Comprehensive Tool Evaluation Report

31 experiments across **30 tools** evaluating **25 extraction dimensions**.

---

1. Capacity Matrix

Which tool covers which extraction dimension.

Tool	File	PDF	Sect	Tbl	Meta	Sent	NER	R.NER	Role	Pred	Trip	Cons	SRL	CNL	QLng	Prod	HSC	Top	Sum	Sent	Disc	Emb	Code	Chnk	Rep	Runtime
CodeGraphContext	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	cpu
CoreNLP OpenIE	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cpu
CoreNLP constituency	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	cpu
Flair	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	gpu
GLiClass	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	gpu
GLiNER	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	gpu
Gensim LDA	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	cpu
HSC Signals	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	cpu
Ingest Heuristic Extractor	.	.	.	.	.	Y	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cpu
Ingest Metadata Extractor	.	Y	.	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cpu
LiteParse	Y	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cpu
Paper Parser (PyMuPDF)	Y	Y	.	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cpu
PyMuPDF	Y	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cpu
QHG Process Extraction	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	gpu

Tool	File	PDF	Sect	Tbl	Meta	Sent	NER	RelNER	Role	Pred	Trip	Cons	SRL	LCNL	QL	Lng	Prod	HSC	Top	Sum	Sent	Disc	Emb	Code	Chnk	Rep	Runtime
QNR2 CNL Parser	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.	.	cpu
Qualtron-4B (Mode A)	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.	gpu
Qualtron-4B (Mode B)	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.	gpu
Qualtron-4B (Mode C)	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.	gpu
SRL (spaCy dep)	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	cpu
Stanza	.	.	.	.	.	Y	.	.	.	.	Y	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	gpu
Sumy	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	.	.	.	.	.	cpu
Sym Pipeline (CPU)	.	Y	.	Y	Y	Y	.	Y	Y	Y	Y	.	Y	Y	.	.	.	.	.	Y	.	.	.	.	.	.	cpu
Sym Pipeline (GPU)	.	Y	.	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	.	.	Y	.	Y	Y	.	.	.	.	.	.	gpu
Tree-sitter (sym-tools)	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	.	cpu
Unstructured	Y	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	cpu
cagent Chunker	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	.	cpu
cagent file_parser	Y	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cpu
gitingest	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	Y	cpu
opendataloader-pdf	Y	Y	Y	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cpu
spaCy NER	.	.	.	.	Y	Y	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	cpu

Legend: Y = capable, . = not applicable

2. Speed Rankings (avg latency per dimension)

file_parse

Rank	Tool	Avg Latency (ms)	Runtime
1	PyMuPDF	23.1	cpu **FASTEST**
2	Paper Parser (PyMuPDF)	46.6	cpu
3	cagent file_parser	184.4	cpu
4	LiteParse	195.0	cpu
5	Unstructured	963.4	cpu
6	opendataloader-pdf	1165.3	cpu

pdf_parse

Rank	Tool	Avg Latency (ms)	Runtime
1	PyMuPDF	23.1	cpu **FASTEST**
2	Paper Parser (PyMuPDF)	46.6	cpu
3	cagent file_parser	184.4	cpu
4	LiteParse	195.0	cpu
5	opendataloader-pdf	1165.3	cpu

sections

Rank	Tool	Avg Latency (ms)	Runtime
1	Ingest MetadataExtractor	13.6	cpu **FASTEST**
2	Paper Parser (PyMuPDF)	46.6	cpu
3	LiteParse	195.0	cpu
4	Unstructured	963.4	cpu
5	opendataloader-pdf	1165.3	cpu
6	Sym Pipeline (CPU)	57728.0	cpu
7	Sym Pipeline (GPU)	58378.0	gpu

tables

Rank	Tool	Avg Latency (ms)	Runtime
1	opendataloader-pdf	1165.3	cpu **FASTEST**

metadata

Rank	Tool	Avg Latency (ms)	Runtime
1	Ingest MetadataExtractor	13.6	cpu **FASTEST**
2	Paper Parser (PyMuPDF)	46.6	cpu
3	Sym Pipeline (CPU)	57728.0	cpu

