Visualizing 30 Years of CHI Research with Generative AI

Sunggyeol Oh
Department of Computer Science
Virginia Tech
Blacksburg, Virginia, USA
sunggyeol@vt.edu

Andrew Katz
Department of Engineering Education
Virginia Tech
Blacksburg, Virginia, USA
akatz4@vt.edu

Scott McCrickard
Department of Computer Science
Virginia Tech
Blacksburg, Virginia, USA
mccricks@cs.vt.edu

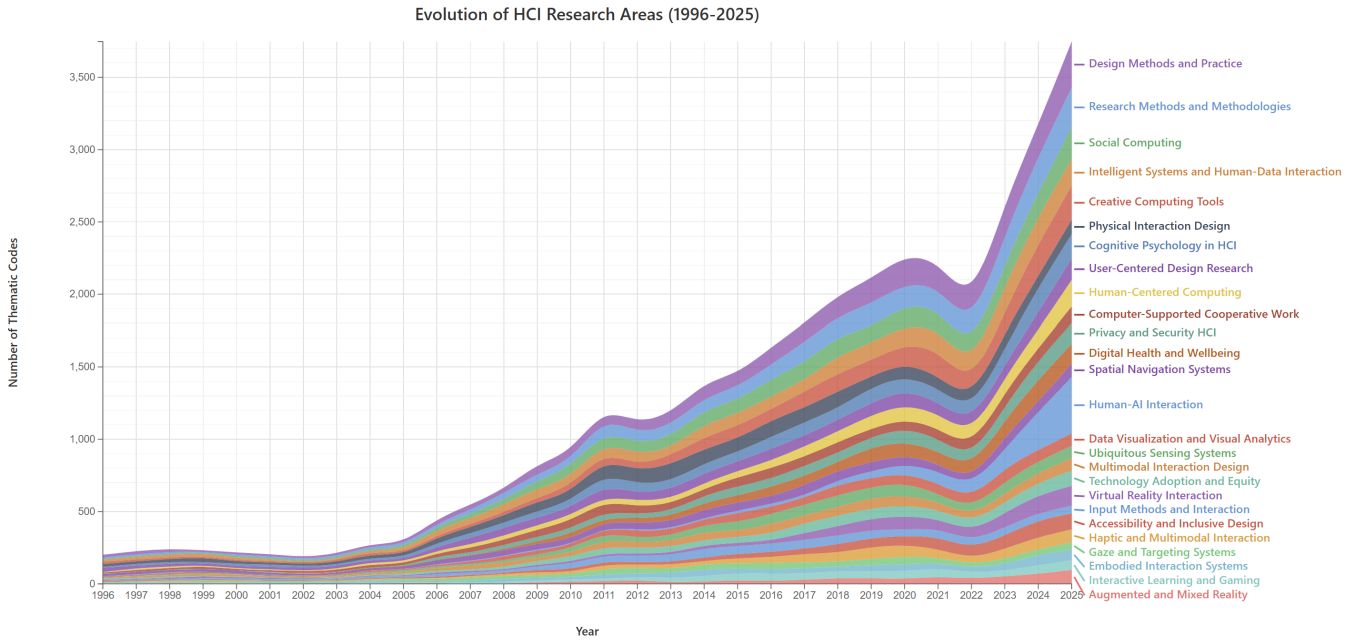


Figure 1: Streamgraph showing the evolution of 26 HCI research areas from 1996 to 2025. The thickness of each colored stream represents the annual volume of thematic codes for a meta-theme.

Abstract

CHI has grown from roughly 60 papers annually in 1996 to over 1,000 in 2025, creating a corpus too large for manual thematic tracking. We present a methodology that combines Generative AI-assisted coding, text embedding, and multi-stage clustering to organize 11,847 CHI paper abstracts into 921 themes under 26 meta-themes. Bootstrap analysis confirms stability (Adjusted Rand Index [ARI] = 0.91) and 100% corpus coverage. Two complementary visualizations reveal distinct aspects of field evolution: a streamgraph showing absolute volume changes and a heatmap showing proportional shifts. Together, they distinguish growth from shifting priorities. Human-AI interaction increased by 60 times in absolute terms and by 5 times proportionally. At the same time, User-Centered Design Research grew modestly in absolute terms while its relative share fell to roughly a third, suggesting maturation into a foundational practice. These patterns, invisible at this scale, demonstrate how Generative AI-powered methods can reveal the dynamics of scholarly evolution.

