

Singing from the Same ANTHEM: An Enriched ACL Anthology for Scholarly Comparison

Moritz Staudinger

TU Wien

Vienna, Austria

moritz.staudinger@tuwien.ac.at

Wojciech Kusa

NASK National Research Institute

Warsaw, Poland

wojciech.kusa@nask.pl

Allan Hanbury

TU Wien

Vienna, Austria

allan.hanbury@tuwien.ac.at

Abstract

Comparative analyses of NLP research, such as identifying who is shaping a subfield, how a topic is shifting geographically, where methodological gaps lie between research communities, require institutional and topical metadata at a density and granularity the ACL Anthology does not itself provide. Large-scale resources offer partial support: OpenAlex provides high-quality institutional metadata where it can match ACL papers (32% of our gold set), but the DOI dependency limits its coverage on older Anthology publications; and the S2ORC bulk release contains 52,013 ACL Anthology fulltexts but populates structured author–affiliation annotations on a small fraction of them in its current state. We present **ANTHEM** (ANTHology with Enriched Metadata), a complementary resource that narrows this gap for NLP publications. ANTHEM links 117,386 ACL Anthology papers to Research Organization Registry (ROR)-normalized author affiliations and Computer Science Ontology (CSO) topic labels via a multi-parser, LLM-based extraction pipeline that retains evidence spans for every decision, achieving $F1_{\text{hier}}$ of 0.760 on a 781-paper decade-stratified gold set, where parent–child ROR identifiers are treated as equivalent (e.g., a department and its parent university). On top of the corpus we release a benchmark of 99 structured comparison questions across five categories, evaluated under a programmatic gold-set construction in which ANTHEM’s own metadata defines the evaluation ground truth. All data, code, and infrastructure are released under CC-BY-SA-NC.

1 Introduction

Suppose you wanted to ask which institutions have driven the recent rise of work on retrieval-augmented generation (RAG), how dialogue-

systems research differs between groups in the UK and Singapore, or how machine translation work has shifted geographically across the last two decades. Each of these questions is comparative, scoped to specific institutions, countries, topics, and year ranges, and each requires structured metadata that the ACL Anthology, the field’s primary archive, does not itself provide. As NLP publication volume grows (Bornmann et al., 2021) and affiliation completeness in scholarly databases declines (Jahn, 2025), the gap between the comparative questions researchers want to ask and the metadata available to answer them widens.

Existing resources address parts of this gap but none composes the pieces these questions need (§2): bibliographic graphs supply institutional metadata through DOI-based joins that are sparse on older ACL entries, full-text corpora supply text rather than structured affiliations, and scholarly-QA systems target factoid QA or related-work generation. Comparative claims about NLP research are consequently reduced to coarse aggregates such as citation counts and publication volume (Alonso et al., 2009; López-Cózar et al., 2012; Mingers and Leydesdorff, 2015), which are blind to collaboration structures, methodological diversity, and the substantive content of contributions.

We introduce ANTHEM (ANTHology with Enriched Metadata). It is a resource paper with a single goal: an enriched ACL Anthology corpus, linking 117,386 papers to Research Organization Registry (ROR)-normalized affiliations and Computer Science Ontology (CSO) topic labels via a multi-parser, LLM-based pipeline that retains evidence spans for every decision. The two remaining contributions serve that goal: a 781-paper decade-stratified gold set that validates the corpus and audits it against three external sources, and a bench-

mark of 99 structured comparison questions that operationalises its intended use, with gold sets computed programmatically from the corpus itself.

2 Related Work

Prior work on this problem comes from three research communities. We describe each below and the role it plays in ANTHEM’s design.

Affiliation extraction and normalization. The first effort treats extraction as a standalone task. Rule-based and shallow-learning systems such as GROBID (Lopez, 2009) and SVM-based extractors (Do et al., 2013) established the baselines, and recent work has shifted toward affiliation *disambiguation*: Lin et al. (2024) presented an end-to-end framework over PubMed, Bitaraf and Jafarpour (2024) combined country, email, and postal-code signals with the Global Research Identifier Database¹ (now ROR), and Duran-Silva et al. (2024) released AffilGood, a multilingual ROR-linked dataset. In the NLP setting, Abdalla et al. (2023) performed affiliation disambiguation as part of a study of industrial presence in ACL papers. These efforts focus on building extractors and disambiguation methods, not on releasing the resulting metadata as a corpus. ANTHEM uses such extraction tooling to construct and release a corpus that integrates affiliations, full text, and topic classification.

Large-scale scholarly resources. The second effort builds durable bibliographic and full-text infrastructure for science as a whole. OpenAlex (Priem et al., 2022), the successor to the now-discontinued Microsoft Academic Graph (Wang et al., 2020), provides broad bibliographic metadata and ROR-linked institutional records. However, OpenAlex’s coverage of older ACL publications depends on DOI availability, which is sparse for pre-2010 Anthology entries; affiliation completeness has also been declining across OpenAlex more broadly (Jahn, 2025). S2ORC (Lo et al., 2020) provides parsed full text across the scholarly literature, combining PDF parsing with L^AT_EX source extraction where available; its structured authoraffiliation annotations are populated densely across many fields, though sparsely on ACL records in the current bulk release (see details in §3.3). Closer to the Anthology itself, AAN (Radev et al., 2009) provides cita-

tion and collaboration graphs updated to 2016, NLP Scholar (Mohammad, 2020) citation and venue statistics but no affiliation metadata, NLPEXplorer (Parmar et al., 2020) coverage to 2019, and D3 (Wahle et al., 2022) affiliation strings at DBLP scale paired via validated name matching but themselves unevaluated and unnormalized. These resources are excellent foundations at the level of the scholarly literature as a whole. ANTHEM provides the NLP-specific composition that brings full text, normalized affiliations, country codes, and topic classification together for the ACL Anthology.

Scientometric analysis and scholarly QA. The third effort asks the comparative questions we want to answer, but with the metadata and tooling each community has at hand. Scientometric studies rely on citation tracking, co-authorship networks, and topic clustering (Luo et al., 2022; Oye-wola and Dada, 2022); in NLP specifically, Rungta et al. (2022) document geographic citation gaps and Abdalla et al. (2023) quantify the growing presence of industrial affiliations. The ACL Anthology publishes its own aggregated statistics² at country-level granularity, but without linked full text or institution-level metadata. On the scholarly-QA side, multi-document QA is established in open-domain (Tang and Yang, 2024; Yang et al., 2018) and scientific (Auer et al., 2023; Wang et al., 2024) settings, and recent synthesis systems including OpenScholar (Asai et al., 2024), Ai2 Scholar QA (Singh et al., 2025), and LitLLM (Agarwal et al., 2025) target factoid QA, single-document grounding, or related-work generation. The most directly related prior system is the Compare framework (Staudinger et al., 2025), which proposes six comparative use cases (University Overview, University Comparison, Multi-University Comparison, Domain Overview, Paper Comparison, Paper QA) and demonstrates RAG-based qualitative synthesis over CORE and OpenAlex; the ANTHEM benchmark adopts a closely related question taxonomy but operationalises it as a quantitative evaluation with programmatic ground truth, adds an agent-mode configuration alongside pipeline RAG, and grounds retrieval in a self-contained corpus rather than live API calls. Neither family directly supports comparative questions structured around institutions, countries, topics, and year ranges as a benchmark, the gap the ANTHEM benchmark (§4

¹<https://www.grid.ac>

²<https://stats.aclrollingreview.org>

is designed to fill.

3 Dataset Construction and Validation

We construct ANTHEM from the complete ACL Anthology and validate it along two axes: an intrinsic benchmark of our extraction pipeline against alternative configurations on a manually annotated gold set, and a cross-source comparison against three external scholarly sources.

3.1 Extraction Pipeline

Figure 1 outlines the pipeline. We process every ACL Anthology entry that ships an associated PDF, retaining Anthology metadata alongside the full text. To handle formatting variability across decades, we run three parsing strategies in parallel: (1) GROBID (Lopez, 2009) produces TEI/XML with structured author blocks; (2) PyMu4LLM converts PDFs to Markdown; (3) and a Tesseract + PyMuPDF fallback yields OCR plain text for documents whose structure cannot be recovered. An LLM (Gemini 2.5 Flash) then extracts author–affiliation pairs into structured JSON from the combined parser outputs, with each affiliation stored alongside its raw string and textual evidence spans tagged with parser source, enabling traceability and treating the LLM as a replaceable component.

Extracted strings are normalized via the ROR API at a 90% confidence threshold; matched affiliations are enriched with canonical name, country code, institution type, and ROR matching score, and we cache the ROR parent–child hierarchy so that downstream evaluation can treat a department and its parent university as equivalent (§3.4). Each paper is additionally classified with the CSO Classifier (Salatino et al., 2019)³ over the parsed full text.

3.2 Corpus Statistics

ANTHEM-202605, the public release described in this paper, processes 117,386 publications from 121,229 ACL Anthology PDFs (96.8% retention), of which 67.6% of extracted raw affiliation instances are matched to ROR above the 90% confidence threshold, yielding 348,057 ROR-linked affiliations across 5,580 institutions in 150 countries; 89.2% of matches are at confidence 1.0. Summary statistics are in Table 1; longitudinal trends are in Appendix A.

³<https://cso.kmi.open.ac.uk>

ANTHEM-202605 Statistics	
Papers processed	117,386
Papers with affiliations	114,210
Papers with CSO topic labels	115,398
Total author entries	536,315
Total ROR-linked affiliations	348,057
Unique ROR institutions	5,580
Unique countries	150
Unique CSO labels	5,893
Mean authors per paper	4.57
Mean CSO labels per paper	8.5
Raw affiliations matched to ROR	67.6%

Table 1: Summary statistics of ANTHEM-202605.

