An Empirical Study of the "Hard-Won Lesson": Two Decades of Research Insights

Abstract

This research investigates the congruence between research in major computer vision conferences and the tenets of the "hard-won lesson" articulated by Rich Sutton. Utilizing large language models (LLMs), we scrutinize twenty years of abstracts and titles from these conferences to evaluate the field's acceptance of these core concepts. Our approach employs cutting-edge natural language processing methodologies to methodically chart the progression of research paradigms within computer vision. The findings indicate notable patterns in the implementation of generalized learning algorithms and the exploitation of enhanced computational capabilities. We analyze the ramifications of these discoveries for the prospective trajectory of computer vision research and its conceivable influence on the broader development of artificial intelligence. This investigation contributes to the persistent discourse regarding the most efficacious methods for propelling machine learning and computer vision forward, furnishing perspectives that could steer forthcoming research orientations and techniques in these domains.

1 Introduction

Rich Sutton's seminal paper, "The Hard-Won Lesson," posits that the most substantial progress in artificial intelligence (AI) has resulted from concentrating on broad methods that utilize computation, as opposed to human-derived representations and knowledge. This concept has been notably apparent in Computer Vision (CV), a domain that has observed a discernible transition from manually engineered features to deep learning frameworks.

In this article, we explore the degree to which the abstracts from a prominent machine learning (ML) conference align with the principles of the "hard-won lesson" across two decades. Our analysis encompasses a randomized selection of 200 papers annually, addressing these research questions:

- How has the emphasis on generalized methodologies and computational approaches developed in major computer vision conference abstracts over the last 20 years?
- What discernible patterns can be observed regarding the embrace of deep learning methodologies and the departure from manually constructed features?
- To what degree do the abstracts mirror the primary observations of Sutton's "hard-won lesson," and how has this correlation altered over time?
- Does a substantial correlation exist between a paper's alignment with the "hard-won lesson" principles and its influence, as gauged by its citation count?

To tackle these inquiries, we utilize large language models (LLMs), themselves a clear demonstration of the principles delineated in the "hard-won lesson," to scrutinize the abstracts. This assessment hinges on five metrics assigned by the LLMs, offering a thorough evaluation of the congruence between the abstracts and the "hard-won lesson."

Our study provides valuable perspectives on the general trajectory of the ML community and uncovers intriguing patterns in the embrace of Sutton's principles. By employing LLMs to analyze a substantial

.

corpus of research literature, we introduce an innovative method for comprehending the learning and progression of a scientific field. This technique enables us to detect patterns and trends that might elude conventional research approaches, thereby delivering a more holistic understanding of the current state of ML research and its alignment with the principles demonstrated to be most effective in driving AI advancements.

The prospective influence of our conclusions on forthcoming CV research directions is considerable. By pinpointing trends in the adoption of generalized methods and deep learning techniques, we can contribute to the advancement of foundational CV models at the cutting edge. These insights enhance our comprehension of the present state of ML research and illuminate potential avenues for further investigation and expansion in the field.

2 Background

2.1 The Hard-Won Lesson

The realm of artificial intelligence (AI) has experienced a fundamental change, eloquently expressed in Rich Sutton's influential essay "The Hard-Won Lesson." Sutton's central idea underscores the importance of generalized methods that utilize computational capability over human-engineered representations and domain-specific expertise. This viewpoint resonates with Leo Breiman's earlier work, which, twenty years prior, outlined the distinction between statistical and algorithmic methods in his paper "Statistical Modeling: The Two Cultures." Breiman's insights, along with subsequent contributions, have significantly influenced our comprehension of data-oriented approaches in AI.

2.2 Evolution of Computer Vision

The discipline of Computer Vision (CV) serves as a prime illustration of the concepts articulated in Sutton's "hard-won lesson." Historically dependent on manually designed features such as SIFT, HOG, and Haar cascades for object recognition and image categorization, CV experienced a transformation with the introduction of deep learning, particularly Convolutional Neural Networks (CNNs). This shift facilitated the automated acquisition of hierarchical features directly from unprocessed image data, thereby bypassing the necessity for manual feature creation and markedly enhancing performance across a range of CV applications.

The emergence of foundational models further aligned CV with Sutton's principles. Models like CLIP, ALIGN, and Florence demonstrate remarkable adaptability across diverse tasks with minimal fine-tuning, leveraging extensive multi-modal datasets to learn rich, transferable representations.