CCS Concepts

• **Human-centered computing** → **Visualization design and evaluation methods**; *HCI theory, concepts and models*; • **Information systems** → *Clustering*.

Keywords

Research trends, Thematic analysis, Large language models, Meta-Research, Visualization, CHI

ACM Reference Format:
Sunggyeol Oh, Andrew Katz, and Scott McCrickard. 2026. Visualizing 30 Years of CHI Research with Generative AI. In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3772363.3798338

This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI EA '26, Barcelona, Spain
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3798338

1 Introduction

The ACM CHI conference on Human Factors in Computing Systems has grown from roughly 60 papers per year in 1996 to more than 1,000 in 2025. Nearly 12,000 publications now span three decades.

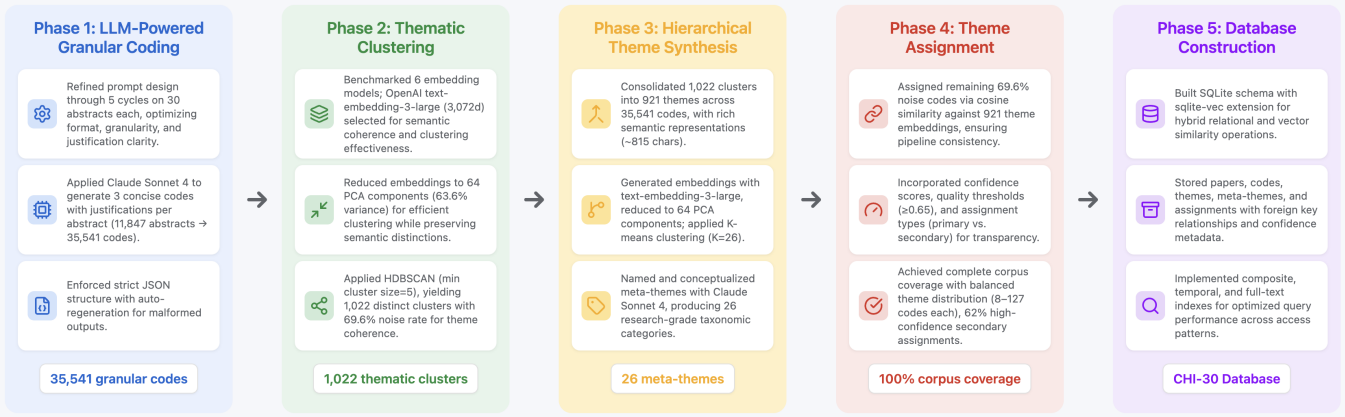


Figure 2: Five-phase framework for building CHI-30 database. The pipeline combines LLM-powered coding, text embedding, and multi-stage clustering to move from abstracts to a hierarchical semantic database.

Manually tracking how themes emerge, peak, and fade across a corpus this size presents significant challenges.

This raises a methodological question: can Generative AI methods produce reliable thematic taxonomies of scholarly corpora at scale, and what patterns emerge when we examine both absolute and proportional change? Large language models (LLMs) offer a path forward. A systematic review of 153 CHI papers found LLMs increasingly used as research tools, though authors consistently raised validity and reproducibility concerns [25]. These concerns are well-founded: LLMs generate plausible but nonfactual content (hallucination) and evaluation practices vary widely, undermining trust in automated analysis [14, 18]. Prior scientometric work on CHI has mapped research themes through co-word analysis [20] and citation network analysis [19], revealing thematic clusters and their temporal trajectories. More broadly, systems such as vitaLITy [23] and vitaLITy 2 [2] have applied text embeddings and LLMs to support semantic literature search across large scholarly corpora. However, these approaches typically produce snapshots at fixed granularities or support individual paper retrieval rather than hierarchical, field-level thematic structures. Neural topic models such as BERTopic [8] embed and cluster whole documents, but produce flat topic distributions without the hierarchical organization or interpretive provenance that LLM-generated codes provide.

We present a methodology addressing these gaps. Our five-phase pipeline processes 11,847 CHI paper abstracts (1996–2025), generating 35,541 codes through LLM analysis and organizing them into 921 themes under 26 meta-themes. Bootstrap analysis confirms stability (ARI = 0.91) with 100% corpus coverage.

We contribute two visualizations that reveal HCI's evolution from complementary angles: a streamgraph shows absolute volume changes; a heatmap shows proportional shifts. We also demonstrate a reproducible pipeline combining LLM-assisted coding, text embedding, and multi-stage clustering, transforming unstructured abstracts into a three-level taxonomy (codes, themes, meta-themes) with explicit provenance at each stage. We validate the approach through bootstrap analysis and parameter sensitivity testing, providing evidence that Generative AI-assisted methods can produce reliable thematic taxonomies at scale.

