Ontology-Guided Triple Extraction from Academic Papers: A Comparative Evaluation of Large Language Models

Sethapong Samaruek[1], Patipon Wiangnak[2], Natthawut Kertkeidkachorn[2], Hideaki Takeda[3] and Rathachai Chawuthai[1,*]

[1]School of Engineering, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
[3]National Institute of Informatics, Tokyo, Japan

Abstract

Open Science requires research findings to be reusable by machines, yet the substance of a paper remains encoded in prose. Semantic Web representations offer a remedy: claims expressed as *(subject, predicate, object)* triples over a shared ontology are machine-readable and linkable across publications, supporting semantic search, comparison against prior results, and agents attempting reproduction. Large language models now generate such triples at negligible cost, relocating the difficulty from extraction to evaluation. This paper compares four language models performing ontology-guided triple extraction from academic papers. Six criteria are applied, covering the correctness, completeness and specificity of concepts, the correctness and completeness of relations, and semantic consistency; as they do not share a unit of analysis, they are judged at three levels by a language model under a fixed rubric. The models were run over 528 paragraphs from 20 papers, yielding 6,014 triples and 404 judgements. Relation correctness spans 41 to 86 points across the four models, against 24 for concept correctness, and no model records more than a third of the concepts in the paragraphs judged. At 40 items per cell, differences below 10 points are not distinguishable. Two limitations of the target representation that no criterion detects are also reported.

Keywords

Knowledge Graph, Ontology, Triple Extraction, Open Science, LLM-as-a-Judge

CEUR-WS.org/Vol-4316/rdm-full-4.pdf

1. Introduction

Open Science asks that research results be open not only to people but to machines, which means machine-readable, linked to other results, and so reusable. Publishing a PDF achieves none of the three, because a paper is prose and prose is where its findings stay.

A knowledge graph is the other thing a paper could be: a set of *(subject, predicate, object)* triples a machine can query, aggregate across thousands of papers and reason over. Expressed with Semantic Web vocabularies the triples are machine-readable, and the shared vocabulary is what links one paper's triples to another's. That link makes three applications possible, and we return to all three in Section 4.11: semantic search that retrieves by what a method did rather than by keyword, comparison of a new result against every earlier result on the same task, and agents that read a method as structured steps and try to reproduce it. Large language models have made the conversion look almost trivial: give a model a paragraph and a list of allowed relations, and triples come back.

The difficulty has moved rather than disappeared. An unconstrained model will emit fluent, plausible triples that the paragraph does not support, and it will invent a fresh relation name for every sentence it reads. Both failures make the resulting graph useless for the thing graphs are for, which is merging evidence across papers. Constraining the relation vocabulary with an ontology addresses the second failure. Measuring the first is what this paper is about.

IJCKG 2026 Posters, Demos and Workshops, November 19–21, 2026, Bangkok, Thailand
*Corresponding author.
66011216@kmitl.ac.th (S. Samaruek); w.patipon@jaist.ac.jp (P. Wiangnak); natt@jaist.ac.jp (N. Kertkeidkachorn); takeda@nii.ac.jp (H. Takeda); rathachai.ch@kmitl.ac.th (R. Chawuthai)
0009-0000-8412-514X (S. Samaruek); 0009-0007-2509-433X (P. Wiangnak); 0000-0003-4527-776X (N. Kertkeidkachorn); 0000-0002-2909-7163 (H. Takeda); 0000-0002-9303-4810 (R. Chawuthai)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-06

The standard measure is per-triple precision: show a judge the triple and the sentence it came from, ask whether the sentence supports it, report the rate. It is a good measure and we use it, but alone it misleads in a specific way. A triple can be perfectly supported by its sentence and still be worthless, because the sentence was an author biography or a publisher watermark. An extractor can score well on precision while ignoring nine tenths of the paper. Two triples can each be correct and contradict each other. A criterion that looks at one triple and one sentence sees none of this.

We therefore evaluated four extraction models under six criteria covering correctness, completeness, specificity and internal consistency. The criteria do not share a unit of analysis, which matters more than it sounds: a completeness question cannot be asked of a single triple, because what is missing is by definition not in front of the judge. We built three judging harnesses accordingly, at the triple, the paragraph and the paper level.

The contributions of this paper are:

1. An evaluation protocol in which each criterion is judged at the unit of analysis it can be answered at, with the stratified sampling and reweighting that design requires (Section 3.2).
2. A four-model, two-ontology comparison, the second ontology on the precision criteria only, in which the extraction model is the only variable because prompt, ontology, guards, deduplication and graph builder are shared code (Sections 3.1 and 3.3).
3. Findings on what separates extractors: typing rather than concept identification, sub-proportional completeness as volume rises, and a consistency effect that survives only at the extreme (Section 4).
4. A sample-size observation for others to check against their own work: at 40 judged items per cell, a fresh draw moved our scores by up to 10 points and reversed a close ranking (Section 4.9).
5. An account of what the criteria cannot see, from an expert review of our output: triples are strings rather than URIs, and measurement results arrive as fused literals for want of an n-ary construct. Both block reuse and both score well on every criterion (Section 4.11).

2. Literature Review

2.1. Knowledge Graphs from Scientific Text

Large scholarly graphs have been built by pipeline. The Semantic Scholar literature graph joins metadata, citations and extracted entities at scale [1], AI-KG derives millions of statements from paper abstracts with a stack of extraction tools [2], and the Open Research Knowledge Graph takes the opposite route, having people curate contributions by hand [3]. Supervised relation extraction over scientific text is its own line of work, with SciERC [4] providing typed relations and generative models such as REBEL [5] covering the general domain.

LLM prompting has largely replaced the supervised route for new projects [6, 7, 8]. Text2KGBench [9] is the closest published work in setting: it benchmarks ontology-driven graph generation from text against a reference graph. Ours is the case that benchmark cannot cover, because no reference graph exists for these papers and building one by hand would cost more than the extraction it checks. How to check output with nothing to compare it against is the gap this paper works in.

2.2. LLM-as-a-Judge Evaluation

Using a strong model as an evaluator is now routine [10], and so are its known weaknesses: position and verbosity biases [11], and a preference for a model's own output [12]. Self-preference is straightforward to design around and we do, since no model here judges its own extraction. Sending the rubric as a system prompt on every call, rather than running one long chat, addresses a third and more mundane problem, namely drift across a batch. We do not claim it addresses position bias. Verbosity bias we do not control at all, and it is a live threat to C6, where the judge sees a model's whole triple set and those sets differ fourfold in size (Section 4.12).

