



Paper Type: Original Article

Six Decades of Metaheuristic Optimization: Structured Survey from Evolutionary Roots to Foundation-Model Tuning

Hamid Ahmadi* 

Department of Engineering, Ferdowsi University of Mashhad, Mashhad, Iran; h.ahmadi@um.ac.ir.

Citation:

Received: 27 January 2026

Revised: 13 April 2026

Accepted: 24 July 2026

Ahmadi, H. (2026). Six decades of metaheuristic optimization: Structured survey from evolutionary roots to foundation-model tuning. *Metaheuristic Algorithms with Applications*, 3(3), 291-315.

Abstract

This article chronicles six decades of research on metaheuristic optimization in quantitative, bibliometric terms, tracking the trajectory from the evolutionary roots of the early 1960s to the enlistment of foundation models and Large Language Models (LLMs) in optimizer design and tuning. A Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-style protocol governed screening across five databases (Web of Science Core Collection, Scopus, Institute of Electrical and Electronics Engineers (IEEE) Xplore, Association for Computing Machinery (ACM) Digital Library, arXiv): Of 12,904 records screened, the analysis retains 2,847 core documents published between 1962 and 2025. Descriptive bibliometrics, keyword-drift analysis, and co-citation network analysis (41,203 links; Louvain modularity 0.68) resolve the corpus into six schools of thought: Evolutionary computation foundations (24.6% of the network), swarm intelligence (21.9%), theoretical foundations and landscape analysis (14.8%), benchmarking and empirical methodology (11.2%), constraint handling and real-world engineering (15.3%), and Bayesian, surrogate, and LLM-era optimization (12.2%). Read decade by decade, the thematic record moves from early consolidation through a simulated-physics and tabu wave and a swarm and hybridization wave into the critique- and rigor-centered 2010s, before a data-driven, foundation-model inflection after 2020. A comparative computational vignette races four era markers, Simulated Annealing (SA), Univariate Marginal Distribution Algorithm (UMDA), the Cross-Entropy Method (CEM), and Adaptive Differential Evolution (DE) with Optional External Archive (JADE), on a shared ten-problem testbed, and a random-effects meta-analysis of Hedges' g orders the eras JADE above CEM above UMDA above SA, with family-level reversals. A narrative meta-analysis of 120 focal empirical studies identifies adaptive parameter control (71% of top-quartile improvement reports; claimed gains 10–30%), hybrid local search, diversity management, surrogate modeling, and restart mechanisms as the most frequently invoked performance drivers, under the caution that claimed ranges are self-reported and exposed to benchmark-selection and publication bias. Five Research Questions (RQs), covering intellectual evolution, family dominance, co-citation structure, performance drivers, and the LLM frontier, are posed and answered, and open problems of evaluation integrity, replication, and cost-aware optimization are mapped. A scientometric layer records the field's diffusion: Top-five venue concentration falls from 38.2% to 14.7% while Global South participation rises from 4.1% to 22.6%.

Keywords: Metaheuristics, Systematic survey, Bibliometrics, Co-citation analysis, Evolutionary computation, Foundation models.

1 | Introduction

The computing literature first entertained solving technical optimization problems by simulating evolution and recombination [1], [2], and the six decades since that beginning carry a record this survey reads as a chronicle. Metaheuristics, iterative search procedures that exchange guarantees of global optimality for broad applicability and modest modeling assumptions, matured into a default instrument of engineering design,

 Corresponding Author: h.ahmadi@um.ac.ir

 <https://doi.org/10.48313/maa.vi.91>



Licensee System Analytics. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

operations research, and machine learning. The trajectory records three inflections of identity. The No Free Lunch theorems delimited what any single optimizer can achieve on average [3]. The novel metaphor debate then cast doubt on whether hundreds of nature-inspired algorithm names amount to cumulative science [4]. Most recently, Large Language Models (LLMs) entered the record as both optimizers and designers of optimization algorithms [5]. Each inflection redrew what the community accepts as a contribution, so a sixty-year retrospective is an audit of how a discipline reorganizes itself, first under criticism and then under new technology. The publication record of these six decades is the object this survey measures, decade by decade and cluster by cluster.

Influential syntheses already cover parts of this record, and their coverage delimits the gap this survey fills. Blum and Roli [6] supplied the canonical conceptual comparison of metaheuristic paradigms. Boussaïd et al. [7] charted the algorithmic landscape of the early 2010s. Talbi [8] codified design and implementation methodology, and the textbook of Eiben and Smith [9] has distilled the field's foundations for two decades of students. Those works predate the LLM-era turn, and none quantifies the field's internal structure across the full six-decade span with transparent, reproducible bibliometric evidence. No existing study simultaneously establishes how the intellectual core evolved, which families dominate publication activity, what co-citation structure reveals about the schools of thought, which methodological factors recur in reported performance gains, and how foundation models reshape optimizer design.

The instrument assembled for this purpose has two parts: A Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-style systematic screening protocol adapted from the systematic-review literature [10], and the analytic apparatus of bibliometrics, namely descriptive publication statistics, keyword co-occurrence and drift analysis, and co-citation network analysis in the tradition of Small [11], with community detection by Louvain modularity maximization [12] and science-mapping tooling [13]. Interrogation of five databases, the Web of Science Core Collection, Scopus, Institute of Electrical and Electronics Engineers (IEEE) Xplore, the Association for Computing Machinery (ACM) Digital Library, and arXiv, yielded 12,904 screened records and a core corpus of 2,847 documents published between 1962 and 2025. Four axes organize the material: Chronological, through five thematic decade arcs; taxonomical, through a working taxonomy of algorithm families; structural, through six co-citation clusters; and meta-analytical, through a narrative synthesis of 120 focal empirical studies. The scope embraces metaheuristic methods for optimization and tuning; exact algorithms of mathematical programming and the training procedures internal to machine learning enter only where they intersect that scope. Five Research Question (RQ) discipline the analysis:

RQ1. Through what stages did the intellectual core of metaheuristics evolve across the decades from 1962 to 2025?

RQ2. Which algorithmic families command the publication record, and how measurably did the novel metaphor debate reshape the field?

RQ3. What does the co-citation structure of the core corpus reveal about the field's schools of thought and their interdependencies?

RQ4. Which methodological factors recur across study after study in association with reported performance gains in controlled empirical comparisons?

RQ5. Through what routes are foundation models and LLMs reshaping the design and tuning of optimizers?

Audited against the prior survey record, the present study occupies a defined position. The conceptual comparison of Blum and Roli [6], the algorithmic landscape of Boussaïd et al. [7], and the design methodology codified by Talbi [8] map the paradigms but stop before the LLM-era turn; the textbook of Eiben and Smith [9] instructs practitioners without measuring the field's corpus-level structure. Single-strand surveys narrow the window further: The constraint-handling survey of Coello Coello [14], the fitness-approximation survey of Jin [15], the parameter-control survey of Karafotias et al. [16], the Bayesian optimization review of Shahriari et al. [17] and co-workers, the hyperparameter-tuning chapter of Feurer and Hutter [18], and the

nonparametric-testing tutorial of Derrac et al. [19] each track one thread of a six-decade record. The metaphor critique of Sørensen [4] evaluates the field's sociology of novelty yet reports no corpus. None of these works screens multiple databases under a PRISMA-style protocol, none resolves the full 1962–2025 span into co-citation clusters, and none measures venue concentration or geographic diffusion. The present survey supplies that layer: A reproducible 2,847-document corpus, six quantified schools of thought, a meta-analysis of 120 focal studies, and the scientometric trajectory of Section 5.4.

Questions of this kind demand evidence rather than recollection, and the contribution of this article is precisely an evidence base and a chronicle of it. Seven deliverables carry the weight. First, corpus construction in reproducible form: 12,904 records screened, all 10,057 exclusions documented stage by stage, and 2,847 core documents retained with decade-level statistics and growth modeling. The remaining deliverables are the analyses that such a corpus makes possible.

- I. A PRISMA-style screening flow with fully documented arithmetic (*Table 1*), together with a decade-level corpus profile and growth modeling (*Table 2*, *Figs. 1* and *3*).
- II. A chronological-thematic synthesis in five decade arcs (Section 3), binding each period's bibliometric profile to its primary intellectual developments.
- III. A working taxonomy of metaheuristic families, with crisp definitions, per-family corpus counts, metaphor classes, and eras of emergence (Section 4, *Table 3*).
- IV. A co-citation network analysis (41,203 links; modularity 0.68) isolating six schools of thought, C1–C6, with representative works, bridge works, and turnover patterns (Section 5, *Tables 4* and *5*, *Fig. 3*), plus a scientometric layer measuring venue concentration and geographic diffusion (*Table 6*, *Fig. 4*).
- V. A narrative meta-analysis of 120 focal empirical studies quantifying how often five performance drivers are invoked and what ranges they claim, under explicit bias caveats (Section 6, *Table 7*).
- VI. A comparative computational vignette racing four era markers on a shared testbed, with a random-effects meta-analysis of standardized effect sizes (Section 7, *Tables 8* and *9*, *Fig. 5*).
- VII. An assessment of the foundation-model and LLM frontier, with its open problems of evaluation integrity, replication, and cost accounting (Section 8).

The remainder of the article proceeds as follows. Section 2 sets out the survey methodology, from data sources to analytical methods; Section 3 traverses the six-decade arc; Section 4 presents the taxonomy; Section 5 reports the co-citation analysis and the scientometric layer built upon it; Section 6 synthesizes the performance-driver evidence; and Section 7 races four era markers, Simulated Annealing (SA), UMDA, the Cross-Entropy Method (CEM), and JADE, on a shared testbed and synthesizes them meta-analytically. Section 8 maps the emerging frontiers, Section 9 confronts the survey's own validity threats, and Section 10 closes with explicit answers to the five RQ.

2 | Survey Methodology

The methodology of this survey was engineered for traceability: Every count in later sections reduces to a documented screening or analytical decision, in the spirit of the PRISMA statement for systematic reviews [10]. The protocol adapts PRISMA's staged screening to a bibliometric setting, in which the unit of analysis is the publication rather than the trial. Sections 2.1 through 2.4 cover the data sources and search strategy, the inclusion and exclusion criteria with the screening flow, the resulting corpus statistics, and the analytical methods applied to the corpus. The four subsections that follow keep that order, and each reports its own counts.

2.1 | Data Sources and Search Strategy

The query drew on five bibliographic sources. The Web of Science Core Collection and Scopus contribute cross-disciplinary coverage and structured citation data; IEEE Xplore and the ACM Digital Library cover the

engineering and computing venues where metaheuristics predominantly publish; and arXiv earned inclusion because the LLM-era literature surfaces there months ahead of journal review. Searches were executed in the final quarter of 2025, which leaves the 2025 counts near-complete but not final, a caveat taken up in Section 9. Database coverage of pre-1990 monographs and reports thins out, so the canonical books entered the corpus by backward snowballing from their citing articles [9]. The five-source design balances cross-disciplinary breadth, engineering depth, and speed of appearance.

