the original release by integrating a redesigned extraction schema, a reasoning-effort-controlled inference configuration with a multi-step retry mechanism, a two-step ESCO occupation-code assignment policy, a five-stage cleaning workflow, and a revised extraction-evaluation protocol scored against three complementary annotation baselines.

This paper makes the following contributions:

- **JobHop v2**: a public dataset of 355,315 career trajectories with ESCO occupation codes, quarter-level temporal annotations, and five-level education attainment, an improved release of the original JobHop benchmark [4] with broader annotation scope and higher extraction quality.
- **An extraction pipeline** based on reasoning-controlled LLM inference with a multi-step retry mechanism, achieving a 100% JSON parse rate after retry over ∼400,000 input resumes and, across three annotation baselines, coming closest of all compared models to the inter-annotator agreement reference.
- **A revised evaluation protocol** with graduated partial-credit scoring and three complementary annotation baselines, disentangling genuine extraction errors from annotation noise.

Ethics. All processing was performed on local infrastructure; resume contents were never transmitted to external services. The released dataset incorporates multi-layered privacy protections: VDAB pseudonymization, removal of

RecSys in HR '26: The 6th Workshop on Recommender Systems for Human Resources, in conjunction with the 20th ACM Conference on Recommender Systems, September 28–October 2, 2026, Minneapolis, MN, United States.

*Corresponding author.

iman.johary@ugent.be (I. Johary); guillaume.bied@ugent.be (G. Bied); alexandru.mara@ugent.be (A. C. Mara); tijl.debie@ugent.be (T. D. Bie)

0009-0007-9223-9995 (I. Johary)

location fields, and coarsening of all dates to quarter-level granularity.

2. Related Work

Career trajectory datasets. Computational career analysis depends on the quality and scale of available trajectory data. Proprietary sources such as LinkedIn [8, 10], Randstad [11], Zippia [6, 3], and large in-house corpora such as UniTRep [7] (1.19M trajectories) and TAPJFNN [12] provide extensive longitudinal histories but remain closed, precluding independent replication. Public survey-based datasets (NLSY79/97 [13], CPS [14]) offer nationally representative samples but are limited in scale or temporal resolution. Among recent public datasets, Decorte et al. [1] released 2,164 career histories from anonymized Kaggle resumes mapped to ESCO [9], establishing the first public benchmark for ESCO-standardized career-path prediction. Senger et al. [2] constructed Karrierewege from 568,888 career paths obtained from the German Federal Employment Agency; occupations were already encoded in the Berufenet taxonomy and mapped to ESCO, with free-text titles *synthesized* by an LLM rather than extracted from raw text. OpenResume [5] provides a small complementary collection of 301 real resumes augmented with synthetic data. No prior public dataset simultaneously combines large scale, full end-to-end extraction from unstructured text, rich temporal metadata, education-level annotations, and ESCO standardization.

JobHop [4] addressed this gap as the first large-scale public dataset to pair ESCO occupation codes with both temporal information and educational attainment, derived end-to-end from real unstructured resumes. Constructed from 361,207 pseudonymized VDAB resumes, it contains 1.67M work experiences with quarter-level temporal annotations and a binary tertiary-degree flag. JobHop v2 retains this real-resume foundation while substantially improving extraction quality, annotation breadth, and evaluation rigor.

LLM-based information extraction. In contrast to earlier rule-based and neural sequence-labeling extraction pipelines [15], recent LLM-based approaches have substantially narrowed the gap for unsupervised occupation extraction. LLM4Jobs [16] and SkillGPT [17] combine prompting with vector-similarity search to extract and standardize occupations and skills without task-specific supervision, while GoLLIE [18] shows that guideline-following instruction tuning substantially improves zero-shot structured extraction under explicit annotation schemas. Within career-trajectory pipelines, Decorte et al. [1] used GPT-3.5 to reformat experience sections into a uniform structure, and Senger et al. [2] employed LLaMA 3.1 to synthesize free-text descriptions from existing structured data. The JobHop setting is more demanding on several axes: the input corpus consists of pseudonymized, multilingual (Dutch, French, English) documents containing `<MASK>` tokens and heterogeneous formatting, and because privacy constraints preclude cloud inference, the extraction pipeline must run entirely on local infrastructure.

3. The JobHop v2 Dataset

This section describes the JobHop v2 dataset, from its underlying resume corpus and ethical safeguards (Section 3.1) to the LLM-based extraction pipeline that converts unstructured text into ESCO-coded career trajectories (Section 3.2), the cleaning (Section 3.3) and normalization (Section 3.4) workflows applied to the raw output, and a summary of the dataset's key statistics (Section 3.5).

3.1. Data Source and Ethical Considerations

JobHop v2 is derived from a corpus of ∼440,000 pseudonymized resumes provided by VDAB, the Flemish Public Employment Service, under a formal research collaboration agreement. Prior to sharing, VDAB converted all resumes to plain text and applied systematic pseudonymization: personally identifiable information and low-frequency terms were replaced with `<MASK>` tokens to reduce indirect re-identification risk. The collection spans multiple languages, predominantly Dutch with smaller portions in English and French, reflecting the multilingual labour market of the Flemish region of Belgium; the composition of this VDAB corpus was reported as 91.4% Dutch, 6.1% English, and 2.4% French in the original JobHop release [4].