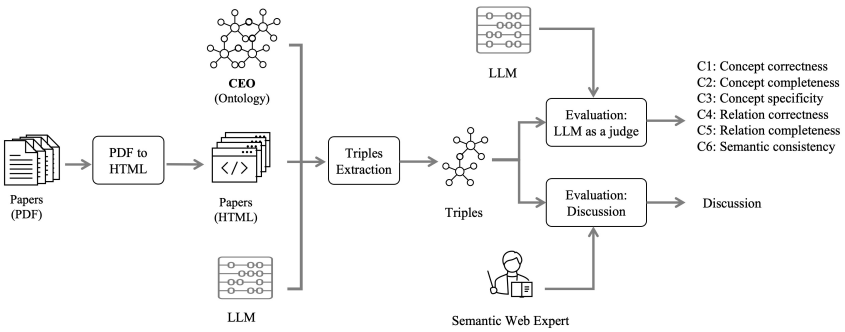


Figure 1: The pipeline and its two evaluation paths. A paper enters as a PDF and is converted to structured HTML. Each paragraph of that HTML is then given to a large language model together with an ontology, here CEO, which fixes the relation vocabulary and the domain and range of every relation. The model returns triples, which pass code-level guards before they enter the graph. The extracted triples are then evaluated twice over. A second language model judges them under the six criteria C1 to C6, at three different units of analysis. Separately, a Semantic Web researcher reviews the same triples and raises the representation questions discussed in Section 4.11. The scinex ontology runs through the identical path; only the box marked CEO changes.

2.3. Ontology and Graph Quality Criteria

Ontology evaluation has a long history of criteria-based assessment [13], and Linked Data quality has been surveyed in similar terms [14]. Paulheim [15] separates the two questions most extraction papers merge: whether what is in the graph is right, and whether what should be in it is there. Wilson et al. [16] give a conceptual model for ontology quality assessment, from which we take completeness and consistency. The correctness and specificity criteria come from a different kind of source. Zhang et al. [17] is an ontology *extraction* paper, not an ontology-quality one: it builds product ontologies of part-whole relations from e-commerce review text. We borrow three of its criteria and change what they are applied to, which is worth stating because the domains are far apart. Concept correctness and specificity transfer intact, since both ask about a concept and the sentence it came from and neither is specific to meronymy. Relation correctness does not: their setting has one relation type, so correctness is a question about whether a part-whole reading holds, whereas ours has 22 or 27 typed relations and correctness becomes domain and range conformance. Our C4 is that adapted version, and the adaptation is ours.

Our contribution is not the criteria themselves but the observation that they need three judging harnesses, and the measurements that follow.

3. Method

3.1. Extraction Pipeline

Figure 1 shows the whole pipeline, and the rest of this section walks through it left to right. Two features deserve comment. The language model appears twice in two roles, as extractor and as judge, and these are never the same model in our experiments. The two arrows leaving the triples are also not two versions of one evaluation: the upper path produces the six scores this paper reports, the lower is a human reading that no scoring rubric would have surfaced.

3.1.1. Document Parsing

Papers arrive as PDFs and are converted to sectioned HTML by a parser built on PyMuPDF and pdfplumber, with a structure builder that recovers headings and a multi-strategy table extractor. The stage is more load-bearing than it looks: several early results that read as “the model found nothing”

were parser bugs, including a heading detector that accepted only numbered headings and so discarded the body of every Nature-style paper, which produces a triple count near zero and looks exactly like an extraction failure. We now verify a parse by body word count. The parsed body is split into paragraphs, each becoming one extraction call carrying its section name.

3.1.2. Ontology-Constrained Extraction

The extractor takes a paragraph and an ontology and returns JSON. The system prompt lists every allowed relation with its domain and range constraint and one worked example each, running to about 12,000 characters. Four constraint regimes are implemented. `llm` leaves subject, relation and object free, `relation` draws the relation from the ontology, `fixed` adds a curated per-paper entity list for the subject, and `pair` takes subject and object from that list while leaving the relation free. Every measurement here uses `relation` mode. The gazetteer available to us, CS-NER [18], is annotated over computer science and NLP papers, while roughly a third of our papers are chemistry, materials science and clinical work, which get generic machine learning vocabulary and no domain terms. `fixed` mode would have failed on them for coverage reasons rather than quality reasons, and keeping subjects free avoids reporting one failure as the other.

3.1.3. Post-Parse Guards

The model's output is not trusted: every proposed triple passes code-level validation before it reaches the graph. The subject must appear literally in the source sentence. Direction guards catch reversed domain and range on the relations where inversion is common. Objects matching an ontology class name are rejected, as are cross-references such as *Table III*. A provenance filter rejects a triple whose source sentence is publisher, biographical or bibliographic matter, keyed on the sentence and not on the section name because front and back matter often has no heading of its own and running headers get glued into body paragraphs by the parser. IS-A relations are rejected because neither ontology declares a subsumption relation: type membership belongs to the schema layer, not to extracted instance data. That is not in tension with the taxonomy called for in Section 4.11, which is a controlled vocabulary maintained as part of the schema rather than something an extractor invents one sentence at a time.

Across the whole study the guards remove between 0.9% and 2.2% of each model's raw output. A late extension of the provenance filter removed 40 triples out of 6,054, and every one of the 40 came from the same model, which we return to in Section 4.10.

3.1.4. Core Ontologies

CEO, the Core Experiment Ontology, is an unpublished schema from a collaborating group with no public identifier we can cite, so its 22 relations are listed here and their domain and range constraints are reproduced in the extraction prompt. It covers bibliographic facts (`cites`, `publishedIn`, `writtenBy`), organisational facts (`affiliatedWith`, `employs`, `locatedIn`) and the experimental core (`uses`, `trainedOn`, `evaluatedOn`, `achieves`, `comparesAgainst`, `splitFrom`, `configures` and others). Each carries a domain to range constraint enforced twice, in the prompt and again in post-parse code: `addresses` is typed *ResearchProcess or Paper* to *ResearchContext*, so a task may never be its subject, because tasks are the research context and cannot address themselves.

scinex is a refined alternative in OWL. It keeps CEO's 22 relations and adds five for 27 in total, `achievesResult`, `producedBy`, `extractedFrom`, `mentions` and `writes`, with tightened domains and ranges throughout, aimed at CEO's known overloads: `achievesResult` is typed *Model* to *ExperimentalResult* where CEO's `achieves` accepts almost anything, and `producedBy` supplies the artefact-to-process direction that CEO's `produces` kept inverting.

