Conference Expansion and Scholarly Impact: A Ten-Year Analysis of Major NLP and AI Venues

Anonymous ACL submission

Abstract

The recent surge of language models (LMs) has rapidly expanded NLP/AI research, driving exponential growth in submissions and acceptances at major conferences. Yet this expansion has also raised concerns about conference quality and review reliability. This raises a fundamental question: *does conference expansion still translate into proportional scientific influence?* We conduct a ten-year empirical study (2014–2024) across seven leading NLP and AI conferences. We build a four-dimensional bibliometric framework and further propose *Quality–Quantity Elasticity* (QQE), a metric measuring whether citation growth keeps pace with publication expansion. Our results reveal three findings. First, most venues remain below unit elasticity (e.g., ICLR 0.82, NeurIPS 0.91), indicating that publication growth generally outpaces citation growth; only ACL (1.20) and NAACL (1.09) exceed unit elasticity. Second, conference scale, impact level, and growth efficiency exhibit substantially different hierarchies: NeurIPS and AAAI dominate in scale, ICLR achieves the strongest normalized influence, while ACL maintains the strongest balance between citation head and tail. Third, QQE captures a complementary dimension beyond conventional citation metrics, showing heterogeneous correlations across venues ($r = 0.88$ vs. -0.87)[1].

1 Introduction

The evolution from neural language models to large language models (LLMs) has propelled NLP research to unprecedented growth, leading to a dramatic increase in submissions and acceptances at major conferences (Naveed et al., 2025). For example, ACL main conference submissions increased from 692 in 2015 to 4,407 in 2024, while accepted long papers grew from 173 to 865 (Ku et al., 2024), as shown in Fig. 1. While this expansion reflects the vitality of the field, it has also raised growing concerns about conference quality and academic governance.

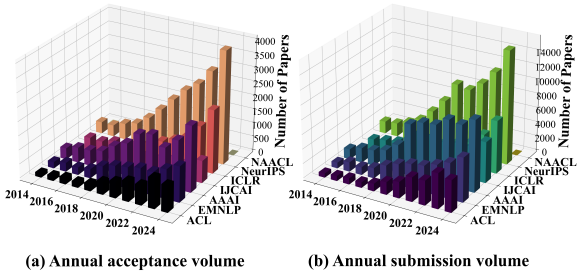

Figure 1: Annual submissions and acceptances across major NLP/AI conferences from 2014 to 2024.

Recent years have witnessed growing concerns over conference quality, including plagiarism, reviewer overload, collusive review practices, and inconsistent peer-review standards. Public controversies such as the "*Who's Adam?*" case[2], the AAAI bid-coordination incident[3], and the ICLR 2026 review dispute[4] have further intensified doubts about the overall quality of major conferences. However, conference quality is inherently multidimensional, involving factors such as review rigor, research novelty, reproducibility, and long-term scientific value, which are difficult to quantify consistently at scale. In practice, scholarly influence is often treated as a measurable proxy, reflecting the extent to which conference publications shape subsequent scientific progress.

From this perspective, a common assumption is that larger conferences should generate proportionally greater scholarly influence, as more

[1] Our codes: https://anonymous.4open.science/r/acl-crown-analysis-38D5.

[2] A NeurIPS 2025 review mistakenly treats the Adam optimizer as a person's name.

[3] A private group of authors and reviewers allegedly coordinated their bid selections.

[4] A public controversy involving an extreme review outcome and subsequent author–reviewer conflict.

accepted papers imply broader participation and greater opportunities for knowledge dissemination. However, scholarly attention is inherently uneven. As conference volume expands, attention may become fragmented, increasing the proportion of low-visibility papers and diminishing the marginal impact of newly published work. Existing studies assess conference quality through peer-review analysis (Rogers and Augenstein, 2020; Wei et al., 2025), citation-based metrics (Meho, 2019), acceptance rates, committee composition (Zhuang et al., 2007), and reputational evaluation (Loizides and Koutsakis, 2017). However, these approaches remain largely fragmented and do not directly examine the long-term relationship between conference expansion and scholarly impact.

To address this gap, we conduct a ten-year empirical analysis (2014–2024) of 60,406 long papers from seven major NLP/AI conferences: ACL, EMNLP, NAACL, AAAI, IJCAI, ICLR, and NeurIPS. We develop a four-dimensional bibliometric framework spanning conference scale, scholarly influence, citation prestige, and *Quality–Quantity Elasticity* (QQE), a new metric that quantifies whether citation growth keeps pace with publication expansion. Our analysis reveals three key findings: (1) most conferences remain below unit elasticity, suggesting that publication growth generally outpaces citation growth; (2) conference scale and scholarly influence become increasingly decoupled under rapid expansion; and (3) QQE captures a complementary efficiency dimension beyond conventional citation metrics. Our work makes the following main contributions:

- We provide a longitudinal ten-year quantitative analysis across seven major NLP/AI conferences.
- We propose *QQE*, an elasticity-based metric linking acceptance growth to citation-impact growth.
- We provide empirical evidence that rapid conference expansion does not necessarily imply proportional scientific growth.

2 Related Work

Conference quality assessment has long been recognized as a critical issue in computer science, dating back to the CRA *Best Practices Memo* (1999). Early work has explored conference evaluation using a variety of bibliometric and reputational features, including citation counts (Meho, 2019), acceptance rates, conference tradition (Bedogni et al., 2023), and program committee characteristics (Zhuang et al., 2007), sometimes combined with machine learning models for automated ranking and classification. Subsequent studies further extend these approaches with composite indicators such as peer reputation (PR) (Loizides and Koutsakis, 2017), CiteScore, and multi-factor classification schemes that integrate citation impact and author-level metrics (Mohammad, 2020; Meho, 2019; Iqbal et al., 2021).

Conference	No. of Papers	Citations
ACL	4,965	488,499
EMNLP	7,106	573,104
NAACL	2,540	414,779
AAAI	12,786	713,125
IJCAI	6,617	245,391
ICLR	7,691	1,771,388
NeurIPS	18,701	2,375,612
Total	**60,406**	**6,581,898**

Table 1: Paper and citation statistics.

Complementary work evaluates conference quality through expert- and community-based mechanisms, including curated ranking systems such as CORE, MSAR, and the CCF recommendations. While such approaches provide domain-informed categorizations, they are inherently subjective and limited in capturing fine-grained variation or temporal dynamics (Rogers and Augenstein, 2020; Kuznetsov et al., 2019; Russo et al., 2025). However, these methods primarily produce static or short-term rankings and do not systematically examine the long-term, cross-venue evolution of conferences.

3 Methodology

3.1 Data Acquisition

(1) Venue Selection. We focus on seven widely recognized leading conferences: the core NLP venues *ACL*, *EMNLP*, and *NAACL*; and the core AI venues *AAAI*, *IJCAI*, *NeurIPS*, and *ICLR*. We restrict our analysis to main-track long papers. Findings papers, short papers, and workshop papers are excluded.

(2) Temporal Scope. The study spans 2014–2024. The year 2014 serves as a pre-expansion baseline, preceding the rapid growth of ACL submissions, while 2024 is chosen as the lat-

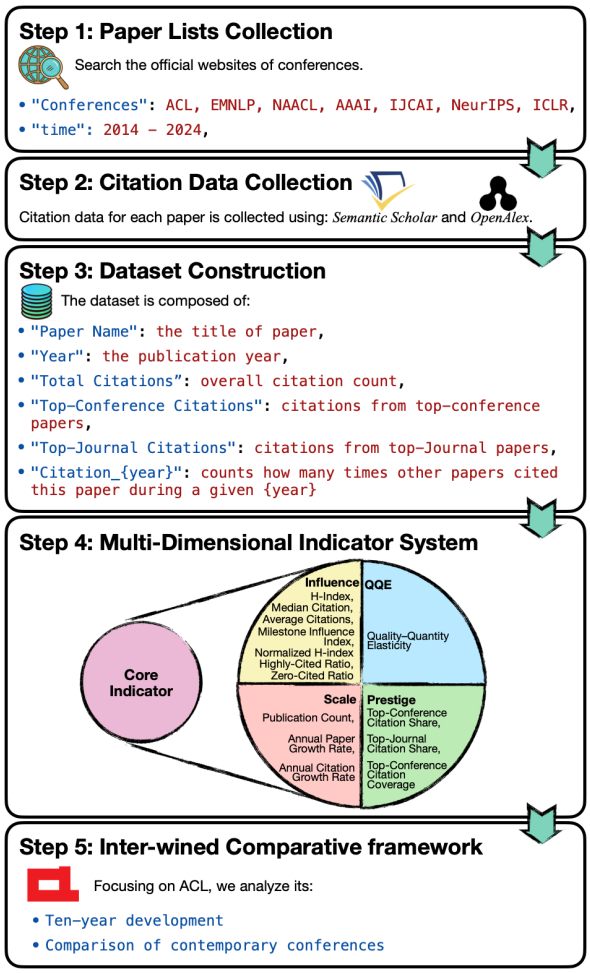

Figure 2: The proposed bibliometric framework.

est year with complete and comparable proceedings across all venues[5].

