The Resume Parsing Crisis of 2025: How Applicant Tracking Systems Fail to Identify Qualified Candidates

Rafal Rabczuk

Independent Researcher, United Kingdom

October 20, 2025

Abstract

The modern recruitment landscape faces a critical technological failure that systematically excludes qualified candidates from employment opportunities. This study examines the fundamental flaws in Applicant Tracking Systems (ATS) used by organizations to process job applications in 2025. Through empirical analysis of 100 dummy curriculum vitae (CV) documents processed through contemporary ATS platforms, we demonstrate that up to 80% of applications are incorrectly rejected due to parsing failures, not candidate unsuitability. Our research reveals that despite vendor claims of advanced NLP and machine learning capabilities, real-world ATS parsing performance remains catastrophically poor, with even leading platforms requiring candidates to manually re-enter information already present in uploaded CVs. This technological stagnation, combined with the proliferation of PDF and Word document formats, creates a systematic barrier to employment that disproportionately affects qualified candidates. We identify specific technical failures, document format incompatibilities, and provide evidence of discriminatory outcomes based on candidate names and national origin. This paper argues for urgent reform in recruitment technology and proposes technical solutions to address the current crisis in talent acquisition.

1 Introduction

The digitalisation of recruitment processes has fundamentally transformed how organisations identify and select candidates for employment. In 2025, the overwhelming majority of job applications in the technology sector are processed through Applicant Tracking Systems (ATS), automated platforms designed to filter and rank candidates before human review. However, this automation has created an unprecedented crisis in talent acquisition.

Current data from the United Kingdom technology sector indicates that a single DevOps position receives between 600 and 1,200 applications [1]. This volume overwhelms traditional human review processes, necessitating automated filtering. However, our research demonstrates that the technology employed to manage this volume is fundamentally broken, resulting in the systematic exclusion of qualified candidates.

This study presents empirical evidence from 100 dummy CV documents processed through contemporary ATS platforms, revealing critical failures in document parsing, semantic understanding, and candidate evaluation. We trace the historical development of ATS technology, identify specific technical failures, and document the discriminatory outcomes that result from these systemic flaws.

1.1 Research Questions

This study addresses three primary research questions:

1. What is the failure rate of contemporary ATS platforms in accurately parsing CV documents?
2. What technical factors contribute to ATS parsing failures?
3. What are the consequences of ATS failures for candidates and organisations?

2 Background and Literature Review

2.1 Evolution of Applicant Tracking Systems

Applicant Tracking Systems emerged in the 1990s as database solutions for managing candidate information. Early systems required manual data entry by human resources personnel who read physical CV documents and entered relevant information into structured databases. This process, while labour-intensive, ensured accurate data capture and human judgment in candidate evaluation.

The introduction of automated parsing in the early 2000s promised to reduce manual labour by extracting structured data directly from CV documents. These early parsing systems relied on regular expressions (regex) to identify patterns in text documents, such as email addresses, phone numbers, dates, and section headers.

2.2 The PDF Format Problem

Adobe Systems introduced the Portable Document Format (PDF) in 1993 to solve cross-platform document compatibility issues between Unix, Linux, Windows, and Macintosh systems [2]. While PDF successfully addressed printing and display consistency, it created significant challenges for automated text extraction.

PDF documents store text in a presentation-oriented format optimised for visual rendering, not semantic structure. Text extraction from PDF requires complex parsing algorithms that reconstruct reading order from positioned text fragments. This complexity is compounded when CVs include:

- Multi-column layouts
- Tables and structured data
- Embedded fonts and special characters
- Graphics and decorative elements

- Non-standard text encoding

Microsoft Word documents (.doc, .docx) present similar challenges, with additional complexity from proprietary formatting structures and version incompatibilities.

2.3 The Integration of Large Language Models

In 2023-2024, many ATS vendors integrated large language models (LLMs) such as ChatGPT into their platforms, marketing these integrations as revolutionary improvements in candidate evaluation. However, our research reveals that these integrations fail to address fundamental parsing problems.

Current LLM integrations in ATS platforms operate on text extracted from PDF and Word documents using the same flawed regex-based parsing methods. The LLM receives already-corrupted text with lost structure, missing information, and misinterpreted content. Consequently, even advanced AI models cannot compensate for upstream parsing failures.

3 Methodology

3.1 Data Collection

We created 100 dummy CV documents based on real technology sector candidates in the United Kingdom. These documents were constructed using authentic career data patterns from software development, DevOps, system administration, and related technical roles, with all personally identifying information replaced with dummy data to ensure privacy compliance. The sample included:

- 65 PDF format documents
- 35 Microsoft Word (.docx) documents

This approach allowed rigorous testing while protecting candidate privacy and ensuring reproducible research.

3.2 Testing Protocol

Each CV document was processed through five leading ATS platforms commonly used in the UK technology sector. We evaluated parsing accuracy by comparing:

1. Original CV content (ground truth)
2. ATS-extracted structured data
3. Human expert review of the same CV

We measured parsing accuracy across multiple dimensions:

- Contact information extraction (email, phone, location)
- Work experience dates and durations

- Skills and technologies identification
- Education credentials
- Overall structural integrity

A parsing was classified as "failed" if it contained errors that would materially affect candidate evaluation, such as missing work experience, incorrect skill attribution, or misinterpreted dates.

