Preprints.org

Article

Not peer-reviewed version

Declared and Undone: An Open Pipeline for Finding Self-Reported Unfinished Work in Research Articles, with a Prospective Test and an Exploratory Case Series

Eirik Botten Nicolaysen *

Posted Date: 7 October 2026

Keywords: research limitations; unfinished research; large language models; text classification; annotation reliability; open science

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0368.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Declared and Undone: An Open Pipeline for Finding Self-Reported Unfinished Work in Research Articles, with a Prospective Test and an Exploratory Case Series

Eirik Botten Nicolaysen

EcoDeco AS, Lillehammer, Norway; eirik@ecodeco.no

Abstract

Research articles often state that part of the reported work was not done and give a reason. We call such statements parked research questions and ask whether they can be found systematically and what obstacles they name. We built an open pipeline — a local language-model sieve followed by an agentic frontier-model reader — and applied it in two phases. In a development phase on 100 open-access works from four fields, the reader returned 432 candidate rows, or 276 distinct declarations; among declarations with a majority class, 71.2% name data that never existed. In a prospective phase on 100 new archaeology works, with gates fixed before the draw, a blind second reading confirmed 85 of 100 sampled reader positives (work-clustered 95% interval 76.2–93.8), and the sieve retained 95 of 104 statements found by cover-to-cover reference reading of 20 works (76.9–96.7). All coders are language models, so these figures measure agreement within a model-based instrument, not validity against human reading; a human validation study is preregistered. Agreement between two model coders on the hit decision was high (Cohen's kappa 0.812), and agentic reading agreed with the reference more than the same model reading one passage per call (paired difference 0.177, 0.057–0.318). An exploratory series of six reopening attempts, selected by judgement, found no obstacle lifted. The pipeline, registers, protocols and a manifest linking every number to its source are openly deposited.

Keywords: research limitations; unfinished research; large language models; text classification; annotation reliability; open science

1. Introduction

A parked research question is a specific speech act: the authors state that a part of *their own* reported work was not done, and give the obstacle as the reason. The sentence is recognisable when quoted — "we were unable to code all data about discharge diagnosis and insurance status" — but it shares no vocabulary and no topic with its neighbours; the class is defined by form.

Three questions follow. First, does the class occur in published text in measurable quantity? Second, can declared obstacles be classified in a way that distinguishes those that might be lifted by new tools from those that cannot? Third, can the class be retrieved systematically, with known retention and agreement? A fourth question — whether specific obstacles have since been lifted — motivated the work but is addressed here only by an exploratory case series.

The contributions are: (i) an operational definition and a nine-class obstacle typology fixed before data; (ii) an open, reproducible pipeline; (iii) measurements of agreement among language-model coders on this task, including the effect of the reader's working mode; (iv) a prospective test of the pipeline on new material with gates fixed in advance; and (v) a deposit in which every reported number is traceable to a file, a hash and the script that generated it.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0368.v1

2. Related Work

Work on extracting limitations from research articles is the closest neighbour. BAGELS provides a benchmark for automated extraction and generation of limitations from scholarly text (Al Azher et al., 2025). LimitGen evaluates whether large language models can identify critical limitations in research papers, using both synthetic and human-written limitations (Xu et al., 2025). Both treat *limitations* — weaknesses of a study as reported or as identified by a reviewer. The class studied here is narrower and different in kind: it requires (a) that an *undone* part of the authors' own work is declared and (b) that a named obstacle is given as the reason, linked to it in the same passage. Many limitations are not declarations of undone work, and many declarations of undone work appear outside limitations sections; 81.4% of passages in our material carry no section label at all, so a section-based approach is not available.

Classification of rhetorical sentence types in scientific text is an established task. Sequential sentence classification with language models has been evaluated against manually annotated corpora, including the biorc800 dataset (Lan et al., 2024), and the SemEval-2021 NLPContributionGraph task structured scholarly contributions for a research knowledge graph (D'Souza et al., 2021). Those studies anchor model output to human annotation; the present study does not yet do so, which is its main limitation (§6).