Data use is governed by the research agreement and complies with applicable ethical and legal requirements. A binding constraint requires all processing to occur on internal infrastructure: resume contents are never transmitted to external services or third-party APIs. This constraint shaped the extraction architecture (Section 3.2), which relies exclusively on locally hosted model inference. The released dataset incorporates additional privacy protections: location information is removed and all dates are coarsened to quarter-level granularity. Together with VDAB's pseudonymization these measures remove direct identifiers and substantially reduce re-identification risk, but they do not eliminate it: rare combinations of occupations, employers, and timings may remain distinctive, so we characterize the release as strongly pseudonymized rather than fully anonymous. We have not conducted a formal residual linkage-risk assessment, and we ask that users of the dataset make no attempt at re-identification. Documents whose formatting produced empty or near-empty text after pseudonymization were excluded prior to extraction, yielding approximately 400,000 input resumes that entered the LLM extraction stage.

Fairness note. Models trained on historical trajectories may perpetuate structural inequities in labour-market access, occupational segregation, and systemic hiring biases [19, 20]. ESCO standardization provides a neutral occupational vocabulary but does not eliminate the biases encoded in the underlying trajectories, so any deployment should be accompanied by rigorous fairness evaluation across demographic groups. In particular, predictions derived from these trajectories should not be used as the sole basis for high-stakes employment decisions such as hiring, promotion, or eligibility determinations. The dataset card accompanying the public release documents intended and discouraged uses, residual privacy risks, and representational limitations in more detail.

3.2. LLM-Based Information Extraction

The extraction pipeline uses a zero-shot, rule-rich instruction prompt with strict structured JSON output. The prompt has two components. The **system message** is a single sentence establishing the model's role (*"You are an expert HR assistant. Extract only what is explicitly present and format as valid JSON."*). The **developer instruction** con-

Extraction Rules (Sections 1–3)

```
You are a strict Resume Parsing Engine. Your goal is to
extract structured data from the resume provided below into
a single valid JSON object.

1. GENERAL CONSTRAINTS
- Input: Resume text captured at timestamp: {timestamp}.
- Output: A single JSON object. No markdown, no explanations,
no <MASK> tokens.
- Language: Do NOT translate descriptions or titles. Keep
original language.
- Missing Data: Use "-" for missing/masked values. Drop
parts containing only <MASK>.
- Current Roles: Replace markers like "Present", "Heden",
"Current" with the
  provided timestamp formatted as "Month Year".

2. LOGIC & PARSING RULES
- MERGE: Consecutive entries for the same role/company with
overlapping or adjacent
  dates must be combined into one entry.
- SPLIT: Single entries listing disjoint dates (e.g., "2008,
2009 and 2015") must be
  split into separate entries for each period.
- GROUPS: Unpack grouped sections (e.g., "Interims",
"Stages") into individual experiences.
- Dates:
  Format: "Month Year" or "Year".
  Do not invent months. If only a year is listed, output
only the year.
  Handle ranges: "May July 2013" → Start: "May 2013", End:
"July 2013".
  Handle masks: "<MASK>-2009" → Start: "-", End: "2009".
  "<MASK>/<MASK>/2015-<MASK>/<MASK>/2018" → Start: "2015",
End: "2018".
  CLEANUP: Ensure no hyphens or slashes remain from the mask.
  Always extract visible years.
- Mappings:
  Internships: "stage", "trainee", "leerling" →
contract_type: "internship".
  Degrees: "Hogeschool" → "Bachelor"; "Lic." → "Master";
"Doctoraat" → "PhD".
  Languages: "moedertaal/native" → "native";
"vloeiend/fluent" → "fluent";
    "goed" → "intermediate"; "basis/notions" → "basic".

3. EXTRACTION SCOPE
- Skills: Extract ONLY technical skills/tools. Ignore soft
skills.
- Experience:
  CRITICAL: Do NOT extract items from "Education" or
"Profile" sections as work experience.
  Be careful with multi-column layouts.
  For the title field, extract the job title exactly as
written.
  For the standard_title field, infer a standardized title
(preferably ESCO).
- Certificates: professional training/courses only.
```

Output Schema (Section 4)

```
{
  "work_experiences": [
    {
      "title": "Original Job Title",
      "standard_title": "Standardized Job Title",
     "company_name": "Organization Name (or '-')",
      "location": "City only (no country)",
      "description": "Original text description",
      "start_date": "Month Year | Year | -",
      "end_date": "Month Year | Year | -",
      "work_schedule_type":
        "full-time | part-time | -",
      "contract_duration":
        "temporary | permanent | seasonal | -",
      "contract_type":
        "employment | freelance | internship |
         volunteer | student job | -",
      "skills_used": ["Skill 1", "Skill 2"]
    }
  ],
  "educations": [
    {
      "degree_level":
        "Primary education | Secondary school |
         Bachelor | Master | PhD",
      "degree_title": "Title (e.g. B.Sc. IT)",
      "institute_name": "School/University Name",
      "start_date": "Month Year | Year",
      "end_date": "Month Year | Year"
    }
  ],
  "languages": [
    {
      "language": "Name",
      "proficiency_level":
        "native | fluent | intermediate | basic"
    }
  ],
  "other_certificates": [
    {
      "certificate_name": "Name",
      "issuing_organization": "Issuer Name",
      "issue_date": "Month Year | Year"
    }
  ]
}
```

Figure 1: Full developer instruction used for resume extraction. **Left:** Extraction rules covering general constraints (§1), date-handling logic and type-mapping conventions (§2), and extraction scope (§3). **Right:** Target JSON output schema; field names shown in blue. Curly braces denote JSON structure; `{timestamp}` is replaced at runtime with the capture date.

tains the full rule set and output schema, reproduced in Figure 1. It addresses challenges characteristic of the corpus: pseudonymization artifacts (`<MASK>` tokens), multilingual content, inconsistent date formats, and heterogeneous multi-column formatting.