Table 1
The six criteria, the unit each is judged at, and where each comes from.

	Criterion	Question	Unit	Ref.
C1	Concept correctness	Are subject and object real, and in the sentence?	triple	[17]
C2	Concept completeness	Are the paragraph's key concepts in the triples?	paragraph	[16]
C3	Concept specificity	Is the granularity right for a graph?	triple	[17]
C4	Relation correctness	Does the relation hold, in that direction?	triple	[17]
C5	Relation completeness	Are its stated relations all extracted?	paragraph	[16]
C6	Semantic consistency	Does the paper's triple set contradict itself?	paper	[16]

3.2. Evaluation Framework

3.2.1. Evaluation Criteria

Table 1 lists the criteria and their sources. Each is graded on three levels. C1, C2, C4, C5 and C6 are ordered, so the levels score 1.0, 0.5 and 0.0. C3 is not ordered, because *too broad* and *too narrow* are two ways of being wrong rather than a better and a worse one, so both score 0.0 while *appropriate* scores 1.0. A reported score is the mean over judged items as a percentage of the maximum. The rubric is in Section 3.2.4.

The decisive rule for C1, C3 and C4 is that the judge decides from the source sentence alone: a triple can be true about the paper and still count as incorrect here, because that sentence is the extractor's evidence and nothing else was in front of it. Showing the judge the predicate's domain and range is what makes C4 a test of typing rather than of whether an English relation name sounds plausible.

3.2.2. Units of Analysis

This split is the part of the design we would defend hardest: completeness cannot be judged from a triple, and consistency cannot be judged from a paragraph. Under a single per-triple rubric these questions are not merely hard to answer, they cannot be posed.

3.2.3. Sampling Design

The triple sampler draws round-robin over predicate buckets so rare relations appear at all, which makes the sampled rate unrepresentative of the full triple set by construction; we say so wherever we quote one. The paragraph sampler stratifies on whether the extractor produced anything, then reweights by each stratum's true share: the sparsest extractor leaves 84% of paragraphs empty, so a uniform draw spends most of the judging budget on empties and can miss the productive stratum entirely, and one seed drew zero productive paragraphs out of twenty. Empty paragraphs are not dropped, because an extractor that silently skips content-rich text is exactly what C2 and C5 exist to catch.

The paper sampler for C6 is a uniform seeded draw over the papers a model produced output for. Because seed and candidate list are the same for every model, the six papers judged are the *same* six for all four, which makes the C6 column like-for-like. It also means that a paper one model found nothing in is still judged for that model, and Section 4 shows this is not harmless. Every triple carries a content-hash identifier including the extractor family, so verdicts survive re-ingestion and a second labeller could later be joined on exactly the same items without re-drawing; this study has one (Section 3.2.4).

3.2.4. LLM as a Judge

All 404 judgements were made by one model, `claude-opus-5`, in a single session, with the rubric sent as the system prompt on every batch so that a long batch could not drift. No model judged its own extraction, the judge belonging to a different family than all four extractors. The rubric grades each criterion on the three levels of Section 3.2.1. On C1 and C4 the best level requires the source sentence to state the triple explicitly and, for C4, the argument types to match the declared domain and range. On

C2 and C5 the judge receives a paragraph together with every triple drawn from it and reports what was omitted. Those lists are the basis of the concept-coverage column in Table 2.

Single-rater judging is a deliberate choice and a real limitation, which we state plainly. There is no second labeller and so no inter-rater agreement figure for C1 to C6. An earlier round of this project, under a different rubric, measured agreement between two model judges at Cohen's κ [19] of 0.34 on three classes and 0.48 on a binary collapse for a weak extractor, rising to 0.46 and 0.64 for a strong one, fair to substantial on the usual reading [20]. Those figures measure a different rubric and do not transfer.

3.3. Experimental Setup

3.3.1. Corpus

Twenty papers, 528 paragraphs, all from one research group, spanning applied machine learning across ophthalmology, chemistry, materials science, traffic, cloud systems, computer vision, natural language processing, sensing, recommendation and linked data, published 2015 to 2026 in venues including *Scientific Reports*, *IEEE Access*, *Sensors*, *Applied Sciences* and *Measurement*. In publication order they are [21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40]. The domain spread is deliberate, so the evaluation is not measuring one narrow vocabulary. Institutional diversity is another matter, and we count it as a limitation.

3.3.2. Extraction Models

Four models were run through OpenRouter: `qwen3-235b` (a mixture-of-experts model with 235B total parameters), `gptoss-120b`, `gemma4-31b` and `ministral-14b`, each over all 528 paragraphs under both ontologies, giving eight runs and no failures. Three counts are easy to confuse. The *raw* output is what the models returned. The *post-guard* count, 6,054, is what survived the guards of Section 3.1.3 as they stood when the samples were first drawn. The *final* count, 6,014, survives the extended provenance filter of Section 4.9 and is the figure used throughout unless stated otherwise: 3,065 under CEO and 2,949 under scinex. Per-model volumes are in Table 2.

3.3.3. Experimental Control

All four extractors were called through OpenRouter with one client, one paragraph per request, temperature 0 and a 3,072-token cap, between 2026-09-03 and 2026-09-07. The identifiers sent on the wire were `openai/gpt-oss-120b`, `google/gemma-4-31b-it`, `mistralai/ministral-14b-2512` and `qwen/qwen3-235b`; the Gemma and Ministral entries are the instruction-tuned releases rather than their base or reasoning siblings, a distinction the short names conceal.

The extractor is split at the model boundary instead of being reimplemented per model: one script renders the real prompts, the model call is the only substituted step, and a second script feeds the replies back through the real parser, guards, deduplication and graph builder. Prompt, ontology, guards and output format are shared code, not copies, which licenses the claim that the extraction model is the only variable below. All eight runs are on one code version, re-ingested from stored replies after the last guard change.

4. Results and Discussion

4.1. Extraction Volume and Coverage

Table 2 gives the volume each model produced, and the spread is large. `ministral-14b` returns 4.2 times as many triples as `qwen3-235b` under CEO and fires on 4.3 times as many paragraphs, while `qwen3-235b` produces nothing at all for 5 of the 20 papers, which none of the others did. Whether that sparsity is caution or failure is what C2 and C5 are for.

