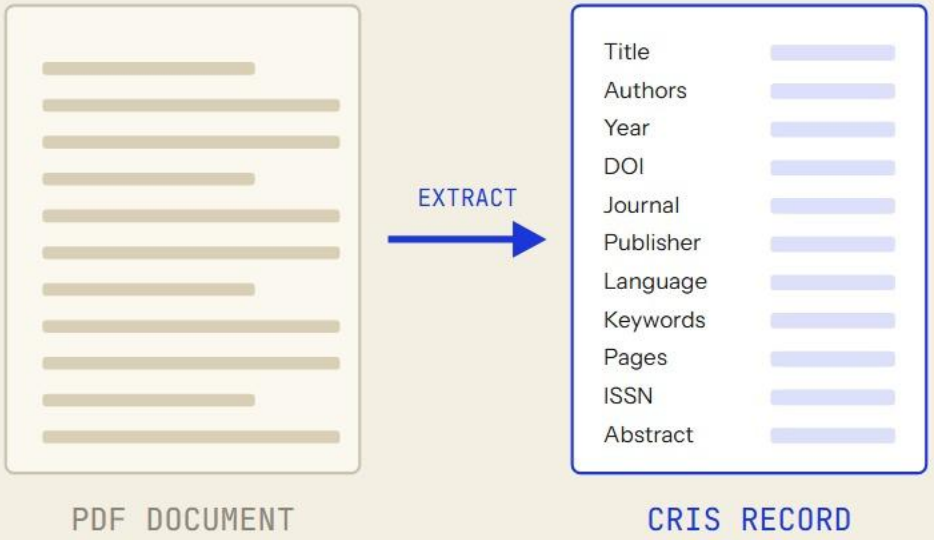
Evaluating LLM-based metadata extraction for CRIS with a multi-criteria framework

Metadata extraction for CRIS - a 960-paper, 12-language, 9-model study, measured by precision and recall.

EUROCRIS 2026

Turning a PDF into a structured record.

- Every paper entering a CRIS must become a **structured record**. Today this is largely manual work for librarians.
- **Classical tools exist.** GROBID parses bibliographic headers with CRF models; Docling reads page layout. Both handle only predefined fields and were not built for multilingual papers.
- A **registry lookup** - Crossref or OpenAlex by DOI - is the most efficient option whenever the record already exists.
- But attribution often happens **for the first time** - a new deposit or preprint has no registry record. Then the PDF is the only source.



How we frame the task - and how we judge it.

Research questions.

RQ1 Which fields are extractable from the PDF text alone - and which need an external source, no matter the model?

RQ2 How do language and research domain affect extraction quality?

The registry as a reference

Querying Crossref or OpenAlex by DOI does not read the document. We use it only to show how much metadata exists for these papers in the first place.

Prompting from MOLE framework; stricter evaluation.

We adopt the schema-driven prompting strategy of MOLE (Alyafeai et al., 2025) - the model fills a given schema. We evaluate more strictly than MOLE and report precision and recall apart.

How we score, and what we compare against.

- **Recall** - of the fields in the document, how many the model recovered. **Precision** - of what it wrote, how much is right and how much is invented. Their harmonic mean is **F1** - the strict score: `F1 = 2 · P · R / (P + R)`
- **Two reference sources** - the publisher page and the OpenAlex record. A field matches if it agrees with either, with fuzzy matching: word overlap for text, name matching for authors, exact match for identifiers.
- **The correct index - judged in two stages.** First: does the value match a reference? If not, a separate audit model decides whether it is still right - another language, script or format - or a genuine error. Both stages together give the correct index, shortly after the study design.

TITLE - CLEAN MATCH

model: **Machine learning in materials science**
reference: **Machine learning in materials science**
✓ correct - matches the reference

TITLE - REFERENCE IN ANOTHER LANGUAGE

model: **Inhibitory kinaz janusowych** (Polish paper)
reference: **Janus kinase inhibitors** (English in OpenAlex)
◆ correct but flagged - strict matching missed it

AUTHOR - DIFFERENT SCRIPT

model: **Gu-Hwan Kim** · reference: **구환 김**
◆ correct but flagged - same person, romanised

AUTHOR - DIFFERENT NAME ORDER

model: **A. Taghavi-Eraghi** · reference: **Taghavi-Eraghi, A.**
◆ correct but flagged - a formatting variant

ISSN - NOT PRINTED IN THIS PDF

model: **2049-3630** · reference: **(none - not in the paper)**
✗ wrong - a genuine hallucination

960 papers, 12 languages, 10 systems.

Corpus - 960 papers = 4 years (2022–2025) × 12 languages × 4 research domains × 5 papers per cell. For each cell, the first OpenAlex records with a PDF and an abstract.