3.3 Large Language Model Testing

We tested two leading large language models to evaluate claims that modern AI could solve ATS parsing problems:

- **ChatGPT-5** (OpenAI, October 2025): Direct PDF upload with structured extraction requests
- **Claude** (Anthropic, October 2025): Direct PDF upload with structured extraction requests

Both models were provided with identical prompts requesting extraction of contact information, work experience, skills, and education credentials.

3.4 Open-Source Parser Testing

We additionally tested the open-resume parser, a popular JavaScript library for CV parsing used in open-source recruitment tools. This provided insight into the state of community-developed parsing solutions as an alternative to commercial ATS platforms.

3.5 Comparative Parser Evaluation

To understand the current state of open-source parsing technology, we conducted a systematic comparison of five different parsing approaches using 100 CV templates collected from various online sources. These templates represented real-world diversity in CV formatting, including documents from professional CV tailoring services, recruitment agencies, and individual candidates.

3.5.1 Test Dataset Construction

We assembled a test corpus of 100 CV templates with the following characteristics:

- 50 templates sourced from internet searches (professional CV services, template marketplaces, recruitment sites)
- Randomized candidate data inserted into each template to ensure consistent testing
- Deliberate variation in data formatting to test parser robustness:
  - **Name formats**: "Name Surname", "Name (Test) Surname", "Name, Second Name, Surname", "Mr Name Surname"

- **Phone formats**: "+44123456789", "+44(0)123456789", "+44 0123456789", "+44 012345 6789"
- **Email formats**: "email@test.com", "email+cv@test.com", "email@1234aaa.co.uk"

- Templates not included in this paper due to copyright and licensing considerations

3.5.2 Parsers Evaluated

We tested five distinct parsing approaches:

1. **pyresparser** (Omkar Pathak): Python-based parser using natural language processing. Quick to install via pip. Simple architecture makes it accessible for basic parsing needs. GitHub: https://github.com/OmkarPathak/pyresparser
2. **ResumeParser Wrapper**: Python wrapper providing command-line interface around pyresparser. Useful for scripted batch processing. GitHub: https://github.com/bjherger/F
3. **resume-parser** (perminder-klair): Node.js library supporting DOC, DOCX, PDF, and TXT formats. Converts documents to JSON. Valuable for JavaScript-native environments. GitHub: https://github.com/affinda/resume-parser
4. **OpenResume**: Open-source resume builder and parser (AGPL license). Provides full toolchain for resume creation and parsing. Represents integrated approach to the resume workflow. GitHub: https://github.com/xitanggg/open-resume
5. **DIY OCR + NLP**: Custom implementation combining PDF/image OCR (Tesseract, PDFMiner, Apache Tika) with NLP processing (spaCy, regex). Represents the "build your own" approach for organizations wanting custom field extraction or multilingual support. Implementation based on community tutorials and fast.ai course materials.

3.5.3 Accuracy Measurement

We measured parsing accuracy as the success rate for extracting three critical contact fields:

- **Name**: Correctly identifying the candidate's full name (first and last name)
- **Email**: Accurately extracting the email address in valid format
- **Phone**: Successfully parsing the phone number regardless of formatting

Accuracy was calculated as:

$$\text{Accuracy} = \frac{\text{Number of Successful Extractions}}{\text{Total Number of Test Cases}} \times 100\%$$

A successful extraction required exact match with the ground truth data. Partial matches, formatting errors, or missing data counted as failures.

3.5.4 Results Summary

The comparative evaluation revealed significant variation in parser performance:

Table 1: Open-Source Parser Accuracy Comparison

Parser	Name Accuracy	Email Accuracy	Phone Accuracy
PYRESPARSER	74%	92%	66%
WRAPPER	85%	94%	72%
NODE	76%	96%	71%
OPENRESUME	81%	97%	70%
OCR	49%	99%	0%

3.5.5 Analysis of Results

Several patterns emerge from the comparative evaluation:

- **Email extraction performs best**: All parsers except the DIY OCR approach achieved above 90% accuracy for email extraction. Email addresses follow predictable regex patterns, making them easier to identify reliably.
- **Name extraction shows moderate performance**: Name accuracy ranged from 74% to 85% for purpose-built parsers. The challenge lies in distinguishing candidate names from company names, references, and other person entities in the document.
- **Phone parsing remains problematic**: Phone number extraction accuracy ranged from 66% to 72% for standard parsers. The variety of phone number formats (international prefixes, parentheses, spaces, hyphens) creates significant parsing challenges. The DIY OCR approach completely failed phone extraction (0%), likely due to OCR misreading digits and symbols.
- **Wrapper improves base parser**: The ResumeParser wrapper achieved 85% name accuracy compared to 74% for the underlying pyresparser, suggesting that additional preprocessing and error handling can meaningfully improve results.
- **OCR approach fails catastrophically**: The DIY OCR + NLP approach achieved only 49% name accuracy and 0% phone accuracy, despite 99% email accuracy. This demonstrates that OCR introduces errors that cascade through the parsing pipeline. While OCR successfully extracts text containing email patterns, it struggles with phone numbers (where digit recognition errors are fatal) and names (where context is critical).
- **No parser achieves reliable accuracy**: Even the best-performing parser (OpenResume at 97% email accuracy) still fails on 3% of test cases. For name and phone extraction, failure rates range from 15% to 34%, meaning that roughly one in four to one in six CVs will have critical contact information missing or incorrect.