(3) Data Sources. Because Google Scholar does not provide an official API due to licensing restrictions, we collect paper metadata and citation statistics from established open sources. Paper titles are obtained from the official conference proceedings. Citation data are primarily sourced from Semantic Scholar, which successfully matches 99.3% of all papers in our dataset, and OpenAlex is used only as a supplemental source to cover the remaining unmatched entries (0.7%). Finally, the dataset contains 60,406 long papers across the seven conferences, with more than 6M citation records, as shown in Tab. 1 (Details please see App. A and B).

3.2 Multi-Dimensional Bibliometric Framework

We develop a four-dimensional bibliometric framework, as shown in Fig. 2.

[5]Up to January 2026, ensuring post-publication citation exposure for 2024 papers.

(1) Scale Dimension. This dimension quantifies the growth of each conference as a scholarly venue. It includes three indicators: *(i) publication count (PC)*, the total number of accepted long papers, reflecting the venue's output capacity; and *(ii) annual paper growth rate (APGR)*, the year-over-year percentage increase in accepted papers, which captures how rapidly a conference expands in size. *(iii) annual citation growth rate (ACGR)*, the year-over-year percentage increase in total citations accumulated by papers from each conference.

(2) Influence Dimension. To capture the magnitude and distribution of scholarly influence, we use six indicators: *(i) normalized average citation (N-AC)* and *(ii) normalized median citations (N-MC)*, representing central tendencies of citation performance; *(iii) highly cited ratio (HCR)* (share of papers with more than 100 citations), measuring the upper tail of impact; and *(iv) zero-cited ratio (ZCR)*, quantifying the share of papers that never attracted citations; *(v) H-index*, which reflects the concentration of highly influential papers; *(vi) milestone influence index (MII)*, defined as the proportion of papers exceeding both an absolute citation threshold ($\geq$1000 citations)[6] (Wang et al., 2021) and a relative threshold (top 0.1% of the annual citation distribution) (Wang et al., 2023).

(3) Prestige Dimension. Beyond citation volume, we analyze citation sources to situate a venue in the scholarly network. We compute: *(i) top-conference citation share (TCS)*, the fraction of citations coming from leading peer conferences; *(ii) top-journal citation share (TJS)*, the fraction coming from elite journals, reflecting broader uptake; and *(iii) top-conference citation coverage (TCC)*, the fraction of the venue's papers that receive at least one citation from a top conference, reflecting penetration breadth. Top conferences are defined as the union of CCF A-level and CORE A* venues, and top journals are represented by *Nature* and *Science* (Ioannidis, 2006), not as exhaustive indicators of domain-specific journal prestige, but as signals of exceptional cross-disciplinary visibility beyond the core AI/NLP community.

(4) Quality–Quantity Elasticity (QQE). While static bibliometric indicators capture snapshots of scale or impact, they fail to track whether expansion is supported by a proportional diffusion of knowledge. We propose *Quality–Quantity Elas-*

[6]Scientific milestones lack a unified definition and are commonly proxied using citation-based indicators in scientometric studies.

ticity (QQE), grounded in scientific growth theory. This framework posits that an efficient venue should exhibit proportional or superlinear growth in citation-based influence relative to its paper volume, as a larger knowledge stock should generate greater spillover effects. We define N_t as the total number of papers published by the conference in year t. We define C_t as the total normalized citations received in year t by all papers published by the conference up to and including year t.

$$QQE_t = \text{sign}(s_t) \cdot \frac{C_t/C_{t-1}}{N_t/N_{t-1}} \quad (1)$$

To distinguish between operational modes, we define the log-growth of scale as $s_t = \ln(N_t/N_{t-1})$, where $s_t > 0$ denotes *expansion* and $s_t < 0$ denotes *contraction*.

Impact-Scale Elasticity Regime. The relationship between citation growth and scale expansion is characterized by QQE_t: *(i)* $|QQE_t| > 1$ *(Effective Elasticity).* Indicates that citation-based influence changes faster than scale. When $QQE_t > 0$, this corresponds to *efficient expansion*; when $QQE_t < 0$, it reflects *quality-concentrating contraction*, where impact density increases despite shrinking volume. *(ii)* $|QQE_t| = 1$ *(Unitary Elasticity).* Indicates proportional change, where citation mass evolves in direct alignment with conference scale. *(iii)* $|QQE_t| < 1$ *(Ineffective Elasticity).* Indicates lagging impact gains, corresponding to *impact dilution* during expansion ($QQE_t > 0$) or *malignant contraction* during scale reduction ($QQE_t < 0$).

Validity and Robustness of QQE. QQE must demonstrate more than mathematical validity. The key issue is whether it provides explanatory power beyond conventional citation-based indicators. Specifically, it should be robust to outliers and yearly fluctuations, provide interpretable scale–impact dynamics, and capture information beyond conventional citation metrics (e.g., total citations and H-index). Therefore, we evaluate QQE in terms of robustness, discriminability, complementarity, and validity in Sec. 4.2.

3.3 Analytical Method

We adopt a combined cross-sectional and longitudinal analysis. At the cross-sectional level, we compare conferences using ten-year averages of bibliometric indicators covering scale, citation impact, citation distribution, citation sources, and QQE. At the longitudinal level, we track annual dynamics using six representative indicators (N-AC, N-MC, HCR, ZCR, MII, and QQE), with additional comparisons among ACL, EMNLP, and NAACL. (Details please see App. D).

3.4 Normalization Controls

Since the selected conferences differ substantially in scale, we apply three normalization controls to adjust for differences in publication volume.

(1) Normalized H-index. The conventional H-index increases with the number of publications, which introduces a systematic bias favoring larger conferences. Hence, we compute a *normalized H-index* defined as:

$$\text{N-H}(\%) = \frac{H\text{-index}}{\text{total publications}} \times 100 \quad (2)$$

This ratio reflects the density of highly influential papers per accepted paper.

(2) Field-normalized Citation Metrics. Direct comparison of citation counts across conferences is biased by differences in research scope, community size, and citation practices. To mitigate these effects, we perform field normalization based on a set of nine main research categories (Sec. 5). Each paper is assigned to a primary category, and its citation count is normalized by the average citation within the same field and publication year:

$$\text{Norm-Citation}_i = \frac{C_i}{\mathbb{E}[C \mid f(i), t(i)]} \quad (3)$$

where C_i denotes the citation count of paper i, and $\mathbb{E}[C \mid f(i), t(i)]$ is the mean citation count of papers in the same field $f(i)$ and year $t(i)$.

To improve robustness under heavy-tailed citation distributions, we report both mean-based (N-AC) and median-based (N-MC) statistics computed from normalized values.

(3) Gini Coefficient of Citation Distribution. Citations in NLP/AI conferences typically follow a long-tail pattern. As conference volume increases, this tail may expand, potentially inflating total citation counts without improving median quality. To measure this concentration, we compute the annual *Gini coefficient* for each venue. Let citation counts be sorted as $c_1 \leq c_2 \leq \cdots \leq c_n$, where n is the paper count and μ is the mean citation count. The Gini coefficient is computed as:

$$\text{Gini} = \frac{1}{n\mu} \sum_{i=1}^{n} (2i - n - 1) c_i \quad (4)$$

	Scale			Influence							Prestige			QQE
Conference	PC	APGR	ACGR	N-AC	N-MC	HCR	ZCR	H	N-H(%)	MII	TCS	TJS	TCC	Avg-|QQE|
ACL	4965	22.06	87.89	0.92	0.35	20.75	0.89	299	6.02	1.09	31.26	2.04	**94.34**	**1.20**
EMNLP	7106	20.68	151.00	0.64	0.22	13.72	2.03	284	4.00	0.73	23.22	1.56	89.39	0.95
NAACL	2540	20.53	**231.63**	1.11	0.26	15.83	1.61	187	**7.36**	0.75	26.63	1.90	91.54	1.09
AAAI	12786	21.54	87.11	0.61	0.26	10.88	2.00	315	2.46	0.48	12.89	1.71	85.53	0.99
IJCAI	6617	5.08	129.11	0.34	0.16	7.42	**3.73**	196	2.96	0.17	8.29	1.07	80.94	0.98
ICLR	7691	**54.63**	92.60	**2.27**	**0.72**	**26.90**	1.52	494	6.42	**2.94**	**45.77**	**3.91**	91.91	0.82
NeurIPS	**18701**	26.44	117.74	1.12	0.39	16.47	2.34	**564**	3.02	1.51	26.06	2.68	85.60	0.91

Table 2: Comparison of major NLP/AI conferences across four dimensions. **Bold** indicates the highest value, and underlining indicate the lowest value.

A value of 0 indicates a uniform distribution, while values approaching 1 signify extreme concentration in a small subset of "blockbuster" papers.

4 Experiment

4.1 Main Results

Tab. 2 reports the ten-year average results of seven conferences.

Scale Dimension. Conference scale is highly uneven. NeurIPS ($PC = 18{,}701$) and AAAI (12,786) are the largest venues, whereas NAACL has the smallest volume (2,540). However, growth rate does not simply follow size: ICLR shows the fastest paper expansion ($APGR = 54.63$), while IJCAI grows the slowest (5.08). Citation growth presents a different pattern, with NAACL ($ACGR = 231.63$) and EMNLP (151.00) leading, whereas AAAI (87.11) and ACL (87.89) are much lower. This suggests that publication expansion and citation growth are not tightly aligned.