`Atomic metadata checks:` **`960 × 10 × 12 = 115,200`**

Four publication years

2022 · 2023 · 2024 · 2025

Four research domains

Life sciences · Social sciences · Physical sciences · Health sciences

Twelve languages

English · German · French · Spanish · Portuguese · Italian · Polish · Russian · Indonesian · Chinese · Japanese · Korean

Metadata fields examined

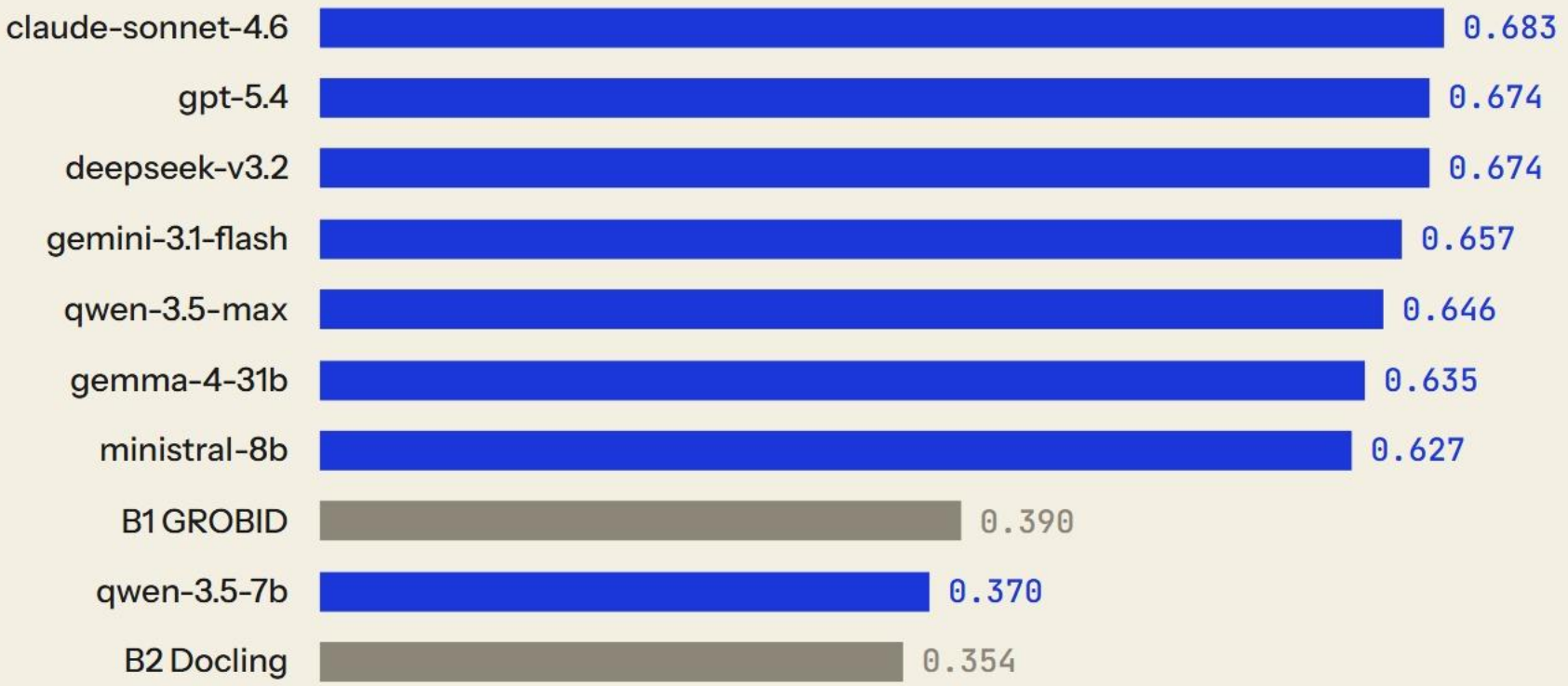
title · authors · year · DOI · journal · publisher · language · keywords · pages · ISSN · abstract

Eight language models + two classical parsers

claude-sonnet-4.6 · gpt-5.4 · gemini-3.1-flash · deepseek-v3.2 · gemma-4-31b · qwen-3.5-max · qwen-3.5-7b · ministral-8b · GROBID · Docling

