

Datasheet: korpus_pl_courts v1.0

Polish court judgments corpus built from the full dump of the SAOS aggregator (System Analizy Orzeczeń Sądowych, <https://www.saos.org.pl/>). One JSONL file, one judgment per line: cleaned plain text with section structure and rich metadata. v1.0 covers the entire SAOS database as of 2026-07-12 (537,455 source records ingested).

Export file: `korpus_pl_courts_v1.0.jsonl` — **505,611 records, 11.6 GB** (11,623 MiB), validated line by line.

Volume

Metric	Value
Source records ingested (full SAOS dump)	537,455
Records in corpus (after cleaning and deduplication)	505,611
Quarantined (kept on disk with reasons, not in corpus)	28,655
Near/exact duplicates removed (kept on disk with provenance)	3,189
Estimated tokens (chars / 3.5)	~3.18 B
Total text characters	11,117,331,392
Judgment date range	1986 – 2026 (peak 2013–2017)
Records with a reasoning section (UZASADNIENIE)	385,925 (76.3%)
Court anonymization placeholders ([ANONIM])	31,253,735
PII maskings applied	4,808 occurrences in 1,904 records

Distribution by court type:

Court type	Records
COMMON (common courts: appellate, regional, district)	436,127
SUPREME (Supreme Court)	38,051
NATIONAL_APPEAL_CHAMBER (public procurement, KIO)	22,106
CONSTITUTIONAL_Tribunal	9,327

Distribution by judgment type: SENTENCE 363,617 · DECISION 81,750 · REASONS 46,315 · REGULATION 11,461 · RESOLUTION 2,468.

Quarantine reasons: empty source text 28,631 · text shorter than 500 chars 21 · judgment date outside 1918–2026 range 3.

Record schema

Each line is a JSON object with the following fields:

Field	Type	Description
<code>id</code>	string	Stable identifier, <code>saos-</code> + SAOS numeric id (e.g. <code>saos-148362</code>)
<code>sections</code>	list	<code>{header, text}</code> objects; header is e.g. <code>WYROK</code> , <code>POSTANOWIENIE</code> , <code>UZASADNIENIE</code> , or <code>null</code> for preamble/unsectioned text. The full document text is not duplicated in the file — reconstruct it as shown below

Field	Type	Description
<code>court</code>	string null	Court name (e.g. "Sąd Apelacyjny w Poznaniu"); common courts only — <code>null</code> for Supreme Court, Constitutional Tribunal and KIO (their division ids live in a different id space than the common-courts dictionary)
<code>division</code>	string null	Division name (e.g. "I Wydział Cywilny"); common courts only
<code>court_type</code>	string	SAOS court class: <code>COMMON</code> , <code>SUPREME</code> , <code>CONSTITUTIONAL_Tribunal</code> , <code>NATIONAL_APPEAL_CHAMBER</code>
<code>judgment_type</code>	string	<code>SENTENCE</code> , <code>DECISION</code> , <code>REASONS</code> , <code>REGULATION</code> , <code>RESOLUTION</code>
<code>judgment_date</code>	string	ISO date of the judgment
<code>docket_numbers</code>	list	Case signatures (e.g. <code>I ACa 957/14</code>); a record may have several
<code>judges</code>	list null	<code>{name, roles}</code> ; roles include <code>PRESIDING_JUDGE</code> , <code>REPORTING_JUDGE</code>
<code>keywords</code>	list null	Editorial keywords from the source
<code>legal_bases</code>	list null	Cited legal bases (statute articles)
<code>language</code>	string	Always <code>"pl"</code>
<code>tokens_est</code>	int	Length estimate: <code>len(text) / 3.5</code>
<code>anon_count</code>	int	Number of <code>[ANONIM]</code> placeholders in the text
<code>pii_masked_count</code>	int	Number of PII maskings applied by the pipeline (<code>[TELEFON]</code> , <code>[EMAIL]</code> , <code>[NR_KONTA]</code> , <code>[NIP]</code> , <code>[NR_DOWODU]</code> , <code>[PESEL]</code>); 0 for untouched records
<code>source_url</code>	string null	URL of the source document at the originating institution
<code>license</code>	string	<code>public_domain_pl_art4_prawo_autorskie</code> (see Legal status)
<code>hash</code>	string	SHA-256 of normalized reconstructed text (lowercased, whitespace collapsed); stable across releases for deduplication

Reconstructing the full text

The export stores each document once, in `sections`. The full text is reconstructed deterministically — section headers become standalone lines, sections are separated by a blank line:

```
def build_text(sections):
    parts = []
    for s in sections:
        parts.append(f"{s['header']}\n{s['text']}" if s["header"] else s["text"])
    return "\n\n".join(parts)
```

`tokens_est` equals `len(build_text(sections)) / 3.5`, and `hash` is the SHA-256 hex digest of that text lowercased with whitespace collapsed (`" ".join(text.lower().split())`) — both can be used to verify a reconstruction is byte-correct.