Influence Dimension. Per-paper influence differs sharply from raw scale. ICLR ranks highest in normalized citation impact ($N\text{-}AC = 2.27$, $N\text{-}MC = 0.72$) and highly cited ratio ($HCR = 26.90$), showing strong impact intensity despite its relatively smaller volume than NeurIPS and AAAI. NeurIPS has the highest absolute H-index ($H = 564$), but its normalized H-index is low ($N\text{-}H = 3.02$), indicating that raw influence is partly driven by scale. By contrast, NAACL obtains the highest normalized H-index (7.36), while ACL has the lowest zero-cited ratio ($ZCR = 0.89$). Larger conferences maintain strong absolute influence, but do not consistently achieve higher impact density.

Prestige Dimension. ICLR shows the strongest citation connection with elite venues and journals ($TCS = 45.77$, $TJS = 3.91$), followed by ACL and NAACL. In contrast, AAAI ($TCS = 12.89$) and IJCAI (8.29) have much weaker top-conference citation shares. ACL achieves the highest top-conference citation coverage ($TCC = 94.34$), while IJCAI remains the lowest (80.94).

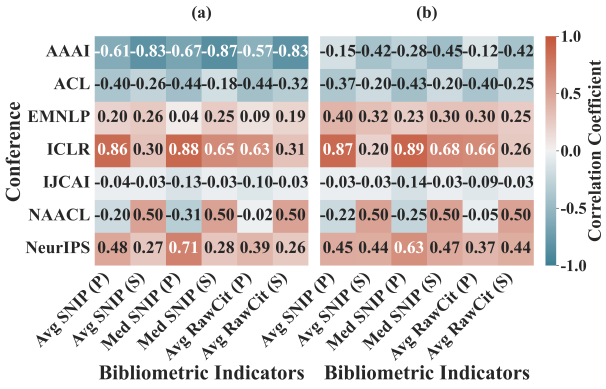

Figure 3: Correlation between QQE and citation-based indicators. (a) Full dataset. (b) After removing the top 1% citation outliers.

QQE Dimension. QQE captures the relative magnitude of citation change with respect to conference expansion. Table 2 reports the ten-year mean absolute QQE values, reflecting average scale–impact responsiveness regardless of direction. Only ACL ($|QQE| = 1.20$) and NAACL (1.09) exceed unit elasticity, indicating stronger citation dynamics relative to scale growth, while other venues remain below the threshold. Notably, ICLR combines the strongest normalized impact with the lowest average absolute QQE (0.82), suggesting weaker scale–impact responsiveness despite its high citation influence. Overall, the four dimensions reveal distinct conference profiles: NeurIPS and AAAI lead in scale, ICLR in normalized influence, ACL in citation penetration, and ACL/NAACL in scale–impact elasticity.

4.2 Validity and Robustness of QQE

Complementarity. As shown in Fig. 3, QQE exhibits heterogeneous correlations with existing ci-

Conf.	Avg(N)-P	Avg(N)-S	Med(N)-P	Med(N)-S	H(N)-P	H(N)-S	Avg(R)-P	Avg(R)-S
AAAI	0.063	**0.0029**	**0.0339**	**0.0011**	0.121	**0.0029**	0.087	**0.0029**
ACL	0.245	0.467	0.189	0.511	0.177	0.365	0.211	0.365
EMNLP	0.580	0.446	0.960	0.467	0.804	0.603	0.806	0.603
ICLR	**0.0013**	0.701	**0.0008**	**0.0376**	0.159	0.603	**0.0476**	0.533
IJCAI	0.925	0.932	0.711	0.932	0.951	0.932	0.802	0.932
NAACL	0.864	0.667	0.842	0.667	0.979	0.667	0.975	0.667
NeurIPS	0.181	0.446	**0.0460**	0.385	0.328	0.467	0.273	0.467

Table 3: P-values of correlations between QQE and normalized and raw citations. Bold values denote statistical significance ($p < 0.05$). P/S: Pearson/Spearman; N/R: normalized/raw.

tation indicators rather than consistent alignment. Here, *avg_SNIP* and *median_SNIP* represent average and median field-normalized cumulative citations. ICLR shows strong positive correlations with normalized citation metrics ($r = 0.86 \sim 0.89$), whereas AAAI shows strong negative correlations (down to -0.87). ACL, EMNLP, and IJCAI remain weakly correlated with most metrics. Table 3 further shows that significant correlations mainly appear in ICLR and AAAI, rather than uniformly across venues.

Discriminability. QQE separates growth efficiency from raw scale and static impact. NeurIPS has the largest scale ($PC = 18{,}701$) and highest H-index ($H = 564$), but remains below unit elasticity ($QQE = 0.91$). ICLR achieves the strongest normalized influence ($Norm\text{-}AC = 2.27$, $HCR = 26.90$), yet has the lowest QQE (0.82).

Robustness. After removing the top 1% most-cited papers, the main correlation patterns remain stable (Fig. 3b). ICLR remains strongly positive ($r = 0.87 \sim 0.89$), AAAI remains negative, and IJCAI remains near zero. The conference ranking after trimming also changes little (App. H), indicating that QQE is not dominated by a few highly cited papers.

Statistical Validity. Kruskal–Wallis tests (App. H) confirm significant citation-distribution differences across venues ($H = 5322.07$, $p < 0.001$). Bonferroni-corrected Mann–Whitney tests further show that ICLR, ACL, NeurIPS, EMNLP, and NAACL significantly outperform AAAI in median citation impact ($p < 0.001$), whereas IJCAI does not differ significantly from AAAI ($p = 0.213$).

4.3 Comprehensive Empirical Decomposition of Conference Trends (RQ1–RQ5)

RQ1. How Have Average and Median Citations Evolved Over Time?

Using a one-year citation window (Fig. 4), we

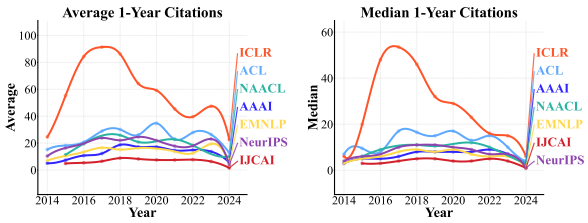

Figure 4: Temporal trends of average and median citations across major Conferences (2014–2024).

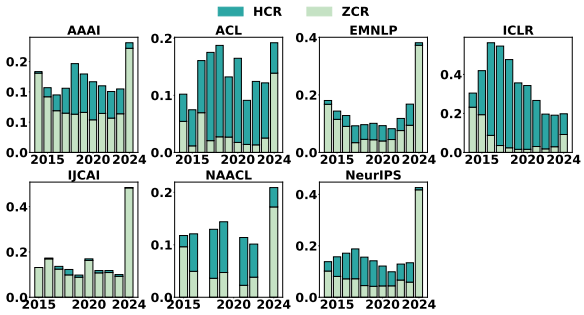

Figure 5: The head and tail of citation impact result.

observe substantial differences in both citation level and temporal volatility across venues. ICLR consistently achieves the highest average and median citations, peaking during 2016–2018. ACL exhibits the strongest fluctuations, with major citation surges around 2017–2020 and again in 2022–2023, temporally coinciding with major paradigm transitions from PLMs to LLMs. NeurIPS remains comparatively stable with moderate peaks, indicating a broader but less volatile influence structure. EMNLP and AAAI show relatively flat trajectories over most years, while IJCAI persistently records the lowest citation levels. Overall, ICLR dominates in short-term citation impact, ACL and NeurIPS form the second tier, and IJCAI remains the weakest venue.

RQ2. How Have the Head and Tail of Citation Impact Shifted?

Fig. 5 further shows that citation impact differs

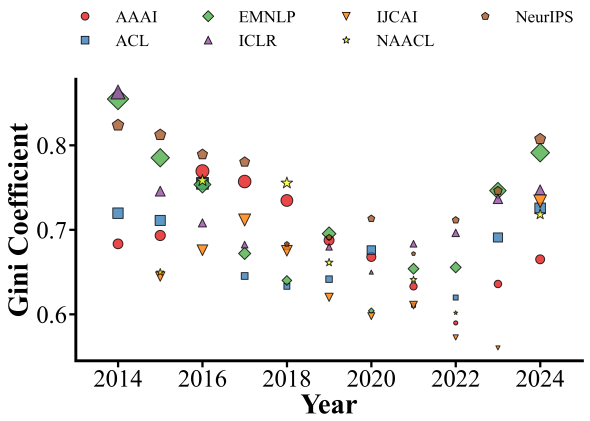

Figure 6: Gini coefficient trends (2014–2024).

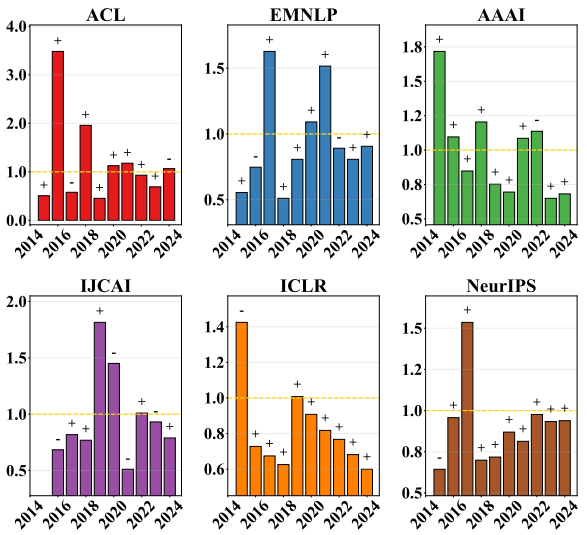

Figure 7: Annual QQE of major NLP/AI conferences (2014–2024).

