Performance Persists, Ethics Emerges: A Diachronic Analysis of Values in 15 Years of NLP Research

Anonymous ACL submission

Abstract

Prior work has scrutinised the values embedded in large language models; far less has measured the values of the NLP community itself, yet these values shape the models developed and, by extension, their societal impact. We present the first large-scale longitudinal analysis of values in NLP research, examining 68,885 ACL Anthology papers with abstracts spanning 2010–2025. Using a lexicon-based detection pipeline with negation filtering, evaluated against four human annotators (F1 = 0.708), we map value mentions to a taxonomy of 19 values across four categories (Technical, Epistemic, Ethical, Governance) synthesised from Birhane et al. (2022), Shneiderman (2022), Zhang and Dafoe (2019), and Hagendorff (2020). Our analysis reveals that (1) the average number of values per paper has more than doubled (1.6 to 3.5), amounting to a roughly 36-fold increase in total value mentions; (2) Performance persists as the dominant value in every year (43.5% to 77.4%), indicating that ethical growth is *additive*, not substitutive; (3) trustworthiness-related values show the fastest growth (Safety: 26% CAGR, Responsibility: 25%, Security: 24%) with a clear inflection point at 2017–2018; and (4) the field has become substantially more value-aware without abandoning its technical core. Our findings offer an empirical baseline for ongoing debates about NLP's research mission.

1 Introduction

Values shape how we make decisions, set priorities, and act; they are also unconsciously embodied in the technologies we build (Friedman and Hendry, 2019). This entanglement of humans, values, and technology renders the thesis of technological neutrality obsolete (Winner, 1980; Friedman and Hendry, 2019). In an increasingly networked, text-based world, the NLP community exerts a profound indirect influence on social, political, and cultural interactions (Bender et al., 2021), and the decisions made in developing NLP systems can have far-reaching societal consequences.

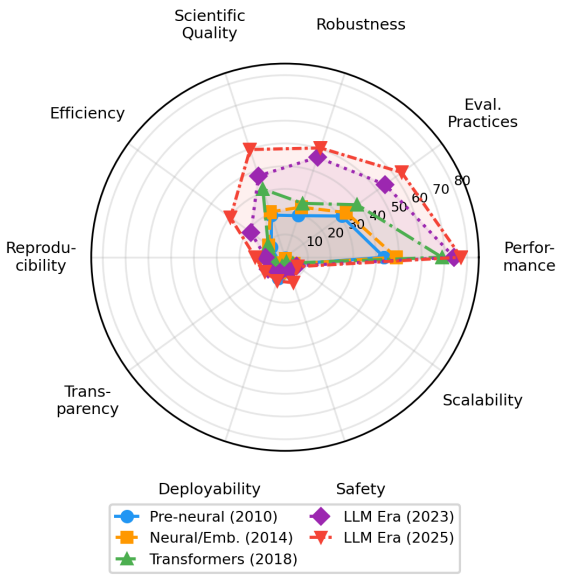

Figure 1: Value profiles across five NLP paradigm years (2010–2025), showing the top 10 values. Each polygon represents the normalised prevalence of values in that year's papers. The concentric expansion reflects the broadening of value discourse.

NLP has undergone successive paradigm shifts over the past fifteen years, from feature engineering through neural embeddings (Mikolov et al., 2013) and the Transformer revolution (Vaswani et al., 2017; Devlin et al., 2019) to today's large language models (Brown et al., 2020). Alongside these technical advances, a less visible transformation has occurred: a broadening of the *values* the community discusses in its publications.

While attention to values embedded in large language models has grown (e.g., Santurkar et al., 2023), far less work has examined the values of the NLP community itself (e.g., Birhane et al., 2022). Prior work identified a terminological blur in moral concepts in NLP research (Vida et al., 2023) and culturally inconsistent moral biases in

LLMs (Ramezani and Xu, 2023; Vida et al., 2024). **We therefore ask: how have the values of the NLP community evolved over time, how do they differ across subfields, and what does this imply for the trajectory of NLP research?**

We address this empirically: we construct a value taxonomy synthesised from research, governance, and public discourse frameworks, and apply an automated extraction pipeline to 68,885 ACL Anthology abstracts (2010–2025). This lets us trace which values dominate and how the community's notion of research progress has evolved across paradigm phases. Specifically, we ask: (1) How has the value profile of NLP research evolved diachronically across paradigm phases (2010–2025)? (2) Which values are growing, stable, or declining, and at what rates? (3) How do value profiles differ across venues (conferences, tracks, workshops)? (4) What structural patterns emerge from value co-occurrence?

Contributions. (1) We present the largest-scale longitudinal value analysis of NLP research to date, covering nearly 69,000 papers over 16 years. (2) We construct a taxonomy of 19 values with an operationalised detection pipeline evaluated against four human annotators. (3) We provide convergent evidence for a 2017–2018 "ethics inflection point." (4) We show that value growth in NLP is *additive*: the community has expanded its value vocabulary while strengthening, not replacing, its performance orientation. Together, these findings provide an empirical baseline for current debates about research priorities, review criteria, and venue policies (Figure 1).

2 Related Work

Values and Norms in AI/NLP. The responsible AI literature has produced numerous ethical frameworks and surveys of guidelines. Jobin et al. (2019) identified 84 AI ethics guidelines worldwide, finding convergence on transparency, fairness, and non-maleficence but divergence on operationalisation. Hagendorff (2020) noted gaps between stated principles and enforceable norms. Bommasani et al. (2022) catalogued the capabilities and risks of foundation models, offering a referent for the governance and safety values whose rise we document. These works prescribe or evaluate norms; they do not measure how the community's *actual discourse* has evolved.

Empirical Studies of Values in AI Research. Birhane et al. (2022) manually annotated 100 highly cited machine learning papers, finding that performance, generalisation, and efficiency dominate while societal considerations remain marginalised. Jiang et al. (2025) scaled this approach to 226,600 computer science abstracts, identifying ten core research values and noting an increasing articulation of values in AI-related subfields since 2013. The *EthiCon* corpus (Karamolegkou et al., 2025) complements these efforts by providing 1,580 ethical concern statements from the ACL Anthology ethics sections. Most closely related to our work, Vida et al. (2023) found significant conceptual blur between values, ethics, and morals in NLP research, motivating the careful operationalisation of our taxonomy. None of these studies provides a longitudinal, cross-value analysis spanning both technical and ethical dimensions across the full corpus.

Value Concepts and Value Detection. Social-science frameworks such as Schwartz's (2012) theory of basic values, Moral Foundations Theory (Haidt and Joseph, 2004), and Value Sensitive Design (Friedman and Hendry, 2019) focus on universal moral or social values and do not capture the technical and epistemic values (e.g., Performance, Reproducibility, Scalability) central to scientific research communities. Computational work has treated values as objects of *detection*: from argumentation (Kiesel et al., 2022) to moral norms encoded in LLMs (Schramowski et al., 2022). We take a complementary perspective: we study values not as detection targets in arbitrary text, but as an analytical lens on a specific scientific community's discourse.

Geographic Disparities in NLP. Rungta et al. (2022) analysed citation patterns across 78,000 ACL Anthology papers, documenting systematic underrepresentation of researchers from the Global South and East Asia in citation networks. Our work is complementary: where they study *whose work gets cited*, we study *what values researchers foreground*. We find that geographic disparities extend into the value layer of scientific discourse, with East Asian papers showing higher emphasis on Scientific Quality and Efficiency, and Global South papers foregrounding Safety, patterns not reducible to citation bias alone.

NLP Meta-Science and Bibliometrics. Individual value dimensions (reproducibility (Belz et al., 2020; Dodge et al., 2019), evaluation practices (Post, 2018; Marie et al., 2021), bias conceptualisation (Blodgett et al., 2020)) and bibliometric trends (Mohammad, 2020) have been studied in isolation. We extend this tradition with the first unified diachronic view across all value categories.

3 Methodology

We examine 68,885 ACL Anthology papers (2010–2025) with available abstracts using an automated extraction pipeline mapped against a value taxonomy.

3.1 Value Canon Construction

Existing frameworks from psychology (e.g., Schwartz, 2012) and computational argumentation (e.g., Kiesel et al., 2022) focus on universal moral values, which are ill-suited for a technical research community where Performance, Efficiency, Reproducibility, and Scalability are central. AI ethics guidelines (Jobin et al., 2019; Hagendorff, 2020) cover governance concerns but omit these technical and epistemic values. We therefore constructed a domain-specific taxonomy through a multi-perspective synthesis.

Starting from 67 values in Birhane et al. (2022), we integrated three further sources: Shneiderman (2022) (15 values, user/developer perspective), Zhang and Dafoe (2019) (20 values, public discourse), and Hagendorff (2020) (28 governance values). Two authors merged the resulting 130 values into 98 unique values, then clustered them into 19 categories operationalised through 255 search terms (Appendix A), organised into four groups.

Technical/Systemic. Performance, Robustness/Generalisation, Reproducibility, Efficiency, Scalability.

Research/Epistemic. Evaluation Practices, Scientific Quality, Transparency/Accountability, Negative Results.

Ethical/Social. Fairness/Bias, Privacy, Responsibility, Safety/Harm Reduction, Human-Centric, Usability & Interpretability.

Governance/Regulatory. Compliance, Security, Deployability, Maintainability.

We deliberately do not provide normative definitions for each value. Our unit of analysis is value *mentions* as operationalised by authors themselves, which intentionally allows for heterogeneous interpretations across papers. Search terms for all 19 values are listed in Appendix A (Table 6).

Phase	Period	Papers	% Det.
Pre-neural	2010–2013	1,878	77.1
Neural/Emb.	2014–2017	5,445	83.0
Transformer	2018–2022	28,716	92.0
LLM Era	2023–2025	32,846	95.6
Total	2010–2025	**68,885**	**92.6**

Table 1: NLP paradigm phases with corpus sizes and detection rates (% of papers with at least one detected value).

3.2 Data Source and Sampling

We target the ACL Anthology[1] as the primary repository for computational linguistics research. Using the `acl-anthology` Python library,[2] we extracted metadata, titles, and abstracts for all papers published between 2010 and December 2025 across all *ACL venues rated B or higher in the CORE rankings, covering the major conferences (ACL, EMNLP, NAACL, EACL, COLING, LREC), the Findings series, and all associated workshops. This yielded a total of 94,606 papers, of which 68,885 (72.8%) have abstracts available for analysis. The remaining 27.2% lack abstracts in the Anthology's metadata and were excluded.

Three years (2011, 2013, 2015) have 93–97% missing abstracts due to an ACL Anthology metadata artefact; we flag them in all visualisations.

Table 1 summarises the four paradigm phases. Boundaries follow landmark shifts: word embeddings (Mikolov et al., 2013) (2014), the Transformer and BERT (Vaswani et al., 2017; Devlin et al., 2019) (2018), and instruction-tuned LLMs (2023).

3.3 Extraction Pipeline

We implemented a two-stage pipeline. **Stage 1** uses NLTK Porter Stemmer keyword matching (Bird et al., 2009; Porter, 1980) with flexible phrase recognition. **Stage 2** discards matches preceded by negation patterns within a 50-character window (e.g., *"we do not address fairness"*). An ablation over seven pipeline variants showed no

[1] https://aclanthology.org/
[2] https://github.com/acl-org/acl-anthology

Metric	Value
Inter-annotator agreement (4 ann.)	
Fleiss' κ (overall)	0.559 (moderate)
Cohen's κ A↔B	0.640 (substantial)
Cohen's κ B↔C	0.491 (moderate)
Cohen's κ C↔D	0.876 (alm. perfect)
Pipeline vs. gold (majority ≥3/4)	
Micro Precision	0.713
Micro Recall	0.702
Micro F1	0.708

Table 2: Evaluation results (n=100). Fleiss' $\kappa = 0.559$ reflects the inherent subjectivity of value attribution (McHugh, 2012). The near-perfect C↔D agreement reflects shared expertise in taxonomy. Pipeline F1 = 0.708 exceeds the average of independent annotator pairs (A↔B: 0.640, A↔C: 0.446).