The visualizations surface patterns invisible to manual analysis. Human-AI Interaction surged 60-fold (11 to 659 thematic codes) and 5-fold proportionally (1.7% to 8.3%), accelerating sharply after 2020. User-Centered Design Research grew modestly (59 to 345 thematic codes) but saw its proportional share fall from 10.3% to 3.8%, suggesting maturation into a foundational practice rather than decline. Design Methods and Practice grew absolutely (55 to 669 thematic codes) while maintaining consistent proportional presence (6.0–9.4%), confirming its role as a stable disciplinary core. These dual perspectives, absolute and proportional, help researchers and program chairs distinguish genuine growth from changing priorities within an expanding field.

Table 1: LLM-Powered Granular Coding samples (three code–rationale pairs per paper).

Code	Rationale
Big or Small, It's All in Your Head: Visuo-Haptic Illusion of Size-Change Using Finger-Repositioning (10.1145/3613904.3642254)	
Visuo-Haptic Illusion	Core concept of creating size-perception illusions by combining visual and haptic feedback in virtual environments.
Finger-Repositioning	Key technical method used to manipulate haptic perception without changing the physical device size.
Virtual Reality Controllers	Primary hardware/interface used to implement and test the size-change illusion technique.
Misty: UI Prototyping Through Interactive Conceptual Blending (10.1145/3706598.3713924)	
Conceptual Blending	Central technique that blends multiple concepts to produce novel UI behaviors and layouts.
UI Prototyping Tools	Toolkit/workflow enabling rapid construction and testing of interactive prototypes.
Developer–Designer Collaboration	System features that support co-creation and iteration across developer and designer roles.
Being The Creek: Mobile Augmented Reality Experience as an Invitation for Exploring More-Than-Human Perspectives (10.1145/3706598.3713713)	
More-Than-Human Perspectives	Central theme exploring non-human viewpoints and challenging human-centric perspectives.
Mobile Augmented Reality	Core technology enabling participants to experience the creek's perspective through mobile AR interface.
Embodied Multisensory Engagement	Key methodological approach using physical positioning and sensory attunement to facilitate perspective-taking.

2 Methodology

Our pipeline transforms 11,847 CHI abstracts (1996–2025) into a hierarchical thematic database through five phases. Although CHI proceedings date back to 1982, we begin at 1996 to capture a 30-year window. The ACM Digital Library provides structured abstracts only from 1999 onward; 1996–1998 required manual extraction, making further backward extension impractical. Figure 2 illustrates the workflow.

2.1 Phase 1: LLM-Powered Granular Coding

Recent work demonstrates that LLMs can support qualitative coding when paired with structured prompts and human oversight [6, 7, 16, 29]. We employ Claude Sonnet 4, selected for its strong instruction adherence and consistent structured output generation at scale (11,847 API calls), to generate three semantic codes per abstract through structured API calls. Each code consists of a concise thematic phrase (typically 2–5 words) paired with a justification explaining its derivation, constrained to 25 words to maintain focus and prevent verbose outputs. This dual structure serves multiple purposes: concise labels facilitate efficient clustering, justifications provide transparency for validation, and the combination enables similarity computations in later phases.

The prompt underwent five refinement cycles: in each cycle, an HCI expert reviewed codes generated from 30 randomly sampled abstracts, assessing code consistency, specificity, and concept coverage, and refined the prompt before the next iteration. This iterative refinement addressed issues including output format consistency, code granularity calibration, and justification clarity. Malformed responses that did not conform to the required JSON format triggered automatic regeneration. After the final iteration yielded satisfactory code quality, the refined prompt was applied to the full corpus of 11,847 abstracts. This process generated 35,541 codes preserving nuanced research concepts while enabling computational analysis. Table 1 shows representative examples.

2.2 Phase 2: Thematic Clustering

We evaluated six embedding models on a 2,000-code sample [27, 28]: OpenAI (text-embedding-3-large, text-embedding-ada-002), Google (gemini-text-embedding-004), Microsoft (multilingual-e5-large-instruct), WhereIsAI (UAE-large-v1), and BAAI (BGE-large-en-v1.5). Testing across minimum cluster sizes of 5, 10, and 15 using Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [21, 22], OpenAI's text-embedding-3-large (3,072 dimensions) with minimum cluster size 5 was selected, identifying 25 distinct clusters on the sample. This model produced the most clusters across all parameter settings (notably 7 clusters at size=10 while other models collapsed to 2), indicating stronger semantic differentiation in the embedding space.