This progression from conventional feature engineering to deep learning and foundational models in CV highlights the significance of employing computational resources and extensive datasets to achieve enhanced performance and generalization.

2.3 Large Language Models in Academic Evaluation

The incorporation of Large Language Models (LLMs) into the assessment of scholarly texts has become a notable area of focus. LLMs, like GPT-4, have shown impressive abilities in swiftly handling and examining vast quantities of data, making them appropriate for numerous uses, including the evaluation of academic papers.

Beyond their analytical abilities, LLMs have been shown to possess a degree of human-like judgment in assessing the quality of text. The G-EVAL framework, which employs LLMs to evaluate the quality of natural language generation outputs, demonstrates that LLMs can closely align with human evaluators in certain contexts. However, deploying LLMs in academic evaluation is not without its challenges. LLMs can exhibit biases similar to those found in human judgments, which may affect the fairness and accuracy of their evaluations.

The function of LLMs in responding to inquiries and formulating hypotheses also deserves consideration. Their capacity to furnish comprehensive answers to intricate queries has been utilized in diverse educational environments, enhancing learning experiences and facilitating knowledge acquisition. In the context of academic research, LLMs can aid in generating hypotheses and guiding exploratory studies, contributing to the advancement of knowledge in various fields.

Despite the promising applications of LLMs in academic evaluation and research, it is crucial to establish ethical guidelines and best practices for their use.

3 Methodology and Evaluation

3.1 LLM Evaluation of Titles and Abstracts

We utilize three large language models to assess the titles and abstracts of papers: GPT-4o-2024-05-13, gpt-4o-mini-2024-07-18, and claude-3-5-sonnet-20240620. The following details are extracted from online sources and stored in a database for each paper: Year of Publication (2005-2024), Title, Authors, and Abstract. Additionally, the citation count for each paper is obtained from the Semantic Scholar API on July 20th, 2024, and recorded alongside the other metadata.

Each LLM is assigned the task of providing a Likert score ranging from 0 to 10, indicating the degree to which a paper corresponds with the principles outlined in Sutton's "hard-won lesson." We employ the Chain-of-Thought Prompting method in conjunction with the Magentic library to interact with the models and accumulate their feedback in a structured manner for subsequent analysis.

We establish five dimensions for alignment with the "hard-won lesson":

1. **Learning Over Engineering:** How much does the idea prioritize using computation through data-driven learning and statistical methods over human-engineered knowledge and domain expertise? 2. **Search over Heuristics:** To what extent does the idea emphasize leveraging computation through search algorithms and optimization techniques instead of relying on human-designed heuristics? 3. **Scalability with Computation:** How much is the idea based on methods that can continuously scale and improve performance as computational resources increase? 4. **Generality over Specificity:** How much does the approach emphasize general, flexible methods that learn from data rather than building complex models of the world through manual engineering? 5. **Favoring Fundamental Principles:** To what extent does the approach adhere to fundamental principles of computation and information theory rather than emulating human cognition?

The prompts were crafted to encapsulate the core of each "hard-won lesson" dimension in a succinct and impartial manner. To standardize the ratings, we furnish examples for the 0, 5, and 10 points on each dimension, elucidating the standards and guaranteeing uniform evaluations.

Given the large number of publications, our research concentrates on a representative random sample of 200 papers from each year. We define the overall alignment score for each paper as the sum of scores across the five dimensions.

3.2 Inter-rater Reliability Measures

Intraclass Correlation Coefficient (ICC): We employ ICC to measure the level of agreement among the models' evaluations. ICC is especially fitting for evaluating reliability when numerous raters assess an identical set of items. Specifically, we utilize the two-way random effects model (ICC(2,k)) to consider both rater and subject influences.

Krippendorff's Alpha: In addition to ICC, we compute Krippendorff's Alpha, a flexible reliability coefficient capable of managing diverse data types (nominal, ordinal, interval, ratio) and resilient to missing data. This metric offers an supplementary viewpoint on inter-rater agreement, particularly beneficial when addressing potential variations in rating scales or absent evaluations.

3.3 Regression Analysis

To examine the connection between alignment scores and a paper's impact, we conduct a regression analysis, using citation count as an indicator of influence. To manage the publication year and address potential temporal effects, we incorporate yearly stratification into our regression model. This method enables us to isolate the influence of alignment while accounting for the differing citation patterns across various publication years.