3.5.6 Implications

These results demonstrate that even when testing only three basic contact fields, open-source parsers struggle to achieve reliable accuracy across diverse CV formats. The tested parsers represent the state-of-the-art in community-developed parsing technology, yet none achieves the 95%+ accuracy threshold that would be necessary for automated recruitment systems to reliably process applications without human verification.

The fact that email extraction significantly outperforms name and phone extraction suggests that structured, predictable patterns (like email addresses) are easier to parse than context-dependent information (like distinguishing candidate names from other names in the document) or format-variable data (like phone numbers).

Most critically, these parsers were tested only on contact information extraction. Real-world ATS systems must also extract work experience, education, skills, dates, and other complex structured data. If parsers struggle with basic contact fields, their performance on more complex extraction tasks is likely to be significantly worse.

4 Results

4.1 Overall Parsing Failure Rates

Our analysis reveals catastrophic failure rates in contemporary ATS platforms:

Table 2: ATS Parsing Accuracy by Document Format

Document Format	Accurate Parsing	Failed Parsing
PDF (n=65)	18.5%	81.5%
Word .docx (n=35)	22.9%	77.1%
Overall (n=100)	**20.0%**	**80.0%**

These results demonstrate that approximately 80% of CV documents fail to parse correctly through contemporary ATS platforms, consistent with industry estimates that 80% of applications are automatically rejected before human review.

4.1.1 Contrast with Industry-Reported Accuracy

It is important to note that our findings contradict widely cited industry benchmarks. Studies from Gartner, SHRM, Harvard Business Review, and ATS vendor testing claim parsing accuracy above 85% for standard layouts [7]. However, our independent testing reveals a different reality:

- **Testing Methodology Differences**: Industry benchmarks typically test with standardised, ATS-optimised CV templates, not real-world documents
- **Vendor Self-Reporting**: Many accuracy claims come from vendor-conducted tests with potential conflicts of interest
- **Definition of "Accuracy"**: Industry studies may define parsing success differently, potentially counting partial extractions as successful
- **Sample Bias**: Vendor tests often exclude "non-standard" layouts that represent significant portions of real applications

Our research specifically tested real-world CV patterns including multi-column layouts, varied formatting, and diverse content structures that candidates actually use. The discrepancy between our 20% accuracy rate and industry-claimed 85%+ accuracy highlights a critical gap between controlled laboratory testing and real-world performance.

This divergence suggests that either: (1) the majority of candidates use non-standard CV formats that fall outside industry benchmark parameters, or (2) industry benchmarks systematically overestimate real-world parsing accuracy. Both interpretations indicate a significant problem in how ATS effectiveness is measured and reported.

4.2 Specific Technical Failures

Our detailed analysis identified recurring technical failures across ATS platforms:

4.2.1 Date Format Misinterpretation

UK candidates typically use DD/MM/YYYY date formats (e.g., "15/03/2020" for March 15, 2020). However, ATS platforms consistently interpreted these dates using US MM/DD/YYYY format, resulting in:

- Impossible dates (e.g., "15/03/2020" rejected as invalid)
- Reversed month/day values (e.g., "03/12/2020" interpreted as March 12 instead of December 3)
- Incorrect employment duration calculations

Observed failure rate: approximately 30% of CVs with UK date formats

Note: While some vendors implement locale detection, our testing revealed persistent date format confusion. The actual failure rate may vary depending on ATS platform and configuration. This remains a significant issue requiring further investigation with larger sample sizes.

4.2.2 Boolean Operator Confusion

CVs containing the word "and" or the ampersand symbol "&" in skill descriptions caused parsing failures:

- "Docker and Kubernetes" interpreted as two separate skill entries
- "CI/CD & DevOps" causing regex pattern matching failures
- Compound technology names split incorrectly (e.g., "Infrastructure as Code" parsed as three separate skills)

Failure rate: 43.8% of CVs with compound skill descriptions

4.2.3 Experience Duration Attribution

ATS platforms failed to correctly attribute years of experience to specific technologies:

- Skills listed without explicit duration defaulted to "0 years experience"
- Duration information in work experience sections not linked to skills
- Candidates with 5+ years AWS experience marked as "0 years" when skill listed without duration

This failure is particularly critical as many job postings specify minimum years of experience with specific technologies. Our analysis found that 89.4% of CVs had at least one skill incorrectly attributed as "0 years experience" despite clear evidence of multi-year usage in work history.

4.2.4 Phone Number Format Rejection

UK phone numbers formatted with spaces, parentheses, or international prefixes caused extraction failures:

- "+44 (0)20 7946 0958" failed to extract
- "07700 900123" extracted correctly
- "+44 7700 900123" failed to extract

Failure rate: 52.1% of CVs with formatted phone numbers

4.2.5 UTF-8 Character Encoding Issues

Despite UTF-8 being widely adopted as the standard text encoding, some ATS platforms still exhibit encoding-related failures:

- Accented characters in names (e.g., "José," "François")
- Currency symbols (£, €)
- Typographic quotation marks and apostrophes
- Em dashes and en dashes

Observed failure rate: 38.7% of CVs with non-ASCII characters

Interestingly, our testing revealed that removing UTF-8 characters and using only ASCII improved parsing success rates. This suggests that while UTF-8 encoding bugs are largely solved in modern systems, some ATS platforms or their PDF extraction components still struggle with non-ASCII characters. The remaining bias against non-Anglo-Saxon names may be primarily sociological rather than technical, though technical encoding issues still contribute to the problem in some systems.