Table 2
Extraction volume over 528 paragraphs, with two different coverage measures. *Paragraph coverage* is the share of paragraphs yielding at least one surviving triple, and is exact. *Concept coverage* is $|\text{extracted}|/(|\text{extracted}| + |\text{missed}|)$ over the 14 to 16 paragraphs sampled per model, computed from the missed-concept lists the judge returns, and carries the sampling caveat of Section 4.9.

	CEO				scinex	
Extractor	Triples	Per para.	Para. cov.	Concept cov.	Triples	Para. cov.
`qwen3-235b`	367	0.70	16.5%	23.0%	563	23.5%
`gptoss-120b`	547	1.04	43.4%	18.0%	323	25.8%
`gemma4-31b`	615	1.16	50.6%	16.7%	525	43.6%
`ministral-14b`	1,536	2.91	71.2%	32.5%	1,538	73.1%

Table 3
Triples produced by `gemma4-31b` under CEO, chosen to span the predicates the ontology uses most. The last two rows are included because they are wrong in ways discussed later: the object of the `achieves` row fuses a value with its unit, and the `configures` row is inverted with respect to the relation's declared direction.

Subject	Predicate	Object	Paper
Deep Learning	`encompasses`	object detection	`parkingyolo2023`
YOLOv8	`uses`	C2f module	`parkingyolo2023`
C2f module	`comprises`	Conv Modules	`parkingyolo2023`
OC-SORT	`evaluatedOn`	MOT17	`parkingyolo2023`
RT-DETR	`trainedOn`	eye landmark dataset	`strabismus2026`
training set	`splitFrom`	TTR dataset	`traveltime2022`
Random Forest classifier	`achieves`	accuracy of 0.95	`strabismus2026`
Random Forest	`configures`	30 estimators	`strabismus2026`

Table 4
C1 to C6 under the CEO ontology. Rows ordered by C4. C1, C3 and C4 rest on $n = 40$ triples per model, C2 and C5 on 16, 14, 16 and 15 paragraphs in row order, and C6 on the same 6 papers for every model. Triples and paragraphs are drawn from all 20 papers, C6 is not. Wilson 95% intervals are given for C1 and C4, the two criteria the paper's claims lean on. All values are percentages.

Extractor	Triples	C1 [95% CI]	C2	C3	C4 [95% CI]	C5	C6
`gptoss-120b`	547	88.8 [75.4, 95.3]	34.4	80.0	**86.2** [72.4, 93.8]	31.2	25.0
`gemma4-31b`	615	**93.8** [81.8, 98.0]	28.6	80.0	73.8 [58.5, 84.9]	21.4	33.3
`qwen3-235b`	367	75.0 [59.8, 85.8]	31.2	60.0	47.5 [32.9, 62.5]	28.1	**50.0**
`ministral-14b`	1,536	70.0 [54.6, 81.9]	26.7	52.5	41.2 [27.4, 56.6]	**30.0**	8.3

4.2. Extracted Triples

Table 3 shows what the extraction produces, taken from `gemma4-31b`, the model we end up recommending (Section 4.4). Each row comes from one sentence of the paper named on the right. *(C2f module, comprises, Conv Modules)*, for instance, comes from "The C2f module comprises two Conv Modules and n BottleNeck". Read as a graph, the first four rows do something a keyword index cannot, placing a module inside a model inside a task inside a field, with the same nodes recurring across papers. The last two rows are the failure modes that matter most for reuse: *accuracy of 0.95* is one string where a graph needs a metric, a number and a unit, and `configures` is declared *ExperimentalSpecification* to *Experiment*, so *(Random Forest, configures, 30 estimators)* points the wrong way. Sections 4.10 and 4.11 take both up.

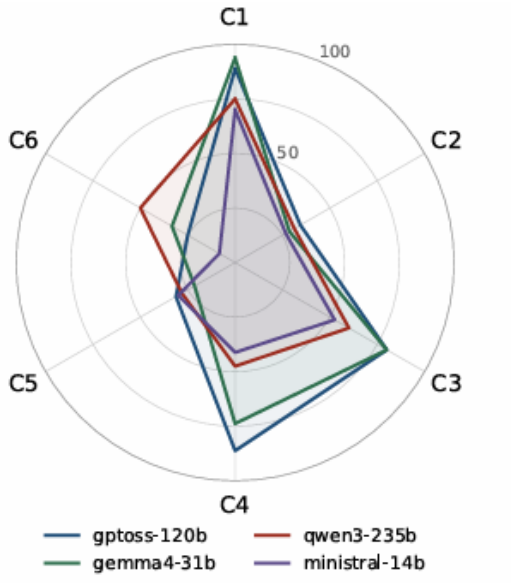


Figure 2: The six scores of Table 4, one polygon per extractor, axes running from 0 to 100. Every polygon is pinched along C2 and C5, the two completeness criteria, and stretched along C1 and C3. No extractor is strong all the way round, and the differences between them lie almost entirely on C4.

4.3. Criterion Scores under CEO

Table 4 is the main result. Before reading rankings out of it, see Section 4.9: at these sample sizes a gap under about 10 points is not decidable. C4 has a spread of 45 points and survived a full redraw unchanged, so its ordering is safe to quote. C1 spans 24 points, and its top pair reversed under a redraw, so its ordering is not.

Figure 2 plots the same numbers as a radar chart, where the shared shape of the four polygons carries more than any single vertex. All four are pulled outward on C1 and C3 and pinched inward on C2 and C5. Whatever separates these models, none is close to recording what a paper actually contains.

4.4. Model Selection

Read as an engineering decision instead of a ranking, the table points at `gemma4-31b`. It scores within the noise band of `gptoss-120b` on five of the six criteria, ahead on two of those five, at roughly a quarter of the parameter count. For a group running extraction over thousands of papers on its own hardware, that is the difference between a job that fits on one GPU and one that does not.

The qualification is C4, where a real gap appears: 86.2% against 73.8%, or 12.4 points. It is the largest gap between these two models, it survived the redraw, and relation typing matters most for merging graphs across papers. It is also not established: the Wilson intervals overlap heavily, [72.4, 93.8] against [58.5, 84.9] (Table 4), so at $n = 40$ this is a direction worth acting on, not a measured difference. Our reading is that the size difference earns the trade for most purposes, and that a group prioritising typed correctness should pay for the larger model and judge more than 40 triples first.