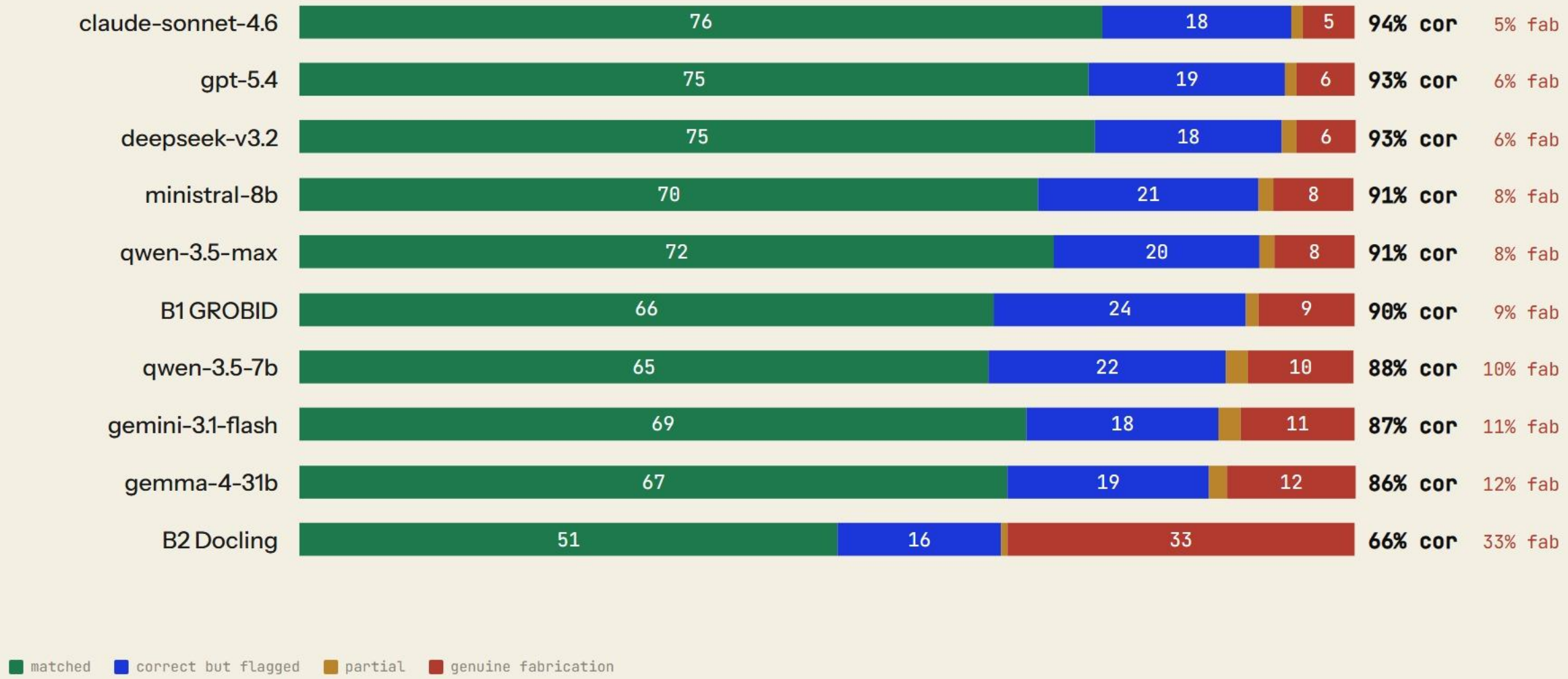
Under strict matching – the F1 score.

F1 - precision and recall combined, with strict string matching, before the audit. Blue = language model · grey = classical parser.



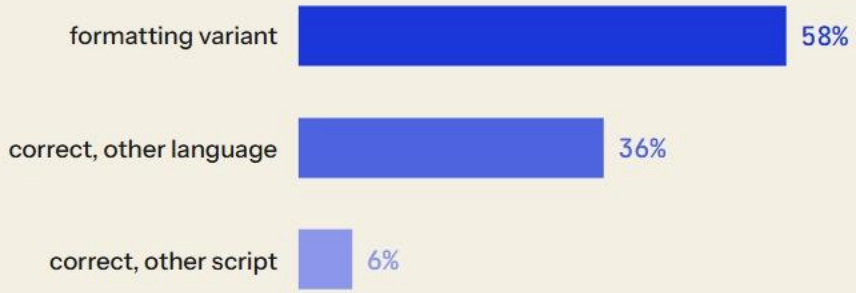
Most extracted metadata is correct, but errors vary sharply by system

Each bar split four ways: matched (green), correct but flagged (blue), partial (amber), genuine fabrication (red). Eight language models plus the two classical parsers. The total **"% correct"** - green + blue - is shown after each bar.