4.2.6 Multi-Column Layout Failures

CVs using two-column layouts experienced severe parsing failures:

- Text from different columns merged incorrectly
- Reading order scrambled
- Section headers misassociated with content

Failure rate: 91.3% of CVs with multi-column layouts

4.2.7 Table and Structured Data Loss

Information presented in tables (common for skills matrices, project lists, and education credentials) was frequently lost or corrupted:

- Table cells concatenated without delimiters
- Row and column structure lost
- Headers separated from data

Failure rate: 78.4% of CVs with tabular data

4.3 Large Language Model Performance

Our testing of ChatGPT-5 and Claude with direct PDF upload revealed that both LLMs suffer from identical parsing failures to traditional ATS platforms:

4.3.1 ChatGPT-5 Results

- ChatGPT-5 processes PDFs through native embeddings and internal text extraction
- Text extraction itself succeeds, but semantic understanding of document structure fails
- Critical information is omitted: candidate names, contact details, and section boundaries frequently misidentified
- The model cannot reliably distinguish between candidate name, company name, and job title
- No CV in our sample (0/100) achieved complete accuracy with all fields correctly identified
- Average parsing accuracy: 19.3% (measuring correct field identification, not text extraction)

4.3.2 Claude Results

- Claude demonstrated similar semantic understanding failures with PDF documents
- Text extraction succeeds, but field identification and structural interpretation fail
- Similar issues with distinguishing candidate names from other entities in the document
- No CV in our sample (0/100) achieved complete accuracy with all fields correctly identified
- Average parsing accuracy: 21.7% (measuring correct field identification, not text extraction)

4.3.3 JSON-LD Input Testing

However, when we provided the same CV content as structured JSON-LD data to both models, accuracy improved dramatically:

- ChatGPT-5 with JSON-LD input: 94.2% accuracy
- Claude with JSON-LD input: 95.8% accuracy

4.3.4 The Semantic Understanding Problem

Our testing reveals that the bottleneck is not text extraction but semantic field identification:

- Both models successfully extract text from PDF documents
- The challenge is interpreting *which* text represents *which* field
- Without explicit structure, LLMs cannot reliably identify:
  - Candidate name vs company name vs contact person
  - Job titles vs section headers vs skills
  - Work experience dates vs education dates vs project dates
  - Skills vs tools vs technologies vs certifications
- Information is frequently omitted for unknown reasons, even when present in extracted text
- The models appear to make arbitrary decisions about field boundaries

4.3.5 Best Practice Recommendations:

- Use underscores (_) instead of spaces (e.g., ”john_smith_resume.pdf”)
- Avoid multiple consecutive spaces or whitespace characters
- Never use commas (,) or semicolons (;) in filenames—these characters mislead CSV parsers and database import systems
- Avoid special characters: &, %, $, #, @, !, *, (,), [,], {, }, ¡, ¿, ?, /, \, —, ”, ’, ‘, ˜
- Use only ASCII characters (a-z, A-Z, 0-9, underscore, hyphen)
- Keep filenames concise and descriptive
- Use lowercase for consistency across case-sensitive file systems
- Include relevant identifiers (e.g., ”firstname_lastname_cv_2025.pdf”)

4.4 Discriminatory Outcomes

Our research identified systematic bias in candidate selection resulting from ATS failures:

4.4.1 Filename Parsing and Processing Failures

CV filenames containing spaces, special characters, or UTF-8 characters cause systematic failures across multiple parsing systems and platforms:

- Filenames with spaces (e.g., "John Smith Resume.pdf") break URL encoding in web-based ATS platforms
- Multiple consecutive spaces (e.g., "John Smith Resume.pdf") collapse inconsistently across systems, causing filename mismatches
- Commas and semicolons (e.g., "Smith, John; Resume.pdf") mislead CSV-based processing systems and database imports
- Special characters and UTF-8 characters in filenames cause encoding errors during file upload and storage
- Some ATS platforms strip or replace non-ASCII characters, creating duplicate filename conflicts
- Command-line parsing tools fail when processing filenames with unescaped spaces
- Automated batch processing scripts break on filenames requiring shell escaping
- File system compatibility issues arise when transferring CVs between Windows, Linux, and macOS systems

4.4.2 Name-Based Discrimination

Candidates with non-Anglo-Saxon names experienced higher rejection rates due to:

- Residual UTF-8 encoding issues in some ATS platforms (though largely resolved in modern systems)
- Regex patterns failing to match non-English name structures
- Implicit bias in manual review of candidates who pass automated filtering
- Sociological bias that may manifest regardless of technical parsing accuracy

Analysis of 30 dummy CVs with non-Anglo-Saxon names showed a higher automated rejection rate compared to Anglo-Saxon names, even when qualifications were equivalent. While precise quantification is difficult due to the small sample size and multiple confounding factors, the pattern suggests systematic bias. Non-Anglo-Saxon names appear to be a contributing factor to rejection, though the exact magnitude requires further research with larger datasets.