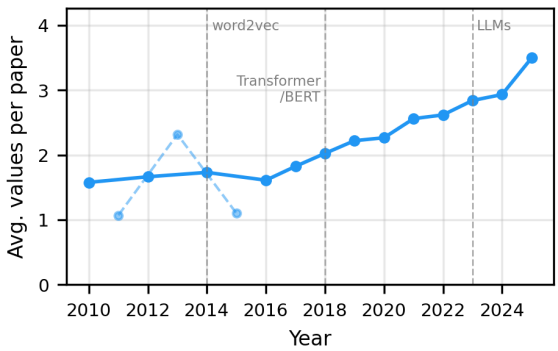

Figure 2: Average number of values per paper (2010–2025). Dips in 2011, 2013, and 2015 reflect gaps in ACL Anthology abstract coverage, not genuine declines. Dashed vertical lines indicate NLP paradigm transitions.

meaningful improvement beyond these two stages (Appendix B), so we favour this simpler design.

3.4 Evaluation

We constructed a stratified test set of 100 papers, sampled proportionally by year with a fixed random seed. The sample includes 89 papers with pipeline-detected values and 11 papers with no pipeline-detected values, enabling assessment of false-negative rates.

Four annotators independently labelled all 100 papers without access to the pipeline-detected values. Annotators C and D are the authors of this paper and co-developers of the value canon; annotators A and B are independent researchers with no prior exposure to the taxonomy. Table 2 reports inter-annotator agreement and pipeline evaluation results.

Overall inter-annotator agreement is moderate (Fleiss' $\kappa = 0.559$; Landis and Koch, 1977; McHugh, 2012), consistent with Birhane et al. (2022) who report $\kappa = 0.45$ for a comparable task. The independent pair A↔B achieves $\kappa = 0.640$; the co-developer pair C↔D achieves $\kappa = 0.876$. Against the majority-vote gold standard (≥3/4 annotators), the pipeline achieves micro F1 = 0.708, exceeding the average of independent annotator pairs. All reported frequencies are conservative lower bounds.

3.5 Author Affiliation Extraction

For geographic analysis (§4.5), we mapped each paper to a world region via a four-stage pipeline: the Rungta et al. (2022) dataset, ACL Anthology XML, the Semantic Scholar API,[3] and PDF parsing. Combined, 95.7% of papers received a country signal; 82% have a known region in the 2016–2025 window. Region mapping follows the WEIRD framework (Henrich et al., 2010), with East Asia, Global South, and *Industry* as separate categories; multi-region papers are attributed to the less-represented region (priority: Global South > East Asia > Industry > WEIRD). Full details in Appendix C.

3.6 Analysis Framework

All comparisons use normalised percentages (base: papers with available abstracts per year) to account for the 16-fold publication volume growth. Trends are quantified as Compound Annual Growth Rate (CAGR) over 2010–2025. Co-occurrence analysis identifies structural value relationships; permutation tests (10,000 iterations) assess significance of venue-profile differences.

4 Results

4.1 Value Expansion: The Overall Picture

The average number of values per paper more than doubled from 1.6 in 2010 to 3.5 in 2025 (Figure 2). Combined with the 16-fold publication volume growth, this amounts to a roughly 36-fold increase in total value mentions. The modal paper in 2025 discusses three distinct values (24.2% of papers), with the distribution right-skewed toward higher counts.

A single value does not drive this growth. Performance remains the most discussed value every year, growing from 43.5% to 77.4%, but its rise

[3] https://api.semanticscholar.org/

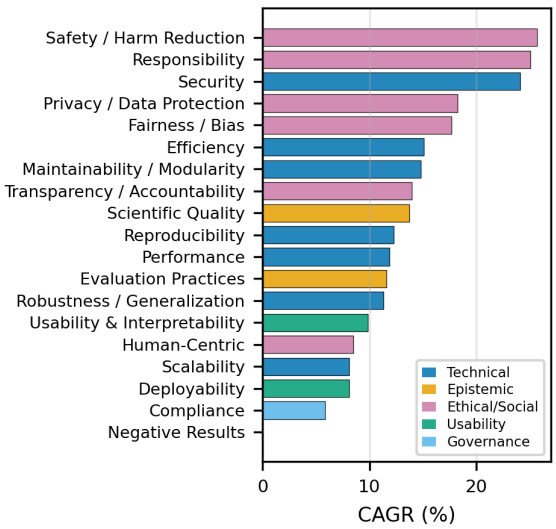

Figure 3: Compound Annual Growth Rate (CAGR) for all 19 values, 2010–2025. Trustworthiness-related values dominate the top positions. Negative Results remains at near-zero prevalence throughout.

has been accompanied by previously marginal values. Detection rates increase monotonically across paradigm phases (77.1% pre-neural to 95.6% LLM era), reflecting a richer and more explicit value vocabulary.

CAGR analysis (Figure 3) shows trustworthiness-related values dominate: Safety (25.7%), Responsibility (25.1%), Security (24.1%), Privacy (18.3%). Efficiency leads technical values at 15.1% CAGR, 2–3× above Performance (11.9%). Figure 4 reveals a clear "wave" of ethical and governance values, nearly absent before 2017 and intensifying through 2025.

4.2 The Ethics Inflection Point

Multiple independent indicators converge on 2017–2018 as the critical inflection point for ethical values in NLP. To contextualise the diachronic trends: in 2025, Performance (77.4%), Evaluation Practices (63.4%), and Robustness (50.5%) form the dominant triad; ethical values remain a minority (Safety: 11.8%, Fairness: 3.5%, Responsibility: 0.9%). Yet these numbers obscure dramatic trajectories: Safety grew from 0.4% to 11.8% (with the sharpest acceleration beginning in 2019), Fairness from 0.9% to 3.5%, and Responsibility roughly eightfold from its 2010 baseline.

Figure 5 shows the Ethical/Social band thin before 2018 (4–6%), then growing to 24% by 2025, while Technical/Systemic maintains dominance throughout. This timing aligns with external catalysts: the Ethics in NLP workshop (EACL 2017), FAccT launch (2018), high-profile bias incidents (e.g., Bolukbasi et al., 2016), the Stochastic Parrots debate (Bender et al., 2021), EU AI Act discussions (European Parliament and Council of the European Union, 2024), and ACL's ethics review requirements.

4.3 Technical Value Dynamics

Efficiency is the most striking shift within technical values, growing from 7.4% to 29.6% with its steepest rise in 2023–2025, reflecting LLM-era discourse on computational costs (Schwartz et al., 2019; Strubell et al., 2019). Reproducibility follows a U-shaped trajectory: elevated in 2010–2011, plateauing through 2022, then rising in 2023–2025, likely driven by debates on open model weights and data documentation. Scalability grows modestly (CAGR 8.1%), treated more as a given than a contested value.

Negative Results is the only value that has not grown, remaining below 0.1% throughout, with near-zero co-occurrence with all other values (Figure 13). Failure reporting is not merely rare but structurally isolated, a paradox we return to in §5.

4.4 Venue Profiles

Conference Homogeneity. ACL, EMNLP, NAACL, COLING, and EACL show near-identical value fingerprints (Figure 6), suggesting a single disciplinary community with shared norms. LREC is the exception: lower Performance emphasis (∼66% vs. ∼80%) with a distinctive focus on Reproducibility and Evaluation Practices, reflecting its language resources orientation and complementing rather than mirroring the flagship conferences.

Main vs. Findings. The Findings track (introduced 2020) is nearly indistinguishable from main conference papers (Figure 14): 14 of 19 values show no significant difference; the five exceptions (Robustness, Performance, Deployability, Scientific Quality, Efficiency) differ by at most 2 percentage points, with Findings papers marginally higher. This pattern suggests thoroughness compensating for lower novelty ratings rather than a qualitatively different type of research.

Workshop Diversity. Workshops show substantial value differentiation: BioNLP/ClinicalNLP emphasise Safety, BlackboxNLP centres on Trans-

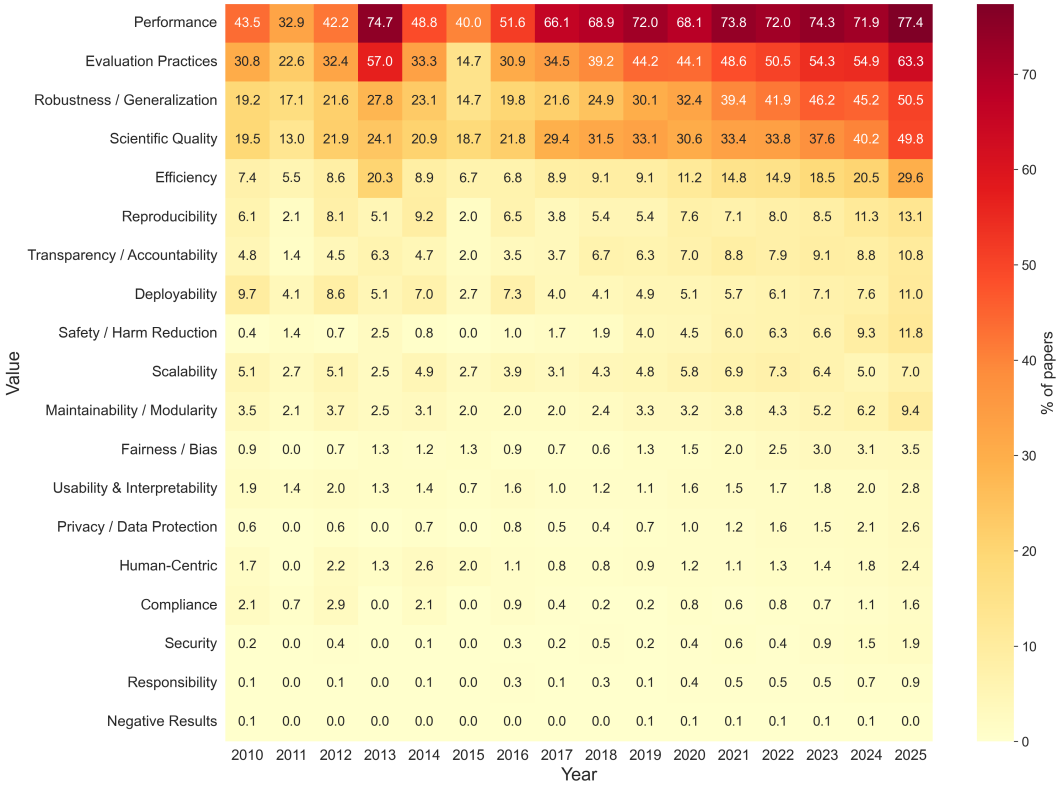

Figure 4: Heatmap of the top 20 values by year (normalised as % of papers). Performance and Evaluation Practices form the consistently "hot" core, while ethical values (Safety, Fairness, Responsibility) emerge visibly from 2018 onward.

