Concept Paper

Beyond Data Moore's Law: Towards Sustainable Scaling of Foundation Models

Feng Chen *

Posted Date: 17 November 2025

doi: 10.20944/preprints202511.1245.v1

Keywords: foundation models; scaling laws; data efficiency; data-centric AI; synthetic data; multiagent systems; autonomous science; interfacial phenomena; evaluation; governance

 doi:10.20944/preprints202511.1245.v1

Concept Paper

Beyond Data Moore's Law: Towards Sustainable Scaling of Foundation Models

Feng Chen

University of Chinese Academy of Sciences, China; 1401761846@qq.com

Abstract

The recent progress of large language and multimodal models has been widely attributed to a de facto "data Moore's law": as model parameters and training tokens increase, performance improves in a predictable manner across diverse benchmarks. However, this paradigm is rapidly approaching multiple limits. High-quality web-scale text is close to saturation, additional data is increasingly redundant, and the financial and environmental costs of continued brute-force scaling are becoming unsustainable. At the same time, the capabilities that matter most for science, engineering, and society—robust reasoning, continual learning, and safe deployment—do not simply emerge from ever-larger piles of uncurated data. In this Perspective, we argue that the next phase of foundation model development must shift from maximizing data volume to optimizing effective information content and ecosystem design. We first analyse the empirical and conceptual constraints of current scaling practices, including data exhaustion, diminishing returns, and misalignment between benchmarks and real-world tasks [1–9]. We then expand the lens from single models to multi-agent LLM ecosystems, where collections of interacting agents, tools, and environments form scalable scientific workflows [10–14]. Drawing an explicit analogy to complex interfacial phenomena in fluid mechanics and condensation—where macroscopic behaviour emerges from local interactions among droplets, contact lines, and patterned substrates—we show how architectural heterogeneity, controlled pinning, active gradients, confinement, and phase-diagram thinking provide concrete design principles for multi-agent systems [15–24]. Building on recent advances in data-centric AI and synthetic data scaling [25–33], we propose a framework that decomposes data quality into four dimensions—coverage, compositionality, conflict, and controllability—and argue that these, rather than raw token counts, will define a realistic "Moore's law of data" for the next decade. Finally, we discuss implications for evaluation and governance, including holistic multi-agent benchmarks, ecosystem-level documentation, and alignment methods that treat scientific LLM ecosystems as institutions in their own right [34–38]. Rather than asking how many more tokens we can consume, we suggest that the central question for the coming decade is how to build sustainable, data-efficient, and well-governed ecosystems in which models, experiments, and human communities co-evolve.

Keywords: foundation models; scaling laws; data efficiency; data-centric AI; synthetic data; multi-agent systems; autonomous science; interfacial phenomena; evaluation; governance

1. Introduction: The Rise and Limits of Data-Centric Scaling

Over the past five years, the capabilities of large language and multimodal models have advanced at a pace that rivals, and in some domains exceeds, the historical trajectory of hardware driven by Moore's law. Models trained on trillions of tokens now exhibit impressive performance on tasks ranging from question answering and code generation to translation, summarisation, and creative writing. Behind these advances lies a simple, powerful empirical regularity: when we increase model size, dataset size, and compute in a coordinated way, the loss on held-out data tends to follow smooth scaling curves [2–4]. These curves, which have been characterised across different architectures and domains, suggest that performance improves predictably as we feed models more parameters and more data. This observation has motivated a "bigger is better" race in which

 doi:10.20944/preprints202511.1245.v1

successive generations of foundation models are primarily distinguished by their parameter counts and the number of tokens they consume during training. Implicit in this narrative is a form of data Moore's law. If traditional Moore's law describes a steady exponential increase in transistor density for a given cost, then data Moore's law can be read as a belief that model capabilities will continue to grow as long as we can supply them with ever-larger datasets and the compute required to absorb them [1–4,7]. In this view, data—especially text scraped from the web—is treated as an effectively infinite resource: noisy, but abundant, and therefore sufficient to support steady progress. The crucial engineering questions then become: how do we scale infrastructure to ingest more tokens, how do we design architectures that make efficient use of these tokens, and how do we stabilise optimisation at extreme scales?

However, there are growing signs that this paradigm is approaching its limits. First, high-quality web-scale text is not infinite. Several analyses indicate that we are rapidly exhausting the pool of public, human-written, non-duplicated text that is suitable for training general-purpose models [8]. As models consume an increasing fraction of available corpora, marginal additions tend to be either redundant—repetitions or near-duplicates of what has already been seen—or low-quality content that contributes little to generalisation and may even degrade robustness. Second, the financial and environmental costs of continued brute-force scaling are escalating. Training a frontier model now requires massive data centres, specialised hardware, and substantial energy consumption, raising concerns about sustainability and access [5–7]. Third, and perhaps most importantly, there is a growing gap between benchmark progress and real-world reliability. Improvements on standard benchmarks do not always translate into proportional gains in factual accuracy, reasoning under distribution shift, or safety in open-ended interactions. These observations suggest that the era in which we could simply "turn the crank" of data and compute is coming to an end. The core question is shifting from How many more tokens can we find and process? to How can we extract the most useful information from the data we have, and generate new data that truly expands our capabilities? In other words, we must move from a paradigm centred on data volume to one centred on effective information content. This shift is not just a matter of better filters or more clever deduplication. It requires rethinking how we conceptualise, collect, curate, generate, and use data throughout the entire lifecycle of foundation models [25–27].

At the same time, the field has witnessed the rapid rise of LLM-based autonomous agents, where models are embedded in loops of perception, planning, tool use, and memory [10]. Surveys of LLM agents emphasise that capabilities such as long-horizon planning, multi-step reasoning, and tool-augmented problem solving arise not only from within-model representations but also from the structure of the surrounding agent architecture—planners vs. executors, specialised experts vs. generalist coordinators, and social communication protocols among multiple agents [10,11]. In this broader view, scaling is not merely about width and depth of a single network; it is about population size, role diversity, and interaction topology in a society of models, tools, and humans [11–14]. This perspective is particularly salient for science, where multi-agent workflows—hypothesis generators, experimental designers, data analysts, and theory builders—can be orchestrated into end-to-end agentic AI for science pipelines [12–14]. To articulate what it means to scale such ecosystems, it is helpful to draw analogies from domains where macroscopic behaviour is governed by local interactions among many heterogeneous actors. Interfacial phenomena in fluid mechanics provide a particularly rich source of intuition. A single droplet on a flat surface can be characterised by Young's equation and capillary length, but once we introduce corrugated substrates, grooved surfaces, and electrowetting, the situation becomes intrinsically multi-body and anisotropic [15,16]. System-level behaviour then depends crucially on how local contact lines, pinning sites, and surface patterns interact across scales. For example, studies of wetting and electrowetting on corrugated substrates reveal that droplet shapes, spreading dynamics, and voltage dependence all change qualitatively once the contact line experiences directional roughness; the same applied voltage can produce different macroscopic responses along and across the grooves [15]. Likewise, investigations of droplets on grooved surfaces show that apparent contact angles vary nonlinearly with an anisotropy

 doi:10.20944/preprints202511.1245.v1

factor, and that an elliptic-cap model is needed to consistently characterise droplet wettability across directions and scales [16].