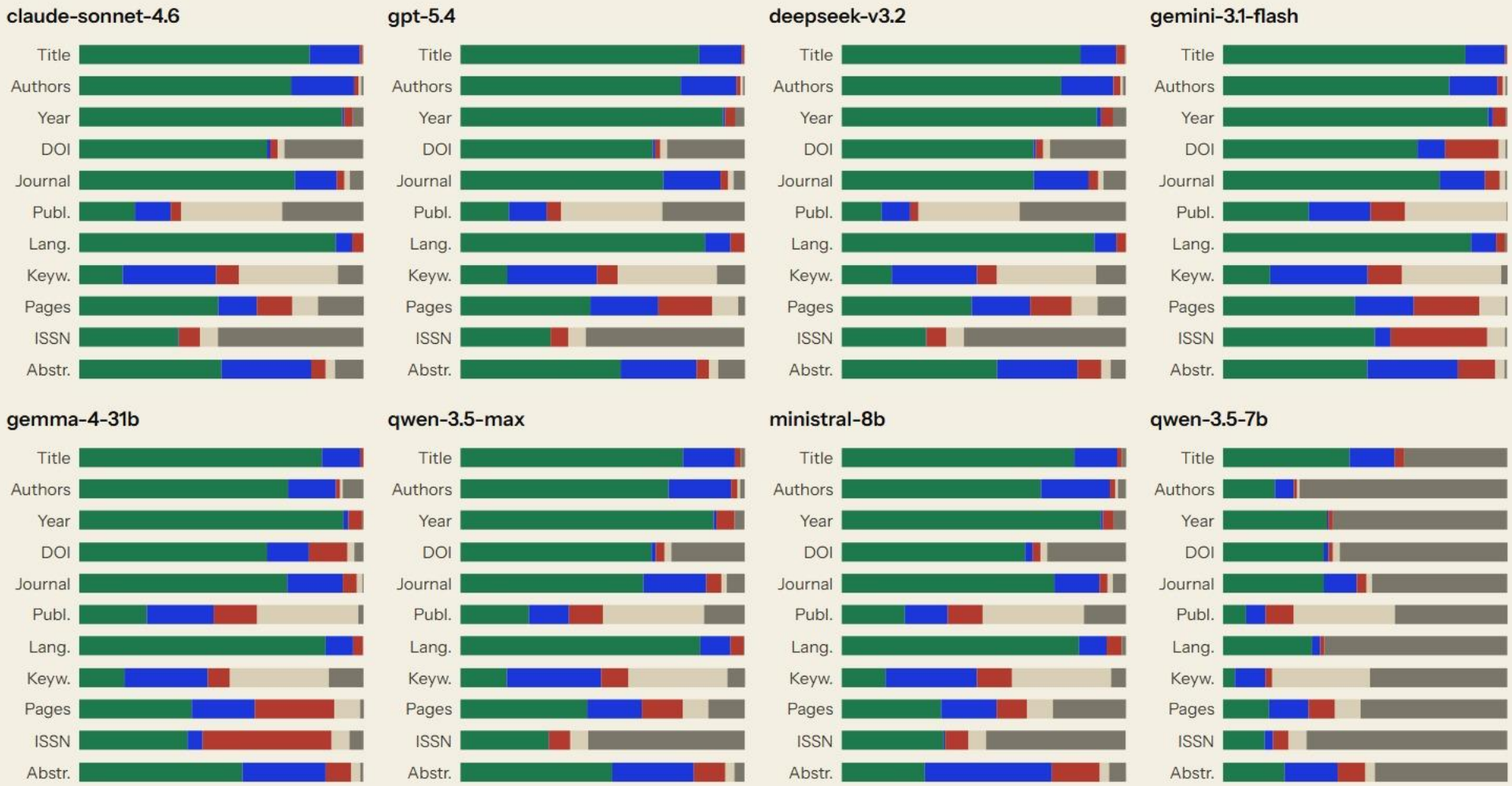
What the "correct but flagged" extractions are.

- **Formatting variants - 58%.** "pp. 21-30" vs "21–30"; "Univ." vs "University"; a different author order. The value is right; the spelling-out differs.
- **Correct, other language - 36%.** The model wrote the title or authors in the paper's own language; the reference stored an English translation.
- **Correct, other script - 6%.** A romanised name against the original script - the same person.