3.3 Comparison with External Sources

We benchmark ANTHEM against three external sources that provide some combination of full text, bibliographic metadata, and structured author affiliations for ACL papers: OpenAlex, the Semantic Scholar Graph API, and the S2ORC bulk release. None of these is designed for full-Anthology per-paper affiliation coverage, and our comparison (Table 2, evaluated in §3.4) characterizes what each contributes complementary to ANTHEM. OpenAlex is the strongest complement for pre-2010 papers, where DOI-based bibliographic backfill produces high-quality institutional metadata and a hierarchical-F1 of 0.635 on the subset it covers, but it is reachable for only 32% of our gold set since DOI availability is sparse on older Anthology publications. The Semantic Scholar Graph API returns each author’s *current* affiliation rather than the affiliation at time of publication, a deliberate design choice that yields hierarchical-F1 of 0.115 on the 24% of papers it covers. S2ORC’s structured authoraffiliation spans are densely populated across the scholarly literature as a whole but sparsely on ACL records in the current bulk release: of the 92,790 ACL papers in the Semantic Scholar papers dataset, 52,013 carry an s2orc fulltext record, and 978 of those carry structured authoraffiliation spans (Appendix C). Each existing resource was optimized for a different layer of the scholarly-information stack; ANTHEM combines full text, normalized affiliations, country codes, and topic classification to allow for comparative analyses.

3.4 Pipeline Evaluation

We construct a 781-paper gold annotation set stratified across five decades to capture historical lay-

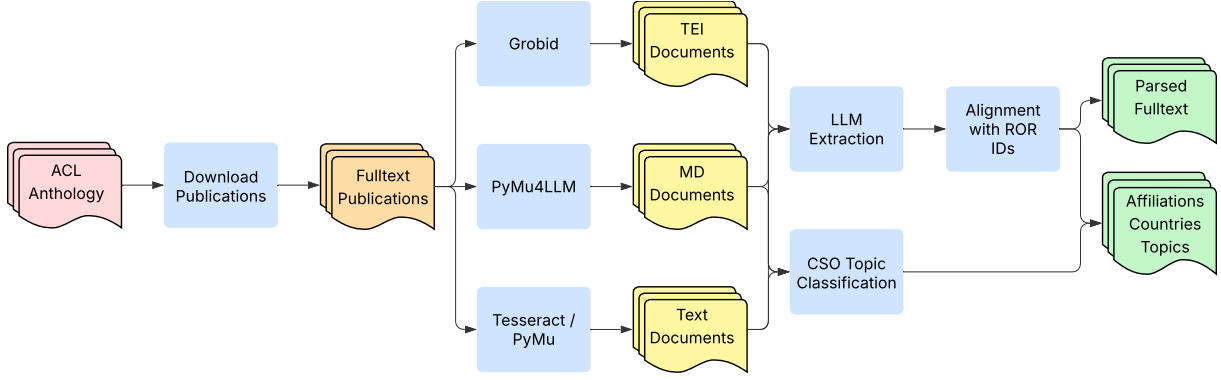


Figure 1: Workflow for extracting and parsing affiliations and topics from ACL Anthology publications.

out heterogeneity, comprising 91/92/187/197/214 papers from the 1980s/1990s/2000s/2010s/2020s, respectively. Annotation was carried out by two of the authors, both with prior experience in scholarly metadata curation, each annotating a disjoint subset of papers by reading the front-matter and recording, for each author, all listed affiliation strings as written and their corresponding ROR identifier obtained from the ROR public lookup. Ambiguous affiliations (e.g., *Meta* without geographic specification) are linked to the closest enclosing entity and flagged. All 802 resulting affiliation–ROR links were re-checked by hand, blind to the pipeline output.

As a reliability check, both annotators independently linked a shared subset of 51 author–affiliation records. Country-code agreement after ISO normalisation was 100% (Cohen’s $\kappa = 1.0$). For ROR identifiers, exact-match agreement was 78.4%. The remaining eleven disagreements were not annotation errors but choices between equally valid annotations of multi-affiliation rows: an author listing both “Systran” and “Autodesk”, or an umbrella research organisation (CNRS) listed alongside its host university (Université Paris-Nord). Disagreements were resolved into a written annotation policy adopted for the remaining annotation work.

We evaluate under three scoring variants. $F1_{\text{raw}}$ is token-Jaccard on raw strings, isolating extraction from linking. $F1_{\text{ROR}}$ requires exact ROR-id set match. $F1_{\text{hier}}$ extends $F1_{\text{ROR}}$ by treating an immediate parent or child ROR id as equivalent, applied non-transitively so siblings never match; $F1_{\text{ROR}}$ is reported in Appendix B.2. We compare our ANTHEM extraction pipeline (Gemini 2.5 Flash over combined parsers, with country-context-aware ROR linking) against LLM and parser abla-

tions, a LoRA-tuned *Qwen2.5-7B-Instruct* baseline (Appendix B.7), GROBID alone, and three external sources (§3.3).

Our pipeline achieves the highest $F1_{\text{hier}}$ of 0.760 at full coverage (Table 2), ahead of Claude Sonnet (0.703) and GPT-5 Nano (0.687), with all pairwise differences against our default pipeline exclude zero under a 10k paired bootstrap. It emits compact normalised strings rather than verbatim source text, costing raw token-Jaccard agreement (0.55 vs. 0.46–0.59) in exchange for more reliable downstream linking. Per-decade performance, scored on each source’s covered subset, varies systematically: our pipeline leads on post-2010 papers ($F1_{\text{hier}}$ 0.761/0.836 on the 2010s/2020s), while OpenAlex leads pre-2010 (0.800/0.691/0.692 on the 1980s/1990s/2000s) where DOI-based backfill compensates for layout heterogeneity that degrades PDF-based extraction; we return to this in §5 and report the full breakdown in Appendix B.1. Error decomposition (Appendix B.5) shows that 90% of our pipeline’s false negatives are ROR *linking* rather than extraction failures: the institution string is present in the extracted text but the ROR linker resolves it incorrectly. The observed errors suggest that improving linking offers greater immediate headroom than improving extraction.

3.5 What the Dataset Reveals

To make concrete what the dataset enables that existing sources do not, Figure 2 renders the institutional and topical composition of ACL Anthology research over 2020–2025 as a four-level treemap (continent, country, institution, topic). The view is produced by joining ROR-linked affiliations, ROR country codes, and CSO topic labels on the released corpus. A reader can see at a glance that Asia (38.4%) has overtaken North America (26.7%)

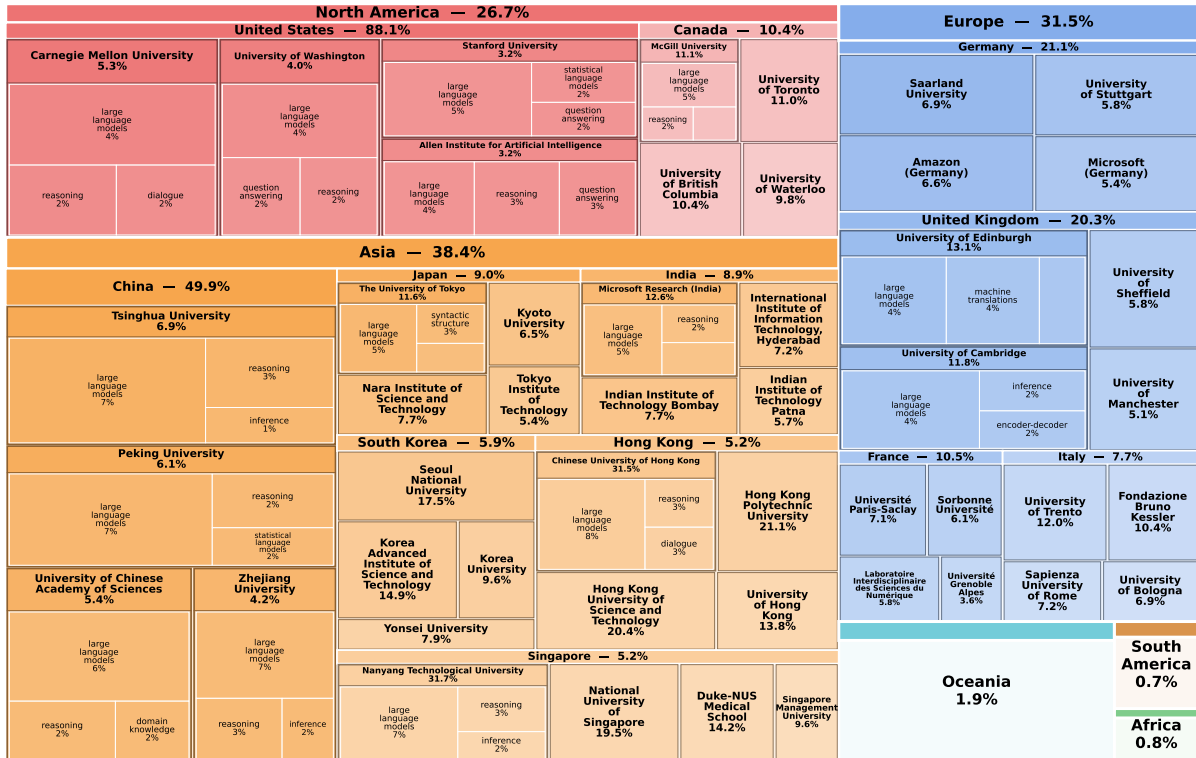


Figure 2: Institutional and topical composition of ACL Anthology research, 2020–2025, computed from ANTHEM-202605. Each rectangle is sized by share of papers; nesting is continent → country → institution → CSO topic. The top four institutions per country and top three topics per institution are shown for readability; the full view, with all institutions above a contribution threshold, is in Appendix A.5.

as the largest contributing continent in the 2020s, with China alone (49.9% of Asia) nearly at parity with the United States; that the United Kingdom and Germany anchor European participation; and that the topical landscape within each institution is dominated by *large language models* and *reasoning* regardless of geography. None of these observations is novel in isolation, but recovering them requires exactly the join of full text, normalized institutional metadata, and topic classification that ANTHEM packages. The same view is reproducible from any released snapshot for longitudinal comparison.