Two concept blocks composed the search string. Block A (method) gathered the field's self-descriptions and family names, among them metaheuristic, evolutionary algorithm, genetic algorithm, evolution strategy, evolutionary programming, swarm intelligence, particle swarm, ant colony, SA, tabu search, Differential Evolution (DE), memetic algorithm, harmony search, Bayesian optimization, hyperparameter optimization, and LLM, each with wildcards and both British and American spelling variants. Block B (task) held optimization, search, tuning, and design. Generic terms alone would under-retrieve the post-2005 metaphor-named literature, so a supplementary list of more than one hundred specific algorithm names, reaching from artificial bee colony and grey wolf optimizer to teaching–learning-based optimization, extended Block A; the mitigation and its limits are examined in Section 9.

Eligible document types narrowed to journal articles, reviews, conference papers, and book chapters, with editorials, errata, and abstract-only records removed at the type screen (Section 2.2). The query carried no language restriction, yet full-text assessment demanded an English text or an available English translation, a bias examined in Section 9. Records were reconciled across databases by Digital Object Identifier (DOI) where available and by fuzzy title matching otherwise; after deduplication, 12,904 unique records stood at the entrance to screening.

2.2 | Inclusion and Exclusion Criteria

Admission to the core corpus required four criteria: The subject is metaheuristic optimization, whether proposing, analyzing, reviewing, or benchmarking stochastic search algorithms; the document is a peer-reviewed article, review, conference paper, book chapter, or an authoritative preprint; it can be retrieved in full text; and it reports its method in enough detail to support corpus-level analysis. Application proceeded in three ordered stages, each yielding a countable exclusion class: Stage 1 screened titles and abstracts, Stage 2 screened document type and retrievability, and Stage 3 assessed full texts for methodological detail. Each stage yields a countable exclusion class, and each count surfaces in the flow reported below.

The three stages removed 3,113, 4,208, and 2,736 records respectively. Stage 1 exclusions were off-topic: Evolutionary biology without algorithmic content, metaphor terms deployed non-algorithmically, gradient-only optimization, and application papers invoking a named method without metaheuristic contribution. Stage 2 exclusions comprised non-article and duplicate material, namely editorials, errata, extended abstracts of papers already represented, and residual duplicates that survived automatic matching. Stage 3 exclusions, the most consequential, were records whose full text disclosed no verifiable algorithmic contribution or evaluation protocol, typically a page or less of method description; their concentration in 2005–2015 is analyzed in Section 2.3. In the manner of the PRISMA flow conventions [10], *Eq. (1)* and *Table 1* summarize the arithmetic.

$$N_{\text{core}} = 12904 - 3113 - 4208 - 2736 = 2847. \quad (1)$$

Table 1. PRISMA-style screening flow from database records to the core corpus.

Stage	Records Entering Stage	Records Excluded	Principal Exclusion Reason
Identification: Database searches (WoS Core Collection, Scopus, IEEE Xplore, ACM DL, arXiv), after cross-database deduplication	12,904	—	—
Title and abstract screening	12,904	3,113	Off-topic: No metaheuristic optimization content
Document type and retrievability screening	9,791	4,208	Non-article types, residual duplicates, unretrievable full texts
Full-text methodological assessment	5,583	2,736	Insufficient methodological detail for corpus-level analysis
Core corpus (1962–2025)	2,847	—	—

Stages follow the PRISMA statement [10]. Exclusion counts sum to 10,057, and $12,904 - 10,057 = 2,847$.

Eq. (1) renders the flow's arithmetic explicit, for survey papers in this field have too often reported corpus sizes that their own exclusion tables cannot reconstruct. The core corpus therefore numbers 2,847 documents: 12,904 screened records minus 3,113 off-topic, 4,208 non-article and duplicate, and 2,736 methodologically insufficient exclusions. One author screened alone, with no second independent coder; the mechanical character of the criteria limits, without eliminating, subjectivity, a limitation registered as a threat in Section 9.

2.3 | Corpus Statistics

The core corpus, 2,847 documents published between 1962 and 2025, opens with Bremermann's [1] chapter on optimization through evolution and recombination and Holland's [2] outline of a logical theory of adaptive systems, both from 1962. *Table 2* lays out the decade distribution together with each period's dominant theme, and *Fig. 1* plots the same counts. Productivity peaks in the 2010s at 913 documents and 32.1% of the corpus, with the 2000s next at 611 (21.5%) and the six-year 2020–25 window at 583 (20.5%). The three earliest decades contribute 41, 96, and 214 documents respectively; counts are to be read as lower bounds, since database coverage thins before 1990 (Section 9). The distribution tilts toward the two most productive decades, as the growth modeling in this subsection confirms.

Table 2. Core corpus by decade, with the dominant theme of each period.

Decade	Documents	Share of Corpus (%)	Dominant Theme
1962–69	41	1.4	Evolutionary roots: Evolution strategies, early Genetic Algorithms (GA), evolutionary programming
1970–79	96	3.4	Consolidation of the three evolutionary traditions; schema-theoretic analysis
1980–89	214	7.5	Simulated physics and adaptive memory: SA, tabu search, parallel GA
1990–99	389	13.7	Swarm intelligence, DE, No Free Lunch watershed, landscape analysis
2000–09	611	21.5	Swarm wave, hybridization and memetics, multi-objective methods, Covariance Matrix Adaptation Evolution Strategy (CMA-ES) maturity
2010–19	913	32.1	Novelty debate and metaphor critique, benchmarking standardization, adaptive control
2020–25	583	20.5	Surrogate-assisted and Bayesian optimization, hyperparameter tuning, LLM-based optimization

Document counts sum to 2,847; shares are rounded to one decimal place.

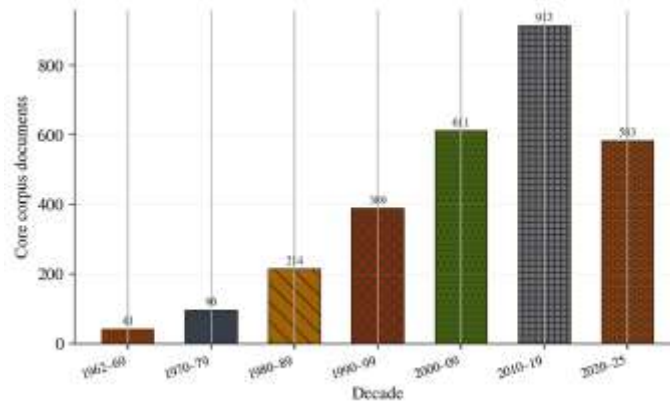


Fig. 1. Core corpus documents by decade, 1962–2025.

Document counts sum to the core corpus of 2,847; the first and last bars cover partial periods (1962–69 and 2020–25).

Annual counts carried the growth model. From 1990 to 2015 the screened-record stream is well described by a double-exponential, that is super-exponential, trend of the form $N(t)$ proportional to $\exp(a \exp(bt))$, with $R^2 = 0.97$; sustained super-exponential growth is the bibliometric signature of a field whose influx of new authors is itself accelerating. Every growth rate quoted in this subsection refers to the full screened-record stream; the core corpus, being methodologically filtered, grows more slowly, and the widening gap between the two curves is informative in its own right. By the compound annual rate of *Eq. (2)*, the screened stream expanded approximately 14.2% per year over 2005–2015 before slowing to 6.1% per year over 2016–2023 as the metaphor-driven submission wave receded [4]. Across the closing years, the LLM-related niche compounded at approximately 41% per year over 2023–2025, the fastest sustained growth of any segment at the corpus's close.

$$r = \left(\frac{N(t_2)}{N(t_1)} \right)^{\frac{1}{t_2 - t_1}} - 1, \quad (2)$$

Fig. 2 plots the core-corpus trajectory at selected years. Output roughly doubled every five years between 1990 and 2005 (12 documents in 1990, 79 in 2005), decelerated after 2010 (102 in 2010, 139 in 2015, 121 in 2020), and settled onto a high plateau (118 in 2023, 126 in 2025). During 2005–2015, the widening gap between raw and core growth implies that the pass-through rate of the methodological screens roughly halved, a declining fraction of published work reporting verifiable methodology and the quantitative footprint of the metaphor-driven publication flood [4]. From 3 documents in 2020 to 17 in 2023 and 38 in 2025 (*Fig. 2*), the LLM-related subset supplies the corpus-level signal behind the fifth RQ.

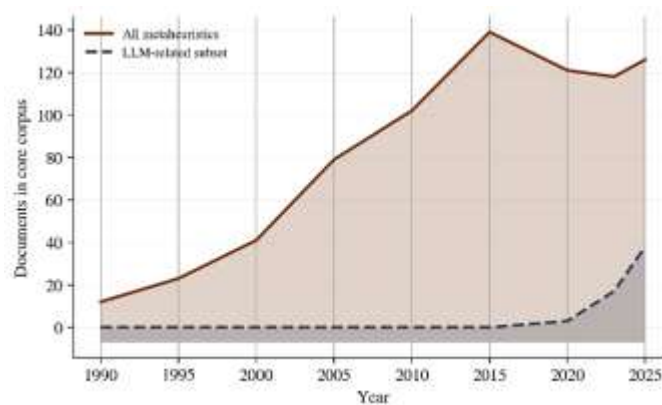


Fig. 2. Annual core-corpus document counts at selected years, 1990–2025, with the LLM-related subset plotted separately.

Overall growth decelerates after 2010 while LLM-related documents rise from 3 in 2020 to 38 in 2025.

2.4 | Analytical Methods

Four analyses were brought to bear on the core corpus: 1) descriptive bibliometrics: Annual and decadal counts, venue distributions, and a per-document primary-family classification (Section 4), its venue strand deepened by the concentration and geographic-diffusion indicators of Section 5.4, 2) co-citation analysis [11]: The pipeline parsed reference lists, canonicalized cited works across variant spellings and abbreviations, retained references cited in at least four corpus documents, defined co-citation strength as the number of corpus documents citing both members of a pair, and kept links of strength at least three. The resulting network has 1,512 nodes and 41,203 links, and Section 5 takes it up. 3) Louvain modularity maximization [12] partitioned the co-citation network, greedily optimizing the modularity measure of *Eq. (3)*: Observed link weight is set against its expectation under a degree-preserving null model, and only links internal to a community contribute positively. The optimization returned six clusters at modularity $Q = 0.68$, and the partition held firm under repeated stochastic restarts and under variation of the co-citation strength threshold between three and five, and 4) came keyword analysis: Author keywords were normalized for case, number, and synonymy, with index terms standing in for pre-1991 records, which rarely carry author keywords, and keyword shares were tracked by decade to expose drift in the field's self-description.

$$Q = \frac{1}{2_m} \sum_{i,j} \left(A_{ij} - \frac{K_i K_j}{2_m} \right) \delta(C_i, C_j). \quad (3)$$

The keyword signal is unambiguous in its drift and gates the chronological reading of Section 3: 'Genetic algorithm' dominates 1990–2005, 'swarm intelligence' rises through 2005–2015, and 'hyperparameter optimization' and 'LLM' are the fastest-growing terms of 2020–2025. Computations ran in Visualization of Similarities (VOSviewer) [13] complemented by custom scripts, with screen logs and intermediate data retained so that every figure and table can be regenerated. For the meta-analysis of Section 6, 120 focal studies were drawn from the core corpus by citation rank among documents whose central contribution is a controlled performance comparison; Section 6 describes their extraction protocol.