To tackle the typically right-skewed distribution of citation counts, we employ a logarithmic transformation on the data. This transformation achieves several objectives in our analysis: it diminishes skewness, yielding a more symmetrical distribution that more closely resembles normality; it stabilizes variance across the data range, reducing the heteroscedasticity often seen in citation count data where variance tends to rise with the mean; and it linearizes potentially multiplicative relationships, converting them into additive ones.

4 Results

4.1 Inter-rater Reliability

The models show consistently strong agreement on all dimensions except "Favoring Fundamental Principles," as indicated by ICC values above 0.5 and Krippendorff's alpha scores exceeding 0.4 on the remaining dimensions. The dimension "Learning Over Engineering" exhibits the highest ICC and Krippendorff's alpha scores.

Although perfect agreement is not achieved, the inter-reliability measures fall within or above common thresholds for "good" reliability, validating the use of AI models for prompt-based research paper evaluation.

4.2 Regression Analysis

Table 1 presents the regression analysis results for each dimension of "hard-won lesson" alignment scores against citation impact, stratified by year of publication. The R-squared values range from 0.027 to 0.306.

In this regression analysis, a multiplicative effect implies that a one-unit change in the alignment score for a particular dimension leads to a proportional change in the original scale of the citation count.

The statistical significance of the regression coefficients is denoted using , , and to represent the 10%, 5%, and 1% significance levels, respectively. Several dimensions, such as "Scalability" and "Learning over engineering," exhibit statistically significant relationships with citation impact across multiple years.

Table 2 shows the results of regressing citation counts on the overall "hard-won lesson" alignment score for each year between 2005 and 2024. The R-squared values are quite low for most years but increase substantially starting in 2015.

4.3 Trends in "Hard-Won Lesson" Alignment

The dimensions of "Scalability with Computation" and "Learning Over Engineering" show a consistent upward trend over the years. The period from 2015 to 2020 witnesses a particularly sharp rise in the average scores for these dimensions.

5 Conclusion

Our study scrutinized the concordance of research with Rich Sutton's "hard-won lesson" over two decades, employing large language models to analyze trends. The results show a steady rise in the adoption of general-purpose learning algorithms and scalability with computational resources, indicating a strong adherence to the core principles of the "hard-won lesson." These trends highlight the machine learning community's inclination towards data-driven and computation-intensive methods over manual engineering and domain-specific knowledge.

However, the "Search over Heuristics" dimension has not shown a similar upward trend, suggesting limited integration of search-based methods in the field. This stagnation contrasts with recent progress in inference-time scaling, exemplified by OpenAI's o1 models, which emphasize the importance of test-time computation in overcoming diminishing returns.

The shift towards scaling inference time, driven by the development of larger and more complex models, has the potential to emulate search-like processes. As computational capabilities continue to expand, it is plausible that future research may increasingly incorporate search techniques, thereby enhancing alignment with this dimension of the "hard-won lesson."

Table 1: Regression analysis results for the relationship between "hard-won lesson" alignment scores and citation impact, stratified by year.

Year	R-squared	N	Learning	Search	Scalability	Generality	Principles
2005	0.027	199	-0.220	0.104	0.139	0.272	-0.171
2006	0.076	200	0.016	-0.042	0.388*	0.199	-0.171
2007	0.035	200	-0.087	0.117	0.350*	-0.006	-0.318*
2008	0.078	200	-0.009	0.096	0.465***	-0.026	-0.463***
2009	0.085	199	-0.073	0.136	0.104	0.378*	-0.631***
2010	0.074	200	0.121	-0.129	0.218	0.016	-0.471**
2011	0.076	200	0.208	-0.036	0.318**	-0.284	-0.423**
2012	0.094	200	0.195	0.077	0.428**	-0.110	-0.517**
2013	0.085	200	0.395***	-0.112	0.013	-0.119	-0.279
2014	0.119	200	0.408***	-0.085	0.308*	-0.348*	-0.266
2015	0.264	200	0.515***	-0.145	0.417**	-0.236	-0.122
2016	0.306	200	0.637***	-0.300**	0.517***	-0.325	-0.372*
2017	0.313	200	0.418***	-0.353**	0.751***	-0.004	-0.508**
2018	0.172	200	0.291*	-0.322*	0.418**	0.156	-0.436**
2019	0.111	200	0.573**	-0.439**	0.229	-0.099	-0.257
2020	0.120	200	0.315	-0.411***	0.179	0.229	0.010
2021	0.090	200	0.269*	-0.381***	0.253	-0.072	-0.265*
2022	0.136	200	0.618***	-0.137	0.110	-0.118	-0.257
2023	0.123	200	0.107	-0.009	0.664***	-0.078	-0.132
2024	0.178	171	-0.619***	0.314	0.808***	0.282	-0.020