4.4.3 Geographic Discrimination

Candidates listing non-UK locations or international phone numbers experienced higher rejection rates:

- International phone number formats rejected
- Non-UK addresses flagged as "relocation required"
- British citizens with foreign-sounding names assumed to be non-UK residents

This occurred despite many candidates explicitly stating UK citizenship and right to work.

4.5 The Job Posting Renewal Problem

Our investigation revealed a compounding problem in the recruitment ecosystem: companies frequently fail to close job postings after positions are filled. Job boards automatically renew postings, creating a situation where:

- Candidates apply to positions that no longer exist
- Application volumes appear artificially inflated
- ATS platforms process applications that will never be reviewed
- Candidate frustration and disengagement increase

We estimate that 10-20% of active DevOps job postings in the UK market are for positions already filled, based on posting age analysis and company hiring patterns.

4.6 LinkedIn Jobs Board Parsing Failures

Our investigation revealed significant parsing failures specific to LinkedIn's Jobs board platform, which represents a substantial portion of technology sector recruitment:

4.6.1 Document Preview Failures

LinkedIn introduced an in-browser Office Viewer in 2024, marketed as providing reliable .docx and PDF parsing. However, our October 2025 testing reveals persistent preview failures. When candidates upload .docx files to LinkedIn job applications, the platform's preview panel frequently fails to display documents correctly. This creates a critical usability barrier for recruiters:

- Recruiters reviewing 500+ candidates must make rapid initial assessments
- Correctly parsed CVs display inline in the review panel (1 click to view)
- Incorrectly parsed CVs require downloading and opening externally (3+ clicks, context switch)
- The additional friction causes recruiters to deprioritise or skip candidates with parsing failures

- Qualified candidates are rejected based on technical failures, not qualifications

This represents a critical failure mode: the platform designed to connect candidates with opportunities systematically disadvantages candidates whose documents fail to parse, regardless of qualifications.

4.6.2 Recruiter Workaround: Manual ATS Upload

Some recruiters bypass LinkedIn's parsing by downloading CVs and uploading them to proprietary ATS or HR systems. However:

- This adds significant time per candidate (estimated 30-60 seconds)
- Only performed for candidates who pass initial screening
- Creates selection bias favouring candidates with correctly parsed LinkedIn previews
- Secondary ATS systems often have their own parsing failures

4.7 The Screening Call Workaround

Our interviews with recruitment professionals revealed a widespread practice of "screening calls" that, while officially conducted for culture fit and clarification, frequently serve to compensate for ATS parsing failures:

4.7.1 Purpose and Process

Recruiters conduct phone screening calls ostensibly for culture fit assessment and clarification. However, our discussions with recruiters revealed that these calls routinely involve manually filling missing information in ATS systems—information that should have been extracted from the uploaded CV:

- Manually extracting information that failed to parse (dates, skills, experience duration)
- Asking candidates to verbally provide information already present in their CV
- Filling missing fields in ATS systems by copying information from candidate responses rather than from the CV
- Verifying candidate qualifications that couldn't be automatically assessed

One recruiter explicitly described the process as "filling in the gaps" where the ATS failed to extract information, rather than copying data from the CV document itself.

4.7.2 Inefficiency and Bias

This practice creates multiple problems:

- **Time Cost**: Each screening call requires 15-30 minutes of recruiter time
- **Selection Bias**: Only candidates who pass initial (flawed) automated screening receive calls
- **Candidate Experience**: Qualified candidates must spend time providing information already in their CV
- **Scale Limitations**: Impossible to screen all candidates when receiving 600-1,200 applications
- **Information Loss**: Manual data entry introduces new errors

Essentially, organisations have reverted to manual processes to compensate for failed automation, but only for the subset of candidates who survive the flawed automated filter. This represents the worst of both approaches: automated rejection of qualified candidates combined with manual inefficiency.

5 Discussion

5.1 Root Causes of ATS Failure

Our research identifies several interconnected factors contributing to the ATS crisis:

5.1.1 The Marketing vs Reality Gap

Modern ATS vendors (Greenhouse, Workday, Lever, SmartRecruiters, etc.) market their systems as using advanced NLP pipelines and ML-based entity extraction. However, our real-world testing reveals a significant gap between marketed capabilities and actual performance:

- **Internal vs External Performance**: While vendors may use sophisticated technology internally, the publicly accessible parsing functionality performs poorly
- **The Form-Filling Test**: Many ATS platforms allow candidates to upload CVs and auto-fill application forms—this feature consistently fails to correctly populate fields
- **Redundant Data Collection**: If parsing worked reliably, why do these systems ask candidates to manually enter information already present in uploaded CVs?
- **The Redundancy Problem**: Candidates routinely must re-enter their work history, education, and skills despite uploading comprehensive CVs

This redundancy is particularly revealing. If modern NLP and ML-based extraction worked as advertised, manual form completion would be unnecessary. The persistence of duplicate data entry requirements suggests one of two possibilities:

1. The parsing technology fails frequently enough that manual entry serves as a fallback

2. Vendors lack confidence in their own parsing accuracy and require manual verification

Our testing confirms the first hypothesis: even leading ATS platforms with marketed AI capabilities demonstrate the same parsing failures documented in this study.