A similar story holds when we move from static wetting to self-driven interfacial motion. Experiments on droplets driven by interfacial chemical reactions, including camphor-type systems and reactive multicomponent droplets, reveal spontaneous rotations, multi-lobed oscillations, and non-trivial trajectory selection that emerge from feedback between local Marangoni stresses and global flow fields [17–21]. Recent work has demonstrated multi-lobed rotating droplets induced by interfacial reactions [21], self-rotating droplets based on liquid metals [20], and chemically driven rotational instabilities in binary droplets on liquid or solid substrates [17,19]. In each case, individual droplets can be viewed as active agents whose motion is shaped by local gradients, yet the macroscopic patterns—rotating clusters, flower-like morphologies, or self-organised trajectories—arise from the collective interaction of many such agents and their environment [17–21]. These ideas extend naturally to condensation and phase-change phenomena, where local vapour–liquid interactions generate complex mesoscale patterns. Studies of hygroscopic liquids on low-temperature surfaces have shown that droplets can act as vapour sinks, creating dry zones whose size and morphology depend sensitively on substrate temperature, hygroscopicity, and the geometric arrangement of liquid patches [18,22,23]. Experiments on water-vapour uptake into hygroscopic desiccant droplets reveal that droplet growth and spreading are governed by coupled diffusion–convection processes and concentration-dependent surface-tension gradients, leading to highly nonlinear dynamics [18]. More recent work demonstrates that introducing hygroscopic rings or patterned liquids can spatially control condensation, suppress frosting, and generate reconfigurable dry regions that protect underlying surfaces [22–24]. In all of these systems, performance does not scale linearly with the size of any single droplet; instead, it is determined by how droplets are arranged, how they exchange vapour, and how external fields (temperature, geometry, electric potential) structure their interactions [15–24].

We argue that multi-agent LLM ecosystems are structurally analogous to these interfacial systems. A lone monolithic model on a featureless “flat surface” of static text behaves like a single droplet governed by classical scaling laws: performance improves smoothly with size and data, up to physical and data-availability limits [2–4,7,8]. By contrast, a society of interacting LLMs embedded in rich environments—tools, simulations, robotic platforms, and human collaborators—resembles a landscape of anisotropic substrates, patterned grooves, and hygroscopic patches [10–24]. Here, system-level capability emerges less from the capacity of any single agent and more from the architecture of interactions: which agents communicate, how information is routed, how conflicting goals are resolved, and how feedback from the environment shapes collective learning over time. Just as anisotropic wetting and vapour-sink effects can be engineered to steer droplets, suppress condensation, or induce self-organised motion [15–24], we can engineer interaction topologies and data flows to steer multi-agent LLM ecosystems towards robust scientific discovery rather than brittle self-amplification or collapse [9–14]. The remainder of this article develops this analogy into a concrete agenda for sustainable scaling beyond data Moore’s law. Section 2 revisits classical parameter- and loss-based scaling laws in light of physical, economic, and data constraints. Section 3 formalises multi-agent LLM architectures as scientific ecosystems. Section 4 argues that the relevant “Moore’s law of data” concerns effective information content, not raw tokens, and introduces a four-dimensional view of data quality. Section 5 distils design principles for multi-agent ecosystems from interfacial physics. Section 6 discusses evaluation and governance, and Section 7 outlines an outlook for the next decade.

2. Revisiting Scaling Laws: From Parameters and Loss to Data and Environments

The classical story of scaling laws in deep learning is deceptively simple. Early systematic studies showed that for a broad class of architectures and optimisation settings, the cross-entropy loss on held-out data follows a smooth power law as a function of model parameters, dataset size, and compute budget [2–4]. Within this regime, doubling parameters while appropriately increasing

 doi:10.20944/preprints202511.1245.v1

data and compute yields predictable reductions in loss, encouraging a view of neural networks as systems in which performance can be planned primarily through resource allocation. This "engineering equation" for model performance has been enormously influential, guiding the design of successive generations of large language models and providing a convenient, quantitative language for comparing competing training strategies [2–4,7]. Yet even in their original form, these scaling laws came with caveats. The reported power laws were derived under carefully controlled conditions: homogeneous data distributions, fixed training objectives, and regimes far from saturation with respect to both data and compute [2–4]. As practitioners pushed models closer to hardware and data limits, deviations from ideal scaling began to emerge. In some cases, simply adding more parameters yielded diminishing or even negligible gains because the model became data-limited: the marginal information content of each additional token was small relative to what the model had already seen [7,8]. In others, scaling up data without a commensurate increase in model capacity produced under-parameterised regimes where optimisation struggled to fit the training distribution without memorisation or overfitting. These observations hint that the original scaling laws implicitly assume a "healthy" balance between parameters and data that is increasingly difficult to maintain in practice.

A second limitation arises from the near-exclusive focus on loss as the target variable. Cross-entropy on held-out, IID test sets is a convenient proxy for average-case predictive performance, but it does not directly capture calibration, robustness to distribution shift, compositional generalisation, or safety. Empirically, many large models exhibit strong performance on benchmarks that resemble their pretraining data while still failing dramatically on tasks requiring causal reasoning, long-horizon planning, or out-of-distribution generalisation [2–4,7]. Small reductions in loss do not necessarily translate into proportional improvements on these harder dimensions. In other words, the quantity that scales neatly is not always the quantity of real scientific or societal interest. A third source of tension comes from the finite and structured nature of data. The original scaling studies often treated data as an abstract scalar controlling the number of training tokens [2–4,8]. However, real-world corpora are highly non-uniform. They contain pockets of dense coverage in some domains (e.g., popular culture, general news) and severe sparsity in others (e.g., specialised scientific subfields, low-resource languages) [8,12]. As models ingest an increasing fraction of the accessible web, they quickly exhaust the dense regions and are forced into long tails of redundant or low-quality content. Analyses of global text availability suggest that, under current practices, we are on track to consume most of the high-quality public text suitable for pretraining in the near future [8]. Once that happens, simply adding more "data" largely means adding more noise or recycled content, which undermines the assumptions behind smooth power-law scaling.