4 The ANTHEM Comparison Benchmark

The ANTHEM benchmark consists of 99 structured comparison questions across five categories spanning institutional, geographic, and topical comparisons over the ACL Anthology; Table 3 shows one example per category. We deliberately exclude paper-level QA categories, which are well covered by existing resources such as QASPER (Dasigi et al., 2021).

Questions are generated from category-specific templates with structured slots filled from the ANTHEM corpus: a CSO topic in its canonical form, one or two institutions referred to by their canonical ROR names, and a year range. Institutions are sampled from the top contributors per topic, country pairs from the geographic distribution per topic, and year ranges randomly within the corpus span. Each candidate is admitted only if its programmatic gold set (§4.1) contains at least 15 papers; all 99 questions were manually verified by the authors. Each question is released with the structured metadata used to compute its gold set: institution and country identifiers, the topic and its CSO closure, and a year range. The CSO closure expands a topic into the full set of equivalent or specialising labels (1 label for *word embedding*, 322 for *speech recognition*). Resulting gold-set sizes range from 15 to 24,340 (median 42), with the largest concentrated in *field_evolution* questions over multi-decade ranges. The reliability of each metadata layer that defines these gold sets is quantified in §4.2 and Appendices B.4 and B.3.

Pipeline	Cov. %	F1 _{raw}	F1 _{hier}
Gemini 2.5 Flash + comb.	100	0.55	0.760 $\pm$.028
<i>LLM ablations</i>			
Claude Sonnet + comb.	99	0.52	0.703 $\pm$.032
GPT-5 Nano + comb.	98	0.46	0.687 $\pm$.033
Claude Sonnet + PyMu	98	0.51	0.676 $\pm$.039
Gemini 2.5 + PyMu	94	0.58	0.653 $\pm$.029
Claude Haiku + comb.	99	0.56	0.650 $\pm$.031
<i>Open / locally-runnable</i>			
Qwen 7B (LoRA) + comb.	99	0.59	0.698 $\pm$.027
<i>Parser-only baseline</i>			
GROBID (header parser)	87	0.47	0.639 $\pm$.030
<i>External sources (§3.3)</i>			
OpenAlex (graph API)	32	0.24	0.431 $\pm$.041
Semantic Scholar Graph	24	0.07	0.126 $\pm$.026
S2ORC (bulk dataset)	2	0.01	0.009 $\pm$.009

Table 2: Affiliation extraction on the 781-paper gold set (blind-verified links). F1_{raw} is token-Jaccard on raw strings; F1_{hier} accepts an exact ROR-id match or an immediate parent/child.

4.1 Programmatic Gold via Constraint Matching

The gold set for a question is computed deterministically against the corpus snapshot: paper p is in $\text{GOLD}(q, \text{ANTHEM-YYYYMM})$ if its publication year falls in $[\text{year_start}, \text{year_end}]$, at least one ROR-linked institution matches the required ROR ids (when specified), at least one ROR-linked institution is in the required country or countries (when specified), and at least one CSO label of p is in the topic closure. Inactive constraints are skipped.

This design has three consequences. First, gold sets are computed rather than annotated per question, which would be structurally infeasible on an evolving corpus of this size, and are reproducible from any released snapshot. Second, the benchmark is naturally longitudinal: the 99 questions can be re-evaluated against any subsequent ANTHEM snapshot, and users can derive subsets by filtering on year or venue. Third, the fields defining the gold answers are the same the extraction pipeline produces, so benchmark reliability tracks corpus reliability across snapshots.

4.2 RAG Configurations and Evaluation

We provide reference results from two RAG configurations evaluated with three frontier LLMs (Claude Sonnet 4.6, Gemini 2.5 Pro, GPT-5 Mini), and three openly available models (DeepSeek v4 Flash, Gemma 4-31B Instruct and Qwen 3.6-35B-

A3B). All configurations share a dense index over all abstracts.

Pipeline mode executes a fixed sequence: (1) the generator LLM extracts structured constraints from the natural-language query (the same model handles classification and final answer generation); (2) the corpus is hard-filtered by year, country, topic via CSO closure, and ROR ids; (3) the question is dense-encoded and the top- K filtered papers are retrieved; (4) the retrieved abstracts and the question are passed to the generator, which produces a grounded answer with inline citations. The classify-filter-retrieve sequence issues five retrieval calls per question, so the effective generator context is $5 \times K$ candidate abstracts. We report both $K=5$ (effective 25 candidates) and $K=15$ (effective 75); larger K broadens the candidate pool for synthesis but stresses generator context handling, and two of the open models failed to produce usable output at $K=15$ (§4.3).

Agent mode exposes the same retrieval and filtering operations as tools, plus full-text fetching for a specific paper, and lets the generator LLM decide autonomously when and how to call them. Unlike pipeline mode, the agent can issue multiple retrieval calls, refine filters mid-trajectory, or skip retrieval on questions it judges answerable directly. We use a single agent prompt across all questions to isolate agentic routing from prompt engineering.

Evaluation. We compare each system’s returned set against the programmatic gold and report set-level precision, recall, and F1. We additionally measure RAGAS Faithfulness, which scores whether each claim in the generated answer is supported by retrieved evidence, and RAGAS Answer Relevancy, which scores whether the answer addresses the question (Es et al., 2024). A single judge (Qwen 3.5-397B) is used across all systems for tractability.

What the benchmark measures. No system receives the structured constraints, each gets only the natural-language question and must recover them, filter, and rank. The benchmark therefore measures *agreement with the corpus metadata*, not absolute correctness. A system retrieving a paper that satisfies the constraints but that ANTHEM mislinked is scored as a false positive. The three defining layers bound how often this happens: extraction 0.760 (§3.4), linking 0.857 (§B.4), and topic precision 0.713 (§B.3), the last being the weakest and active

Category (#Q)	Input	Example query
overview_uni (20)	one university + topic + years	What did Tianjin University contribute to word embedding research between 2015 and 2021 ?
uni_to_uni (20)	two universities + topic + years	What did Carnegie Mellon University contribute to parallel corpora research between 2014 and 2022 compared to University of Illinois Urbana-Champaign ?
uni_country (13)	one university + country + topic + years	What did Heriot-Watt University contribute to spoken dialogue research between 2014 and 2023 compared to other universities in the United Kingdom ?
cross_country (28)	two countries + topic + years	Which approaches to deep learning dominated in research from Egypt versus Greece between 2020 and 2023 , and what does the gap suggest?
field_evolution (18)	topic + years	How has machine translation research evolved between 2004 and 2025 ?

Table 3: One example question per category from the ANTHEM benchmark, with parenthesised question counts. Colours mark the structured slots: **single institution**, **paired institution**, **topic**, **country**, **year range**.

in every question. The noise is shared across systems, so relative comparisons at a fixed snapshot are less affected than absolute values.

4.3 Results

Retrieval: two operating points. Set-level F1 ranges from 0.207 (pipeline Claude Sonnet 4.6) to 0.293 (pipeline Gemini 2.5 Pro). Two qualitatively distinct operating points emerge in agent mode: GPT-5 Mini returns broad candidate sets and achieves the highest recall (0.468) at the cost of lower precision (0.337), while Gemini 2.5 Pro returns tighter sets, achieving the highest precision (0.470) but recovering only 30.4% of the gold. Pipeline systems at K=5 span the full band (F1 0.207–0.293), with the strongest ahead of any agent configuration. The confidence intervals overlap substantially across the leading systems, so we report Gemini 2.5 Pro as leading rather than as significantly best.

Open models: competitive retrieval, weaker synthesis. Gemma 4-31B Instruct and Qwen 3.6-35B-A3B (served via OpenRouter) outperform the commercial GPT-5 Mini and Claude Sonnet pipelines on retrieval F1 (0.277 and 0.254 against 0.228 and 0.207), but diverge on synthesis: Gemma reaches RAGAS faithfulness of 0.413, comparable to several commercial systems, while Qwen reaches only 0.191. For institutions that cannot rely on commercial APIs, this suggests open-model synthesis quality is the more pressing direction for improvement than open-model retrieval.

K=5 vs K=15. Enlarging the candidate pool changes little for retrieval: Gemini scores 0.293 at both settings and the other systems shift by at most 0.036, trading precision for recall (0.446/0.401 at

K=5 against 0.424/0.425 at K=15). Faithfulness at K=15 is judged on only 41–56 questions per system, and on a non-random subset, against 73–99 at K=5, so we do not compare the two settings on that metric. Within the range we test, K is therefore not decisive for retrieval quality.

Pipeline vs. agent. Within Gemini 2.5 Pro, the comparison least confounded by output failures, agent mode improves faithfulness (0.323 under our default judge, 0.134 under a second, see Appendix D.2) and precision by 0.024 over the K=5 pipeline, while pipeline mode improves recall by 0.097 and F1 by 0.042. The pattern is consistent with the agent’s ability to analyze a paper’s full text before citing it: a citation following a `drill_into_papers_fulltext` call rests on the paper’s actual content, whereas pipeline citations rest only on the retrieved abstract. The choice between modes is application-dependent: agent mode is preferable when citation fidelity matters; pipeline mode is preferable when comprehensive recovery is the primary goal. Relevancy clusters tightly across all systems (0.81–0.95) and is the least discriminating of our three metrics.