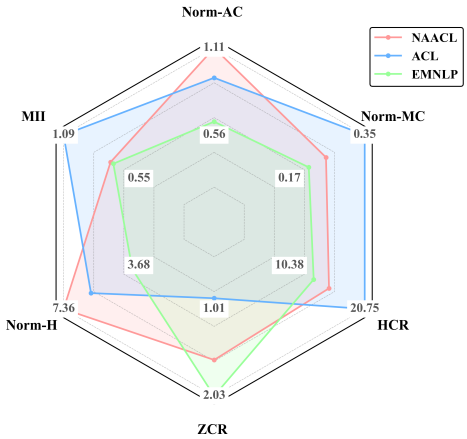

Figure 8: Ecology of ACL, EMNLP, and NAACL.

not only in central tendency but also in head–tail structure. ICLR has the strongest citation head, with HCR peaking at 51.01% in 2017, while maintaining very low ZCR during 2018–2020 ($< 2\%$). ACL also shows a clear increase in high-impact papers after 2017, reaching 15.46%, together with a consistently small citation tail. NeurIPS presents a more stable head–tail structure, without the sharp concentration observed in ICLR. EMNLP remains relatively stable but shows a visible tail expansion in the most recent year. In contrast, IJCAI has the weakest citation head, peaking at only 2.40%, and the heaviest tail, with ZCR rising to 16.27% in 2020 and nearly 50% by 2024.

RQ3. Is Citation Inequality Increasing Over Time?

Fig. 6 shows that Gini coefficients across all venues remain consistently high (0.6–0.8), indicating persistent citation concentration in conference publications. NeurIPS and ICLR generally exhibit the highest inequality, suggesting that citations are concentrated in a relatively small subset of influential papers. In contrast, IJCAI maintains the lowest Gini values, consistent with its overall weak citation impact. ACL and NAACL remain comparatively lower and more stable over time, whereas EMNLP and AAAI occupy the middle range. Most venues become relatively more balanced around 2019–2021, followed by renewed concentration in recent years.

RQ4. When Does Scale–Impact Decoupling Emerge?

Fig. 7 shows clear heterogeneity in scale–impact dynamics across venues. NeurIPS remains predominantly within the $|QQE| < 1$ regime over time, indicating stable but sub-unit elasticity. In contrast, ACL, EMNLP, and IJCAI frequently enter the $|QQE| > 1$ regime (e.g., ACL: 2.0, -1.1; IJCAI: 1.8, -1.5), where the magnitude of citation change exceeds that of scale. AAAI exhibits a transitional pattern, with multiple super-unit years in the early period (e.g., 1.7), but shifting toward $|QQE| < 1$ in recent years. Overall, scale–impact decoupling appears as a persistent long-term pattern, especially in rapidly expanding venues.

RQ5. How Has the Internal Ecology of NLP Conferences Evolved?

Fig. 8 shows clear differences among NLP venues. NAACL achieves the highest normalized average citations ($Norm\text{-}AC = 1.11$) and normalized H-index ($Norm\text{-}H = 7.36$), indicating the strongest impact density despite its smaller scale. ACL records the highest normalized median citations ($Norm\text{-}MC = 0.35$), highly cited ratio ($HCR = 20.75\%$), and milestone influence

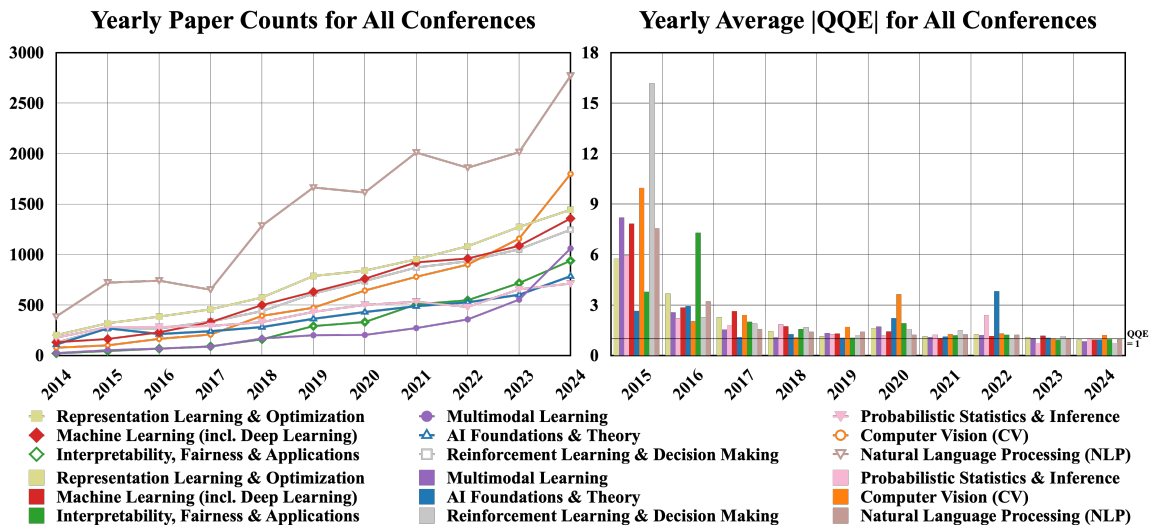

Figure 9: Historical trends in Paper Counts and QQE across different categories.

Category	r	p
Representation Learning & Optimization	-0.915	0.000
Multimodal Learning	-0.818	0.004
Probabilistic Statistics & Inference	-0.842	0.002
Machine Learning (incl. Deep Learning)	-0.939	0.000
AI Foundations & Theory	-0.455	0.187
Computer Vision (CV)	-0.673	0.033
Interpretability, Fairness & Applications	-0.867	0.001
Reinforcement Learning & Decision Making	-0.948	0.000
Natural Language Processing	-0.879	0.001

Table 4: Spearman rank correlations between annual publication counts and the QQE across categories. **r** represents **Spearman r** and **p** represents **Spearman p**.

($MII = 1.09$), while maintaining the lowest zero-cited ratio ($ZCR = 0.89\%$). In contrast, EMNLP has the highest zero-cited ratio (2.03%), suggesting a more diffuse impact distribution after expansion. Overall, ACL shows the strongest balance between citation head and tail, where NAACL achieves the highest per-paper impact density.

5 Publication Growth and Diminishing Impact: A Field-Level Extension

We conduct a field-level analysis by first classifying all papers into one of nine major research fields (Fig. 9, left) and computing the annual average absolute QQE for each field (Fig. 9, right). We then examine the Spearman rank correlation between annual publication volume and absolute QQE within each field over 2015–2024 (Tab. 4). Field classification is performed using the DeepSeek model. To assess reliability, we conduct a human evaluation by randomly sampling 500 papers and comparing human annotations with DeepSeek predictions. Results show strong agreement, with a Fleiss' kappa of 0.76 among annotators and a Cohen's kappa of 0.72 between human and model labels. DeepSeek also achieves 0.85 accuracy and 0.82 macro-F1, indicating near human-level consistency. Detailed setup and analysis are provided in App. G.1.

We find that annual publication volume is negatively correlated with the absolute value of QQE: eight of the nine fields show statistically significant correlations ($p \leq 0.05$). The only exception is the **AI Foundations & Theory** field, where the correlation is weaker and not statistically significant ($p > 0.05$), possibly indicating that paper growth in foundational research does not dilute scholarly impact. Our results demonstrate that the decline in absolute QQE is a robust cross-field phenomenon, suggesting diminishing marginal returns to scholarly impact as publication volumes expand in these conferences (c.f. App. G.2).

6 Conclusion

This work presents a ten-year empirical analysis of seven major NLP and AI conferences (2014–2024) using a four-dimensional bibliometric framework and the proposed QQE. The results reveal clear divergence between conference scale, citation impact, and growth efficiency: NeurIPS and AAAI dominate in publication volume, ICLR achieves the strongest normalized influence, ACL maintains the healthiest citation structure, and NAACL exhibits the highest impact density despite its smaller scale. Most venues remain below unit elasticity, indicating that publication growth generally outpaces citation growth.

7 Limitations

Our study has several limitations. First, although we aggregate data from multiple high-quality sources (ACL Anthology, Semantic Scholar, and OpenAlex), citation records may still contain minor inconsistencies due to update delays or incomplete indexing. These issues are unlikely to affect cross-venue comparisons but may influence fine-grained year-to-year values. Second, our analysis focuses exclusively on long papers; short papers, findings papers, and workshop papers are excluded due to inconsistent archival standards across venues. While this choice improves comparability, it may omit certain forms of early-stage or exploratory research. Third, citation-based indicators inherently lag behind real-time research influence and may partially reflect field-specific citation practices or short-term topic popularity rather than long-term scientific contribution. Finally, while our framework captures aggregate and distributional patterns, it does not explicitly model venue-specific policy changes (e.g., review procedures, submission formats, or acceptance strategies), which may also shape conference trajectories.