Processing pipeline

1. **Ingest** — SAOS dump API (`/api/dump/judgments`), 100 records/page, 1 s politeness delay, custom User-Agent, atomic page writes, resumable via a state file (resume point reconstructible from page files). Transient network errors retried with exponential backoff; a runner loop restarts the ingest after hard failures. Court/division dictionary fetched separately (374 courts / 2,179 divisions).

2. **Cleaning & sectioning** — HTML parsed with a custom parser: `<h2>` headers become section boundaries; judge-panel tables become plain text lines; HTML comments dropped; whitespace normalized. Records without HTML headers (e.g. Constitutional Tribunal texts sourced from .doc files) are sectioned by keyword heuristics with a 100-char minimum section guard. `<span class="anon-block">` (court anonymization) → `[ANONIM]`.
3. **Validation / quarantine** — records with empty or <500-char source text, missing id, or judgment date outside 1918–2026 are diverted to a quarantine file with reasons; nothing is silently dropped.
4. **Deduplication** — exact stage: SHA-256 of normalized text (181 groups, 199 removed); near-duplicate stage: MinHash/LSH (datasketch), Jaccard threshold 0.9 over 5-word shingles (1,461 groups, 2,990 removed). The longest record of each group is kept; removed records are archived with provenance and similarity scores.
5. **Export** — single JSONL written to a temp file, validated line by line (parse + count), then atomically renamed.
6. **Quality control** — pattern-based residual-PII detection (PESEL/NIP checksums, ID cards, bank accounts, e-mails, phone numbers) with regression self-tests against legal-text false-positive classes (docket numbers, dates, amounts, statute citations, case-file page references, LEX citations, share/vote counts, DNA probabilities). Detected occurrences are masked with typed placeholders **iterated to a fixpoint** (masking one number can unshadow an adjacent one in enumerations/ranges); every replacement is logged with before/after context in an append-only audit log. **Release gate: a verification scan of the released file must report zero hits — v1.0 passed with 0 hits across 505,611 documents.** Corpus statistics report generated per release.

Design decisions

- **Near-duplicate removal is deliberate.** Courts issue batches of template decisions differing only in dates/docket numbers. For LM training, keeping dozens of ~identical texts skews the distribution, so groups above 0.9 Jaccard are collapsed to one representative. Removed records are preserved with provenance notes — reversible.
- **[ANONIM] is a preserved marker, not noise.** Polish courts anonymize published judgments; the placeholder keeps that signal explicit (`anon_count` lets you filter by anonymization density).
- **Empty records are quarantined, not deleted.** 5.3% of SAOS records carry metadata but no text (the share rises to ~23% for 2020+ common-court records). They are excluded from the corpus yet retained with reasons, so coverage gaps are auditable and texts can be backfilled from the Ministry of Justice portal (`source_ur1`).

Legal status

Under Article 4(2) of the Polish Copyright Act (Ustawa z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych, art. 4), official documents and official materials — including court judgments — are not subject to copyright. The `license` field records this basis as `public_domain_pl_art4_prawo_autorskie`.

Note that copyright status is separate from personal-data obligations: judgments are officially anonymized by courts before publication, and this pipeline additionally masks pattern-detectable residual PII, but pattern-based detection cannot guarantee completeness. Downstream users are responsible for GDPR compliance in their jurisdiction and use case.

Known limitations

- **Coverage skew.** 86% common courts; Supreme Court, KIO and Constitutional Tribunal are present but smaller; administrative courts (NSA/WSA) are absent (not covered by SAOS). Yearly coverage peaks in 2013–2017 and thins out both earlier (portal adoption) and later (publication lag) — this mirrors the source, not sampling.

- **court / division names only for common courts.** SN/TK/KIO records have `court: null` (their SAOS division ids are a different id space than the common-courts dictionary); the court class is always available in `court_type`.
- **~5.3% of source records have no text** and are quarantined; the share is concentrated in recent common-court records.
- **Source data errors.** OCR-era typos, malformed dates (3 quarantined), concatenated judge names. Textual errors are preserved as-is.
- **Residual PII.** All pattern-detected occurrences are masked and the release gate reports zero; PII in formats outside the pattern set (unformatted addresses, names missed by court anonymization, phone numbers inside long comma-separated enumerations rejected by the number-list guard) may remain.
- **Sectioning is heuristic** for records without HTML structure; section headers there depend on keyword matching.
- **Token counts are estimates** (chars/3.5), not tokenizer-specific.

Version history

- **v0.1** — 10k pilot (superseded, deleted)
- **v0.2** — 10k pilot + PII masking; schema included a redundant `text` field (superseded by v1.0)
- **v1.0** (2026-07-12) — full SAOS dump: 505,611 records, ~3.18 B tokens; sections-only schema; fixpoint PII masking; zero-hit release gate passed

Provenance & contact

Built with a Python pipeline (standard library + datasketch) from the public SAOS API. Pipeline code and architecture decisions are documented alongside the data (`PIPELINE.md`). Contact: licensing@aioil.ai.