Tracing the Evolution of Scientific Discourse: A Corpus-Driven Analysis of Scientific Cultural Heritage Publications in the Age of LLM

Simone Pio Barbagallo, Dario Allegra, and Filippo Stanco

Department of Mathematics and Computer Science, University of Catania, Italy
simone.barbagallo@phd.unict.it, {dario.allegra, filippo.stanco}@unict.it

Abstract. We present a corpus-driven study of diachronic language change in the *Journal on Computing and Cultural Heritage* (JOCCH), covering 590 articles (4.7M tokens, 2008–2026) across three macro-periods (Pre-Transformer, Transformer/pre-LLM, Post-LLM). Using Moving-Average Type-Token Ratio (MATTR), hapax ratio, Part-of-Speech (POS) distributions, log-likelihood keyness, and term frequency–inverse document frequency (TF-IDF) cosine similarity, we test three hypotheses about discourse shifts associated with Large Language Model (LLM)-centred methodologies. H1 (lexical standardisation) is not supported: the raw hapax decline is a Zipfian corpus-size artefact; once paper count is controlled, hapax legomena increase, and MATTR and cosine similarity likewise contradict homogenisation. H2 (methodological-linguistic coupling) is strongly confirmed: LLM-era terms explode after 2022, proper-noun density rises by 40%, and a gerund/past-participle shift signals a move from result-reporting to process-oriented phrasing. H3 (discursive reframing) is confirmed: the Generation verb group grows significantly, driven by *predict* and *generate*; *render* and *reconstruct* contract sharply. The overall picture is one of selective lexical absorption: scientific writing rapidly imports paradigm-specific vocabulary without converging on a uniform style.

Keywords: corpus linguistics, scientific discourse, diachronic analysis, NLP, large language models, digital humanities

1 Introduction and Related Work

The growing availability of full-text scientific archives makes it possible to study the evolution of academic writing at scale. The rapid diffusion of Large Language Models (LLMs) since late 2022, and the broader paradigm shift towards generative deep learning, have raised questions about possible standardisation, homogenisation and reshaping of scientific prose, whether through direct LLM-assisted writing, through the adoption of new methodological vocabularies, or through topical evolution driven by the new computational tools available to

researchers. Cultural Heritage Computing, a field at the intersection of engineering, the humanities and conservation science, offers a particularly rich test bed, bringing together communities with different writing traditions and a rapid adoption of new computational methods.

The present work analyses the articles published in the *Journal on Computing and Cultural Heritage* (JOCCH) as its empirical object and asks how the language of the discipline has changed over roughly two decades, with a specific focus on the pre- vs. post-LLM divide. This journal spans the full transition from classical photogrammetry and 3D reconstruction to deep-learning pipelines, providing an ideal longitudinal window on paradigm-driven language change. While diachronic corpus studies of scientific English exist for longer timescales [1], the specific window of LLM adoption has only recently received attention [2,3]. Three primary hypotheses are tested, each targeting a different level of linguistic organisation, lexical (H1), lexico-grammatical (H2), and verb-semantic (H3), so as to assess at which level LLM-centred methodological change leaves detectable traces in a discipline's prose:

H1 Lexical standardisation. LLM adoption is associated with a measurable reduction in lexical variety (lower MATTR, lower hapax ratio), consistent with homogenisation of scientific style.

H2 Methodological-linguistic coupling. Methodological shifts leave distinct lexical and grammatical signatures, observable through the diachronic distribution of Part-of-Speech (POS) tags and domain keywords.

H3 Discursive reframing. The dominant action verbs shift from *reconstruction* and *description* towards *generation*, *classification* and *prediction*.

Diachronic corpus studies of scientific English have established that disciplinary writing evolves measurably over time, with genre conventions, syntactic choices, and lexical profiles shifting across decades [1]. The rapid diffusion of LLMs since late 2022 has prompted a growing body of quantitative work on their impact on academic prose. Kobak et al. [2] identified LLM-processed biomedical abstracts through excess vocabulary analysis, estimating that at least 13.5% of 2024 publications were affected. Bao et al. [3] documented systematic lexical shifts in arXiv preprints following ChatGPT's launch, finding increased use of certain adjectives and decreased syntactic variability. Padmakumar and He [4] showed experimentally that writing with InstructGPT reduces lexical diversity across authors, suggesting a homogenising pressure at the inter-author level. Kendro and Maloney [5] further confirmed that ChatGPT-generated prose exhibits measurably distinct lexical diversity patterns compared with human-authored academic text. These studies share two limitations relative to the present work: they focus on biomedical or broad multidisciplinary preprint corpora, and they do not examine the lexico-grammatical and verb-semantic dimensions of discourse change simultaneously. Corpus-driven analyses of specialised scientific communities remain largely absent from the literature, particularly those at the intersection of engineering, the humanities, and conservation science. To the best of our knowledge, no prior study has investigated diachronic

language change in Computing and Cultural Heritage publications, nor has any corpus analysis examined how the LLM transition affects this interdisciplinary domain across lexical, grammatical, and verb-semantic levels simultaneously.

This paper makes three contributions: **(i)** a 590-article longitudinal corpus of Computing and Cultural Heritage publications (2008–2026, 4.7M tokens) as a reproducible empirical benchmark; **(ii)** the first quantitative assessment of the LLM transition's linguistic footprint in Cultural Heritage Computing across three levels of linguistic organisation; while the broad direction of discourse shift (H3) is predictable from the field's trajectory, the present work provides the first measurement of its rate and selectivity, identifying lemma-level displacement patterns and distinguishing paradigm-driven reframing from mere topical expansion; **(iii)** a reusable methodological framework combining MATTR, bootstrap resampling, log-likelihood keyness, and verb-semantic profiling applicable to comparable domain-specific corpora. Section 2 describes the corpus and methodology; Section 3 establishes the lexical and grammatical baseline; Section 4 reports hypothesis evaluation; Section 5 presents conclusions, limitations, and future directions.