On language models as annotators, a recent comparison found models effective as assistants to human annotators but not reliable as independent annotators against experts (Gu et al., 2026). We report agreement among model coders as a property of the instrument and do not treat it as evidence of validity. For agreement on clustered data we follow the cluster bootstrap described by Vanbelle (2019).

3. Design

3.1. Two Phases and a Case Series

The study has two phases with different status, and they are reported separately.

Phase 1 (development, 100 works). The protocol, typology and passage unit were locked internally before data on 12 September 2026, with a hash as the first repository commit, but were not timestamped by a third party until the first deposit on 25 September 2026. During phase 1, three rule clarifications were written after data were seen, the sieve was revised, a gate failed and a second gate was added afterwards. Phase 1 is therefore reported as development and exploratory analysis. Its gates are reported with the timing of each.

Phase 2 (prospective, 100 new works). The addendum governing phase 2 was locked before the draw; it fixed the frame, the draw, the reference procedure, both gates and the analysis. Phase 2 is the confirmatory test of the pipeline's properties. Three deviations occurred and are reported in §5.5.

Case series (exploratory). Six attempts to reopen a parked question were selected by judgement from phase 1 output (§5.7). The criterion for calling an obstacle "lifted" was written after the six attempts. The series is descriptive.

3.2. Frame and Unit

The frame is OpenAlex, publication years 2015–2020, open access with retrievable full text. Phase 1 drew 25 works from each of energy modelling, archaeology, clinical epidemiology and historical textual scholarship, chosen to span the sciences and humanities; phase 2 drew 100 works from the frozen archaeology frame, none overlapping phase 1. The retrievability requirement makes the frame skewed by publisher delivery format.

The passage is a candidate sentence with two sentences on each side; passages are generated at a stride of three sentences, so neighbouring passages share text. A *hit* requires that an undone part of the reported work and an obstacle both lie in the passage and that the obstacle is given as the reason. Because one statement can appear in more than one passage, results are reported both per register

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0368.v1

row and per **distinct declaration**, obtained by merging rows from the same work with overlapping sentence spans.

3.3. Typology

Each hit is assigned one of nine obstacle classes fixed before data: human reading or coding at scale (H1), legibility of a damaged or handwritten source (H2), language (H3), pattern detection in images or signals at volume (H4), simulation and compute (H5), structural inference from sequence (H6), data that did not exist (H7), access, law, ethics and consent (H8), and a missing concept or theory (H9). When a passage does not say whether the obstacle was workload or missing data, the class is *undecided* (H1/H7). Three non-hit forms are recorded separately: answered in the same article (N1), a novelty claim (N2), and a scope choice without a named obstacle (N3). Whether a class is liftable by current tools was assigned per class in the protocol (H1–H6 liftable, H7–H9 not) and is never inferred by a model; the undecided class is reported apart from both.

3.4. Pipeline

The pipeline runs as one command: retrieve → passages → judge → extract → union → blind files → read → work-level summary → citation check → register.

- **Sieve.** A local model (`gemma2:9b`, temperature 0, fixed seed) judges each passage against the typology, and the same model extracts candidate sentences per document. The union of the two forms the candidate set: 2 844 of 22 243 passages (12.8%) in phase 1, 4 107 of 21 361 (19.2%) in phase 2.
- **Reader.** An agentic frontier-model instance (`claude-opus-5`, Anthropic), given the rule file and a block list, reads candidates in batches of at most 356 and returns a verdict, class and doubt flag per passage. Verdicts are dated and not repeatable: the model has no seed and is updated by its vendor. Phase 2 model identity was verified per session from transcripts.
- **Register.** One row per reader-positive passage, with class, doubt flag, sieve provenance, liftability from the protocol table and section provenance.