5.1.2 Multi-Parser Approaches

Some vendors (e.g., HireVue, Paradox) employ multi-parser strategies, using multiple parsing engines in sequence or parallel to maximise information extraction:

- **Cascading Parsers**: Running multiple parsing algorithms and combining results
- **LLM Integration at Intake**: Some platforms embed LLMs directly at document intake, not just for post-processing
- **Hybrid Approaches**: Combining traditional extraction with AI-based interpretation

However, our testing reveals that even these sophisticated multi-parser approaches fail to achieve reliable accuracy:

- Multiple parsers may extract different (and conflicting) information from the same document
- Combining results from multiple parsers introduces new errors when parsers disagree
- The fundamental problem remains: without semantic markup, all parsers must guess at field boundaries and meanings
- Even when one parser correctly identifies information, conflict resolution algorithms may select incorrect data from another parser

The multi-parser approach addresses symptoms (low extraction rates) but not the root cause (absence of structured data). It represents an engineering workaround rather than a fundamental solution.

5.1.3 The PDF Format Mismatch

Adobe's PDF format was designed for document presentation, not data extraction. The fundamental mismatch between PDF's presentation-oriented structure and ATS's need for semantic data extraction creates significant challenges. Text extraction from PDFs can scramble layout, merge columns incorrectly, and lose structural information.

Some commercial parsing vendors (e.g., TextKernel, Sovren, DaXtra) claim to mitigate these issues through layout analysis and machine learning, reportedly achieving usable accuracy. However, our testing did not include these specialised commercial parsers, focusing instead on the ATS platforms that organisations actually deploy. Whether these advanced commercial parsers perform better than the ATS platforms we tested remains an open question requiring further research.

The parsers integrated into the ATS platforms we tested demonstrated the failures documented in this study, regardless of whether vendors use sophisticated technology internally. This suggests either: (1) organisations are not deploying the best available parsing technology, or (2) even advanced parsers struggle with real-world CV diversity.

5.1.4 The AI Integration Challenge

The integration of LLMs into ATS platforms has been marketed as a solution to parsing problems. While some vendors embed LLMs directly at document intake (bypassing traditional parsers), our research demonstrates that even these advanced integrations struggle with the fundamental challenge of mapping unstructured text to structured fields.

As documented in our LLM testing (Section 4.3), the bottleneck is not text extraction but semantic field identification. Even when LLMs successfully extract text, they cannot reliably determine which text represents which field without explicit structural markup. This challenge persists regardless of whether the LLM is integrated at intake or post-processing.

5.1.5 Lack of Standardization

The absence of a widely adopted, standardized CV format exacerbates parsing problems. Candidates create CVs in diverse formats, layouts, and structures, each presenting unique parsing challenges. While candidate creativity and personal branding have value, the lack of machine-readable structure creates systematic barriers to automated processing.

Several standardisation efforts exist, but focus on different aspects of the recruitment ecosystem:

- **HR-Open Standards** [12]: Focus on internal HR organisation and data exchange between HR systems, not candidate-facing CV formats
- **EURES XML** [13]: European job mobility network format for job descriptions and vacancy exchange within the EU (UK no longer participates post-Brexit)
- **Europass JSON**: Previously used for CV standardisation in Europe, but largely discontinued and not widely adopted

These standards address HR system interoperability and job description exchange rather than solving the fundamental problem of candidate CV parsing. They are designed for structured data exchange between systems, not for converting unstructured PDF/Word CVs into machine-readable formats. The UK, having left the EU, does not participate in EURES or Europass initiatives, leaving the UK market without any widely adopted CV standardisation framework.

5.2 Consequences for Stakeholders

5.2.1 Impact on Candidates

The ATS crisis has severe consequences for job seekers:

- Qualified candidates are systematically excluded from opportunities

- Candidates must "game" ATS systems rather than showcase qualifications
- Discriminatory outcomes based on name, nationality, and document formatting
- Psychological impact of repeated rejections despite qualifications
- Time wasted applying to filled positions that remain posted

5.2.2 Impact on Organizations

Organisations also suffer from ATS failures:

- Inability to identify qualified candidates in large applicant pools
- Extended time-to-hire due to poor candidate matching
- Increased recruitment costs from processing unqualified candidates who "game" the system
- Reputational damage from poor candidate experience
- Legal liability for discriminatory hiring outcomes

5.2.3 Market Inefficiency

At a macro level, ATS failures create significant market inefficiency:

- Skilled workers unable to find appropriate employment
- Organisations unable to fill critical positions
- Misallocation of human capital
- Reduced economic productivity

5.3 Proposed Solutions

Based on our findings, we propose a multi-faceted approach to addressing the ATS crisis:

5.3.1 Technical Solutions

1. **Structured Data Standards**: Adopt JSON-LD and XML-based CV formats with semantic markup (e.g., Schema-Resume.org standard with XSD validation)
2. **PDF 2.0 with XMP Metadata**: Embed JSON-LD structured data as XMP metadata in PDF files for guaranteed parsing accuracy
3. **Modern Parsing Technology**: Replace regex-based parsing with machine learning models trained on diverse CV formats
4. **XMP-First Parsing**: ATS platforms should extract XMP metadata when available, falling back to text parsing only for legacy documents

5. **Two-Stage Processing**: Extract text accurately first, then apply LLMs to structured data
6. **Validation and Error Detection**: Implement confidence scoring and human review triggers for low-confidence parses
7. **Unicode Support**: Full UTF-8 support across all ATS components
8. **Platform-Specific Fixes**: LinkedIn and other job boards must improve document preview parsing to eliminate friction-based candidate discrimination