2 Research Design

2.1 Corpus construction

The corpus consists of 590 full-text JOCCH articles retrieved from the ACM Digital Library on 29 April 2026 (42 articles were inaccessible due to access restrictions). A custom Python pipeline (PyMuPDF extraction, boilerplate removal via line-frequency thresholding, line-break dehyphenation, and alphabetic-token filtering) yields 4,745,952 alpha tokens across three macro-periods aligned with the LLM cut-off: *Pre-Transformer* (2008–2017, 167 papers), *Transformer/pre-LLM* (2018–2022, 186 papers), and *Post-LLM* (2023–2026, 237 papers). Throughout this paper, *Post-LLM* denotes the publication period following ChatGPT's public release (November 2022); it does not imply that papers in this period were written with LLM assistance, as the two phenomena are empirically distinct and cannot be disentangled from publication metadata alone. Alpha tokens are tokens composed exclusively of alphabetic characters. Numerical strings, punctuation, URLs, and mixed alphanumeric sequences (e.g. *ResNet-50*, *3D*) are discarded. The raw (unfiltered) token count is approximately 6.4M; alpha tokens represent the subset used for all linguistic analyses. Table 1 summarises the corpus. The post-LLM sub-corpus already contains more tokens than either predecessor despite covering only four publication years, reflecting both higher publication volume (237 papers in four years vs. 167 in ten) and a tendency towards longer articles. These unequal sub-corpus sizes are a structural feature of the archive rather than a sampling choice, and directly motivate the subsampling bootstrap used to control for corpus-size effects in the H1 analysis.

Table 1. Corpus overview. Tokens and Types from the Python pipeline (alpha tokens only). MATTR: sliding window $w = 50$ [6].

Sub-corpus	Years	Files	Tokens	Types	MATTR
Pre-Transformer	2008–2017	167	1,304,869	36,740	0.7819
Transformer / pre-LLM	2018–2022	186	1,506,008	39,765	0.7775
Post-LLM	2023–2026	237	1,935,075	46,138	0.7835
Full corpus	2008–2026	590	4,745,952	81,321	0.7812

2.2 Methodology and tooling

Following Tognini-Bonelli's distinction [7], the study combines a *corpus-based* component (testing explicit hypotheses against statistical evidence) and a *corpus-driven* component (exploring unexpected patterns). Lexical diversity is measured via Moving-Average Type-Token Ratio (MATTR; primary measure, robust to text length), Measure of Textual Lexical Diversity (MTLD) and normalised hapax ratio [6,8]. POS tagging uses the Natural Language Toolkit (NLTK) averaged-perceptron tagger (Penn Treebank tags, abbreviated for display: VB→VV, VBG→VVG, VBN→VVN, PRP→PP, NNP→NP). Log-likelihood (G^2) is used for keyness and frequency comparisons [9,10]; sign convention: positive LL = overuse in Pre-Transformer, negative LL = overuse in Post-LLM. All relative frequencies are reported per 10,000 tokens; raw counts are retained for rare items only. All scripts are deterministic (seed = 42). MATTR is preferred over the raw Type-Token Ratio (TTR) because TTR is inversely correlated with text length, making cross-document comparisons unreliable; the sliding window of fixed size w removes this dependency. The 1,000-resample bootstrap allows direct confidence interval estimation without parametric assumptions about the distribution of per-paper scores. Three filtering levels are used depending on the analysis: no filtering for lexical-diversity and POS measures; a standard 179-item NLTK list for frequency and collocation analyses; and a domain-augmented extension (adding academic boilerplate such as *paper, study, result*) for topic modelling.

3 Lexical and Grammatical Baseline

3.1 Lexical diversity

Table 2 reports the main diversity indicators per macro-period. The normalised hapax ratio declines monotonically (−14%, 119.63 → 102.97/10k), though confounded by Zipf's-law corpus-growth effects (word frequency is inversely proportional to rank, so hapax ratios fall mechanically as corpus size grows) and Optical Character Recognition (OCR) quality bias (see §4.1); MATTR, by contrast, is remarkably stable across all three periods (0.7819, 0.7775, 0.7835). MTLD dips in the Transformer period (68.65, its minimum) before recovering to 71.17 in Post-LLM, suggesting transient lexical consolidation that gradually resolved as the community diversified its LLM-adjacent vocabulary (Fig. 1).

Table 2. Lexical diversity per sub-corpus. TTR: raw types-to-tokens ratio; Standardised Type-Token Ratio (STTR): 100-token segments [11]; MATTR: $w = 50$; MTLD: $\tau = 0.72$ [8]. Hpx/10k: hapax legomena per 10,000 tokens.

Subcorpus	Tokens	Types	TTR	STTR	MATTR	MTLD	Hapax	Hpx/10k
Full Corpus	4,745,952	81,321	0.0171	0.6889	0.7812	70.35	36,310	76.51
Pre-Transformer	1,304,869	36,740	0.0282	0.6903	**0.7819**	**71.16**	15,610	**119.63**
Transformer / pre-LLM	1,506,008	39,765	0.0264	0.6847	**0.7775**	**68.65**	16,820	**111.69**
Post-LLM	1,935,075	46,138	0.0238	0.6918	**0.7835**	**71.17**	19,925	**102.97**

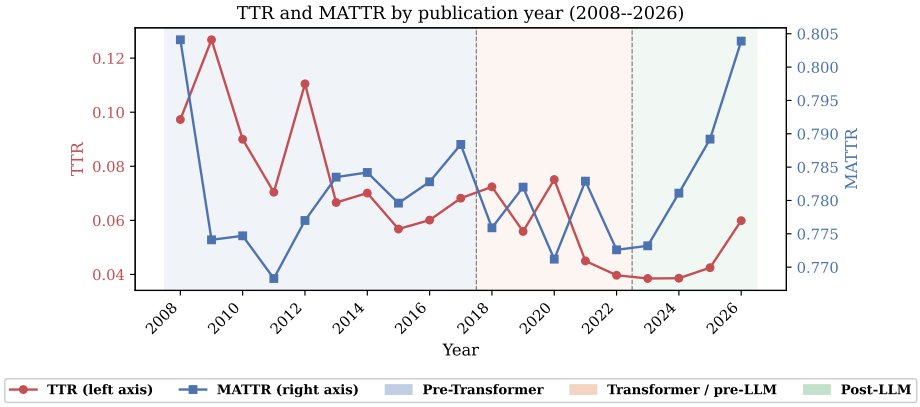

Fig. 1. TTR and MATTR by publication year (2008–2026). TTR falls as yearly volume grows; MATTR oscillates within a narrow band (0.768–0.804), validating it as the primary length-robust diversity measure.