3.5. Coders and Terminology

All coders are language models. *Coder 1* produced the phase 1 reference reading; *coder 2* is an agentic claude-opus-5 instance given the rule file, a blind file and a block list; *coder 4* is a chat instance of a different model from the same vendor (product Claude Fable 5.1; the underlying model identifier was not recorded), reading the same blind file with the same rule file; seven of its 320 rows were not coded blind.

Because no human criterion exists yet, we use the following terms. A *reference-positive* is a passage a reference reading coded as a hit. *Agreement* is Cohen's κ, reported with raw agreement, PABAK and Gwet's AC1. *Confirmation rate* is the share of sampled reader positives that a separate blind model instance codes as hits. *Retention* is the share of reference-positives that pass the sieve; *reader agreement with the reference* is the share the reader also codes as hits. None of these is precision or recall against human reading.

3.6. Statistics

The work is the cluster unit. All intervals are percentile intervals from a work-clustered bootstrap (10 000 replicates, seed 734248) unless stated otherwise (Vanbelle, 2019); shares of distinct declarations carry Wilson intervals; passage-level intervals are given where they were the original report. Differences between agreement coefficients are tested with a paired bootstrap on the difference.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0368.v1

3.7. Use of Large Language Models

Language models were used in two roles. As research instruments, they judged and extracted candidate passages (local model), read candidates and served as reference coders (frontier models), and executed steps in the case series under locked instructions; their roles, model identities and agreement are reported in §3–§5. As writing tools, a large language model (Claude, Anthropic) assisted in drafting and editing the manuscript text under the author's direction. The author designed the study, wrote the protocol and every criterion, made every recorded decision, interpreted the results, and revised and approved the final text. Every number was checked against a hash-anchored fact file, and every reference against Crossref or a verified URL, before submission. The author takes full responsibility for the content.

4. Human Validation (Preregistered, Not yet Run)

A validation study was locked before any human coding: a stratified sample of 149 passages from 75 works, covering sieve-positive and sieve-negative passages, reader doubt and no doubt, all four phase 1 fields and phase 2 positives and misses, drawn after the lock; a blind file without model output; a codebook written for human annotators from the rule file; adjudication rules; and an analysis plan reporting human–human, model–human and model–adjudicated agreement and precision and recall against the adjudicated reference. Results will be reported in a revision or a separate paper.

5. Results

5.1. The Class Occurs, and Most Declared Obstacles Are Missing Data (Phase 1)

The reader returned 432 rows from 60 of the 100 works. Of these rows, 269 share a sentence with another row in the same work; merged, they are **276 distinct declarations**. Among declarations with a majority class, **183 of 257 (71.2%, 65.4–76.4)** (Wilson) name data that did not exist (H7), and among decided declarations **42 of 245 (17.1%, 12.9–22.4)** (Wilson) fall in a liftable class (H1–H6); 19 declarations had a tied class and are held out of class shares. Per row, H7 is 289 of 432 (work-clustered 58.2–79.7) and the liftable share 77 of 406 decided rows (9.0–26.5); the width of the clustered intervals reflects that one work contributes 118 rows. Row-level class counts: H7 289, H5 50, H9 29, undecided 26, H2 14, H8 11, H1 10, H3 2, H4 1, H6 0.

The share of reader positives among sieve candidates differs by field — archaeology 274 of 990 (27.7%), clinical epidemiology 53 of 347 (15.3%), textual scholarship 37 of 378 (9.8%), energy modelling 68 of 1 129 (6.0%) — and so does the share of passages that become sieve candidates (6.8–17.5%). These are rates within the sieve's output, not field prevalences: the sieve's retention per field was not measured in phase 1.