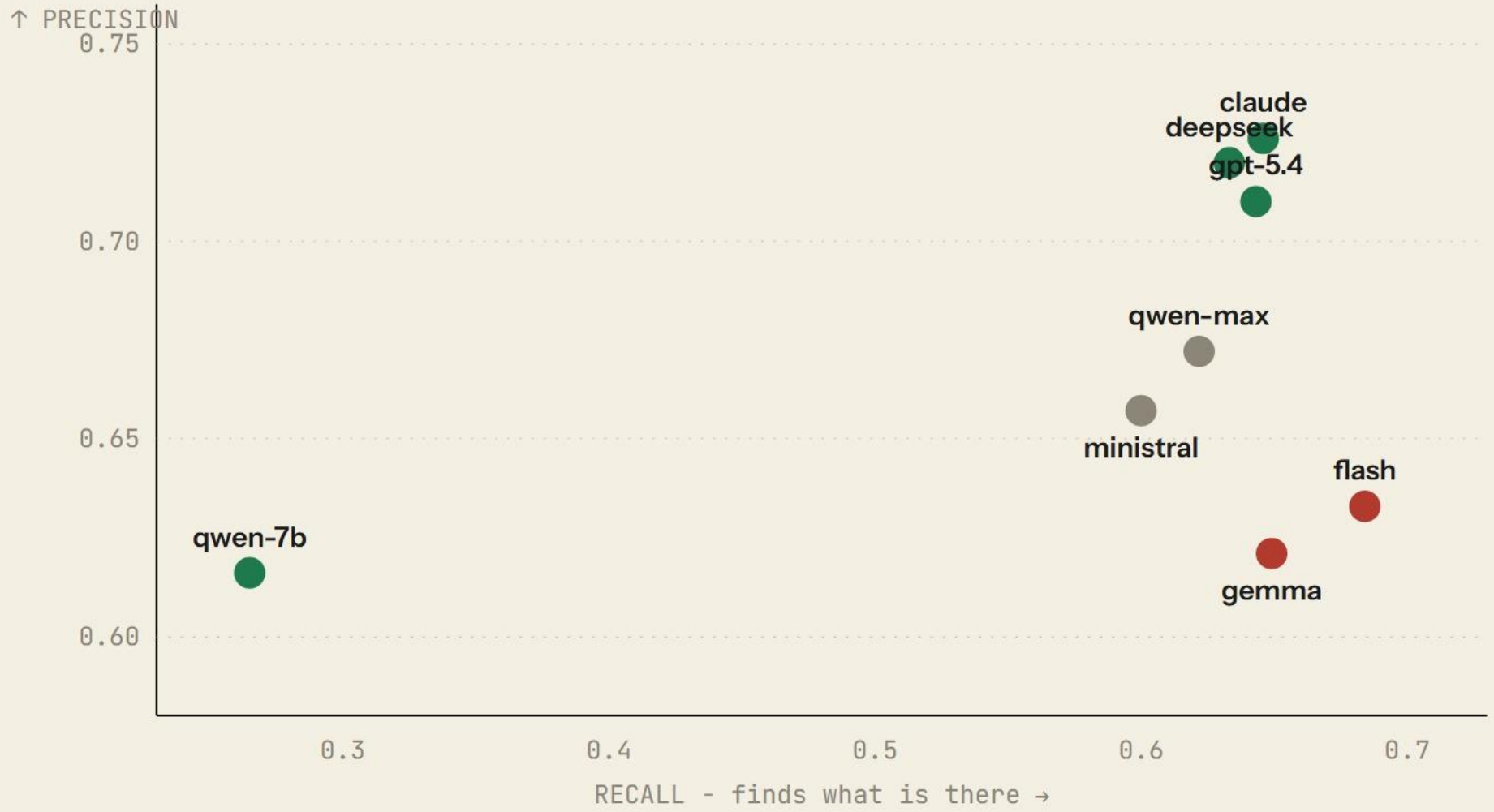
Per field, per model – what is correct, what is genuine error.

Each panel is one model, each bar one field. Green = matched · blue = correct but flagged · red = genuine error · sand = not in the document · grey = no answer.



The model that answers the most invents the most.

Recall on the horizontal axis, precision on the vertical - the ideal is the top right. Red dots over-fill (invent) often; green dots rarely. The careful models sit higher: they answer a little less, and invent about half as often.



When a model is wrong

7,178 genuine-error cells from the audit (eight language models plus the two classical parsers), classified into five error types, and the concentration heatmap of those errors across model and field