3.2 POS distribution

Table 3 reports the ten most frequent POS tags in the full corpus with relative frequency, range across yearly sub-corpora and Deviation of Proportions (DP) [12]. The profile confirms the nominal, impersonal character of scientific writing in the corpus: singular common nouns (NN, 1877/10k) are the dominant category, followed by prepositions (IN, 1371/10k) and determiners (DT, 1027/10k). Personal pronouns (PP, 138/10k) are the rarest tag, consistent with impersonal academic prose. All tags appear in every one of the 19 yearly sub-corpora (Range = 19/19) with uniformly low DP values (< 0.08); the highest DP belongs to proper nouns (NP, 0.0762), reflecting how the mix of named entities shifts as the field's topical focus moves from traditional topics like photogrammetry and 3D reconstruction to deep-learning applications. The overall nominal and prepositional dominance, combined with low personal-pronoun frequency, confirms the genre as technical-scientific discourse with a strong preference for objectified knowledge claims and minimal authorial presence, a profile consistent across all 19 yearly sub-corpora.

Table 3. Most frequent POS tags in the full corpus (4,745,952 alpha tokens). Penn Treebank tagset [13], abbreviated: VV=verb base form, VVG=gerund/present participle, VVN=past participle, NP=proper noun, PP=personal pronoun. Rel. freq. per 10k tokens; Range = yearly sub-corpora in which the tag occurs (out of 19); DP = Deviation of Proportions [12] (0 = perfectly uniform, 1 = concentrated in one year).

POS	Description	Rel. freq. (/10k)	Range	DP
NN	noun, singular or mass	**1877.39**	19/19	0.0091
NNS	noun, plural	887.40	19/19	0.0201
NP	proper noun, singular	831.99	19/19	**0.0762**
JJ	adjective	929.25	19/19	0.0193
IN	preposition / sub. conjunction	**1371.05**	19/19	0.0090
DT	determiner	**1027.04**	19/19	0.0206
VV	verb, base form	292.31	19/19	0.0256
VVG	verb, gerund / pres. participle	253.79	19/19	0.0347
VVN	verb, past participle	353.76	19/19	0.0232
PP	personal pronoun	**138.29**	19/19	0.0375

Table 4. POS-tag relative frequency (/10k) by macro-period. ΔRel% = change from Pre-Transformer to Post-LLM. LL: positive = Pre-Transformer overuse; negative = Post-LLM overuse.

POS	**Description**	**Pre-T**	**Trans/pre-LLM**	**Post-LLM**	**ΔRel%**	**LL**
NN	noun, singular or mass	1932.91	1872.68	1844.55	−4.6%	396.57
NNS	noun, plural	847.66	880.56	918.67	+8.4%	−485.45
NP	proper noun, singular	702.24	742.44	**985.13**	**+40.3%**	−8000.74
JJ	adjective	929.02	911.88	942.59	+1.5%	−16.88
IN	preposition / sub. conjunction	1383.36	1394.87	1344.88	−2.8%	97.77
DT	determiner	1064.66	1059.35	977.80	−8.2%	641.13
VV	verb, base form	303.77	302.45	277.07	−8.8%	197.10
VVG	verb, gerund / pres. participle	236.86	247.52	**269.65**	**+13.8%**	−337.24
VVN	verb, past participle	373.93	358.45	336.95	−9.9%	311.27
PP	personal pronoun	145.25	147.79	126.51	−12.9%	204.69

Table 4 and Fig. 2 report diachronic shifts across macro-periods. All differences are highly significant; discussion focuses on effect size (ΔRel%). The most striking change is the rise in proper nouns (NP: +40.3%, 702 → 985/10k), driven by the proliferation of named deep-learning architectures (*CLIP*, *SAM*), datasets (*ImageNet*) and heritage sites. Verb morphology shifts consistently with H2 and H3: past participles (VVN, −9.9%) decrease while gerunds (VVG, +13.8%) increase, suggesting a move from result-reporting (*was computed*) to process-oriented phrasing (*training*, *generating*). Personal pronouns (PP, −12.9%) decline, consistent with a shift towards more impersonal constructions, whether through passive voice, nominalisations, or third-person framing, compatible with disciplinary writing conventions, multi-author consortia dynamics, or possibly LLM-assisted writing, though the data do not allow these mechanisms to be disentangled. Fig. 3 shows both the year-by-year and macro-period POS profiles through correspondence analysis; the three macro-periods separate along the NP and VVG axes (right panel).