Rank	Tool	Avg Latency (ms)	Runtime
4	Sym Pipeline (GPU)	58378.0	gpu

sentence_split

Rank	Tool	Avg Latency (ms)	Runtime
1	spaCy NER	489.4	cpu **FASTEST**
2	Sym Pipeline (CPU)	57728.0	cpu
3	Sym Pipeline (GPU)	58378.0	gpu

ner

Rank	Tool	Avg Latency (ms)	Runtime
1	Ingest Heuristic Extractor	1.0	cpu **FASTEST**
2	Flair	158.2	gpu
3	spaCy NER	489.4	cpu
4	Stanza	671.7	gpu
5	Sym Pipeline (CPU)	57728.0	cpu
6	Sym Pipeline (GPU)	58378.0	gpu

ner_custom

Rank	Tool	Avg Latency (ms)	Runtime
1	GLiNER	318.2	gpu **FASTEST**
2	Sym Pipeline (GPU)	58378.0	gpu

role_classify

Rank	Tool	Avg Latency (ms)	Runtime
1	GLiClass	215.9	gpu **FASTEST**
2	Qualtron-4B (Mode B)	4531.1	gpu
3	Qualtron-4B (Mode C)	9880.4	gpu
4	Qualtron-4B (Mode A)	15766.4	gpu
5	Sym Pipeline (CPU)	57728.0	cpu
6	Sym Pipeline (GPU)	58378.0	gpu

predicates

Rank	Tool	Avg Latency (ms)	Runtime
1	Ingest Heuristic Extractor	1.0	cpu **FASTEST**

Rank	Tool	Avg Latency (ms)	Runtime
2	Sym Pipeline (CPU)	57728.0	cpu
3	Sym Pipeline (GPU)	58378.0	gpu

triples

Rank	Tool	Avg Latency (ms)	Runtime
1	CoreNLP OpenIE	847.0	cpu **FASTEST**
2	Sym Pipeline (CPU)	57728.0	cpu
3	Sym Pipeline (GPU)	58378.0	gpu

constituency

Rank	Tool	Avg Latency (ms)	Runtime
1	Stanza	671.7	gpu **FASTEST**
2	CoreNLP constituency	27641.0	cpu
3	Sym Pipeline (CPU)	57728.0	cpu
4	Sym Pipeline (GPU)	58378.0	gpu

srl

Rank	Tool	Avg Latency (ms)	Runtime
1	SRL (spaCy dep)	118.3	cpu **FASTEST**
2	Sym Pipeline (GPU)	58378.0	gpu

cnl_parse

Rank	Tool	Avg Latency (ms)	Runtime
1	QNR2 CNL Parser	2.0	cpu **FASTEST**
2	Sym Pipeline (CPU)	57728.0	cpu
3	Sym Pipeline (GPU)	58378.0	gpu

qlang_extract

Rank	Tool	Avg Latency (ms)	Runtime
1	Qualtron-4B (Mode B)	4531.1	gpu **FASTEST**
2	Qualtron-4B (Mode C)	9880.4	gpu
3	Qualtron-4B (Mode A)	15766.4	gpu
4	Sym Pipeline (CPU)	57728.0	cpu
5	Sym Pipeline (GPU)	58378.0	gpu

hsc_signals

Rank	Tool	Avg Latency (ms)	Runtime
1	HSC Signals	1.5	cpu **FASTEST**

topics

Rank	Tool	Avg Latency (ms)	Runtime
1	Gensim LDA	68.9	cpu **FASTEST**
2	Sym Pipeline (GPU)	58378.0	gpu

summary

Rank	Tool	Avg Latency (ms)	Runtime
1	Sumy	18688.4	cpu **FASTEST**

sentiment

Rank	Tool	Avg Latency (ms)	Runtime
1	Flair	158.2	gpu **FASTEST**
2	Stanza	671.7	gpu
3	Sym Pipeline (GPU)	58378.0	gpu

discourse

Rank	Tool	Avg Latency (ms)	Runtime
1	Sym Pipeline (CPU)	57728.0	cpu **FASTEST**
2	Sym Pipeline (GPU)	58378.0	gpu

code_ast

Rank	Tool	Avg Latency (ms)	Runtime
1	Tree-sitter (sym-tools)	30.5	cpu **FASTEST**
2	CodeGraphContext	672.8	cpu