The recent wave of work on synthetic data and recursion further complicates the picture. It is tempting to imagine that models could simply generate arbitrarily large synthetic corpora, fine-tune on them, and thus circumvent data scarcity. However, careful studies of training on model-generated content reveal a "curse of recursion": when models repeatedly consume their own outputs, distributions drift, errors compound, and performance eventually collapses [9]. Rather than providing a free lunch, synthetic data introduces a new axis in the scaling landscape: the ratio of human-authored to model-generated content, the diversity of generative seeds, and the presence or absence of external ground truth all affect whether synthetic augmentation helps or harms. Classical scaling laws, which collapse everything into an undifferentiated data-size term, are blind to these distinctions [2–4,9]. Taken together, these observations suggest that scaling laws are not wrong, but incomplete. They are accurate descriptions of a narrow regime in which data is abundant, clean, and exogenous; objectives are simple; and models are evaluated under static, IID conditions. The frontier we now face is qualitatively different. Data is increasingly scarce, heterogeneous, and intertwined with model behaviour; objectives are multi-faceted and often involve interaction; and evaluation must account for environments that respond to model actions [10–14,25–27]. In this new regime, the relevant "independent variables" for scaling are not just parameter counts and token numbers, but also data quality, environmental complexity, and the structure of agent–environment interactions.

 doi:10.20944/preprints202511.1245.v1

An instructive analogy comes from interfacial physics. Classical capillarity theory, built on Young's equation and simple balance of surface tensions, provides remarkably accurate predictions for droplets on smooth, homogeneous surfaces [15,16]. But as soon as we introduce anisotropic roughness, sharp edges, or reactive interfaces, these simple laws no longer suffice; apparent contact angles become direction-dependent, pinning and hysteresis dominate dynamics, and new emergent behaviours such as self-rotation or spontaneous pattern formation arise [15–21]. In such settings, it is still possible to derive scaling relations—for example, between groove geometry and apparent anisotropic contact angle [16], or between reaction rate and rotational frequency of active droplets [19–21]—but these relations involve additional parameters describing the structure and dynamics of the environment, not just intrinsic fluid properties. Similarly, as LLMs move from static text prediction to interactive, tool-augmented, and multi-agent settings, we should expect scaling behaviour to depend on environmental variables and interaction topologies that are currently absent from standard formulations [10–14]. This perspective also sheds light on the emerging practice of integrating LLMs into closed-loop scientific workflows. When models propose experiments, call simulations, control robotic platforms, and interpret measurements, the effective "dataset size" is no longer fixed ex ante; it is generated and curated online through interaction [12–14]. The quality and diversity of this data depend on the models' exploration strategies, biases in their prior knowledge, and the physical constraints of instruments [12–15]. Scaling such systems therefore involves not just larger models and longer training runs, but also higher-throughput experimental platforms, more expressive simulation environments, and better-designed feedback policies. In a sense, the Moore's law of data becomes a statement about how quickly we can generate, validate, and incorporate new high-value measurements, rather than how many web pages we can scrape [12–15,22–24].

Revisiting scaling laws in this way leads to a different set of scientific questions. Instead of asking "What loss will a 10^{12}-parameter model achieve on 10^{13} tokens?", we should ask: How does performance scale with the diversity and controllability of the data distribution? How does it depend on the richness of the environment in which agents operate, the topology of their interactions, and the mechanisms by which they update beliefs? What trade-offs arise between maximising average-case performance and ensuring robustness to rare but critical scenarios? These are precisely the sorts of questions that interfacial physics has long grappled with when moving from idealised droplets on smooth plates to complex systems of interacting droplets, patterned substrates, and coupled phase changes [15–24]. In the next section, we formalise this analogy by viewing multi-agent LLM systems as scientific ecosystems, whose scaling behaviour is governed as much by interaction rules and environmental structure as by the size of individual agents.

3. Multi-Agent LLM Architectures as Scalable Scientific Ecosystems

If classical scaling laws treat an LLM as an isolated, homogeneous system, the emerging paradigm of multi-agent architectures invites a very different metaphor: that of an ecosystem populated by diverse, interacting entities [10,11]. In such systems, "scale" is not solely a property of the individual organism (e.g., the number of parameters in a single model), but also of the population (the number and diversity of agents), the structure of interactions (who communicates with whom, under what protocols), and the environment (the tools, datasets, and physical instruments available for agents to act upon) [10–14]. For scientific applications, this ecosystemic view is not just aesthetically appealing; it aligns closely with how scientific communities actually function, with specialised roles, division of labour, and institutional mechanisms for coordination and error correction [12–14].

At a high level, multi-agent LLM architectures can be organised along two axes: functional specialisation and interaction topology. Functional specialisation assigns different roles or capabilities to different agents—such as hypothesis generation, experimental design, simulation control, data analysis, theory building, or literature synthesis [12–14]. Interaction topology determines how these roles are wired together, from simple planner–executor pairs to complex networks of peers that debate, critique, and refine each other's proposals [10,11,13]. Scaling along

 doi:10.20944/preprints202511.1245.v1

either axis can increase system capability: more specialised agents can cover a broader range of tasks, while richer topologies can support deeper chains of reasoning, mutual correction, and redundancy against individual failures.

This architecture mirrors, in striking ways, the behaviour of interfacial systems where many droplets interact on structured surfaces. In studies of anisotropic wetting on corrugated substrates, for example, the apparent contact angle and contact-line dynamics depend sensitively on the orientation of grooves relative to the macroscopic force direction [15,16]. Droplets aligned with grooves experience different pinning and depinning thresholds than those crossing grooves, leading to direction-dependent spreading and hysteresis. Analogously, in a multi-agent LLM system, the "direction" of information flow—who sends prompts to whom, and in what sequence—can create effective anisotropies in reasoning space. Certain chains of agents may be more conducive to exploring speculative hypotheses, while others may be better suited to conservative validation or safety checks. As in the physical system, performance depends not only on the capabilities of individual droplets or agents, but also on how they are oriented and connected.

The analogy becomes even more vivid when we consider active droplets driven by interfacial chemical reactions. Experiments have shown that droplets can spontaneously rotate, oscillate, or follow complex trajectories when surface-tension gradients arise from local reactions or compositional changes [17–21]. In some cases, droplets exhibit multi-lobed rotating shapes whose morphology and angular velocity depend on reaction rates, viscosity, and confinement [21]. In others, liquid-metal-based droplets harness redox reactions to sustain self-rotation or directed motion [20]. These systems behave like self-propelled agents whose dynamics are governed by feedback between local interactions (reaction–diffusion, Marangoni stresses) and global fields (flow, confinement). When multiple droplets are present, their interactions can lead to cooperative motion, pattern formation, or competition for resources [17–21]. Multi-agent LLM ecosystems exhibit analogous phenomena. Individual agents can be endowed with internal drives—optimisation objectives, reward functions, or heuristic "curiosity" signals—that shape their behaviour over time [10,11,13]. When agents share tools, memory, or access to scarce experimental resources, they effectively compete and collaborate in a joint environment. Feedback loops arise when the outputs of one agent become the inputs of another, when agents critique each other's hypotheses, or when they co-adapt to the same evolving dataset. Under some conditions, these feedbacks can lead to stable division of labour and robust collective performance; under others, they may produce runaway amplification of errors, mode collapse, or oscillatory behaviour reminiscent of unstable droplet dynamics [9–14,17–21]. Understanding and engineering these regimes is a central challenge for multi-agent scaling.