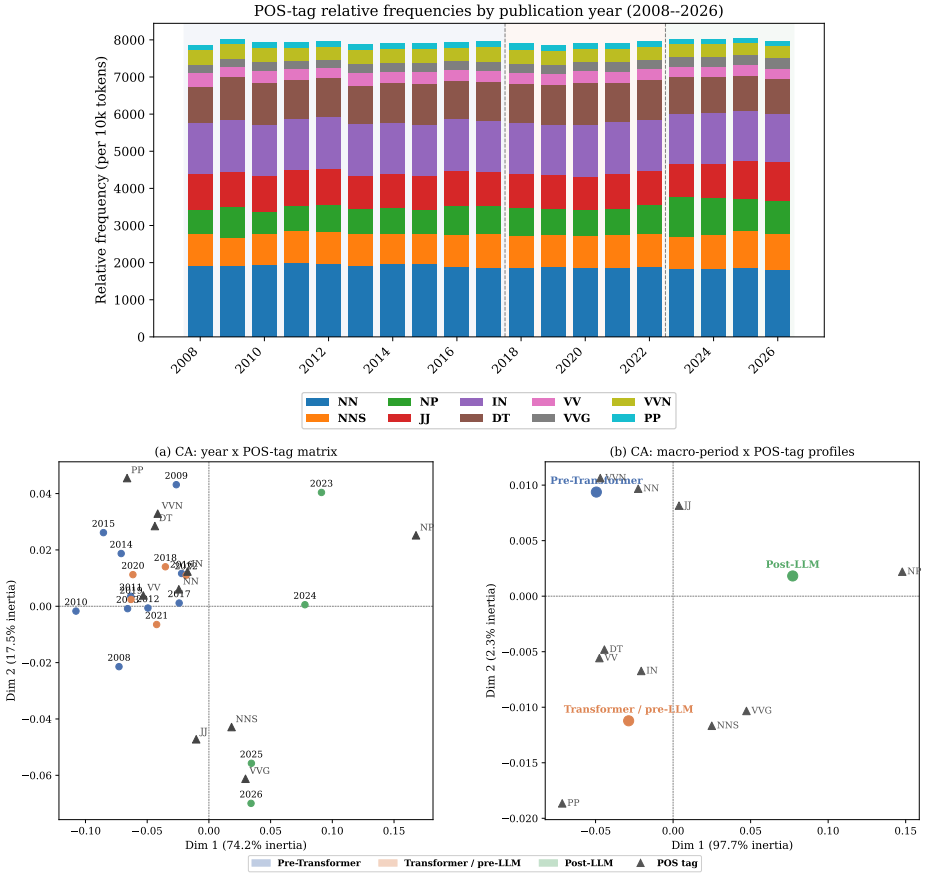

Fig. 3. Correspondence analysis of POS profiles. **Left**: year-by-year projection; years cluster into temporal bands, with 2023–2026 separating on the NP and VVG axes. **Right**: macro-period projection; the three periods separate along the first axis, driven by NP and VVG contrasts.

4 Hypothesis Evaluation

4.1 Lexical Standardisation (H1)

The hypothesis is motivated by converging evidence: Padmakumar and He [4] showed that writing with InstructGPT reduces lexical diversity between authors; Kobak et al. [2] estimated at least 13.5% of 2024 biomedical abstracts were LLM-processed; and Kendro and Maloney [5] found that text generated by ChatGPT exhibits lexical patterns measurably distinct from human-authored academic prose. We test H1 via three complementary indicators (Table 5): (1) per-paper MATTR with bootstrap confidence interval (CI), assessed at window $w = 100$ tokens (larger than the subcorpus $w = 50$ to reduce per-paper variance; the two variants are not directly comparable numerically); (2) normalised hapax ratio;

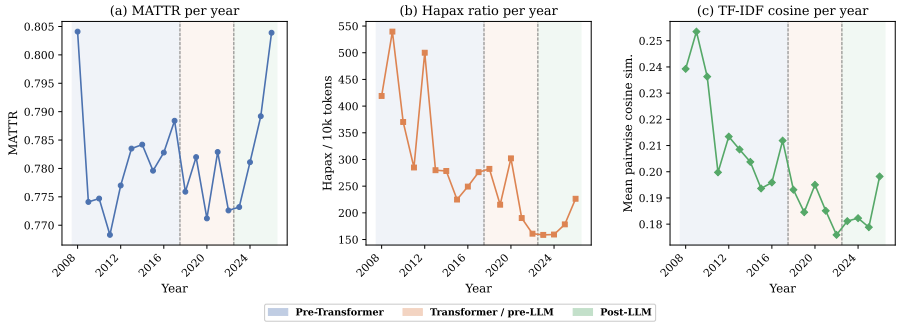

Fig. 4. Diachronic indicators of lexical standardisation (H1): hapax ratio, MATTR, and mean pairwise cosine similarity by year.

(3) mean pairwise term frequency–inverse document frequency (TF-IDF) cosine similarity between paper-level vectors.

Table 5. Lexical-diversity indicators by macro-period. $\text{MATTR}_{\text{sub}}$ and Hpx/10k from Table 2 (sub-corpus level, $w = 50$). $\text{MATTR}_{\text{paper}}$: mean per-paper MATTR ($w = 100$) with 95% bootstrap CI (1,000 resamples, seed = 42). Cosine CIs are upward-biased due to pseudo-self-pairs under replacement sampling; observed means are the primary estimator.

Macro-period	$\text{MATTR}_{\text{sub}}$	Hpx/10k	$\text{MATTR}_{\text{paper}}$ [95% CI]	Mean cosine sim. [95% CI]
Pre-Transformer	0.7819	119.63	**0.6901** [0.6860, 0.6943]	**0.1867** [0.1879, 0.1953]
Transformer / pre-LLM	0.7775	111.69	**0.6814** [0.6771, 0.6857]	0.1793 [0.1795, 0.1882]
Post-LLM	0.7835	102.97	0.6910 [0.6869, 0.6950]	**0.1774** [0.1772, 0.1844]

The evidence for H1 is mixed. The hapax ratio declines monotonically (−14%, 119.63 → 102.97/10k), but two confounders operate in the *same* direction as the prediction: (1) a Zipf's-law corpus-growth effect mechanically lowers hapax ratios as sub-corpora grow; (2) an OCR quality bias inflates Pre-Transformer hapax counts via spurious ligature fragments (*elds*, *lter*, *nding*) in pre-2018 PDFs, which are born-digital in the Post-LLM period. A sub-sampling bootstrap (1,000 resamples, 10 papers/year) confirms the confounding: when paper count is held constant, the hapax ratio *reverses*: 413.1 [340.7, 485.8] → 497.1 [413.5, 580.1] → 562.9 [473.5, 645.2]/10k, with non-overlapping CIs between Pre-T and Post-LLM. Per-paper MATTR is non-monotonic: the Transformer/pre-LLM period dips to 0.6814 [0.6771, 0.6857], while Pre-T (0.6901) and Post-LLM (0.6910) are statistically indistinguishable; sub-sampled values confirm this pattern. Mean pairwise TF-IDF cosine similarity moves *opposite* to H1 (0.1867 → 0.1793 → 0.1774), indicating lexico-topical *divergence* consistent with additive vocabulary expansion rather than homogenisation. All three indicators (Fig. 4) converge: the LLM transition has not homogenised writing at the level of individual documents.