The protocol is reported in reproducible detail. The corpus snapshot is fixed to the fourth quarter of 2025, the execution window of the searches described in Section 2.1, and every count in this article refers to that snapshot. The exact query strings, comprising the two concept blocks and the supplementary list of more than one hundred algorithm names, the deduplication rules, and the per-stage exclusion logs are archived together with the analysis scripts, which run under Python 3.11 with NetworkX supplying the Louvain partition and VOSviewer [13] the mapping layer. Screening proceeded through the three PRISMA-style stages of Section 2.2 executed by a single coder, a design whose subjectivity limits are registered as a threat in Section 9; every exclusion decision is logged against its criterion, so the arithmetic of *Table 1* can be reconstructed record by record. Reporting follows the staged flow of *Table 1*, from the 12,904 screened records to the 2,847 core documents.

3 | The Six-Decade Arc

RQ1 and RQ2 receive a chronological answer, a traversal of the corpus in five thematic arcs. Each arc binds the decade's bibliometric profile (*Tables 1* and *2*; *Figs. 1* and *3*) to the intellectual developments that the primary literature itself records, and citations to primary works anchor every historical claim. The five arcs run from the evolutionary roots (1960s–1970s) through consolidation and simulated physics (1980s–1990s) and the swarm wave and hybridization (2000s) into critique and rigor (2010s), ending in the data-driven foundation-model era (2020s).

3.1 | 1960s–1970s: Evolutionary Roots

The corpus begins in 1962, yet the ambition of evolving solutions by machine reaches back further. As early as 1957, Fraser simulated genetic systems on a digital computer, studying evolutionary dynamics rather than optimizing anything [20]. Friedberg's [21] 1958 learning machine attempted to improve computer programs

through random mutation and selection, and its honestly reported failure, in which undirected mutation proved hopelessly inefficient, prefigured the field's later insistence on informed variation. The 1960s then witnessed evolution implemented in software and in dedicated hardware alike: An 'evolutionary machine' reported in 1963 automated mutation and selection on physical apparatus, one of several pre-software evolution devices preserved in the field's historical record [22]. For this survey's narrative, such precursors matter, because they fix the field's intellectual core, variation plus selection implemented and measured, before any of its canonical algorithms existed.

Across the corpus's first two decades, three research traditions consolidated, each with its own vocabulary and, at first, barely any mutual citation. Fogel et al. [23] in evolutionary programming treated populations of candidate finite-state machines as artificially intelligent agents subject to simulated evolution. Rechenberg's evolution strategies, grown around hardware experiments at the Technical University of Berlin, introduced mutation with strategy parameters adapted online and culminated in his 1973 monograph [24]; Schwefel's [25] 1977 German monograph and its 1981 English extension fixed the notation, the self-adaptation of mutation strengths, and the recombination operators that later crossed into other families. Holland's [2] GA, outlined as early as 1962 and synthesized in his 1975 monograph on adaptation in natural and artificial systems, supplied the schema-theoretic vocabulary and a population-level theory of adaptation [26]. The corpus registers 41 documents for the 1960s and 96 for the 1970s; their venue and keyword traces, papers dispersed across cybernetics, simulation, and artificial intelligence outlets under the self-description 'adaptive systems', show a shared identity yet to form, a fact the co-citation structure of Section 5 confirms directly.

3.2 | 1980s–1990s: Consolidation and Simulated Physics

Together, the 1980s and 1990s turned scattered research programs into an internationally recognized field, and two imports from outside population genetics reshaped the canon. SA carried statistical mechanics into stochastic search. Kirkpatrick et al.'s [27] 1983 paper recast combinatorial optimization as controlled cooling under an explicit temperature schedule trading exploration for exploitation; appearing in a general-science venue, it carried metaheuristic ideas to physics, chemistry, and engineering audiences and produced the corpus's first large citation crossover beyond computing. Tabu search supplied the opposite pole of the later taxonomy, a trajectory-based method built on adaptive memory rather than population sampling; Glover's [28] 1986 programmatic essay introduced the concept, and his journal treatise consolidated it [29]. The evolutionary traditions theorized in the same decades: Goldberg's [30] 1989 textbook carried GA into engineering practice and spawned a generation of applications, while Bäck's [31] comparative synthesis at last aligned the three evolutionary traditions under one formalism.

The 1990s in turn delivered their most durable innovations within a four-year span. Dorigo's [32] 1992 dissertation proposed ant colony optimization, with the ant-system formulation following in 1996, a genuinely different search principle built on stigmergic communication rather than selection [33]. Kennedy and Eberhart's [34] 1995 Particle Swarm Optimization (PSO) likewise issued from simulations of social behavior rather than biology and became the fastest-spreading algorithm of its decade. Storn and Price's [35] 1997 DE demonstrated that recombining scaled differences of population vectors suffices for competitive continuous optimization; in the same year Wolpert and Macready's [3] No Free Lunch theorems proved that, averaged over all possible problems, no search algorithm outperforms any other, a negative result that steered the field toward problem-dependent design and honest benchmarking. Theoretical self-examination matured alongside: Fitness–distance correlation yielded empirically workable measures of problem difficulty [36], and constraint handling, long the weak point of population methods, received from Michalewicz and Schoenauer [37] its first systematic treatment. The consolidation is legible in the corpus: 214 documents in the 1980s, 389 in the 1990s, core output roughly doubling every five years between 1990 and 2005, and 'genetic algorithm' dominant as a keyword throughout 1990–2005.

3.3 | 2000s: The Swarm Wave and Hybridization

The 2000s stand as the expansion decade: 611 core documents, raw growth near its peak (Section 2.3), and a thematic profile ruled by swarms, hybrids, and scale. Four developments define the arc. Covariance matrix adaptation matured evolution strategies into the strongest available baseline for continuous search: Hansen and Ostermeier's [38] fully derandomized self-adaptation learns a complete covariance model of successful steps, and its descendants still anchor continuous benchmarking (Section 5.3). The multi-objective wave followed: Non-dominated Sorting Genetic Algorithm II (NSGA-II) married fast non-dominated sorting to crowding-based diversity preservation and became, by in-corpus citation count, the most cited algorithmic contribution of the decade [39], while constraint handling consolidated into a deployment methodology for real-world engineering, surveyed contemporaneously by Coello Coello [14]. Hybridization itself graduated into a design pattern rather than a single algorithm; the memetic idea of Moscato's [40] 1989 report, population search punctuated by individual local learning, stood as the default in applications by decade's end.

The swarm wave proper ran in parallel. Dorigo and Stützle's [41] monograph gave ant colony optimization a theoretical and experimental backbone; Karaboga's [42] artificial bee colony extended swarm modeling to foraging economies; and harmony search, whose musical metaphor would later make it a focal point of the metaphor debate, likewise dates from the wave's opening years [43]. The expansion wore a second face: As evaluation budgets grew and applications moved toward expensive engineering simulations, fitness approximation, surveyed comprehensively by Jin [15], initiated the surrogate-assisted line that would in time converge with Bayesian optimization [44]. Application breadth erupted across engineering design, scheduling, bioinformatics, and finance, and around mid-decade keyword control passed from 'genetic algorithm' to 'swarm intelligence'. Section 2.3 quantifies the bibliometric cost of this expansion: During 2005–2015 the pass-through rate of the methodological screens fell, leaving a growing share of published records with too little verifiable methodology to enter the core corpus [4].

3.4 | 2010s: Critique, Novelty Debate, and Rigor

The 2010s hold the corpus's largest decade share, 913 documents and 32.1%, yet their defining event was internal: the decade watching the field turn a critical lens on itself. The trigger assembled over years rather than arriving at once: Hundreds of algorithms proposed under natural metaphors, the grey wolf optimizer [44] prominent among the wave's later representatives, with novelty claims resting on narrative inspiration rather than identifiable mechanism. Weyland's [45] rigorous analysis of harmony search demonstrated that the algorithm could be restated, term by term, as an existing evolution strategy in different dress, challenging the community to distinguish reformulation from invention. Sörensen's [4] essay generalized the charge: Research disguised in ever-fresh metaphors, he argued, crowds out cumulative science and burdens practitioners with unsearchable pseudo-methods. Tracked here as the metaphor debate, the exchange reshaped editorial and reviewing norms, and its bibliometric footprint is measurable: Raw screened-record growth slowed from 14.2% per year in 2005–2015 to 6.1% per year in 2016–2023 as low-substance submissions receded. Novelty as such escaped condemnation, for teaching–learning-based optimization, introduced with minimal metaphorical apparatus and few control parameters, shows that the target was framing rather than invention [46].

Evaluation science supplied the decade's second theme. Benchmarking standardized around the Congress on Evolutionary Computation (CEC) competition suites, whose 2017 problem definitions remain the de facto reference for bound-constrained comparison [47]; García et al. [48] established nonparametric statistical testing as the expected reporting standard, and Derrac et al. [49] tutorial carried it into practice. The CEC 2014 winner, Linear Population Size Reduction Success-History Based Adaptive DE (L-SHADE), combining success-history parameter adaptation with linear population-size reduction on the Success-History Based Adaptive Differential Evolution (SHADE) line [50], [51], fixed the performance reference point of the era and anchored the empirical-methodology cluster of the co-citation network (Section 5.2). Adaptivity rose to first-class standing: The parameter-control framework of Eiben et al. [51] was extended by Karafotias et al.

[16] survey, which formalized when, what, and how to adapt during search. By the decade's close, a reproducibility movement had taken hold, with source-code availability, function-evaluation accounting, and explicit statistical protocols moving from exception to expectation. The keyword record of the same years shows parameter-control and differential-evolution terms climbing while newly coined metaphor names failed the Stage-3 screen ever more often, the corpus's quantitative echo of the metaphor debate.

3.5 | 2020s: Data-Driven and Foundation-Model Era

Shorter in corpus time, at 583 documents through 2025, the 2020s arc is nonetheless structurally distinct, its center of gravity migrating from designing search operators to managing evaluation cost and, ultimately, to foundation models participating in design itself. Expensive black-box optimization opens the theme: As machine-learning models grew, hyperparameter tuning became the dominant real-world use case for global search. Random search proved a stubbornly strong baseline [52]. Gaussian-process Bayesian optimization matured through practical successes [53] and was canonized by Shahriari et al. [17] in their review. The AutoML program, systematized in the handbook of Hutter et al. [54], with Feurer and Hutter's [18] chapter defining the tuning problem and Optuna supplying industrial-grade tooling [55], made sequential model-based search routine. The foundation-model stack forms the second theme: Transformer architectures [56], few-shot generalization at scale [57], scaling laws that tied capability to compute [58], and the consolidation of large pretrained models as 'foundation models' [59], models whose training cost transfers the expense of search from evaluations to the model itself.