Output schema. The schema defines four entity groups. **Work experiences** carry a job title, an ESCO-aligned standardized title (a secondary signal for downstream code assignment), company name, location, description, start and end dates, work schedule (full- or part-time), contract duration (temporary, permanent, or seasonal), contract type (employment, freelance, internship, volunteer, or student job), and explicitly mentioned technical skills. **Education entries** carry a degree level (PhD, Master, Bachelor, Secondary, Primary), degree title, institution name, and start and end dates. **Languages** carry a name and proficiency level (native, fluent, intermediate, or basic). **Certificates** carry a name, issuing organization, and issue date. This schema is substantially richer than the original JobHop

Table 1
Inference configuration for the production extraction pipeline.

Parameter	Value
Model	`openai/gpt-oss-120b`
Precision	bfloat16
Tensor parallelism	2 (one shard per GPU)
Prefix caching	Enabled
Decoding	Greedy (T=0)
Repetition penalty	1.1 (initial), 1.15 (retry)
Max generation tokens	24,000 (initial), 32,000 (retry)
Reasoning effort	HIGH (both passes)
Batch size	200 resumes per batch
GPU memory utilization	0.90

scope, which was limited to work experiences and a basic educational qualification; the language, certificate, contract, work-schedule, technical-skill, and normalized education-level fields are all new in v2. Technical skills are restricted to tools and technologies explicitly present in the resume text; soft skills are excluded by instruction to limit hallucination.

Inference configuration and throughput. Extraction runs on vLLM [21] with `openai/gpt-oss-120b` [22], a 120B-parameter reasoning model, at HIGH reasoning effort. The complete inference configuration is listed in Table 1. All computation was performed on a single workstation with two NVIDIA H200 NVL GPUs (140 GB HBM3e each), a dual-socket AMD EPYC 9535 CPU, and 1.5 TiB of system RAM, satisfying the on-premises processing constraint. A persistent retry loop reprocesses samples failing JSON validation using an extended generation budget and an increased repetition penalty: of $\sim$400,000 input resumes, $\sim$50,500 (12.6%) failed the initial pass and were all recovered through retry, yielding a final corpus with a 100% JSON parse rate. Figure 2 illustrates the end-to-end behavior of the pipeline on a synthetic resume representative of the VDAB corpus.

3.3. Data Cleaning Pipeline

The raw JSON outputs pass through a five-stage cleaning pipeline, applied sequentially, each operating on the output of the previous one.

1. **Parsing-artifact correction.** Dates containing extraction artifacts are identified via regular expressions targeting residual <MASK> tokens within date fields, future-year placeholders (years > 2025), and malformed strings with dangling hyphens or slashes. Affected dates are corrected to the extractable year component or set to missing.
2. **Invalid-range correction.** Entries where the parsed start date is later than the end date are flagged. Where the discrepancy is small ($\leq$ 1 quarter) the dates are swapped; entries with implausible durations ($>$ 40 years) are removed.
3. **Exact deduplication.** Entries identical across all core fields (title, company, start date, end date) within the same resume are collapsed to a single entry, addressing artifacts where multi-column layouts cause the same position to be extracted twice.
4. **Quality filtering.** Resumes are removed if they contain more than 20 work-experience entries (likely parsing errors on non-resume documents) or zero meaningful entries across all entity groups after earlier stages.
5. **Consecutive-job merging.** Adjacent entries within the same resume are merged if they share the same normalized job title, the same normalized company name, and temporally contiguous or overlapping date ranges (gap $\leq$ 1 quarter); descriptions are concatenated. This stage consolidates approximately 19,432 entries split during extraction due to multi-line or multi-column resume formatting.

3.4. Normalization Pipeline

After cleaning, the extracted fields are normalized into controlled representations suited to cross-resume analysis and modeling. Normalization comprises three independent components: assigning ESCO occupation codes to job entries, converting dates to quarter-year granularity, and mapping education entries to a common degree-level taxonomy.

Occupation-code assignment. Individual job titles are highly variable, so cross-resume analysis requires mapping them to a controlled vocabulary [16]. We adopt the ESCO taxonomy (v1.1.2) [9], organized hierarchically with up to 3,007 leaf-node occupations. Occupation-code assignment uses a commercial classifier provided by Nobl.ai,[1] applied via a two-step hybrid policy that exploits the richer JobHop v2 extraction schema. *Primary assignment* passes the concatenation of the extracted `title` and `description` fields to the classifier, accepting the top-1 prediction if its confidence exceeds $\tau_1 = 0.45$. *Rescue assignment* submits the model-generated `standard_title` for entries below τ_1, accepting the top-1 prediction only if confidence exceeds a stricter threshold $\tau_2 = 0.85$. Entries unassigned after both steps are labeled `unknown`. The asymmetric thresholds are deliberate: primary assignment maximizes coverage at moderate confidence, whereas rescue assignment acts as a high-precision fallback activated only when the standardized title provides a strong signal. Both thresholds were set by manual inspection of the resulting assignments rather than by optimization against labelled data. Manual review of several hundred assigned entries found the labels to be of consistently high quality, but we report no quantitative audit of ESCO assignment accuracy; such an audit, including a characterization of the `unknown` cases, is left to future work.

Temporal normalization. All dates are converted to quarter-year format (Q# YYYY): a multilingual mapping first converts Dutch, French, and English month names to a canonical numeric representation; regular expressions then extract year and optional month from heterogeneous formats ("MM-YYYY", "YYYY-MM", "Month YYYY", abbreviated forms such as "jan. 2015"); months are mapped to quarters (1–3 $\rightarrow$ Q1, …, 10–12 $\rightarrow$ Q4, month-only dates default to Q1); and years outside $[1950, 2025]$ are rejected.