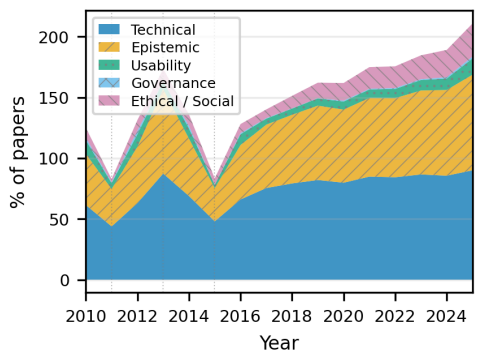

Figure 5: Evolution of value categories (2010–2025). The Ethical/Social band emerges from near-zero after 2017, while Technical values remain dominant. The total exceeds 100% because the papers discuss multiple categories.

parency, WASSA shows elevated Fairness. Performance drops to 30–45% at ethically oriented workshops, the lowest among all venue types (see Figures 18 and 15), confirming workshops as incubators for values underrepresented in flagship conferences.

4.5 Geographic Distribution of Values

We mapped each paper to one of four regions: WEIRD, East Asia, Global South, and Industry.[4] 82% of papers in the 2016–2025 window received a region assignment. All regions share the same dominant triad (Performance, Evaluation Practices, Robustness), with moderate regional differences (max Cramér's $V = 0.161$; Figure 7).

Three patterns emerge from 52 significant pairwise comparisons ($p < .05$, chi-squared with Bonferroni correction):

Technical emphasis in East Asia and Industry. East Asian papers show higher Performance (79% vs. 67% WEIRD, $V = 0.13$), Efficiency (24% vs. 14%, $V = 0.14$), and Evaluation Practices (58% vs. 47%, $V = 0.11$). Industry papers amplify this further: highest Performance (80%), Efficiency (28%, $V = 0.15$), and Robustness (51%, $V = 0.11$).

Scientific Quality gap. The largest difference is Scientific Quality: East Asian papers at 49% vs.

[4]Priority order: Global South > East Asia > Industry > WEIRD; *Industry* denotes papers with at least one author at a major technology company.

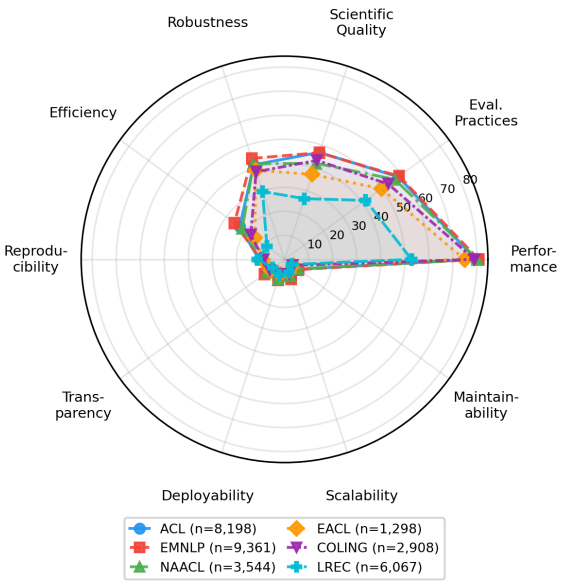

Figure 6: Value fingerprints for six major ACL Anthology venues. ACL, EMNLP, and NAACL overlap substantially, while LREC shows a distinctive profile with less emphasis on Performance.

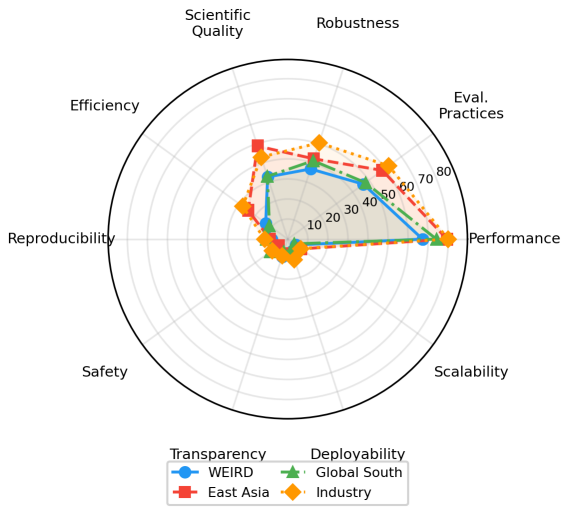

Figure 7: Value profiles by world region (top-10 values by mean prevalence, 2016–2025, $n = 54{,}317$ papers with known region). All regions share the same dominant triad; differences are most pronounced for Scientific Quality and Efficiency.

33% WEIRD ($V = 0.161$). This gap also holds against Global South papers (49% vs. 33%, $V = 0.13$). It may reflect higher implicit standards for empirical rigour in competitive venues, or different surface forms our lexicon captures more readily in East Asian writing styles.

Safety and Reproducibility in the Global South. Global South papers show elevated Safety (11% vs. 6% East Asia, $V = 0.08$) and Reproducibility (11% vs. 8% WEIRD, $V = 0.03$), alongside the lowest Efficiency rate (11%), consistent with application-oriented research where deployment safety is pressing.

These findings complement Rungta et al. (2022), who document geographic disparities in citation impact. Geographic disparities extend beyond whose work gets cited: they reach into the *values* researchers foreground. Complete profiles for all 19 values are in Table 8 and Appendix C.

5 Discussion

Our findings reveal **a research community in transition**: technically oriented but increasingly value-aware, expanding its normative vocabulary without abandoning its performance core. We interpret these patterns with respect to the additive nature of value growth, the structural isolation of negative results, and implications for NLP's research.

5.1 The Additive Growth Pattern

The 2017–2018 inflection point we document coincides with external catalysts (FAccT, bias incidents (e.g., Bolukbasi et al., 2016), the EU AI Act (European Parliament and Council of the European Union, 2024)) and institutional mechanisms (ARR responsible NLP checklist, ethics review requirements, broader impact statements). The most important finding is *how* value discourse expanded: the community added ethical values to its vocabulary without displacing Performance. Performance grew from 43.5% to 77.4%, a stronger absolute increase than any ethical value. The difference is that Performance started high and grew steadily, while ethical values started near zero and grew exponentially.

Whether this breadth translates into depth remains an open question. Our pipeline detects value *mentions*, not substantive engagement. A paper noting "we acknowledge potential fairness implications" registers a Fairness detection, but may reflect compliance with institutional norms rather than genuine concern. The rapid growth of ethical mentions after mandatory ethics statements and broader impact sections raises the possibility that some growth is norm-driven rather than conviction-driven.

5.2 The Negative Results Blind Spot

The persistent near-absence of negative result reporting amid otherwise expanding value awareness is paradoxical. As the community becomes more reflexive about fairness, safety, and transparency, it has not increased its engagement with one of the most fundamental principles of scientific self-correction: reporting what does not work.

Publication incentives provide a likely explanation (Fanelli, 2012): positive results are more likely to be published, and the growth in NLP's publication volume has not been accompanied by proportional growth in venues for negative results. The dedicated Insights from Negative Results in NLP workshop (hosted at EMNLP since 2020) is a notable exception, yet its impact on the broader community remains limited. Negative Results also shows near-zero association with every other value in our co-occurrence network, which suggests that failure reporting is not merely rare but structurally disconnected from the broader value discourse. This structural isolation is arguably more concerning than the low absolute numbers: it implies that the community does not conceptually link failure reporting to other forms of scientific rigour.

5.3 Implications

Our findings carry several practical implications. The near-identical value profiles of ACL, EMNLP, and NAACL suggest that venue differentiation would require deliberate programmatic choices. Workshop selection emerges as a particularly consequential lever: since workshops are the primary incubators for underrepresented values, the decisions of workshop chairs and programme committees about which workshops to accept implicitly shape the community's value pluralism. Making these selection criteria more transparent could strengthen the community's normative diversity.

What NLP research does not discuss. Comparing our empirical taxonomy against AI ethics guidelines (Jobin et al., 2019) reveals a systematic gap: values that feature prominently in policy and governance discourse are largely absent from NLP abstracts. Jobin et al. (2019) identify *autonomy*, *beneficence*, *non-maleficence*, and *justice* as the most frequently invoked principles across 84 global AI ethics guidelines, yet none of these map onto high-prevalence NLP values. While Safety and Fairness represent partial coverage, the broader notion of justice (distributional, procedural, or structural) does not appear as an articulable research value in our corpus. Human autonomy, a central concern in human-computer interaction and AI governance, remains largely invisible outside narrow contexts such as user-centric design (<2% of papers). This gap is not merely a labelling difference: it reflects a structural disconnect between the values that NLP researchers operationalise in their work and the values that society's regulatory frameworks demand from AI systems. Bridging it requires more than lexicon expansion; it requires changes in how NLP frames its research questions.

The conservative nature of our pipeline means that any value trend we detect is a lower bound: the community's actual engagement with ethical values is likely stronger. The 2017–2018 inflection point demonstrates that a combination of external pressure and institutional mechanisms can measurably shift a research community's value discourse within a few years.

6 Conclusion

We have presented the first large-scale diachronic analysis of values in NLP research, examining how 68,885 papers across 16 years articulate 19 distinct values. Value density per paper has more than doubled (1.6 to 3.5) against a 16-fold publication growth, a roughly 36-fold increase in total value mentions. This growth is additive: Performance remains dominant (43.5% to 77.4%) while epistemic, ethical, and governance dimensions have expanded alongside it. The 2017–2018 inflection point marks the most significant structural shift, driven by a convergence of external events and institutional mechanisms.

Geographic and venue analyses reveal further structure: all regions share the same dominant triad but diverge in secondary emphases, with East Asia and Industry foregrounding efficiency and Global South papers showing elevated Safety. Workshops emerge as the primary incubators for underrepresented values, making workshop acceptance an implicit lever for the community's normative diversity.

Negative result reporting remains a community-wide blind spot, structurally isolated from all other value discourse. A deeper gap lies between NLP's operationalised values and AI governance principles: autonomy, beneficence, and justice remain largely absent from research abstracts. Closing this distance is the work ahead.

7 Limitations

Evaluation coverage. Our test set of 100 stratified papers provides reliable evaluation for frequent values (Performance: F1 = 0.86) but is too small to assess pipeline performance on rare values. Seven of 19 values (including Fairness, Security, Scalability, and Responsibility) have zero instances in the majority-vote gold standard ($\geq$3/4 annotators), despite being present in the full corpus at rates of 1–7%. However, these seven values collectively account for less than 15% of all value detections, meaning that the pipeline's overall F1 of 0.708 is primarily driven by the 12 values for which we do have gold-standard instances. A larger, value-stratified evaluation set would strengthen per-value reliability estimates but is unlikely to change the aggregate picture substantially. Lexicon-based detection further introduces false positives through polysemy and false negatives for implicitly expressed values; however, since these errors are not systematically linked to year or venue, they are unlikely to produce the directional trends we report.

Per-value annotation difficulty. While overall inter-annotator agreement is moderate (Fleiss' κ = 0.559; Landis and Koch, 1977), it varies considerably across values. Subjective values like Responsibility ($\kappa = -0.010$), Usability & Interpretability ($\kappa = -0.010$), and Maintainability ($\kappa = -0.008$) show near-chance agreement. These three values account for only a small fraction of total detections. The values that drive our main findings (Performance, Efficiency, Safety, Robustness, and Evaluation Practices) all achieve moderate to substantial agreement. We note that value annotation is inherently difficult: Birhane et al. (2022) report Fleiss' κ = 0.45 for a comparable task (annotating values in ML papers with 75 categories), confirming that moderate agreement is typical for value attribution in scientific texts.