5 Discussion

Where retrieval ends and synthesis begins. The benchmark results separate two failure modes that conventional retrieval evaluation tends to conflate. Among pipeline systems at K=5, Gemma 4-31B attains both the highest precision (0.462) and the highest faithfulness (0.413), while Gemini 2.5 Pro leads on F1 (0.293); the open Qwen 3.6-35B-A3B is close behind on retrieval (precision 0.445, F1 0.254) but reaches only 0.191 faithfulness. The agent advantage on Gemini 2.5 Pro, a 0.323-point

Model	n	Retrieval (set-level)				RAGAS		
		P	R	F1	95% CI	n _j	Faith.	Rel.
Pipeline mode, K=5								
Gemini 2.5 Pro	98	0.446	0.401	0.293	[0.250, 0.336]	98	0.394	0.923
Gemma 4-31B Instr.	99	0.462	0.368	0.277	[0.234, 0.320]	99	0.413	0.932
DeepSeek v4 Flash	99	0.451	0.353	0.264	[0.222, 0.306]	98	0.304	0.940
Qwen 3.6-35B-A3B	98	0.445	0.344	0.254	[0.212, 0.295]	96	0.191	0.922
GPT-5 Mini	90	0.404	0.311	0.228	[0.187, 0.270]	90	0.347	0.920
Claude Sonnet 4.6	73	0.248	0.267	0.207	[0.164, 0.252]	73	0.332	0.866
Pipeline mode, K=15								
Gemini 2.5 Pro	99	0.424	0.425	0.293	[0.249, 0.337]	47	0.463	0.917
DeepSeek v4 Flash	99	0.430	0.377	0.269	[0.227, 0.313]	56	0.316	0.927
GPT-5 Mini	96	0.390	0.384	0.264	[0.224, 0.303]	43	0.479	0.924
Claude Sonnet 4.6	99	0.371	0.317	0.221	[0.182, 0.263]	41	0.345	0.810
Agent mode								
Gemini 2.5 Pro	96	0.470	0.304	0.251	[0.210, 0.292]	62	0.717	0.948
GPT-5 Mini	98	0.337	0.468	0.211	[0.174, 0.250]	60	0.633	0.906

Table 4: Set-based benchmark results over all 99 ANTHEM comparison questions. P, R, and F1 are means over the 99 questions of the corresponding per-question value, with failed or unparseable outputs scored as zero; F1 is therefore the mean of per-question F1 rather than the harmonic mean of the columns to its left. n is the number of questions with parseable IR-scored output, reported as a diagnostic. Intervals are 95% percentile CIs from a 10k paired bootstrap over questions.

faithfulness gain over the matched pipeline, comes from the same architectural lever in the opposite direction: analyzing a paper’s full text before citing it grounds claims in the paper’s actual content rather than in its abstract. The implication for practitioners building comparative-QA systems on ANTHEM is that retrieval competence and citation grounding are separable engineering targets; the architectural choice between modes determines which one is easier to optimise.

Programmatic gold and longitudinal benchmarking. The benchmark’s gold sets are computed against the corpus snapshot rather than annotated separately (§4.1), which makes them reproducible, naturally longitudinal, and as reliable as the underlying corpus. A corollary is that absolute metric values are anchored to the snapshot they were computed against, so cross-system comparisons should reference a common archived snapshot. The corpus has grown across internal snapshots between January and May 2026, and users can filter the corpus on year or venue. The benchmark is one operationalisation of the corpus among several the composition supports: topic-conditioned institutional trajectories, geographic shifts within a topic, methodological gaps between countries, and hybrid extraction systems that route by publication year are all downstream uses.

6 Conclusion

We introduced ANTHEM, a dataset and benchmark for institutional analysis of NLP research. ANTHEM-202605 links 117,386 ACL Anthology papers to ROR-normalized affiliations across 5,580 institutions in 150 countries, and to CSO topic labels covering 98.3% of the corpus. Our extraction pipeline achieves $F1_{\text{hier}}$ of 0.760, treating parent-child ROR identifiers as equivalent, on a 781-paper decade-stratified gold set, leading on post-2010 papers while OpenAlex complements it most effectively on pre-2010 records. We additionally document that S2ORC’s structured author–affiliation spans, densely populated across the scholarly literature as a whole, are sparsely populated on ACL records in the current bulk release. On top of the corpus we release a comparative-QA benchmark of 99 structured questions across five categories, evaluated under a programmatic gold-set construction in which ANTHEM’s own metadata defines the ground truth, enabling both absolute comparisons against any archived snapshot and longitudinal re-evaluation as the corpus evolves. All data, evaluation infrastructure, and code are released under CC-BY-SA-NC. Our data⁴ and our code⁵ are available online.

⁴<https://huggingface.co/datasets/Molarix/anthem>

⁵<https://github.com/MoritzStaudinger/anthem>

Limitations

Scope of analysis. ANTHEM is designed as an infrastructure dataset and benchmark for institutional analysis of NLP research, not as a complete scientometric study. The longitudinal statistics in Appendix A are intentionally descriptive and serve primarily to validate the normalized meta-data; deeper findings about institutional research patterns (topic-conditioned trajectories, causal analyses of geographic shifts, finer-grained community structure) are downstream uses we leave to future work building on ANTHEM.

Source data and affiliation ambiguity. ANTHEM inherits structural limitations from the ACL Anthology. Some entries lack accessible or valid full-text PDFs,⁶ contain damaged files,⁷ or use non-standard affiliation formats that complicate automated extraction (see Appendix B.6). Even with access to full text, some affiliation strings lack sufficient information to uniquely identify an institution, particularly for organizations with multiple geographic branches (e.g., *Meta* or *Microsoft Research*). Rather than enforcing potentially incorrect assignments, ANTHEM preserves unmatched and ambiguous affiliations alongside normalized ROR matches, prioritizing transparency over forced completeness and allowing downstream users to control precision-recall trade-offs. A complementary direction would be linking authors to their ORCID IDs (already stored on OpenReview) and extending the Responsible NLP Research guidelines to require ROR identifiers at submission time; with such information, data quality would improve substantially.

Validation sample size. Our 781-paper gold set, while substantially larger and more temporally diverse than prior affiliation-extraction validations, still represents under 1% of the corpus. The decade stratification captures historical layout heterogeneity, but precision on rare layout patterns or low-frequency institution types remains correspondingly uncertain. Annotation was carried out by the authors, which is standard for resource papers but does not provide an estimate of how a wider annotator pool would resolve genuinely ambiguous cases.

Runtime characteristics. The extraction pipeline is not designed for low-latency online

use, but the experimental configuration (Gemini 2.5 Flash over combined parser outputs) is highly parallelisable: requests against commercial LLM APIs can be batched concurrently, allowing corpus-wide extraction to complete in wall-clock hours. The LoRA-tuned Qwen 2.5 7B baseline (Appendix B.7) takes 20–30 seconds per paper on a single A100; extending it to the full corpus would require roughly $30 \times 117,000$ GPU-seconds, on the order of a thousand A100-hours, which is infeasible to block a shared cluster for as a periodic operation. This asymmetry between commercial API parallelism and local-GPU throughput is the main practical reason the experimental ANTHEM pipeline uses a commercial LLM rather than the locally-runnable open baseline, and is a structural consideration for any group planning to reproduce ANTHEM on a fully self-hosted stack.

Benchmark scope. The 99 questions target five comparative-analysis categories but do not exhaust the space of useful comparative questions over a scientific corpus. The size is a deliberate, cost-driven design choice: a single sweep across the systems we evaluate already requires several hundred long-form generation runs, with RAGAS faithfulness and relevancy scoring at least doubling the LLM call count. A benchmark an order of magnitude larger would be straightforward to generate from our rule-based templates but would multiply the cost of every system evaluation accordingly. We treat 99 questions as the practical balance between per-category coverage and tractability of running new systems. Per-question gold summaries are not provided: a single field-evolution question can have a gold set of more than 24,000 papers, and what constitutes a “correct” synthesis of a research landscape is genuinely contested and would require multiple expert annotators to assess. The programmatic constraint-based gold set (§4.1) sidesteps this by evaluating against structured criteria checkable from the corpus.

Retrieval configuration. We report results under a single dense retrieval configuration (all-MiniLM-L6-v2 over abstracts, at $K=5$ and $K=15$). Stronger retrievers (E5, BGE, scientific-domain embeddings) and longer-context generation are likely to raise absolute numbers, and we do not claim the system ordering in Table 4 is invariant to the choice of encoder. The structural ceiling, however, is: recall is bounded by the size of the merged candidate set relative to $|\text{GOLD}|$ (§D.1), so

⁶Such as 2023.splurobonlp-1.0, W18-4422, or L14-1626.

⁷Such as 2023.acl-short.85, L14-1636.

field-evolution questions, whose gold sets run into the thousands, measure ranking quality rather than retrieval coverage under any encoder.

Judge robustness. Faithfulness and relevancy are RAGAS metrics scored by an LLM judge, with Qwen 3.5-397B as our default. Re-scoring the matched Gemini 2.5 Pro pipeline/agent pair under a second judge (§D.2) leaves the direction of the agent’s faithfulness advantage intact but changes its magnitude by more than a factor of two, and pass/fail agreement between the two judges is poor (κ 0.093). Absolute faithfulness values should therefore be read as judge-relative, and faithfulness figures from papers using different judges are not comparable to ours. We validate one matched pair under a second judge and no configuration against human ratings; multi-judge consensus and a human-agreement study would strengthen these metrics and are a natural extension.