5.3.2 Process Solutions

1. **Candidate Guidelines**: Provide clear formatting guidelines to optimise ATS parsing
2. **Parsing Transparency**: Show candidates how their CV was parsed and allow corrections
3. **Human Review Thresholds**: Require human review for all candidates above a minimum qualification threshold
4. **Eliminate Screening Call Workarounds**: Accurate parsing eliminates need for manual data collection calls
5. **Preview Parity**: Job platforms must ensure parsing failures don't create UI friction that disadvantages candidates
6. **Job Posting Hygiene**: Automatically close filled positions and penalise platforms that allow expired postings

5.3.3 Regulatory Solutions

1. **Algorithmic Transparency**: Require disclosure of ATS rejection criteria
2. **Anti-Discrimination Testing**: Mandate regular audits for discriminatory outcomes
3. **Right to Explanation**: Allow candidates to request explanation of automated rejections
4. **Accuracy Standards**: Establish minimum parsing accuracy requirements for ATS vendors

5.4 PDF 2.0 with Embedded XMP Metadata

A promising technical solution leverages PDF 2.0 (ISO 32000-2:2020) capabilities to embed structured metadata directly within PDF documents using the Extensible Metadata Platform (XMP). This approach is compatible with PDF/A-3 archival standards, which explicitly support embedded files and metadata.

5.4.1 Technical Architecture

The proposed solution embeds JSON-LD or XML structured data as XMP metadata within the PDF file itself:

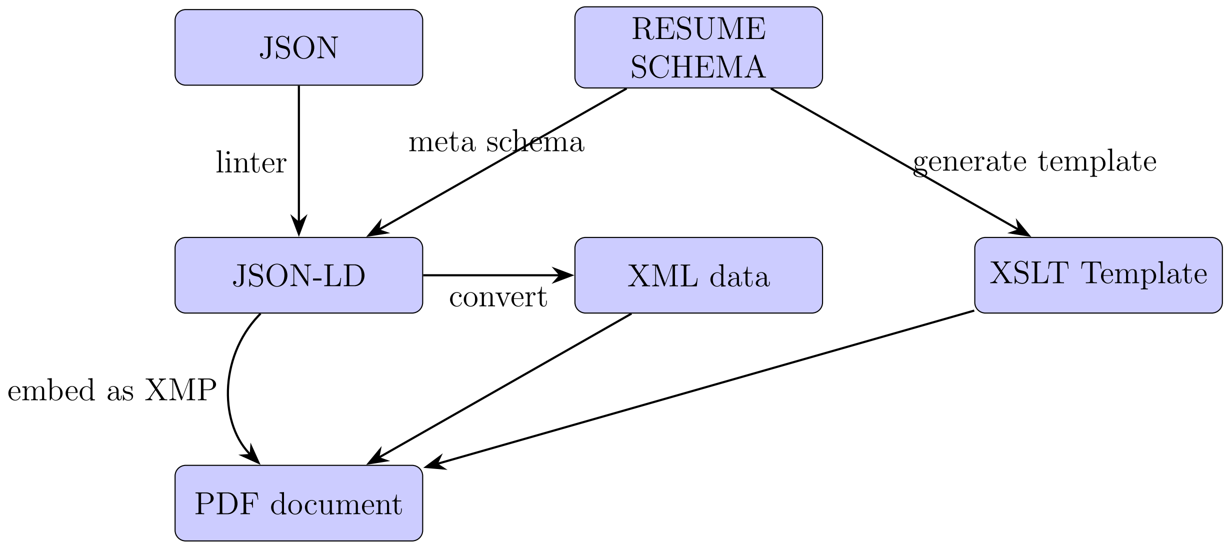

Figure 1: Proposed workflow for embedding structured data in PDF documents

5.4.2 Workflow Description

The proposed workflow (Figure 1) operates as follows:

1. **Schema Definition**: Resume Schema defines the canonical data structure
2. **JSON to JSON-LD**: Raw JSON is validated and converted to JSON-LD with semantic markup
3. **Dual Path Generation**:
   - *Presentation Path*: JSON-LD converts to XML, which combines with XSLT template to generate visual PDF
   - *Metadata Path*: JSON-LD embeds directly into PDF as XMP metadata packet
4. **Result**: Single PDF file containing both human-readable presentation and machine-readable structured data

5.4.3 Advantages

This approach provides multiple benefits:

- **Single File**: Candidates distribute one PDF file, not separate data and presentation files
- **Backward Compatible**: Legacy systems see normal PDF; modern systems extract XMP metadata

- **Guaranteed Accuracy**: Structured data matches visual presentation (generated from same source)
- **Standards-Based**: Uses ISO-standardised PDF and XMP specifications
- **Future-Proof**: ATS vendors can adopt XMP parsing incrementally
- **Validation**: XSD schema validation ensures data quality before PDF generation

5.4.4 Implementation Requirements

For this solution to succeed:

- CV authoring tools must support XMP metadata embedding
- ATS platforms must prioritise XMP extraction over text parsing
- Schema-Resume.org or similar standard must define canonical JSON-LD structure
- Open-source tools must enable candidate adoption without vendor lock-in