Abstract-only analysis. We analyse abstracts rather than full texts. This is a deliberate methodological choice: abstracts are the author-curated, standardised summary of a paper's contribution, and thus represent what researchers choose to foreground as central. Boilerplate methodology sections and acknowledgement text introduce noise in full-text analyses that abstracts avoid. This approach is consistent with Birhane et al. (2022) and Jiang et al. (2025), enabling direct comparability. The main consequence is that values expressed only in ethics statements or conclusions (most notably Safety and Responsibility) may be underestimated in absolute terms, though not differentially across years or venues. Three years (2011, 2013, 2015) additionally have 93–97% missing abstracts due to ACL Anthology metadata artefacts; we flag these data points consistently throughout.

English-only scope. The ACL Anthology is predominantly English-language, and our analysis targets this community as its subject rather than as a sample of global NLP. Non-English papers ($\sim$15% of no-value papers, primarily French and Chinese) are outside scope. One consequence for §4.5 is that Chinese-language CCL papers, which are not included in the Anthology, are absent from the East Asia region, meaning East Asian value profiles reflect primarily internationally published, English-language research from that region rather than the full domestic research landscape.

Geographic affiliation coverage. Our affiliation pipeline achieves 95.7% coverage overall, but coverage is uneven by time period: 75–87% for 2016–2021 (primarily from Rungta et al., 2022) and 80–87% for 2022–2025 (from ACL XML, Semantic Scholar, and PDF parsing). The remaining 4.3% consists largely of pre-2010 proceedings and papers with non-Latin-script or image-only affiliations. Additionally, our *Industry* category conflates multinational companies (Google, Meta, Microsoft) that operate across regions; their geographic distribution does not map cleanly onto the WEIRD/East Asia/Global South framework. We report them separately and advise against interpreting Industry as a geographically homogeneous group.

References

Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2020. ReproGen: Proposal for a shared task on reproducibility of human evaluations in NLG. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 232–236, Dublin, Ireland. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623. ACM.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Nat-*

ural Language Processing with Python. O'Reilly Media.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 173–184. ACM.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the opportunities and risks of foundation models.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. 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 2185–2194, Hong Kong, China. Association for Computational Linguistics.

European Parliament and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI act). Official Journal of the European Union. OJ L 2024/1689.

Daniele Fanelli. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3):891–904.

Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.

Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. The MIT Press.

Thilo Hagendorff. 2020. The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 30(1):99–120.

Jonathan Haidt and Craig Joseph. 2004. Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus*, 133(4):55–66.

Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3):61–83.

Hang Jiang, Tal August, Luca Soldaini, Kyle Lo, and Maria Antoniak. 2025. Automatic detection of research values from scientific abstracts across computer science subfields.

Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

Antonia Karamolegkou, Sandrine Schiller Hansen, Ariadni Christopoulou, Filippos Stamatiou, Anne Lauscher, and Anders Søgaard. 2025. Ethical concern identification in NLP: A corpus of ACL Anthology ethics statements. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11618–11635, Albuquerque, New Mexico. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Martin Potthast, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of ACL 2022*, pages 4459–4471.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Mary L. McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica*, 22(3):276–282.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26.

Saif M. Mohammad. 2020. NLP scholar: An interactive visual explorer for natural language processing literature. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 232–255, Online. Association for Computational Linguistics.

Martin F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Aida Ramezani and Yang Xu. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 428–446, Toronto, Canada. 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.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect?

Patrick Schramowski, Cigdem Turan, Nico Andersen, Constantin A. Rothkopf, and Kristian Kersting. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4:258–268.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green ai.

Shalom H. Schwartz. 2012. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1).

Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press, London, England.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of ACL 2019*, pages 3645–3650.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Karina Vida, Florian Damken, and Anne Lauscher. 2024. Decoding multilingual moral preferences: Unveiling LLM's biases through the moral machine experiment. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7.

Karina Vida, Judith Simon, and Anne Lauscher. 2023. Values, ethics, morals? on the use of moral concepts in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5534–5554, Singapore. Association for Computational Linguistics.

Langdon Winner. 1980. Do artifacts have politics? *Daedalus*, 109(1):121–136.

Baobao Zhang and Allan Dafoe. 2019. Artificial intelligence: American attitudes and trends. *SSRN Electronic Journal*.

A Value Canon: Construction and Search Terms

Construction Process

Figure 8 summarises the multi-stage construction of the value canon. We established an initial catalogue of 67 values from Birhane et al. (2022), who investigated values in the broader ML community. We then integrated three further sources: 15 values from Shneiderman (2022) capturing user and developer perspectives, 20 values from a public discourse survey on AI (Zhang and Dafoe, 2019), and 28 governance-related values from Hagendorff (2020). The resulting 130 values were manually filtered for duplicates and merged by two authors of this paper in a joint discussion, yielding 98 unique values. These were clustered into 19 categories across four groups, and operationalised through 255 search terms.

Inter-Annotator Agreement on Paper Annotation

Table 3 reports pairwise Cohen's κ for all six annotator pairs on the 100-paper test set. Annotators C and D are the paper authors and co-developers of the taxonomy; A and B are independent researchers with no prior exposure to the canon.

Pair	Cohen's κ	Interpretation
A ↔ B	0.640	Substantial
A ↔ C	0.446	Moderate
A ↔ D	0.414	Moderate
B ↔ C	0.491	Moderate
B ↔ D	0.468	Moderate
C ↔ D	0.876	Almost perfect
Fleiss' κ (all 4)		0.559 (Moderate)

Table 3: Pairwise inter-annotator agreement (Cohen's κ) for all six annotator pairs on the 100-paper test set. The near-perfect C↔D agreement reflects shared taxonomy expertise; A↔B represents fully independent agreement.

Source	Values	Perspective
Birhane et al. (2022)	67	ML research community
Hagendorff (2020)	44	Governance & regulatory
Zhang and Dafoe (2019)	27	Public discourse on AI
Shneiderman (2022)	18	User & developer
Raw union	130	
After dedup & merge	98	2 authors, joint discussion
Final taxonomy	**19**	4 groups, 255 terms

Table 4: Sources and consolidation steps for the value canon. Values were independently extracted by two annotators (82% agreement).[5] Deduplication performed by two paper authors in joint discussion.

Value	B	H	Z	S
Performance	✓		✓	
Robustness / Generalisation	✓			
Reproducibility	✓			
Efficiency	✓			✓
Scalability	✓			
Evaluation Practices	✓			
Scientific Quality	✓			
Transparency / Accountability		✓	✓	
Negative Results	✓			
Fairness / Bias	✓	✓	✓	
Privacy / Data Protection	✓	✓	✓	✓
Responsibility		✓		
Safety / Harm Reduction	✓	✓	✓	✓
Human-Centric		✓		✓
Usability & Interpretability		✓	✓	✓
Compliance		✓	✓	
Security	✓	✓	✓	✓
Deployability	✓			✓
Maintainability / Modularity	✓			✓

Table 5: Source mapping for the 19 final values. **B** = Birhane et al. (2022), **H** = Hagendorff (2020), **Z** = Zhang and Dafoe (2019), **S** = Shneiderman (2022).

Search Terms

Table 6 lists all 19 values with their categories and the complete set of search terms. Morphological variants (stemmed forms, plurals, hyphenation alternatives) are generated automatically by the Porter Stemmer and are not listed separately.

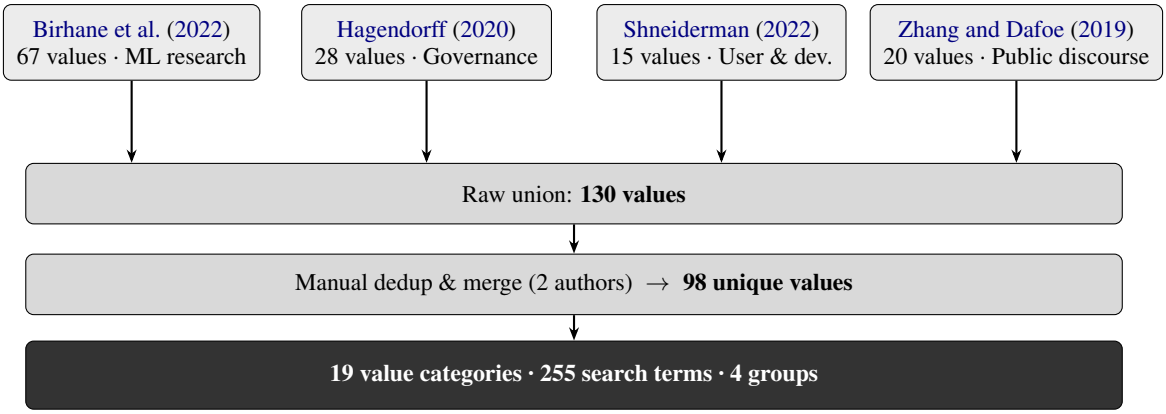

Figure 8: Multi-stage construction of the value canon. Four source perspectives were combined into a raw union of 130 values, manually deduplicated and merged into 98 unique values, and clustered into the final 19-category taxonomy.

Category	Value	Search Terms
Technical / Systemic	Performance	accuracy, precision, recall, F1, macro-F1, micro-F1, BLEU, ROUGE, METEOR, perplexity, EM score, hits@k, state-of-the-art, SOTA, improvement, outperform, superior, gain, better than, surpass
	Robustness / Generalisation	robustness, robust, adversarial, perturbation, noise, noisy, resilience, stable, stability, domain shift, out-of-distribution, OOD, generalisation, transfer, cross-lingual, multilingual, zero-shot, few-shot, fine-tuning, adaptation
	Reproducibility	reproducibility, reproducible, replication, replicate, replicable, code available, open source, dataset release, benchmark release, hyperparameter, seed, experiment logs, configuration, setup details
	Efficiency	efficiency, latency, throughput, runtime, inference time, memory footprint, parameter count, compression, pruning, distillation, quantisation, optimisation, speedup, compute cost, computational cost, energy, carbon footprint, green AI, sustainability
	Scalability	scalability, large-scale, scaling laws, distributed training, parallelism, batch size, sharding
Research / Epistemic	Evaluation Practices	evaluation, benchmark, baseline, ablation, error analysis, case study, limitation, limitations, failure case, failure mode, test set, validation set
	Scientific Quality	novelty, novel, innovative, innovation, original contribution, significant improvement, new approach, new methodology, paradigm, framework, architecture
	Transparency / Accountability	transparency, explainability, explainable, understandable, accountable, accountability, audit, auditing, traceability, provenance
	Negative Results	negative results, null results, unexpected findings, failure analysis
Ethical / Social	Fairness / Bias	fairness, fair, unbiased, bias, debias*, de-bias*, equitable, discrimination, disparate impact, demographic parity, group fairness, representativeness, inclusivity, inclusiveness, equality of opportunity
	Privacy / Data Protection	privacy, private, confidential, anonymity, anonymisation, de-identification, pseudonymisation, differential privacy, secure data, secure storage, GDPR, data protection, privacy protection
	Responsibility	social responsibility, responsibility, responsible
	Safety / Harm Reduction	safety, safe, harmful, harm, risk, toxicity, toxic, offensive, hate speech, misinformation, disinformation, robustness to misuse, safe deployment, risk mitigation, safeguard
	Human-Centric	human-centred, human-centered, human in the loop, participatory design, co-design, user feedback, stakeholder input, stakeholder engagement, collaborative
	Usability & Interpretability	usability, user experience, UX, human factors, ease of use, interpretability, user study, user evaluation, task performance, satisfaction, cognitive load
Governance / Regulatory	Compliance	compliance, regulation, regulated, policy, law, legal, legal framework, AI Act, high-risk, data governance, consent, licence, licensing, terms of use, copyright, intellectual property, IP rights, standardisation, standard, ISO
	Security	security, secure, secure system, cybersecurity, vulnerability, vulnerabilities, attack, adversarial attack, data security, network security, encryption, authentication, authorisation, access control, threat, defence, protection, secure communication, malware, intrusion detection, security breach, data breach, secure by design
	Deployability	deployability, deployment, real-world use, production, integration, system integration, plug-and-play
	Maintainability / Modularity	maintainability, maintainable, modular, modularity, reusability, reusable, updatable, extensibility, incremental learning, continual learning, lifelong learning

Table 6: Complete value taxonomy with 19 values across four categories and all search terms. Morphological variants are generated automatically by the Porter Stemmer and not listed separately.