The 3,072-dimensional embedding space required dimensionality reduction for efficient clustering. Principal Component Analysis (PCA) reduced the full 35,541×3,072 embedding matrix to 64 components [1, 5]. Systematic variation (32–128 components) confirmed this configuration provides stable clustering (ARI = 0.819–1.000 vs. baseline) while enabling efficient full-corpus processing. The transformation matrix is preserved to allow consistent projection of new codes during future database updates.

HDBSCAN was applied to the PCA-reduced space with minimum cluster size 5 and Euclidean distance. The algorithm identified 1,022 clusters covering 30.4% of codes, with 69.6% marked as noise. The high noise rate is expected and strategically advantageous: rather than forcing all codes into potentially incoherent clusters, HDBSCAN's conservative approach ensures that identified themes maintain high internal consistency, while noise points represent unique contributions or emerging topics not yet forming coherent clusters.

2.3 Phase 3: Hierarchical Theme Synthesis

Claude Sonnet 4 generated labels for each of the 1,022 clusters through structured prompts. Each cluster received three components: a concise theme name capturing the research area, a concept definition explaining scope and boundaries, and a coherence rationale justifying thematic unity. Clusters with identical names were systematically merged, reducing 1,022 clusters to 921 unique themes spanning 6–207 papers each. The largest theme, "Participatory Design," encompasses 207 papers, while specialized themes contain as few as 6, demonstrating the taxonomy's ability to capture both mainstream areas and focused specializations.

We then generated semantic representations for each theme by combining repeated theme names with concept definitions to weight primary semantic signals, averaging 815 characters per representation. These were embedded using text-embedding-3-large, and K-means clustering [11] with K ranging from 2 to 30 revealed K=26 as optimal, achieving the highest silhouette score (0.103; Figure 3). Meta-theme sizes range from 18–69 themes. Claude Sonnet 4 generated descriptive meta-theme names aligned with established HCI research traditions.

2.4 Phase 4: Theme Assignment

To achieve complete corpus coverage, we implemented two-tier assignment. The 30.4% of codes clustered by HDBSCAN received primary assignments with methodological attribution to density-based clustering, distinguishing them from subsequent similarity-based assignments. The remaining 69.6% received secondary assignments through vectorized cosine similarity computation, comparing each unassigned code's embedding against all 921 theme concept embeddings simultaneously.

Each secondary assignment includes confidence classification: high (similarity > 0.75) represents semantically strong matches, often codes that narrowly missed HDBSCAN inclusion due to density requirements; medium (0.65–0.75) typically indicates interdisciplinary research bridging multiple thematic areas; and low (< 0.65) frequently identifies emerging topics not well represented by existing thematic structures. This achieves 100% corpus coverage while maintaining algorithmic provenance, enabling researchers to apply differential quality thresholds based on their analytical requirements.

2.5 Phase 5: Database Construction

In the final phase, the CHI-30 database was built on SQLite with the sqlite-vec extension, enabling vector similarity operations alongside

