How duplicate attribution changes database coverage in two reviews of AI in engineering education

Lluis Eriksson

Abstract

Which search sources would have recovered all studies retained by a review? The database printed on each accepted row can give a misleading answer when another source supplied a duplicate. We reconstruct the search-source incidences of 22 accepted titles in one education review and three final studies in another. Accounting for duplicates reduces the loss attributed to omitting IEEE from eight titles to one in the first register. Its three sources remain necessary. In the second, the ACM export alone contains all three final studies, while the set of 15 studies advancing to the preceding screening stage requires both ACM and ScienceDirect. A reference marked for inclusion during backward searching resolves to one of the original three. These are retrospective results for frozen search exports. They identify the consequences of attribution and denominator choice for these reviews.

1 What it means for a search source to cover a study

Let T be the retained titles at a specified screening stage and let $S(t)$ contain the sources with a matching record for t. For a portfolio P of sources, the observed coverage is $C(P) = \sum_{t \in T} \mathbf{1}\{S(t) \cap P \neq \emptyset\}$. An omitted source costs one title for every title it alone supplied in the selected portfolio. Source membership refers to a deposited search export. A paper published by IEEE may have been retrieved through ACM or Scopus.

Bramer et al. [1] measured single-database and joint coverage in systematic reviews. We apply that established question to two publicly auditable engineering-education registers [2,3]. The first distinguishes accepted, duplicate and rejected search results within the same table. The second preserves exports from six sources and records intermediate and final screening decisions. These differences let us separate the effect of attributing duplicate records from the effect of changing the target set. The two reviews used different search strategies and inclusion criteria.

2 Attribution within the requirements-education register

Rodrigues de Lima's requirements-engineering education register [2] lists 1,259 rows: 22 accepted, 94 duplicated and 1,143 rejected. We reconstructed the three columns from PDF character positions, checking all accepted entries on the first two pages visually. Only accepted or duplicated entries could add a source to a retained title. We joined titles first after Unicode normalization, case folding and whitespace normalization, then after tokenizing letters and digits. The latter joined one additional pair. The two spellings differ only by the apostrophe in the paper by Guardado et al. [5]. The additional Scopus row appears on page 104. The publisher's title corroborates the identification.

Source omitted	Accepted row only	Strict title join	Resolved join
ACM Digital Library	2	1	1
IEEE Digital Library	8	2	1
Scopus	12	8	8

Each cell counts the titles missing when the other two sources are retained. The resolved join credits a second source for 12 of the 22 titles. All three sources remain necessary under both title rules. If unrecorded incidences were later established, one new link outside ACM for its sole exclusive title could make ACM dispensable. Removing Scopus would require at least eight such links, one for each of its exclusive titles. This bound is attained in the abstract missing-link model by adding a link for each missing title.

3 Changing the screening stage in the modeling review

Chakraborty et al. [3,4] searched ACM, IEEE Xplore, Scopus, ScienceDirect, SpringerLink and Web of Science. We reconciled the 15 electronically identified studies advancing in the workbook with the 15 studies receiving a final decision. Of those, three were retained. Their publisher-registered DOIs correspond to the three studies named in the authors' Table II. We checked the original export files against the archive checksums, matched identifiers before titles, and parsed the DOI parameter of Scopus links separately from tracking parameters. We compared a bibliographic parser with direct field extraction. Where the export contains additional search records absent from the parser's output we retained the union of supported matches.

Target set	Titles	Covered by ACM	Smallest full portfolio
Electronic shortlist	15	13	ACM and ScienceDirect
Final electronic selections	3	3	ACM
Final plus distinct backward additions	3	3	ACM

The two shortlisted records outside ACM are workbook identifiers 897 and 1383. ScienceDirect supplied both. Final selections 2 and 1387 occur in the ACM export, and final selection 822 occurs in both ACM and Scopus. The one positive backward-search row has the DOI of selection 822, the paper by Morales et al. Counting it as a fourth distinct study would disagree with both the deposited identifiers and the authors' account [4]. In the union of the three sources common to both reviews, ACM covers all 13 visible shortlisted titles. Restricting the denominator to this union would silently discard the two remaining shortlisted titles.

4 Losing access to one source

For a fixed portfolio P, its worst coverage after losing any selected source is $\min_{s \in P} C(P \setminus \{s\})$. With all sources selected, this equals $|T|$ minus the largest number of titles exclusive to one source, since removing one source loses exactly its exclusive titles. In the first register the worst loss is eight, leaving 14 of 22. In the second the worst loss is two at each stage, leaving 13 of 15 on the shortlist and one of three among final selections. Enumerating all portfolios on the 15-title shortlist gives worst-case coverage of 8 with two sources, 11 with three and 13 with four. A fifth or sixth source adds no worst-case coverage in these files. This fixed-export sensitivity bound requires no assumptions about the probability of an outage. It does not predict which studies another search conducted today would retrieve. Alternative title versions without publisher identifiers and records absent from the deposited exports remain possible. The source-specific counts in the article and the archived export files need not coincide.

Model assessments and review context

Codex assisted with extraction, analysis and drafting. The author administers AIRR. Any assessment of the released PDF belongs with its version-specific report and objections. No independent peer review is claimed.

1. W. Bramer et al., Optimal database combinations for literature searches in systematic reviews (2017). doi:10.1186/s13643-017-0644-y.
2. G. Rodrigues de Lima, LLMs in requirements engineering education, selection register (2026). Zenodo 24407666.
3. S. Chakraborty, GenAI in software modeling education, search and screening files (2025). Zenodo 16185787.
4. S. Chakraborty et al., Mind the Ethics! (2025), arXiv:2509.13896v1, Sec. II and Table II.
5. S. Guardado et al., Students' perceptions of the use of LLMs in requirements engineering education (2025). doi:10.1109/RE63999.2025.00022.