Use of large language models. ANTHEM relies on LLMs for affiliation extraction, benchmark question generation, and answer synthesis. To mitigate reproducibility and auditability concerns, extracted affiliations are stored together with their raw textual evidence spans and parser representations, and benchmark questions are released with their structured metadata. This treats LLMs as a replaceable component rather than a fixed dependency, allowing decisions to be audited or reprocessed as models improve.

Ethics Statement

ANTHEM supports aggregate, descriptive analysis. It should not be used to evaluate individual researchers or to rank institutions: publication counts derived from it reflect front-matter text after automated extraction, not research quality.

Residual matching errors are systematic, not random. Extraction failures concentrate in pre-2000 papers with degraded OCR, linking failures concentrate in branched corporate entities, restructured institutions, and national conventions that differ from the Anglophone norm (§B.6). ROR registry coverage is itself thinner for older institutions and less well-resourced research systems. Error rates therefore correlate with era, region, and institution type, so a geographic analysis is least reliable for precisely the under-represented institutions it is usually about, and an apparent gap between countries may partly reflect linking quality. We report

per-decade performance (§B.1) and per-year ROR coverage (§A.4) to bound this, and recommend checking any institution-specific claim against the released raw strings and evidence spans rather than the ROR-linked view alone. Every decision is stored with its provenance so records can be audited; a linking error found during review (*Wrocław University of Science and Technology* resolved to *University of Wrocław*) is corrected.

The corpus reproduces author names and affiliations already public in the ACL Anthology, without augmentation: we infer no attribute not stated in the source document and do not disambiguate authors across papers. Affiliations are recorded as of publication.

Use of AI Assistants

We used AI-based writing assistants (Claude Code with Claude Opus 4.8 and Claude Opus 5) for language editing and for support in writing and debugging analysis and evaluation code. All research design, experimental decisions, analysis, and claims are the authors’ own, and all generated text and code were reviewed and verified by the authors.

References

- Mohamed Abdalla, Jan Philip Wahle, Terry Lima Ruas, Aurélie Névél, Fanny Ducel, Saif Mohammad, and Karen Fort. 2023. *The Elephant in the Room: Analyzing the Presence of Big Tech in Natural Language Processing Research*. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13141–13160, Toronto, Canada. Association for Computational Linguistics.
- Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2025. *LitLLMs, LLMs for Literature Review: Are we there yet?* *arXiv preprint*. ArXiv:2412.15249 [cs].
- Sergio Alonso, Francisco Javier Cabrerizo, Enrique Herrera-Viedma, and Francisco Herrera. 2009. h-index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of informetrics*, 3(4):273–289.
- Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D’arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, and 6 others. 2024. *OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs*. *arXiv preprint*. ArXiv:2411.14199 [cs].

- Sören Auer, Dante A. C. Barone, Cassiano Bartz, Eduardo G. Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, Ivan Shilin, Markus Stocker, and Eleni Tsalapati. 2023. [The SciQA Scientific Question Answering Benchmark for Scholarly Knowledge](#). *Scientific Reports*, 13(1):7240. Publisher: Nature Publishing Group.
- Ehsan Bitaraf and Maryam Jafarpour. 2024. [MIE Articles Dataset](#). Artwork Size: 11453824 Bytes Pages: 11453824 Bytes.
- Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. [Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases](#). *Humanities and Social Sciences Communications*, 8(1):1–15. Publisher: Palgrave.
- Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*.
- Huy Hoang Nhat Do, Muthu Kumar Chandrasekaran, Philip S. Cho, and Min Yen Kan. 2013. [Extracting and matching authors and affiliations in scholarly documents](#). In *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '13*, page 219–228, New York, NY, USA. Association for Computing Machinery.
- Nicolau Duran-Silva, Pablo Accuosto, Piotr Przybyła, and Horacio Saggion. 2024. [AffilGood: Building reliable institution name disambiguation tools to improve scientific literature analysis](#). In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 135–144, Bangkok, Thailand. Association for Computational Linguistics.
- Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. [RAGAs: Automated evaluation of retrieval augmented generation](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.
- Najko Jahn. 2025. [Decreasing affiliation metadata coverage in OpenAlex](#). *Scholarly Communication Analytics*.
- Renars Liepins, Ulrich Germann, Guntis Barzdins, Alexandra Birch, Steve Renals, Susanne Weber, Peggy van der Kreeft, Hervé Bourlard, João Prieto, Ondřej Klejch, Peter Bell, Alexandros Lazaridis, Alfonso Mendes, Sebastian Riedel, Mariana S. C. Almeida, Pedro Balage, Shay B. Cohen, Tomasz Dwojak, Philip N. Garner, and 27 others. 2017. [The SUMMA platform prototype](#). In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–119, Valencia, Spain. Association for Computational Linguistics.
- Lianghong Lin, Wenxiu Xie, Zili Chen, and Tianyong Hao. 2024. [An end-to-end entity recognition and disambiguation framework for identifying Author Affiliation from literature publications](#). In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 120–129, Bangkok, Thailand. Association for Computational Linguistics.
- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. 2020. [S2ORC: The Semantic Scholar Open Research Corpus](#). *arXiv preprint ArXiv:1911.02782* [cs].
- Patrice Lopez. 2009. [GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications](#). In *Research and Advanced Technology for Digital Libraries*, pages 473–474, Berlin, Heidelberg. Springer.
- Emilio Delgado López-Cózar, Nicolas Robinson-Garcia, and Daniel Torres-Salinas. 2012. Manipulating google scholar citations and google scholar metrics: Simple, easy and tempting. *arXiv preprint arXiv:1212.0638*.
- Wenkai Luo, Malindu Sandanayake, Lei Hou, Yongtao Tan, and Guomin Zhang. 2022. [A systematic review of green construction research using scientometrics methods](#). *Journal of Cleaner Production*, 366:132710.
- John Mingers and Loet Leydesdorff. 2015. A review of theory and practice in scientometrics. *European journal of operational research*, 246(1):1–19.
- Saif M. Mohammad. 2020. [NLP scholar: A dataset for examining the state of NLP research](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 868–877, Marseille, France. European Language Resources Association.
- David Opeoluwa Oyewola and Emmanuel Gbenga Dada. 2022. [Exploring machine learning: a scientometrics approach using bibliometrix and VOSviewer](#). *SN Applied Sciences*, 4(5):143.
- Monarch Parmar, Naman Jain, Pranjali Jain, P. Jayakrishna Sahit, Soham Pachpande, Shruti Singh, and Mayank Singh. 2020. [Nlpexplorer: Exploring the universe of nlp papers](#). In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part II*, page 476–480, Berlin, Heidelberg. Springer-Verlag.
- Jason Priem, Heather Piwowar, and Richard Orr. 2022. [OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts](#). *arXiv preprint ArXiv:2205.01833* [cs].
- Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. [The ACL Anthology network corpus](#). In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLP4DL)*, pages 54–61, Suntec City, Singapore. Association for Computational Linguistics.

- Mukund Rungta, Janvijay Singh, Saif M. Mohammad, and Diyi Yang. 2022. [Geographic Citation Gaps in NLP Research](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1371–1383, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Angelo A. Salatino, Francesco Osborne, Thiviyan Thanapalasingam, and Enrico Motta. 2019. [The CSO Classifier: Ontology-Driven Detection of Research Topics in Scholarly Articles](#), page 296–311. Springer International Publishing.
- Amanpreet Singh, Joseph Chee Chang, Chloe Anastasiades, Dany Haddad, Aakanksha Naik, Amber Tanaka, Angele Zamarron, Cecile Nguyen, Jena D. Hwang, Jason Dunkleberger, Matt Latzke, Smita Rao, Jaron Lochner, Rob Evans, Rodney Kinney, Daniel S. Weld, Doug Downey, and Sergey Feldman. 2025. [Ai2 Scholar QA: Organized Literature Synthesis with Attribution](#).
- Moritz Staudinger, Wojciech Kusa, Matteo Cancellieri, David Pride, Petr Knöth, and Allan Hanbury. 2025. [Compare: A framework for scientific comparisons](#). In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25*, page 6703–6707, New York, NY, USA. Association for Computing Machinery.
- Yixuan Tang and Yi Yang. 2024. [MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries](#). *arXiv preprint*. ArXiv:2401.15391 [cs].
- Rong-Cheng Tu, Lei Ji, Huaishao Luo, Botian Shi, Heyan Huang, Nan Duan, and Xian-Ling Mao. 2021. [Hashing based efficient inference for image-text matching](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 743–752, Online. Association for Computational Linguistics.
- Jan Philip Wahle, Terry Ruas, Saif Mohammad, and Bela Gipp. 2022. [D3: A massive dataset of scholarly metadata for analyzing the state of computer science research](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2642–2651, Marseille, France. European Language Resources Association.
- Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. [Microsoft Academic Graph: When experts are not enough](#). *Quantitative Science Studies*, 1(1):396–413.
- Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024. [Leave No Document Behind: Benchmarking Long-Context LLMs with Extended Multi-Doc QA](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5627–5646, Miami, Florida, USA. Association for Computational Linguistics.
- Penghui Wei, Nan Xu, and Wenji Mao. 2019. [Modeling conversation structure and temporal dynamics for jointly predicting rumor stance and veracity](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4787–4798, Hong Kong, China. Association for Computational Linguistics.
- Leonie Weissweiler, Valentin Hofmann, Anjali Kantharuban, Anna Cai, Ritam Dutt, Amey Hengle, Anubha Kabra, Atharva Kulkarni, Abhishek Vijayakumar, Hao-fei Yu, Hinrich Schuetze, Kemal Oflazer, and David Mortensen. 2023. [Counting the bugs in ChatGPT’s wugs: A multilingual investigation into the morphological capabilities of a large language model](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6508–6524, Singapore. Association for Computational Linguistics.
- Jiashu Xu, Fei Wang, Mingyu Ma, Pang Wei Koh, Chaowei Xiao, and Muhao Chen. 2024. [Instructional fingerprinting of large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3277–3306, Mexico City, Mexico. Association for Computational Linguistics.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering](#). *arXiv preprint*. ArXiv:1809.09600 [cs].