Neither model is good enough on the criteria where every model is weak. `gemma4-31b` scores 28.6% on C2 and 21.4% on C5, the lowest of the four on both, and lifting those two numbers is the work we would prioritise next. Only concept identification approaches saturation, C1 standing at 88.8% and 93.8% for these two models. Relation typing does not, at 86.2% and 73.8%, and completeness is not close for any model in the study.

4.5. Concept Identification versus Relation Typing

C1 runs from 70.0% to 93.8%, a factor of 1.3. C4 runs from 41.2% to 86.2%, a factor of 2.1. The concepts these models pull out of a sentence are mostly real and mostly present in it, whatever the model. What separates a good extractor from a weak one is whether the relation it picks respects that relation's domain and range. Under a single correct-or-not rubric these were one number and the split was invisible, which matters practically: the failures that most damage a merged graph are typing failures, and a wrongly typed triple reads as fluent English about the right two concepts.

4.6. Volume and Completeness

A C2 score is the mean of a three-level grade, so it cannot be read as the fraction of a paper's concepts that were captured: a model graded *partial* on every paragraph scores 50% while saying nothing about how much it missed. The judge returns the concepts it considered missed, though, so the fraction can be computed directly. It appears as the concept-coverage column of Table 2.

Measured this way, `ministral-14b` captures 32.5% of the important concepts in its sampled paragraphs against `qwen3-235b`'s 23.0%. So 4.2 times the triples and 4.3 times the paragraph coverage buy about 1.4 times the concept coverage: sub-proportional by a wide margin rather than absent. The extra volume goes mostly on restating what was already captured, enumerating pairwise combinations of things already linked, and mining boilerplate.

Two caveats. The direct count and the C2 rubric disagree in sign for this pair. C2 puts `ministral-14b` below `qwen3-235b` (26.7% against 31.2%) where coverage puts it above, because the rubric penalises a *poor* grade on a content-rich paragraph regardless of what was caught elsewhere. We report both and prefer the count. And every model leaves between two thirds and five sixths of the important concepts unrecorded, which is the more robust half of the result because it does not depend on comparing models.

An earlier draft added that the sparse and the dense extractor leave *different* thirds. That is withdrawn. Each model has its own paragraph draw, five of the six pairs share no sampled paragraph, and the one pair sharing two paragraphs missed exactly the same concepts on both.

4.7. Semantic Consistency

C6 runs 50.0%, 25.0%, 33.3% and 8.3% for the four models ordered by volume, a Spearman ρ of -0.80. An earlier draft reported that as a directional result. It does not survive the obvious control, and the correction is more informative than the original claim.

The same six papers were judged for every model (Section 3.2.3), so the comparison is fair. Two of those six are papers from which `qwen3-235b` extracted nothing, and an empty triple set cannot contradict itself, so both scored *consistent*, vacuously. Removing them drops `qwen3-235b` from 50.0% to 25.0%, level with `gptoss-120b` and below `gemma4-31b`, and ρ falls to -0.25. With four extractors, $\rho = -0.80$ carries a p of about 0.33 in any case.

What survives is narrower and we state only that. The densest extractor is the least internally consistent by a wide margin: `ministral-14b` is graded *contradictory* on five of its six papers and scores 8.3%, against 25.0 to 33.3% for the other three, whose volumes differ by a factor of two and a half without separating them. Volume at the extreme appears to cost coherence. Across the middle of the range it orders nothing. The mechanism we had offered, that more triples means more chances to conflict, is true by construction. What needs measuring is whether contradictions rise faster than volume, and a per-paper grade cannot answer that.

Three conflict types recur. Superlatives are extracted with the scope qualifier stripped, so a best result is asserted of three different methods. Metric nodes are bound to two values because the class label was dropped. Surface variants of one entity are split across nodes, which lets incompatible claims attach to what should be a single thing. C6 must never be reported without C2 and C5 beside it. It rewards extracting nothing, rests on six papers per model, and the vacuity above moved its headline by 25 points.

4.8. Ontology Comparison

Under scinex the same four models score C1/C3/C4 of 86.25/72.50/90.00 (`gptoss-120b`), 88.46/79.49/58.97 (`gemma4-31b`), 72.50/57.50/46.25 (`qwen3-235b`) and 72.50/60.00/46.25 (`ministral-14b`), on $n = 40$ per cell except `gemma4-31b` at 39. Against their CEO scores that is a change of at most ± 7.5 points everywhere except `gemma4-31b`'s C4, which falls 14.78. Volumes move more than scores, the scinex/CEO ratios being 0.59, 0.85, 1.53 and 1.00 in the same order, and those counts are exact and not sampled. Every delta but one sits inside the noise band described below. The honest summary is that no consistent ontology effect was detectable at this sample size, which claims less than calling the effect model-specific.

What the richer schema does is visible case by case, not in the aggregate. Where CEO forced `employs`, typed organisation to person, onto a model consuming an architectural block, scinex offered `uses` and the triple became correct; `splitFrom` came out inverted under CEO and correct under scinex for the same model and sentence. Against that, scinex adds failure modes CEO cannot have: `extractedFrom`, whose range is *AcademicPaper*, was misused five times, pointed at a dataset, an architecture component, OpenStreetMap, a vehicle fleet and a data-availability statement, because its name reads generically enough that the English always seems to fit. The refinements also pay only when a model reaches for them, and `gemma4-31b` kept using plain `achieves` for metric outcomes, losing `achievesResult` entirely.

4.9. Sampling Noise

Late in the study we extended the provenance filter and re-ingested all eight runs, removing 40 triples of the 6,054 that had survived the guards, or 0.7%. Disturbing the pool reshuffles a round-robin draw, so the samples were redrawn and re-judged at $n = 40$, against $n = 39$, 39, 37 and 39 before. C1 moved $79.5 \to 75.0$, $89.7 \to 88.8$, $85.1 \to 93.8$ and $65.4 \to 70.0$ for `qwen3-235b`, `gptoss-120b`, `gemma4-31b` and `ministral-14b`; C3 and C4 moved comparably, the largest single change being `gemma4-31b`'s C4 at $63.5 \to 73.8$. Movement reached 10.3 points and the C1 ranking reversed at the top.

We had predicted movement of about 4.5 points and no change in ordering, and we were wrong. The reason is not the filter: removing 0.7% of a pool cannot move a score ten points. It is that the two draws barely overlap, sharing 6 to 14 items out of roughly 40, so they behave like two independent samples of the same pool and the movement is ordinary sampling variance.