4.2 Methodological-Linguistic Coupling (H2)

Seed keyword trends The ten seed terms were selected a priori to cover the three successive methodological strata documented in Cultural Heritage Computing literature: classical 3D acquisition (*photogrammetry*, *point cloud*), the discriminative deep-learning transition (*neural network*, *convolutional*), and the generative/LLM paradigm (*transformer*, *diffusion model*, *embedding*, *fine-tuning*, *prompt*). *Ontology* is included as a cross-cutting semantic-web term expected to remain stable across periods, serving as an implicit stability control. This selection is theory-driven; data-driven keyword extraction by log-likelihood keyness is reported separately in Table 8. A seed list of ten method-related terms is tracked diachronically. Table 6 reports absolute frequency (FA), relative frequency (FR/10k), coefficient of variation across 19 yearly sub-corpora (CV), and per-period relative frequencies.

Table 6. Seed method-related keywords by macro-period. CV = coefficient of variation across 19 yearly sub-corpora (high CV = bursty, uneven distribution). $\Delta\%$ relative to Pre-Transformer; n/a if Pre-T count is zero. LR = $\log_2$(Post-LLM / Pre-T): effect size (Hardie 2014; |LR| > 1 large, > 0.5 medium); n/a† if Pre-T≈0.

Term	FA	FR/10k	CV	Pre-T	Trans/pre-LLM	Post-LLM	$\Delta\%$	LR
photogrammetry	876	1.85	0.834	1.24	**2.51**	1.74	+40.2%	+0.52
point cloud	1196	2.52	0.699	**2.98**	2.56	2.20	−26.4%	−0.40
ontology	3453	7.29	0.965	7.43	4.07	**9.64**	+29.7%	+0.41
neural network	715	1.51	0.868	0.42	1.91	**1.92**	+356.8%	+2.22
convolutional	328	0.69	0.518	0.20	0.56	**1.12**	+452.5%	+2.48
transformer	236	0.50	0.757	0.00	0.16	**1.08**	n/a	n/a†
diffusion model	41	0.09	1.869	0.00	0.02	**0.19**	n/a	n/a†
embedding	454	0.96	0.990	0.13	0.68	**1.71**	+1196.5%	+3.70
fine-tuning	401	0.85	0.813	0.02	0.83	**1.41**	+8952.5%	+6.22
prompt	444	0.94	1.593	0.12	0.35	**1.92**	+1441.6%	+3.94

The LLM-era terms are effectively absent before 2018 and grow rapidly afterwards (Fig. 5): *fine-tuning* posts the highest relative growth (+8,952%), *prompt* the highest Post-LLM frequency (1.92/10k, equalling *neural network*). The CV column confirms that these terms are the most bursty (*diffusion model*: CV = 1.869; *prompt*: 1.593), consistent with rapid post-tipping-point diffusion rather than gradual growth. Counter-intuitively, *point cloud* declines (−26%), likely reflecting the shift from explicit 3D acquisition pipelines to image-based and end-to-end learned representations that reduce the need to name intermediate data structures; *ontology* shows a U-shaped trajectory (7.43 → 4.07 → 9.64/10k).

To test substitution dynamics, Pearson's r is computed on the yearly relative-frequency series ($n = 19$ years) for four candidate pairs (Table 7). No pair shows a significant negative correlation, ruling out clean substitution at annual granularity. The one significant result is a positive correlation between *convolutional*

and *diffusion model* ($r = 0.521$, $p < .05$): both grow together, consistent with co-diffusion of generative and discriminative deep-learning vocabularies. The absence of negative correlations confirms that vocabulary expansion has been predominantly *additive*, consistent with the lexico-topical divergence observed under H1.

Table 7. Pearson correlations between candidate substitution pairs (H2). $n = 19$ yearly sub-corpora; 95% CI via Fisher's z-transform (a variance-stabilising transformation for r). Positive r = co-occurrence; negative r = substitution.

Pair	r	95% CI	p-value
photogrammetry vs. neural network	0.216	[−0.264, 0.610]	n.s.
point cloud vs. embedding	−0.065	[−0.504, 0.401]	n.s.
ontology vs. transformer	−0.041	[−0.486, 0.421]	n.s.
convolutional vs. diffusion model	**0.521**	[0.088, 0.789]	$< .05$

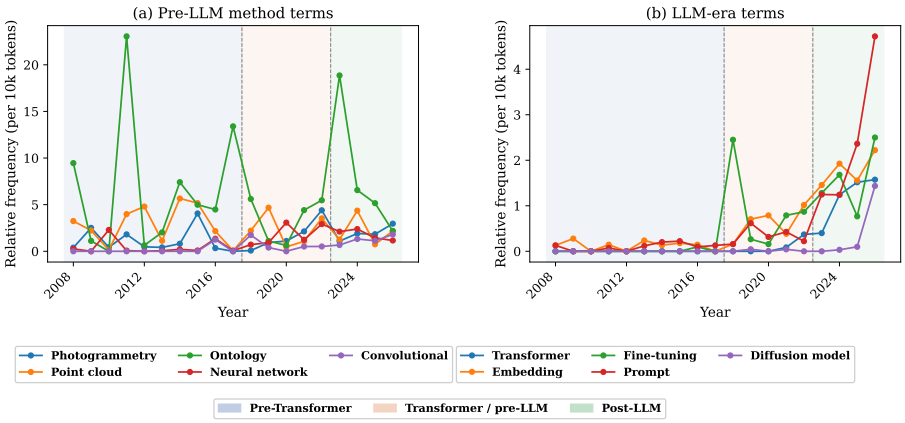

Fig. 5. Diachronic relative frequency (/10k) of the ten seed method-related keywords (H2). LLM-era terms (*prompt*, *fine-tuning*, *embedding*) are effectively flat until 2018–2020, then rise sharply; *point cloud* declines steadily from 2016 onward.