Education-level normalization. Education entries are mapped to a five-level taxonomy via a multilingual term mapping: *PhD* (Doctoraat, PhD, Doctorate); *Master* (Master, Licentiate, Lic., Ma.); *Bachelor* (Bachelor, Hogeschool,

[1] https://nobl.ai/. The classifier is commercially available: access is arranged by contacting the vendor through their website and may be subject to a fee.

(a) Pseudonymized resume (input)

Curriculum Vitae

<MASK> <MASK>
<MASK>, 9000 Ghent
<MASK>@<MASK>.com +32 <MASK>
Belgian, DOB <MASK>/1988

WORK EXPERIENCE

Sept. 2020 -- **Present**:
  Data Engineer
  <MASK> Solutions Inc., Brussels
  Responsible for ETL pipelines ...
  - Tools: Python, Spark, dbt

Jan. 2016 -- Aug. 2020:
  Junior Developer
  InnoSoft Ltd., Ghent
  - Back-end Java/Spring Boot; REST APIs
  - Tools: Java, Spring Boot, PostgreSQL

Mar. 2015 -- Dec. 2015:
  Internship Software Developer
  TechStart SA, Antwerp

EDUCATION
2013--2015 Master in Computer Science
            Ghent University
2010--2013 Bachelor in Applied Informatics
            University College Ghent

LANGUAGES
Dutch: **Native**
English: **Fluent** | French: **Good**

CERTIFICATIONS
Nov. 2021: AWS Certified Solutions Architect
           Amazon Web Services

***Metadata**: Captured at February 2022*

(b) Structured JSON output (extracted)

```
{"work_experiences": [{
   "title": "Data Engineer",
   "standard_title": "Data engineer",
   "company_name": "-", // masked → "-"
   "location": "Brussels",
   "start_date": "Sept. 2020",
   "end_date": "Feb. 2022", // Present → timestamp
   "description": "Responsible for...",
   "contract_type": "employment",
   "skills_used": ["Python", "Spark", "dbt"]
 }, {
   "title": "Junior Developer",
   "standard_title": "Software developer", // ESCO-aligned
   "company_name": "InnoSoft Ltd.",
   "start_date": "Jan. 2016",
   "end_date": "Aug. 2020",
   "contract_type": "employment",
   "skills_used": ["Java", "Spring Boot", ...]
 }, {
   "title": "Internship Software Developer",
   "standard_title": "Software developer",
   "contract_type": "internship", // inferred from title
   "start_date": "Mar. 2015",
   "end_date": "Dec. 2015",
   "skills_used": ["Django"]
 }],
 "educations": [{
   "degree_level": "Master",
   "degree_title": "Master in Computer Science",
   "institute_name": "Ghent University",
   "start_date": "2013", "end_date": "2015"
 }, {
   "degree_level": "Bachelor", // level from title
   "degree_title": "Bachelor in Applied Informatics",
   "institute_name": "University College Ghent",
   "start_date": "2010", "end_date": "2013"
 }],
 "languages": [
   {"language": "Dutch",
    "proficiency_level": "native"},
   {"language": "English",
    "proficiency_level": "fluent"},
   {"language": "French",
    "proficiency_level": "intermediate"} // Good → intermediate
 ],
 "other_certificates": [{
   "certificate_name": "AWS Cert. Solutions Architect",
   "issuing_organization": "Amazon Web Services",
   "issue_date": "Nov. 2021"
 }]}
```

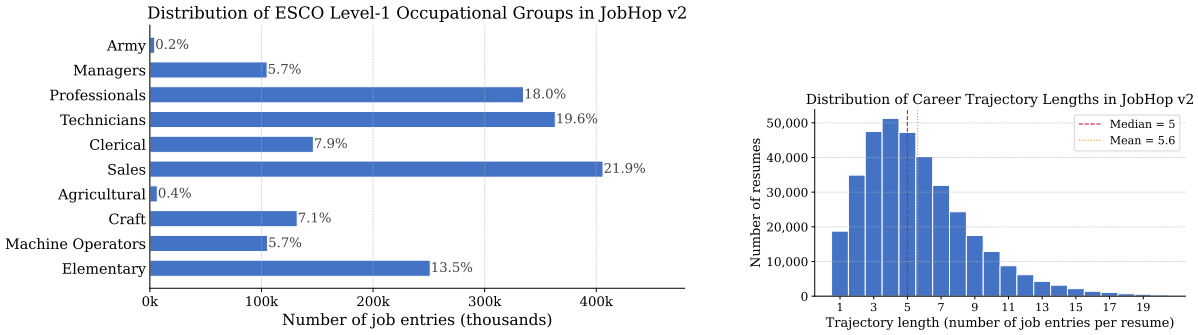
Figure 2: Extraction illustration using a synthetic resume representative of the VDAB corpus. For readability the example is shown in English; in practice resumes may be in any language, predominantly Dutch and French in our corpus, and the model preserves the original language rather than translating it. **(a)** Pseudonymized input: <MASK> tokens replace removed personally identifiable information. Bold terms mark surface cues that the prompt normalizes to controlled schema labels. **(b)** Structured JSON produced by the extraction prompt, demonstrating five pipeline behaviors: (*i*) masked fields resolved to "-"; (*ii*) present-tense marker *Present* resolved to the capture timestamp; (*iii*) free-text descriptions extracted and preserved (shown truncated); (*iv*) job titles mapped to ESCO-aligned `standard_title` values; (*v*) surface terms normalized to controlled schema labels (*Internship* → `internship`, *Good* → `intermediate`).

Professional Bachelor, Ba.); *Secondary* (Secondary, Middelbaar, Secondaire, ASO, TSO, BSO); and *None* (Primary, Lager onderwijs, Primaire, and other degrees). When the `degree_level` field is missing or uninformative, keyword matching is applied to the `degree_title` field, recovering degree-level information for approximately 3,194 entries. The highest attained degree level per resume is computed by taking the maximum across all education entries.