5.2. Agreement Among Language-Model Coders

On the 320 passages read by coders 1 and 2 (92 works; largest work 33 passages; Table 1):

Table 1. Agreement on the hit decision between language-model coders, 320 passages.

pair	κ	work-clustered 95%	raw	PABAK	AC1	positives
coder 1 – coder 2	0.812	0.711–0.909	96.2%	0.925	0.953	39 / 33
coder 4 – coder 2	0.899	0.833–0.974	98.1%	0.963	0.977	33 / 33
coder 4 – coder 1	0.781	0.675–0.877	95.6%	0.912	0.945	33 / 39

Excluding the seven rows not coded blind by coder 4 lowers κ by 0.024–0.025, against 0.018 for the control pair (coder 1 – coder 2), which was never exposed; the drop reflects the loss of reference-positives from the positive class, and the ordering of the three pairs is unchanged. Agreement on class among passages both coders call hits is lower (coders 1–2: κ 0.722); the weakest boundaries are

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0368.v1

two non-hit classes, N3 (κ 0.530) and N2 (κ 0.535). The three coders recorded N3 54, 34 and 13 times on the same passages. A decision rule for the N3 boundary has since been written; its effect has not been measured. Of the 27 phase 1 reference-positives, only 3 are coded as hits by both coders without a doubt flag.

5.3. The Reader's Working Mode Matters

With the same rule file, agreement with coder 1 rises with model capability when each passage is a separate call: Haiku 4.5 κ 0.031, Sonnet 5 κ 0.294, Opus 5 κ 0.636 (0.503–0.744). The same Opus model working agentically, with the rules and a batch of passages in one session, reaches κ 0.812. The paired difference, 0.177, has a work-clustered interval of **0.057–0.318** and excludes zero. A local 7B model run agentically reached κ 0.248, so working mode alone does not close the gap.

5.4. Phase 1 Gates

The preregistered gate required a third reading of the candidate set to confirm at least 22 of 27 reference-positives; it confirmed 21 and the gate failed. All six losses were doubt-marked by the reference coder. The threshold (81%) had been set above the agreement of the two reference coders with each other (19 of 27, 70.4%). A second gate, written after the first failed, sampled 100 reader positives for blind confirmation: **85 of 100 (work-clustered 75.7–92.3; 33 works, largest 21 rows)**. The phase 1 list is therefore named a candidate list.

5.5. Phase 2: Prospective Test

Before the sieve ran, 20 of the 100 new works were read cover-to-cover by twenty agentic reference instances, one per work, with an empty context, the rule file and a block list; the key was secured with its hash before the sieve started. The reference reading found **104 reference-positives in 12 of the 20 works; one work holds 51** (Table 2).

Table 2. Prospective phase: confirmation, retention and reader agreement.

measure	value	work-clustered 95%
confirmation rate (100 sampled reader positives, 38 works)	85 / 100	76.2–93.8
retention by the sieve	95 / 104 (91.3%)	76.9–96.7
reader agreement with the reference	85 / 104 (81.7%)	60.9–89.7
retention, excluding the 51-hit work	45 / 53 (84.9%)	—
reader agreement, excluding that work	38 / 53 (71.7%)	—

Confirmation separates on the reader's doubt flag: 37 of 38 without doubt, 48 of 62 with doubt (post hoc subgroup). The reader returned 673 rows from 73 of 100 works, which deduplicate to **424 declarations**; among declarations with a majority class H7 is **265 of 402 (65.9%, 61.2–70.4)** (Wilson), and among decided declarations the liftable share **53 of 375 (14.1%, 11.0–18.0)** (Wilson). The reader consumed 118 185 context tokens per reader positive (phase 1: 164 805).

The confirmation gate (≥ 0.70) was met, and the phase 2 list is named a worklist under a prospective gate. Retention and reader agreement were reported without a threshold, as locked. Deviations: the reference order was enforced by procedure when the sieve started and in code from later the same day; the register step was corrected after the lock (row content unchanged, hash for hash); two reader sessions hit a usage limit and were completed in a second instance without rewriting verdicts.