A Detailed Data Analysis

This appendix presents longitudinal statistics derived from the ANTHEM-202605 corpus. We note upfront that the aggregate trends shown here, growth in publication volume, authors per paper, institutional collaboration, and geographic diversification, are not novel scientific findings. They have been documented in the ACL Anthology’s own summary statistics, in scientometric studies of NLP (Rungta et al., 2022; Abdalla et al., 2023), and in broader bibliometric work (Bormmann et al., 2021). We include them for two reasons: they validate that the normalized metadata in ANTHEM reproduces the qualitative trends one would expect from independent sources, and the same analyses become tractable once full-text content is linked to normalized institutional and topical metadata in a single resource.

Data before 1990 is sparse (between 100 and 400 publications per year; Figure 3), so we restrict the analyses below to 1990–2025. Annual publication

counts grow from roughly 500 in the early 1990s to over 10,000 per year in the early 2020s.

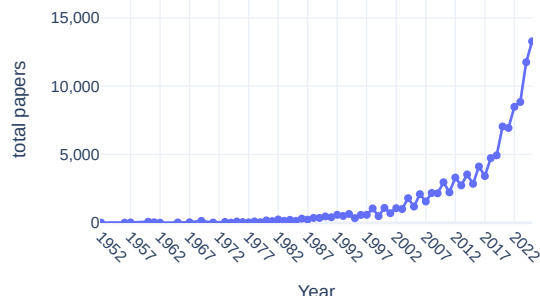


Figure 3: Annual ACL Anthology publications, 1952–2025.

A.1 Authors and Affiliations per Paper

The growth in publication volume between 1990 and 2025 is accompanied by an increase in average team size. The average number of authors per paper rises from approximately 3 in 1990 to over 7 in 2025 (Figure 4), while the average number of distinct ROR-linked institutions per paper increases more modestly, from roughly 1.5 to 2 (Figure 5). The asymmetry suggests that individual institutions are now contributing larger teams to single papers, rather than the collaboration pattern shifting toward many institutions contributing one author each.

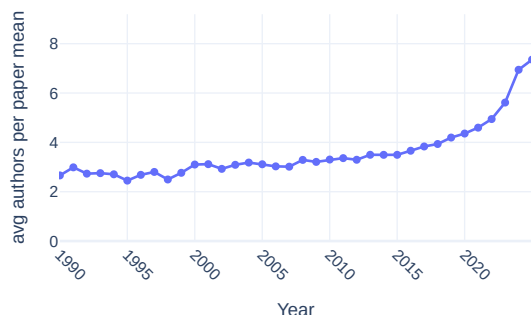


Figure 4: Average authors per publication, 1990–2025.

A.2 Multi-Institution and Multi-Country Collaboration

Figure 6 shows the proportion of publications involving multiple institutions and multiple coun-

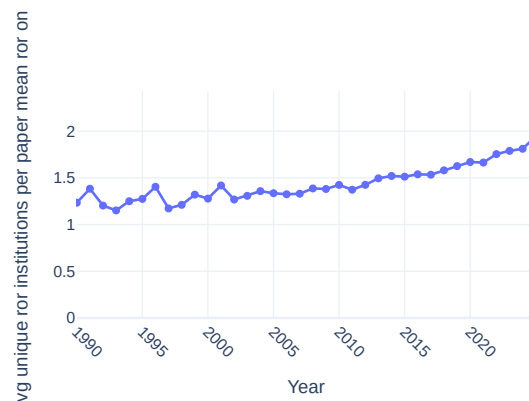


Figure 5: Average institutions per publication, 1990–2025 (ROR-matched institutions only).

tries.⁸ Multi-institution collaboration grows from 6.4% in 1990 to 46.7% in 2025, and multi-country collaboration from 3% to 26.6% over the same period. Identifying multi-institution and multi-country publications at scale is only possible with normalized affiliations: raw strings such as “University of California, Berkeley” and “UC Berkeley” would otherwise be counted as distinct institutions, which ANTHEM’s ROR-linked metadata avoids.

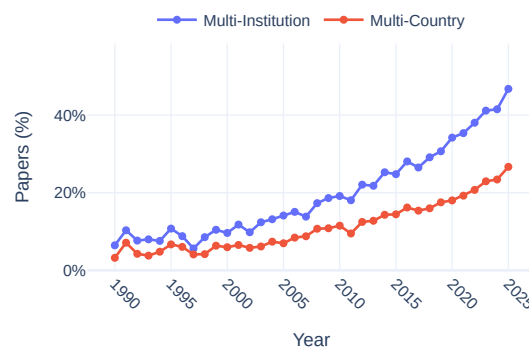


Figure 6: Trends in multi-institution and multi-country collaboration in NLP research, 1990–2025.

A.3 Country-Level Participation Over Time

Figure 7 shows the proportion of publications associated with the five most frequent countries in the corpus. NLP research in the early 1990s is strongly concentrated in the US and GB, which together account for approximately 80% of yearly contribu-

⁸Percentages are computed over publications with at least one extracted affiliation. Earlier publications in the ACL Anthology occasionally bundle multiple papers in a single PDF file, which can slightly affect early-decade collaboration counts.

tions between 1990 and 1993. Participation diversifies over time, with the most pronounced change being the rise of China from roughly 10% in 2015 to around 30% by 2025. Other high-frequency countries maintain relatively stable participation below 10% of yearly output.

The total number of contributing countries grows substantially over the same period (Figure 8): only 50 countries contributed in 2000, rising to 121 in 2024 and 119 in 2025. This expansion provides empirical motivation for the cross-country method comparison category in the ANTHEM benchmark (§4).

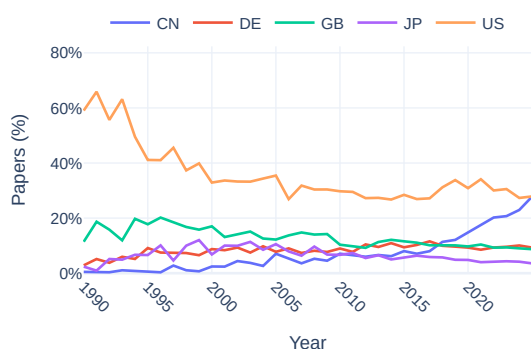


Figure 7: Country-level participation in NLP research over time. Shown are the proportions of publications associated with the five countries with the most publications overall.

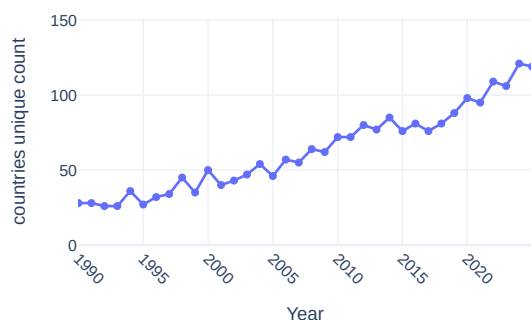


Figure 8: Number of distinct countries contributing at least one publication, 1990–2025.

A.4 ROR Match Coverage Over Time

A natural concern with corpus-wide affiliation analysis is whether the fraction of raw affiliation strings that do not link to ROR is systematically biased toward specific time periods, which would distort

longitudinal analyses. Figure 9 shows the percentage of raw affiliation instances per year matched to a ROR identifier above the 90% confidence threshold.

ROR coverage rises from below 40% in 1990 to nearly 80% in 2025. Two factors explain this pattern: ROR itself has more complete registry coverage for institutions active in recent decades than for historical or restructured institutions, and layout heterogeneity in older ACL publications increases extraction errors upstream of linking. The error analysis in §B.5 corroborates the second factor: extraction failures, as distinct from linking failures, are concentrated in pre-2000 papers. Users planning institutional analyses on older NLP research should account for this elevated extraction-side noise and supplement the ROR-linked view with raw-string search where the linker is known to be brittle.

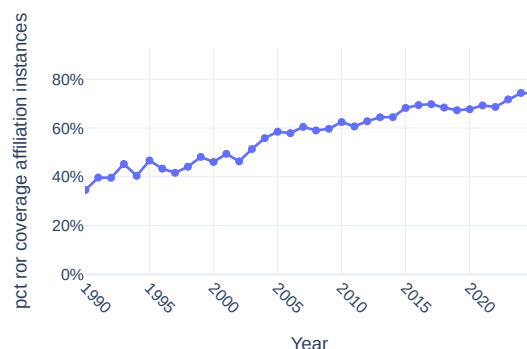


Figure 9: ROR match coverage of raw affiliation instances per year, 1990–2025.

A.5 Institutional and Topical Composition

Figure 10 shows a larger version of the four-level institutional-topical treemap summarised in Figure 2, covering all institutions with contribution share above the rendering threshold over 2020–2025. The dense view supports investigation of individual institutions (e.g., second-tier contributors within a country, region-specific clusters such as the Hong Kong universities, or industry labs in specific countries) and the topical specialisation of each. As with the main-paper figure, the view is reproducible from the released corpus via the analysis code accompanying ANTHEM, and re-running it against future snapshots tracks how the institutional and topical composition of NLP research evolves over time.