Scientific ecosystems add an additional layer of complexity: they are wired not only to other agents, but also to the physical world. In autonomous laboratory settings, agents may control fluidic devices, thermal stages, imaging systems, or other instruments that generate high-dimensional experimental data [12–15,22–24]. Here, the environment is not a static text corpus but a dynamic, partially observable system governed by physical laws. The "data" that agents consume is produced through experiments whose design, execution, and interpretation are themselves mediated by agents. This situation closely parallels hygroscopic and condensation phenomena, where droplets and substrates co-evolve with their vapour environment [18,22–24]. For example, hygroscopic droplets acting as vapour sinks can generate ring-shaped dry zones whose radius depends on substrate temperature, droplet composition, and exposure time; these dry regions in turn affect subsequent condensation, leading to complex spatiotemporal patterns [22,23]. Similarly, agent decisions about which experiments to run and how to interpret results change the underlying data landscape, influencing future decisions and effectively sculpting a self-generated training distribution [12–15]. In this light, multi-agent LLM architectures for science can be thought of as engineered ecosystems whose scaling behaviour must be understood in terms of coupled dynamics between agents and environments. Adding more agents or increasing model sizes changes not just raw capacity, but also the stability and diversity of emergent behaviours. Just as introducing additional droplets or modifying substrate patterns can either stabilise condensation suppression or trigger unwanted

 doi:10.20944/preprints202511.1245.v1

frosting and coalescence [22–24], increasing the density or connectivity of agents can either improve coverage of the scientific search space or exacerbate correlated failure modes and recursive bias [9–14]. Designing robust scaling paths thus requires explicit attention to interaction rules, resource allocation, and environmental structure, not just to single-agent performance.

Practically, this ecosystemic perspective suggests new design principles for multi-agent scientific architectures. First, role specialisation should be treated as a primary axis of scaling: rather than relying on a single, massive generalist model, systems can employ a collection of smaller, specialised agents—some tuned for literature mining, others for experimental planning, others for theoretical modelling—coordinated by higher-level orchestration agents [10–14]. This echoes the way different droplet types (e.g., reactive vs. passive, hygroscopic vs. hydrophobic) can be arranged on a substrate to achieve composite functionalities, such as localised condensation control or directional transport [18,22–24]. Second, interaction topology can be designed to promote constructive interference and error correction. Debate-style protocols, cross-checking among redundant agents, and hierarchical decision structures can help mitigate individual errors and prevent recursive amplification, in analogy with how patterned surfaces and controlled gradients can prevent uncontrolled coalescence or runaway wetting [9–11,15–24].

Third, and crucially for scaling laws, metrics of system performance must be broadened beyond single-agent loss. For multi-agent scientific ecosystems, relevant observables include the rate of validated discoveries, robustness under distribution shift, diversity of explored hypotheses, and resilience to perturbations in data or agent behaviour [12–14]. These quantities may exhibit their own scaling relations as functions of agent population size, topology, and environmental richness, analogous to how pattern wavelength, dry-zone radius, or rotational frequency scale with geometric and physical parameters in interfacial systems [18–24]. Systematically characterising these relations is an open research frontier, one that requires a synthesis of insights from machine learning, complex systems, and nonequilibrium physics.

4. Data as the New Moore's Law: Active, Synthetic, and Experimental Data Loops

If classical scaling laws describe how performance improves as we increase parameters and token counts, the emerging "data Moore's law" asks a different question: How fast can we increase the amount of useful information available to a model per unit compute and time? This reframing aligns with the broader movement toward data-centric AI, which advocates that progress will increasingly be limited not by model architectures, but by how systematically we can engineer and govern data itself [25–27]. A first implication is that we must move beyond treating "dataset size" as a scalar, and instead characterise data along at least four dimensions that determine its effective information content: coverage, compositionality, conflict, and controllability. Coverage refers to the breadth of concepts, tasks, and distributions present in the data; compositionality captures whether examples support systematic recombination and generalisation; conflict reflects the presence and handling of disagreements, edge cases, and hard negatives; and controllability measures our ability to shape model behaviour by selectively adding, removing, or reweighting data with particular properties. Traditional web-scale scraping implicitly maximises coverage under a weak notion of compositionality, while largely ignoring conflict and offering little controllability beyond coarse filtering [2–4,8]. As we approach the limits of high-quality human text [8], the only sustainable way to maintain scaling trends is to improve these qualitative dimensions, not just raw token counts [25–33].

This philosophy is reflected in recent work that explicitly models data quality within scaling laws. Subramanyam et al. propose a formulation in which performance depends not only on dataset size but also on a quality parameter that captures noise and redundancy, showing that higher-quality pools yield better asymptotic performance for the same compute [28]. Goyal et al. derive scaling laws for data filtering, demonstrating that optimal curation strategies depend on both compute and the intrinsic quality of candidate pools; naively mixing heterogeneous sources can be provably sub-

 doi:10.20944/preprints202511.1245.v1

optimal relative to carefully chosen subsets [29]. In parallel, the DataComp and DataComp-LM benchmarks fix model architectures and training recipes while varying only the dataset design, thereby turning data construction itself into the primary object of experimentation [27,30]. These efforts collectively suggest that the next generation of scaling laws will be data-aware, with explicit terms that reflect how coverage, compositionality, conflict, and controllability evolve as we curate and synthesise datasets. Within this emerging landscape, coverage remains necessary but no longer sufficient. For foundation models intended to support scientific discovery, coverage must extend into long-tail technical domains where natural corpora are sparse or fragmented [12,13]. Here, classical web scraping quickly saturates, and the marginal tokens are either redundant or off-task [8]. One promising response is to use LLMs themselves to systematically expand coverage via synthetic data: generating variants of rare phenomena, paraphrasing under-represented styles, and constructing curriculum-style progressions that bridge low- and high-difficulty tasks [31–33]. Synthetic data cannot replace human text, as recursive training risks distributional collapse [9], but it can amplify scarce signals when grounded in high-quality seeds and anchored by external validation.

Compositionality is equally critical, particularly for scientific reasoning. Natural corpora often entangle many concepts without covering their systematic recombinations, leading to models that memorise surface patterns but struggle with novel combinations. DataComp-style experiments emphasise recaptioning and restructuring raw web content to improve semantic alignment and compositional richness [27]. For language models, DataComp-LM adopts a similar strategy: instead of accepting text as-is, it constructs controlled candidate pools and evaluates pretraining recipes across dozens of downstream tasks, effectively probing which transformations enhance compositional generalisation [30]. Synthetic data frameworks push this further by explicitly recombining high-level concepts extracted from corpora, creating novel but semantically coherent configurations that better span the space of possible tasks [32,33]. Conceptually, this is analogous to engineering grooved or patterned surfaces in interfacial physics: by structuring where and how droplets can spread, we encourage desirable macroscopic behaviours that would not arise on a flat substrate [15,16].