5.6. A Secondary Base-Rate Estimate (Phase 1)

A cell-weighted estimator over the strata of the 320-passage reading (judge-flagged passages by field and judged class; non-flagged passages by judged non-hit class) gives 760 expected reference-positive passages, **3.42% of passages (work-clustered 1.3–7.5)**. The estimate rests on 2 positives

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0368.v1

among 100 read non-flagged passages and on one assumption (cells without read rows take the field's rate); it describes reference-positive passages in this frame, not prevalence in the literature.

5.7. Exploratory Case Series

Selection. Case A was selected on 21 September 2026 from the four phase 1 reference-positives in a liftable class, as the only one whose input was available. Cases B–F were selected on 27 September from the 50 phase 1 rows in a liftable or undecided class without a doubt flag: these were grouped into 19 cases (work × obstacle), 10 were set aside as requiring physical material, fieldwork or people, and a short list was made of those for which a criterion could be locked and the resource appeared to exist. Every step after the first was judgement. The series therefore favours cases where reopening looked feasible and cannot estimate how often any obstacle type blocks reopening.

Outcomes. "Lifted" was defined afterwards by three conditions: the named resource exists in an open source; the source's own numbers can be reproduced first; a criterion locked before computation is met. Criteria were locked before computation for cases A, B and C (Table 3).

Table 3. Exploratory reopening attempts.

case	work	obstacle	outcome
A	Pring (2016)	H5, model would not accept measured density	not lifted; model inconsistency reproduced for the four horizons named
B	Haouachi (2016)	H3, Greek sources read only in translation	not lifted; 79 of 90 places resolved to open Greek text (87.8%, threshold 90%)
C	Rodrigues Aragao (2018)	H5, tool could not force equal partitions	not lifted by the named means; the named package (Žiberna, 2018, 2025) lacks the mechanism
D	Rodrigues Aragao (2018)	interview material for automatic concept-network generation	stopped: material withheld under consent
E	Riris (2018)	H5, model runs split for time	stopped: model code not found in six checked routes; one route unresolved
F	Bynum et al. (2021)	H5, full scenario enumeration	stopped: scenario set and distribution parameters not reported

In case B, the 90-place denominator is the one used; a narrower citation form gives 55 of 60 (91.7%), reported as a post hoc sensitivity. Two of eight checkable references in the thesis point to a different passage than the one cited. Both open text services answered HTTP 200 with a different passage when the requested one did not exist; without address-level comparison, all eleven misses would have been recorded as resolved. In case C, the thesis's reported values were reproduced exactly (12 of 12), and its total error of 14 was reproduced in two independent implementations; the thesis's own workaround is a constrained solution with lower error (14) than the best constrained search found from 120 random starts (18). In case F, the article's statement that 100 scenarios are below 0.001% of the space was confirmed ($C(41,7) = 22\,481\,940$; 0.000445%), with the 41 lines of MATPOWER's case30 (MATPOWER, n.d.; Zimmerman et al., 2011).

For three further parked items (Hansen et al., 2019; Karatas, 2018; Mecking et al., 2017), the citation graph since publication shows no later work doing what was declared undone.

6. Discussion

What the evidence supports. The class can be retrieved at scale with a known pipeline, and the pipeline's behaviour is stable across two applications: confirmation rates of 85 of 100 in both phases, sieve retention above 90% against reference reading in phase 2, and high agreement between model coders on the hit decision. Among declarations with a majority class in both phases, between 66% and 71% name data that never existed — an obstacle that later analytical tools cannot supply. In the exploratory series, no obstacle was shown to be lifted.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0368.v1

What it does not support. None of these figures is validity against human reading: every coder and reference reader is a language model, the phase 2 reference and the reader share vendor and rule file, and agreement among model coders can coexist with shared errors (Gu et al., 2026). The phase 2 retention estimate is dominated by one work and has a wide clustered interval. The case series was selected by judgement and favours feasible cases; that three of six stopped on unavailable input is an observation about these six, not an estimate of how often access limits reopening, and the series contains no comparison with model capability.