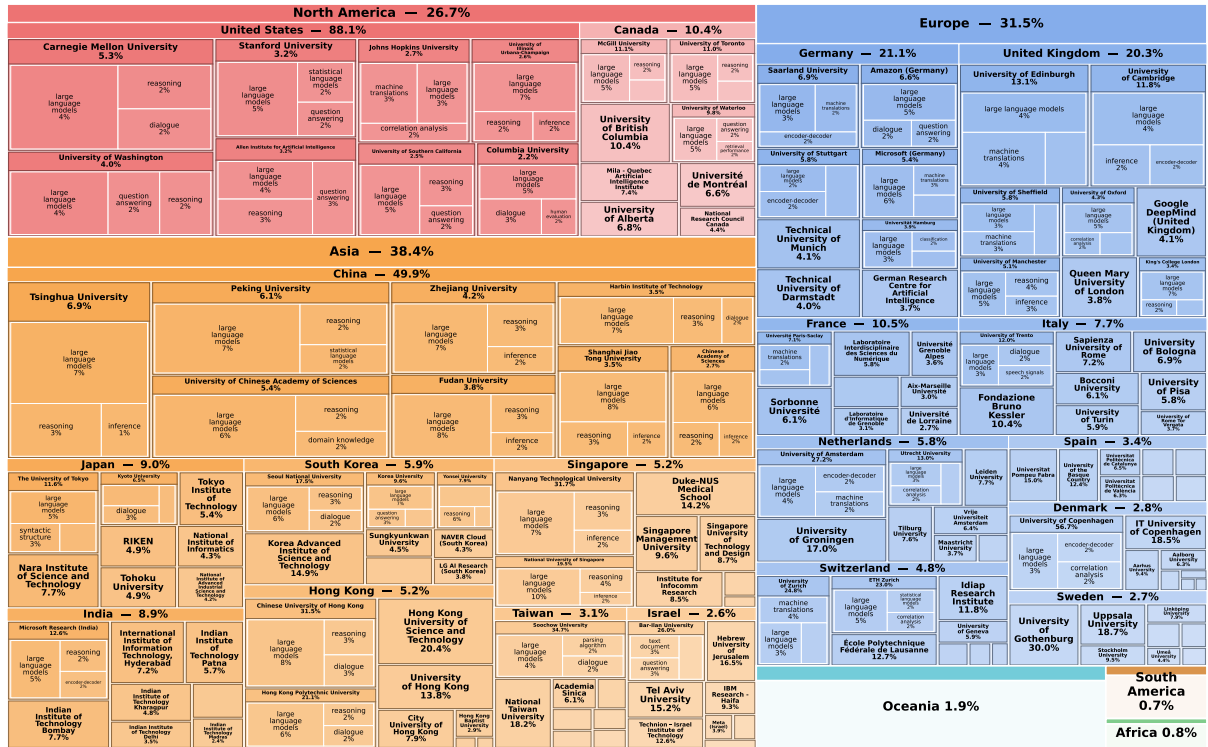


Figure 10: Full institutional and topical composition of ACL Anthology technical research, 2020–2025, computed from ANTHEM-202605. All institutions above the rendering threshold are shown, with top topics per institution; cf. the abridged main-paper version in Figure 2.

B Affiliation Extraction Performance

This appendix reports per-decade performance (§B.1), strict ROR matching (§B.2), topic-label reliability (§B.3), a comparison of ROR linking strategies (§B.4), an error-type decomposition (§B.5), and qualitative failure modes (§B.6); and it documents the LoRA-tuned open baseline (§B.7).

Two conventions apply throughout. First, §B.1 and §B.5 report scores on each source’s covered subset, so a low-coverage external source is scored only where it produces output; scores in §B.2 and §B.4 are computed over the full gold set and are comparable to Table 2.

Second, §B.2 and §B.4 are scored against the blind-verified gold (§3.4), while §B.1 and §B.5 are scored against the pre-verification gold; the two differ by at most 0.03 $F1_{\text{hier}}$ where both are available, and we report the verified figures wherever a number appears in the main text.

B.1 Per-Decade Performance

Table 5 reports covered $F1_{\text{hier}}$ per decade. The ANTHEM pipeline dominates on modern papers (0.761/0.836 on the 2010s/2020s) but is middling on pre-2000 papers, where layout het-

erogeneity and OCR errors degrade PDF-based extraction. OpenAlex shows the opposite pattern, leading pre-2010 (0.800/0.691/0.692 on the 1980s/1990s/2000s) and dropping below 0.65 from the 2010s onward. No single source is optimal across the full historical Anthology; year-based routing (OpenAlex for pre-2010, our pipeline for newer) is a natural extension that we leave to future work, releasing the per-decade evaluation infrastructure to support it.

Pipeline	80s	90s	00s	10s	20s	All
Gemini 2.5 + comb.	0.60	0.66	0.65	0.76	0.84	0.75
Claude Sonnet + comb.	0.58	0.70	0.73	0.79	0.71	0.72
GPT-5 Nano + comb.	0.55	0.71	0.72	0.79	0.70	0.71
Gemini 2.5 + PyMu	0.63	0.63	0.71	0.70	0.66	0.68
OpenAlex	0.80	0.69	0.69	0.65	0.60	0.64
GROBID	0.55	0.68	0.59	0.66	0.61	0.62
S2 Graph API	0.02	0.07	0.10	0.14	0.13	0.12
S2ORC	0.04	0.07	0.00	0.00	0.00	0.01
Papers per decade	91	92	187	197	214	781

Table 5: Covered $F1_{\text{hier}}$ per decade on the 781-paper gold set.

B.2 Strict ROR Matching

Table 6 reports $F1_{\text{ROR}}$, which requires an exact ROR-id set match and admits no parent-child equivalence, alongside the exact interval bounds underlying the half-widths in Table 2. Both variants are scored against the blind-verified gold. The difference between them isolates the cost of the hierarchy rule: for our pipeline it is 0.058 (0.702 against 0.760), and the ordering at the top of the table is unchanged. The one system whose position depends on the rule is the LoRA-tuned Qwen 2.5 7B baseline, which places third under $F1_{\text{hier}}$ but fourth under $F1_{\text{ROR}}$. S2ORC scores identically under both variants, as its coverage is too sparse for the rule to apply.

Pipeline	$F1_{\text{ROR}}$	$F1_{\text{hier}}$
Gemini 2.5 Flash + comb.	0.702 $\pm$.029	0.760 $\pm$.028
Claude Sonnet + comb.	0.656 $\pm$.033	0.703 $\pm$.032
GPT-5 Nano + comb.	0.638 $\pm$.034	0.687 $\pm$.033
Qwen 7B (LoRA) + comb.	0.637 $\pm$.029	0.698 $\pm$.027
Claude Sonnet + PyMu	0.629 $\pm$.039	0.676 $\pm$.039
Gemini 2.5 + PyMu	0.605 $\pm$.030	0.653 $\pm$.029
Claude Haiku + comb.	0.603 $\pm$.031	0.650 $\pm$.031
GROBID (header parser)	0.587 $\pm$.032	0.639 $\pm$.030
OpenAlex (graph API)	0.400 $\pm$.040	0.431 $\pm$.041
Semantic Scholar Graph	0.104 $\pm$.024	0.126 $\pm$.026
S2ORC (bulk dataset)	0.009 $\pm$.009	0.009 $\pm$.009

Table 6: Strict and hierarchical ROR matching on the 781-paper gold set, scored against the blind-verified gold. $F1_{\text{ROR}}$ requires an exact ROR-id set match; $F1_{\text{hier}}$ additionally accepts an immediate parent or child of the gold identifier. The $\pm$ term is the half-width of a 95% percentile interval from a 10k paired bootstrap over papers; intervals are near-symmetric (largest asymmetry 0.003). Rows are sorted by $F1_{\text{ROR}}$.

B.3 Topic-Label Reliability

We annotated 104 paper-CSO label pairs manually (single annotator, Wilson 95% intervals): precision is 0.713 [0.610, 0.797] overall, 0.737 [0.610, 0.834] for labels with in-text evidence and 0.667 [0.488, 0.808] for semantically inferred labels; 16% of pairs were judged uncertain. Counting uncertain cases as errors gives a lower bound of 0.596 [0.500, 0.685]. We report this as the reliability bound on the topic layer and on all topic-conditioned gold sets in §4.

B.4 ROR Linking Strategies

Holding extraction fixed at the gold affiliation strings, we compare four ROR linking strategies against the 802 hand-verified links (Table 7). Our

country-context-aware linker outperforms all three alternatives, every paired difference excluding zero: +0.037 [+0.017, +0.057] over ROR’s chosen-only mode and +0.085 [+0.065, +0.106] over top-1-by-score. ROR’s chosen-only mode is the higher-precision operating point (P 0.975 at R 0.708), preferable where false links are costlier than missing ones; the next snapshot releases full candidate lists so users can select either. The spread across strategies (0.037–0.085) is comparable to our end-to-end margin over the strongest LLM ablation (+0.058 [+0.028, +0.088], Table 2), consistent with the linking layer rather than extraction accounting for the advantage.

Strategy	P	R	$F1_{\text{hier}}$
Ours (country-aware)	0.845	0.870	0.857 $\pm$.020
ROR chosen-only	0.975	0.708	0.820 $\pm$.020
Chosen (legacy)	0.808	0.794	0.801 $\pm$.026
Top-1 by score	0.757	0.787	0.772 $\pm$.024

Table 7: ROR linking strategies on the 802 hand-verified affiliation-ROR links, extraction held fixed at gold strings. $\pm$ is the half-width of a 95% percentile interval (10k paired bootstrap); all differences from our default exclude zero.

