Supplementary Information

Supplementary Table 1: Criteria and scoring scheme for expert audit of numerical (mechanical properties and composition) and textual (processing route) fields.

Attribute	**Dimension**	**Scoring Rule**	**Score Range**
Numerical	Property	1 point if YS, UTS, and EL perfectly match the source; 0 points if any field is missing or incorrect.	0 or 1
Numerical	Composition	1 point if percentages of 36 elements perfectly match; 0 points if any field is missing or incorrect.	0 or 1
Textual	Processing	1.0: Complete steps, correct sequence, and faithful semantics. 0.8: Minor missing of non-essential information. 0.6: Slight sequence deviation without affecting understanding. 0.2–0.4: Significant omissions or excessive simplification. 0: Content contradicts the source or is untraceable.	0 to 1

Supplementary Table 2: Consistency scoring rules for benchmarking automated extraction against manual records.

Dimension	**Sub-item**	**Max Score**	**Scoring Rules**	**Objective**
Coverage	Total Coverage	40	0: No corresponding candidate found. 20: Weak match (similar content, missing key attributes). 30: Clear correspondence (all key attributes covered). 40: Strong match (high consistency across all fields).	Detect if manual records are recovered.
Accuracy	Properties	20	20: Numerical values (YS, UTS, EL) perfectly match manual records. 0: Any discrepancy or missing value.	Evaluate numerical recovery quality.
Accuracy	Composition	20	20: Elements, symbols, and ratios perfectly match manual records after normalization. 0: Incorrect symbols or inconsistent ratios.	Evaluate extraction and standardization.
Accuracy	Process	20	20: Complete steps, correct conditions, and semantic consistency in manufacturing routes. 0: Significant omissions or incorrect parameters.	Evaluate natural language recovery.
Overall	**Total Score**	**100**	**Total = Coverage (40) + Accuracy (60).** Scores ≥ 80 are defined as *high-confidence matches*.	Integrated assessment.

Supplementary Table 3: Literature search strategy. Search parameters for corpus retrieval (2000–2026).

Parameter	Description
Search period	2000 – 2026 (publication year)
Search fields	Title or abstract
Databases	**Elsevier** (via APIs, full-text access)
Total retained articles	18,019 DOI-traceable metallurgical articles (after duplicate removal, metadata validation, full-text availability screening and steel-relevance filtering)
Keywords	**Steel-specific terms (85+):**

CFB Steel	Aluminum-Silicon Coated	Wear-Resistant Steel
Carbide-Free Bainitic Steel	Hot-Stamped Steel	Creep-Resistant Steel
Nanostructured Bainitic Steel	Alloyed Steel	Heat-Resistant Steel
Low-Temperature Bainitic Steel	Low-Alloy Steel	Cryogenic Steel
Super Bainitic Steel	High-Strength Low-Alloy Steel	Electrical Steel
Bainitic Ferrite Steel	HSLA Steel	Silicon Steel
Austempered Bainitic Steel	Microalloyed Steel	Additively Manufactured Steel
Advanced High-Strength Steel	Ultrahigh-Strength Steel	3D-Printed Steel
AHSS	UHSS	Laser Powder Bed Fusion Steel
Third-Generation Advanced High-Strength Steel	Tool Steel	Directed Energy Deposition Steel
3rd Gen AHSS	Bearing Steel	Gradient Nanostructured Steel
Dual-Phase Steel	Spring Steel	Nanocrystalline Steel
DP Steel	Stainless Steel	Ultrafine-Grained Steel
Transformation-Induced Plasticity Steel	Austenitic Stainless Steel	Heterostructured Steel
TRIP Steel	Ferritic Stainless Steel	Laminated Steel
Twinning-Induced Plasticity Steel	Martensitic Stainless Steel	Clad Steel
TWIP Steel	Duplex Stainless Steel	Composite Steel
Quenching and Partitioning Steel	Super Duplex Stainless Steel	Medium-Entropy Steel
Q&P Steel	Maraging Steel	High-Entropy Steel
Quenched and Tempered Steel	Precipitation-Hardened Steel	Compositionally Complex Steel
Martensitic Steel	PH Steel	Lightweight Steel
Tempered Martensitic Steel	Age-Hardened Steel	Low-Density Steel
Hot-Stamped Steel	Secondary-Hardening Steel	Fe-Mn-Al-C Steel
Press-Hardened Steel	Oxide-Dispersion-Strengthened Steel	Fe-Mn-Si Steel
Boron Steel	ODS Steel	Fe-Cr-Ni Steel
	Hydrogen-Resistant Steel	Fe-Co-Ni Steel
	Hydrogen-Embrittlement-Resistant Steel	Fe-Cr-Mn-Ni Steel
	Hydrogen-Storage Steel	
	Corrosion-Resistant Steel	

General materials terms (45+):

metallic materials	superalloys	materials	precipitation-strengthened alloys
metals and alloys	nickel-based alloys	multi-principal-element alloys	age-hardenable alloys
engineering alloys	cobalt-based alloys	MPEAs	dispersion-strengthened alloys
structural metals	titanium alloys	high-entropy alloys	oxide-dispersion-strengthened alloys
ferrous alloys	aluminum alloys	high-entropy alloys (HEA)	additively manufactured metals
iron-based alloys	magnesium alloys	medium-entropy alloys	additively manufactured alloys
advanced metallic materials	copper alloys	compositionally complex alloys	laser powder bed fusion alloys
structural alloys	bulk metallic glasses	refractory high-entropy alloys	directed energy deposition alloys
lightweight alloys	metal matrix composites	lightweight high-entropy alloys	
high-strength alloys	nanostructured alloys	eutectic high-entropy alloys	
corrosion-resistant alloys	ultrafine-grained metals		
wear-resistant alloys	gradient nanostructured metals		
heat-resistant alloys	heterostructured		
creep-resistant alloys			
refractory alloys			

Note: Only articles with at least one keyword in title or abstract and published after 2000 were considered.