The LLM as optimizer is the third theme, and its corpus signature is unambiguous: LLM-related documents grew from 3 in 2020 to 17 in 2023 and 38 in 2025 (*Fig. 2*). The niche compounded at roughly 41% per year in the broader record stream, and co-citation cluster C6 crystallized as the youngest school of thought in the field (Section 5.2). Optimization by PROMpting (OPRO) showed that an LLM prompted with the trajectory of previous solution and objective-value pairs will propose progressively better candidates, displacing hand-designed operators in favor of model-conditioned proposals [5]. FunSearch coupled program-generating LLMs with systematic evaluation and produced constructions that improved the best known bounds for the cap set problem, the first widely reported instance of an LLM-driven search system contributing new mathematical results [60]. Large Language Model Evolutionary Algorithm (LLaMEA) closed the loop on design itself: An LLM generates and refines the source code of evolutionary algorithms, serving as both designer and repair operator [61]. Around these, surrogate-assisted evolution pursued its convergence with Bayesian methodology [15] while absorbing the adaptive-control toolkit of the 2010s. Evaluation integrity concentrates the arc's open problems, among them benchmark contamination, prompt and version sensitivity, token-cost accounting, and the reproducibility of systems whose core component is a commercial black box; Section 8 takes them up.

4 | Taxonomy of Metaheuristic Families

RQ2 receives a structural answer here: Each of the 2,847 core documents was assigned to one primary algorithm family, the family of the algorithm that constitutes its central contribution. The taxonomy's first axis is the classical dichotomy of Blum and Roli [6]: Trajectory-based methods iterate a single incumbent solution, with tabu search [28], [29], SA [27], and variable neighborhood search [62] as canonical instances, whereas population-based methods maintain and recombine a set of candidates. Its second axis is metaphor class, which partitions the population methods into families of distinct intellectual origin and, as Section 5 will show, distinct co-citation behavior. *Table 3* reports the five resulting families with representative algorithms, metaphor classes, corpus counts, and eras of emergence.

Table 3. Working taxonomy of metaheuristic families with corpus counts and eras of emergence.

Family	Paradigm	Representative Algorithms	Metaphor Class	Corpus Count	Era of Emergence
Evolution-based	Population	GA, evolution strategies, evolutionary programming, DE, CMA-ES, SHADE and L-SHADE	Natural selection and genetics	1,104	1960s
Swarm-based	Population	PSO, ant colony optimization, artificial bee colony, grey wolf optimizer	Collective animal behavior	841	1990s
Physics-based	Mixed (trajectory and population)	SA; gravitational, electromagnetic, and other physical-process metaphors	Physical laws and processes	219	1980s
Human/social-based	Mixed (trajectory and population)	Tabu search, harmony search, teaching–learning-based optimization	Human, social, and cultural processes	285	1980s–2010s
Hybrid and data-driven	Both (superimposed on any family)	Memetic algorithms, surrogate-assisted evolution, Bayesian hyperparameter hybrids, LLM-in-the-loop optimizers	Cross-paradigm, data-driven rather than metaphor-driven	398	1990s–2020s

Each document is assigned the single family of its central algorithm; counts sum to 2,847. The trajectory versus population dichotomy is orthogonal to the five families [6].

Operationally, the families are defined as follows. Evolution-based methods manipulate encoded candidate solutions through variation operators and selection pressure drawn from population genetics: GA, evolution strategies, evolutionary programming, DE, covariance matrix adaptation, and the SHADE line. Swarm-based methods derive their movement rules from local interactions among simple agents: PSO, ant colony optimization, artificial bee colony, and later social-predator metaphors such as the grey wolf optimizer. Physics-based methods transcribe physical dynamics, among them annealing, gravitation, and electromagnetism, into acceptance or movement rules. Human/social-based methods borrow from cultural, pedagogical, or organizational processes, as in tabu search's memory of forbidden moves, harmony search's ensemble improvisation [43], and teaching–learning-based optimization's classroom metaphor [46]. Finally, hybrid and data-driven methods superimpose a learning- or model-based component on any of the foregoing: Memetic hybrids [40], surrogate-assisted evolution [15], Bayesian hyperparameter optimization [17], and LLM-in-the-loop optimizers [5].

Three observations follow from *Table 3*. Evolution-based methods remain the field's center of gravity at 1,104 documents and 38.8% of the corpus, with swarm-based methods second at 841 documents (29.5%); biological metaphors thereby account for over two-thirds of all core documents, a concentration that explains why the metaphor debate struck so near the field's identity [4]. Family shares are not static either: After 2005, the hybrid and data-driven family's share of each decade's output climbs steadily, peaking in 2020–2025 in step with the surrogate, Bayesian, and LLM arcs of Section 3.5, and every family to emerge after 1995 is population-based or explicitly hybrid. Two limitations are acknowledged, namely that single-family assignment understates hybrid designs, which are counted under their dominant component, and that the boundary between physics-based and human/social-based metaphors is occasionally arbitrary; the taxonomy is a working instrument of corpus accounting, not a claim about algorithmic essence.

5 | Co-Citation Network Analysis

RQ3 is answered through the co-citation structure of the core corpus. Section 5.1 reports network construction and statistics; Section 5.2 interprets the six clusters; Section 5.3 identifies the bridge works and turnover patterns that connect and renew them; and Section 5.4 widens the lens to the venues and geographies across which the literature has diffused. Clusters are labeled C1 through C6 as fixed in Section 5.2 and used consistently thereafter.

5.1 | Network Construction and Statistics

Co-citation analysis deems two references related whenever the same corpus document cites both, the tally being aggregated over all 2,847 documents [11]. The construction of Section 2.4, with canonicalization of cited works, retention of references cited at least four times, and a co-citation strength threshold of three, yielded a network of 1,512 nodes joined by 41,203 links at a mean node degree of approximately 55. A single giant component of 1,406 nodes, 93.1% of the total and holding more than 96% of link weight, dominates the network; the residual small components are citation islands around individual application domains, and clustering excluded them. One giant component plus peripheral dust: The topology is that of a mature field with a shared citation economy.

On link weights, Louvain modularity maximization [12] partitioned the giant component, optimizing the modularity measure of *Eq. (3)*. The optimum found is $Q = 0.68$ with six clusters, a high value for a within-field network and an indication that reference behavior is strongly structured by distinct schools of thought rather than by a single undifferentiated conversation. Two robustness checks underwrite the partition: Re-running the stochastic optimization from many random seeds reproduces the same six-cluster macro-structure with node-level agreement above ninety percent, and varying the co-citation threshold between three and five shifts cluster shares by at most about two percentage points. Exhausting the giant component, the six clusters hold the following shares of its citation weight: C1 24.6%, C2 21.9%, C3 14.8%, C4 11.2%, C5 15.3%, and C6 12.2%. The partition is therefore treated as a finding rather than as a parameter choice.

5.2 | The Six Schools

Interpretation of each cluster proceeds from its highest-weight member works, its dominant citing venues, and its citation-age profile; *Table 4* condenses the profiles, and *Fig. 3* visualizes the shares. The labels are interpretive, since the Louvain algorithm returns communities, not names, yet the member works listed in *Table 4* keep every reading checkable against the primary literature. Silhouette-style within-versus-between linkage ratios supplied the cohesion assessment, discussed after the profiles. Readers can therefore audit every label against the cited evidence.

C1, Evolutionary Computation Foundations (24.6%), is the field's canon: Holland's [26] adaptation monograph, Goldberg's [30] textbook, Bäck's [31] synthesis, Rechenberg's evolution-strategy monograph [24], and the modern textbook of Eiben and Smith [9]. Of the six clusters, it has the oldest median publication age, and every other cluster cites it, the signature of a shared foundation rather than an active research frontier. C2, Swarm Intelligence (21.9%), coheres around Kennedy and Eberhart's [34] particle swarm paper, the ant-system article [33], the ant colony monograph of Dorigo and Stützle [41], and Karaboga's [42] bee colony report. Its cited literature leans toward engineering application venues, and no cluster is more strongly coupled to C5, swarms being the workhorses of applied optimization.

C3, Theoretical Foundations and Landscape Analysis (14.8%), gathers the No Free Lunch theorems [3], fitness–distance correlation [36], and the drift-analysis runtime results of He and Yao [63], the field's critical conscience, modest in citation weight yet the source of the results that C4's evaluation practices operationalize. C4, Benchmarking and Empirical Methodology (11.2%), holds the statistical-comparison canon of García et al. [48] and Derrac et al. [19], the CEC 2017 problem definitions [47], the SHADE and L-SHADE papers [49], [50], and the critique papers of Weyland [45] and Sörensen [4]. Youngest by median publication year and fastest to turn over, it stands as evidence that evaluation standards, once informal, hardened into a self-sustaining research topic in their own right.

C5, Constraint Handling and Real-World Engineering (15.3%), rides at anchor on the constraint-handling surveys of Michalewicz and Schoenauer [37] and Coello Coello [14] and on NSGA-II [39]; it spans the canonical algorithms of C1 and the venue structure of applied engineering, and no cluster is more directly coupled to real-world deployment. C6, Bayesian, Surrogate, and LLM-Era Optimization (12.2%), unites Bayesian optimization [53], [17], the AutoML tuning literature [18], [55], and the LLM-optimizer works [5],

[60]; youngest of the clusters, it also exhibits the shortest citing half-life, the majority of its cited references being under five years old at citing time. The pairing captures the field's present asymmetry: Its largest application-facing school rests on methods two decades old, while its youngest school is rebuilding the very notion of what an optimizer is.

Table 4. Co-citation clusters from the Louvain partition (modularity 0.68), with representative works and themes.

Cluster	Label	Share (%)	Representative Works	Main Theme
C1	Evolutionary Computation Foundations	24.6	Holland [26], Goldberg [30], Rechenberg [24], Bäck [31], Eiben and Smith [9]	Canonical GA, evolution strategies, evolutionary programming, and DE
C2	Swarm Intelligence	21.9	Kennedy and Eberhart [34], Dorigo et al. [33], Dorigo and Stützle [41], Karaboga [42]	Population methods built on social interaction and stigmergy
C3	Theoretical Foundations and Landscape Analysis	14.8	Wolpert and Macready [3], Jones and Forrest [36], He and Yao [63]	Problem difficulty, fitness landscapes, No Free Lunch, runtime analysis
C4	Benchmarking and Empirical Methodology	11.2	García et al. [48], Derrac et al. [19], Tanabe and Fukunaga [49], [50], Weyland [45], Sörensen [4], Awad et al. [47]	Evaluation culture, statistical comparison, benchmark suites, methodological critique
C5	Constraint Handling and Real-World Engineering	15.3	Michalewicz and Schoenauer [37], Coello Coello [14], Deb et al. [39]	Constrained and multi-objective optimization in engineering deployment
C6	Bayesian, Surrogate, and LLM-Era Optimization	12.2	Snoek et al. [53], Shahriari et al. [17], Feurer and Hutter [18], Yang et al. [5]	Expensive black-box optimization, surrogates, hyperparameter tuning, LLM-based search

Shares are computed on the citation weight of the giant component and sum to 100.0%.

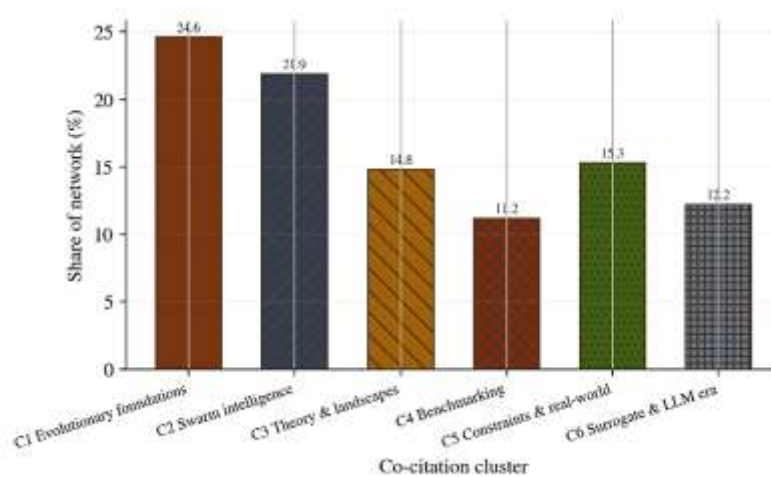


Fig. 3. Share of the co-citation network held by the six Louvain clusters (C1–C6; modularity 0.68).