The reader. The size of the paired difference between agentic and per-passage reading by the same model, together with the failure of a smaller local model in agentic mode, indicates that for this task the reader must be both capable and given the rules and context together. That is a practical constraint on anyone reproducing the pipeline.

Relation to limitations research. Declared-undone work with a named obstacle overlaps with limitations as studied by BAGELS and LimitGen but is neither a subset nor a superset: it requires an undone component and a stated reason, and it is distributed throughout articles rather than in limitations sections. A direct comparison on a shared corpus is left for future work.

7. Limitations

1. No human annotation yet; a validation study is preregistered (§4).
2. All coders are language models; coder 4's underlying model identifier was not recorded, and coders 1, 2 and the phase 2 reference readers share vendor and rule file.
3. Phase 1 is development and exploratory analysis; the protocol was not timestamped by a third party before data, and three rule clarifications followed data.
4. "Lifted" was defined after the case series; the series is exploratory and judgement-selected.
5. Four purposively chosen fields in phase 1 and one field in phase 2; frames limited to open-access, retrievable texts with publisher skew.
6. Sieve retention was measured only in phase 2, on 20 works, one of which holds half the reference-positives.
7. Reader verdicts are dated and not repeatable.
8. Assessability of obstacles without a domain expert was not operationalised; two readings within the protocol's wording gave 88% and 32%.

8. Conclusions

Declarations of undone work with a named obstacle can be retrieved from research articles with an open pipeline whose behaviour was stable across a development and a prospective phase. Most such declarations, in the fields studied, name data that were never collected. Whether the pipeline's agreement figures hold against human reading is the next question, and the study to answer it is locked.

Funding: No funding was received for conducting this study.

Institutional Review Board Statement: Not applicable: the study analyses published text and involved no human participants.

Data Availability Statement: Code is Apache-2.0 at https://github.com/avalyset/gjenopptak. Own data are CC BY 4.0; third-party text is not redistributed. The deposit (concept DOI 10.5281/zenodo.22959326; version v0.5.0 for this manuscript, 10.5281/zenodo.23120143) contains the protocol and all addenda with lock hashes and OpenTimestamps receipts (Todd, 2024), the rule files and block lists every locked measurement read, blind files and both registers as indices (work identifier, sentence span, text hash, class, doubt, sieve provenance; no passage text), the phase 2 measurements and reference key, the reanalysis output, the case files, and a number manifest that links each number in this paper to its file, hash and generating script.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0368.v1

Conflicts of Interest: The author is the founder and CEO of EcoDeco AS, which developed the pipeline as open-source software; the author has no other relevant financial or non-financial interests.

References

Al Azher, I., Mokarrama, M. J., Guo, Z., Ray Choudhury, S., & Alhoori, H. (2025). BAGELS: Benchmarking the automated generation and extraction of limitations from scholarly text. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 19279–19294). https://doi.org/10.18653/v1/2025.findings-emnlp.1050

Botten Nicolaysen, E. (2026). *Gjenopptak: Declared-and-undone work in research articles* (Version 0.5.0) [Data set and software]. Zenodo. https://doi.org/10.5281/zenodo.23120143

Bynum, M., Staid, A., Arguello, B., Castillo, A., Knueven, B., Laird, C. D., & Watson, J.-P. (2021). Proactive operations and investment planning via stochastic optimization to enhance power systems' extreme weather resilience. *Journal of Infrastructure Systems*, *27*(2), Article 04021004. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000603

D'Souza, J., Auer, S., & Pedersen, T. (2021). SemEval-2021 Task 11: NLPContributionGraph — Structuring scholarly NLP contributions for a research knowledge graph. In *Proceedings of the 15th International Workshop on Semantic Evaluation* (pp. 364–376). https://doi.org/10.18653/v1/2021.semeval-1.44

Gu, F., Li, Z., Colon, C. R., Evans, B., Mondal, I., & Boyd-Graber, J. L. (2026). Large language models are effective human annotation assistants, but not good independent annotators. In *Findings of the Association for Computational Linguistics: ACL 2026* (pp. 71–89). https://doi.org/10.18653/v1/2026.findings-acl.4