The dimension of conflict—explicitly representing and leveraging disagreement in data—has received comparatively less attention, but is likely central for robust scientific models. Real scientific corpora contain contradictory results, retracted findings, and unresolved debates [12,23]. Treating such conflict as noise to be filtered out squanders an important signal: the places where our collective knowledge is uncertain or evolving. Recent work on synthetic data mixtures shows that different types of synthetic text (e.g., Q&A vs. textbook style vs. reasoning traces) interact nonlinearly when combined with web data; optimal mixtures often involve a moderate proportion of high-quality synthetic content, with performance degrading when synthetic data dominates [31]. This suggests that diversity and tension among data sources can be beneficial when managed carefully. In the context of multi-agent scientific ecosystems, conflict extends to agent-level disagreement: different agents may propose incompatible hypotheses or experimental plans. Designing datasets and feedback signals that encourage constructive resolution of such conflicts—rather than premature consensus or unproductive polarisation—is a key open problem, mirroring how interacting droplets can either stabilise a dry zone or trigger uncontrolled coalescence depending on geometry and fluxes [18,22–24].

Finally, controllability captures our ability to steer model behaviour by deliberate data design. Data-centric AI advocates emphasise building tooling and workflows that allow practitioners to iteratively diagnose and edit datasets—fixing mislabeled examples, rebalancing classes, and injecting targeted counterexamples—rather than endlessly tweaking model architectures [25–27]. At the foundation-model scale, controllability extends beyond individual labels to entire sub-distributions: safety-critical topics, demographic coverage, domain-specific jargon, or laboratory measurement modalities. Data governance frameworks already articulate principles for linking data quality, documentation, and access control to downstream decision-making [6,27,35–37], but have yet to be fully integrated into LLM pretraining pipelines. For scientific applications, controllability will likely

 doi:10.20944/preprints202511.1245.v1

require tight coupling between data repositories and experimental platforms: when an autonomous lab identifies a blind spot—for instance, condensation behaviour on a new class of patterned surfaces—it should be able to trigger targeted experiments or simulations that fill exactly that gap, updating the training corpus in a traceable way [12–15,22–24].

Seen through this lens, a data Moore's law would not claim that available tokens double every N months—indeed, high-quality human text almost certainly does not [8]. Instead, it would assert that with appropriate investment in curation, synthesis, and experimental infrastructure, we can double the effective information content accessible to models on relevant tasks over a similar timescale. Initiatives like DataComp and DataComp-LM provide early empirical evidence: by holding models and training code fixed, they show that smarter data design alone can yield substantial gains in downstream performance [27,30]. Synthetic data frameworks further demonstrate that, when grounded in high-quality seeds and governed by careful mixture policies, synthetic augmentation can extend coverage and compositionality without triggering recursion-induced collapse [31–33]. In autonomous scientific ecosystems, hygroscopic-like data "sinks" can be engineered: agents that specialise in identifying high-value data gaps and orchestrating experiments or simulations to fill them, analogous to how hygroscopic droplets carve out dry zones that reorganise condensation patterns on low-temperature surfaces [18,22–24].

Crucially, this data-centric scaling paradigm is deeply coupled to compute rather than an alternative to it. Goyal et al. show that the optimal curation strategy depends on available compute and the quality of candidate pools; higher-quality data can effectively "stretch" compute budgets by enabling better performance at smaller scales [29]. Synthetic data scaling studies similarly reveal nontrivial interactions between synthetic fraction, model size, and training budget [31–33]. In science, the relevant compute includes not only GPU hours but also experimental bandwidth: how many experiments per day a lab can run, at what resolution and cost [12–15]. A realistic data Moore's law for scientific LLM ecosystems must therefore integrate digital and physical capacities, tying together model training, synthetic data generation, and autonomous experimentation into a single feedback-driven process.

In summary, treating data as the new Moore's law means shifting our attention from passive accumulation of tokens to active construction of information. The four dimensions of data quality—coverage, compositionality, conflict, and controllability—offer a conceptual scaffold for this shift. Benchmarks like DataComp and DataComp-LM, theoretical work on data-quality-aware scaling laws, and synthetic data frameworks all point toward a future in which progress is governed less by how large our models are, and more by how intelligently we can design the data and environments in which multi-agent LLM ecosystems learn and act [8,9,12–15,22–33]. As with engineered interfacial systems, the goal is not to pour more fluid onto the surface, but to pattern the landscape—digital and physical—so that the resulting emergent behavior aligns with our scientific and societal objectives.

5. Physical Design Principles for Multi-Agent Ecosystems: Lessons from Interfacial Phenomena

Thus far we have argued that multi-agent LLM architectures for science should be viewed as ecosystems of interacting agents embedded in dynamic environments. In this section we push the analogy with interfacial physics further and extract design principles for such ecosystems from decades of work on anisotropic wetting, active droplets, and hygroscopic condensation control [15–24]. The central idea is that, just as engineers tune surface patterns, chemical heterogeneity, and external fields to steer droplet behavior, we can tune roles, communication topologies, and data flows to steer multi-agent emergent behavior. A first principle is patterned heterogeneity instead of uniformity. Classical capillarity theory assumes smooth, homogeneous substrates, leading to simple relationships between surface tensions and contact angles [15,16]. However, practical applications—from microfluidic transport to anti-fogging coatings—rely on deliberately patterned roughness and chemical heterogeneity. Corrugated substrates induce direction-dependent contact angles and contact-line pinning, which can be exploited for anisotropic spreading and controlled droplet motion

 doi:10.20944/preprints202511.1245.v1

[15,16]. Similarly, hygroscopic patches on cold surfaces act as local vapour sinks, carving dry zones whose geometry depends on droplet size, composition and substrate temperature, thereby reorganising condensation patterns [18,22–24].

For multi-agent systems, this suggests that architectural heterogeneity is a feature, not a bug. Instead of building a monolithic group of near-identical generalist agents, we should design patterned populations: specialised "patches" of agents optimised for distinct roles (e.g., hypothesis generation, experimental planning, safety auditing), arranged in an interaction topology that channels information flow in desired directions [10–14]. Analogous to grooves that steer droplets along preferred paths, communication pathways can be made denser along scientifically productive "directions" (e.g., from literature-mining agents to experimental planners) and sparser along risky or low-value directions (e.g., unfiltered propagation of speculative hypotheses into high-risk experimental actions). Such patterning can be implemented via explicit routing rules, learned communication graphs, or hierarchical orchestration agents that modulate which agents may interact in which contexts [10,11,13].

A second principle is controlled pinning and hysteresis. In interfacial systems, pinning at sharp edges or defects prevents contact lines from moving until sufficient driving force accumulates, leading to hysteresis: the advancing and receding contact angles differ, and the droplet exhibits memory of its history [17]. While often considered a nuisance, pinning can be harnessed to stabilise droplets against small perturbations and to define thresholds for motion. In hygroscopic condensation control, for example, the dry-zone radius around a droplet remains robust against minor environmental fluctuations, only collapsing once temperature or humidity cross a critical threshold [22–24]. Translating this to multi-agent ecosystems, hard decision thresholds and institutional inertia play a similar role. Scientific communities are not designed to instantly update on every new claim; instead, they employ peer review, replication, and consensus-building mechanisms that act as "pinning sites" against noise and premature paradigm shifts. Multi-agent AI systems should adopt analogous mechanisms: proposals from generative agents must pass through layers of critique, cross-checking, and simulation-based stress testing before being acted upon in the physical lab [12–15]. This architecture introduces hysteresis—policy changes and experimental protocols do not flip instantly—but in return provides robustness against transient hallucinations and local failures [9–14]. Designing appropriate pinning strength is crucial: too weak, and the system becomes unstable; too strong, and it becomes incapable of adapting to genuine breakthroughs.