B Pipeline Ablation Study

To validate our two-stage pipeline design (§3.3), we conducted an ablation study over seven pipeline variants on the 100-paper evaluation set. Table 7 reports the results.

Pipeline Variant	Micro F1
Stage 1 only (lexicon matching)	0.706
Stages 1–2 (+ negation filtering)	**0.708**
+ spaCy dependency parsing	0.709
+ co-keyword disambiguation	0.707
+ TF-IDF semantic fallback	0.710
+ spaCy + co-keyword + TF-IDF	0.710
+ SPECTER neural embeddings	0.708

Table 7: Ablation study results. No variant improves F1 by more than 0.002 over the two-stage baseline (bold). The dominant factor for pipeline quality is lexicon precision, not post-processing complexity.

The 100-paper evaluation set was stratified by year with a fixed random seed, and includes 89 papers with pipeline-detected values and 11 papers with no pipeline-detected values, enabling assessment of false-negative rates. Of the 11 papers with no detections, 7 were confirmed by majority-vote annotation to genuinely discuss no taxonomy value, while 4 were false negatives; this yields an estimated false-negative rate of 4% at the paper level.

The results confirm that post-processing additions provide negligible improvements. Negation filtering contributes a marginal +0.002 over raw lexicon matching, while more complex elements (dependency parsing, disambiguation, semantic fallback) add at most another +0.002. We therefore adopted the simplest effective design: lexicon matching with negation filtering (Stages 1–2).

C Geographic Distribution: Full Results

Affiliation Extraction Pipeline

The four-stage pipeline summarised in §3.5 is applied in priority order:

1. **Rungta et al. dataset** (Rungta et al., 2022): provides institution-level affiliation data for ACL Anthology papers up to 2021, covering 45,388 papers (48%) via ACL ID matching.
2. **ACL Anthology XML**: the Anthology's structured metadata includes `<affiliation>` tags for many post-2020 papers. We extracted country signals from these tags for an additional 24,793 papers (26%).
3. **Semantic Scholar API**: for remaining papers, we queried author affiliations via the Semantic Scholar Academic Graph API, adding 4,461 papers (5%).
4. **PDF parsing**: for all remaining papers without affiliation data, we downloaded PDFs from the ACL Anthology and extracted country signals from first-page text using three heuristics in order: (a) email domain TLDs (e.g. `.cn` → China), (b) institution name substring matching against the CSRankings database, and (c) country name and city regular expressions. We used `pymupdf` as the primary parser with `pdfminer` as fallback to handle font encoding issues in EMNLP and NAACL proceedings. This stage recovered 15,935 papers (17%).

Region Definitions

Country-to-region mapping follows the WEIRD framework (Henrich et al., 2010), with East Asia (China, Japan, South Korea, Taiwan, Hong Kong, Singapore) and Global South (South/Southeast Asia, Middle East, Africa, Latin America) as separate categories, and an *Industry* category for papers with at least one author at a major technology company. Papers with authors from multiple regions are assigned to the most underrepresented region (priority: Global South > East Asia > Industry > Other > WEIRD).

Quantitative Results

The analysis covers 54,317 papers from 2016–2025 for which at least one country signal could be extracted (82% of papers in that period).

Table 8 shows the complete value profiles for all four regions, sorted by mean prevalence across regions. Bold entries indicate values where a region exceeds WEIRD by more than 5 percentage points.

The Scientific Quality gap between East Asia and WEIRD ($V = 0.161$, the strongest effect in the entire analysis) admits two interpretations: East Asian researchers may face higher implicit standards for empirical rigour in internationally competitive venues, or the value may be expressed through different surface forms that our lexicon captures more readily in East Asian writing styles. These interpretations are not mutually exclusive.

Figure 9 provides an overview of the full value–region matrix. The colour intensity makes it immediately apparent which values drive regional differentiation (Scientific Quality, Efficiency, Robustness) and which are universally rare (Negative Results, Responsibility, Compliance).

Figure 10 lists the top-7 values per region. Performance and Evaluation Practices appear in all four top-7 lists, confirming a shared disciplinary core. Safety/Harm Reduction enters the Global South top-7 but not those of WEIRD, East Asia, or Industry, pointing to a regionally distinctive weighting of ethical concerns.

Figure 11 tracks the six most regionally discriminative values over time. The Efficiency gap between WEIRD and East Asia / Industry widens noticeably after 2022, coinciding with the LLM era's emphasis on inference cost. The convergence of Safety trajectories after 2018 suggests that the ethical turn is a community-wide rather than region-specific phenomenon.

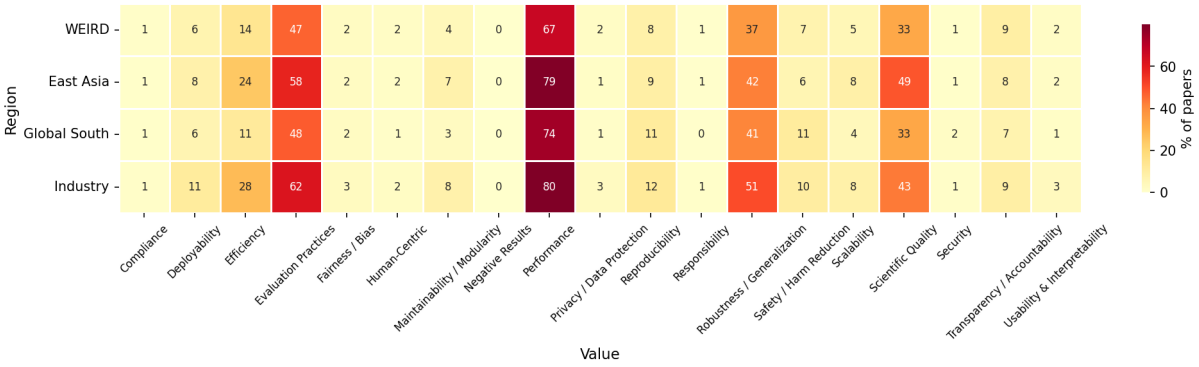

Figure 9: Heatmap of all 19 values by world region (% of papers, 2016–2025). Empty or near-zero cells indicate values that are rare across all regions.

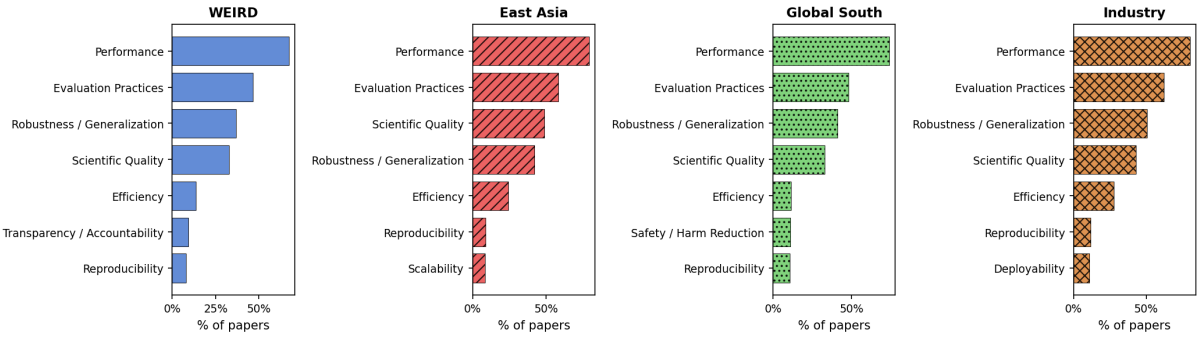

Figure 10: Top-7 values per region. All four regions share Performance and Evaluation Practices as dominant values; Safety/Harm Reduction appears in the Global South top-7 but not elsewhere.

Figure 11: Temporal trends for the six most regionally discriminative values (2016–2025). Only region–year cells with $n \geq 20$ papers are plotted. The Efficiency gap between WEIRD and East Asia/Industry widens noticeably after 2022.

Value	**WEIRD** *n=26,536*	**East Asia** *n=16,422*	**Global South** *n=4,206*	**Industry** *n=5,928*
Performance	67.2	**79.4**	**74.1**	**80.0**
Evaluation Practices	46.6	**58.3**	47.9	**62.2**
Robustness / Generalization	36.8	42.1	41.1	**50.5**
Scientific Quality	32.6	**48.8**	32.9	**43.0**
Efficiency	13.6	**24.4**	11.3	**27.8**
Transparency / Accountability	9.2	7.8	6.7	8.8
Reproducibility	8.0	8.9	**10.7**	**11.6**
Safety / Harm Reduction	6.7	5.7	**10.9**	9.7
Deployability	6.4	7.6	6.1	**10.8**
Scalability	5.0	8.3	4.0	8.0
Maintainability / Modularity	4.1	7.3	2.8	7.7
Fairness / Bias	2.5	2.4	2.4	2.9
Usability & Interpretability	1.9	1.6	1.5	3.2
Privacy / Data Protection	1.6	1.4	1.1	2.7
Human-Centric	1.6	1.5	1.0	1.9
Compliance	1.0	0.7	0.8	1.0
Security	0.7	0.9	1.7	1.1
Responsibility	0.6	0.5	0.4	0.8
Negative Results	0.1	0.0	0.0	0.0

Table 8: Complete value profiles by world region (% of papers mentioning each value, 2016–2025). Sorted by mean prevalence across regions. **Bold**: region exceeds WEIRD by >5 percentage points. Bottom rows (Compliance–Negative Results) are uniformly rare across all regions, indicating community-wide rather than region-specific blind spots.

D Additional Visualisations

This appendix collects supplementary figures that provide additional detail on the results discussed in the main paper. Figures are grouped thematically: corpus overview, value co-occurrence structure, publication track comparisons, workshop profiles, diachronic radar charts, and per-venue breakdowns.

Corpus overview. Figure 12 shows the raw paper counts underlying our analysis. The tenfold growth in annual publications between 2010 and 2025 motivates our use of normalised percentages throughout. The three low-coverage years (2011, 2013, 2015) reflect a metadata artefact in the ACL Anthology and are flagged in all trend plots.

Value co-occurrence structure. Figure 13 shows pairwise co-occurrence correlations across all 19 values. Two structural findings stand out: (1) Negative Results is almost entirely uncorrelated with all other values, confirming its isolation as a niche discourse rather than part of mainstream value bundles; (2) positive clusters emerge within the ethical group (Fairness, Safety, Responsibility) and within the technical group (Performance, Robustness, Evaluation Practices), suggesting that authors tend to invoke values within their cluster rather than across group boundaries.