These limitations suggest several directions for future work, including incorporating acceptance-rate dynamics, policy changes, and field-normalized impact measures to achieve a more fine-grained understanding of conference evolution. Citations are an imperfect proxy for scholarly value and do not directly capture review rigor, originality, or ethical integrity. Our analysis is therefore limited to citation-based influence rather than holistic conference quality.

Potential risks. A potential risk is that reliance on citation-based indicators may overemphasize short-term visibility and popular research topics, potentially underrepresenting slower-burning or practice-oriented contributions.

References

Luca Bedogni, Giacomo Cabri, Riccardo Martoglia, and Francesco Poggi. 2023. Does the venue of scientific conferences leverage their impact? a large scale study on computer science conferences. *Library Hi Tech*, 41(2):501–515.

John PA Ioannidis. 2006. Concentration of the most-cited papers in the scientific literature: analysis of journal ecosystems. *PloS one*, 1(1):e5.

Sehrish Iqbal, Saeed-Ul Hassan, Naif Radi Aljohani, Salem Alelyani, Raheel Nawaz, and Lutz Bornmann. 2021. A decade of in-text citation analysis based on natural language processing and machine learning techniques: an overview of empirical studies. *Scientometrics*, 126(8):6551–6599.

Lun-Wei Ku, André FT Martins, and Vivek Srikumar. 2024. Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Ilia Kuznetsov, Iryna Gurevych, Yusuke Miyao, Yang Gao, and Steffen Eger. 2019. Does my rebuttal matter? insights from a major. In *Proceedings of the 2019 Conference of the North*. Association for Computational Linguistics (ACL).

Orestis-Stavros Loizides and Polychronis Koutsakis. 2017. On evaluating the quality of a computer science/computer engineering conference. *Journal of Informetrics*, 11(2):541–552.

Lokman I Meho. 2019. Using scopus's citescore for assessing the quality of computer science conferences. *Journal of Informetrics*, 13(1):419–433.

Saif Mohammad. 2020. Examining citations of natural language processing literature. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5199–5209.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Anna Rogers and Isabelle Augenstein. 2020. What can we do to improve peer review in NLP? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262, Online. Association for Computational Linguistics.

Giuseppe Russo, Manoel Horta Ribeiro, Tim Ruben Davidson, Veniamin Veselovsky, and Robert West. 2025. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–28.

Shiyun Wang, Yaxue Ma, Jin Mao, Yun Bai, Zhentao Liang, and Gang Li. 2023. Quantifying scientific breakthroughs by a novel disruption indicator based on knowledge entities. *Journal of the Association for Information Science and Technology*, 74(2):150–167.

Xue Wang, Xuemei Yang, Jian Du, Xuwen Wang, Jiao Li, and Xiaoli Tang. 2021. A deep learning approach for identifying biomedical breakthrough discoveries using context analysis. *Scientometrics*, 126(7):5531–5549.

Qiyao Wei, Samuel Holt, Jing Yang, Markus Wulfmeier, and Mihaela van der Schaar. 2025. The ai imperative: Scaling high-quality peer review in machine learning. *arXiv preprint arXiv:2506.08134*.

Ziming Zhuang, Ergin Elmacioglu, Dongwon Lee, and C Lee Giles. 2007. Measuring conference quality by mining program committee characteristics. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, pages 225–234.

A Data Collection and Processing

To ensure the comprehensiveness and accuracy of our longitudinal analysis, we curated a large-scale dataset of publications and their corresponding citation metrics from seven premier AI and NLP conferences. We restrict our analysis to main-track long papers. Findings papers, short papers, and workshop papers are excluded to ensure cross-year and cross-venue comparability. The data collection process was conducted in two primary stages: metadata extraction and citation retrieval.

Temporal Scope: 2014–2024. Our analysis covers 2014–2024. Since ACL submission volumes were essentially stable from 2009–2014, e.g., 634 (2011), 571 (2012), 664 (2013), and 572 (2014), but began accelerating sharply in 2015 (692), 2014 provides a natural pre-expansion baseline. The year 2024 is chosen as the endpoint to ensure complete and comparable proceedings across all venues[7].

In the first stage, we extracted the official metadata for all published papers, specifically focusing on publication years and paper titles. The source repositories were selected based on the official hosting platforms of each venue:

- ACL, EMNLP, and NAACL: Data were retrieved from the ACL Anthology, the comprehensive digital archive for conference proceedings in computational linguistics.
- IJCAI: Data were collected from the IJCAI Proceedings archive.
- AAAI: Metadata were sourced from the AAAI Digital Library.
- NeurIPS: Data were extracted from the NeurIPS Proceedings.
- ICLR: Given its open-reviewing model, data were retrieved from OpenReview.net.

Due to the absence of an official Google Scholar API, we rely on Semantic Scholar and OpenAlex to obtain high-quality citation data. In our pipeline, paper titles were first harvested directly from the official websites of each conference and subsequently matched with external databases using exact string matching—augmented by venue and year constraints—to minimize ambiguity. Semantic Scholar served as the primary data source due to its superior coverage and its provision of time-sliced citation records essential for calculating dynamic metrics like Quality-Quantity Elasticity (QQE). This approach achieved a 99.3% successful match rate, with the remaining 0.7% (approximately 430 entries) supplemented via OpenAlex. Together, these sources provided granular, programmatically retrieved metadata, including annual and total citation counts, necessary for our longitudinal analysis.Due to indexing limitations in the source databases, the number of collected papers is slightly lower than the total number of published papers. However, the missing data accounts for less than 1% of the total volume and has a negligible impact on the overall analysis.

[7]Up to January 2026.

Conference	No. of Papers	Citations
ACL	4,965	488,499
EMNLP	7,106	573,104
NAACL	2,540	414,779
AAAI	12,786	713,125
IJCAI	6,617	245,391
ICLR	7,691	1,771,388
NeurIPS	18,701	2,375,612
Total	**60,406**	**6,581,898**

Table 5: Paper and citation statistics.

The final dataset underwent a cleaning process to resolve potential title mismatches and to ensure that only main conference research papers (excluding workshops and tutorials) were included in the longitudinal study.

B Raw Dataset Structure and Schema

The initial dataset constructed for this study comprises raw records for publications from the seven major conferences spanning 2014 to 2024. Each record represents an individual paper and includes the following fields, designed to support multi-dimensional and longitudinal analysis of academic impact:

- **Core Metadata Fields**:
  - *paperId*: The unique identifier assigned to each publication (sourced from Semantic Scholar).
  - *title*: The full official title of the paper.
  - *year*: The formal year of publication.
  - *ai_category*: The specific sub-field or category classification of the paper within the AI domain.

- *notes*: Supplementary information regarding paper types or special session inclusions.

- **Aggregated Citation Fields**:
  - *citationCount*: The cumulative citation count at the time of data collection.
  - *top_conf_citations*: The total number of citations received from the peer top-tier conferences defined in this study.
  - *top_journal_citations*: The total number of citations received from high-impact top-tier journals.

- **Annual Time-Series Citation Fields**:
  - *citations_2014* to *citations_2024*: A series of annual fields recording the incremental citations received by the paper within each specific calendar year from 2014 to 2024. This granular time-series data serves as the fundamental basis for calculating dynamic metrics such as Quality-Quantity Elasticity (QQE), annual highly-cited ratios, and Gini coefficients.

Table 5 summarizes the total volume of data collected from the various meetings.

C List of Identified Top-Tier Venues

The following lists define the venues used to calculate the top-tier citation metrics . These categories ensure that the cross-disciplinary impact of the research is accurately captured.

Top Conferences

- **Artificial Intelligence & ML**: AAAI, IJCAI, ICML, NeurIPS (including NIPS).
- **Natural Language Processing**: ACL, EMNLP.
- **Computer Vision**: CVPR, ICCV, ECCV.
- **Data Mining & Web**: KDD, SIGIR, WWW.
- **Software Engineering**: ICSE, FSE, ASE.
- **Systems & Networking**: SOSP, OSDI, NSDI, SIGCOMM, INFOCOM, MOBICOM.
- **Security**: CCS, S&P (Oakland), USENIX Security.
- **Databases**: SIGMOD, VLDB, ICDE.
- **Theory**: STOC, FOCS, SODA.
- **Computer Graphics**: SIGGRAPH.
- **Human-Computer Interaction**: CHI.

Top Journals

- **General Science**: Nature, Science.

D Analytical Method

We adopt a combined cross-conference and longitudinal analytical design.

(1) Cross-Conference Comparison. We compute the ten-year averages of fourteen bibliometric indicators for ACL, EMNLP, NAACL, AAAI, IJCAI, ICLR, and NeurIPS. This provides a macro-level snapshot of each venue's overall scale, citation performance, impact distribution, citation composition, and growth efficiency (QQE). The analysis characterizes the long-term static landscape of major NLP/AI conferences and their relative positions.

(2) Longitudinal Evolution Analysis. We then track year-to-year changes using a one-year computation window, focusing on six core dynamic indicators: average and median citations (N-AC, N-MC), the proportions of highly cited and zero-cited papers (HCR, ZCR), QQE, milestone influence (MII), and within-NLP comparisons among ACL, EMNLP, and NAACL. This longitudinal component examines the temporal evolution of citation levels, impact distribution, and growth efficiency across the seven venues over the past decade, revealing the shifts in their developmental trajectories.