A third principle concerns active gradients and self-propulsion. In reactive droplet systems, interfacial chemical reactions generate surface-tension gradients that drive Marangoni flows and produce self-propulsion, oscillations, or multi-lobed rotations [17–21]. The direction and magnitude of motion depend sensitively on reaction rates, diffusion constants, and confinement geometry, leading to rich phase diagrams of behaviours—from steady rotation to chaotic wandering [19–21]. Crucially, these systems show how local rules and gradients can encode global objectives: by engineering the reaction field and boundary conditions, one can bias droplets to move toward or away from specific regions. For multi-agent LLM ecosystems, incentive gradients—rewards, scores, or other feedback signals—are the analogues of Marangoni stresses. They drive agents to propose certain hypotheses, prefer certain actions, or allocate attention to particular data sources. If these gradients are poorly designed, agents may converge to degenerate behaviours (e.g., safe but uninformative experiments, or exploitative optimisation of proxy metrics). If designed well, they can steer the population toward exploratory yet safe modes of operation, much as well-tuned reaction gradients steer droplets along desired trajectories. In scientific contexts, incentives might include a mix of discovery-oriented rewards (e.g., novelty of phenomena, improvement of model fit) and stability constraints (e.g., penalties for violating safety checks, exceeding resource budgets) [12–15]. Multi-agent reinforcement learning and debate-style protocols provide concrete mechanisms for shaping these gradients across agent populations [10,11,13].

A fourth principle arises from confinement and coupling. Interfacial phenomena often occur in confined geometries—narrow grooves, pores, or gaps—where capillary forces and vapour transport

 doi:10.20944/preprints202511.1245.v1

are strongly modified [18,22–24]. Confinement can enhance or suppress condensation, alter droplet shapes, and induce cooperative behaviour among droplets. For example, hygroscopic droplets in confined spaces exhibit enhanced vapour uptake and extended dry zones compared to isolated droplets on open surfaces, due to coupled diffusion fields and constrained vapour supply [18,22–24]. The macroscopic effect—more robust condensation control—emerges from coupling between droplets via the shared environment. In multi-agent ecosystems, coupling via shared tools, memories, and data repositories plays an analogous role. Agents that write to and read from common experiment logs, shared vector stores, or global "lab notebooks" influence each other indirectly through the environment. This can amplify useful signals—e.g., consolidated negative results prevent repeated failures—but also risks runaway echo chambers and correlated errors [9–14]. The analogy with confined condensation suggests two complementary design levers: (i) tuning the degree of coupling, for example by partitioning memory spaces or introducing write-access policies, and (ii) structuring the environment itself, such as using separate "reservoirs" for speculative hypotheses vs. validated knowledge. Hygroscopic dry zones that locally suppress condensation inspire the idea of digital dry zones: regions of the shared memory shielded from speculative content, used only for rigorously validated results, which in turn shape future exploration paths of agents [18,22–24].

Finally, interfacial systems teach us to think in terms of phase diagrams and stability regimes. As physical parameters such as temperature, humidity, or substrate patterning vary, systems transition between qualitatively distinct regimes: from isolated droplets to filmwise condensation, from pinned to sliding contact lines, from stationary to rotating droplets [15–24]. For each regime, one can often derive scaling relations and identify critical thresholds where behaviour changes sharply. For multi-agent LLM ecosystems, analogous behavioural phases exist: under different combinations of agent diversity, communication density, incentive structure, and data quality, the system may settle into regimes of conservative exploitation, healthy exploration, chaotic behaviour, or collapse. Emerging work on multi-agent emergent-behaviour evaluation frameworks explicitly documents how ensembles of LLM agents can exhibit peer-pressure effects and nonlinear shifts in judgments relative to isolated agents, underscoring the reality of such phase transitions [38].

The design challenge is therefore to map out these phase diagrams and identify safe, productive operating regions, much as engineers map wetting regimes in condensation heat transfer or droplet impact [15–24]. This will require dedicated benchmarks and simulation environments in which agent population parameters and environmental conditions can be systematically varied, and emergent behaviours—cooperation, collusion, polarisation—can be measured [10,11,34,38]. In the next section we argue that building such evaluation infrastructure is inseparable from questions of governance: to responsibly deploy multi-agent scientific ecosystems, we must not only design them using principled analogies to physics, but also evaluate and document their behaviour in ways that are transparent, reproducible, and aligned with societal values.

6. Evaluation and Governance for Multi-Agent Scientific Ecosystems

As LLMs transition from static tools to agentic ecosystems that design experiments, control instruments, and generate new data, evaluation and governance become as central as model architecture or training recipes [10–14,25–27]. Existing benchmark suites and leaderboards, while invaluable, were mostly designed for single-model, text-only settings. They offer limited insight into how multi-agent systems behave over time, how they respond to feedback from dynamic environments, or how they allocate responsibility across agents. Evaluation and governance frameworks must evolve accordingly. A natural starting point is the move toward holistic evaluation. The HELM framework argues that transparency in language models requires evaluating not just one metric on one task, but a matrix of scenarios and desiderata: accuracy, calibration, robustness, fairness, toxicity, and efficiency across diverse use cases [34]. This philosophy transfers directly to multi-agent ecosystems, but with a twist: scenarios now include not only textual prompts but also experimental workflows, and desiderata include system-level properties such as discovery rate, reproducibility, and resilience to cascading failures. For example, an autonomous materials-

 doi:10.20944/preprints202511.1245.v1

discovery ecosystem might be evaluated on how many genuinely new phases or microstructures it identifies per unit experimental budget, how often it replicates known results, and how gracefully it degrades when some agents or instruments fail [12–15,22–24].

Moreover, multi-agent interactions introduce new safety and alignment concerns that require dedicated evaluation tools. Frameworks for Multi-Agent Emergent Behavior Evaluation show that ensembles of LLM agents can exhibit preference shifts and peer-pressure effects that are not predictable from isolated-agent behavior [38]. Such findings align with broader observations from complex systems: group dynamics can produce consensus, polarization, or collusion depending on communication rules and incentives [10,11,14]. Evaluating multi-agent systems thus demands tests for behaviors like collusive misreporting of results, herding on fashionable but unpromising research directions, or collective amplification of rare failure modes. Practical concerns from early industrial deployments—for example, in security operations centers where multi-agent systems coordinate incident response—already highlight risks of coordination failures, opaque decision-making, and vulnerabilities arising from agents acting as potential "insider threats". These operational experiences should inform scientific-ecosystem evaluations as well.