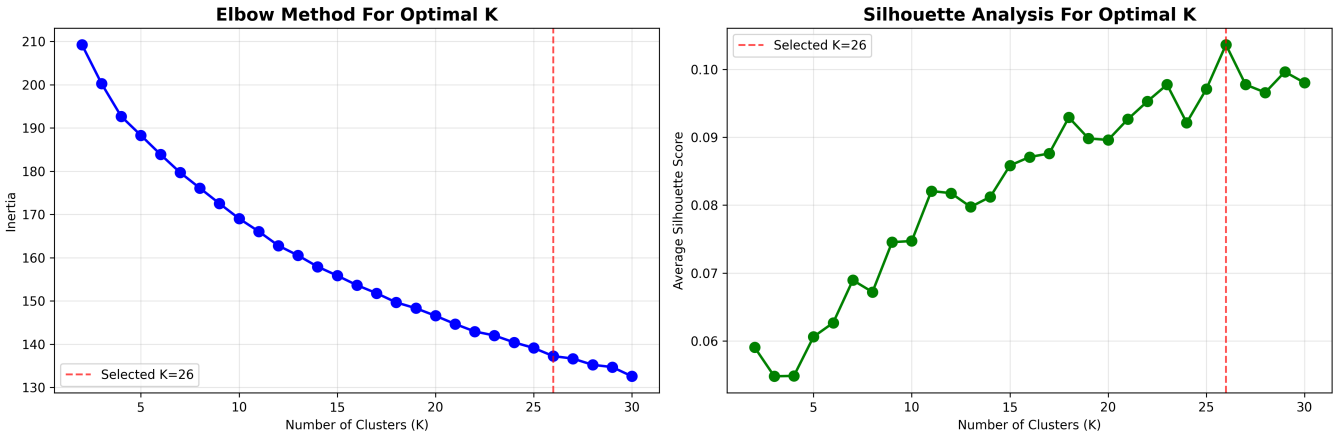


Figure 3: K-means optimization analysis. The elbow method (left) shows diminishing returns in inertia reduction, while silhouette analysis (right) identifies K=26 as optimal for cluster separation and coherence.

relational queries in a single portable file. The schema includes tables for papers, codes, themes, meta-themes, and assignments, with indexing on theme IDs, confidence scores, and temporal metadata.

The resulting single-file architecture requires no external dependencies, supporting both programmatic access through SQL queries and semantic search through vector similarity operations.

2.6 Validation

Cluster stability was assessed through bootstrap analysis: each of 10 iterations resampled 80% of codes and repeated the full PCA-HDBSCAN pipeline, with resulting cluster assignments compared against the full-corpus baseline. The mean ARI of 0.91 ($\sigma = 0.0038$) [15, 26] indicates consistent recovery of thematic structures across subsamples; the low observed variance suggests additional iterations would not materially change the estimate. Because HDBSCAN assigns 69.6% of codes as noise, this stability primarily reflects the robustness of the 30.4% that form coherent clusters; noise-point consistency may contribute to the high agreement.

PCA component variation across 32–128 dimensions produced ARI = 0.819–1.000 against the 64-component baseline, with cluster counts and noise rates remaining stable throughout. This confirms the 64-component configuration as an optimal balance between computational efficiency and clustering quality [17]. Additional HDBSCAN parameter sensitivity testing (minimum cluster size: 3–10, minimum samples: 3–10) showed clustering remains stable near the selected configuration (ARI > 0.8 within ±1 of chosen values), with expected divergence at more distant parameter settings (ARI = 0.586–1.000).

All 921 themes map to exactly one of 26 meta-themes, with no orphaned themes. Temporal coverage spans all years (1996–2025) with multi-theme representation preventing historical blind spots. The average 35.4 themes per meta-theme (range: 18–69) maintains navigational usability while preventing fragmentation.

3 Visualization Design

We present two visualizations that reveal distinct aspects of HCI's evolution: a heatmap and a streamgraph.

Prior work has visualized CHI's structure through co-word analysis [20], citation networks [19], and author metadata [13]. These approaches typically present either absolute counts or network relationships. We instead pair absolute and proportional views, enabling researchers to distinguish field expansion from compositional change.

3.1 Streamgraph: Absolute Volume

The streamgraph (Figure 1) shows absolute research activity across 26 meta-themes from 1996 to 2025. Each colored stream represents a meta-theme, with stream thickness indicating annual paper volume. Streamgraphs extend stacked area charts by using a centered baseline that emphasizes organic flow [3], building on ThemeRiver's river metaphor for temporal document analysis [12] where width encodes thematic strength over time. The centered layout reduces the distortion that fixed baselines impose on middle layers, making cross-theme comparison easier.

The streamgraph answers questions about which areas attract the most papers, when domains emerged, and how total activity has grown. Human-AI Interaction surges from 11 thematic codes (1996–1998) to 659 codes (2023–2025), a 60-fold increase visible as a rapidly widening stream after 2020.

3.2 Heatmap: Proportional Shifts

The heatmap (Figure 4) reveals proportional shifts in research focus. Each cell represents the percentage of total activity devoted to a meta-theme during a 3-year period, with color intensity encoding relative prominence. Temporal binning (3-year periods) reduces noise while preserving trends, and percentage-based encoding normalizes for CHI's growth, revealing whether areas gain or lose prominence independent of field expansion.

The heatmap surfaces which areas are becoming central, which domains are maturing, and how compositional balance has shifted. Human-AI Interaction grows from 1.7% to 8.3% of total research. User-Centered Design Research drops from 10.3% to 3.8% despite growing absolute counts, indicating maturation rather than decline.