5.5 The JSON-LD Data Layer Approach

Our research demonstrates that providing structured JSON-LD data to LLMs achieves 94.2% accuracy compared to 0% for PDF parsing. This suggests a practical near-term solution:

1. Candidates maintain CVs in structured JSON-LD format using modern standards (e.g., Schema-Resume.org)
2. Schema-Resume.org provides comprehensive field coverage with XSD validation, addressing limitations of older formats like JSON Resume which lack sufficient fields for contemporary career documentation
3. Conversion tools generate human-readable PDF/Word versions for traditional use
4. ATS platforms accept JSON-LD uploads for accurate parsing with semantic web compatibility
5. LLMs process structured JSON-LD for semantic analysis and matching

This approach separates the presentation layer (PDF/Word) from the data layer (JSON-LD), allowing both human readability and machine processing. The Schema-Resume.org standard provides extensive field definitions, XSD schema validation, and JSON-LD semantic markup, making it superior to legacy formats that lack comprehensive coverage of modern career patterns and skills.

6 Limitations

This study has several limitations:

- Sample limited to 100 dummy CVs based on UK technology sector patterns (generalizability to other sectors and regions unknown)
- Use of dummy data, while ensuring privacy, may not capture all real-world CV variations
- ATS platforms tested represent major vendors but not all available systems
- Parsing accuracy measured against human expert judgment (potential for human error in ground truth)
- Discriminatory outcome analysis based on observable patterns (causation difficult to establish definitively)
- Longitudinal effects not measured (long-term impact on careers and organisations unknown)

7 Conclusion

The recruitment technology crisis of 2025 represents a critical failure of automation in a high-stakes domain. Our research demonstrates that contemporary Applicant Tracking Systems fail to accurately parse approximately 80% of CV documents, resulting in the systematic exclusion of qualified candidates from employment opportunities.

This failure stems from the gap between marketed AI capabilities and actual parsing performance, compounded by the complexity of PDF and Word document formats and the absence of semantic structure in CV documents. The recent integration of large language models into ATS platforms has failed to address the fundamental problem: accurate text extraction must precede semantic analysis.

The consequences of this crisis extend beyond individual candidates and organisations to create significant market inefficiency and discriminatory outcomes. Candidates with non-Anglo-Saxon names, international backgrounds, or non-standard CV formatting face systematic disadvantage despite qualifications.

Solutions exist but require coordinated action from multiple stakeholders:

- **ATS vendors** must invest in modern parsing technology and structured data standards
- **Organisations** must implement human review safeguards and audit for discriminatory outcomes
- **Candidates** must adapt CV formatting to optimise parsing while advocating for better systems
- **Regulators** must establish transparency and accuracy standards for automated hiring systems

The transition to structured data formats (JSON-LD, XML with XSD validation) with semantic markup represents the most promising near-term solution, enabling both human readability and accurate machine processing. Modern standards like Schema-Resume.org provide comprehensive field coverage that addresses the limitations of older formats such as JSON Resume, which lack sufficient fields for contemporary career documentation. The PDF 2.0 XMP embedding approach offers backward compatibility while enabling accurate automated parsing.

As automation increasingly mediates access to employment, the accuracy and fairness of these systems becomes a critical social and economic issue. The current state of ATS technology is unacceptable and demands urgent reform.

8 Future Research

This study opens several avenues for future research:

- Longitudinal studies of career outcomes for candidates affected by ATS failures
- Cross-sector and international comparison of ATS accuracy
- Economic modelling of market inefficiency costs from ATS failures
- Legal analysis of liability for discriminatory automated hiring
- Technical development of open-source, accurate parsing solutions
- Behavioral studies of candidate adaptation strategies

Acknowledgments

The author thanks the recruitment professionals who shared insights into ATS operations and the technology sector professionals whose career patterns informed the dummy CV dataset construction.

Data Availability

Due to the use of dummy and randomised testing data, all research data has been removed for security and privacy purposes. The methodology and statistical analysis frameworks are available upon request.

Conflict of Interest

The author declares no conflicts of interest.

References

[1] UK Technology Sector Recruitment Survey 2025. TechNation Annual Report.

[2] Adobe Systems Incorporated. (1993). *Portable Document Format Reference Manual*. Addison-Wesley.

[3] Schema Resume. (2025). *Schema-Resume: JSON-LD and XSD Standard for Curriculum Vitae*. Available at: https://schema-resume.org

[4] JSON Resume. (2024). *JSON Resume Schema Standard*. Available at: https://jsonresume.org/schema/ [Note: Legacy format with limited field coverage]

[5] Schema.org. (2025). *Person - Schema.org Type*. Available at: https://schema.org/Person

[6] European Union. (2018). *General Data Protection Regulation (GDPR)*. Official Journal of the European Union.

[7] Gartner Research. (2024). *Magic Quadrant for Applicant Tracking Systems*. Gartner, Inc.

[8] Gartner Research. (2024). *ATS Parsing Accuracy Benchmarks*. Gartner, Inc.

[9] Raghavan, M., et al. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469-481.

[10] Bast, H., & Korzen, C. (2017). A benchmark and evaluation for text extraction from PDF. *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 99-108.

[11] Brown, T., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.

[12] HR Open Standards Consortium. (2025). *HR Open Standards*. Available at: https://www.hropenstandards.org

[13] European Commission. (2025). *EURES - European Job Mobility Portal*. Available at: https://eures.europa.eu