chunking

Rank	Tool	Avg Latency (ms)	Runtime
1	cagent Chunker	24.1	cpu **FASTEST**

Rank	Tool	Avg Latency (ms)	Runtime
2	Unstructured	963.4	cpu

repo_digest

Rank	Tool	Avg Latency (ms)	Runtime
1	gitingest	95.0	cpu **FASTEST**

---

3. CPU vs GPU Comparison

Tool	Runtime	Layer	Dimensions
CodeGraphContext	cpu	python-sibling	code_ast
CoreNLP OpenIE	cpu	java-service	triples
CoreNLP constituency	cpu	java-service	constituency
Flair	gpu	python-service	ner, sentiment
GLiClass	gpu	python-service	role_classify
GLiNER	gpu	python-service	ner_custom
Gensim LDA	cpu	python-service	topics
HSC Signals	cpu	typescript	hsc_signals
Ingest Heuristic Extractor	cpu	typescript	ner, predicates
Ingest MetadataExtractor	cpu	typescript	metadata, sections
LiteParse	cpu	typescript	file_parse, pdf_parse, sections
Paper Parser (PyMuPDF)	cpu	python-service	file_parse, pdf_parse, sections, metadata
PyMuPDF	cpu	python-service	file_parse, pdf_parse
QHG Process Extraction	gpu	typescript	process_extract
QNR2 CNL Parser	cpu	typescript	cnl_parse
Qualtron-4B (Mode A)	gpu	sglang-llm	qlang_extract, role_classify
Qualtron-4B (Mode B)	gpu	sglang-llm	qlang_extract, role_classify
Qualtron-4B (Mode C)	gpu	sglang-llm	qlang_extract, role_classify
SRL (spaCy dep)	cpu	python-service	srl
Stanza	gpu	python-service	ner, constituency, sentiment
Sumy	cpu	python-service	summary

Tool	Runtime	Layer	Dimensions
Sym Pipeline (CPU)	cpu	typescript	sentence_split, ner, predicates, triples, constituency, role_classify, qlang_extract, cnl_parse, discourse, sections, metadata
Sym Pipeline (GPU)	gpu	typescript	sentence_split, ner, ner_custom, predicates, triples, constituency, role_classify, qlang_extract, cnl_parse, discourse, sections, metadata, srl, sentiment, topics
Tree-sitter (sym-tools)	cpu	python-service	code_ast
Unstructured	cpu	python-service	file_parse, sections, chunking
cagent Chunker	cpu	python-sibling	chunking
cagent file_parser	cpu	python-sibling	file_parse, pdf_parse
gitingest	cpu	python-sibling	repo_digest
opendataloader-pdf	cpu	java-cli	file_parse, pdf_parse, sections, tables
spaCy NER	cpu	python-service	ner, sentence_split

CPU-only tools: 21 (PyMuPDF, LiteParse, opendataloader-pdf, cagent file_parser, Unstructured…) **GPU-required tools**: 9 (GLiNER, Stanza, Flair, GLiClass, Sym Pipeline (GPU), QHG Process Extraction, Qualtron-4B (Mode A), Qualtron-4B (Mode B), Qualtron-4B (Mode C))

---

4. Coverage Analysis

Tool count per dimension

Dimension	Tool Count	Tools
file_parse	6	PyMuPDF, LiteParse, opendataloader-pdf…
pdf_parse	5	PyMuPDF, LiteParse, opendataloader-pdf…
sections	7	LiteParse, opendataloader-pdf, Unstructured…
tables	1	opendataloader-pdf
metadata	4	Paper Parser (PyMuPDF), Sym Pipeline (CPU), Sym Pipeline (GPU)…

Dimension	Tool Count	Tools
sentence_split	3	spaCy NER, Sym Pipeline (CPU), Sym Pipeline (GPU)
ner	6	spaCy NER, Stanza, Flair…
ner_custom	2	GLiNER, Sym Pipeline (GPU)
role_classify	6	GLiClass, Sym Pipeline (CPU), Sym Pipeline (GPU)…
predicates	3	Sym Pipeline (CPU), Sym Pipeline (GPU), Ingest Heuristic Extractor
triples	3	CoreNLP OpenIE, Sym Pipeline (CPU), Sym Pipeline (GPU)
constituency	4	Stanza, CoreNLP constituency, Sym Pipeline (CPU)…
srl	2	SRL (spaCy dep), Sym Pipeline (GPU)
cnl_parse	3	Sym Pipeline (CPU), Sym Pipeline (GPU), QNR2 CNL Parser
qlang_extract	5	Sym Pipeline (CPU), Sym Pipeline (GPU), Qualtron-4B (Mode A)…
process_extract	1	QHG Process Extraction
hsc_signals	1	HSC Signals
topics	2	Gensim LDA, Sym Pipeline (GPU)
summary	1	Sumy
sentiment	3	Stanza, Flair, Sym Pipeline (GPU)
discourse	2	Sym Pipeline (CPU), Sym Pipeline (GPU)
embeddings	0	
code_ast	2	Tree-sitter (sym-tools), CodeGraphContext
chunking	2	Unstructured, cagent Chunker

GAPS (no tool coverage): embeddings

Single-tool coverage (risk of single point of failure): - tables: only `opendataloader-pdf` - summary: only `Sumy` - hsc_signals: only `HSC Signals` - process_extract: only `QHG Process Extraction` - repo_digest: only `gitingest`