B.5 Error-Type Decomposition

We decompose predictions into true positives (TP, gold match including parent/child equivalence), wrong-organisation false positives (FP_{org}), and two kinds of false negatives: *linking failures* (FN_{link} , institution extracted but linker resolves incorrectly or returns no match) and *extraction failures* (FN_{extract} , institution absent from the extracted text). Table 8 reports the counts.

The dominant failure mode for LLM-based pipelines is linking, not extraction: 90% of AN-THM’s false negatives are linking failures (96% for Claude Sonnet and GPT-5 Nano). The LoRA-tuned Qwen 2.5 7B exhibits the opposite imbalance (20% linking failures, extraction failures dominating), and GROBID sits between (54% linking). Decade-stratified, linking failures dominate across all decades but peak in modern papers (87% in 2020s, driven by geographically branched corporate entities and joint-affiliation conventions), while extraction failures are relatively more common in older decades (33% in 1980s, driven by OCR errors). We additionally identify 31 papers where no pipeline in our evaluation reached $F1 \geq 0.5$; these concentrate in pre-2000 papers with

severe OCR degradation and modern papers with very large author lists (60+ authors across 20+ institutions).

B.6 Qualitative Failure Modes

Manual annotation of the gold set surfaced recurring failure patterns, illustrated with concrete examples from the ACL Anthology.

Symbol-based author-affiliation linking. When papers use unusual superscript symbols (daggers, double daggers, custom glyphs), the LLM’s ability to recover author-affiliation alignment degrades, as illustrated by [Liepins et al. \(2017\)](#); [Tu et al. \(2021\)](#) with complex symbol vocabularies over large author lists.

Numeric ranges as affiliation markers. Some recent papers use compressed notation such as “2-5” to indicate affiliations 2 through 5 simultaneously. All extractors in our evaluation treat “2-5” as a single token rather than expanding the range, causing middle affiliations to be silently dropped (e.g., [Weissweiler et al. \(2023\)](#)).

Images as affiliation markers. A growing number of papers use institutional logos in the affiliation block; none of our parsing strategies recovers affiliation information from embedded images (e.g., [Xu et al. \(2024\)](#)).

Missing or inconsistent affiliation symbols. Symbol-based linking presumes that every author and every affiliation carries a marker. When markers are omitted on one side, alignment can only be recovered via positional heuristics that are unreliable for long author lists (e.g., [Wei et al. \(2019\)](#)).

Ambiguous and multi-affiliation conventions. “Meta” without geographic specification could refer to Meta US, Meta Ireland, or Meta UK; “IBM Research” can refer to several country-level entities or a corporate parent. French academic papers frequently list a host institution alongside CNRS or partner-laboratory units without a clear canonical affiliation. City-level clusters (e.g. “University of Paris” versus “Sorbonne University,” both with multiple ROR successors after restructuring) and the boundary between institution and location (“MIT” versus “MIT, Cambridge, MA”) produce annotator-pipeline disagreement even when the text is identical.

These failure modes are not uniformly distributed: encoding issues in PDFs concentrate in

pre-2000 papers, while images-as-affiliations, complex symbol conventions, and very large author lists concentrate in 2017+ papers. Users should treat per-decade F1 in Table 5 as a guide to where the metadata is most reliable, and supplement ROR-linked queries with raw-string search where the linker is known to be brittle.

B.7 LoRA-Tuned Open Baseline

The Qwen 2.5 7B (LoRA) row characterises what an institution that cannot or will not send PDFs to commercial APIs can achieve with a single A100. Training data: 9,000 papers (1,000 held out for validation), constructed by joining the PyMu4LLM Markdown parse of each PDF with ACL Anthology DOI metadata and OpenAlex works/doi : lookups for ROR-linked author-affiliation gold labels; the 781 evaluation papers were excluded from the candidate pool before sampling to guarantee zero train/test contamination. The training corpus spans 1979–2025 with a strong recency skew ($\geq 90\%$ from 2014 onward) driven by DOI availability in OpenAlex.

The base model is Qwen2.5-7B-Instruct, loaded with 4-bit nf4 double quantisation (bfloat16 compute). LoRA adapters are applied to all linear projections in attention and MLP blocks ($r=16$, $\alpha=32$, dropout 0.05) and trained for 3 epochs (effective batch size 16, learning rate 2×10^{-4} cosine, max sequence length 2,048), in 3h 31min on one A100 80GB, reaching final validation loss 0.083 and 98.17% eval-token accuracy. Inference uses the same 4-bit base plus adapter and produces compact JSON via greedy single-shot generation; JSON parse failures emit an empty author list rather than silently skipping.

The trained model reaches $F1_{\text{hier}}$ of 0.698, placing third among the configurations we evaluate and ahead of both PyMu-based variants and Claude Haiku, though it falls behind on strict matching ($F1_{\text{ROR}}$ 0.637, Table 6) and its errors are dominated by extraction rather than linking failures (Table 8). The training-data recency skew is consistent with the model’s relatively weak per-decade scores on older papers (Table 5). All scripts, configurations, training data, OpenAlex query cache, and the trained adapter are released alongside ANTHEM.

C S2ORC Coverage Breakdown

Section 3.3 reports that S2ORC’s structured authoraffiliation spans, densely populated

Pipeline	TP	FP _{org}	FN _{link}	FN _{extr}	% link
Gemini 2.5 Flash + comb.	1809	538	730	85	90
Claude Sonnet + comb.	2109	712	496	19	96
GPT-5 Nano + comb.	2090	929	515	19	96
Claude Haiku + PyMu	2084	937	483	57	89
Gemini 2.5 + PyMu	1951	795	436	237	65
GROBID	1456	634	634	534	54
Qwen-7B (LoRA)	1336	332	258	1030	20

Table 8: Error-type decomposition on the 781-paper gold set. % link is the share of false negatives caused by linking failures.

across the scholarly literature as a whole, are sparsely populated on ACL Anthology records in the current bulk release (Lo et al., 2020). This appendix provides the per-stage coverage breakdown.

We measured the S2ORC bulk release version 2026-05-12 on the ACL Anthology subset. Table 9 reports the count of records remaining at each stage of the lookup pipeline, from initial paper-level coverage in the Semantic Scholar papers dataset through to the structured authoraffiliation annotations.

Stage	Count	% of ACL
ACL papers in S2 papers dataset	92,790	76%
With s2orc fulltext record	52,013	43%
With content.annotations	1,570	1.3%
With authoraffiliation spans	978	0.8%

Table 9: S2ORC structured-spans coverage on the ACL Anthology subset (release 2026-05-12, approximately 121,000 ACL papers).

The drop from 92,790 ACL papers in the Semantic Scholar papers dataset to 978 papers with structured affiliation spans is concentrated in two transitions. First, 56% of ACL papers known to Semantic Scholar have a corresponding s2orc full-text record (52,013 of 92,790). Second, of those, 3% (1,570 of 52,013) carry any non-empty content.annotations, and 62% of those (978 of 1,570) contain the specific authoraffiliation span type.

D Benchmark: Additional Details

D.1 Oracle Retrieval Ceiling

Set-level recall is structurally capped by the size of the merged retrieved set relative to |GOLD|, which ranges from 15 to 24,340 (median 42). Table 10 reports an oracle handed the gold constraints directly: precision is 1.0 by construction, so the ceiling is entirely recall-bounded. Each question issues the

original query plus five expansions, so a run at K retrieves up to $6K$ papers before deduplication; at $K=5$ this leaves a median of 22 unique papers after merging. Overall oracle recall@15 is 0.616 — 1.000 where the gold set holds at most 15 papers, but 0.006 where it exceeds 1,000, and evaluated systems recover roughly half of the attainable ceiling. Categories therefore measure different things: *overview_uni* and *uni_country* questions (small gold sets) test coverage, while *field_evolution* questions (gold sets in the thousands) test ranking quality alone. This is why field-evolution scores should not be read as coverage failures, and why we report the size slices above rather than a single aggregate.

GOLD slice	n	R@5	R@15	R@25	R@50
All	99	0.334	0.616	0.698	0.750
≤ 15	43	0.629	1.000	1.000	1.000
16–100	30	0.193	0.578	0.834	0.965
101–1000	10	0.020	0.059	0.097	0.195
> 1000	16	0.002	0.006	0.010	0.019

Table 10: Oracle recall ceiling by gold-set size, given the gold constraints directly. Precision is 1.0 by construction; the ceiling is bounded by the size of the merged retrieved set, up to $6K$ papers per question, relative to |GOLD|.

D.2 Judge Robustness

Faithfulness and relevancy were originally scored by a single judge (Qwen 3.5-397B). We re-scored the matched Gemini 2.5 Pro pipeline/agent pair with a second judge (GPT-4o-mini) on the 95 questions both runs share. The agent’s faithfulness advantage holds under both judges and the paired difference excludes zero, but its magnitude is judge-dependent: +0.32 under Qwen (0.717 vs 0.394) and +0.134 [+0.072, +0.198] under GPT-4o-mini (0.871 vs 0.736). The judges also disagree on absolute level — pooled across both runs, GPT-4o-mini averages 0.805 faithfulness against Qwen’s 0.519 — and their pass/fail agreement at a 0.5 threshold is

poor (κ 0.093). Answer relevancy is stable across judges (ρ 0.655, mean $|\Delta|$ 0.041, paired Δ -0.033 $[-0.056, -0.017]$). We therefore state the pipeline-vs-agent faithfulness claim as directionally stable with judge-dependent magnitude, and treat absolute faithfulness values as judge-relative throughout.