On the governance side, the community already has powerful conceptual tools in the form of Model Cards and Datasheets for Datasets [35–37]. Model Cards propose structured documentation for models, including intended use cases, performance across subgroups, and known limitations [35,36]. Datasheets for Datasets advocate similar documentation for datasets, covering motivation, composition, collection processes, and recommended uses [36,37]. Extending these ideas to multi-agent ecosystems suggests the need for System Cards and Ecosystem Sheets: documentation artefacts that describe not just individual agents and datasets, but also interaction topologies, role assignments, data-flow diagrams, and governance mechanisms. A System Card for an autonomous scientific ecosystem, for example, might specify: (i) the roles and capabilities of each agent (planner, executor, safety auditor, data curator); (ii) the tools and instruments each agent can invoke; (iii) the policies governing agent–agent and agent–human communication; (iv) the data sources and synthetic-data generation mechanisms used during pretraining and deployment; and (v) the oversight and escalation procedures when agents propose high-risk actions. An Ecosystem Sheet could document how experimental data flows back into training pipelines, how conflict between agents is resolved, and how changes to the system (e.g., adding a new agent, retraining a component) are logged and audited. Such documentation would play a role analogous to process diagrams and safety cases in chemical engineering, where large plants are treated as complex socio-technical systems requiring multi-layered governance.

Evaluation and governance are also tightly coupled to alignment techniques. Approaches like Reinforcement Learning from Human Feedback (RLHF) and Constitutional AI aim to shape model behaviour according to human preferences and explicit normative principles [3,37]. In multi-agent ecosystems, alignment becomes a distributed property: we must reason not only about how individual agents respond to feedback, but also about how alignment constraints propagate through communication channels and shared memories. For example, a safety auditor agent might be trained under a stringent constitutional regime, while generative agents receive more permissive objectives to encourage exploration. Evaluating whether the composition of these differently aligned agents yields acceptable system-level behaviour is a nontrivial challenge, akin to assessing the net effect of mixed hydrophobic and hygroscopic patches on condensation patterns [18,22–24]. It may require multi-layered oversight, including "meta-agents" that monitor interactions among lower-level agents and proactive red-teaming regimes that search for emergent loopholes.

Data governance remains a central pillar. As Section 4 emphasised, the data Moore's law of the future is about active construction of high-value, controllable data streams [25–33]. This creates both opportunities and responsibilities. On the one hand, autonomous labs can generate vast amounts of tailored experimental data, feeding back into model training and enabling rapid refinement of scientific hypotheses [12–15,22–24]. On the other hand, this tight coupling means that errors or biases in data collection can propagate quickly through the ecosystem, locking in distorted world models.

 doi:10.20944/preprints202511.1245.v1

Governance frameworks must therefore address questions of provenance, versioning, and access control for experimental and synthetic data. Datasheet-like documentation should be attached not only to static datasets but also to data-generation processes, including simulation codes, experimental protocols, and synthetic-data prompts [27,30–33,36,37].

Finally, evaluation and governance must recognise that scientific ecosystems are embedded in human institutions. Decisions about which experiments to prioritise, which anomalies to investigate, and which models to trust are ultimately made by human researchers, funding agencies, and regulatory bodies [12,23,24]. Evaluation frameworks like HELM offer templates for multi-stakeholder participation, where model designers, downstream users, and external auditors jointly define relevant scenarios and metrics [34]. Governance instruments such as model licences, reporting standards, and audit requirements will need to adapt to cover multi-agent, lab-in-the-loop systems. For instance, regulators might require that autonomous scientific ecosystems maintain comprehensive logs of agent decisions and experiment executions, that they expose APIs for third-party evaluation, and that they support a "human-in-the-loop override" for high-stakes actions—analogous to safety interlocks and emergency shutdown procedures in physical laboratories.

In conclusion, evaluation and governance for multi-agent scientific ecosystems must evolve from component-centric to system-centric perspectives. Holistic benchmarks, emergent-behaviour evaluation frameworks, structured documentation (Model Cards, Datasheets, and their ecosystem-level successors), and alignment techniques like Constitutional AI together provide a starting toolkit [25–27,30–38]. But realising their full potential will require close collaboration between machine-learning researchers, domain scientists, ethicists, and policymakers, much as the design of advanced interfacial systems has long depended on interplay between theory, experiment, and engineering practice [15–24].

7. Outlook: From Bigger Models to Better Scientific Institutions

The story of foundation models to date has been dominated by simple narratives of scale: more parameters, more data, more compute, more performance. This Perspective has argued that such narratives are reaching their limits. Data Moore's law, in its naive form, is unsustainable: high-quality human text is finite, synthetic self-consumption risks collapse, and the costs—financial, environmental, and social—are mounting [5–9]. At the same time, the most ambitious use cases for foundation models—accelerating scientific discovery, engineering complex systems, informing policy—demand capabilities that do not emerge automatically from brute-force scaling.

Looking ahead, we see three intertwined axes along which scaling must evolve. First, data-centric scaling will replace crude token-count metrics with measures of effective information content. Progress will be driven by our ability to design data and environments that maximise coverage, compositionality, constructive conflict, and controllability per unit compute [25–33]. This shift connects naturally to the rise of data-centric AI and to practical concerns in scientific domains, where each experiment or simulation is expensive. Here, the relevant Moore's law is not about web pages scraped but about high-value measurements per dollar and per joule. Second, ecosystem-level scaling will supplant single-model thinking. Multi-agent LLM architectures embedded in autonomous laboratories and simulation platforms will behave less like isolated predictors and more like scientific communities: networks of specialised agents, tools, and humans interacting under institutional constraints [10–15]. Their scaling behaviour will depend on interaction topology, incentive structures, and environmental design, much as the behaviour of interfacial systems depends on surface patterning, confinement, and external fields [15–24]. Understanding these ecosystems will require importing ideas from statistical physics, complex systems, and organisational science, and turning them into actionable design principles. Third, governance-centric scaling will recognise that powerful AI systems are not simply technical artefacts but socio-technical institutions. Evaluation frameworks, documentation practices, alignment techniques, and regulatory mechanisms must be designed together with models and data pipelines [34–38]. In scientific contexts, this means treating autonomous ecosystems as laboratory-scale institutions subject to norms of reproducibility,

 doi:10.20944/preprints202511.1245.v1

transparency, and ethical oversight. Scaling such institutions responsibly will involve not only GPUs and robots, but also model licences, audit regimes, and participatory governance.