3.5. Dataset Summary

After extraction, normalization, and cleaning, JobHop v2 comprises 355,315 unique resumes, 1,993,291 work-experience entries, and 923,981 education entries.[2] ESCO assignment is dominated by the primary title-plus-description path (93.0%); the rescue mechanism adds 1,484 labels (0.1%); the remaining 6.9% are labeled unknown, re-

[2]The dataset is publicly available at https://huggingface.co/datasets/aida-ugent/JobHop, and the code at https://github.com/aida-ugent/Step.



(a) ESCO level-1 distribution.

(b) Trajectory lengths.

Figure 3: Occupational diversity and career-history depths in JobHop v2. Trajectory lengths range from 1 to over 20 jobs, with a median of 5.

flecting deliberate conservatism to prioritize label quality over coverage.

Reproducibility. Every stage of the pipeline except ESCO code assignment—LLM extraction, the five-stage cleaning workflow, temporal and education normalization, and the evaluation protocol of Section 4—relies only on openly available models and is released as open code, so it can be re-run, inspected, and modified independently. ESCO assignment is the single proprietary component and cannot be reproduced without access to the Nobl.ai classifier. The assigned ESCO codes are nevertheless distributed as part of the dataset, so downstream users can work with the labels without re-running that step; users who require a fully open stack can substitute an alternative occupation classifier at that point in the pipeline without altering any other stage.

Compared with prior public benchmarks, JobHop v2 remains distinctive in being constructed through full end-to-end extraction from real, unstructured resumes, rather than from pre-standardized occupational codes or LLM-synthesized text. Decorte et al. [1] extracted 2,164 histories from Kaggle resumes at considerably smaller scale and without temporal metadata or education annotations; OpenResume [5] releases only 301 real English trajectories; and Karrierewege [2] provides 568,888 paths built from Berufenet occupational codes rather than raw resume text, with free-text descriptions synthesized rather than extracted.

Figure 3 characterizes the dataset along two complementary dimensions. The occupational distribution (Figure 3a) shows that Service & Sales Workers (ISCO 5), Professionals (ISCO 2), and Technicians & Associate Professionals (ISCO 3) account for the largest share of entries, reflecting the white-collar composition of the underlying resume corpus; elementary occupations (ISCO 9) are next most frequent, while primary-sector occupations (ISCO 6) are rare. The trajectory-length distribution (Figure 3b) is right-skewed: most individuals have between two and six recorded positions, but a non-trivial fraction contain ten or more entries, providing rich sequential context for downstream prediction models. Together, the two panels confirm that JobHop v2 captures a broad occupational spectrum and a realistic range of career-history depths, supporting both occupational embedding evaluation and next-occupation prediction.

4. Extraction Evaluation

Reliable evaluation of LLM-based extraction requires both a scoring protocol that faithfully reflects quality for downstream use and reference annotations that minimize labeling noise as a confound. Section 4.1 describes the revised evaluation protocol; Section 4.2 the three labeled reference sets; and Section 4.3 the model comparison.

4.1. Evaluation Protocol

We evaluate extraction quality on 200 benchmark resumes from the original JobHop annotation study [4], using a revised protocol that replaces binary date scoring and unweighted text matching with graduated partial-credit scoring, length-ratio-gated string comparison, and field-importance weighting. The protocol normalizes each field, scores it with a string-similarity or date-specific function, weights fields by importance, and composes these into a single sample-level score.

String similarity. Operating on normalized text (lowercased, whitespace-collapsed, abbreviations unified), the similarity function returns 100 for an exact match or two empty fields, 10 when one field is meaningful and the other missing, a length-ratio-gated score for substring containment (90/75/50 for length ratio $r > 0.7$, $0.5 < r \leq 0.7$, $r \leq 0.5$), and a graduated Levenshtein-based score otherwise (full credit at ratio $s \geq 90$, decreasing penalties below).

Date scoring. Dates are parsed flexibly, supporting English, Dutch, and French month names, with present/ongoing markers mapped to a reference date. Scores range from 85 (year and month match) and 60 (year-only match) down to 15 (no match), with intermediate values for partial parses and missing fields.

Field-importance weights. Fields are weighted by downstream importance: Title and Company/Institute name $1.5\times$; Start and End dates $1.3\times$; Description $1.2\times$; Place/Location and Type $1.0\times$; Degree level $0.9\times$; other fields $1.0\times$ by default.

Sample-level composition. The overall score per sample is

$$S = P_{\text{count}} \cdot \left(0.6 \cdot \bar{M} + 0.2 \cdot R_{\text{entry}} + 0.2 \cdot P_{\text{entry}}\right),$$

where $\bar{M}$ is the mean matched-entry weighted similarity (computed via the Hungarian algorithm), R_{entry} and P_{entry} are entry-level recall and precision, and P_{count} is an entry-count penalty,

$$P_{\text{count}} = \max\big(0.5,\ 1.0{-}0.3{\cdot}|n_{\text{gt}}{-}n_{\text{pred}}|/\max(n_{\text{gt}}, n_{\text{pred}})\big).$$

4.2. Labeled Reference Datasets

The 200 benchmark resumes were annotated by the original JobHop authors and external annotators over approximately 60 cumulative hours. During development of the revised protocol we observed that these annotations contain occasional inconsistencies that can systematically bias accuracy estimates. Four recurring categories account for most of the disagreements we encountered: (*i*) *date granularity*, where one annotation records a month and year for an experience and the other only a year; (*ii*) *section-boundary conventions*, where an item lying near an Education/Experience boundary, or inside a multi-column block, is treated as work experience by one annotation and not by the other; (*iii*) *description scope*, where the amount of free-text detail copied into the description field differs between annotations; and (*iv*) other *subjective judgment calls* that vary across annotators. These are judgment calls rather than unambiguous errors, which is why we treat agreement between reference sets as an interpretive reference rather than a hard bound. To disentangle genuine extraction errors from annotation noise, we evaluate against three annotation sets:

1. **Hand annotations**: the original multi-annotator labels, reflecting genuine human judgment but exhibiting inter-annotator variability.
2. **Edited labels**: a systematic revision by an AI agent (Claude Sonnet) instructed to resolve inconsistencies and enforce uniform conventions while preserving correct annotations; lower noise, same structure.
3. **Agent labels**: an independent annotation set generated by the same AI agent *without access* to the hand or edited annotations, providing an internally consistent reference unanchored to the conventions or errors of the original process.