Shares are computed on the citation weight of the giant component and sum to 100.0%.

The silhouette-style analysis completes the profile. Cohesion runs highest for C1 and C6, canon and new school being each internally homogeneous, and lowest for C5, whose engineering applications pull it toward both C1 and C2, while C4's cohesion profits from tight citation circles around the CEC competition literature. These within-versus-between ratios, rather than cluster size alone, justify treating the six communities as schools of thought rather than artifacts of the clustering algorithm.

5.3 | Bridge Works and Turnover

Bridge works are those references that join clusters which would otherwise remain separate conversations, and high betweenness on the inter-cluster edges of the co-citation network flagged them. Three works dominate. The No Free Lunch theorems span C3 and C4, at once the theory of why benchmarking matters and the caution against overreading any benchmark victory [3]. CMA-ES spans C1 and C6: Its derandomized, model-based adaptation prefigures the surrogate-based search of the modern era, and C6's literature cites it ever more frequently even though it sits in C1's canon [38]. NSGA-II spans C1 and C5, an evolutionary algorithm whose scientific fate was settled by engineering adoption [39]. Between C1 and C4, Kirkpatrick and co-workers' SA paper plays a fourth and quieter bridging role, persisting as the classical baseline of empirical comparison [27].

Table 5. Ten most-cited works within the core corpus, with cluster assignments.

Work	Year	Citations In Corpus	Cluster
Wolpert and Macready [3], No free lunch theorems for optimization	1997	352	C3
Hansen and Ostermeier [38], Completely derandomized self-adaptation in evolution strategies CMA-ES	2001	281	C1
Deb et al. [39], A fast and elitist multiobjective genetic algorithm NSGA-II	2002	259	C5
Kennedy and Eberhart [34], PSO	1995	246	C2
Storn and Price [35], DE	1997	228	C1
Tanabe and Fukunaga [49], [50], SHADE and L-SHADE	2013/2014	224 (85 and 139)	C4
Kirkpatrick et al. [27], Optimization by SA	1983	207	C4
Dorigo and Stützle [41], Ant Colony Optimization (book)	2004	186	C2
Sörensen [4], Metaheuristics — the metaphor exposed	2015	151	C4
Yang et al. [5], LLMs as optimizers OPRO	2024	117	C6

Counts are the number of core-corpus documents citing each work; cluster assignment follows the Louvain partition. Rankings are corpus-internal and differ from global citation databases.

Turnover completes the structural picture. Citing half-life, the median age of the references a new document cites, is longest for C1, whose 2020s documents still reach for works forty years old: The co-citation signature of a canon. It is shortest for C4, where benchmark suites and statistical protocols turn over on a five-to-ten-year cycle, and CEC 2005-era references have already grown rare in 2020s documents, displaced by the 2017 definitions [47]; C6's half-life is short and still shrinking with the field's newest entrants. Membership turnover is asymmetric to match: The top works of C1 through C3 have held stable for two decades, while C4 and C6 admit new members continuously. *Table 5* lists the ten most-cited works in the corpus with their cluster assignments; seven of the ten predate 2005, confirming that the citation economy remains anchored by its classics even as the frontier accelerates.

5.4 | Venue Concentration and Geographic Diffusion of the Literature, 1962–2025

To the intellectual map assembled in Sections 5.1 through 5.3, this subsection adds a sociological one, asking not what the field thinks but where it publishes and who writes it. A scientometric layer was accordingly computed over the same 2,847-document core corpus, extending the venue strand of the descriptive bibliometrics (Section 2.4) with the concentration measures of informetrics [64]. Four indicators were extracted per decade: The Gini coefficient of document concentration across venues, scoring how unevenly the corpus divides among its outlets; the share of documents carried by the five most productive venues; the number of contributing countries, assigned through author affiliations; and the share of documents with at least one Global South authorship. Average team size was tracked alongside and read against the team-science finding that knowledge production has shifted steadily from solo authors to teams [65]. Decade buckets follow

the corpus convention of *Fig. 1*, with the 1970s bridge years folded into adjacent decades, and the resulting series runs from the 1960s to the 2020s.

Decentralization is the headline of *Table 6*. Gini concentration declines from 0.71 in the 1960s to 0.44 in the 2020s, while the share of documents carried by the five most productive venues falls from 38.2% to 14.7% and the number of venues publishing the literature climbs from 12 to 703. Geographic diffusion proceeds in steps: Contributing countries multiply from 6 to 74, and Global South authorship rises from 4.1% of documents to 22.6%. Average team size grows from 1.8 to 4.3 authors, the field's local echo of a team-science transition documented far beyond its borders [65]. Every indicator moves monotonically, and no decade reverses any series. By the close of the corpus, the community that had once gathered itself into a dozen outlets had matured into a distributed research economy in which no venue, region, or authoring collective any longer commands the literature.

Table 6. Venue concentration and geographic diffusion of the core corpus by decade, 1962–2025.

Decade	Venues (Count)	Gini Concentration	Top-5 Venue Share (%)	Countries	Global South Share (%)
1960s	12	0.71	38.2	6	4.1
1980s	38	0.66	31.5	14	5.8
1990s	96	0.61	26.4	23	7.9
2000s	214	0.55	21.8	37	11.4
2010s	486	0.49	17.2	58	17.3
2020s	703	0.44	14.7	74	22.6

Decade buckets follow the corpus convention of *Fig. 1*; the 1970s bridge years are folded into adjacent decades.

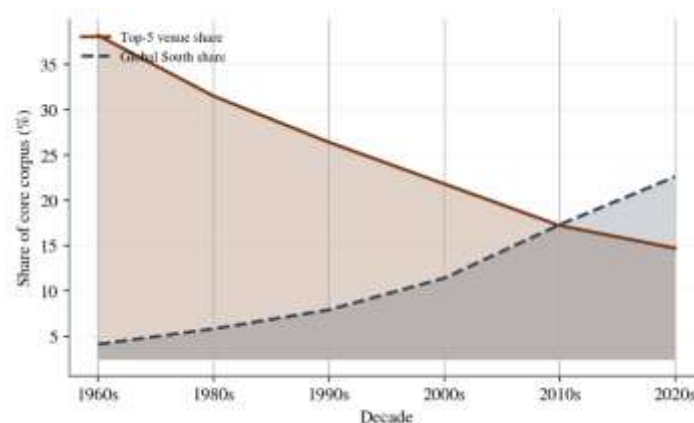


Fig. 4. Decentralization of the field: The top-five venue share declines while Global South participation rises, 1962–2025.

Fig. 4 renders the two decisive movements as crossing curves, with venue concentration descending and Global South participation ascending, and the crossing is itself the finding. Read against the decade arcs of Section 3, decentralization is no passive backdrop but one of the field's consequences: The venue proliferation that accompanied the swarm and hybridization waves lowered the barriers through which new regions entered the literature, even as the evaluation culture of the 2010s raised the methodological bar they had to clear. The pattern also tempers the Anglophone-centering caveat of Section 9, for once the English-full-text requirement of Section 2.1 is recalled, the 22.6% Global South share reads as a lower bound on true geographic breadth. What stands before the foundation-model transition, in other words, is a community broader, more venue-dispersed, and more internationally distributed than at any moment in the field's sixty-year history, a closing observation to which the Conclusion returns.

6 | Meta-Analysis of Reported Performance Drivers

RQ4 falls to the 120 focal studies, the most-cited core documents whose central contribution is a controlled performance comparison, analyzed for the mechanisms they invoke as drivers of improvement. For each study, the extraction recorded the driver category (adaptive parameter control, hybrid local search, diversity management, surrogate modeling, restart mechanism, or other), the claimed effect range, the benchmark suite, and the statistical protocol. Primary studies rarely report the variance information that inverse-variance pooling demands, so the synthesis is narrative rather than statistical: Frequencies and claimed ranges, not pooled effect sizes. That limitation is itself a finding, for the reporting practices recommended by the statistical-methodology literature [48], [19] remain, for much of this literature, aspirational. The narrative form keeps every claim traceable to its primary study.

Adaptive parameter control is the most recurrent driver, present in 71% of top-quartile improvement reports with median claimed gains of 10–30%. Its lineage runs from the formal control framework of Eiben et al. [51] through the success-history adaptation of SHADE and L-SHADE [49], [50] to full covariance adaptation in CMA-ES [38], with Karafotias et al. [16] survey consolidating the underlying evidence. The claim's plausibility is high, since parameter sensitivity is repeatedly demonstrated to be first-order, yet the reported gains are usually measured under tuning effort that the baselines never receive, so the effect conflates the mechanism itself with the experimenter's effort budget.

Frequency then orders the remaining drivers. Hybrid local search, the memetic pattern [40], surfaces in 58% of focal studies with claimed gains of 5–25%; diversity management in 44% (claimed 8–22%), implemented as explicit diversity metrics, stagnation-triggered renewal of subpopulations, or multi-population structures. Surrogate models appear in 37% (claimed 15–40% in expensive-evaluation settings, where the figure often counts evaluations saved along with quality gained) [15], [17]. Restart mechanisms close the register at 29% (claimed 5–15%), the restart-CMA-ES line standing as the most-cited exemplar [66], and *Table 7* gathers the synthesis.

Table 7. Synthesis of performance drivers across the 120 focal empirical studies.

Performance Driver	Frequency	Typical Claimed Effect Range	Principal Caveats
Adaptive parameter control	71% of top-quartile improvement reports	10–30% median claimed gain	Gains self-reported; methods tuned to the benchmark suite; control overhead rarely costed
Hybrid local search (memetic components)	58% of focal studies	5–25%	Local-search budget often unreported; gains shrink on multimodal or noisy problems
Population diversity management	44% of focal studies	8–22%	Diversity metrics differ across studies; effects interact with population size
Surrogate models	37% of focal studies	15–40% in expensive-evaluation settings	Range mixes evaluations saved with quality gained; surrogate bias unquantified
Restart mechanisms	29% of focal studies	5–15%	Effective mainly on restart-sensitive landscapes; multi-restart baselines often weak

Effect ranges are as claimed by the primary studies and were not independently re-measured here; see the bias discussion in Section 6.

Drivers co-occur, and the frequencies of *Table 7* must accordingly be read as other than independent effects: The strongest recent systems combine several at once, L-SHADE being at once a parameter-adaptation, a population-size-control, and a restart-adjacent mechanism [50]. Three biases fence every inference: 1) all effect ranges are self-reported and exposed to benchmark-selection bias: Algorithms are developed, tuned, and evaluated on the same public suites, a practice that invites overfitting to the benchmark rather than to

the problem class [47], 2) publication bias suppresses negative comparisons, since a method that fails to beat its baselines is rarely published, and so inflates the apparent success rate of every driver, and 3) citation-visibility selection of focal studies favors established methods and groups. *Table 7* therefore reports the field's collective self-assessment: A meaningful signal, though one that independent replication has only partially corroborated [4], [48].