E Experimental Setup and Hyperparameters

Our study does not involve model training or learning-based optimization; therefore, no conventional hyperparameters are required. All experiments are conducted within a descriptive and longitudinal bibliometric analysis framework. The experimental setup consists of (1) selecting seven major NLP/AI conferences, (2) restricting analysis to main-track long papers, (3) defining a fixed temporal window (2014–2024), and (4) computing citation-based indicators using consistent rules across venues and years.

Key analytical settings, including the one-year citation window, the definition of highly cited papers, the construction of milestone indicators, and the computation of Quality–Quantity Elasticity (QQE), are fixed a priori and applied uniformly throughout all experiments. No parameter tuning or post-hoc threshold adjustment is performed. No stochastic experiments are involved, so error bars are not applicable. This design ensures that all reported results are fully reproducible and not sensitive to hyperparameter choices.

F Distributional Dynamics and Milestone-Level Papers

In this appendix, we further address two additional research questions (RQ6–RQ7). We first analyze how citation distributions evolve across conferences using per-year distributional visualizations , and then compare milestone-level papers across venues.

RQ6. Is the Change Driven by Head Papers, Tail Papers, or the Full Distribution?

The per-year violin plots (Fig. 10) utilize a logarithmic scale to compress multi-order citation differences, revealing a clear structural divergence between venues. ML conferences (ICLR, NeurIPS) exhibit elevated centers and stable heads, indicating consistently high visibility and tight alignment with community attention. Conversely, NLP venues (notably EMNLP and NAACL) maintain stable centers but show widening tails after 2022, reflecting increasing heterogeneity and a growing share of minimally cited papers during rapid expansion. ICLR experienced the most dramatic shift, moving from an exceptionally long head during its 2014–2017 breakthrough era to post-2020 contraction, while IJCAI shows recent head compression. Overall, these patterns demonstrate that citation evolution is driven by differentiated movements across the distribution's head, body, and tail rather than uniform shifts.

RQ7. Which Venues Produce the Most Milestone-Level Papers in the LLM Era?

Fig. 11 shows substantial variation across conferences in their production of milestone-level research. Although NeurIPS has the highest absolute number of $\geq$ 1000-citation papers (283), it also has the largest overall volume, resulting in a proportion of 1.51%. In contrast, ICLR, despite having far fewer total papers, achieves the highest proportion of milestone papers at 2.94%. Within the NLP community, ACL leads with a proportion of 1.09%, while NAACL (0.75%) and EMNLP (0.73%) exhibit similar levels, indicating that high-impact work remains a significant component of these language-focused venues. AAAI and IJCAI show much lower proportions (0.48% and 0.17%, respectively), suggesting that they produce fewer ultra-high-impact papers relative to their total volume in the LLM era.

G QQE Analysis Across Different Domains

G.1 Validation of DeepSeek-based Field Classification

To ensure the reliability of the DeepSeek-based field classification in our study, we perform a systematic human evaluation, and this section describes its sampling strategy, annotation protocol, evaluation metrics, and results. We randomly sample 500 papers from our full dataset of 60,406 papers published between 2014 and 2024 at seven major conferences (ACL, EMNLP, NAACL, AAAI, IJCAI, ICLR, NeurIPS), using stratified sampling to ensure roughly 70 papers per conference and uniform coverage across years. This design ensures that the evaluation covers diverse venues and temporal phases of research development. We employ **three independent human annotators** with backgrounds in natural language processing and machine learning.

Each paper is labeled according to its title and abstract, which supply sufficient information to determine its primary research field, with annotators instructed to assign each paper to one of the nine predefined research categories used in the main analysis; to ensure annotation consistency, annotators adhere to a shared guideline containing category definitions and examples, conduct all annotations independently in a blind setting, and are prohibited from discussion or reconciliation during the initial labeling phase.

We evaluate both inter-annotator agreement and the consistency between human annotations and DeepSeek predictions.

(1) Inter-Annotator Agreement We measure agreement among human annotators using Fleiss' kappa:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

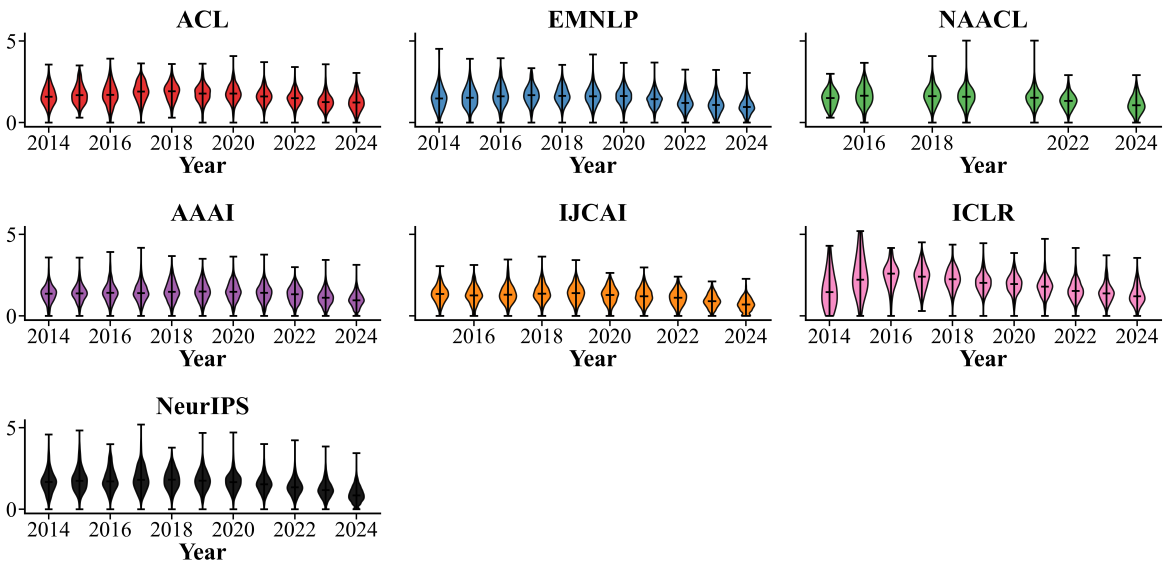

Figure 10: Evolution of citation distribution in major NLP/AI conferences.

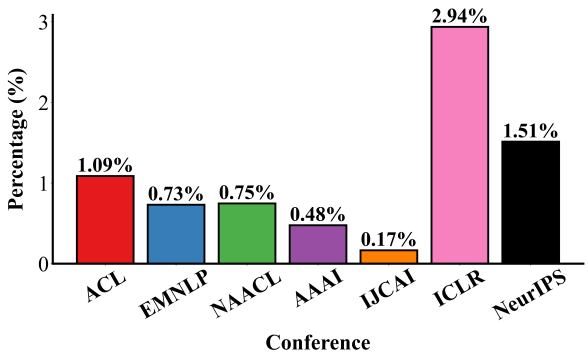

Figure 11: Percentage of Milestone Papers.

where $\bar{P}$ denotes the observed agreement and $\bar{P}_e$ denotes the expected agreement by chance.

(2) Human–DeepSeek Agreement We compute Cohen's kappa to measure agreement between DeepSeek predictions and human annotations:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed agreement and p_e is the expected agreement under random labeling.

(3) Classification Performance We further report standard classification metrics:

Accuracy:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Samples}}$$

Macro-F1 Score:

$$\text{Macro-F1} = \frac{1}{C}\sum_{i=1}^{C} F1_i$$

where C is the number of categories and $F1_i$ is the F1-score of class i.

Table 6 summarizes the evaluation results.

Metric	Value
Fleiss' Kappa (Human–Human)	0.76
Cohen's Kappa (Human–DeepSeek)	0.72
Accuracy	0.85
Macro-F1	0.82

Table 6: Agreement between human annotators and DeepSeek classification.

The inter-annotator agreement (Fleiss' $\kappa = 0.76$) indicates substantial agreement among human annotators, confirming that the category definitions are clear and consistent. DeepSeek achieves a Cohen's κ of 0.72 with respect to human annotations, demonstrating strong alignment with human judgment. In addition, the model attains an accuracy of 0.85 and a macro-F1 score of 0.82, indicating robust and balanced performance across categories. Notably, the agreement between DeepSeek and human annotators is close to the level of human–human agreement, suggesting that the classification performance is approaching the upper bound imposed by inherent ambiguity in the task. Overall, these results validate the use of DeepSeek

as a reliable and scalable proxy for field classification in large-scale bibliometric analysis. We note that disagreements between DeepSeek and human annotators primarily arise in semantically adjacent categories (e.g., Representation Learning vs. General Machine Learning), where boundaries are inherently fuzzy. This suggests that most errors are due to task ambiguity rather than systematic bias in the model.

G.2 Field-Level Decomposition of Publication Growth and QQE Dynamics

Our main experiment considers all papers in aggregate and reveals a systematic decline in the absolute value of QQE as publication volume expands. However, these aggregate trends may potentially be influenced by changes in the mix of fields, as different research fields exhibit different growth rates and citation dynamics.

To separate these effects, we conduct a field-level analysis by first categorizing papers into major research areas using DeepSeek-V3, with the annual publication counts and the absolute values of QQE shown in Tab. 10. We then examined the relationship between annual publication counts and the absolute values of QQE over the period 2015–2024 within each field. Specifically, for each field, we compute the Spearman rank correlation between annual publication volume and the absolute values of QQE, allowing us to assess whether the observed negative association reflects a field-specific phenomenon or a more general structural pattern.