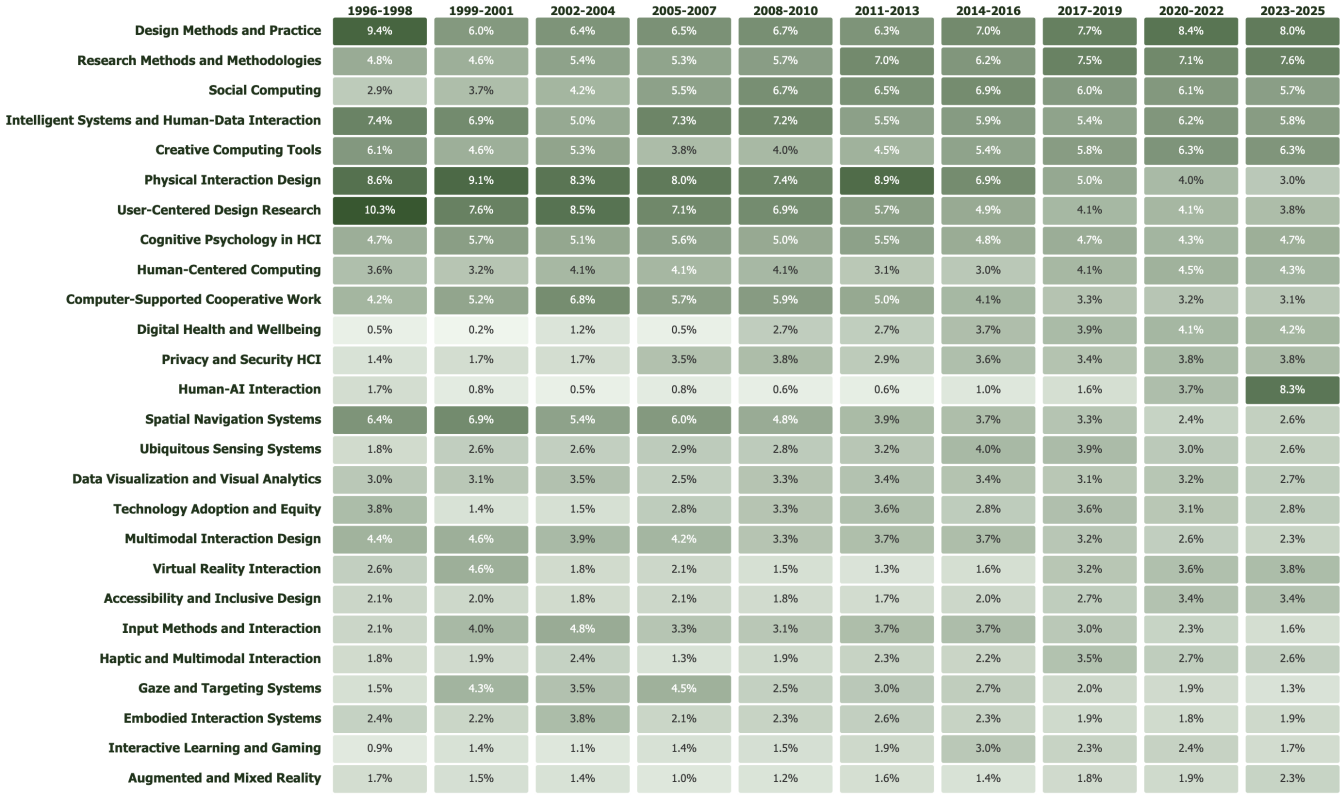


Figure 4: Proportional representation of 26 HCI meta-themes across 3-year periods (1996–2025). Cell intensity indicates the percentage of total research activity per meta-theme.

Design Methods and Practice holds steady at 6.0–9.4%, confirming its role as a disciplinary core.

3.3 Complementary Perspectives

Absolute volume conflates field expansion with thematic priority: a topic doubling in papers while the field triples has lost relative ground. The streamgraph shows growth; the heatmap shows centrality, helping researchers distinguish genuine surges from artifacts of CHI's expansion. Applying this dual lens to the full corpus reveals both expected and surprising patterns, though we note these are derived from abstracts and may not capture nuances present in full texts.

Human-AI Interaction shows the most dramatic trajectory: a 60-fold absolute increase and 5-fold proportional surge, both accelerating sharply after 2020 with the public release of large language models.

User-Centered Design Research presents the opposite pattern. Absolute volume grew modestly (59 to 345 thematic codes), far slower than the field's overall expansion, while its proportional share fell from 10.3% to 3.8%, signaling maturation rather than decline. User-centered methods have become foundational assumptions rather than novel contributions, aligning with accounts of HCI's paradigm shifts [4, 10]: what was once a research frontier has become methodological baseline.

