BookMyJobInterview.ai

ORIGINAL BENCHMARK - REPRODUCIBLE - EVERY FIGURE COMPUTED, NONE TYPED

The ATS Parse Benchmark

What actually happens when resume structures meet a parser: a constructed corpus of 12 test documents - identical content, different structures - run through the real parsing path our free tools and paid audits use. Corpus, measurement script and results file all published, so every figure can be independently re-derived.

bookmyjobinterview.ai/ats-parse-benchmark

WHAT THIS BENCHMARK MEASURES - AND WHAT IT DOESN'T

This benchmark measures the parsing path built into bookmyjobinterview.ai — the same code the site's free tools and paid audits run — against a constructed corpus of 12 documents. It does NOT measure any proprietary vendor ATS (Workday, Greenhouse, iCIMS, Taleo or others), and no claim on this page is about any specific vendor's parser.

Why it still matters: the parsing problem is the same one candidates hit everywhere - text extraction, field mapping, keyword search - and this is the one parsing path anyone can inspect and re-run. Where a vendor's parser behaves differently, we say so rather than guess.

6 of 11

HAZARD-STRUCTURE DOCUMENTS LOST VERIFIED CONTENT OUTRIGHT AT PARSE (54.5%)

We planted verified "canary" strings - names, contact details, roles, achievements, skills - in 12 constructed documents, confirmed each string's presence in the raw file, then measured which survived the parse. 6 of the 11 hazard documents lost content outright; most of the rest lost something subtler - field mapping, reading order, or the employment timeline. The control document lost nothing and filled every field: the proof the harness itself is sound.

100%

OF CONTACT DETAILS PLACED IN THE PAGE HEADER OR FOOTER WERE LOST AT PARSE (10 OF 10 CANARIES, 5 DOCUMENTS)

1. Contact details in the header or footer simply vanish

Across the 5 corpus documents that carry their email address and phone number in the Word header/footer region - a layout most template marketplaces ship - not one of those 10 contact strings survived text extraction. The body text arrived intact; the person it belongs to became unreachable.

The parsed record makes the failure concrete: on the header-contact document the extractor found no email address at all, and the phone-number field filled itself with "2013 - 2017" - the education date range, misread as a phone number. A recruiter exporting contact details from that record would dial a degree.

100%

OF CONTENT CARRIED AS IMAGES WAS LOST (4 OF 4 PIXEL-BORNE SKILLS, BOTH IMAGE DOCUMENTS)

2. Anything drawn instead of typed does not exist

Two corpus documents carry part of their skills list only inside an embedded image - a skills chart, the kind design-led templates render as bars or badges. Every skill that existed only as pixels (4 of 4) was absent from the extracted text: text extraction does not OCR images; it discards them. The structure scan did detect the embedded image, and the record's skills field was left holding a heading with nothing under it.

79.2%

OF THE CREATIVE-HEADINGS DOCUMENT'S LINES MAPPED TO NO FIELD (19 OF 24 LINES; 4 OF 8 FIELDS FILLED)

3. Unusual section headings lose nothing - and map almost nothing

The most instructive failure, because nothing is lost: every canary survived extraction. But with "My Journey" in place of Work Experience and "What I Bring" in place of Skills, the field-mapper recognised no section to file the content under - 19 of 24 lines (79.2%) landed in no database field, and only 4 of 8 standard fields filled, against 8 of 8 for identical content under conventional headings. The resume reads beautifully to a human, extracts perfectly, and produces a database record that is mostly empty.

---

5 of 8

RECORD FIELDS FILLED FROM THE SEED TWO-COLUMN CV (2 OF 9 CANARIES LOST; SUMMARY FIELD EMPTY)

4. Tables and columns scramble reading order more than they delete text

On this parsing path, table layouts mostly kept their text: the constructed tables and two-column documents each lost 0 canaries. What broke was order and mapping - the two-column document's sections were read sidebar-first (SKILLS, EDUCATION, CERTIFICATIONS, SUMMARY, WORK EXPERIENCE), and the seed documents, built the way real templates actually are, fared worse: the two-column CV filled 5 of 8 record fields with its summary empty and its header-borne contact details gone; the table-based resume filled 6 of 8, its structure scan flagging 3 tables and its free health score landing at 44 - the corpus's lowest.

Honest caveat: Different extraction engines linearise tables differently - some scramble cell order far more aggressively than this path does. Our figures describe our extractor; the hazard flags exist precisely because behaviour varies.

---

2 of 4

DATE RANGES RECOVERED FROM THE NON-STANDARD-DATES DOCUMENT (CONVENTIONAL DATES: 4 OF 4)

5. Non-standard date formats halve the recoverable timeline

The audit engine's chronology parser recovered every range on the control document (4 of 4) and only 2 of 4 on the same document with its role dates rewritten into three non-standard formats - tenure, gaps and role order became uncomputable for two of the three roles. Subtler still: the free health score's date-format check awarded the document 10 of 10 points, because the odd formats were internally consistent enough to pass a consistency check. A document can look fine on a format check and still not yield a timeline.