Main vs. Findings track. Figures 14 and 15 address the question of whether the Findings series represents a qualitatively different tier of NLP research. The bar chart (Figure 14) shows that differences across all 19 values are below 5 percentage points, and the heatmap (Figure 15) confirms that the full track-level picture is dominated by Demo and Industry papers rather than any Main/Findings split. Figure 16 quantifies this as average value density per paper across tracks.

Findings by parent conference. Figure 17 breaks down Findings papers by their parent conference (ACL, EMNLP, NAACL). The near-identical profiles across all three confirm that the Findings format itself, rather than the venue's thematic culture, determines the value profile.

Workshop profiles. Figure 18 illustrates how thematically focused workshops diverge from the main-conference value profile. BioNLP and ClinicalNLP emphasise Safety (due to high-stakes clinical applications); BlackboxNLP centres on Transparency and Interpretability; WASSA shows elevated Fairness. Notably, Performance drops to

30–45% at ethically-oriented workshops, the only context in the corpus where Performance is not the dominant value.

Paradigm radar charts. Figures 19 and 20 visualise value profiles as radar charts for adjacent paradigm transitions and for the full 2010–2025 span. The 2014→2018 transition (right panel of Figure 19) shows the first visible expansion of ethical values; the 2010→2025 comparison (right panel of Figure 20) captures the full magnitude of change, with every dimension of the radar chart expanding outward.

Per-venue top and bottom values. Figures 21 and 22 show the top-5 and bottom-5 values for each of the six major venues. The three flagship conferences (ACL, EMNLP, NAACL) share nearly identical rankings, confirming the disciplinary homogeneity discussed in §4. LREC shows a distinctive profile with lower Performance emphasis, consistent with its focus on resource and evaluation papers rather than model development.

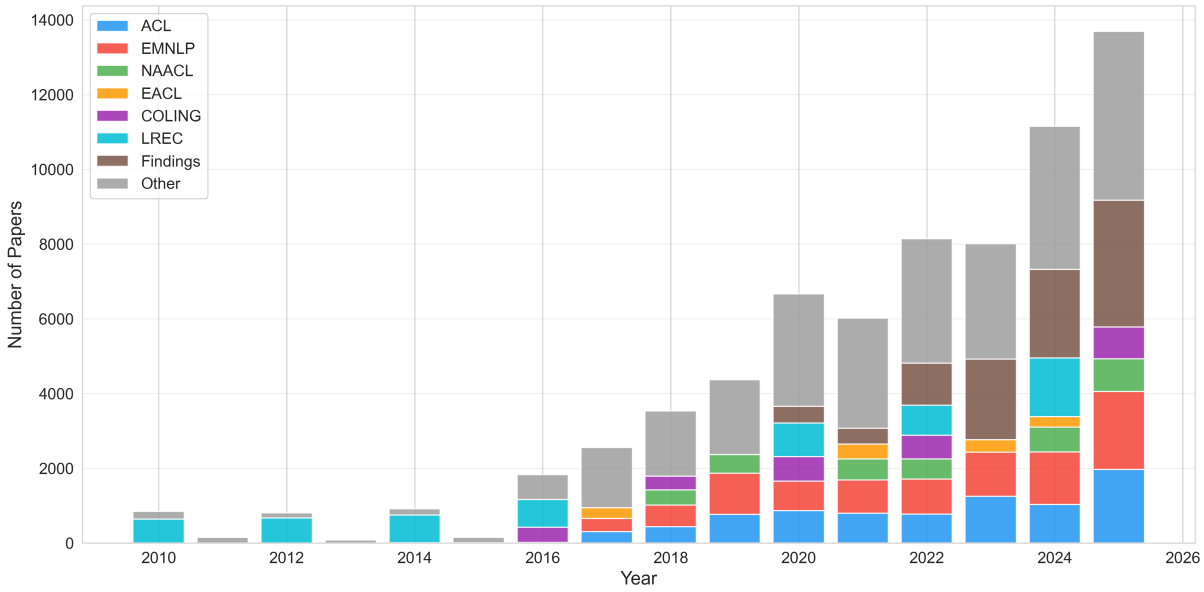

Figure 12: Number of analysable papers per year (2010–2025). The sharp increase from 2018 onward reflects the growth of the NLP community. Low counts in 2011, 2013, and 2015 are due to missing abstracts in the ACL Anthology metadata.

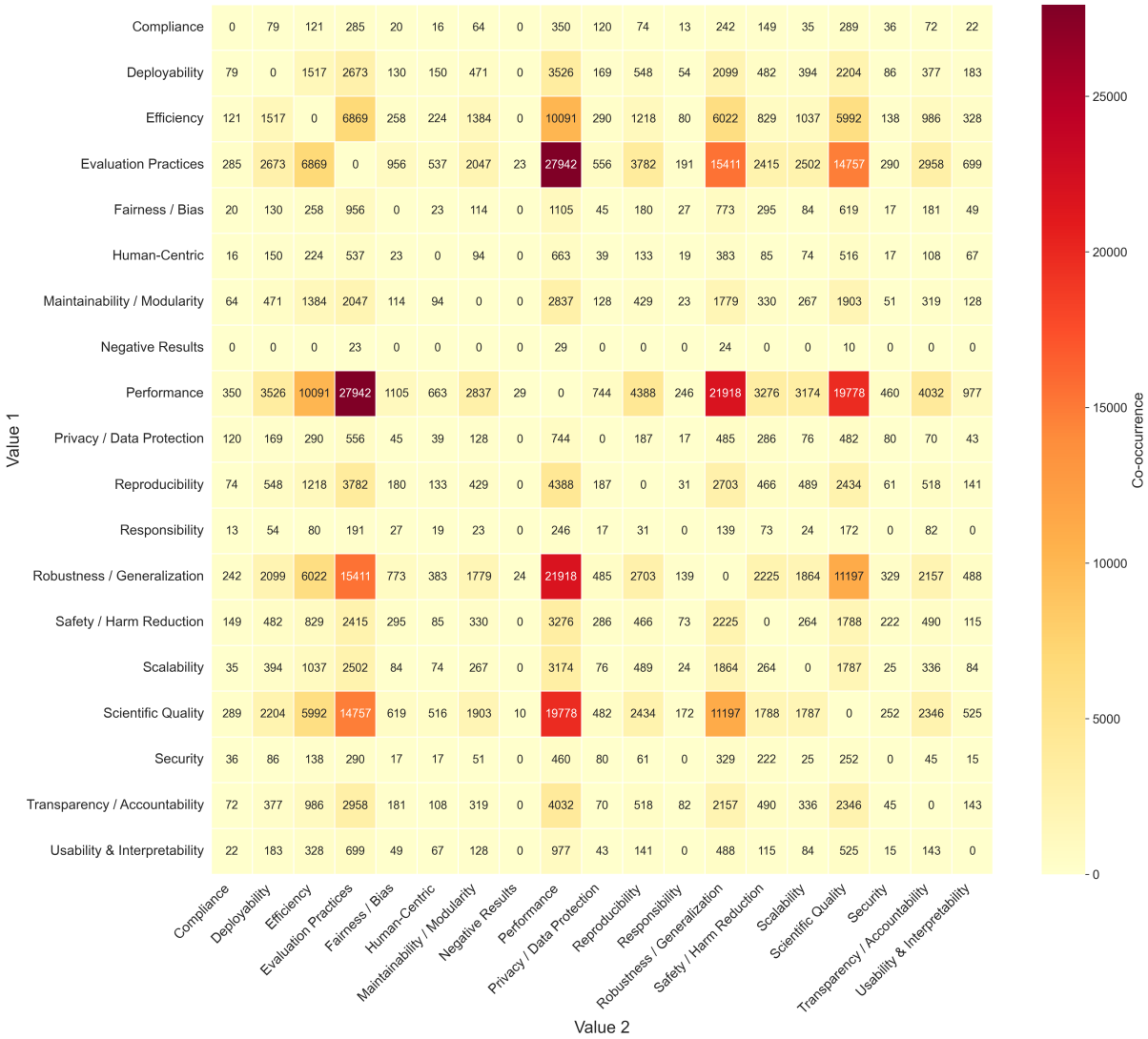

Figure 13: Pairwise co-occurrence correlation matrix for all 19 values. Negative Results shows near-zero correlation with all other values, confirming its structural isolation. Strong positive clusters emerge among ethical values (Fairness, Safety, Responsibility) and among technical values (Performance, Robustness, Evaluation Practices).

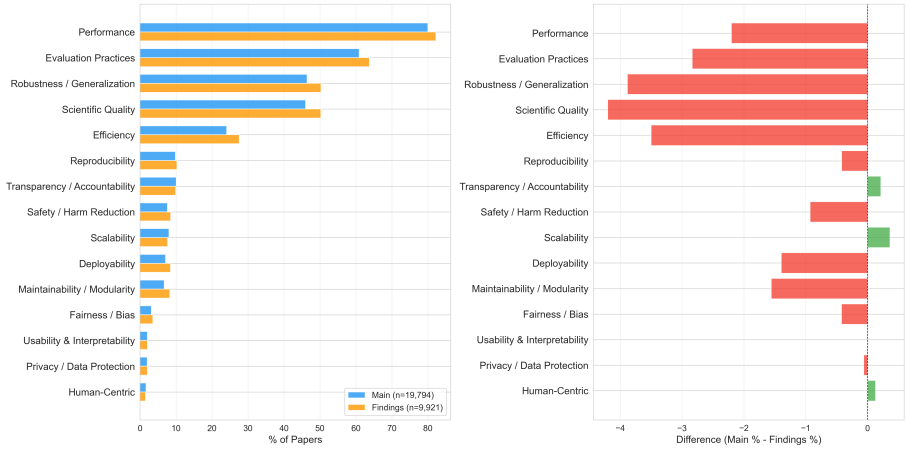

Figure 14: Value profiles of Main vs. Findings papers. Differences are below 5 percentage points across all values, challenging perceptions of Findings as a qualitatively different tier.

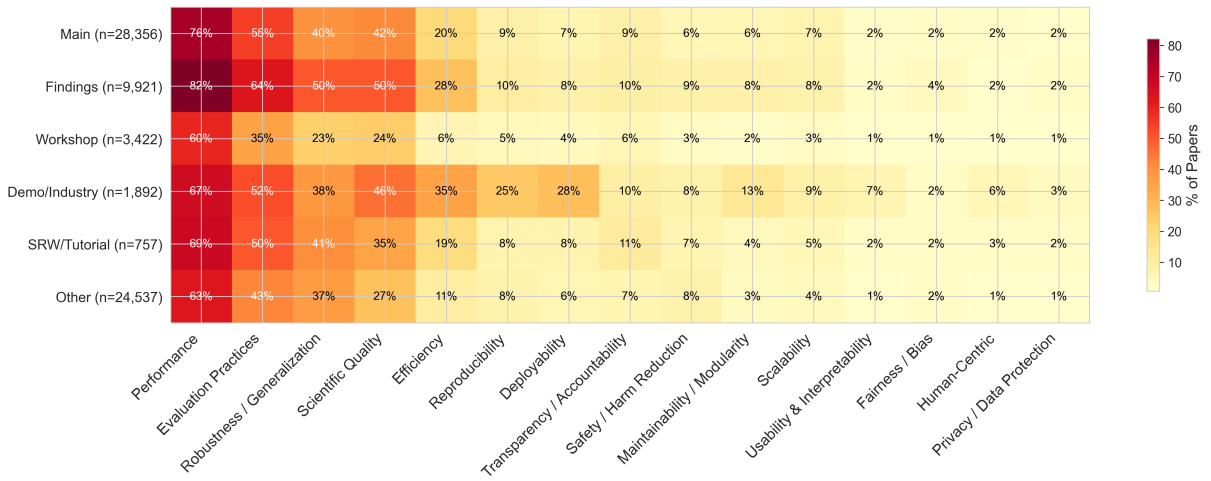

Figure 15: Value distribution across publication tracks (Main, Findings, Demo, Industry, Workshop, etc.). Demo and Industry tracks show the highest engagement with Deployability and Maintainability.