Emergent keywords by period Table 8 reports the top 25 distinctive keywords per macro-period by log-likelihood keyness (G^2, word vs. rest of corpus), with residual OCR ligature artefacts and extraction boilerplate marked †.

Ignoring † items, the Pre-Transformer list is dominated by computer-vision and geometry vocabulary (*descriptors*, *sift*, *rendering*, *computed*). The Transformer/pre-LLM period is topically the most heterogeneous, reflecting a transitional moment in which the field simultaneously explored game-based her-

Table 8. Top 25 distinctive keywords per macro-period (keyness = G^2). FA = absolute frequency; FR = /10k; LL = log-likelihood. † = OCR ligature artefact or extraction boilerplate (e.g. email addresses, journal name strings).

Pre-Transformer (2008–2017)					**Transformer / pre-LLM (2018–2022)**					**Post-LLM (2023–2026)**				
#	Word	FA	FR	LL	#	Word	FA	FR	LL	#	Word	FA	FR	LL
1	cant†	331	2.57	840	1	tango	219	1.47	507	1	e-mail†	571	2.91	924
2	descriptors	448	3.48	461	2	game	1806	12.12	500	2	china	741	3.78	472
3	modi†	169	1.31	441	3	inpainting	414	2.78	495	3	significant	1071	5.46	459
4	simpli†	166	1.29	433	4	participants	1826	12.25	447	4	emotion	445	2.27	412
5	spiral	183	1.42	410	5	obis	164	1.10	379	5	emotions	537	2.74	379
6	stroke	247	1.92	404	6	projectile	159	1.07	368	6	learning	2949	15.03	365
7	elds†	152	1.18	396	7	gloss	172	1.15	309	7	engagement	968	4.93	350
8	atayal	150	1.17	391	8	authoring	309	2.07	308	8	facial	333	1.70	341
9	penn	154	1.20	390	9	puppetry	117	0.79	245	9	specific	1794	9.15	338
10	plaza	156	1.21	379	10	tweets	131	0.88	234	10	authenticity	431	2.20	327
11	lter†	140	1.09	365	11	fingerprints	104	0.70	222	11	experience	3457	17.62	323
12	ciently†	139	1.08	362	12	disc	114	0.76	216	12	proceedings†	416	2.12	320
13	nding†	138	1.07	360	13	players	442	2.97	215	13	dataset	2244	11.44	307
14	computed	516	4.01	351	14	fingerprint	102	0.68	211	14	historical	2098	10.70	301
15	tting†	130	1.01	339	15	listening	260	1.74	209	15	accessibility	453	2.31	295
16	ectance†	129	1.00	336	16	wedge	110	0.74	202	16	loss	617	3.15	293
17	warcbase	125	0.97	326	17	custodian	151	1.01	201	17	ovis†	163	0.83	288
18	dept†	162	1.26	325	18	minecraft	90	0.60	198	18	paradata	192	0.98	280
19	rendering	488	3.79	324	19	munsell	94	0.63	193	19	artwork	600	3.06	276
20	kunqu	128	0.99	323	20	bbc	101	0.68	192	20	tourism	615	3.14	273
21	sift	197	1.53	321	21	craft	177	1.19	189	21	diverse	552	2.81	268
22	descriptor	277	2.15	319	22	floor	307	2.06	185	22	environmental	446	2.27	261
23	argumentation	166	1.29	304	23	lublin	79	0.53	183	23	island	226	1.15	256
24	ltering†	116	0.90	302	24	cemomemo†	78	0.52	180	24	immersive	664	3.39	254
25	coef†	122	0.95	299	25	hieroglyph	103	0.69	178	25	parks	168	0.86	251

itage (*game*, *players*, *minecraft*), social media (*tweets*), user studies (*participants*), and colour science (*munsell*, *gloss*) before converging on deep-learning approaches. The Post-LLM period shifts towards affective and experiential vocabulary (*emotion*, *experience*, *engagement*, *immersive*) alongside learning infrastructure (*learning*, *dataset*, *loss*) and policy terms (*accessibility*, *paradata*). The prominence of *china* (3.78/10k) suggests increased Chinese institutional presence from 2023, though confirmation requires author-affiliation data not available in the corpus.

4.3 Discursive Reframing (H3)

We track three semantic verb groups by relative frequency per 10k tokens and log-likelihood across macro-periods (Tables 9 and 10; Fig. 6): **Analysis** (*analyse, describe, measure, classify, compare*), **Reconstruction** (*reconstruct, model, simulate, render, recover*), and **Generation** (*generate, predict, produce, synthesise, create*). The operationalisation of discursive reframing through verb semantics rests on the assumption that dominant action verbs encode a discipline's epistemological stance, what it does to its objects of study, and therefore track paradigm shifts more directly than content nouns, which merely name new topics [14]. Counts are computed by regex matching all inflected forms of each lemma and are not POS-disambiguated (see caveat for *model* in Table 10).

Table 9. Verb-group relative frequency (/10k) by macro-period. ΔLL: positive = Pre-Transformer overuse; negative = Post-LLM overuse. p-values derived from the χ^2 distribution of G^2 (1 d.f.).

Verb group	Pre-Transformer	Trans. / pre-LLM	Post-LLM	ΔRel%	ΔLL	p
Analysis	54.62	54.92	**57.73**	+5.7%	−13.48	< 0.001
Reconstruction	**59.12**	49.22	54.87	−7.2%	**24.76**	< 0.001
Generation	33.44	37.08	**41.41**	+23.8%	**−131.50**	< 0.001

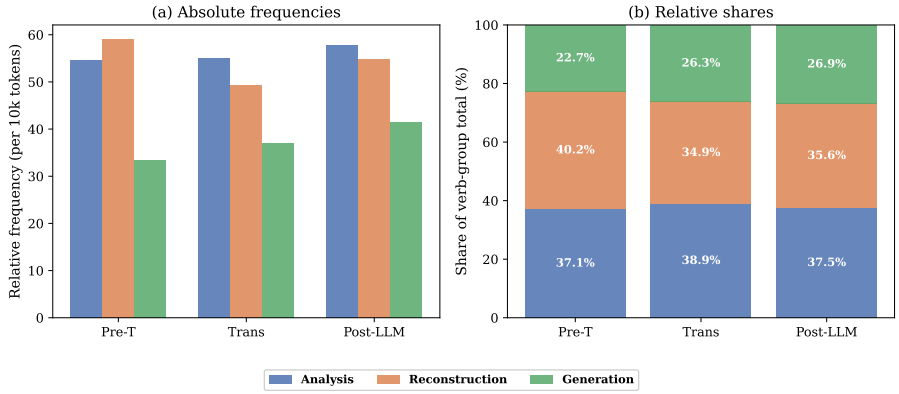

Fig. 6. Verb-group shares by macro-period (H3). Generation (bottom band) grows steadily; Reconstruction (middle) contracts; Analysis (top) remains broadly stable with an upward inflection post-2022.