38.5%

OF VERIFIED CONTENT LOST ON THE COMPOUND DOCUMENT (5 OF 13 CANARIES; 2 OF 8 FIELDS FILLED)

6. Hazards compound: the heavily designed template fails on every axis at once

One document combines the hazards the way real designed templates do - two-column table layout, contact only in the header, skills only as an image, creative headings, non-standard dates. It lost 38.5% of its canaries, filled 2 of 8 record fields, mapped 17 of 20 extracted lines to no field, and scored 65 on the free health check against the control's 93 - same words, same person, same experience. Structure is the only variable in this corpus, so every gap from the control is attributable to structure alone.

The hazard that didn't bite (here): text boxes

We report the non-failures too. The text-box document - its skills list inside a floating text box - lost 0 canaries on this parsing path: the extractor read the box's content. The damage was subtler: the box's text merged into the surrounding line and swallowed the Education heading that followed it, leaving that record field empty - but the skills survived. Many extraction engines do skip text-box content entirely; ours, on this corpus, did not. A benchmark that only published failures would be marketing - this line is what keeps the figures above believable.

Verified content lost at parse, per document

Full results

One row per corpus document. Canaries lost = verified content absent after extraction; fields = standard parse-record fields filled; unmapped = extracted lines mapped to no field; timeline = date ranges recovered by the audit engine's chronology parser against the ranges the document carries; health = the free Resume Health Score for the identical text.

Document	Canaries lost	Fields	Unmapped	Timeline	Health
Control: clean single-column	0/14	8/8	0/24	4/4	93
Content laid out in tables	0/14	8/8	0/47	4/4	85
Two-column layout	0/14	8/8	0/31	4/4	83
Contact details in the page header	3/14	6/8	0/22	4/4	84
Contact details in the page footer	2/14	7/8	0/23	4/4	84
Skills inside a floating text box	0/13	7/8	0/23	4/4	88
Skills carried as a graphic	2/13	8/8	0/23	4/4	89
Unusual section headings	0/14	4/8	19/24	4/4	78
Non-standard date formats	0/14	8/8	0/24	2/4	93
Compound: several hazards at once	5/13	2/8	17/20	2/4	65
Seed: table-based layout (existing fixture)	2/9	6/8	1/28	4/4	44
Seed: two-column layout (existing fixture)	2/9	5/8	1/27	3/3	50

Methodology

The corpus: 12 .docx documents. Ten are generated with identical fictional content - one candidate, one career - varying only the structure under test (clean single-column control, table layout, two-column layout, contact in the page header, contact in the page footer, skills in a floating text box, skills as an embedded image, unusual section headings, non-standard date formats, and a compound document). Two are the pre-existing hazard test documents that seeded the corpus. Corpus, generator and ground-truth manifest are stored in the repository with each file's hash pinned.

Ground truth before measurement: every canary string is re-verified against the raw document XML - header and footer parts included - before anything is measured, and any canary claimed to sit outside the body is rejected if it also appears in the body. Image-borne canaries are validated against the text file the image is deterministically rendered from.

The measurement: each document runs through the same parsing path this site runs on real documents - the .docx text extraction used by every free tool and the paid audit pipeline (mammoth 1.12.0), the audit engine's structure scan, the simulator's field-mapping record, the audit engine's chronology parser, and the health-score composite with published weights. Canary survival is whitespace-stripped, case-insensitive substring matching - a bias that runs toward survival, i.e. against overclaiming loss. The control document must come through perfectly or the harness refuses to emit results. Benchmark version 1.0.

LIMITATIONS, STATED PLAINLY

This measures our parsing path, not Workday's, Greenhouse's, iCIMS's or any other vendor's - vendor parsers differ, and none of these figures should be quoted as a claim about a specific vendor. The corpus is constructed to isolate structures, not a survey of real resumes: the percentages describe these 12 documents under this parser, and their value is exact reproducibility, not population estimation. Text-box handling genuinely splits across parsers (this path read the box; many don't). Only .docx documents are measured - PDF extraction is a different pipeline.

Cite This Benchmark

BookMyJobInterview.ai (2026). The ATS Parse Benchmark: how document structures survive resume parsing - a reproducible corpus study. bookmyjobinterview.ai/ats-parse-benchmark - reuse any figure or chart with a link back; the corpus, measurement script and results file are published, so every figure can be independently re-derived.

Parse failure is silent - nothing tells you your contact details vanished or your experience mapped to no field. BookMyJobInterview.ai rewrites your resume and LinkedIn profile so the structure parses cleanly and the wording matches how recruiters in your field actually search - checked by the same engine measured above, then verified by an expert reviewer. Free tools: resume health score, job match score, ATS simulator, sameness score at bookmyjobinterview.ai.