These form a progression in annotation consistency, from multi-annotator hand labels (highest variability), through edited labels (reduced noise), to agent labels (most internally consistent). Evaluating against all three assesses whether comparative conclusions are robust to the choice of reference.

4.3. Results

Five extraction systems are compared under the revised protocol: GPT-OSS-120B with HIGH and MEDIUM reasoning, GPT-OSS-20B, Llama-3.3-70B [23], and Gemma-2 [24]. GPT-OSS-120B is evaluated across 10 independent extraction runs; 95% confidence intervals are computed via the t-distribution with 9 degrees of freedom.

Table 2 presents the results. The *Label agreement* row reports the symmetric pairwise agreement of each reference set with the other two, and acts as the *annotator-agreement reference* for its column: because the human and agent labels themselves disagree at this level, we do not expect an extractor to substantially exceed it, and the objective for each model is to come as close to it as possible. We call this a reference rather than a ceiling because it is an empirical agreement level between imperfect annotation sets, not a proven upper bound on achievable accuracy. Read this way, GPT-OSS-120B at HIGH reasoning is the strongest system: it comes closest to this reference on all three reference sets, at 83.9 vs. 86.6 (Hand), 86.2 vs. 88.5 (Edited), and 88.0 vs. 89.1 (Agent), a gap of only 1.1–2.7 percentage points, and does so by a smaller margin than any other model, while also outperforming MEDIUM reasoning by 0.6–1.8 pp, which justifies the additional inference cost. Llama-3.3-70B performs competitively, approaching GPT-OSS-120B (MEDIUM) on agent labels; Gemma-2 achieves comparable accuracy on edited and agent labels despite a lower JSON parse rate ($\approx$86% vs. $\geq$98.5%), indicating that its errors concentrate in format compliance rather than extraction quality.

That the best extractor sits within 1.1–2.7 pp of the agreement reference is consistent with a substantial share of the residual error reflecting annotation ambiguity of the kind catalogued in Section 4.2 rather than systematic model failure. We stress that this is an interpretation rather than a measurement: we did not carry out an independently adjudicated error analysis, so we cannot quantify how the residual gap divides between genuinely ambiguous cases and true extraction errors. Reported accuracy also increases from hand annotations through edited to agent labels for every model; this is consistent with the intended progression in annotation consistency, though we note that the agent labels were themselves produced by an LLM and may share inductive biases with the extractors, so scores against them should not be read as a fully independent measure of quality.

4.4. Head-to-Head Comparison against JobHop v1

The evaluation above measures agreement with fixed annotations on the 200 benchmark resumes. To test whether the redesigned pipeline yields better extractions than the original JobHop pipeline on the corpus at large, we additionally run a blind pairwise comparison judged by an LLM.

Protocol. For each of 1,000 resumes sampled from the shared corpus, we render the JobHop v1 and JobHop v2 extractions into a common plain-text schema exposing only the fields present in *both* releases (work experiences, education, and certificates), so that neither side is identifiable by formatting; fields unique to v2 (standardized titles, skills, contract and work-schedule type) are omitted. The two extractions are presented as "Extraction A" and "Extraction B" in a randomized order, alongside the *original* resume as reference. A judge model (Kimi K2.6) scores each extraction on *accuracy* (fidelity of titles, companies, dates, and descriptions to the resume) and *completeness* (whether all experiences, educations, and certificates are captured) on a 0–10 scale, with accuracy weighted more heavily and pure formatting differences disregarded; it also selects an overall winner. The presentation order is remapped afterward for analysis.

Results. Table 3 summarizes the outcome. JobHop v2 is preferred overall on 68.3% of resumes versus 29.9% for v1 (1.8% ties; 69.6% vs. 30.4% excluding ties), a highly significant margin. The advantage is concentrated in work experiences (v2 preferred on 59.7% vs. 29.4%), whereas education is effectively at parity (v2 33.4% vs. v1 31.2%,

Table 2
Extraction accuracy (%) across the three labeled reference sets. **Sim.**: weighted text similarity (60% component weight). **Rec.**/**Prec.**: entry-level recall/precision (20% each). **Final**: full weighted score after the entry-count penalty.