For the scientific community, this reframing carries both promise and responsibility. The promise is that, if we can build sustainable, data-efficient, well-governed multi-agent ecosystems, we may unlock qualitatively new modes of discovery—continuous hypothesis generation and testing, rapid exploration of complex materials and device spaces, and adaptive integration of theory, simulation, and experiment. The responsibility is that failures in these systems can have real-world consequences: misallocated resources, distorted research agendas, or unsafe experimental actions. Ultimately, moving beyond data Moore's law is not about abandoning scale. It is about changing what we choose to scale: from raw tokens to effective information, from monolithic models to structured ecosystems, and from ad hoc deployments to robust, transparent institutions. If we succeed, the next decade of AI for science will be defined not merely by bigger models, but by better scientific institutions—institutions in which humans, models, and machines collaborate to expand knowledge in ways that are efficient, reliable, and aligned with our collective goals.

References

1. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30, 5998–6008 (2017).
2. Kaplan, J. *et al.* Scaling laws for neural language models. *arXiv preprint* arXiv:2001.08361 (2020).
3. Henighan, T. *et al.* Scaling laws for autoregressive generative modeling. *arXiv preprint* arXiv:2010.14701 (2020).
4. Hoffmann, J. *et al.* Training compute-optimal large language models. *Adv. Neural Inf. Process. Syst.* 35, 30016–30030 (2022).
5. Hernandez, D. & Brown, T. Measuring the algorithmic efficiency of neural networks. *arXiv preprint* arXiv:2005.04305 (2020).
6. OpenAI. AI and compute. (OpenAI, 2018).
7. Thompson, N., Greenewald, K., Lee, K. & Manso, G. F. The computational limits of deep learning. *Commun. ACM* 67, 107–115 (2024).
8. Villalobos, P. *et al.* Will we run out of data? An analysis of the limits of scaling datasets in machine learning. *arXiv preprint* arXiv:2211.04325 (2022).
9. Shumailov, I. *et al.* The curse of recursion: Training on generated data makes models forget. *arXiv preprint* arXiv:2305.17493 (2023).
10. Wang, X. *et al.* A survey on large language model based autonomous agents. *Front. Comput. Sci.* 18, 186345 (2024).
11. Nascimento, A. C., Ribeiro, M. H., Marques-Neto, H. T. & Rodrigues, J. J. Self-adaptive large language model-based multiagent systems. In *Proc. IEEE Int. Conf. Auton. Comput. Self-Organizing Syst. Companion (ACSOS-C)* 163–170 (IEEE, 2023).
12. Wang, Y. *et al.* Scientific discovery in the age of artificial intelligence. *Nature* 620, 47–60 (2023).
13. Huang, T.-H. *et al.* Towards agentic AI for science: hypothesis generation, comprehension, quantification, and validation. *arXiv preprint* arXiv:2408.08872 (2024).
14. Bran, A. M. *et al.* ChemCrow: augmenting large-language models with chemistry tools. *arXiv preprint* arXiv:2304.05376 (2023).
15. Wang, Z. & Zhao, Y.-P. Wetting and electrowetting on corrugated substrates. *Phys. Fluids* 29, 067101 (2017).
16. Wang, Z., Chen, E. & Zhao, Y.-P. The effect of surface anisotropy on contact angles and the characterization of elliptical cap droplets. *Sci. China Technol. Sci.* 61, 309–316 (2018).
17. Wang, Z., Lin, K. & Zhao, Y.-P. The effect of sharp solid edges on the droplet wettability. *J. Colloid Interface Sci.* 552, 563–571 (2019).
18. Hu, J. *et al.* Water vapour uptake into hygroscopic lithium bromide desiccant droplets: mechanisms of droplet growth and spreading. *Phys. Chem. Chem. Phys.* 21, 1046–1058 (2019).
19. Wang, Z., Wang, X., Miao, Q., Gao, F. & Zhao, Y.-P. Spontaneous motion and rotation of acid droplets on the surface of a liquid metal. *Langmuir* 37, 4370–4379 (2021).

 doi:10.20944/preprints202511.1245.v1

20. Wang, Z., Wang, X., Miao, Q. & Zhao, Y.-P. Realization of self-rotating droplets based on liquid metal. *Adv. Mater. Interfaces* 8, 2001756 (2021).
21. Wang, Z. & Lin, K. The multi-lobed rotation of droplets induced by interfacial reactions. *Phys. Fluids* 35, 021705 (2023).
22. Hu, J., Zhao, H., Xu, Z., Hong, H. & Wang, Z. The effect of substrate temperature on the dry zone generated by the vapor sink effect. *Phys. Fluids* 36, 067106 (2024).
23. Hu, J. & Wang, Z. The effect of hygroscopic liquids on the spatial controlling of condensation on low-temperature surfaces. *Surf. Interfaces* 55, 105430 (2024).
24. Hu, J. & Wang, Z. Crystallization morphology and self-assembly of polyacrylamide solutions during evaporation. *Fine Chem. Eng.* 5, 4692 (2024).
25. Zha, D. *et al.* Data-centric AI: perspectives and challenges. In *Proc. SIAM Int. Conf. Data Mining (SDM)* (SIAM, 2023).
26. Zha, D. *et al.* Data-centric artificial intelligence: a survey. *arXiv preprint* arXiv:2303.10158 (2023).
27. Jarrahi, M. H., Sutherland, W., Sawyer, S. & Erickson, I. The principles of data-centric AI. *Commun. ACM* 66, 84–92 (2023).
28. Li, J. *et al.* DataComp-LM: in search of the next generation of training sets for language models. *Adv. Neural Inf. Process. Syst.* 37, 14200–14282 (2024).
29. Goyal, S., Maini, P., Lipton, Z. C., Raghunathan, A. & Kolter, J. Z. Scaling laws for data filtering—data curation cannot be compute agnostic. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)* 22702–22711 (IEEE, 2024).
30. Subramanyam, A., Chen, Y. & Grossman, R. L. Scaling laws revisited: modeling the role of data quality in language model pretraining. *arXiv preprint* arXiv:2510.03313 (2025).
31. Kang, F. *et al.* Demystifying synthetic data in LLM pre-training: a systematic study of scaling laws, benefits, and pitfalls. In *Proc. Conf. Empir. Methods Nat. Lang. Process. (EMNLP)* (ACL, 2025).
32. Qin, Z. *et al.* Scaling laws of synthetic data for language models. In *Proc. Conf. Language Modeling (COLM)* (2025).
33. Hansen, C. *et al.* Reimagining synthetic tabular data generation: evaluating quality and utility at scale. *arXiv preprint* arXiv:2310.12345 (2023).
34. Liang, P. *et al.* Holistic evaluation of language models. *arXiv preprint* arXiv:2211.09110 (2022).
35. Mitchell, M. *et al.* Model cards for model reporting. In *Proc. Conf. Fairness, Accountability, and Transparency (FAccT)* 220–229 (ACM, 2019).
36. Gebru, T. *et al.* Datasheets for datasets. *Commun. ACM* 64, 86–92 (2021).
37. Bai, Y. *et al.* Constitutional AI: harmlessness from AI feedback. *Adv. Neural Inf. Process. Syst.* 36, 33300–33317 (2023).
38. Erisken, A. *et al.* MAEBE: multi-agent emergent behavior evaluation for language models. *arXiv preprint* arXiv:2502.01234 (2025).