The arithmetic agrees, which turns one redraw into a rule. Treating a cell as a proportion at $n = 40$, the standard error $\sqrt{p(1-p)/n}$ is 6.3 points at $p = 0.8$ and 7.9 at $p = 0.5$; the difference between two near-independent draws carries $\sqrt{2}$ times that, so 8.9 and 11.2. An observed maximum movement of 10.3 points sits inside that, making the redraw a demonstration of a quantity we could have written down rather than a surprise.

The Wilson intervals in Table 4 should be read before any pair of models is compared. They are 16 to 30 points wide. On C4 the extremes separate cleanly, `gptoss-120b` at [72.4, 93.8] against `ministral-14b` at [27.4, 56.6], which is what licenses the claim in Section 4.5. No adjacent pair on either criterion separates, including the `gptoss-120b`/`gemma4-31b` gap on C4 that Section 4.4 treats as the one real trade.

Two caveats on the arithmetic. C1, C2, C4, C5 and C6 are graded on three levels and not two, so a binomial standard error is only an approximation for them. For C3, which is binary, it is exact, but its variance is not comparable with the others. And an interval of this kind describes sampling from the judged pool, not disagreement between judges, which this study does not measure. The practical rule, which we would suggest to anyone running an evaluation of this shape, is that at 40 items per cell a gap under about 10 points is not distinguishable. Raising n to 120 would halve the intervals, and costs judging time only.

4.10. Error Analysis

Boilerplate is mined as content, and the per-triple criteria cannot see it. Triples scoring at the best level on C1, C3 and C4 together were drawn from author biographies, an IEEE download watermark, bibliographies, acknowledgements and a classification block. *(SOKENDAI) locatedIn (Japan)* is a correct triple about a correct sentence with no business in a graph of the paper's content: the criteria judge a triple against its source sentence and never ask whether that sentence belonged to the paper. The provenance filter now handles this, and it is the clearest case we found of a per-triple criterion satisfied by something worthless. The behaviour is not general, though: the final filter extension removed 40 triples and all 40 were `ministral-14b`'s, 12 under CEO and 28 under scinex, the other three models having declined that material unprompted.

A relation direction that every model gets wrong. `configures` is defined *ExperimentalSpecification* to *Experiment*. Across the whole study, exactly three correctly directed instances exist, all from `ministral-14b`, for example *(K-Fold Cross-Validation) configures (each experiment)*. Every other instance from every model has the form *Model* to *setting-value*, as in *(Random Forest) configures (30 estimators)*. No guard can repair this, because it is not a simple inversion: the object is a value rather than an experiment, so swapping the arguments yields nothing valid either.

An earlier draft read this as four independent models converging on one shape, and so as evidence the schema is wrong. That does not hold. The four models share one system prompt, and a shared prompt is a common cause. They all saw the line

> `configures` (ExperimentalSpecification → Experiment | use for explicit hyperparameter or setup relationships)

which cuts both ways. The prompt states the correct direction explicitly and supplies no inverted example, so the models had the declared domain and range in front of them and wrote the reverse anyway. But the usage hint, *use for explicit hyperparameter or setup relationships*, names the hyperparameter first and may be doing the damage by itself. We cannot separate the two explanations from the runs we have. The test is cheap and unrun: re-extract with that hint reworded, and see whether the inversion persists. Until then we record the observation and the confound together, and leave the ontology unchanged.

Inversions and contradictions. `splitFrom` is typed *Dataset* to *Dataset*, so both sides pass any type check, yet three of the four models wrote it backwards consistently and inverted the partition hierarchy. Our guard keys on which side names a partition, and misses the case where the subject is the partition word, as in *(training set) splitFrom (five segments)*. The clearest contradiction found is `gemma4-31b` asserting both *(VGG16) achieves (F1 w-avg 0.23)* and *(VGG16) achieves (over 90% accuracy)* in one paper, and the clearest structural violation is `ministral-14b`'s *(route prediction model) comprises (route prediction model)*, a node containing itself.

4.11. Towards Reusable Linked Data

The evaluation above asks whether each triple is right. A Semantic Web researcher who reviewed our output asked whether the triples are usable as linked data, and the answer is that they are not yet. The distinction matters for the Open Science motivation of Section 1: a correct triple that no other dataset can join to is machine-readable but not reusable. What follows are limits of the target representation rather than of the extractors, and none of the six criteria can see them.

4.11.1. Benefits of the Ontology Constraint

The ontology constraint visibly improves how the triples are structured: relation names come from a fixed vocabulary instead of being invented per sentence, so triples from different papers share a predicate space, and each relation carries a domain and range, giving a triple a declared shape that code can check. That is what makes C4 answerable at all, and none of what follows undoes it.

4.11.2. Absence of a URI Layer

Every element of every triple is a literal string. A representative example from `traveltime2022` is *(SA-LSTM, uses, Butterworth low-pass filter)*. Nothing in it resolves, so two graphs from two papers cannot merge on a shared identifier. They merge only where their strings coincide, and mostly they do not.

The scale is visible inside one paper. In `traveltime2022`, `gemma4-31b` denotes one model with six surface forms: *SA-LSTM*, *SA_LSTM*, *Self-Attention LSTM*, *self-attention LSTM*, *self-attention LSTM (SA-LSTM) model* and *Self-Attention Long Short-Term Memory (SA-LSTM) model*. All six are correct readings of their own sentences, so all six pass C1 and the extractor is not at fault. They are six nodes where the graph needs one.

What is missing is a post-processing stage that assigns URIs, and behind it a taxonomy deep enough for known terms to be looked up instead of guessed at. This connects to a result above. Surface-variant splitting is one mechanism behind low C6 scores, so C6 partly measures the absence of the identifier layer and not a defect in the extraction.

4.11.3. Non-Atomic Measurement Values

Our objects are sometimes not single values. Take *(LSTM, achievesResult, "MAE of 27.12 min per 100 km")*, again from `traveltime2022`: that object packs the metric, the number and the unit into one slot, so no query can ask for the value alone, compare it against another paper's MAE, or convert the unit. The packing is not even consistent within one paper. The same measurement also appears as *(LSTM, achievesResult, "error of ±27.12 min per 100 km")*, dropping the metric name, and as *(MAE, evaluates, "27.12 min per 100 km")*, separating metric from model but still fusing number with unit. Three triples, one fact, three decompositions.