Supplementary Table 4: Fixed element set for composition vector alignment. The 36 chemical elements used for composition vector alignment.

No.	Element	Remarks
1	H	Hydrogen
2	B	Boron
3	C	Carbon
4	N	Nitrogen
5	O	Oxygen
6	F	Fluorine
7	Na	Sodium
8	Mg	Magnesium
9	Al	Aluminium
10	Si	Silicon
11	P	Phosphorus
12	S	Sulfur
13	Cl	Chlorine
14	Ca	Calcium
15	Ti	Titanium
16	V	Vanadium
17	Cr	Chromium
18	Mn	Manganese
19	Fe	Iron
20	Co	Cobalt
21	Ni	Nickel
22	Cu	Copper
23	Zn	Zinc
24	As	Arsenic
25	Y	Yttrium
26	Zr	Zirconium
27	Nb	Niobium
28	Mo	Molybdenum
29	Sn	Tin
30	Sb	Antimony
31	La	Lanthanum
32	Ce	Cerium
33	Ta	Tantalum
34	W	Tungsten
35	Pb	Lead
36	Bi	Bismuth

Supplementary Table 5: Metallurgy-aware validation rules and tolerances. Key validation rules for numerical fields.

Field	Acceptable range	Logical consistency
UTS / YS (MPa)	0 – 3000	YS $\leq$ UTS
Elongation (%)	0 – 100	–
Element (wt.%)	0 – 100	sum $\approx 100 \pm 5\%$
Temperature (°C)	-273 – 2000	–
Time (min/h)	> 0	–

Note: All values are converted to standard units (MPa, %, °C, min/h) with a tolerance $\leq 0.1\%$ after conversion.

Supplementary Table 6: Hyperparameters for deduplication and record selection. Similarity thresholds and weights.

Parameter	Value	Description
α	0.6	Weight for composition similarity
τ	0.85	Composite similarity threshold
λ	0.7	Confidence vs. completeness trade-off

Supplementary Methods 1: Complete prompt template and extraction instructions

```
You are an expert in materials-science information extraction. You will receive the text of one
DOI-traceable steel-related paper, its bibliographic metadata, and tables parsed from the paper
in JSON format. Extract experimentally reported steel material-state records that link material
identity, chemical composition, processing route and mechanical properties.

Return only a valid JSON object with the following top-level structure:
{
  "records": [
    {
      "source": {
        "doi": "...",
        "title": "...",
        "year": null
      },
      "material": {
        "material_name": "...",
        "grade_name": null,
        "state_or_condition": null
      },
      "composition": [
        {
          "element": "C",
          "value": 0.20,
          "unit": "wt.%",
          "evidence": "..."
        }
      ],
      "processing_route": {
        "process_description": "...",
        "ordered_steps": [
          {
            "step_type": "solution treatment",
            "temperature_C": 900,
            "time": 30,
            "time_unit": "min",
            "deformation": null,
            "atmosphere_or_cooling": "water quench",
            "evidence": "..."
          }
        ]
      },
      "mechanical_properties": {
        "YS": {"value": null, "unit": "MPa", "evidence": null},
        "UTS": {"value": null, "unit": "MPa", "evidence": null},
        "EL": {"value": null, "unit": "%", "evidence": null}
      },
      "provenance": {
        "section_or_table": "...",
        "evidence_spans": ["..."]
      }
    }
```

```
  ]
}

Extraction rules:
1. Extract one record for each experimentally reported material state, not merely for each
article. A material state is defined by material identity, composition, processing condition and
measured mechanical properties.
2. Use the DOI provided in the metadata exactly as the source DOI for every extracted record.
3. Extract only values supported by the supplied text or tables. Do not infer, estimate or
complete missing values from external knowledge.
4. Use null for missing, uncertain or unsupported record-level fields. Do not fill unreported
composition elements with 0 during extraction; zero filling is applied only later during
composition-vector alignment.
5. Standardize units where the conversion is explicit and unambiguous: strengths to MPa,
elongation to %, temperature to degree Celsius, and time to min or h. If a numerical value lacks
a confirmable unit, set that field to null.
6. Preserve evidence spans for numerical values and processing descriptions whenever the source
text or table provides localized evidence.
7. Reconstruct processing routes as ordered steps. Include temperatures, times, deformation
levels, cooling methods and atmospheres when reported. Resolve vague pronouns such as "the
sample" to the specific material or condition when the context supports it.
8. If a table is referenced by text, integrate the table values directly into the corresponding
record rather than writing phrases such as "as shown in Table".
9. Return valid JSON only. Do not include explanations, markdown, comments or text outside the
JSON object.

Input metadata:
{metadata}

Input text:
{text}

Input tables:
{tables}
```

In the actual execution, `{metadata}` was replaced with bibliographic metadata including DOI, `{text}` with the raw text of the target paper, and `{tables}` with the JSON representation of tables extracted from that paper. The prompt was provided to each LLM exactly as shown above.