The Generation group shows the sharpest rise (+23.8%, ΔLL= −131.50, $p < 0.001$), driven by *generate* (+37.1%), *create* (+34.2%) and the remarkable ascent of *predict* (+226.9%, ΔLL= −256.22), a diagnostic marker of the predictive-modelling paradigm imported from computer vision and Natural Language Processing (NLP). Two internal counter-trends temper this picture: *produce* falls −33.3% (likely displaced by the more technically specific *generate*) and *synthesise* declines modestly (n.s.). Despite these, the net group effect remains a significant +23.8%.

To test direct substitution between *generate* and *produce*, Pearson and Spearman correlations were computed on the 19 annual relative-frequency series (Table 11). The pairwise correlation is negligible ($r = -0.044$, $p = 0.855$), ruling out abrupt annual substitution. Trend tests reveal a more nuanced picture: *produce* declines monotonically ($\rho = -0.677$, $p < 0.001$) while *generate* grows non-significantly ($\rho = +0.303$, $p = 0.189$). The generate/produce ratio nonetheless more than doubles across macro-periods ($1.03 \rightarrow 1.56 \rightarrow 2.12$), pointing to *secular displacement* rather than direct lexical substitution.

Within Analysis, *classify* records the strongest single-lemma effect (+509.7%, ΔLL= −865.55), signalling the wholesale adoption of classification-framed tasks from machine learning. Conversely, *describe* retreats (−26.9%), marking a shift from the observational rhetoric of pre-2018 heritage documentation: together,

Table 10. Per-lemma relative frequency (/10k). ΔLL convention as above. p: derived from $\chi^2(1)$ distribution of G^2. LR = $\log_2$(Post-LLM / Pre-T): effect size (Hardie 2014); LR values in bold when |LR| > 1. Frequency values in bold = maximum across the three periods for that lemma. **model* counts include nominal uses. POS recount (NLTK, n = 21,308): nominal share 91.5 % → 91.4 % → 93.3 %; verbal RF 3.69 → 3.34 → 3.07/10k (declining). LR (+0.21) is nominal-driven (*foundation model*, *language model*).

Group	Lemma	Pre-T	Trans/pre-LLM	Post-LLM	ΔRel%	ΔLL	p	LR
Analysis	*analyse/analyze*	15.45	18.81	**18.94**	+22.6%	−54.82	< 0.001	+0.33
	describe	**19.01**	11.79	13.90	−26.9%	**125.21**	< 0.001	−0.41
	measure	**8.48**	7.29	6.35	−25.1%	48.32	< 0.001	−0.38
	classify	1.44	8.04	**8.78**	+509.7%	**−865.55**	< 0.001	**+2.65**
	compare	**10.24**	8.98	9.77	−4.6%	1.76	0.185	−0.03
Reconstruction	*reconstruct*	**10.47**	6.19	6.39	−39.0%	**157.86**	< 0.001	−0.67
	*model**	38.97	37.31	**43.96**	+12.8%	−46.70	< 0.001	+0.21
	simulate	**2.90**	2.23	2.08	−28.3%	21.27	< 0.001	−0.44
	render	**5.63**	2.80	1.78	−68.4%	**340.49**	< 0.001	**−1.62**
	recover	**1.14**	0.68	0.66	−42.1%	20.43	< 0.001	−0.74
Generation	*generate*	9.12	9.43	**12.50**	+37.1%	**−81.48**	< 0.001	+0.49
	predict	1.34	4.17	**4.38**	+226.9%	**−256.22**	< 0.001	**+1.75**
	produce	**8.86**	6.06	5.91	−33.3%	93.64	< 0.001	−0.55
	synthesise/synthesize	**0.63**	0.56	0.51	−19.0%	1.97	0.160	−0.27
	create	13.49	16.86	**18.10**	+34.2%	**−104.10**	< 0.001	+0.46

Table 11. Substitution analysis for *generate* vs *produce* (n = 19 annual sub-corpora). Rows 1–2: pairwise correlation between the two verbs' yearly relative frequencies; 95% CI via Fisher's z-transform; p-values from the standard Pearson t-test (n = 19). Rows 3–4: Spearman monotonic trend against publication year; p-values from the standard Spearman t-approximation; CI not reported for trend rows as the year variable is fixed, not randomly sampled.

Test	Value	95% CI	p-value	Interpretation
Pairwise r (gen vs prod)	−0.044	[−0.489, +0.418]	0.855	no substitution
Pairwise ρ (gen vs prod)	−0.044	[−0.489, +0.418]	0.856	no substitution
Trend ρ: produce vs year	−0.677	n/a	< 0.001	secular decline
Trend ρ: generate vs year	+0.303	n/a	0.189	non-significant rise

these complement each other as signals of an epistemological transition from descriptive cataloguing to predictive categorisation. Within Reconstruction, *render* collapses by −68.4% and *reconstruct* by −39.0%, consistent with the displacement of explicit 3D pipelines by end-to-end learned representations and a fundamental reorientation of the field's core technical practices.