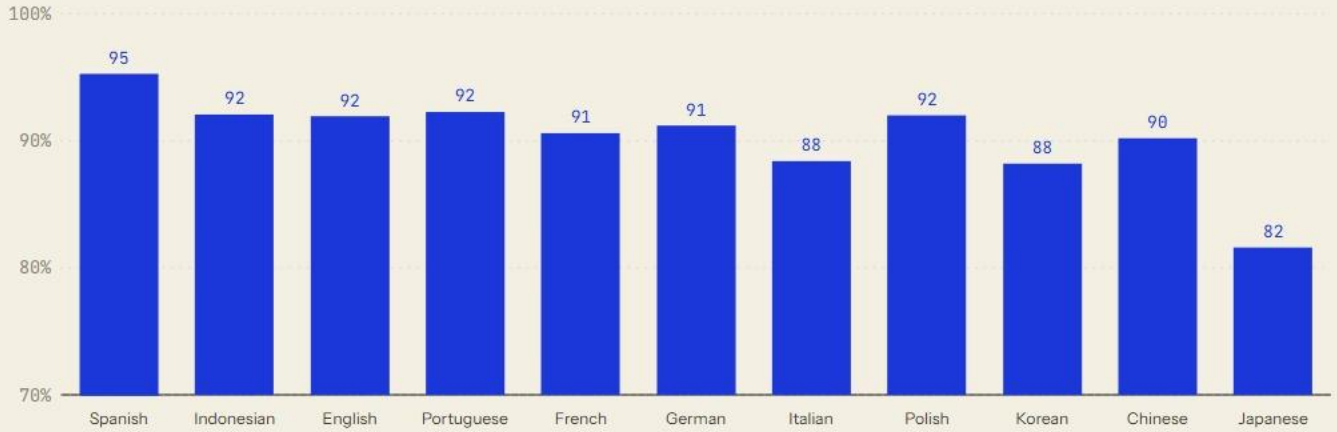
THE FIVE ERROR TYPES

TYPE	WHAT IT MEANS	CELLS	SHARE
Fabricated identifier	An invented DOI / ISSN / pages / year - looks valid, but is not the real value.	1963	**27.3%**
From the wrong place	A real value lifted from the wrong part of the document - a cited reference, a running header, or a different field.	3051	**42.5%**
Invented (plausible)	A plausible-sounding journal / publisher / title that is simply made up.	943	**13.1%**
Inferred, not stated	Guessed from context rather than read - e.g. publisher from the journal name, language guessed.	1139	**15.9%**
Garbled	Malformed, truncated, encoding-broken or nonsense output.	82	**1.1%**

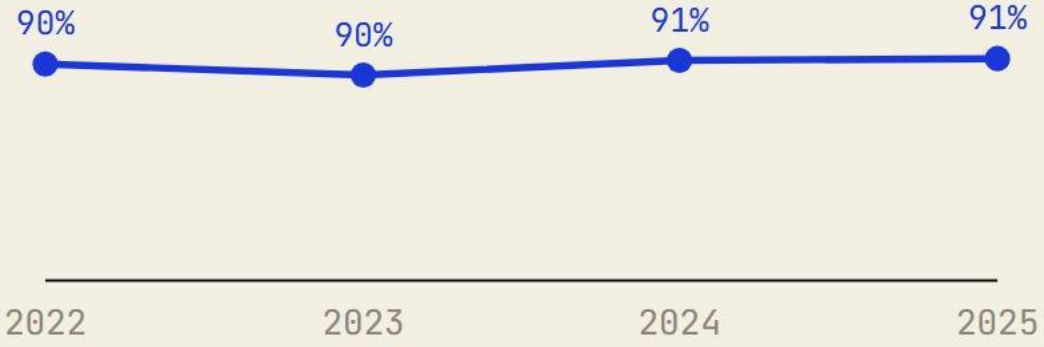
WHERE ERRORS CONCENTRATE - MODEL × FIELD

MODEL	TITLE	AUTHO	YEAR	DOI	JOURN	PUBLI	LANGU	KEYWO	PAGES	ISSN	ABSTR	TOTAL
claude-sonnet-4.6	8	12	30	25	25	35	38	60	48	71	45	**397**
gpt-5.4	8	10	34	16	25	48	47	54	115	60	37	**454**
deepseek-v3.2	28	18	39	25	31	28	33	51	63	67	62	**445**
gemini-3.1-flash	6	15	45	178	50	117	31	87	85	326	90	**1030**
gemma-4-31b	11	10	47	129	48	147	34	55	149	429	67	**1126**
qwen-3.5-max	19	17	61	30	51	114	44	77	68	72	74	**627**
qwen-3.5-7b	32	9	16	15	32	94	14	20	44	51	58	**385**
ministral-8b	16	18	36	26	26	118	51	94	58	79	103	**625**
B1-grobid	82	62	11	10			63	33		60	53	**374**
B2-docling	128	140	465	192			46	49	152	517	26	**1715**