The standard remedy is an n-ary construct, a reified measurement node linking the model, the metric, the value, the unit and the conditions it was taken under. Neither CEO nor scinex offers one, so the extractors had no way to produce it and the guards nothing to enforce. This also accounts for a failure mode listed earlier without explanation. Metric nodes bound to two values, counted as C6 contradictions, are often two measurements the binary form cannot keep apart.

4.11.4. Implications for Schema and Criteria

Two lines of work follow. The schema needs n-ary support for measurement results and experimental conditions, which is a modelling change and not a prompting one. The vocabulary needs an extensible taxonomy covering the domains in scope, because URI assignment is only as good as the term list behind it. Both are prerequisites for the three applications of Section 1 rather than refinements of them. Semantic search needs the same method to be one node, not six strings, or a query returns a sixth of the evidence. Comparison against earlier results needs value, metric and unit in separate slots, or two papers' numbers cannot go on one axis. A reproducing agent needs steps and settings typed well enough to execute. Our triples are machine-readable and, for all three, not yet linked or reusable.

The consequence for the evaluation is uncomfortable. Under C1 and C3 as written, *"MAE of 27.12 min per 100 km"* is a good object: specific, unambiguous and present in the source sentence, so it scores at the best level on both. The criteria cannot express that it should have been three nodes. Atomicity of values and resolvable identity of concepts fall outside the six criteria, and a framework judging graphs intended as linked data would need them.

4.12. Limitations

1. **Sample size sizes the bands, not the cells.** Roughly 40 triples, 15 paragraphs and 6 papers per model; Section 4.9 quantifies the cost, and no point estimate here should be read to a decimal place.

2. **Single rater, by design.** One model judged all 404 items, so there is no agreement statistic for C1 to C6. Criteria such as C4's domain and range test involve genuine judgement calls whose reproducibility across raters is unmeasured here.
3. **C6 rewards an empty graph**, rests on six papers per model, and must never be reported without C2 and C5 beside it. Completeness and consistency were judged for CEO only, so the ontology comparison is a precision comparison.
4. **One research group.** Domain diversity is real, institutional diversity is not, and writing conventions travel with an institution. The triple sampler also over-represents rare predicates by design, so these are sampled rates, not rates weighted by each predicate's true share.
5. **Verbosity bias is uncontrolled.** C6 shows the judge a model's whole triple set, and those sets differ fourfold in size across the models compared. A judge grading longer input more harshly, or more leniently, would produce the pattern we observe, and nothing here separates the two.
6. **Completeness is measured on per-model paragraph draws**, so the coverage figures in Table 2 are within-model measurements and the cross-model comparison is confounded with the draw. A shared paragraph sample would fix this and was not taken.
7. **The output is not linked data.** Nodes are strings and not URIs, and measurement results are stored as fused literals for want of an n-ary construct. Section 4.11 works through both, and neither is visible to any of the six criteria.

5. Conclusion

We evaluated four LLM extractors of ontology-guided triples under six criteria, judged at three units of analysis, with the extraction model as the only variable. Models separate on typing far more than on concept identification, so the errors that most damage a merged graph are the ones a precision figure is least able to show. Volume buys completeness sub-proportionally. The densest extractor covered 4.3 times as many paragraphs as the sparsest and captured about 1.4 times the concepts, and every model left two thirds or more of each paper unrecorded.

The methodological result may be the more useful one. At about 40 judged items per cell, redrawing the sample moved scores by up to 10 points and reversed a close ranking, which the binomial arithmetic predicts. Evaluations at that scale are common, our own earlier rounds included, and a comparison drawn from one may be less stable than it looks. Our next round raises n to roughly 120, runs the completeness and consistency harnesses for scinex, and batches the remaining guard gaps into one round instead of patching mid-measurement, which is what forced the redraw described here.

On the representation side, Section 4.11 sets out two changes that have to happen before this output is usable as linked data: a URI layer backed by an extensible taxonomy of domain vocabulary, and n-ary constructs in the schema so that a measurement can hold its metric, value, unit and conditions in separate slots. Both are modelling work rather than prompting work, and both would need the evaluation framework extended to see them.

The wider goal is the one stated at the outset. Open Science needs a paper's contents machine-readable, linked across papers, and therefore reusable by the search, comparison and reproduction tools built on top. Ontology-guided extraction with a language model delivers the first of those three today, at the quality this paper measures. It does not yet deliver the second, and the second is what the third depends on. Saying which of the three is done, and what the remaining two cost, seems to us more useful than another precision figure.

Acknowledgment

The research was supported by ROIS NII Open Collaborative Research 2026 (261S03-24185).

Declaration on Generative AI

During the preparation of this work, the authors used Claude Opus for grammar checking, paraphrasing, and rewording. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] W. Ammar, et al., Construction of the literature graph in semantic scholar, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Industry Papers, 2018, pp. 84–91. doi:10.18653/v1/N18-3011.

[2] D. Dessì, et al., AI-KG: An automatically generated knowledge graph of artificial intelligence, in: The Semantic Web – ISWC 2020, Lecture Notes in Computer Science, Springer, 2020, pp. 127–143. doi:10.1007/978-3-030-62466-8_9.

[3] M. Y. Jaradeh, et al., Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge, in: Proceedings of the 10th International Conference on Knowledge Capture (K-CAP), ACM, 2019, pp. 243–250. doi:10.1145/3360901.3364435.

[4] Y. Luan, et al., Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018, pp. 3219–3232. doi:10.18653/v1/D18-1360.

[5] P.-L. Huguet Cabot, R. Navigli, REBEL: Relation extraction by end-to-end language generation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 2370–2381. doi:10.18653/v1/2021.findings-emnlp.204.

[6] S. Carta, et al., Iterative zero-shot LLM prompting for knowledge graph construction, arXiv (2023). ArXiv:2307.01128.

[7] Y. Zhu, et al., LLMs for knowledge graph construction and reasoning: Recent capabilities and future opportunities, World Wide Web 27 (2024). doi:10.1007/s11280-024-01297-w.

[8] M. Hofer, et al., Construction of knowledge graphs: Current state and challenges, Information 15 (2024) 509. doi:10.3390/info15080509.

[9] N. Mihindukulasooriya, et al., Text2KGBench: A benchmark for ontology-driven knowledge graph generation from text, in: The Semantic Web – ISWC 2023, Lecture Notes in Computer Science, Springer, 2023, pp. 247–265. doi:10.1007/978-3-031-47243-5_14.