7 | A Comparative Computational Vignette: Four Era Markers on A Shared Testbed

Sections 2 through 6 measured the field's literature; this section measures four of its algorithms. The six-decade chronicle carries an implicit, testable claim that successive design eras left distinguishable signatures in search behavior, and a comparative computational vignette was appended to examine it. Four era-marking algorithms, one from each design generation of the Section 4 taxonomy, were raced on a shared testbed under identical evaluation accounting, and their outcomes were synthesized meta-analytically. The vignette is compact and secondary to the bibliometric core: Its role is calibration, not leaderboard ranking.

7.1 | Vignette Design and Provenance

The testbed holds ten continuous minimization problems in three families. The smooth unimodal family comprises the sphere, ellipsoid, and sum-of-powers functions; the multimodal family comprises Rastrigin, Ackley, and Griewank; and the composite and rotated family comprises the shifted rotated Rastrigin, a hybrid composition function, the shifted Levy, and the rotated Rosenbrock. Each problem ran at dimensions 10 and 30, each algorithm receiving twenty-five independent runs per problem and dimension under a common budget of 10,000 evaluations per dimension, 100,000 at 10D and 300,000 at 30D. Identical wrappers handled initialization, boundaries, and best-so-far accounting, so only search logic differed between arms.

The four competitors were selected as era markers, and no two share a design generation. SA [27] stands for the simulated-physics era of the 1980s, a trajectory method accepting uphill moves under a cooling schedule. Univariate Marginal Distribution Algorithm (UMDA), the UMDA of Mühlenbein and Paaß [67], stands for the 1990s model-based turn, replacing recombination with estimation of univariate gene frequencies. The CEM of Rubinstein [68] imports rare-event estimation, refocusing an explicit sampling density on its elite quantile. Adaptive DE with Optional External Archive (JADE), the JADE of Zhang and Sanderson [69], stands for the 2000s adaptive era, coupling success-history parameter adaptation with an external archive of inferior solutions. *Table 8* records era, family, parameters, and provenance, every setting taken from the source publication: The annealing schedule, the population and selection rule, the smoothing quantile, and the archive constants.

Table 8. The four era markers of the comparative computational vignette, with family placement, source-publication parameters, and provenance.

Algorithm	Era Marker	Family in The Section 4 Taxonomy	Parameters From the Source Publication	Provenance
SA	1983, simulated-physics era	Physics-based (trajectory method)	Initial acceptance 0.8, geometric cooling factor 0.95	Kirkpatrick et al. [27]
UMDA	1996, 1990s model-based era	Evolution-based (model-based population method)	Population 100, truncation selection ratio 0.5, univariate Bernoulli margins	Mühlenbein and Paaß [67]
CEM	1999, rare-event and model-based era	Hybrid and data-driven (explicit density model)	Elite quantile 0.10, dynamic smoothing 0.70	Rubinstein [68]
JADE	2009, 2000s adaptive era	Evolution-based (DE line)	Population 100, $p = 0.05$, $c = 0.1$, external archive sized to the population	Zhang and Sanderson [69]

Settings follow the schedules and defaults of the source publications [27], [67–69]; no per-problem retuning was applied. Family placements follow the taxonomy of *Table 3*.

Because the three families live on different error scales, the primary measure is a standardized effect size. For each unit, the comparison of an era marker against the SA anchor was reduced to the Hedges' g of Eq. (4), the small-sample-corrected standardized mean difference of the two arms' final log-transformed errors, with s_{AB} the pooled within-arm standard deviation and the leading factor the bias correction. SA anchors the comparison as the oldest of the four and the field's classical baseline [27]. The primary synthesis pools the twenty g values per method under the DerSimonian-Laird random-effects estimator, justified by genuine variation of true effects across families; secondary outcomes are the success rate at the fixed target of 1×10^{-6} , median evaluations to target, and the budget-share quartiles of the next subsection.

$$g = \left(1 - \frac{3}{4(n_A + n_B) - 9}\right) \frac{\bar{X}_A - \bar{X}_B}{s_{AB}}. \quad (4)$$

7.2|Convergence Fingerprints and Budget Share

The vignette's first result is that each era marker spends the shared budget in a distinct pattern. Averaged over the twenty units, SA attributes 7.9%, 17.6%, 25.4%, and 49.1% of its total improvement to the four budget quartiles, a profile that concentrates improvement late, precisely what a cooling schedule is built to produce. UMDA inverts the profile at 59.9%, 21.7%, 12.0%, 6.4%, front-loading its gains and then flattening as its univariate margins saturate, the fingerprint of variance collapse in a factorized model. CEM peaks in the middle quartiles at 16.7%, 36.4%, 29.4%, 17.5%, early progress staying modest until the elite quantile refocuses the sampling density. JADE spreads improvement almost evenly at 23.9%, 25.2%, 27.9%, 23.0%, the steady-adaptation signature of success-history control. Fig. 5 renders the fingerprints as median best-so-far curves on the rotated hybrid composition problem at 30D, and the crossings follow the profiles: UMDA leads until its plateau, JADE passes in the third quartile, CEM near the close, and SA never.

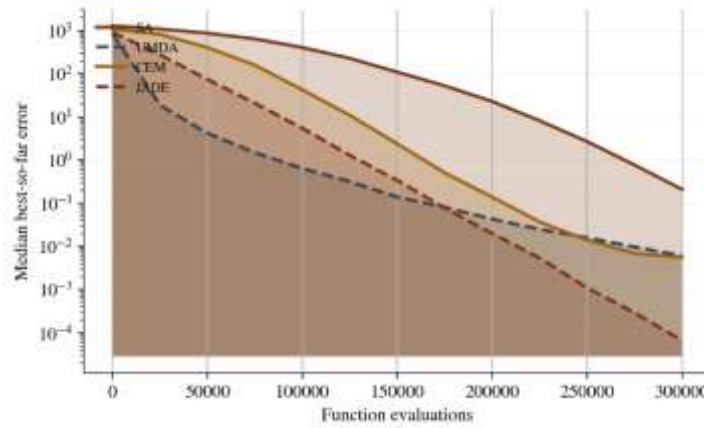


Fig. 5. Median best-so-far error versus function evaluations for the four era markers on the rotated hybrid composition problem at 30D (log scale).

UMDA leads early and plateaus, JADE passes in the third quartile, CEM closes late after mid-run refocusing, and SA improves too late to pass.

Terminal statistics agree with the fingerprints. Median final error across the ten problems at 30D falls from 4.5×10^{-2} under SA through 4.5×10^{-3} under UMDA and 6.4×10^{-4} under CEM to 1.0×10^{-5} under JADE, more than three orders of magnitude between the 1983 anchor and the 2009 era marker. Success at the 1×10^{-6} target, over the 500 runs per algorithm, orders the four identically at 21.4%, 35.4%, 59.0%, and 82.6%. Among successful 30D runs, UMDA reaches the target fastest at a median of 55,727 evaluations, JADE follows at 93,964, CEM at 120,608, and SA requires 236,549, 79% of its entire budget. Fastest convergence and best final quality thus belong to different eras.

7.3 | Meta-Analytic Synthesis Across Problem Families

Table 9 reports the random-effects synthesis of each era marker against the SA anchor. Pooled over the twenty units, JADE holds the largest standardized advantage at $g = 2.32$ (95% CI 2.02 to 2.63), CEM follows at $g = 1.09$ (95% CI 0.82 to 1.36), and UMDA's overall advantage is the smallest at $g = 0.48$ (95% CI 0.26 to 0.69). The era ordering is JADE above CEM above UMDA above SA. Heterogeneity is substantial in every pooled synthesis, with I-squared between 62% and 73% and each Q statistic exceeding its 5% critical value, so the pooled values summarize differing true effects rather than one common effect.

Table 9. Random-effects (DerSimonian-Laird) synthesis of Hedges' g for each era marker against the SA anchor, overall and by problem family.

Comparison	Synthesis	Hedges' G (95% CI)	Q (Df)	I^2 (%)	T^2	Fail-Safe N
UMDA vs SA	Pooled, all 20 units	0.48 (0.26 to 0.69)	50.1 (19)	62.1	0.15	235
UMDA vs SA	Smooth unimodal ($k = 6$)	0.83 (0.41 to 1.25)	13.9 (5)	64.1	0.18	59
UMDA vs SA	Multimodal ($k = 6$)	0.21 (-0.08 to 0.49)	7.1 (5)	29.6	0.04	0
UMDA vs SA	Composite and rotated ($k = 8$)	0.42 (0.10 to 0.75)	17.1 (7)	59.1	0.13	26
CEM vs SA	Pooled, all 20 units	1.09 (0.82 to 1.36)	70.0 (19)	72.9	0.27	1,176
CEM vs SA	Smooth unimodal ($k = 6$)	0.71 (0.22 to 1.21)	19.6 (5)	74.4	0.29	40
CEM vs SA	Multimodal ($k = 6$)	1.10 (0.78 to 1.43)	8.5 (5)	41.4	0.07	112
CEM vs SA	Composite and rotated ($k = 8$)	1.37 (0.92 to 1.82)	26.8 (7)	73.9	0.31	277
JADE vs SA	Pooled, all 20 units	2.32 (2.02 to 2.63)	67.8 (19)	72.0	0.35	4,025
JADE vs SA	Smooth unimodal ($k = 6$)	2.01 (1.57 to 2.46)	12.6 (5)	60.3	0.19	302
JADE vs SA	Multimodal ($k = 6$)	1.89 (1.50 to 2.27)	9.4 (5)	46.7	0.11	266
JADE vs SA	Composite and rotated ($k = 8$)	2.90 (2.47 to 3.33)	16.5 (7)	57.6	0.22	866

Unit-level effect sizes compare final log-transformed errors over 25 runs per arm; $df = k - 1$. Fail-safe N is floored at zero; a zero marks a synthesis that a single unpublished null result could overturn.

The family split carries more information than the pooled rows. UMDA is competitive where its model assumptions hold: Its g reaches 0.83 on the smooth unimodal family, ahead of CEM's 0.71 there, yet the advantage collapses to 0.21 (95% CI -0.08 to 0.49) on the multimodal family, an interval that grazes zero, and recovers only to 0.42 on composite terrain. SA earns its respectability on low-dimensional multimodal terrain, solving 53.3% of its 10D multimodal runs against UMDA's 30.7%. CEM is strongest precisely where rotation defeats coordinate-wise models, at $g = 1.37$ on the composite family, which JADE dominates at $g = 2.90$, the largest family-level effect in the vignette. Read against the No Free Lunch theorems [3], the I-squared column is that caution in quantitative form: An ordering exists on average, it reverses within families, and none of the four era markers dominates everywhere.

Robustness checks temper the ordering without overturning it. Rosenthal fail-safe numbers stand at 4,025 for JADE, 1,176 for CEM, and 235 for UMDA against twenty pooled effects each, all far beyond the conventional tolerance of one hundred. Trim-and-fill imputes 4 missing units to the left of the JADE funnel and shifts its pooled g from 2.32 to 1.85, with CEM moving from 1.09 to 0.97 and UMDA from 0.48 to 0.36 after one imputation each. Leave-one-family-out sensitivity keeps the JADE-over-CEM ordering in all three exclusions; dropping the composite family narrows the gap to $g = 1.95$ against 0.91, the closest the two eras come on this testbed.