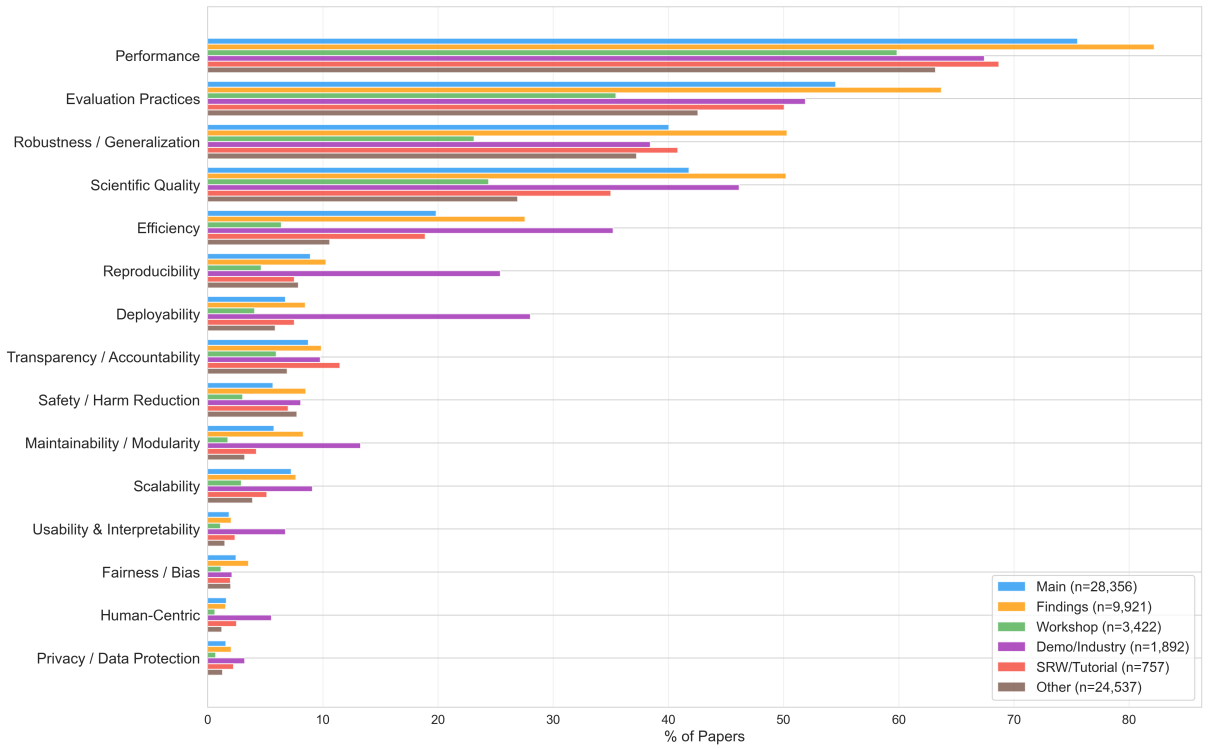

Figure 16: Average values per paper across different publication tracks. Industry papers show the highest value density (4.2 values/paper), followed by Demo (3.5).

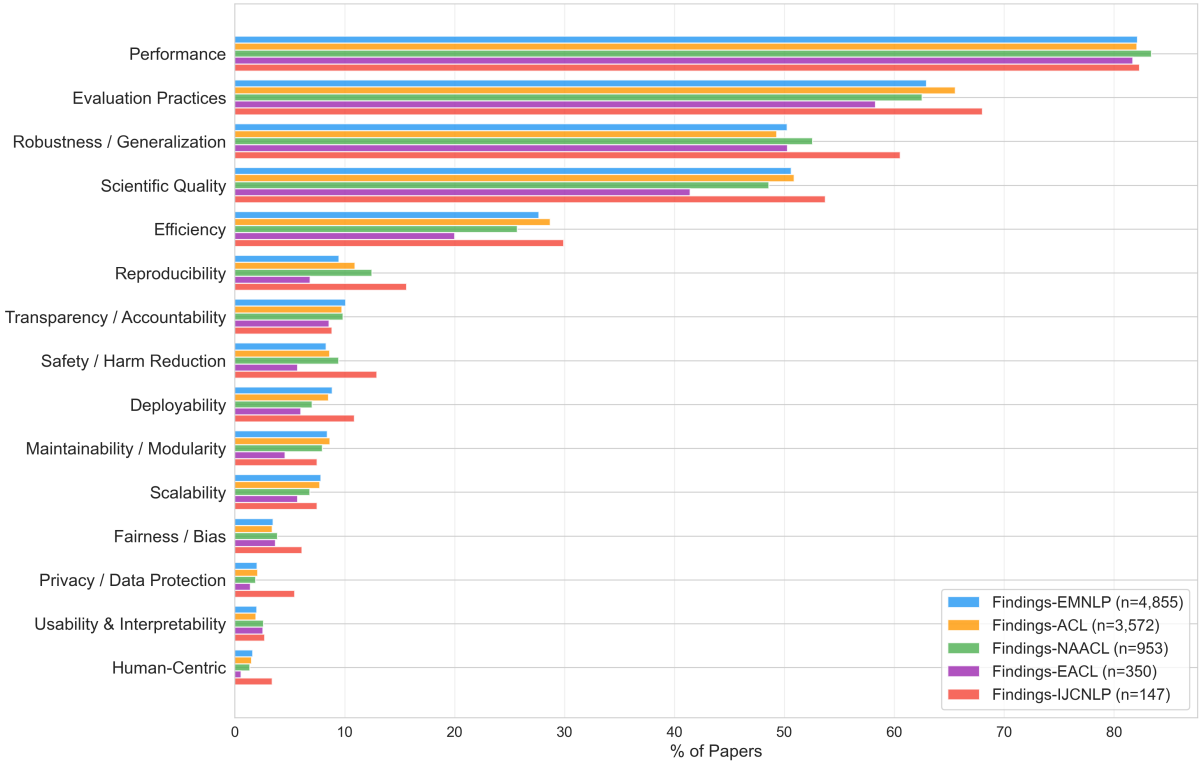

Figure 17: Value profiles of Findings papers broken down by parent conference (ACL Findings, EMNLP Findings, NAACL Findings). Differences between parent conferences are minimal, reinforcing the homogeneity finding.

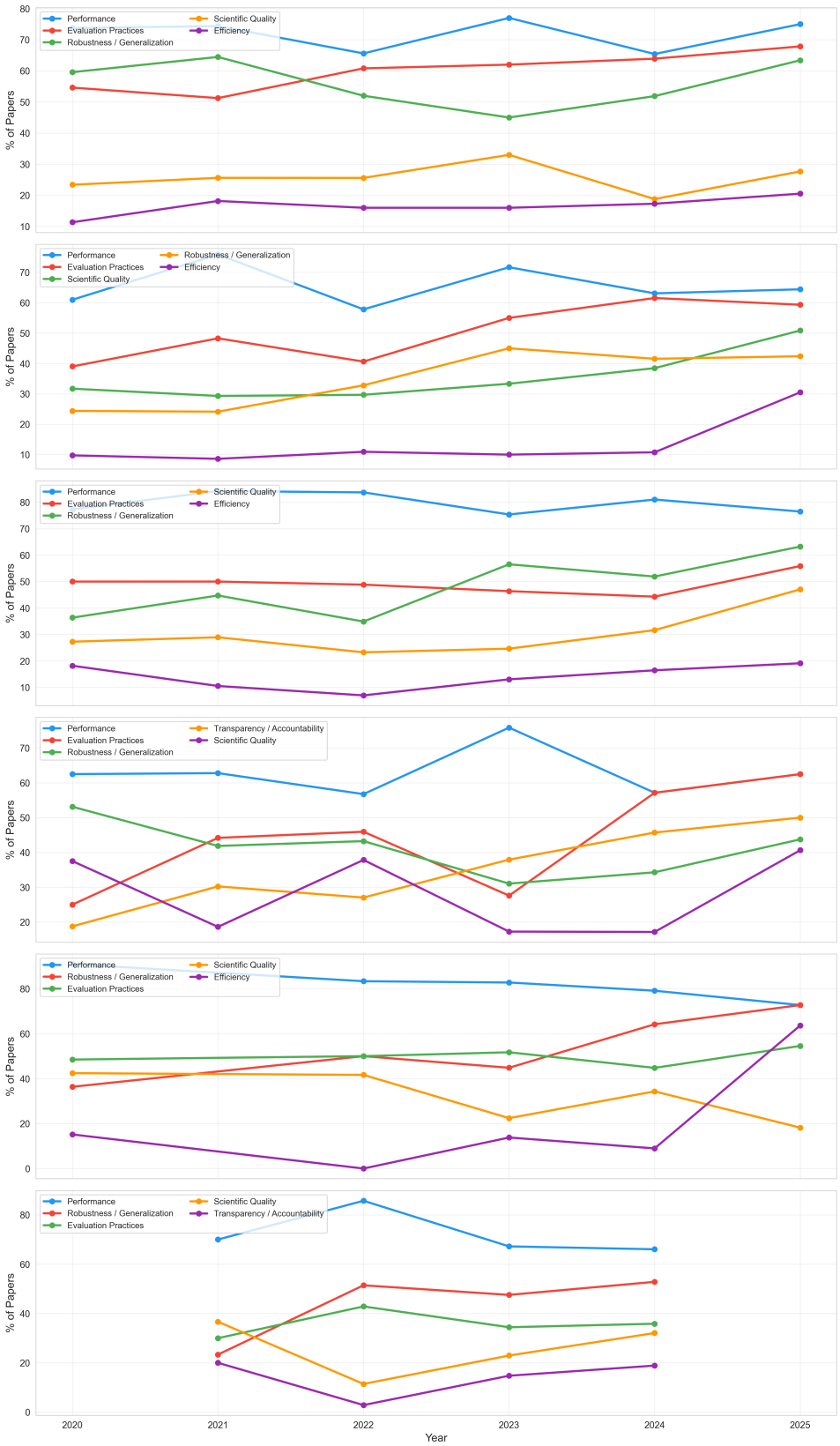

Figure 18: Value profiles of selected workshops. BioNLP and ClinicalNLP emphasise Safety; BlackboxNLP centres on Transparency; WASSA shows elevated Fairness. Performance drops to 30–45% at ethically-oriented workshops.