[10] L. Zheng, et al., Judging LLM-as-a-judge with MT-Bench and chatbot arena, in: Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Datasets and Benchmarks Track, 2023. ArXiv:2306.05685.

[11] P. Wang, et al., Large language models are not fair evaluators, arXiv (2023). ArXiv:2305.17926.

[12] A. Panickssery, S. R. Bowman, S. Feng, LLM evaluators recognize and favor their own generations, in: Advances in Neural Information Processing Systems 37 (NeurIPS 2024), 2024. ArXiv:2404.13076.

[13] J. Brank, M. Grobelnik, D. Mladenić, A survey of ontology evaluation techniques, in: Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD), 2005. URL: https://aile3.ijs.si/dunja/SiKDD2005/Papers/BrankEvaluationSiKDD2005.pdf.

[14] A. Zaveri, et al., Quality assessment for linked data: A survey, Semantic Web 7 (2016) 63–93. doi:10.3233/SW-150175.

[15] H. Paulheim, Knowledge graph refinement: A survey of approaches and evaluation methods, Semantic Web 8 (2017) 489–508. doi:10.3233/SW-160218.

[16] R. S. I. Wilson, et al., A conceptual model for ontology quality assessment: A systematic review, Semantic Web 14 (2023) 1051–1097. doi:10.3233/SW-233393.

[17] D. Zhang, S. Conia, A. Rago, Meronymic ontology extraction via large language models, in: Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics

(IJCNLP-AACL 2025), Mumbai, India, 2025, pp. 423–433. ACL Anthology 2025.ijcnlp-short.34; arXiv:2510.13839.

[18] J. D'Souza, S. Auer, Computer science named entity recognition in the open research knowledge graph, in: From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries (ICADL 2022), Lecture Notes in Computer Science, Springer, 2022. doi:10.1007/978-3-031-21756-2_3.

[19] J. Cohen, A coefficient of agreement for nominal scales, Educational and Psychological Measurement 20 (1960) 37–46. doi:10.1177/001316446002000104.

[20] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174. doi:10.2307/2529310.

[21] R. Chawuthai, et al., A machine learning approach for predicting osmotic coefficients and deriving activity coefficients in alkyl ammonium salts, Scientific Reports (2026). doi:10.1038/s41598-026-36758-x.

[22] R. Chawuthai, et al., Utilizing deep learning from mobile phone photos for early detection of horizontal strabismus: a screening approach, Scientific Reports (2026). doi:10.1038/s41598-026-48893-6.

[23] R. Chawuthai, et al., Integration of genetic algorithm with machine learning for properties prediction, Chemical Engineering Transactions (2025). doi:10.3303/CET25117170.

[24] R. Chawuthai, et al., Assessing the effects of corrupted parameters in a large language model: A case study of LLAMA 3.2 1b, in: ASSE 2025, ACM, 2025. doi:10.1145/3775030.3775040.

[25] R. Chawuthai, et al., Novel method for predicting the cracks of oxide scales during high temperature oxidation of metals and alloys by using machine learning, Scientific Reports (2025). doi:10.1038/s41598-025-91449-3.

[26] R. Chawuthai, et al., Development of machine learning enhanced low-cost spectrophotometer for pesticide prediction, Measurement (2025). doi:10.1016/j.measurement.2025.114890.

[27] W. Puisamlee, R. Chawuthai, Minimizing model size of CNN-based vehicle make recognition for frontal vehicle images, IEEE Access (2025). doi:10.1109/ACCESS.2025.3574187.

[28] C. Chungmarisakul, R. Chawuthai, A comparative study of video segmentation techniques for graduate detection in thai graduation ceremonies, in: ECTI-CON 2025, 2025. doi:10.1109/ECTI-CON64996.2025.11100560.

[29] N. Chairatana, R. Chawuthai, Stateless system performance prediction and health assessment in cloud environments: Introducing cSysGuard, an ensemble modeling approach, IEEE Access (2024). doi:10.1109/ACCESS.2024.3406670.

[30] C. Boonnithititikul, T. Jaknamon, R. Chawuthai, Eye landmarks detection using RT-DETR with rules, in: ECTI-CON 2024, 2024. doi:10.1109/ECTI-CON60892.2024.10594871.

[31] N. Treesoonrat, et al., U-GMo: Individual clip detection from a graduation ceremony video, in: ITC-CSCC 2024, 2024. doi:10.1109/ITC-CSCC62988.2024.10628160.

[32] N. Sharma, et al., Parking time violation tracking using YOLOv8 and tracking algorithms, Sensors (2023). doi:10.3390/s23135843.

[33] T. Raksachat, R. Chawuthai, Improving a text classifier using text augmentation: road traffic content from Twitter, in: ECTI-CON 2023, 2023. doi:10.1109/ECTI-CON58255.2023.10153191.

[34] R. Chawuthai, et al., Travel time prediction on long-distance road segments in Thailand, Applied Sciences (2022). doi:10.3390/app12115681.

[35] R. Chawuthai, K. Pruekwangkhao, T. Threepak, Spatial-temporal traffic speed prediction on Thailand roads, in: ICEAST 2021, 2021. doi:10.1109/ICEAST52143.2021.9426257.

[36] R. Chawuthai, P. Omarak, V. Thaiyingsombat, A recommender system for trip planners, in: Intelligent Information and Database Systems (ACIIDS 2020), Springer, 2020. doi:10.1007/978-981-15-3380-8_43.

[37] R. Chawuthai, R. Sakdanuphab, The analysis of a microwave sensor signal for detecting a kick gesture, in: ICEAST 2018, 2018. doi:10.1109/ICEAST.2018.8434455.

[38] R. Chawuthai, N. Chankaew, T. Threepak, A hybrid method for predicting a potential next rest stop of commercial vehicles, Transportation Research Procedia (2018). doi:10.1016/j.trpro.

2018.11.011.

[39] R. Chawuthai, Monitoring roadway lights and pavement defects for nighttime street safety assessment by sensor data analysis and visualization, Sensors and Materials (2018). doi:10.18494/SAM.2018.1842.

[40] R. Chawuthai, H. Takeda, T. Hosoya, Link prediction in linked data of interspecies interactions using hybrid recommendation approach, in: Semantic Technology (JIST 2014), Springer, 2015. doi:10.1007/978-3-319-15615-6_9.