	Hand Ann.				Edited				Agent			
Model	**Sim.**	**Rec.**	**Prec.**	**Final**	**Sim.**	**Rec.**	**Prec.**	**Final**	**Sim.**	**Rec.**	**Prec.**	**Final**
Gemma-2 [24]	79.4	96.2	92.6	81.6	80.3	95.6	97.1	84.8	83.1	96.6	95.7	86.0
Llama-3.3-70B [23]	79.1	**97.2**	93.9	82.5	79.0	**96.5**	97.7	84.8	83.4	**97.8**	95.7	86.6
GPT-OSS-20B	78.9	94.9	94.6	82.0	78.2	93.6	97.8	83.2	82.7	95.1	96.3	85.5
GPT-OSS-120B (MED.)	80.3	94.7	**95.7**	83.3	79.3	93.5	**98.9**	84.4	84.2	94.9	**97.0**	86.7
GPT-OSS-120B (HIGH)	**81.9**	96.4	93.9	**83.9**	**81.7**	96.0	97.6	**86.2**	**86.4**	97.0	95.5	**88.0**
Label agreement[a]	84.4	96.6	95.8	86.6	86.2	97.1	96.7	88.5	87.4	97.0	96.6	89.1

Similarity, Recall, and Precision are sample-level component means under the revised protocol; Final is the weighted score after the entry-count penalty.
GPT-OSS-120B values are means over 10 independent runs (± 0.2–0.5 pp 95% CI). Bold = best model result per column (label agreement excluded).
[a] Symmetric pairwise agreement between label sets, providing an interpretive reference (not a hard upper bound).

Table 3
Blind pairwise comparison of JobHop v2 vs. JobHop v1 extractions, judged by Kimi K2.6 over 1,000 resumes with the original resume as reference. Win rates (%) and mean judge scores (0–10); bold marks the preferred release.

	JobHop v2	**JobHop v1**	**Tie**
Win rate			
Overall	**68.3**	29.9	1.8
Work exp.	**59.7**	29.4	10.9
Education	33.4	31.2	35.4
Mean score			
Accuracy	**8.60**	7.81	–
Completeness	**8.24**	7.38	–

35.4% ties), consistent with the schema and prompt changes chiefly targeting work-experience fields. Mean judge scores favor v2 on both axes (accuracy 8.60 vs. 7.81; completeness 8.24 vs. 7.38). The preference is invariant to presentation order (v2 wins 66.0% and 70.3% of decisions depending on its slot), and the judge's declared winner agrees with the sign of its score difference on 99.8% of resumes, indicating internally consistent verdicts. While configuring the judge we additionally reviewed several tens of its pairwise decisions by hand against the underlying resumes and found its verdicts to agree with our own reading; this is an informal sanity check during development, not a quantified human baseline, and we return to this point in Section 4.5. The judge's free-text rationales most often attribute v2's advantage to cleaner handling of redacted (<MASK>) text, certificate capture, and multilingual date normalization.

4.5. Limitations

Several limitations qualify these results. First, occupational coverage is deliberately conservative: 6.9% of work experiences remain unknown because the two-step assignment policy favors label quality over completeness. That step also depends on a proprietary classifier, making it the one stage of the pipeline that cannot be reproduced without vendor access, and we report no quantitative audit of ESCO label accuracy. Second, the pairwise comparison relies on a single judge model that was not validated against a human baseline. We verified that its verdicts are order-invariant and internally consistent and reviewed several tens of decisions informally during development, but we did not collect human preference labels on a held-out sample and therefore cannot report judge–human agreement. The win rates should accordingly be read as a relative comparison under one automatic judge rather than as an absolute measure of extraction quality, and they may still encode model-specific preferences; the observed effect size is also only small-to-moderate ($d \approx 0.48$) despite the very small p-values afforded by the large sample. Third, the pairwise judge sees only the schema fields shared by both releases, so the additional v2 fields (skills, languages, contract and work-schedule type) do not contribute to the comparison, which likely *understates* v2's added value. Finally, the reference-based evaluation uses 200 resumes from the original annotation study, and both evaluations are confined to the VDAB corpus and its three languages, so generalization to other labour markets remains to be established.

5. Conclusion

We presented JobHop v2, an improved large-scale career-trajectory dataset constructed through end-to-end LLM-based extraction from a corpus of ∼440,000 pseudonymized multilingual resumes provided by VDAB. The released dataset comprises 355,315 unique career trajectories annotated with ESCO occupational codes, quarter-level temporal information, and normalized five-level education attainment. It improves on the original JobHop release in extraction quality, evaluation rigor, and annotation breadth, while preserving its public, real-resume foundation.

The accompanying extraction pipeline, based on reasoning-controlled LLM inference with a multi-step retry mechanism, achieves a 100% JSON parse rate after retry and, across three annotation baselines, comes closest of all compared models to the inter-annotator agreement reference, trailing it by only 1.1–2.7 pp. In a blind pairwise evaluation with the original resume as reference, an LLM judge preferred JobHop v2 extractions over the original JobHop pipeline on 68.3% of resumes versus 29.9%. The revised evaluation protocol, with graduated partial-credit scoring and three complementary annotation baselines, provides a more faithful assessment of extraction quality than prior approaches.

JobHop v2 is intended as a shared resource for the career-trajectory research community. Immediate applications include next-occupation prediction, occupational embedding evaluation, and labour market analysis. Future directions include extending the extraction pipeline to additional labour markets and languages, and supporting fairness evaluation

across education levels and occupational sectors.

Acknowledgments

This research was funded by the BOF of Ghent University (BOF20/IBF/117), the Flemish Government (AI Research Program), the FWO (G073924N), and the EU (ERC, VIGILIA, 101142229). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the EU or the ERC Executive Agency. Neither the EU nor the granting authority can be held responsible for them. For the purpose of Open Access the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. Part of the experiments were conducted on pseudonimized HR data generously provided by VDAB.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-OSS-120B, Claude Sonnet, and Kimi K2.6 in order to: build the dataset extraction, annotation, and evaluation pipeline described in the paper (a core research contribution), and, separately, for grammar and spelling checking of the manuscript. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J.-J. Decorte, J. Van Hautte, J. Deleu, C. Develder, T. Demeester, Career path prediction using resume representation learning and skill-based matching, arXiv preprint arXiv:2310.15636 (2023).