Design Methods and Practice demonstrates both absolute growth (55 to 669 thematic codes) and consistent proportional presence (6.0–9.4%), tracking CHI's overall expansion almost exactly. This stability suggests a disciplinary anchor defining HCI's identity regardless of application domains.

Other meta-themes show instructive dynamics. Digital Health and Wellbeing rose from 0.5% to 4.2%, accelerating during pandemic years (2020–2022). Privacy and Security HCI doubled its share by the mid-2000s (1.4% to 3.5% in 2005–2007), generally stabilizing around 3%–4% in subsequent periods. Cognitive Psychology in HCI held between 4.3% and 5.7% across the full period before settling to 4.7%, likely reflecting integration into mainstream methodology [9].

The dual-visualization approach proves essential: a rising stream with shrinking heatmap cell signals growth slower than the field, while a thin stream with darkening cells indicates concentration without volume. These distinctions help program chairs balance sessions, and researchers identify genuine gaps.

4 Limitations

Several limitations bound our claims. Abstract-only analysis trades depth for scale: methodological details and theoretical frameworks live in full text, beyond our taxonomy's reach. The 26 meta-themes impose boundaries that some papers straddle awkwardly. LLM-generated codes, despite validation (ARI = 0.91), inherit training biases [24] and may miss expert-level nuance. Our multi-stage

pipeline also risks compounding errors: biased codes may produce skewed embeddings, and information loss during PCA may obscure meaningful distinctions. We mitigate these risks through prompt refinement with manual review at the coding stage, systematic PCA component variation testing (ARI = 0.819–1.000) to verify preserved cluster structure, and bootstrap resampling (ARI = 0.91) to confirm stability. These measures validate algorithmic consistency but not semantic correctness; hallucination remains an inherent limitation of LLMs [30], and expert review of generated themes would further strengthen validity. Finally, we analyze CHI only: findings characterize this venue specifically, not the HCI field as a whole.

5 Conclusion

We present a validated pipeline for transforming large scholarly corpora into navigable thematic structures. The CHI-30 database (11,847 paper abstracts, 921 themes, 26 meta-themes) offers the HCI community a new lens on three decades of research evolution, with complementary visualizations that distinguish genuine surges from artifacts of field expansion. The concerns raised about LLM validity in HCI research [25] are real; our bootstrap analysis provides evidence that automated coding can complement human judgment at scale, though careful validation reduces rather than resolves the complexities of using LLMs as research tools. We see this work as a step toward scalable, transparent methods for scholarly self-reflection, not a replacement for expert interpretation but a scaffold for it.

Acknowledgments

Generative AI models (Anthropic Claude series) were used to assist with text revisions. All AI-generated content was subsequently reviewed to ensure accuracy.

References

[1] Hervé Abdi and Lynne J. Williams. 2010. Principal component analysis. *WIREs Computational Statistics* 2, 4 (July 2010), 433–459. doi:10.1002/wics.101

[2] Hongye An, Arpit Narechania, Emily Wall, and Kai Xu. 2024. vitaLITy 2: Reviewing Academic Literature Using Large Language Models. doi:10.48550/ARXIV.2408.13450 Version Number: 1.

[3] L. Byron and M. Wattenberg. 2008. Stacked Graphs – Geometry & Aesthetics. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (Nov. 2008), 1245–1252. doi:10.1109/TVCG.2008.166

[4] Susanne Bødker. 2015. Third-wave HCI, 10 years later—participation and sharing. *Interactions* 22, 5 (Aug. 2015), 24–31. doi:10.1145/2804405

[5] Leo H. Chiang, Evan L. Russell, and Richard D. Braatz. 2001. Principal Component Analysis. In *Fault Detection and Diagnosis in Industrial Systems*, Michael J. Grimble and Michael A. Johnson (Eds.). Springer London, London, 35–55. doi:10.1007/978-1-4471-0347-9_4

[6] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 9993–10001. doi:10.18653/v1/2023.findings-emnlp.669

[7] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. CollabCoder: A Lower-barrier, Rigorous Workflow for Inductive Collaborative Qualitative Analysis with Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–29. doi:10.1145/3613904.3642002

[8] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. doi:10.48550/arXiv.2203.05794

[9] Fatih Gurcan, Nergiz Ercil Cagiltay, and Kursat Cagiltay. 2021. Mapping Human–Computer Interaction Research Themes and Trends from Its Existence to Today: A Topic Modeling-Based Review of past 60 Years. *International Journal of Human–Computer Interaction* 37, 3 (Feb. 2021), 267–280. doi:10.1080/10447318.2020.1819668

[10] Steve Harrison, Deborah Tatar, and Phoebe Sengers. 2007. The Three Paradigms of HCI. (Jan. 2007).