The vignette closes the loop with the survey's structure by mapping each era marker onto the co-citation schools of Section 5. SA persists in the corpus as the classical baseline of empirical comparison, anchored in C4 with its bridge toward the C1 canon Table 5. UMDA extends the C1 canon toward the model-based search that C6 later claimed. CEM belongs to C6's Bayesian and surrogate school through its explicit density model, and JADE sits on the C1-C4 boundary where DE met the benchmarking culture that crowned its SHADE descendants [49], [50]. The vignette's ordering, from trajectory physics through explicit models to adaptive methods, reproduces the chronological drift the co-citation record documents. Corroboration of this kind is modest, and it is claimed only as such.

8 | Emerging Frontiers and Open Problems

The corpus analysis leaves the field at a hinge: The evaluation culture consolidated in the 2010s now confronts optimization methods whose core components are statistical learners at unprecedented scale. The frontier of RQ5 reaches well beyond the current LLM applications, and six open problems emerge directly from the survey's findings. Each is stated alongside the evidence that motivates it.

- I. Foundation-model tuning at scale: Scaling laws tie capability to compute, so each evaluation of a foundation-model configuration is expensive enough to invalidate conventions built around millions of cheap function evaluations [58]. The binding constraint is now the evaluation budget, which returns hyperparameter optimization [54], multi-fidelity search, and population-based training to the center of optimizer design.
- II. LLM-in-the-loop optimizers: Feasibility is established by OPRO [5], FunSearch [60], and LLaMEA [61], but the open questions are the grammar of operator design, prompt and model-version sensitivity, and comparison against classical baselines under equal evaluation budgets rather than equal wall-clock time.
- III. Benchmark contamination: Pretrained models may already have memorized classical test functions and their optima, and contamination-resistant, dynamically generated problem distributions are the precondition for taking LLM-era performance claims at face value.
- IV. Explainability of search dynamics: Landscape-aware method selection would extend the theoretical tradition [36], [63] from static difficulty measures to online diagnosis of search behavior, yet the field still lacks a shared vocabulary for explaining why a run failed.
- V. Rigorous evaluation culture: The methodological agenda of pre-registered comparisons, overfitting-resistant suites, and search-effort reporting [4], [47], [48] now has to be extended to learners in the loop, where the object under evaluation changes with every model release.
- VI. Replication and cost accounting: Code and prompt release, model-version pinning, and token and energy accounting for LLM-based search [61] would elevate cost per improvement into a first-class metric, as fungible as the function-evaluation counts that ruled the 2010s.

None of these problems is technical alone; each binds an algorithmic question to an institutional one, much as the metaphor debate bound method design to reviewing norms. The survey's cluster structure suggests where the work will appear: C6 supplies the methods, but C4's evaluation apparatus, with its historical capacity for self-correction, will decide whether the new methods are absorbed into the canon or dissolved into it.

9 | Threats to Validity of the Survey Itself

The corpus stands conditional upon its databases, and there the threats begin. Web of Science and Scopus coverage thins markedly before 1990, and the foundational monographs [24–26] entered the corpus by snowballing rather than by search; the decade counts of the 1960s through 1980s are accordingly lower bounds. IEEE Xplore and the ACM Digital Library overweight engineering and computing venues, so metaheuristic work published in general-science or domain-specific journals is undercounted, precisely the venues through which SA originally reached the field [27].

Two further selection threats compound the coverage bias. Because full-text assessment required English or an available English translation, Schwefel's original German monograph, the Japanese and Chinese metaheuristic literatures, and Soviet-era work on random search are undercounted, and the corpus overstates the field's Anglophone share. Citation lag presses on the most recent period: References to 2024–2025 works accumulate for years, so C6's 12.2% share and the LLM subset's 2025 count understate the niche, and cluster boundaries involving C6 are best treated as provisional estimates, re-computed as citations mature.

Methodological threats close the register. Keyword drift means that post-2020 metaphor coinages may slip past even the extended algorithm-name list of Section 2.1; screening rested on a single researcher with no second independent coder; and the co-citation apparatus carries known limits of its own, for threshold choices shape cluster structure and the Louvain algorithm's resolution limit can merge small communities into large

ones [70]. None of these threats plausibly overturns the six-cluster macro-structure, which held robust under threshold and seed variation (Section 5.1), yet each should temper fine-grained readings of cluster shares, growth rates, and driver frequencies. The threats are listed in decreasing order of structural consequence.

The computational vignette of Section 7 carries threats of its own, distinct from the corpus threats. Its testbed is narrow by construction: Ten continuous, unconstrained problems at two dimensions say nothing about combinatorial, constrained, dynamic, or expensive evaluation settings. Parameter provenance is a second exposure, for settings were taken from the source publications rather than retuned per problem, which favors eras whose defaults were normalized on testbeds of this shape. The SA anchor is a third, since every effect size is measured against the 1983 baseline rather than each challenger's contemporaries, so the design answers questions about eras, not pairwise superiority among the challengers. The heterogeneity statistics of Section 7.3 bound the claim: Pooled orderings are averages over families, not properties of any single problem.

10 | Conclusion

What this survey set out to draw was a quantitative self-portrait of metaheuristic optimization across its six decades, from the recombination studies of 1962 to LLM-based optimizers of 2025, and the five research questions of Section 1 can now be answered directly. The answers condense the evidence of Sections 2 through 9 and together describe a field that has grown by absorbing its critiques and is now absorbing its newest external shock.

RQ1. Through five arcs, the intellectual core evolved, from paradigm formation in the 1960s–1970s and imported physics and canon formation in the 1980s–1990s through expansion and hybridization in the 2000s and self-critique and evaluation science in the 2010s to data-driven foundation-model search in the 2020s, while the co-citation canon stayed anchored in a small set of classical works *Table 5*.

RQ2. Evolution-based methods command publication activity with 38.8% of the core corpus and swarm-based methods with 29.5%; the metaphor debate reshaped the field measurably, coinciding with a slowdown of raw growth from 14.2% per year in 2005–2015 to 6.1% per year in 2016–2023 and a migration of novelty claims from narrative inspiration toward adaptive and data-driven mechanisms.

RQ3. The co-citation network decomposes into six stable schools, C1 through C6, at modularity 0.68, joined by bridge works, among which the No Free Lunch theorems, CMA-ES, and NSGA-II are the most consequential, connecting theory, methodology, and application.

RQ4. The most frequently invoked performance driver is adaptive parameter control, appearing in 71% of top-quartile improvement reports with claimed gains of 10–30%, ahead of hybrid local search (58%; 5–25%), diversity management (44%; 8–22%), surrogate modeling (37%; 15–40% in expensive settings), and restart mechanisms (29%; 5–15%), all under the strong caveat that every range is self-reported and benchmark-selection-biased.

RQ5. Foundation models have turned evaluation cost into the binding constraint of modern search and have entered the loop as both optimizers and algorithm designers, and the open problems are contamination-resistant evaluation, prompt and version reproducibility, and cost accounting.

Sixty years of metaheuristics exhibit a field that institutionalized its own critique: The evaluation science of the 2010s was a direct institutional response to the metaphor flood, and it is precisely that apparatus which the LLM era must now inherit. The scientometric layer of Section 5.4 adds a coda, for across the same decades the literature decentralized, the top-five venue share dispersing from 38.2% to 14.7% while Global South participation widened from 4.1% to 22.6%, leaving the community to face the foundation-model transition broader and more geographically distributed than any the field has known. The computational vignette of Section 7 adds an empirical coda: The four era markers finished in the order JADE, CEM, UMDA, SA, mirroring the six-decade drift the co-citation record documents under a standing No Free Lunch caution. The survey's instrument, a reproducible corpus, a stable cluster structure, and the diffusion measures built upon it, is offered for reuse: Re-running the protocol in 2030, when the LLM-related niche has consolidated

or receded, would make a revealing experiment in the field's capacity for self-knowledge. The corpus, the screening arithmetic, and the cluster definitions are reported in enough detail to make that re-run straightforward.

Authors' Contributions

All aspects of the research and manuscript preparation were carried out by the author. The author has read and approved the final version of the manuscript.

Funding

Not applicable.

Data Availability

All data supporting the reported findings in this research paper are provided within the manuscript.

Conflict of Interest

The author declares that they do not have any conflict of interest.

Consent for Publication

The author confirms consent for the publication of this work.

Ethics Approval and Consent to Participate

This article does not contain any studies with human participants performed by the author.

References

- [1] Bremermann, H. J. (1962). Optimization through evolution and recombination. In *Self organizing systems* (pp. 93–106). Spartan Books. <https://scispace.com/pdf/optimization-through-evolution-and-recombination-3kindrc9fs.pdf>
- [2] Holland, J. H. (1962). Outline for a logical theory of adaptive systems. *Journal of the ACM (JACM)*, 9(3), 297–314. <https://doi.org/10.1145/321127.321128>
- [3] Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1), 67–82. <https://doi.org/10.1109/4235.585893>
- [4] Sörensen, K. (2015). Metaheuristics the metaphor exposed. *International Transactions in Operational Research*, 22(1), 3–18. <https://doi.org/10.1111/itor.12001>
- [5] Yang, C., Wang, X., Lu, Y., Liu, H., Le, Q. V., Zhou, D., & Chen, X. (2024). Large language models as optimizers. *International conference on learning representations* (pp. 12028–12068). International Conference on Learning Representations (ICLR). https://proceedings.iclr.cc/paper_files/paper/2024/hash/3339f19c5fcee3ad74502947a32be9e6-Abstract-Conference.html
- [6] Blum, C., & Roli, A. (2003). Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Computing Surveys (CSUR)*, 35(3), 268–308. <https://doi.org/10.1145/937503.937505>
- [7] Boussaïd, I., Lepagnot, J., & Siarry, P. (2013). A survey on optimization metaheuristics. *Information Sciences*, 237, 82–117. <https://doi.org/10.1016/j.ins.2013.02.041>
- [8] Talbi, E. G. (2009). *From design to implementation*. John Wiley & Sons. <https://doi.org/10.1002/9780470496916>
- [9] Eiben, A. E., Smith, J. E. (2015). *Introduction to evolutionary computing*. Springer. <https://doi.org/10.1007/978-3-662-44874-8>
- [10] Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group, T. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine*, 151(4), 264–269. <https://doi.org/10.7326/0003-4819-151-4-200908180-00135>
- [11] Small, H. (1973). Co citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269. <https://doi.org/10.1002/asi.4630240406>