Hansen, S., Montero-Ruiz, I., Rovira, S., Steiniger, D., & Toderaş, M. (2019). The earliest lead ore processing in Europe. 5th millennium BC finds from Pietrele on the Lower Danube. *PLOS ONE*, *14*(4), Article e0214218. https://doi.org/10.1371/journal.pone.0214218

Haouachi, D. (2016). *Les personnages féminins chez Tite-Live : idéologie et art de la mise en scène* [Doctoral dissertation, Université de Strasbourg]. https://doi.org/10.70675/8513d3c9ze385z4b4bz910cz221c7718984d

Karatas, A.-M.-S. (2018). The significance of money for the cults and sanctuaries of Demeter and Kore: The shaping of the cults by commercial transactions. *Mythos*, *12*, 47–88. https://doi.org/10.4000/mythos.297

Lan, M., Zheng, L., Ming, S., & Kilicoglu, H. (2024). Multi-label sequential sentence classification via large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 16086–16104). https://doi.org/10.18653/v1/2024.findings-emnlp.944

MATPOWER. (n.d.). *case30.m* [Data set]. GitHub. Retrieved September 28, 2026, from https://raw.githubusercontent.com/MATPOWER/matpower/master/data/case30.m

Mecking, O., Hohle, I., & Wolfram, S. (2017). The development of pottery technology in Eythra from the Early Linear Pottery culture to the Late Stroke Ornamented Pottery culture. *Archeologické rozhledy*, *69*(2). https://doi.org/10.35686/ar.2017.12

Pring, L. J. (2016). *Soil-based analysis methods to aid the detection of cropmarks over buried features* [Doctoral dissertation, University of Birmingham]. https://etheses.bham.ac.uk/id/eprint/7069/

Riris, P. (2018). Assessing the impact and legacy of swidden farming in neotropical interfluvial environments through exploratory modelling of post-contact Piaroa land use (Upper Orinoco, Venezuela). *The Holocene*, *28*(6), 945–954. https://doi.org/10.1177/0959683617752857

Rodrigues Aragao, R. (2018). *Using network theory to manage knowledge from unstructured data in construction projects: Application to a collaborative analysis of the energy consumption in the construction of oil and gas facilities* [Doctoral dissertation, University of Toronto]. https://hdl.handle.net/1807/92034

Todd, P. (2024). *opentimestamps-client* (Version 0.7.2) [Computer software]. Python Package Index. https://pypi.org/project/opentimestamps-client/0.7.2/

Vanbelle, S. (2019). Asymptotic variability of (multilevel) multirater kappa coefficients. *Statistical Methods in Medical Research*, *28*(10–11), 3012–3026. https://doi.org/10.1177/0962280218794733

Xu, Z., Zhao, Y., Patwardhan, M., Vig, L., & Cohan, A. (2025). Can LLMs identify critical limitations within scientific research? A systematic evaluation on AI research papers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (pp. 20652–20706). https://doi.org/10.18653/v1/2025.acl-long.1009

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 October 2026 doi:10.20944/preprints202610.0368.v1

Žiberna, A. (2018). *blockmodeling* (Version 0.3.1) [Computer software]. CRAN. https://cran.r-project.org/src/contrib/Archive/blockmodeling/

Žiberna, A. (2025). *blockmodeling: Generalized and classical blockmodeling of valued networks* (Version 1.1.8) [Computer software]. CRAN. https://CRAN.R-project.org/package=blockmodeling

Zimmerman, R. D., Murillo-Sanchez, C. E., & Thomas, R. J. (2011). MATPOWER: Steady-state operations, planning, and analysis tools for power systems research and education. *IEEE Transactions on Power Systems*, *26*(1), 12–19. https://doi.org/10.1109/TPWRS.2010.2051168.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.