5 Conclusions

This study traces linguistic change in JOCCH (2008–2026) across three hypotheses. **H1** (lexical standardisation) is *not* supported. The 14% raw hapax decline is entirely a Zipfian corpus-size artefact: once paper count is held constant, hapax legomena *increase* by 36%, with non-overlapping bootstrap CIs between the

Pre-Transformer and Post-LLM periods. Per-paper MATTR is non-monotonic and TF-IDF cosine similarity declines, indicating lexico-topical *divergence*, not convergence. The field is not homogenising; individual writing style has remained stable while shared vocabulary has expanded. **H2** (methodological-linguistic coupling) is strongly confirmed. LLM-era keywords grow by one to three orders of magnitude after 2022, proper-noun density rises by 40% driven by named architectures and datasets, and the gerund/past-participle shift signals a systematic move from result-reporting to process-oriented phrasing. These co-occurring lexical and grammatical changes are consistent with a rapid paradigm shift rather than incremental drift. **H3** (discursive reframing) is confirmed. The Generation verb group grows by 24% and *classify* by 510%, reflecting the wholesale adoption of predictive and classification tasks from computer vision and NLP. Conversely, *render* (−68%) and *reconstruct* (−39%) contract sharply, consistent with the displacement of explicit 3D pipelines by end-to-end learned representations. The generate/produce ratio more than doubles (1.03→2.12), pointing to secular displacement rather than direct lexical substitution. The generalisability of these findings beyond JOCCH remains an open question. JOCCH is the flagship ACM venue for Cultural Heritage Computing and covers the field's full methodological spectrum, but it represents a single publication culture with specific editorial conventions. We conjecture that the pattern of selective lexical absorption may generalise to other interdisciplinary venues at the engineering/humanities interface, where community diversity and editorial pluralism similarly resist convergence; however, this remains to be verified empirically. Some limitations remain. OCR artefacts in pre-2018 PDFs make the +36% hapax increase a conservative lower bound. The 42 inaccessible articles (6.6% of the target corpus) introduce potential selection bias if non-randomly distributed across periods, topics, or author affiliations; this effect is unquantifiable without access to the missing texts and may be non-negligible for lexical diversity measures and keyness analysis. The Post-LLM sub-corpus spans only four years (2023–2026); observed shifts may intensify, stabilise, or partially reverse as the community's engagement with generative AI matures, and results for this period should be interpreted as provisional. Finally, 34 of 36 tests reach significance ($p < 0.05$) and 34 survive Benjamini–Hochberg False Discovery Rate correction [15] ($q = 0.05$); no formal inter-annotator agreement study was conducted, with robustness argued from effect magnitudes throughout. The overall picture is one of *selective absorption*: authors in this domain rapidly adopt new paradigm vocabularies (H2, H3) without converging on a uniform style (H1). This pattern suggests that computational innovation reshapes *what* the community talks about and *how it frames actions*, but not the underlying stylistic diversity of individual authors. Observed shifts are described as post-2022 discourse shifts without asserting a specific causal mechanism, whether direct LLM assistance, topical evolution, or community demographic change remain plausible and non-exclusive explanations. Among the possible future directions: extension to comparable venues, analysis of methodological complexity over time (H4), BERTopic modelling (H5), author-affiliation analysis (H6), and function-word stylometry.

References

1. Bizzoni, Y., Degaetano-Ortlieb, S., Fankhauser, P., Teich, E.: Linguistic Variation and Change in 250 Years of English Scientific Writing: A Data-Driven Approach. Frontiers in Artificial Intelligence 3, 73 (2020). https://doi.org/10.3389/frai.2020.00073
2. Kobak, D., González-Márquez, R., Horvát, E., Lause, J.: Delving into LLM-Assisted Writing in Biomedical Publications through Excess Vocabulary. Science Advances 11, eadt3813 (2025). https://doi.org/10.1126/sciadv.adt3813
3. Bao, T., Zhao, Y., Mao, J., et al.: Examining Linguistic Shifts in Academic Writing Before and After the Launch of ChatGPT: A Study on Preprint Papers. Scientometrics 130, 3597–3627 (2025). https://doi.org/10.1007/s11192-025-05341-y
4. Padmakumar, V., He, H.: Does Writing with Language Models Reduce Content Diversity? In: International Conference on Learning Representations (ICLR 2024). https://doi.org/10.48550/arXiv.2309.05196
5. Kendro, K., Maloney, J.: Do Large Language Models Produce Texts With "Human-Like" Lexical Diversity? Evidence From Four ChatGPT Models. International Journal of Applied Linguistics (2026). https://doi.org/10.1111/ijal.70115
6. Covington, M.A., McFall, J.D.: Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR). Journal of Quantitative Linguistics 17(2), 94–100 (2010). https://doi.org/10.1080/09296171003643098
7. Tognini-Bonelli, E.: Corpus Linguistics at Work. John Benjamins, Amsterdam (2001). https://doi.org/10.1075/scl.6
8. McCarthy, P.M., Jarvis, S.: MTLD, vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. Behaviour Research Methods 42(2), 381–392 (2010). https://doi.org/10.3758/BRM.42.2.381
9. Rayson, P., Garside, R.: Comparing Corpora using Frequency Profiling. In: Proceedings of the Workshop on Comparing Corpora (ACL 2000), pp. 1–6 (2000). https://doi.org/10.3115/1117729.1117730
10. Brezina, V.: Statistics in Corpus Linguistics: A Practical Guide. Cambridge University Press (2018). https://doi.org/10.1017/9781316410899
11. Scott, M.: PC Analysis of Key Words – and Key Key Words. System 25(2), 233–245 (1997). https://doi.org/10.1016/S0346-251X(97)00011-0
12. Gries, S.Th.: Dispersions and Adjusted Frequencies in Corpora. International Journal of Corpus Linguistics 13(4), 403–437 (2008). https://doi.org/10.1075/ijcl.13.4.02gri
13. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics 19(2), 313–330 (1993). https://aclanthology.org/J93-2004/
14. Hyland, K.: Disciplinary Discourses: Social Interactions in Academic Writing. University of Michigan Press, Ann Arbor (2004). ISBN 978-0-472-03024-2 https://doi.org/10.3998/mpub.6719
15. Benjamini, Y., Hochberg, Y.: Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B 57(1), 289–300 (1995). https://doi.org/10.1111/j.2517-6161.1995.tb02031.x