[2] E. Senger, Y. Campbell, R. van der Goot, B. Plank, Karrierewege: A large scale career path prediction dataset, 2024. URL: https://arxiv.org/abs/2412.14612. arXiv:2412.14612.

[3] T. Du, A. Kanodia, H. Brunborg, K. Vafa, S. Athey, Labor-llm: Language-based occupational representations with large language models, 2024.

[4] I. Johary, R. Romero, A. C. Mara, T. De Bie, Jobhop: A large-scale dataset of career trajectories, in: 2025 IEEE International Conference on Big Data (BigData), 2025, pp. 2184–2191. URL: https://arxiv.org/abs/2505.07653. doi:10.1109/BigData66926.2025.11402454.

[5] M. Yamashita, T. Tran, D. Lee, Openresume: Advancing career trajectory modeling with anonymized and synthetic resume datasets, in: 2024 IEEE International Conference on Big Data (BigData), 2024, pp. 6697–6706. doi:10.1109/BigData62323.2024.10825519.

[6] K. Vafa, E. Palikot, T. Du, A. Kanodia, S. Athey, D. Blei, CAREER: A foundation model for labor sequence data, 2024. URL: https://openreview.net/forum?id=4i1MXH8Sle.

[7] R. Zha, Y. Sun, C. Qin, L. Zhang, T. Xu, H. Zhu, E. Chen, Toward unified representation learning for career mobility analysis with trajectory hypergraph, ACM Transactions on Information Systems 42 (2024). URL: https://doi.org/10.1145/3651158. doi:10.1145/3651158.

[8] L. Li, H. Jing, H. Tong, J. Yang, Q. He, B.-C. Chen, Nemo: Next career move prediction with contextual embedding, in: Proceedings of the 26th International Conference on World Wide Web Companion, WWW '17 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2017, p. 505–513. URL: https://doi.org/10.1145/3041021.3054200. doi:10.1145/3041021.3054200.

[9] European Commission, Directorate-General for Employment, Social Affairs and Inclusion, M. Le Vrang, A. Papantoniou, E. Pauwels, P. Fannes, D. Vandensteen, J. De Smedt, ESCO: Boosting job matching in Europe with semantic interoperability, Computer 47 (2014) 57–64. doi:10.1109/MC.2014.283.

[10] Q. Meng, H. Zhu, K. Xiao, L. Zhang, H. Xiong, A hierarchical career-path-aware neural network for job mobility prediction, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 14–24.

[11] R. Schellingerhout, V. Medentsiy, M. de Rijke, Explainable career path predictions using neural models, in: Proceedings of the RecSys in HR 2022 Workshop (co-located with RecSys 2022), CEUR Workshop Proceedings, 2022. URL: https://ceur-ws.org/Vol-3218/RecSysHR2022-paper_6.pdf.

[12] C. Qin, H. Zhu, T. Xu, C. Zhu, C. Ma, E. Chen, H. Xiong, An enhanced neural network approach to person-job fit in talent recruitment, ACM Transactions on Information Systems 38 (2020). URL: https://doi.org/10.1145/3376927. doi:10.1145/3376927.

[13] W. Moore, S. Pedlow, P. Krishnamurty, K. Wolter, I. Chicago, National longitudinal survey of youth 1997 (nlsy97), National Opinion Research Center, Chicago, IL 254 (2000) 22.

[14] U. C. Bureau, U. B. of Labor Statistics, Current population survey, U.S. Census Bureau, 2023. URL: https://www.census.gov/programs-surveys/cps.html.

[15] C. H. Ayishathahira, C. Sreejith, C. Raseek, Combination of neural networks and conditional random fields for efficient resume parsing, in: 2018 International CET Conference on Control, Communication, and Computing (IC4), IEEE, Thiruvananthapuram, India, 2018, pp. 388–393.

[16] N. Li, B. Kang, T. De Bie, Llm4jobs: unsupervised occupation extraction and standardization leveraging large language models, 2023. doi:10.48550/arXiv.2309.09708, arXiv:2309.09708 [cs].

[17] N. Li, B. Kang, T. De Bie, Skillgpt: a restful API service for skill extraction and standardization using a large language model, 2023. doi:10.48550/arXiv.2304.11060, arXiv:2304.11060 [cs].

[18] O. Sainz, I. García-Ferrero, R. Agerri, O. L. de Lacalle, G. Rigau, E. Agirre, GoLLIE: Annotation guidelines improve zero-shot information-extraction, in: International Conference on Learning Representations (ICLR), 2024. URL: https://openreview.net/forum?id=Y3wpuxd7u9.

[19] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, A. T. Kalai, Bias in bios: A case study of semantic representation bias in a high-stakes setting, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, Association for Computing Machinery, New York, NY, USA, 2019, pp. 120–128. URL: https://doi.org/10.1145/3287560.3287572. doi:10.1145/3287560.3287572.

[20] A. Salinas, P. Shah, Y. Huang, R. McCormack, F. Morstatter, The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama, in: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '23, Association for Computing Machinery, New York, NY, USA, 2023. URL: https://doi.org/10.1145/3617694.3623257. doi:10.1145/3617694.3623257.

[21] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, I. Stoica, Efficient memory management for large language model serving with PagedAttention, in: Proceedings of the 29th Symposium on Operating Systems Principles (SOSP '23), Association for Computing Machinery, New York, NY, USA, 2023. doi:10.1145/3600006.3613165.

[22] OpenAI, gpt-oss-120b and gpt-oss-20b model card, https://openai.com/index/introducing-gpt-oss/, 2025. Model release and accompanying technical report.

[23] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, et al., The llama 3 herd of models, 2024. arXiv:2407.21783.

[24] Gemma Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, et al., Gemma 2: Improving open language models at a practical size, 2024. arXiv:2408.00118.