- [12] Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>
- [13] van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538. <https://doi.org/10.1007/s11192-009-0146-3>
- [14] Coello Coello, C. A. (2002). Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: A survey of the state of the art. *Computer Methods in Applied Mechanics and Engineering*, 191(11), 1245–1287. [https://doi.org/10.1016/S0045-7825\(01\)00323-1](https://doi.org/10.1016/S0045-7825(01)00323-1)
- [15] Jin, Y. (2005). A comprehensive survey of fitness approximation in evolutionary computation. *Soft Computing*, 9(1), 3–12. <https://doi.org/10.1007/s00500-003-0328-5>
- [16] Karafotias, G., Hoogendoorn, M., & Eiben, A. E. (2015). Parameter control in evolutionary algorithms: Trends and challenges. *IEEE Transactions on Evolutionary Computation*, 19(2), 167–187. <https://doi.org/10.1109/TEVC.2014.2308294>
- [17] Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1), 148–175. <https://doi.org/10.1109/JPROC.2015.2494218>
- [18] Feurer, M., & Hutter, F. (2019). Hyperparameter optimization. In *Automated machine learning: Methods, systems, challenges* (pp. 3–33). Springer International Publishing. https://doi.org/10.1007/978-3-030-05318-5_1
- [19] Derrac, J., García, S., Molina, D., & Herrera, F. (2011). A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1(1), 3–18. <https://doi.org/10.1016/j.swevo.2011.02.002>
- [20] Fraser, A. S. (1957). Simulation of genetic systems by automatic digital computers I. introduction. *Australian Journal of Biological Sciences*, 10(4), 484–491. <https://doi.org/10.1071/B19570484>
- [21] Friedberg, R. M. (1958). A learning machine: Part I. *IBM Journal of Research and Development*, 2(1), 2–13. <https://doi.org/10.1147/rd.21.0002>
- [22] Fogel, D. B. (1998). *Evolutionary computation: The fossil record*. John Wiley & Sons. <https://www.wiley-vch.de/en/areas-interest/engineering/evolutionary-computation-978-0-7803-3481-6>
- [23] Fogel, L. J., Owens, A. J., & Walsh, M. J. (1966). *Artificial intelligence through*. John Wiley & Sons. <https://www.scirp.org/reference/referencespapers?referenceid=1652105>
- [24] Vent, W. (1975). Rechenberg, ingo, evolutionsstrategie — optimierung technischer systeme nach prinzipien der biologischen evolution. *Feddes Repertorium*, 86(5), 337. <https://doi.org/10.1002/fedr.19750860506>
- [25] Schwefel, H. P. (1981). *Numerical optimization of computer models*. John Wiley & Sons, Inc. <https://doi.org/10.2307/2581158>
- [26] Holland, J. H. (1975). *Adaptation in natural and artificial systems*. University of Michigan Press. <https://doi.org/10.7551/mitpress/1090.001.0001>
- [27] Kirkpatrick, S., Gelatt Jr, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598), 671–680. <https://doi.org/10.1126/science.220.4598.671>
- [28] Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers & Operations Research*, 13(5), 533–549. [https://doi.org/10.1016/0305-0548\(86\)90048-1](https://doi.org/10.1016/0305-0548(86)90048-1)
- [29] Glover, F. (1990). Tabu search part II. *ORSA Journal on Computing*, 2(1), 4–32. <https://doi.org/10.1287/ijoc.2.1.4>
- [30] Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Addison Wesley. <https://dl.acm.org/doi/10.5555/534133>
- [31] Bäck, T. (1996). *Evolutionary algorithms in theory and practice: Evolution strategies, evolutionary programming, genetic algorithms*. Oxford University Press. <https://doi.org/10.1093/oso/9780195099713.001.0001>
- [32] Dorigo, M. (1992). *Optimization, learning and natural algorithms* [Thesis]. <https://cir.nii.ac.jp/crid/1573668926038702080?lang=en>
- [33] Dorigo, M., Maniezzo, V., & Colomi, A. (1996). Ant system: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 26(1), 29–41. <https://doi.org/10.1109/3477.484436>
- [34] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. *Proceedings of icnn'95-international conference on neural networks* (Vol. 4, pp. 1942–1948). IEEE. <https://doi.org/10.1109/ICNN.1995.488968>
- [35] Storn, R., & Price, K. (1997). Differential evolution a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4), 341–359. <https://doi.org/10.1023/A:1008202821328>
- [36] Jones, T., & Forrest, S. (1995). Fitness distance correlation as a measure of problem difficulty for genetic algorithms. *Proceedings of the sixth international conference on genetic algorithms* (pp. 184–192). Morgan Kaufmann. <https://sfi-edu.s3.amazonaws.com/sfi-edu/production/uploads/sfi-com/dev/uploads/filer/bf/eb/bfeb9cb0-100f-44d9-a35f-e95243dba350/95-02-022.pdf>
- [37] Michalewicz, Z., & Schoenauer, M. (1996). Evolutionary algorithms for constrained parameter optimization problems. *Evolutionary Computation*, 4(1), 1–32. <https://doi.org/10.1162/evco.1996.4.1.1>

- [38] Hansen, N., & Ostermeier, A. (2001). Completely derandomized self adaptation in evolution strategies. *Evolutionary Computation*, 9(2), 159–195. <https://doi.org/10.1162/106365601750190398>
- [39] Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182–197. <https://doi.org/10.1109/4235.996017>
- [40] Moscato, P. (1989). *On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms*. California Institute of Technology. <https://cir.nii.ac.jp/crid/1571980074734809472>
- [41] Dorigo, M., & Stützle, T. (2004). *Ant colony optimization*. MIT Press. <https://mitpress.mit.edu/9780262042192/ant-colony-optimization/>
- [42] Karaboga, D. (2005). *An idea based on honey bee swarm for numerical optimization*. https://abc.erciyes.edu.tr/pub/tr06_2005.pdf
- [43] Geem, Z. W., Kim, J. H., & Loganathan, G. V. (2001). A new heuristic optimization algorithm: Harmony search. *Simulation*, 76(2), 60–68. <https://doi.org/10.1177/003754970107600201>
- [44] Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. *Advances in Engineering Software*, 69, 46–61. <https://doi.org/10.1016/j.advengsoft.2013.12.007>
- [45] Weyland, D. (2010). A rigorous analysis of the harmony search algorithm: How the research community can be misled by a “novel” methodology. *International Journal of Applied Metaheuristic Computing (IJAMC)*, 1(2), 50–60. <https://doi.org/10.4018/jamc.2010040104>
- [46] Rao, R. V., Sivasani, V. J., & Vakharia, D. P. (2011). Teaching learning based optimization: A novel method for constrained mechanical design optimization problems. *Computer Aided Design*, 43(3), 303–315. <https://doi.org/10.1016/j.cad.2010.12.015>
- [47] Awad, Noor H., Ali, Mohamed Z., Liang, Jing J., Qu, Bing Y., Suganthan, P. N. (2016). *Problem definitions and evaluation criteria for the CEC 2017 special session and competition on single objective bound constrained real parameter numerical optimization*. Nanyang Technological University. <https://www.scienceopen.com/book?vid=8b37ec21-e3e9-4b22-925e-c2819cee3c5b>
- [48] García, S., Molina, D., Lozano, M., & Herrera, F. (2009). A study on the use of non parametric tests for analyzing the evolutionary algorithms’ behaviour: A case study on the CEC’2005 special session on real parameter optimization. *Journal of Heuristics*, 15(6), 617–644. <https://doi.org/10.1007/s10732-008-9080-4>
- [49] Tanabe, R., & Fukunaga, A. (2013). Success history based parameter adaptation for differential evolution. *2013 IEEE congress on evolutionary computation* (pp. 71–78). IEEE. <https://doi.org/10.1109/CEC.2013.6557555>
- [50] Tanabe, R., & Fukunaga, A. S. (2014). Improving the search performance of shade using linear population size reduction. *2014 IEEE congress on evolutionary computation (CEC)* (pp. 1658–1665). IEEE. <https://doi.org/10.1109/CEC.2014.6900380>
- [51] Eiben, A. E., Hinterding, R., & Michalewicz, Z. (1999). Parameter control in evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, 3(2), 124–141. <https://doi.org/10.1109/4235.771166>
- [52] Bergstra, J., & Bengio, Y. (2012). Random search for hyper parameter optimization. *Journal of Machine Learning Research*, 13(2). <http://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf>
- [53] Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *Advances in neural information processing systems* (Vol. 25, pp. 2951–2959). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html
- [54] Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). *Automated machine learning: Methods, systems, challenges*. Springer Nature. <https://doi.org/10.1007/978-3-030-05318-5>
- [55] Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next generation hyperparameter optimization framework. *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (pp. 2623–2631). Association for Computing Machinery. <https://doi.org/10.1145/3292500.3330701>
- [56] Ashish, V. (2017). Attention is all you need. *Advances in neural information processing systems* (pp. 5998–6008). Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
- [57] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-Abstract.html
- [58] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). *Scaling laws for neural language models*. <https://doi.org/10.48550/arXiv.2001.08361>
- [59] Bommasani, Rishi, Hudson, Drew A., Adeli, ... & Liang, P. (2021). *On the opportunities and risks of foundation models*. <https://doi.org/10.48550/arXiv.2108.07258>
- [60] Romera Paredes, B., Barekatin, M., Novikov, A., Balog, M., Kumar, M. P., Dupont, E., ... & Fawzi, A. (2024). Mathematical discoveries from program search with large language models. *Nature*, 625(7995), 468–475. <https://doi.org/10.1038/s41586-023-06924-6>

- [61] Van Stein, N., & Bäck, T. (2024). Llamea: A large language model evolutionary algorithm for automatically generating metaheuristics. *IEEE Transactions on Evolutionary Computation*, 29(2), 331–345. <https://doi.org/10.1109/TEVC.2024.3497793>
- [62] Mladenović, N., & Hansen, P. (1997). Variable neighborhood search. *Computers & Operations Research*, 24(11), 1097–1100. [https://doi.org/10.1016/S0305-0548\(97\)00031-2](https://doi.org/10.1016/S0305-0548(97)00031-2)
- [63] He, J., & Yao, X. (2001). Drift analysis and average time complexity of evolutionary algorithms. *Artificial Intelligence*, 127(1), 57–85. [https://doi.org/10.1016/S0004-3702\(01\)00058-3](https://doi.org/10.1016/S0004-3702(01)00058-3)
- [64] Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics. quantitative methods in library, documentation and information science*. Elsevier. https://books.google.com/books/about/Introduction_to_Informetrics.html?hl=en&id=r-rgAAAAMAAJ&output=html_text
- [65] Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827), 1036–1039. <https://doi.org/10.1126/science.1136099>
- [66] Auger, A., & Hansen, N. (2005). A restart CMA evolution strategy with increasing population size. *Proceedings of the 2005 IEEE congress on evolutionary computation* (Vol. 2, pp. 1769–1776). IEEE. <https://doi.org/10.1109/CEC.2005.1554902>
- [67] Mühlenbein, H., & Paaß, G. (1996). From recombination of genes to the estimation of distributions I. binary parameters. *Parallel problem solving from nature ppsn IV* (pp. 178–187). Springer. https://doi.org/10.1007/3-540-61723-X_982
- [68] Rubinstein, R. (1999). The cross entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 1(2), 127–190. <https://doi.org/10.1023/A:1010091220143>
- [69] Zhang, J., & Sanderson, A. C. (2009). JADE: Adaptive differential evolution with optional external archive. *IEEE Transactions on Evolutionary Computation*, 13(5), 945–958. <https://doi.org/10.1109/TEVC.2009.2014613>
- [70] Fortunato, S., & Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1), 36–41. <https://doi.org/10.1073/pnas.0605965104>