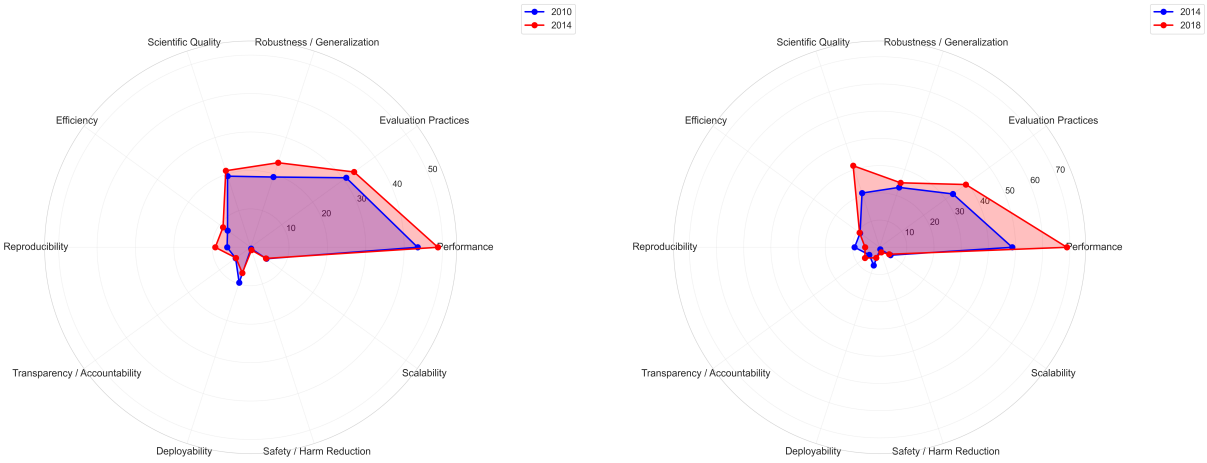

Figure 19: Pairwise paradigm radar charts: Pre-neural → Neural (left) and Neural → Transformer (right). The 2014→2018 transition shows the first visible expansion of ethical values.

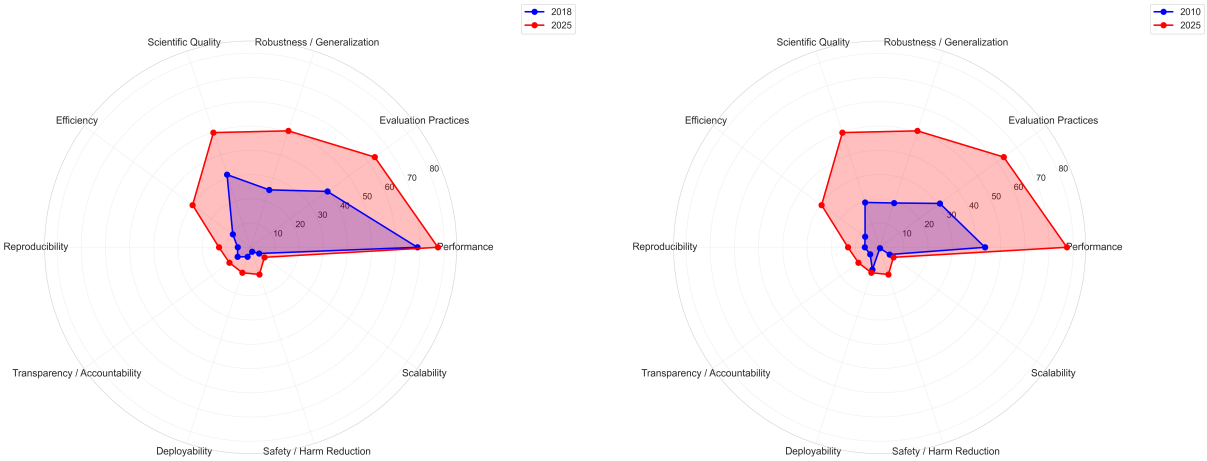

Figure 20: Pairwise paradigm radar charts: Transformer → LLM (left) and full span 2010→2025 (right). The overall expansion across all value dimensions is clearly visible in the 15-year comparison.

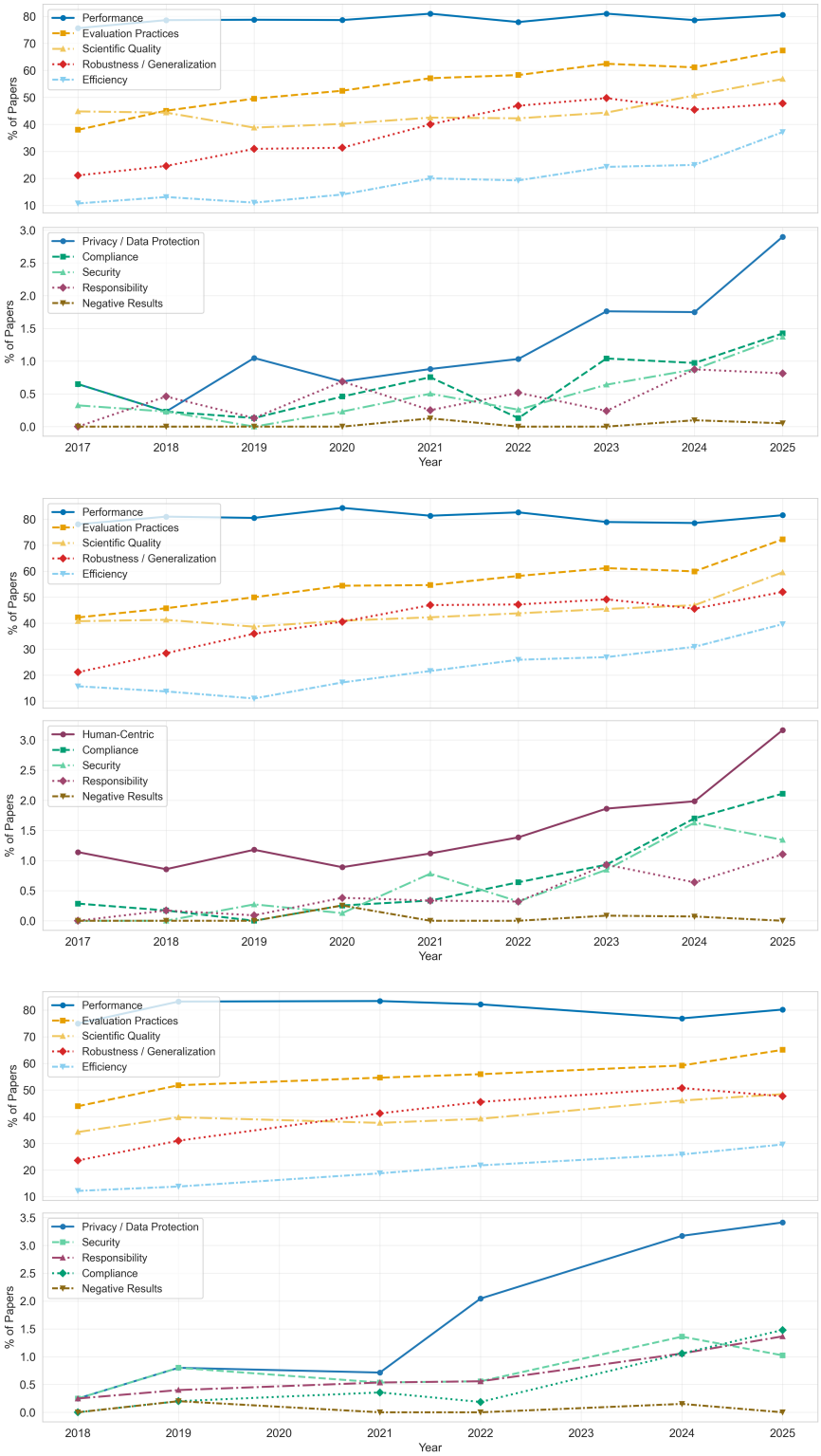

Figure 21: Top and bottom values per venue (ACL, EMNLP, NAACL). The three flagship venues share nearly identical top and bottom rankings, confirming disciplinary homogeneity.

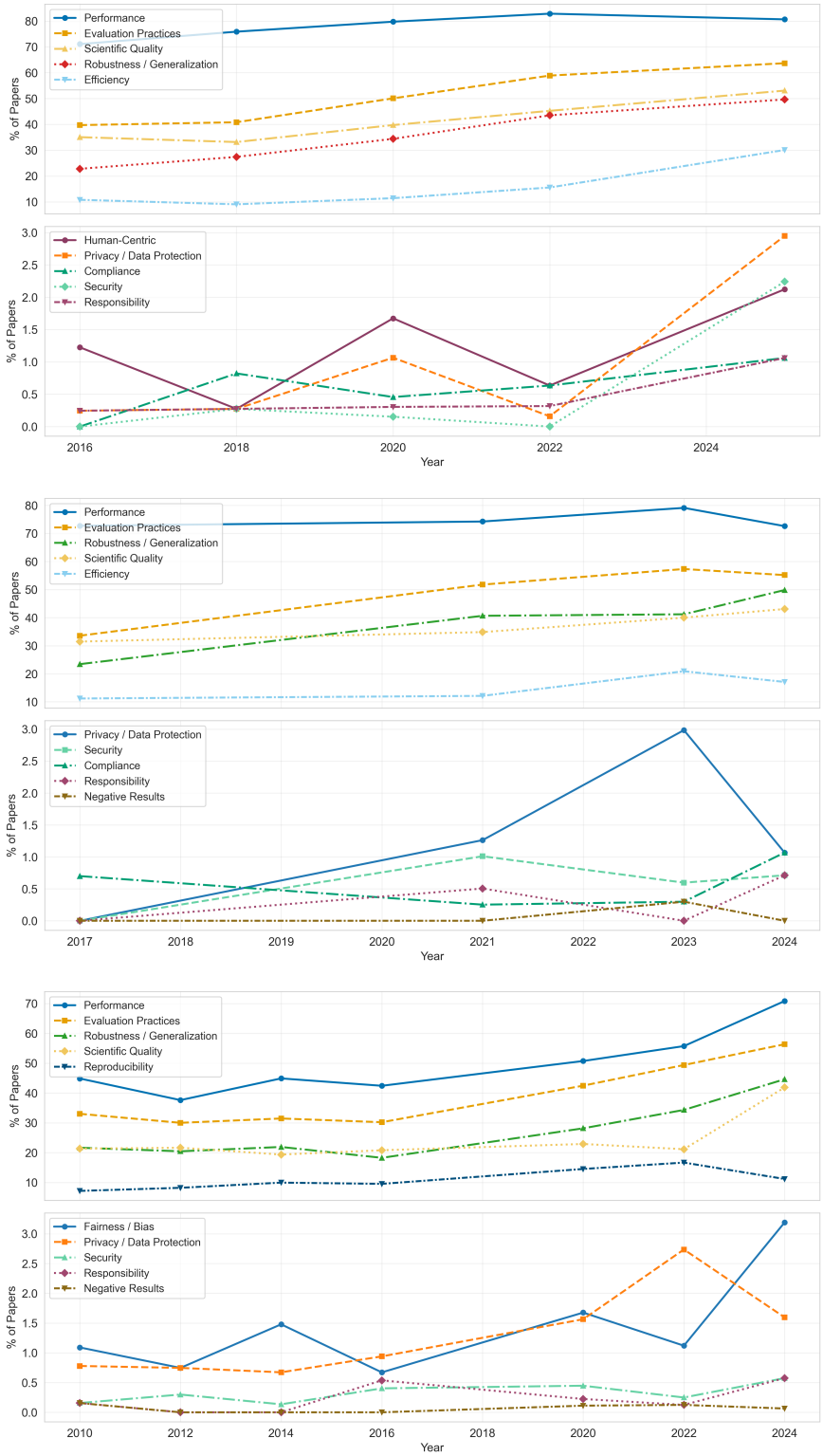

Figure 22: Top and bottom values per venue (COLING, EACL, LREC). COLING and EACL align closely with the ACL cluster; LREC shows a distinctive profile with lower Performance emphasis.