*** indicates significance at the 1% level, ** indicates significance at the 5% level, and * indicates significance at the 10% level.

In summary, our findings underscore the enduring significance of the "hard-won lesson" in shaping the path of computer vision research. By emphasizing generality and scalability, the field is well-positioned to leverage emerging computational advancements. Future work should explore the integration of search methodologies and assess their impact on research impact and innovation within computer vision, particularly in light of recent breakthroughs in inference-time scaling.

6 Limitations

This study has several limitations. First, our reliance on large language models (LLMs) for evaluating research abstracts introduces potential biases inherent to these models. Second, the absence of human expert evaluation as a ground truth is a significant limitation.

Furthermore, our analysis is limited to the information contained in titles and abstracts, which may not capture the full depth and nuance of the methodologies and findings presented in the full papers. Lastly, while our study spans two decades of proceedings, it does not account for research published in other venues or unpublished work that may have influenced the field.

Despite these limitations, we believe our study provides valuable insights into broad trends in computer vision research and its alignment with the principles of the "hard-won lesson." Future work could address these limitations by incorporating human expert evaluations, analyzing full paper contents, and expanding the scope to include a wider range of publication venues.

7 Ethics Statement

This study adheres to ethical guidelines. Our use of large language models (LLMs) for analyzing trends in academic literature raises important ethical considerations. We acknowledge that LLMs may introduce biases when used for direct evaluation of academic work. However, our study focuses solely on using LLMs to analyze broad trends rather than to assess individual papers' quality or merit.

All data were collected in accordance with applicable privacy and intellectual property laws. No personally identifiable information was collected from human subjects. Our methodology aims to

Table 2: Regression analysis results for the relationship between overall "hard-won lesson" alignment scores and citation impact, stratified by year.

Year	R-squared	N	F-statistic	Prob (F-statistic)	Overall Alignment Score
2005	0.007	199	1.409	0.237	0.029 [-0.019, 0.076]
2006	0.050	200	10.335	0.002	0.083*** [0.032, 0.134]
2007	0.003	200	0.554	0.457	0.019 [-0.031, 0.068]
2008	0.010	200	1.993	0.160	0.031 [-0.012, 0.075]
2009	0.015	199	2.998	0.085	0.045* [-0.006, 0.097]
2010	0.000	200	0.033	0.856	0.005 [-0.049, 0.059]
2011	0.000	200	0.000	0.993	-0.000 [-0.051, 0.051]
2012	0.024	200	4.898	0.028	0.057** [0.006, 0.109]
2013	0.005	200	0.944	0.333	0.022 [-0.023, 0.067]
2014	0.030	200	6.023	0.015	0.056** [0.011, 0.101]
2015	0.170	200	40.618	0.000	0.141*** [0.097, 0.184]
2016	0.128	200	29.114	0.000	0.129*** [0.082, 0.176]
2017	0.133	200	30.338	0.000	0.182*** [0.117, 0.248]
2018	0.066	200	13.996	0.000	0.098*** [0.047, 0.150]
2019	0.021	200	4.241	0.041	0.061** [0.003, 0.119]
2020	0.040	200	8.325	0.004	0.079*** [0.025, 0.133]
2021	0.002	200	0.407	0.524	-0.017 [-0.068, 0.035]
2022	0.062	200	13.054	0.000	0.097*** [0.044, 0.149]
2023	0.063	200	13.416	0.000	0.099*** [0.046, 0.153]
2024	0.092	171	17.040	0.000	0.127*** [0.066, 0.188]

*** indicates significance at the 1% level, ** indicates significance at the 5% level, and * indicates significance at the 10% level.

minimize risks by using multiple models and focusing on aggregate trends rather than individual assessments.