---

5. Recommended Pipeline Composition

Based on the evaluation results, the optimal pipeline uses the fastest tool per dimension:

Stage	Dimension	Recommended Tool	Latency (ms)	Runtime
1	file_parse	PyMuPDF	23.1	cpu
2	pdf_parse	PyMuPDF	23.1	cpu
3	sections	Ingest MetadataExtractor	13.6	cpu
4	tables	opendataloader-pdf	1165.3	cpu
5	metadata	Ingest MetadataExtractor	13.6	cpu
6	sentence_split	spaCy NER	489.4	cpu
7	ner	Ingest Heuristic Extractor	1.0	cpu
8	ner_custom	GLiNER	318.2	gpu
9	role_classify	GLiClass	215.9	gpu
10	predicates	Ingest Heuristic Extractor	1.0	cpu
11	triples	CoreNLP OpenIE	847.0	cpu
12	constituency	Stanza	671.7	gpu
13	srl	SRL (spaCy dep)	118.3	cpu
14	cnl_parse	QNR2 CNL Parser	2.0	cpu
15	qlang_extract	Qualtron-4B (Mode B)	4531.1	gpu
16	process_extract	*no tool*	-	-
17	hsc_signals	HSC Signals	1.5	cpu
18	topics	Gensim LDA	68.9	cpu
19	summary	Sumy	18688.4	cpu
20	sentiment	Flair	158.2	gpu
21	discourse	Sym Pipeline (CPU)	57728.0	cpu
22	embeddings	*no tool*	-	-
23	code_ast	Tree-sitter (sym-tools)	30.5	cpu
24	chunking	cagent Chunker	24.1	cpu
25	repo_digest	gitingest	95.0	cpu

Suggested Pipeline Flow

```
Input File
  |
  v
[1. File Parse] -- LiteParse (PDF) / cagent (other) / opendataloader (tables)
```

```
  |
  v
[2. Structure] -- Ingest MetadataExtractor + StructureParser
  |
  v
[3. Sentence Split] -- spaCy via sym-tools
  |
  v
[4. NLP Extraction (parallel)]
  |- spaCy NER (CPU)
  |- GLiNER custom NER (GPU)
  |- CoreNLP OpenIE triples (CPU)
  |- SRL frames (CPU)
  |- Stanza deps + sentiment (GPU)
  |
  v
[5. Role Classification] -- GLiClass (GPU) or Sym Pipeline role assigner
  |
  v
[6. QLang Assembly] -- Sym Pipeline assembleQLang()
  |
  v
[7. Post-processing]
  |- CNL generation + QNR2 parsing
  |- HSC intelligence signals
  |- Negation / discourse detection
  |
  v
[8. Optional: LLM Enhancement]
  |- Qualtron-4B single-sentence refinement
  |- GPT-5.2 for complex documents
  |- QHG process extraction (FSM/DFG/BPMN)
  |
  v
[9. Storage] -- embeddings + Supabase
```

---

6. Key Findings

PDF Parsing

- **LiteParse** (327ms for 50-page paper) vs **PyMuPDF** (72ms) vs **opendataloader-pdf** (520-1983ms)
- PyMuPDF fastest for raw text; opendataloader-pdf best for layout/tables; LiteParse best for spatial/OCR

NER

- **spaCy** (CPU, broad coverage, 836ms for full doc) — standard entity types
- **GLiNER** (GPU, zero-shot, 346ms for 20 sentences) — custom domain labels
- **Stanza** (GPU, 477ms for 10 sentences) — multilingual, dep parsing bonus
- **Flair** (GPU, 108ms NER + 42ms sentiment) — fastest GPU NER

QLang Extraction

- **Sym Pipeline** (CPU, 57s for 229 sentences) — deterministic, full coverage, no LLM needed
- **Qualtron-4B** (GPU, ~5-9s) — fast but role diversity issues (defaults to Claim)
- **GPT-5.2** (via Ingest pipeline) — highest quality, 1-3 LLM calls, slower

Code Analysis

- **Tree-sitter** (25ms) vs **CodeGraphContext** (650ms) — tree-sitter 25x faster for AST
- CodeGraphContext adds graph DB / symbol resolution not available in tree-sitter

CNL Parsing

- **QNR2** parser: 100% success rate on well-formed CNL (8/8 samples, <2ms)
- Deterministic, no LLM needed — ideal for rule round-tripping

HSC Intelligence Signals

- Detects conflicts, overlaps, redundancy purely from predicate structure
- No embeddings needed for basic signals — embedding-enhanced for relevance scoring