As shown in Tab. 4, we observe predominantly strong and statistically significant negative correlations across most research fields. In particular, Representation Learning & Optimization, Machine Learning (incl. Deep Learning), Interpretability, Fairness & Applications, Reinforcement Learning & Decision Making, and Natural Language Processing exhibit Spearman correlation coefficients below -0.85 ($p \leq 0.01$), indicating that sustained increases in publication volume are systematically associated with declining absolute QQE. These results suggest that as these fields expand, marginal increases in publication output yield diminishing returns in scholarly impact.

An exception is AI Foundations & Theory area, where the correlation is weaker and statistically insignificant ($p > 0.05$), possibly reflecting the fact that publication growth in foundational research does not necessarily dilute research impact. Overall, the persistence of negative correlations within individual fields demonstrates that the decline in the absolute values of QQE represents a robust, cross-fields phenomenon in these ML, AI, and NLP conferences. While spillover effects from accumulated knowledge remain present, their marginal contribution to scholarly impact appears to weaken as research fields scale up.

H Statistical Analysis of QQE

H.1 Significance of QQE Correlations

Table 3 reports the p-values of Pearson and Spearman correlations between QQE and citation-based indicators. The results show that significant correlations are concentrated in a few venues rather than uniformly observed across all conferences.

For ICLR, QQE is significantly correlated with normalized average citation ($p = 0.0013$), normalized median citation ($p = 0.0008$), normalized median citation under Spearman correlation ($p = 0.0376$), and raw average citation ($p = 0.0476$). This indicates that variations in ICLR's QQE are systematically associated with changes in citation impact. However, this does not imply super-unit elasticity, since ICLR's overall QQE remains below 1. Rather, it shows that QQE and citation impact move together within ICLR, even though impact growth does not fully keep pace with its rapid scale expansion.

AAAI presents a different pattern. Its Spearman correlations are significant for normalized average citation ($p = 0.0029$), normalized median citation ($p = 0.0011$), normalized H-index ($p = 0.0029$), and raw average citation ($p = 0.0029$), while the corresponding correlation coefficients are negative in Fig. 3. This suggests that higher QQE periods in AAAI are not aligned with stronger static citation performance, reflecting a decoupling between growth efficiency and citation magnitude.

For ACL, EMNLP, IJCAI, and NAACL, most correlations are not statistically significant. NeurIPS only shows a significant Pearson correlation with normalized median citation ($p = 0.0460$). These results indicate that QQE is not consistently reducible to conventional citation indicators. Instead, its relationship with citation metrics varies by venue, supporting its role as a complementary measure of scale–impact efficiency.

H.2 Robustness to Outliers

To test whether the results are driven by a small number of highly cited papers, we remove the

Conf.	Avg(N)-P	Avg(N)-S	Med(N)-P	Med(N)-S	H(N)-P	H(N)-S	Avg(R)-P	Avg(R)-S
AAAI	0.063	**0.0029**	**0.0339**	**0.0011**	0.121	**0.0029**	0.087	**0.0029**
ACL	0.245	0.467	0.189	0.511	0.177	0.365	0.211	0.365
EMNLP	0.580	0.446	0.960	0.467	0.804	0.603	0.806	0.603
ICLR	**0.0013**	0.701	**0.0008**	**0.0376**	0.159	0.603	**0.0476**	0.533
IJCAI	0.925	0.932	0.711	0.932	0.951	0.932	0.802	0.932
NAACL	0.864	0.667	0.842	0.667	0.979	0.667	0.975	0.667
NeurIPS	0.181	0.446	**0.0460**	0.385	0.328	0.467	0.273	0.467

Table 7: P-values of correlations between QQE and normalized citations. Bold values indicate statistical significance ($p < 0.05$). (P: Pearson, S: Spearman, N: Normalized, R: Raw)

Conference	Median (Full)	Trimmed Mean	Δ
ACL	0.348	0.719	+0.371
EMNLP	0.421	0.612	+0.191
NAACL	0.389	0.655	+0.266
AAAI	0.260	0.495	+0.235
IJCAI	0.274	0.522	+0.248
ICLR	0.724	1.683	+0.959
NeurIPS	0.512	0.846	+0.334

Table 8: Robustness analysis before and after removing the top 1% highest-cited papers. Δ denotes the difference between trimmed mean and original median.

Comparison	Median Diff (%)	p-value
ICLR vs AAAI	178.3%	< 0.001
ACL vs AAAI	33.8%	< 0.001
ICLR vs ACL	108.5%	< 0.001
NeurIPS vs AAAI	96.7%	< 0.001
EMNLP vs AAAI	62.1%	< 0.001
IJCAI vs AAAI	5.4%	0.213
NAACL vs AAAI	48.9%	< 0.001

Table 9: Pairwise comparisons using Mann–Whitney U-test with Bonferroni correction.

top 1% highest-cited papers and recompute citation statistics. Table 8 shows that the main cross-conference ordering remains stable after trimming.

ICLR still has the highest trimmed mean citation impact (1.683), followed by NeurIPS (0.846), ACL (0.719), NAACL (0.655), EMNLP (0.612), IJCAI (0.522), and AAAI (0.495). Thus, ICLR and NeurIPS remain the strongest venues in trimmed citation impact, while AAAI remains the lowest. This indicates that the observed impact hierarchy is not solely caused by a small number of extreme citation outliers.

Notably, the robustness results should be interpreted as evidence for the stability of citation-impact patterns, rather than as a direct validation of QQE values. Together with Fig. 3b, however, they show that the correlation structure between QQE and citation indicators remains broadly consistent after removing top-cited outliers.

H.3 Distributional Significance Tests

We further test whether citation distributions differ significantly across conferences. The Kruskal–Wallis H-test shows a strong cross-venue difference ($H = 5322.07$, $p < 0.001$), confirming that the conferences should not be treated as samples from the same citation distribution.

Bonferroni-corrected Mann–Whitney U-tests further identify the main pairwise differences. Compared with AAAI, ICLR shows a much higher median citation performance ($+178.3\%$, $p < 0.001$), followed by NeurIPS ($+96.7\%$, $p < 0.001$), EMNLP ($+62.1\%$, $p < 0.001$), NAACL ($+48.9\%$, $p < 0.001$), and ACL ($+33.8\%$, $p < 0.001$). In contrast, IJCAI does not differ significantly from AAAI ($+5.4\%$, $p = 0.213$). ICLR also significantly outperforms ACL ($+108.5\%$, $p < 0.001$).

These tests validate the statistical separability of conference-level citation distributions. They provide a necessary empirical basis for cross-venue comparison, while QQE further examines whether such citation impact grows proportionally with conference scale.

H.4 Validity and Robustness of QQE

We evaluate QQE in terms of complementarity, discriminability, robustness, and statistical validity.

Complementarity. Fig. 3 shows that QQE is not reducible to conventional citation indicators. The correlations vary substantially across venues and metrics. ICLR shows strong positive correlations with normalized citation indicators ($r = 0.86$ for Avg-SNIP and $r = 0.88$ for Med-SNIP), while AAAI shows strong negative correlations (down to $r = -0.87$). ACL and IJCAI remain weakly correlated with most indicators, and EMNLP stays close to zero. The significance tests in App. H further support this pattern: only part of these correla-

tions are statistically significant, mainly for ICLR and AAAI, rather than uniformly across all venues. This indicates that QQE is related to citation impact, but does not simply replicate citation magnitude.

Discriminability. QQE separates growth efficiency from raw scale and static influence. For example, NeurIPS has the largest publication volume ($PC = 18{,}701$) and the highest H-index ($H = 564$), but its QQE is below unit elasticity (0.91). ICLR has the strongest normalized influence ($Norm\text{-}AC = 2.27$, $Norm\text{-}MC = 0.72$, $HCR = 26.90$), yet has the lowest QQE (0.82). In contrast, ACL and NAACL are much smaller in scale but exceed unit elasticity (1.20 and 1.09). Thus, QQE captures whether impact growth keeps pace with conference expansion, not how large or highly cited a venue is.

Robustness to Outliers. To test whether QQE-related patterns are driven by a few highly cited papers, we remove the top 1% citation outliers and recompute the correlations. As shown in Fig. 3b, the main structure remains stable: ICLR remains strongly positive ($r = 0.87$ for Avg-SNIP and $r = 0.89$ for Med-SNIP), AAAI remains negative, and IJCAI remains close to zero. The correlations are generally weakened after trimming, especially for AAAI, but the signs and venue-level patterns are largely preserved. The robustness table in App. H also shows that the relative impact ordering remains stable after outlier removal: ICLR still has the highest trimmed mean (1.683), followed by NeurIPS (0.846), ACL (0.719), NAACL (0.655), EMNLP (0.612), IJCAI (0.522), and AAAI (0.495). Therefore, the observed patterns are not solely caused by a small number of extremely cited papers.

Statistical Validity. The distributional tests in App. H confirm that cross-venue impact differences are statistically meaningful. The Kruskal–Wallis test shows significant differences among conferences ($H = 5322.07$, $p < 0.001$). Bonferroni-corrected Mann–Whitney tests further show that ICLR, ACL, NeurIPS, EMNLP, and NAACL significantly outperform AAAI in median citation performance ($p < 0.001$), while IJCAI does not differ significantly from AAAI ($p = 0.213$). These results justify treating venues as statistically distinct citation populations rather than assuming that observed differences are random fluctuations.