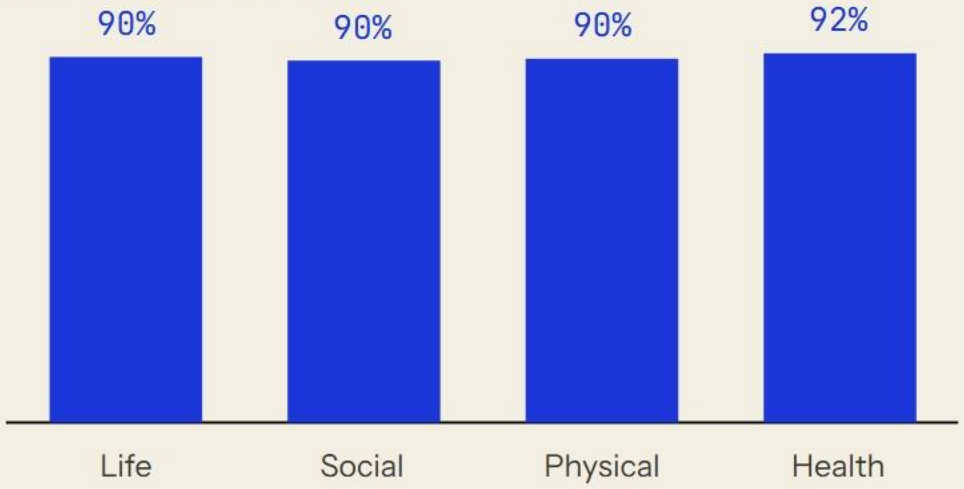
Accuracy is stable across years and domains, while language still matters



Year of publication



Research domain



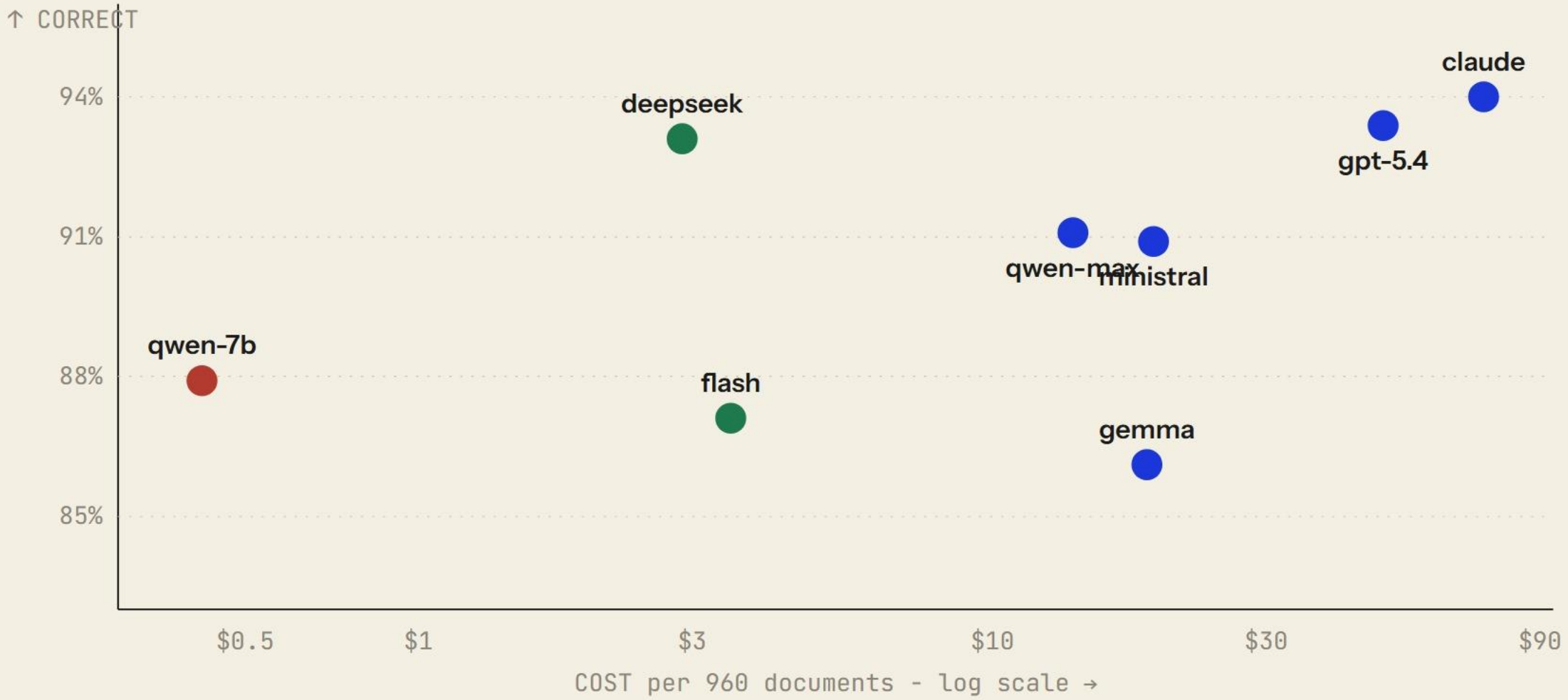
Careful models stay reliable across languages.