[11] J. A. Hartigan and M. A. Wong. 1979. Algorithm AS 136: A K-Means Clustering Algorithm. *Applied Statistics* 28, 1 (1979), 100. doi:10.2307/2346830

[12] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. 2002. ThemeRiver: visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics* 8, 1 (March 2002), 9–20. doi:10.1109/2945.981848

[13] Nathalie Henry, Goodell Howard, Elmqvist Niklas, and Jean-Daniel Fekete. 2007. 20 Years of Four HCI Conferences: A Visual Exploration. *International Journal of Human–Computer Interaction* 23, 3 (Dec. 2007), 239–285. doi:10.1080/10447310701702402

[14] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* 43, 2 (Jan. 2025), 42:1–42:55. doi:10.1145/3703155

[15] Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2, 1 (Dec. 1985), 193–218. doi:10.1007/BF01908075

[16] Andrew Katz, Gabriella Coloyan Fleming, and Joyce B. Main. 2026. Thematic analysis with open-source generative AI and machine learning: a new method for inductive qualitative codebook development. *Humanities and Social Sciences Communications* 13, 1 (Jan. 2026), 209. doi:10.1057/s41599-026-06508-5

[17] Tilman Lange, Volker Roth, Mikio L. Braun, and Joachim M. Buhmann. 2004. Stability-Based Validation of Clustering Solutions. *Neural Computation* 16, 6 (June 2004), 1299–1323. doi:10.1162/089976604773717621

[18] Md Tahmid Rahman Laskar, Sawsan Alqahtani, M. Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, Enamul Hoque, Shafiq Joty, and Jimmy Huang. 2024. A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations. doi:10.48550/arXiv.2407.04069

[19] Clement Lee, Andrew Garbett, Junyan Wang, Bingzhang Hu, and Dan Jackson. 2019. Weaving the Topics of CHI: Using Citation Network Analysis to Explore Emerging Trends. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–6. doi:10.1145/3290607.3312776

[20] Yong Liu, Jorge Goncalves, Denzil Ferreira, Bei Xiao, Simo Hosio, and Vassilis Kostakos. 2014. CHI 1994-2013: mapping two decades of intellectual progress through co-word analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Toronto Ontario Canada, 3553–3562. doi:10.1145/2556288.2556969

[21] Leland McInnes and John Healy. 2017. Accelerated Hierarchical Density Based Clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, New Orleans, LA, 33–42. doi:10.1109/ICDMW.2017.12

[22] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* 2, 11 (March 2017), 205. doi:10.21105/joss.00205

[23] Arpit Narechania, Alireza Karduni, Ryan Wesslen, and Emily Wall. 2022. VITALITY: Promoting Serendipitous Discovery of Academic Literature with Transformers & Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 486–496. doi:10.1109/TVCG.2021.3114820

[24] Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in Large Language Models: Origins, Inventory, and Discussion. *Journal of Data and Information Quality* 15, 2 (June 2023), 1–21. doi:10.1145/3597307

[25] Rock Yuren Pang, Hope Schroeder, Kynnedy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–20. doi:10.1145/3706598.3713726

[26] Douglas Steinley. 2004. Properties of the Hubert-Arable Adjusted Rand Index. *Psychological Methods* 9, 3 (2004), 386–396. doi:10.1037/1082-989X.9.3.386

[27] Tomasz Walkowiak and Mateusz Gniewkowski. 2019. Evaluation of vector embedding models in clustering of text documents. In *Proceedings - Natural Language Processing in a Deep Learning World*. Incoma Ltd., Shoumen, Bulgaria, 1304–1311. doi:10.26615/978-954-452-056-4_149

[28] Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. 2019. Evaluating word embedding models: methods and experimental results. *APSIPA Transactions on Signal and Information Processing* 8, 1 (2019). doi:10.1017/ATSIP.2019.12

[29] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *28th International Conference on Intelligent User Interfaces*. ACM, Sydney NSW Australia, 75–78. doi:10.1145/3581754.3584136

[30] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2025. Hallucination is Inevitable: An Innate Limitation of Large Language Models. doi:10.48550/arXiv.2401.11817