Taken together, QQE passes three key checks. It is non-redundant because it is not consistently aligned with existing citation indicators; it is discriminative because it separates scale, impact level, and growth efficiency; and it is robust because the main patterns remain after removing top-cited outliers. Therefore, QQE should be interpreted as a complementary metric for scale–impact efficiency, rather than a replacement for conventional citation indicators.

I Statement on AI Assistance

ChatGPT was used for grammar checking and language polishing only. Additionally, in the supplementary analysis of Section 5, DeepSeek-V3 (a 671B parameter Mixture-of-Experts model) was employed to assist in the automated classification of paper topics. All AI-generated classifications were manually audited and verified by the authors. All core analyses, results, and interpretations were conducted by the authors, who maintain full responsibility for the contents of this paper.

Paper Count	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**	**2024**
Representation Learning & Optimization	202	319	385	457	574	787	840	953	1082	1273	1443
Multimodal Learning	23	52	68	87	168	200	203	271	357	553	1060
Probabilistic Statistics & Inference	176	281	282	292	330	432	501	532	480	655	712
Machine Learning (incl. Deep Learning)	131	163	229	330	499	630	760	922	960	1086	1357
AI Foundations & Theory	108	269	212	240	281	363	430	488	526	601	785
Computer Vision (CV)	77	100	164	207	392	473	643	779	900	1157	1798
Interpretability, Fairness & Applications	21	44	67	90	160	291	331	506	547	717	938
Reinforcement Learning & Decision Making	128	267	267	335	439	612	737	871	933	1052	1245
Natural Language Processing (NLP)	386	721	741	651	1286	1664	1614	2010	1859	2014	2770
Avg. of the Absolute Values of QQE	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**	**2024**
Representation Learning & Optimization	/	5.74	3.68	2.28	1.44	1.15	1.63	1.14	1.26	1.09	0.93
Multimodal Learning	/	8.20	2.57	1.53	1.07	1.33	1.72	1.08	1.22	1.04	0.83
Probabilistic Statistics & Inference	/	5.95	2.20	1.76	1.86	1.26	1.18	1.22	2.40	0.74	1.02
Machine Learning (incl. Deep Learning)	/	7.84	2.84	2.63	1.72	1.30	1.42	1.03	1.16	1.18	0.93
AI Foundations & Theory	/	2.64	2.94	1.09	1.27	1.05	2.21	1.12	3.83	1.04	0.94
Computer Vision (CV)	/	9.96	2.03	2.39	1.06	1.69	3.64	1.27	1.30	1.01	1.21
Interpretability, Fairness & Applications	/	3.78	7.30	2.00	1.58	1.05	1.90	1.19	1.24	0.93	0.96
Reinforcement Learning & Decision Making	/	16.19	2.27	1.90	1.69	1.17	1.56	1.47	1.02	1.16	0.74
Natural Language Processing (NLP)	/	7.55	3.21	1.55	1.40	1.40	1.22	1.25	1.23	1.04	0.98

Table 10: Historical trends in Paper Counts and the absolute values of QQE across different categories.

Conference	Year	N_t	C_t	P_t	g_t	QQE	$\ln(P_t)$
AAAI	2014	397	2437.41	-	-	-	-
AAAI	2015	533	5618.78	1.343	2.305	1.717	0.295
AAAI	2016	537	6203.25	1.008	1.104	1.096	0.007
AAAI	2017	632	6192.09	1.177	0.998	0.848	0.163
AAAI	2018	934	11019.40	1.478	1.780	1.204	0.391
AAAI	2019	1119	9929.34	1.198	0.901	0.752	0.181
AAAI	2020	1559	9614.09	1.393	0.968	0.695	0.332
AAAI	2021	1645	11014.09	1.055	1.146	1.086	0.054
AAAI	2022	1322	10062.15	0.804	0.914	-1.137	-0.219
AAAI	2023	1707	8438.68	1.291	0.839	0.650	0.256
AAAI	2024	2401	8092.07	1.407	0.959	0.682	0.341
ACL	2014	147	1008.76	-	-	-	-
ACL	2015	174	606.20	1.184	0.601	0.508	0.169
ACL	2016	230	2789.58	1.322	4.602	3.481	0.279
ACL	2017	194	1361.55	0.843	0.488	-0.579	-0.170
ACL	2018	256	3517.82	1.320	2.584	1.958	0.277
ACL	2019	446	2783.29	1.742	0.791	0.454	0.555
ACL	2020	570	4007.35	1.278	1.440	1.127	0.245
ACL	2021	570	4720.68	1.000	1.178	1.178	0.000
ACL	2022	603	4648.11	1.058	0.985	0.931	0.056
ACL	2023	911	4854.43	1.511	1.044	0.691	0.413
ACL	2024	864	4906.97	0.948	1.011	-1.066	-0.053
EMNLP	2014	227	3086.05	-	-	-	-
EMNLP	2015	313	2362.27	1.379	0.765	0.555	0.321
EMNLP	2016	265	1495.27	0.847	0.633	-0.748	-0.166
EMNLP	2017	324	2976.16	1.223	1.990	1.628	0.201
EMNLP	2018	549	2577.03	1.694	0.866	0.511	0.527
EMNLP	2019	682	2584.77	1.242	1.003	0.807	0.217
EMNLP	2020	752	3111.81	1.103	1.204	1.092	0.098
EMNLP	2021	848	5320.65	1.128	1.710	1.516	0.120
EMNLP	2022	829	4639.22	0.978	0.872	-0.892	-0.023
EMNLP	2023	1048	4736.14	1.264	1.021	0.808	0.234
EMNLP	2024	1269	5199.08	1.211	1.098	0.907	0.191
ICLR	2014	69	5171.77	-	-	-	-
ICLR	2015	31	3310.71	0.449	0.640	-1.425	-0.800
ICLR	2016	80	6216.55	2.581	1.878	0.728	0.948
ICLR	2017	198	10374.09	2.475	1.669	0.674	0.906
ICLR	2018	336	11015.90	1.697	1.062	0.626	0.529
ICLR	2019	502	16580.79	1.494	1.505	1.007	0.401
ICLR	2020	685	20540.54	1.365	1.239	0.908	0.311

Continued on next page

Conference	Year	N_t	C_t	P_t	g_t	QQE	$\ln(P_t)$
ICLR	2021	859	21058.43	1.254	1.025	0.818	0.226
ICLR	2022	1094	20581.45	1.274	0.977	0.767	0.242
ICLR	2023	1576	20208.56	1.441	0.982	0.682	0.365
ICLR	2024	2261	17373.73	1.435	0.860	0.599	0.361
IJCAI	2014	0	1419.32	-	-	-	-
IJCAI	2015	571	3815.45	-	2.688	-	-
IJCAI	2016	549	2510.69	0.961	0.658	-0.684	-0.039
IJCAI	2017	660	2471.36	1.202	0.984	0.819	0.184
IJCAI	2018	709	2041.97	1.074	0.826	0.769	0.072
IJCAI	2019	847	4426.92	1.195	2.168	1.815	0.178
IJCAI	2020	590	4474.04	0.697	1.011	-1.451	-0.362
IJCAI	2021	585	2269.59	0.992	0.507	-0.512	-0.009
IJCAI	2022	678	2657.19	1.159	1.171	1.010	0.148
IJCAI	2023	639	2332.97	0.942	0.878	-0.932	-0.059
IJCAI	2024	789	2274.36	1.235	0.975	0.790	0.211
NAACL	2014	0	2898.25	-	-	-	-
NAACL	2015	187	2165.01	-	0.747	-	-
NAACL	2016	182	2669.88	0.973	1.233	-1.267	-0.027
NAACL	2017	0	2652.26	0.000	0.993	-	-
NAACL	2018	332	2695.90	-	1.016	-	-
NAACL	2019	424	2749.39	1.277	1.020	0.799	0.245
NAACL	2020	0	1941.86	0.000	0.706	-	-
NAACL	2021	483	2003.46	-	1.032	-	-
NAACL	2022	444	2191.05	0.919	1.094	-1.190	-0.084
NAACL	2023	0	2219.87	0.000	1.013	-	-
NAACL	2024	488	1896.15	-	0.854	-	-
NeurIPS	2014	412	7350.43	-	-	-	-
NeurIPS	2015	407	4673.58	0.988	0.636	-0.644	-0.012
NeurIPS	2016	572	6286.78	1.405	1.345	0.957	0.340
NeurIPS	2017	681	11493.49	1.191	1.828	1.536	0.174
NeurIPS	2018	1013	11948.97	1.488	1.040	0.699	0.397
NeurIPS	2019	1432	12123.50	1.414	1.015	0.718	0.346
NeurIPS	2020	1903	14004.30	1.329	1.155	0.869	0.284
NeurIPS	2021	2342	14017.09	1.231	1.001	0.813	0.208
NeurIPS	2022	2674	15624.84	1.142	1.115	0.976	0.133
NeurIPS	2023	3229	17615.35	1.208	1.127	0.934	0.189
NeurIPS	2024	4036	20663.64	1.250	1.173	0.938	0.223

Table 12: Conference growth rates and QQE