Correct index for 8 models against 11 languages. Claude, GPT and DeepSeek hold near 90% almost everywhere; the always-answering models drop, and drop most in Japanese.

	es	id	en	pt	fr	de	it	pl	ko	zh	ja
claude-sonnet-4.6	98	95	95	97	94	94	92	94	91	94	90
gpt-5.4	97	95	94	93	92	94	91	95	91	95	91
deepseek-v3.2	98	94	95	96	93	94	92	95	90	93	81
ministral-8b	94	93	91	92	91	91	89	94	89	90	85
qwen-3.5-max	96	93	94	93	93	93	90	92	88	91	77
qwen-3.5-7b	91	90	93	89	88	86	87	88	83	90	72
gemini-3.1-flash	93	88	90	90	87	89	84	89	86	86	75
gemma-4-31b	94	89	85	89	86	86	83	87	86	85	78

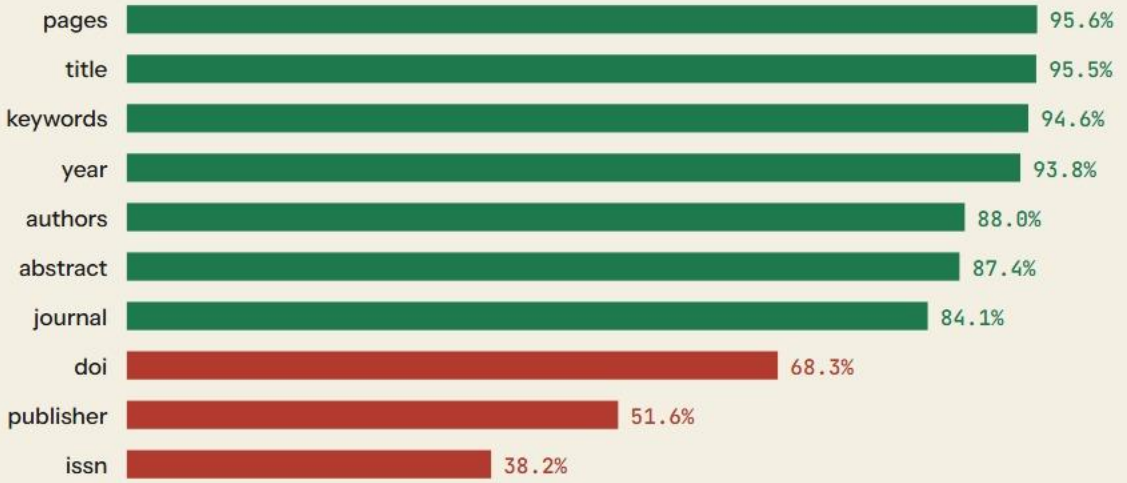
Spending more does not buy better extraction.

Cost of processing all 960 documents against the correct index. DeepSeek (green) costs under $3 and scores within a point of Claude, which costs 25× more.



Unrecovered metadata

- **Recoverability** = share of papers where at least one system produced a value matching the registry. The complement is what no model found - but we **cannot tell** whether the value was missing from the PDF or every model simply failed to pick it up.
- DOI **68%**, publisher **52%**, ISSN **38%** sit at the bottom - these are fields publishers rarely print on the page itself, so source-absence is the likely driver.
- Even **title** is unrecovered for ~**4.5%** of papers. Inspection shows these are **non-standard documents** - conference programs, session schedules, multi-abstract pages - where the DOI-registered title is genuinely not on the rendered page.



The raw model is a baseline. The workflow is the method.

All results so far are raw single-prompt models - a lower bound. Validation, normalization and registry linking raise precision and close the absent-field ceiling.

Extraction
LLM reads the trimmed PDF

→

Schema validation
type, format, cross-field checks

→

Normalization
DOI / ISSN / language canonicalised

→

Registry linking
ORCID · ROR · Crossref - closes absent fields

→

Human evaluation
expert review of flagged records

→

Provenance log
source snippet, model, tool calls

What this means for building a metadata autofill pipeline

Choose models for precision

A cautious model that abstains is safer than one that always answers. And cost is not a barrier - the cheapest capable models match the most expensive ones.

Add a registry-linking step

Extraction alone cannot reach ISSN, DOI and publisher - fields often absent from the PDF. A registry lookup closes them.

Wrap the model in a workflow

Validation and normalization catch errors a raw model leaves; log provenance so every value can be audited.

Flag high-risk fields

Affiliations and grant identifiers carry institutional weight - route them to targeted human review.

Dedicated workflow inside our **DSpace-CRIS** instance, clearly **speeds up the editors' work**